Journal of Organizational and End User Computing
Volume 34 • Issue 2

Text Similarity Measurement Method and Application of Online Medical Community Based on Density Peak Clustering

Mingyang Li, School of Management, Jilin University, China
Xinhua Bi, School of Management, Jilin University, China*
Limin Wang, Information Science School, Guangdong University of Finance and Economics, China
Xuming Han, College of Information Science and Technology, Jinan University, China
Lin Wang, School of Management, Jilin University, China
Wei Zhou, School of Computer Science and Technology, Changchun University of Science and Technology, China

ABSTRACT

Text similarity measurement is a link between basic research such as text modeling and upper-level application research of text potential information. In order to improve the accuracy of text similarity measurement, this paper proposes a semantic similarity calculation method integrating word2vec model and TF-IDF and applies it to the density peak clustering of Chinese text data consulted by patients in the online medical community. Experimental results show that the proposed similarity measurement method is superior to the traditional method. Furthermore, the study is among the first to apply the density peak clustering algorithm to the online medical community, which offers a reference for how to find user demands from medical text data in the big data environment.

KEYWORDS

Clustering Analysis, Density Peak Clustering, Hot Topic, Online Medical Community, Semantic Similarity Calculation, Text Mining, Text Similarity, User Demands

INTRODUCTION

In the medical and health industry, problems such as difficulty and high cost of getting medical service and unequal distribution of medical resources are widespread, while patients' demand for medical and health services keeps increasing (Shen et al., 2019). With the continuous development of social media and the popularization of mobile smart terminals, online medical communities have entered people's lives, and more and more patients have begun to conduct medical consultations in online medical communities (Hajli, 2014). Online medical communities refer to the use of Internet information technology to present the medical ecosystem including patients, doctors and hospitals in the form of community network, providing a medical information exchange platform for patients and doctors, and providing patients with a series of services such as seeking medical treatment, evaluating medical treatment, registration, health consultation and so on (Julian, 2018). There are a large number of different types of participants in the online medical community, and the focus of people's attention is different for different types of diseases (Liu et al., 2018; Sung et al., 2020). Taking the relevant

DOI: 10.4018/JOEUC.302893 *Corresponding Author

information they consulted in the online medical community as the research object, through the analysis of these text contents, we can find the hot issues that community members are concerned about, so as to provide theoretical and practical basis for the research direction of medical researchers and better meet the medical needs of patients (Lu et al., 2013).

With the advent of the era of big data and the gradual maturity of artificial intelligence, the analysis and utilization of large-scale Internet data and the extraction of valuable information have become a research hotspot in the academic field (Tandel et al., 2019). Text mining technology can help people discover knowledge and rules from text data, thereby generating value (Yu et al., 2015; Zhang et al., 2021). Commonly used text mining techniques include text structure analysis, text summarization, text classification, text clustering, text association analysis, etc (Ngai et al., 2016). Text similarity, as a method to measure the differences and commonalities between texts, is the core of these technical tasks (Zaher et al., 2020). Ensuring the rationality of similarity measurement is particularly critical in text mining, and even plays a decisive role in the performance of the algorithm (Lin et al., 2014). According to different research fields, the specific content of text similarity calculation is different, and the nature of text content characterized by text similarity is also different. Text similarity computation needs to be studied according to the specific application field, so as to improve the accuracy of text similarity calculation and better serve the corresponding context (Wang & Dong, 2020). At present, there are more and more methods to calculate the text similarity, but there are still some gaps and limitations in the text similarity research on the text characteristics of online medical community.

In view of the foregoing, according to the characteristics of text data in online medical community, this paper proposes a new method to calculate the semantic similarity of Chinese text, and applies it to text clustering to verify the effectiveness of the proposed method and discover hot topics of patient consultation in online medical community. Text clustering is based on clustering algorithm. Due to the complexity and diversity of data, many domestic and foreign scholars have proposed a large number of clustering algorithms with outstanding performance based on special problems in different fields, but they all have their own inconsistent disadvantages (Abbas, 2007). Rodriguez & Laio (2014) proposed a concise clustering method, density peak (DPeak) clustering algorithm, in the journal Science. The algorithm has the advantage of effectively processing complex-shaped data sets and can transform the clustering of high-dimensional data into clustering of data in a two-dimensional space, thereby realizing its efficient operation. Therefore, aiming at the high-dimensional characteristics of text data, this paper uses DPeak algorithm for text clustering. Specifically, this paper conducts centralized preprocessing on the corpus, abstracts it into computable elements, adopts the proposed method to calculate the text similarity, and converts the text similarity into the distance and density of data sample, so as to realize text clustering based on DPeak algorithm.

The rest of the paper is arranged as follows: Background section elaborately studies the current text similarity measurement methods, applications and improvements. Theory and Methodology section describes the text similarity measurement method proposed in this paper in detail, and introduces DPeak clustering algorithm and its effectiveness evaluation index. Experiment and Result Analysis section verifies the effectiveness of the proposed method from multiple angles. Conclusion section concludes this paper with future enhancements.

BACKGROUND

Text similarity, as the name suggests, is to measure the similarity of a group of text content. The calculation of text similarity is the basis of text mining, and the research on it includes research on text representation models and research on text similarity calculation methods. The purpose of the text representation model is to find a better representation method to present the text content; the purpose of the text similarity calculation method is to find a more accurate way to measure the similarity between text content (Li, et al., 2017). The calculation methods of text similarity range from the initial similarity calculation based on text surface information, to the similarity calculation

based on text vector distance in vector space, to the similarity calculation based on topic modeling based on text content, and the similarity calculation based on word vector representation in text, etc (Prakoso et al., 2021).

At present, text similarity calculation methods are mainly divided into four categories: string-based method, knowledge-based method, corpus-based method and mixed method. The string-based method is also called the "literal similarity method", which measures the similarity by the degree of string co-occurrence and repetition. Typical methods include Levenshtein distance (LD), longest common substring (LCS), Jaccard coefficient and so on (Alqahtani et al., 2021). LD is usually used in the field of fast fuzzy matching of sentences to represent the minimum number of editing operations required to convert one text into the other. Edit operations include add, delete and change, which can be either at character level or word level. LCS method uses the co-occurrence of two texts and the length of the longest substring to represent the similarity between texts. Compared with LD, LCS method not only considers the composition of text characters, but also adds character order. Jaccard coefficient measures the similarity between two texts by the ratio of the number of the same words in two texts to the number of all non-repeated words. It is based on the idea of set and only pays attention to the number of common elements in two sets, but does not pay attention to the difference between set elements. The string-based method is a direct text comparison at the literal level, which is simple in principle and convenient in implementation, and is the basis of many current algorithms. For example, Taeho (2018) proposed an improved KNN (k nearest neighbor) algorithm, which takes the string vector as its input data and applies it to text summarization. Text summarization can be regarded as a binary classification, in which each paragraph is classified as abstract or non-abstract. The experiment shows that the text similarity algorithm based on string vector can be applied to text classification and get good experimental results. In the case of spelling errors, the efficiency of extracting information from free text data is low. Many existing methods use string methods to search for valid words in the text. Tissot & Dobson (2019) used the method of combining character strings and language-related phonetic similarity to identify the misspelled names of drugs in a group of medical records written in Portuguese, and obtained good experimental results. Because the string-based method does not consider the semantic information of text and the relationship between words, the calculation effect is limited to a certain extent. In order to solve this problem, scholars began to study semantic similarity methods, resulting in the emergence of knowledge-based methods, corpus-based methods and hybrid methods.

The semantic similarity calculation method based on knowledge uses the knowledge base with standard organization system to calculate the text similarity. According to the types of knowledge base, it can be divided into two categories, one is based on ontology, the other is based on network knowledge. The former uses structured semantic dictionaries for computing, and its basic idea is to use the concept information contained in these semantic dictionaries and the hierarchical relationship between concepts to calculate the semantic text similarity. The method based on network knowledge mainly uses the large-scale knowledge base resources. Compared with the ontology-based method, the network knowledge-based method has a wider coverage, a more comprehensive knowledge description, and a faster update of information content (Jorge, 2014). The most widely used network knowledge is Wikipedia and Baidu Encyclopedia.

The corpus-based semantic similarity calculation method is to use the information obtained from corpus to calculate text similarity. Corpus-based methods can be divided into two categories: based on distributional representation and based on search engines. The method based on distributional representation mainly uses corpus to transform text into vector representation with semantic information, and then determines semantic or distributed similarity based on vector similarity. This method is a major category of corpus-based methods, and it is also the most mainstream research direction at present. Zhao et al. (2021) used the geometric structure of sentences to improve the rendering effect of biomedical texts, and introduced manifold learning into the expression of biomedical sentences, making the similarity of sentences in Euclidean space closer to sentence

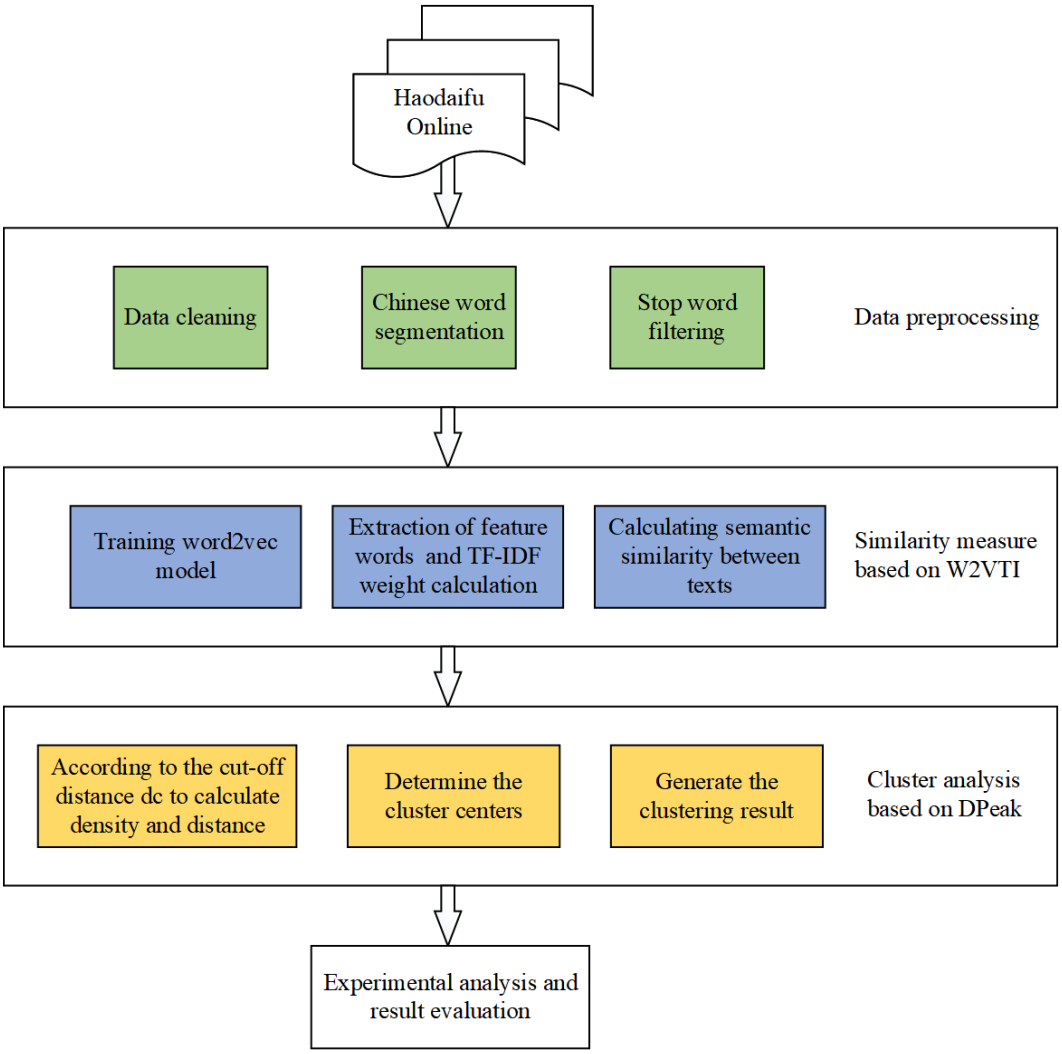
semantics, and then mining geometric structure information from sentence expression. In order to capture word-level semantics, Shajalal & Aono (2019) used distributed representation of words in two different languages, and proposed a method to calculate semantic similarity between sentences using bilingual word-level semantics. Experimental results show that this method can effectively measure the similarity of semantic text and is superior to some known related methods. Sogancioglu et al. (2017) proposed several sentence level semantic similarity calculation methods in biomedical field, including string similarity measurement and sentence similarity measurement based on distributed vector representation, and established a model based on supervised regression, which effectively combined different similarity calculation indexes. Experiments show that the supervised semantic sentence similarity calculation method achieves the best performance among many methods. The web search engines-based method regards the entire web as a dynamic corpus, and the search engine provides an efficient interface for massive amounts of information. It uses page counts-based, text snippets-based, or a combination of the two methods to calculate semantic similarity. Biondi et al. (2016) proposed an emotion recognition method based on network similarity, which extracted basic emotions from network sentences containing emotional content and performed a network-based quantitative evaluation of the semantic proximity between each word of the analyzed sentence and each emotion of the mental model.

Throughout the current researches on text similarity calculation methods, most researches focus on the method itself, either to propose a high-performance computing method, or to improve the original method to obtain higher performance. Few researches start from the characteristics of the text in the application environment, and even fewer researches put forward a text similarity calculation method suitable for the text mining of online medical community according to the particularity of the text. Therefore, based on the text characteristics of online medical community, this paper proposes a new text similarity measurement method and applies it to the text mining of online medical community. It provides a reference for the similarity measurement of such texts and how to discover user needs from medical text data in a big data environment.

THEORY AND METHODOLOGY

Text clustering is to divide text objects into several clusters, whose goal is to make the text within the cluster as similar as possible and the text between the clusters as different as possible. Judging the degree of similarity and dissimilarity between text objects needs a quantitative scale, which is similarity calculation (Zhou et al., 2019). The research process is shown in Figure 1. Firstly, the text data of patient consultation on the platform of Haodaifu Online is obtained through the crawler. Then, the text data is preprocessed, that is, data cleaning, word segmentation, stop word filtering. On this basis, a text similarity measurement method based on word2vec and TF-IDF including word weights and semantic relations (W2VTI) is proposed. The proposed method is applied to the density peak clustering algorithm, and cluster analysis is performed on the patient consultation text data of the Haodaifu Online platform. Next, the text data is divided into different categories according to the similarity results. In the experimental verification, the accuracy F-Measure is used to evaluate the effectiveness of the method.

Figure 1. Research flow of text similarity measure and clustering analysis



Similarity Measure Based on W2VTI

The text similarity measurement method proposed in this paper is based on word2vec and TF-IDF. Therefore, in the following subsections, word2vec and TF-IDF are systematically described, and then the proposed method is elaborated in detail.

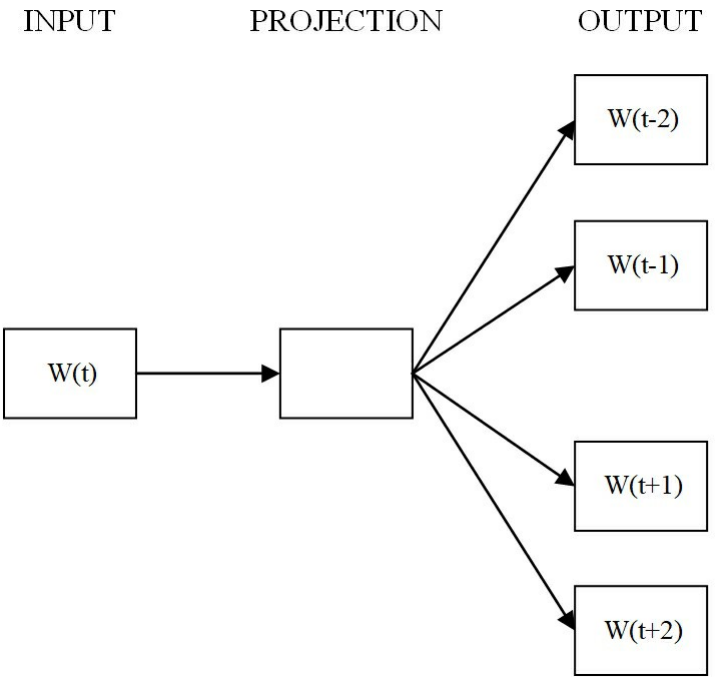
Word2vec

Text is an important way to carry semantic information, the quality of text representation directly affects the performance of the whole natural language processing system. Traditional text representation uses vector space model to express semantic information, which does not consider the order of feature words and contextual semantic understanding, resulting in high-dimensional sparsity and low computational efficiency. Word2vec is an implementation of the model proposed by Mikolov et al. (2013). It is an efficient natural language processing algorithm that represents words as low-dimensional real-valued vectors, which is convenient for measuring and mining the relationship between words (Khatua et al.,

2019). The algorithm learns semantic knowledge to obtain word vectors from a large number of text corpus in an unsupervised manner, words with similar meanings will be mapped to close positions in the vector space.

The word vector obtained by word2vec training can simplify the text into vector in k-dimensional space. Word2vec includes two models: Continuous Bag-of-Words (CBOW) model and Skip-Gram model. Among them, the CBOW model uses *m* words before and after the word *W(t)* to obtain the occurrence probability of the current word; Skip-Gram, on the contrary, uses the current word to predict the occurrence probability of *m* words before and after it (in Figure 2, $m = 2$). In comparison, the CBOM model is faster, but the Skip-Gram model has more advantages in processing rare words. Due to the particularity of online medical field, user comments often contain uncommon professional terms. Therefore, this paper selects the Skip-Gram model to train word vectors, and its model architecture is shown in Figure 2.

Figure 2. Architecture diagram of Skip-Gram model



For the whole corpus, the objective function of Skip-Gram model is shown in Eqs. (1) and (2).

$$\arg\max_{\theta} \sum_{w \in text} \sum_{c \in context(w)} \log p\left(c \mid w;\theta\right) \tag{1}$$

$$p\left(c \mid w;\theta\right) = \frac{e^{u_c v_w}}{\sum_{c'} e^{u_{c'} v_w}} \tag{2}$$

Where *w* is the target word, *c* is the context of the target word. θ is the inner product of the word vector, including *u* and *v*. *u* is the word vector when the word is used as the context, and *v* is the word vector when the word is used as the center word. That is $\theta = u_c \cdot v_w$. $c' \in thesaurus$.

After getting the training result of the word vectors, the distance between the word vectors can be calculated, and then the similarity between the two words can be obtained. The cosine similarity is a common and effective method to express the distance between word vectors. The calculation method of cosine similarity between vector $\bar{X} = \left(x_1, x_2, \cdots, x_n\right)$ and vector $\bar{Y} = \left(y_1, y_2, \cdots, y_n\right)$ is shown in Eq. (3).

$$S\left(\bar{X} \quad \bar{Y}\right) = \cos\left(\bar{X} \quad \bar{Y}\right) = \frac{\sum_{k=1}^{n} x_k \times y_k}{\sqrt{\sum_{k=1}^{n} x_k^{\,2}} \times \sqrt{\sum_{k=1}^{n} y_k^{\,2}}} \tag{3}$$

TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) algorithm is a statistical calculation method, which is often used to evaluate the importance of a word to a document in a document set. TF-IDF algorithm consists of two parts: TF algorithm and IDF algorithm (Aizawa, 2003).

TF algorithm is to count the frequency of a word appearing in a document. The basic idea is that the more times a word appears in the document, the stronger its ability to express the document will be. The calculation method is shown in Eq. (4). Among them, n_{ij} represents the occurrence frequency of the word *i* in document *j*, and the denominator is the sum of the occurrence times of each word in document *j*.

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{4}$$

IDF algorithm is to count how many documents a word appears in the document set. The basic idea is that if a word appears in fewer documents, its ability to distinguish documents will be stronger. The calculation method is shown in Eq. (5). Among them, |*D*| is the total number of documents in the document set, and |*Di*| is the number of documents in which the word *i* appears in the document set.

$$IDF_i = \log\left(\frac{\left|D\right|}{1 + \left|D_i\right|}\right) \tag{5}$$

TF algorithm and IDF algorithm can be used separately, but in the process of using, both of these algorithms have some weakness. TF only measures the occurrence frequency of words, but does not consider the ability of words to distinguish documents; IDF, on the contrary, emphasizes the ability to distinguish words, but ignores that if a word can appear frequently in a document, it means that the word can well express the characteristics of the document. TF-IDF algorithm is a comprehensive use of TF and IDF, the calculation method is shown in Eq. (6).

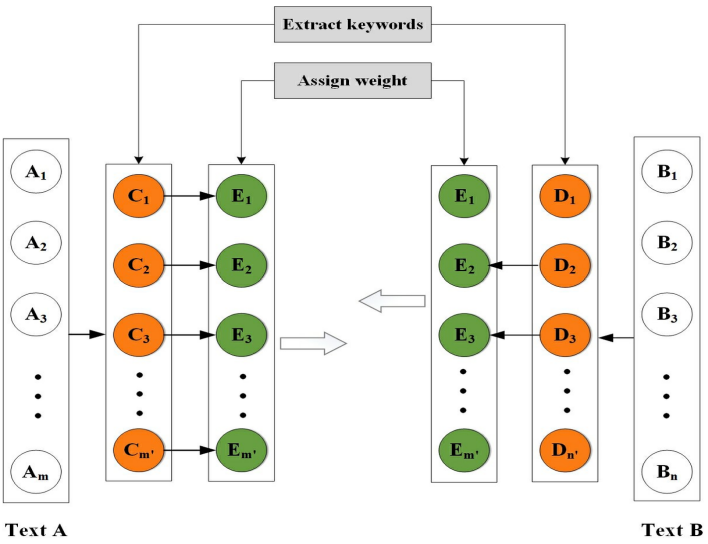
$$TF\_IDF_{(ij)} = TF_{ij} \times IDF_i = \frac{n_{ij}}{\sum_k n_{kj}} \times \log\left(\frac{|D|}{1+|D_i|}\right) \tag{6}$$

The proposed semantic similarity of Chinese text

A Chinese text can be regarded as a set of orderly arranged words, and a Chinese document is composed of several words. It cannot be said that a certain word can completely represent a document. It can only be said that the word can represent the document to a certain degree. The evaluation criterion of this degree is the weight. The greater the weight of the word, the more the word can represent this document, and vice versa. Of course, when the weight is less than a certain value (i.e. a given threshold), it can be considered that the word does not represent the document and should be removed.

The calculation of text similarity is influenced by a lot of information. The more complete the information is, the more accurately the potential connection between words and words, text and text can be calculated. The original TF-IDF method only focuses on the occurrence frequency of words, ignores the internal semantic relationship between words in the document, and has poor effect in dealing with short text. The semantic information of words can be well represented by word2vec model, but the weight information of words cannot be reflected in word2vec model. The mean word2vec model is to average the word2vec of all words in the text to obtain the text vector, without considering the weight information of the keywords in the text. In order to calculate the similarity between texts more accurately, this paper proposes a text similarity calculation method based on word2vec and TF-IDF(W2VTI). A schematic diagram of the text similarity calculation based on W2VTI is shown in Figure 3.

Figure 3. W2VTI model



Step 1: Calculate the *TF_IDF* of each word in text A and B according to Eq. (6), treat it as the corresponding weight of each word, and extract the words whose *TF_IDF* is greater than the threshold in text A and B. That is, after text A extracts keywords, it is composed of *n* words whose *TF_IDF*

value is greater than the threshold, $A\left(A_1, A_2, \cdots, A_n\right)$, $W_A = \left(w_{A1}, w_{A2}, \cdots, w_{An}\right)$, the weight of each word is w_{Ai}, $i = 1, 2, \cdots, n$; after text B extracts keywords, it is composed of *m* words whose *TF_IDF* value is greater than the threshold, $B\left(B_1, B_2, \cdots, B_m\right)$, $W_B = \left(w_{B1}, w_{B2}, \cdots, w_{Bm}\right)$, the weight of each word is w_{Bj}, $j = 1, 2, \cdots, m$.

For keywords $A\left(A_1, A_2, \cdots, A_n\right)$ in text A and keywords $B\left(B_1, B_2, \cdots, B_m\right)$ in text B, according to Eq. (3), calculate the similarity between the keywords and generate a similarity matrix as shown in Eq. (7).

$$S = \begin{bmatrix} S\left(A_1, B_1\right) & S\left(A_1, B_2\right) & \cdots & S\left(A_1, B_m\right) \\ S\left(A_2, B_1\right) & S\left(A_2, B_2\right) & \cdots & S\left(A_2, B_m\right) \\ \vdots & \vdots & \ddots & \vdots \\ S\left(A_n, B_1\right) & S\left(A_n, B_2\right) & \cdots & S\left(A_n, B_m\right) \end{bmatrix}_{n \times m} \tag{7}$$

Step 3: Find out the keywords with the highest similarity in text B for each keyword in text A and the keywords with the highest similarity in text A for each keyword in text B, respectively. $\underset{j}{\operatorname{argmax}} S\left(A_i, B_j\right)$, $i = 1, 2, \cdots, n$, get the maximum value $u_i = \max\ S\left(A_i, B_j\right)$ of each row in *S* matrix and the corresponding word B_j; $\underset{i}{argmax}\, S\left(A_i, B_j\right)$, $j = 1, 2, \cdots, m$, get the maximum value $u_j = \max\ S\left(A_i, B_j\right)$ of each column in *S* matrix and the corresponding word A_i.

On the basis of the first three steps, a weighted method is used to calculate the similarity between text A and B. The calculation method is shown in Eq. (8).

$$S_{A,B} = \frac{1}{m+n} \times \left(\sum\nolimits_{i=1}^{n} w_{Ai} \times w_{Bj} \times u_i + \sum\nolimits_{j=1}^{m} w_{Ai} \times w_{Bj} \times u_j \right) \tag{8}$$

Cluster Analysis Based on DPeak

In order to verify the effectiveness of the proposed method, this paper applies it to cluster analysis to verify its effectiveness from multiple angles. In the following subsections, the DPeak clustering algorithm and the evaluation method of clustering effectiveness are introduced respectively.

Density Peak Clustering Algorithm

Density peak (DPeak) clustering algorithm is a clustering method which can quickly find the density distribution of data points. The DPeak algorithm is derived from two simple facts in the clustering process: (1) the cluster center point is highly dense, that is, it is surrounded by neighbors whose density does not exceed its; (2) the distance between cluster center points is relatively far, that is, the "distance" between a cluster center point and data points with greater density than it is large. This is the key for candidate points to become cluster centers (Ni et al., 2019). The steps of DPeak algorithm are as follows: first, determine the cluster centers according to the characteristics of the cluster centers; after that, divide other data points into the clusters of data points that are closest to them and have a higher density than themselves.

In the data set, the data set to be clustered is $s = \left\{x1, x2, \cdots, xn\right\}$, $dij = dist\left(xi, xj\right)$ represents a certain distance between data point x_i and x_j. DPeak algorithm introduces the local density ρ_i of the sample *i* and the distance δ_i from the sample *i* to the nearest sample *j* with higher density than sample *i*. Their calculation methods are shown in Eqs. (9), (10) and (11).

$$\rho_i = \sum_{j \neq i} \chi\left(dij - dc\right) \tag{9}$$

Where the function $\chi\left(x\right)$ is:

$$\chi\left(x\right) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases} \tag{10}$$

The parameter d_c is the cut-off distance, which needs to be set manually. It can be seen from the calculation equation that in the DPeak algorithm, the local density of the sample is affected by the cut-off distance. Moreover, the authors of DPeak algorithm point out that when the number of samples in the data set is large, the clustering result of DPeak algorithm is less affected by the cut-off distance, otherwise it is larger.

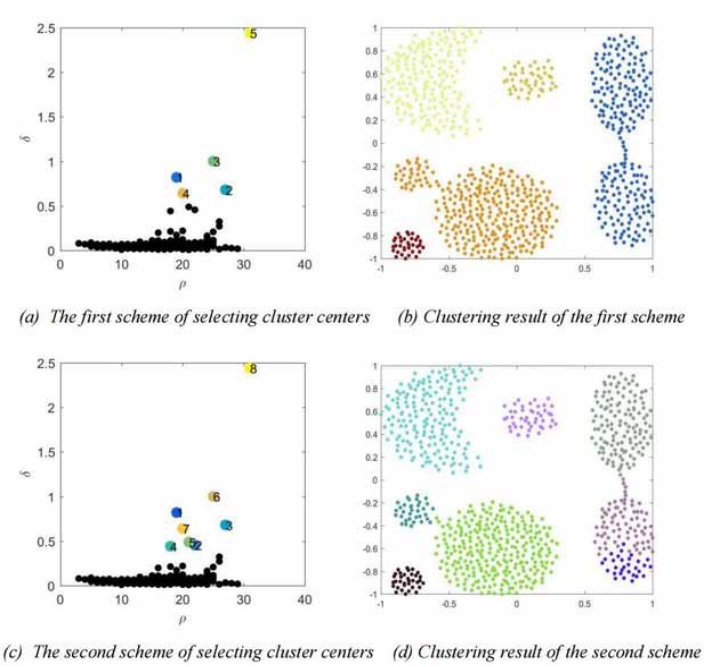
$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{11}$$

For the data point with the highest local density ρ_i in the data set, the distance calculation method needs to be processed separately. For this point, the calculation method of the distance δ_i is shown in Eq. (12).

$$\delta_i = \max_j d_{ij} \tag{12}$$

DPeak algorithm points out that according to the above calculation equations, a decision graph corresponding to local density ρ_i and distance δ_i can be drawn. On the basis of the characteristics of cluster center points, the sample points with large local density ρ_i and distance δ_i are circled as the cluster centers in the decision graph. However, when selecting the cluster centers, qualitative analysis rather than quantitative analysis is used to determine the cluster centers, which contains subjective factors. Different people may get different results in the same decision graph. Some people may think that these are the cluster centers, while others think that those are the cluster centers. Taking the UCI data set Twomoon as an example, choosing different cluster center points will have completely different clustering results, as shown in Figure 4.

Figure 4. Example of different clustering results with different clustering centers in the decision graph



(a) The first scheme of selecting cluster centers (b) Clustering result of the first scheme

(c) The second scheme of selecting cluster centers (d) Clustering result of the second scheme

(a) The first scheme of selecting cluster centers (b) Clustering result of the first scheme
(c) The second scheme of selecting cluster centers (d) Clustering result of the second scheme

Evaluation With F-Measure

F-Measure, also known as F-Score, is a comprehensive evaluation index commonly used in the field of information retrieval. It comprehensively considers recall and precision for cluster evaluation (Hripcsak & Rothschild, 2005). In theory, to achieve the optimal clustering effect, the precision and recall of the algorithm should be improved as much as possible. However, in practice, because these two indexes are mutually exclusive to some extent (increasing one index value will often reduce the other index value), it is impossible for any clustering algorithm to give consideration to both of them. Therefore, in order to comprehensively evaluate the accuracy of clustering analysis, this study uses F-Measure index to evaluate the clustering quality.

For data set $D = \left\{x_1, x_2, ..., x_n\right\}$, assuming that the cluster obtained by the clustering algorithm is divided into $C = \left\{c_1, c_2, ..., c_k\right\}$, and the correct cluster that is known in advance is divided into $C^* = \left\{c_1^*, c_2^*, ..., c_s^*\right\}$, then for a certain category c_i in the cluster, the precision and recall are defined as Eqs. (13) and (14) respectively.

$$P\left(i,j\right) = precision\left(i,j\right) = E_i / \left(E_i + F_i\right) \tag{13}$$

$$R\left(i,j\right) = recall\left(i,j\right) = E_i / \left(E_i + H_i\right) \tag{14}$$

Where, E_i represents the number of samples correctly classified into class c_i* in the actual clustering class c_j; F_i represents the number of samples that do not belong to class c_i* but are wrongly classified to class c_i* in the actual clustering class c_j; H_i represents the number of samples that should be classified into class c_i* but are not wrongly classified into class c_i* in the actual clustering class c_j.

Based on precision and recall, F-Measure of c_j relative to c_i* can be calculated by Eq. (15). It can be seen that F-Measure is the weighted harmonic average of precision and recall. In this study, the parameter α is set to 1, which is the most common setting value of α. At this time, Eq. (16) can be obtained.

$$F\_Measure\left(i,j\right)=\frac{\left(\alpha^2+1\right)P\left(i,j\right)\times R\left(i,j\right)}{\alpha^2\left(P\left(i,j\right)+R\left(i,j\right)\right)} \tag{15}$$

$$F\_Measure\left(i,j\right)=\frac{2\times P\left(i,j\right)\times R\left(i,j\right)}{P\left(i,j\right)+R\left(i,j\right)} \tag{16}$$

The F-Measure of global clustering is defined as Eq. (17).

$$F\_Measure=\sum\nolimits_{i=1}^{s}\frac{\left|c_i^*\right|}{n}\underset{j=1,\cdots,k}{max}F\_Measure\left(i,j\right) \tag{17}$$

EXPERIMENT AND RESULT ANALYSIS

To prove the validity of the proposed method of text similarity measure and its effect in clustering, this research obtains text data from online medical community and carries out a series of experiments from data acquisition to result analysis.

Experimental Corpus

The corpus used in this experiment comes from Haodaifu Online, one of the most famous online medical communities in China. Founded in 2006, Haodaifu Online currently includes nearly 10 000 regular hospitals and up to 610 000 doctors, with abundant medical resources and services. Its main services include: online diagnosis and treatment, electronic prescription, door-to-door drug delivery, remote consultation, expert surgery, etc. It is one of the more mature doctor-patient interaction platforms with typical characteristics. Patients with endocrine diseases have complicated conditions, unclear mechanisms, and usually have a long time of illness, so they are easy to browse and release information in the medical community. Therefore, this study makes use of a self-compiled program with the help of Jupyter, to crawl the online consultation data in department of endocrinology, and mines the text information of Haodaifu Online patient consultation in the department. The crawling date is April 4 and 5, 2020, and a total of 18000 pieces of data are captured. After cleaning up, a total of 14671 pieces of valid data are obtained.

Data Preprocessing

When clustering Chinese texts, the problem that needs to be faced is the preprocessing of Chinese texts, including text noise processing, word segmentation, removing stop words (including punctuations,

Arabic numerals and other meaningless words), etc. Text noise processing refers to the removal of noise data such as texts that interfere with text processing, links, emoticons, URLs, special characters, and texts less than 5 words, in the collected text data set. The removal of these useless noise data is conducive to the smooth running of subsequent text processing. In the crawled text data of patient consultation, the form such as "2018-03-05填写" (filled in on March 5, 2018) refers to the time when the patient fills in this content; "病历资料仅医生和患者本人可见" (medical records are only visible to the doctor and the patient) means that the patient's medical records are not open to the public and cannot be viewed. These contents have no meaning for text analysis, and will interfere with subsequent word vector training and text similarity calculation. Therefore, it is necessary to remove the texts in advance. The principle of regular expression is string matching, which can be used to extract the substrings that meet a certain condition in a string, or to replace some substrings in a string. This paper adopts the method of regular expression to identify and filter useless information in patient consultation texts. Table 1 gives some regular expressions and examples of the fields to be filtered.

Table 1. Fields to be filtered and regular expressions

regular expression	format of the field to be filtered	example
\d{4}-\d{2}-\d{2}填写 (\d{4}-\d{2}-\d{2}filled in)	年-月-日填写 (year-month-day filled in)	2018-03-05填写 (2018-03-05 filled in)
已开检查|暂时没开|谢谢医生 (diagnostic test has been prescribed|hasn't been prescribed temporarily|Thank you, doctor)	无意义的词组 (meaningless phrase)	仅医生和患者本人可见 (only visible to the doctor and the patient)

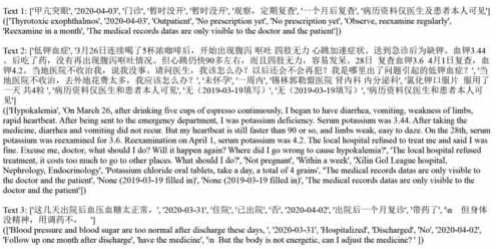
The preprocessing process of Chinese text has some special features compared with that of English text. The word segmentation technique of Chinese text preprocessing is much more complicated than that of English, the main difficulty is that the structure of Chinese is very different from the languages of the Indo-European system, and it is difficult to define the boundaries of words. For example, in English, the word itself is the expression of "word", and an English article is represented by "words" plus separators (spaces). In Chinese, word is based on characters, but the semantic expression of an article is still divided by words. Therefore, in the preprocessing of Chinese text, word segmentation is needed to convert sentences into word representations. In recent years, with the increasing maturity of natural language processing technology, there are more and more word segmentation tools of open source implementation, such as Ansj, Jieba, Pangu segmentation, etc. At present, Jieba has become the most popular Chinese word segmentation tool in China, and supports several kinds of modes. In view of the characteristics of active community and rich functions of Jieba, this paper adopts the precise pattern of Jieba word segmentation tool, which is suitable for text analysis and aims at the most accurate division of sentence units in text.

Stop word filtering is to filter out the high-frequency words that do not work on the content of text information. Its purpose is to reduce text redundancy and improve the accuracy of text clustering. Aiming at the online medical community text corpus selected in this paper, this paper merges and removes the duplication of multiple stop word sets such as the stop words list of Harbin Institute of Technology, stop words list of machine learning intelligent laboratory of Sichuan University, Baidu stop words list and so on, and then simply expands the stop words to extract a more comprehensive stop words list, with a total of 1887 stop words. Specifically, the following three kinds of stop words are involved: blank string in the text, all kinds of punctuation for full and half angle, and special symbols; words that are widely used and cannot well reflect the semantic information of the Chinese

text, such as "当时" ("then"), "但是" ("but"), etc; words with high probability but no specific meaning, such as "啊" ("ah"), "哦" ("oh"), "嗯" ("um"), "的" ("of"), etc.

After the above text preprocessing steps, the text data of consultation for patients in endocrinology department under Haodaifu Online platform, an online medical community, is compared before and after processing, as shown in Figures 5 and 6.

Figure 5. Texts before data preprocessing


Text 1: ['甲亢 突眼', '2020-04-03', '门诊', '暂时没开', '暂时没开', '观察，定期复查', '一个月后复查', '病历资料仅医生及患者本人可见']
(['Thyrotoxic exophthalmos', '2020-04-03', 'Outpatient', 'No prescription yet', 'No prescription yet', 'Observe, reexamine regularly', 'Reexamine in a month', 'The medical records datas are only visible to the doctor and the patient'])

Text 2: ['低钾血症', '3月26日连续喝了5杯浓咖啡后，开始出现腹泻 呕吐 四肢无力 心跳加速症状，送到急诊后为缺钾，血钾3.44，后吃了药，没有再出现腹泻呕吐情况，但心跳仍快90多左右，而且四肢无力，容易发呆，28日 复查血钾3.6 4月1日复查，血钾4.2，当地医院不收治我，说我没事，请问医生，我该怎么办？以后还会不会再犯？我是哪里出了问题引起的低钾血症？', '当地医院不收治，去外地花费太多，我应该怎么办？', '未怀孕', '一周内', '锡林郭勒盟医院 肾内科 内分泌科', '氯化钾口服片 服用了一天 共4粒', '病历资料仅医生和患者本人可见', '无（2019-03-19填写）', '无（2019-03-19填写）', '病历资料仅医生和患者本人可见']
(['Hypokalemia', 'On March 26, after drinking five cups of espresso continuously, I began to have diarrhea, vomiting, weakness of limbs, rapid heartbeat. After being sent to the emergency department, I was potassium deficiency. Serum potassium was 3.44. After taking the medicine, diarrhea and vomiting did not recur. But my heartbeat is still faster than 90 or so, and limbs weak, easy to daze. On the 28th, serum potassium was reexamined for 3.6. Reexamination on April 1, serum potassium was 4.2. The local hospital refused to treat me and said I was fine. Excuse me, doctor, what should I do? Will it happen again? Where did I go wrong to cause hypokalemia?', 'The local hospital refused treatment, it costs too much to go to other places. What should I do?', 'Not pregnant', 'Within a week', 'Xilin Gol League hospital, Nephrology, Endocrinology', 'Potassium chloride oral tablets, take a day, a total of 4 grains', "The medical records datas are only visible to the doctor and the patient', 'None (2019-03-19 filled in)', 'None (2019-03-19 filled in)', 'The medical records datas are only visible to the doctor and the patient'])

Text 3: ['这几天出院后血压血糖太正常，', '2020-03-31', '住院', '已出院', '否', '2020-04-02', '出院后一个月复诊', '带药了', "n　但身体没精神，用调药不，", ']
(['Blood pressure and blood sugar are too normal after discharge these days, ', '2020-03-31', 'Hospitalized', 'Discharged', 'No', '2020-04-02', 'Follow up one month after discharge', 'have the medicine', "n　But the body is not energetic, can I adjust the medicine?'])

Figure 6. Texts after data preprocessing

Text 1: 甲亢 突眼 观察 定期 复查 一个月 复查 病历 资料
(thyrotoxicosis, exophthalmos, observe, regular, reexamine, a month, reexamine, medical records, datas)

Text 2: 低钾血症 3 月 26 日 连续 喝 5 杯 浓咖啡 腹泻 呕吐 四肢无力 心跳 加速 症状 送到 急诊 后为 缺钾 血钾 3.44 吃 药 腹泻 呕吐 情况 心跳 仍快 90 四肢无力 发呆 28 日 复查 血钾 3.6 4 月 1 日 复查 血钾 4.2 医院 收治 说 没事 请问 我该 还会 再犯 低钾血症 医院 收治 外地 花费 太 一周 锡林郭勒盟 医院 肾内科 内分泌科 氯化钾 口服 片 服用 4 粒
(hypokalemia, 3, month, 26, day, continuously, drink, 5, cup, espresso, diarrhea, vomit, weakness of limbs, heartbeat, accelerate, symptom, send to, emergency department, after, potassium deficiency, serum potassium, 3.44, eat, medicine, diarrhea, vomit, situation, heartbeat, still fast, 90, weakness of limbs, daze, 28, day, reexamine, serum potassium, 3.6, 4, month, 1, day, reexamine, serum potassium, 4.2, hospital, treat, say, fine, excuse me, I should, will, happen again, hypokalemia, hospital, treat, cost, too, a week, Xilin Gol League, hospital, nephrology, endocrinology, potassium chloride, oral, tablet, take, 4, grain)

Text 3: 几天 出院 血压 血糖 太 住院 出院 否 出院 一个月 复诊 带药 身体 精神 调药
(days, discharge, blood pressure, blood sugar, too, hospitalized, discharge, no, discharge, one month, follow up, have medicine, body, energetic, adjust medicine)

Word Vector Training on Experimental Corpus

Related researches show that there are differences between the two models CBOW and Skip-Gram included in word2vec: the training results of the CBOW model are more focused on the description of the grammatical information of words, which can have a higher accuracy of grammar testing; the Skip-Gram model can better distinguish the semantics of words and describe the semantic features of words more accurately, so the obtained semantic calculation accuracy is higher. In this paper, the Skip-Gram model of word2vec is used for word vector learning, and word vectors with dimensions of 100, 150, 200, 250, 300 are obtained respectively, and the final experiment selects the training results of word vectors with dimension 200. Table 2 shows examples of word vector training results arranged in decreasing similarity.

Table 2. Examples of word vector training results in experimental corpus

word	near-synonym	cosine similarity	word	near-synonym	cosine similarity	word	near-synonym	cosine similarity
高血压 (hypertension)	高血脂 (hyperlipidemia)	0.758025	癌 (cancer)	乳头状 (papillary)	0.955317	饮食 (diet)	饮食习惯 (dietary habit)	0.662150
	心脏病 (heart disease)	0.731171		微小 (tiny)	0.855288		严格控制 (strictly control)	0.651480
	高血压病 (hypertension)	0.712235		乳头 (papilla)	0.850766		忌口 (on a diet)	0.649291
	脑梗塞 (cerebral infarction)	0.709198		转移 (metastasize)	0.833712		适量 (appropriate amount)	0.636538
	三高 (three tenors)	0.701040		全切 (total resection)	0.811079		刻意 (deliberate)	0.636243
	冠心病 (coronary disease)	0.697877		微癌 (microcarcinoma)	0.810809		食谱 (recipe)	0.631736
	脑出血 (cerebral hemorrhage)	0.695403		半切 (hemisection)	0.807132		清淡 (light)	0.626768

Similarity Calculation

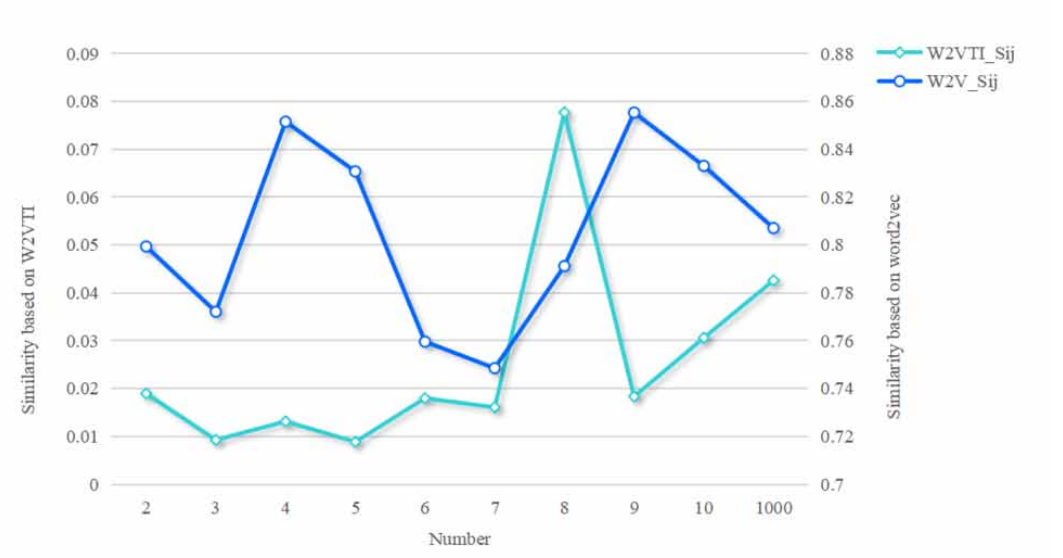
The used word2vec training model is backed by the captured Chinese text data set of large-scale patient consultation in online medical community, and the text similarity calculation is based on W2VTI and word2vec.

$W2VTI\_S_{ij}(i,j = 1,2,3...1000)$ represents the similarity value based on W2VTI between texts, $W2V\_S_{ij}(i,j = 1,2,3...1000)$ represents the cosine similarity of the text vector calculated by the average word vector based on word2vec. For example, $W2VTI\_S_{12}$ represents the similarity between text 1 and text 2 based on W2VTI, $W2V\_S_{12}$ represents the cosine similarity between text 1 and text 2. Several of them are list in Table 3.

Table 3. Similarity based on W2VTI and word2vec

Number	$W2VTI\_S_{ij}$	Similarity	$W2V\_S_{ij}$	Similarity
1	$W2VTI\_S_{1\,1}$	1	$W2V\_S_{1\,1}$	1
2	$W2VTI\_S_{1\,2}$	0.018956	$W2V\_S_{1\,2}$	0.799338
3	$W2VTI\_S_{1\,3}$	0.009263	$W2V\_S_{1\,3}$	0.772080
4	$W2VTI\_S_{1\,4}$	0.013134	$W2V\_S_{1\,4}$	0.851452
5	$W2VTI\_S_{1\,5}$	0.008861	$W2V\_S_{1\,5}$	0.830697
6	$W2VTI\_S_{1\,6}$	0.017981	$W2V\_S_{1\,6}$	0.759594
7	$W2VTI\_S_{1\,7}$	0.016088	$W2V\_S_{1\,7}$	0.748523
8	$W2VTI\_S_{1\,8}$	0.077659	$W2V\_S_{1\,8}$	0.791143
9	$W2VTI\_S_{1\,9}$	0.018325	$W2V\_S_{1\,9}$	0.855283
10	$W2VTI\_S_{1\,10}$	0.030599	$W2V\_S_{1\,10}$	0.832998
...	...	...	...	...
1000	$W2VTI\_S_{1\,1000}$	0.042587	$W2V\_S_{1\,1000}$	0.807020

Figure 7. Comparison of similarity coefficients between texts



It can be seen from Figure 7 that the similarity based on W2VTI is different from the cosine similarity based on word2vec in the concentrated area, and the similarity value based on W2VTI is generally smaller. On the whole, the trend of similarity between the texts obtained by the two calculation methods is similar, but there are also inconsistencies.

Cluster Results Based on DPeak

In this paper, the first 5000 pieces of data are selected from the 14671 pieces of valid data after cleaning. Among them, every 1000 pieces of data are used as a group of data for experiments. The clustering results of text similarity measurement based on W2VTI and word2vec are shown in Tables 4 to 7 with different number of clusters. In Tables 4 to 7, "NAS" means number of actual samples in the clusters.

Table 4. Clustering results with 3 clusters

Clustering number		1	2	3
First group	NAS	620	159	221
	W2VTI	620	130	250
	Word2vec	608	106	286
Second group	NAS	261	246	493
	W2VTI	250	234	516
	Word2vec	396	14	590
Third group	NAS	912	37	51
	W2VTI	955	3	42
	Word2vec	804	152	44

Table 4 continued on next page

Table 4 continued

Clustering number		1	2	3
Fourth group	NAS	252	694	54
	W2VTI	272	700	28
	Word2vec	353	633	14
Fifth group	NAS	43	885	72
	W2VTI	26	972	2
	Word2vec	152	843	5

Table 5. Clustering results with 4 clusters

Clustering number		1	2	3	4
First group	NAS	598	159	221	22
	W2VTI	608	130	250	12
	Word2vec	570	106	286	38
Second group	NAS	261	122	493	124
	W2VTI	250	109	516	125
	Word2vec	396	121	469	14
Third group	NAS	852	37	51	60
	W2VTI	954	3	42	1
	Word2vec	803	152	44	1
Fourth group	NAS	250	252	54	444
	W2VTI	261	272	28	439
	Word2vec	219	414	14	353
Fifth group	NAS	43	885	33	39
	W2VTI	26	937	2	35
	Word2vec	152	834	5	9

Table 6. Clustering results with 5 clusters

Clustering number		1	2	3	4	5
First group	NAS	541	159	221	22	57
	W2VTI	569	130	250	12	39
	Word2vec	560	106	286	38	10
Second group	NAS	240	122	493	124	21
	W2VTI	247	109	516	125	3
	Word2vec	396	121	397	14	72
Third group	NAS	813	37	51	60	39
	W2VTI	895	3	42	59	1
	Word2vec	802	152	44	1	1
Fourth group	NAS	159	252	54	444	91
	W2VTI	170	272	28	439	91
	Word2vec	219	374	14	353	40
Fifth group	NAS	43	860	33	39	25
	W2VTI	26	922	2	35	15
	Word2vec	152	757	5	9	77

Table 7. Clustering results with 6 clusters

Clustering number		1	2	3	4	5	6
First group	NAS	541	159	221	22	37	20
	W2VTI	569	121	250	12	39	9
	Word2vec	560	106	158	38	10	128
Second group	NAS	240	122	493	77	21	47
	W2VTI	246	109	516	125	3	1
	Word2vec	396	121	397	14	71	1
Third group	NAS	772	37	51	60	39	41
	W2VTI	894	3	42	59	1	1
	Word2vec	802	151	44	1	1	1
Fourth group	NAS	159	252	54	346	91	107
	W2VTI	170	272	28	321	91	118
	Word2vec	219	374	14	314	40	39
Fifth group	NAS	43	833	33	39	25	27
	W2VTI	26	922	2	35	14	1
	Word2vec	152	583	5	9	77	174

Comparative Analysis with F-Measure

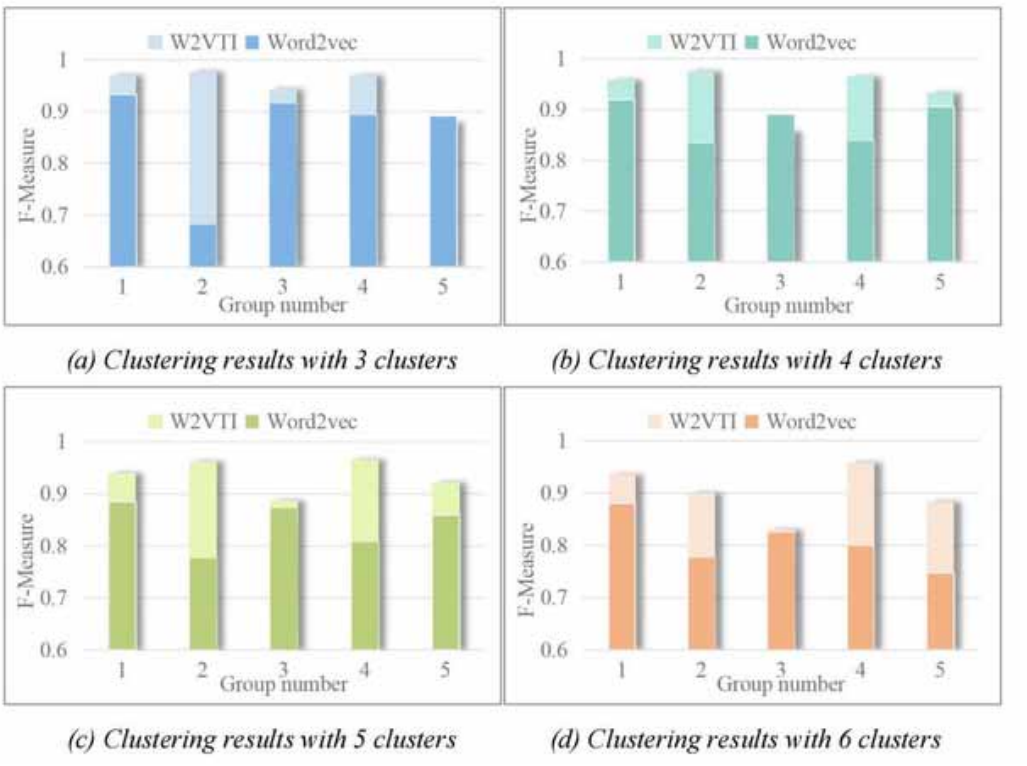
Here, this paper selected 3–6 as the numbers of classes, and we determined $\alpha = 1$ for Eq. (15). For

different cluster numbers, the clustering results obtained by the two semantic similarity calculation methods are different. The evaluation index F-Measure is shown in Table 8, and the clustering results obtained by W2VTI method are obviously better than those obtained by cosine similarity based on word2vec.

Table 8. Evaluation index F-Measure of clustering results

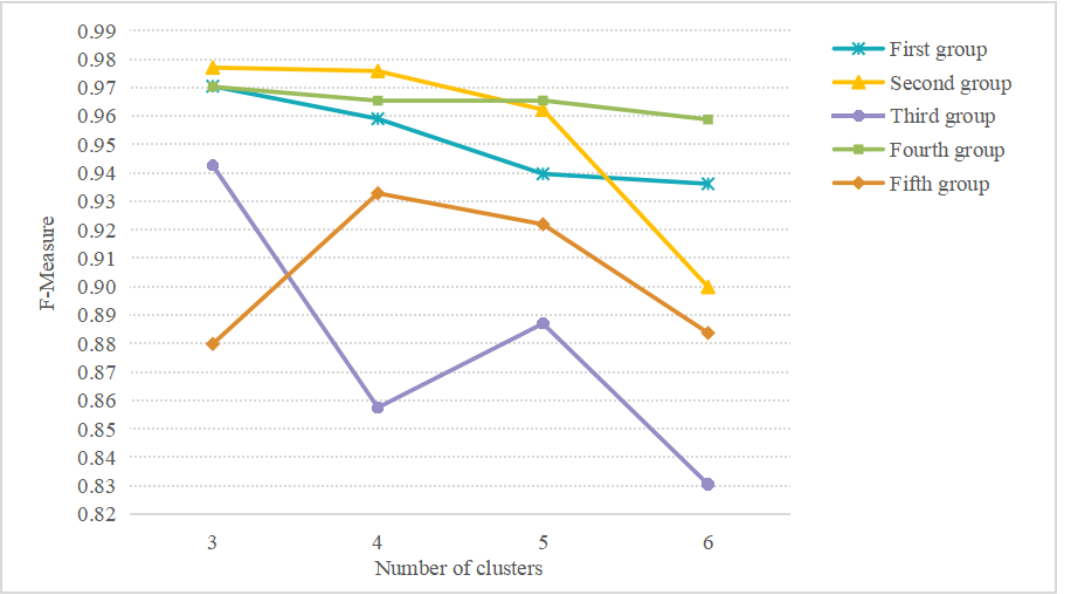
Number of clusters	Method	First group	Second group	Third group	Fourth group	Fifth group
3	W2VTI	0.9704	0.9770	0.9426	0.9703	0.8798
	Word2vec	0.9338	0.6827	0.9163	0.8943	0.8918
4	W2VTI	0.9590	0.9758	0.8575	0.9654	0.9328
	Word2vec	0.9197	0.8347	0.8905	0.8397	0.9058
5	W2VTI	0.9396	0.9622	0.8870	0.9654	0.9219
	Word2vec	0.8847	0.7771	0.8731	0.8078	0.8597
6	W2VTI	0.9361	0.8998	0.8305	0.9587	0.8837
	Word2vec	0.8804	0.7777	0.8250	0.8008	0.7477

Figure 8. Comparison of clustering results based on two methods



(a) Clustering results with 3 clusters (b) Clustering results with 4 clusters
(c) Clustering results with 5 clusters (d) Clustering results with 6 clusters

Figure 9. Clustering results based on W2VTI with different numbers of clusters



As can be seen from Figure 8, for the same number of clusters, the increasing extent of F-Measure by W2VTI method is different on different data sets; For different cluster numbers, on the same data set, the increasing extent of F-Measure by W2VTI method is not equal, but it is relatively close. As can be seen from Figure 9, for different data sets, the optimal cluster number obtained by W2VTI method calculating semantic similarity is also different. For the first, second, third and fourth data sets, the best clustering results are obtained when the number of clusters is 3, and for the fifth data set, the best clustering result is obtained when the number of clusters is 4. Among all the clustering results obtained by W2VTI method, the second data set obtains the clustering result with the largest F-Measure when the number of clusters is 3.

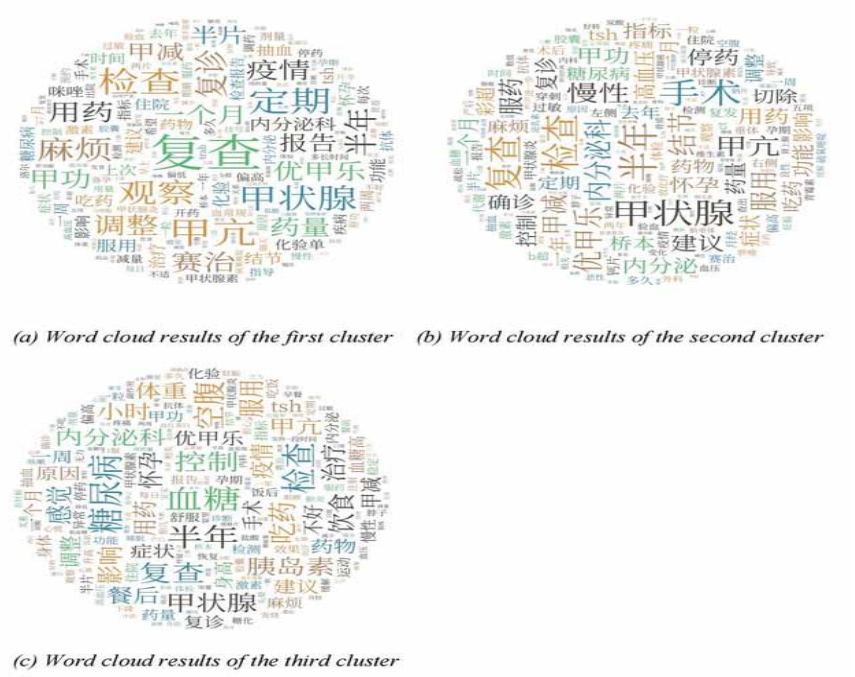
Cluster Analysis of Online Medical Inquiry Texts

Word clouds are an intuitive visualization technology, which is usually used to provide the first impression of text documents. They display the most frequent words in the text as a weighted list of words in a specific spatial layout. The font size of a word indicates its relevance or frequency of occurrence, while other visual characteristics (such as color, position, orientation) are often varied for aesthetic reasons or visually encoding additional information (Lohmann et al., 2015). Word clouds are used as a more in-depth text analysis method. This paper takes the second set of data as an example, further analyzes the clustering results based on W2VTI and DPeak with cluster number of 3, and obtains the word clouds as shown in Figure 10.

The size of the keywords in the word clouds is displayed according to the importance of the words in the clustering results. In order to enhance the display of effective information in the word clouds, the words of entity person, place and so on that have nothing to do with the condition inquiry (such as doctor, professor, teacher, central hospital, etc.) are deleted. The subgraphs (a), (b) and (c) in Figure 10 all show words such as “thyroid”, “diabetes”, “hyperthyroidism”, “follow-up visit”, “chronic” and

"dosage", indicating that for endocrinology patients, no matter what kind of problems are consulted, the topics of "thyroid", "diabetes" and "hyperthyroidism" are inseparable. Words such as "follow-up visit", "chronic" and "dosage" also reflect that endocrine diseases generally have the characteristics of a long disease cycle, and certain drugs need to be taken. The differences of the three subgraphs are as follows: subgraph (a) takes "take medicine", "dosage", "drug", "dose" and other problems of thyroid disease patients taking medicine as the hot topic of inquiry. Subgraph (b) is different from subgraph (a), in the word cloud of the second cluster, there are many keywords for consultation of patients with "thyroid cancer", such as "puncture", "diagnosis", "resection", "operation", "postoperation", "radiotherapy and chemotherapy". Compared with the first cluster, patients in the second cluster are more severely ill. Subgraph (c) is also different from subgraph (a) and (b), the word cloud of the third cluster shows "insulin", "diet", "fasting", "postprandial", "monitoring", "saccharification", "glucose meter" and other subject words related to "diabetes" patients with blood sugar control. In general, the above three word clouds have their own obvious characteristics, which indicates that the text similarity measurement method proposed in this paper based on cluster analysis is feasible and effective to a certain extent. The visualization of hot topic discovery results quickly discovers and displays the hot topics of patient consultation, and realizes the summary of the inquiry content.

Figure 10. Word clouds of the second group of data clustering results



(a) Word cloud results of the first cluster (b) Word cloud results of the second cluster
(c) Word cloud results of the third cluster

CONCLUSION

Online medical community text has the characteristics of large amount of data, high sparsity, strong professionalism and complex context. Due to individual different language habits, there are a lot

of abbreviations and vague words in doctor-patient question and answer text, which brings huge challenges to text mining. According to the characteristics of text data in online medical communities, this paper proposes a text similarity measurement method, which aims to obtain text information from multiple directions and improve the effect of text data mining. This paper applies the density peak clustering algorithm to the discovery of hot topics in the online medical community for the first time, and applies the proposed semantic similarity measurement method to the clustering algorithm. The experimental results show that the similarity measurement method proposed in this paper is better than the cosine similarity method based on the mean word2vec, and in all the experimental results of the proposed method, the highest F-Measure reaches 97.7%, and the lowest F-Measure still reaches 83.05%, which means that this method is effective. Hence, the paper provides a reference for mining hot topics of patient consultation in online medical community. According to the cluster analysis results of this study, the hot topics of consultation for endocrinology patients in the online medical community mainly focus on three categories: medication for chronic thyroid diseases; surgical treatment of thyroid diseases; glycemic control in diabetes. Although the method proposed in this paper has achieved good recognition results, the data involved in the research only includes the department of endocrinology under the Haodaifu Online platform, and the text content is relatively not extensive enough. Future works will be carried out in combination with more types of diseases and medical community platforms.

REFERENCES

Abbas, O. A. (2007). Comparisons between data clustering algorithms. *The International Arab Journal of Information Technology*, *5*(3), 320–325. doi:10.1007/s10796-008-9081-8

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, *39*(1), 45–65. doi:10.1016/S0306-4573(02)00021-3

Alqahtani, A., Alhakami, H., Alsubait, T., & Abdullah, B. (2021). A survey of text matching techniques. *Engineering, Technology &. Applied Scientific Research*, *11*(1), 6656–6661. doi:10.48084/etasr.3968

Biondi, G., Franzoni, V., Li, Y., & Milani, A. (2016). Web-based similarity for emotion recognition in web objects. *Proceedings of 2016 IEEE/ACM 9th International Conference on Utility and Cloud Computing*, 327-332. doi:10.1145/2996890.3007883

Hajli, M. N. (2014). Developing online health communities through digital media. *International Journal of Information Management*, *34*(2), 311–314. doi:10.1016/j.ijinfomgt.2014.01.006

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the F-Measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association: JAMIA*, *12*(3), 296–298. doi:10.1197/jamia.M1733 PMID:15684123

Jorge, M. G. (2014). An overview of textual semantic similarity measures based on web intelligence. *Artificial Intelligence Review*, *42*(4), 935–943. doi:10.1007/s10462-012-9349-8

Sims, J. M. (2018). Communities of practice: Telemedicine and online medical communities. *Technological Forecasting and Social Change*, *126*, 53–63. doi:10.1016/j.techfore.2016.08.030

Khatua, A., Khatua, A., & Cambria, E. (2019). A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. *Information Processing & Management*, *56*(1), 247–257. doi:10.1016/j.ipm.2018.10.010

Li, X., Liu, N., Yao, C., & Fan, F. (2017). Text similarity measurement with semantic analysis. *International Journal of Innovative Computing, Information, & Control*, *13*(5), 1693–1708.

Lin, Y. S., Jiang, J. Y., & Lee, S. J. (2014). A similarity measure for text classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, *26*(7), 1575–1590. doi:10.1109/TKDE.2013.19

Liu, X., Wu, D., Peng, H., & Wang, R. (2018). Health topics mining in online medical community. *2018 IEEE Global Communications Conference*, 1-6. doi:10.1109/GLOCOM.2018.8647970

Lohmann, S., Heimerl, F., Bopp, F., Burch, M., & Ertl, T. (2015). Concentri cloud: word cloud visualization for multiple text documents. *2015 19th International Conference on Information Visualisation*, 114-120. doi:10.1109/iV.2015.30

Lu, Y. J., Zhang, P. Z., Liu, J. F., Li, J., & Deng, S. S. (2013). Health-related hot topic detection in online communities using text clustering. *PLoS One*, *8*(2), e56221. doi:10.1371/journal.pone.0056221 PMID:23457530

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* Retrieved from https://arxiv.org/abs/1301.3781

Ngai, E. W., Lam, S., Poon, J., Shen, B., & Moon, K. K. (2016). Design and development of intelligent decision support prototype system for social media competitive analysis in fashion industry. *Journal of Organizational and End User Computing*, *28*(2), 13–32. doi:10.4018/JOEUC.2016040102

Ni, L., Luo, W., Zhu, W., & Liu, W. (2019). Clustering by finding prominent peaks in density space. *Engineering Applications of Artificial Intelligence*, *85*, 727–739. doi:10.1016/j.engappai.2019.07.015

Prakoso, D. W., Abdi, A., & Amrit, C. (2021). Short text similarity measurement methods: A review. *Soft Computing*, *25*(6), 1–25. doi:10.1007/s00500-020-05479-2

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, *344*(6191), 1492–1496. doi:10.1126/science.1242072 PMID:24970081

Shajalal, M., & Aono, M. (2019). Semantic textual similarity between sentences using bilingual word semantics. *Progress in Artificial Intelligence*, *8*(2), 263–272. doi:10.1007/s13748-019-00180-4

Shen, X., Yang, W., & Sun, S. (2019). Analysis of the impact of China's hierarchical medical system and online appointment diagnosis system on the sustainable development of public health: A case study of Shanghai. *Sustainability*, *11*(23), 6564. doi:10.3390/su11236564

Soganciogl u, G., Öztürk, H., & Özgür, A. (2017). BIOSSES: A semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics (Oxford, England)*, *33*(14), i49–i58. doi:10.1093/bioinformatics/btx238 PMID:28881973

Sung, S. F., Lee, P. J., Hsieh, C. Y., & Zheng, W. L. (2020). Medication use and the risk of newly diagnosed diabetes in patients with epilepsy: A data mining application on a healthcare database. *Journal of Organizational and End User Computing*, *32*(2), 93–108. doi:10.4018/JOEUC.2020040105

Taeho, J. (2018). Automatic text summarization using string vector based K nearest neighbor. *Journal of Intelligent & Fuzzy Systems*, *35*(6), 6005–6016. doi:10.3233/JIFS-169841

Tandel, S. S., Jamadar, A., & Dudugu, S. (2019). A survey on text mining techniques. *5th International Conference on Advanced Computing & Communication Systems*, 1022-1026. doi:10.1109/ICACCS.2019.8728547

Tissot, H., & Dobson, R. (2019). Combining string and phonetic similarity matching to identify misspelt names of drugs in medical records written in Portuguese. *Journal of Biomedical Semantics*, *10*(S1), 17. doi:10.1186/s13326-019-0216-2 PMID:31711534

Wang, J., & Dong, Y. (2020). Measurement of text similarity: A survey. *Information (Switzerland)*, *11*(9), 421. doi:10.3390/info11090421

Yu, Z., Chen, M., & Liu, L. (2015). A review on text mining. *6th IEEE International Conference on Software Engineering and Service Science*, 681-685. doi:10.1109/ICSESS.2015.7339149

Zaher, M., Shehab, A., Elhoseny, M., & Farahat, F. F. (2020). Unsupervised model for detecting plagiarism in internet-based handwritten Arabic documents. *Journal of Organizational and End User Computing*, *32*(2), 42–66. doi:10.4018/JOEUC.2020040103

Zhang, N., Zhang, R., Pang, Z. L., Liu, X., & Zhao, W. F. (2021). Mining express service innovation opportunity from online reviews. *Journal of Organizational and End User Computing*, *33*(6), 3. doi:10.4018/JOEUC.20211101.oa3

Zhao, D., Wang, J., Lin, H., Chu, Y., Wang, Y., Zhang, Y., & Yang, Z. (2021). Sentence representation with manifold learning for biomedical texts. *Knowledge-Based Systems*, *218*, 106869. doi:10.1016/j.knosys.2021.106869

Zhou, S., Xu, X., Liu, Y., Chang, R., & Xiao, Y. (2019). Text similarity measurement of semantic cognition based on word vector distance decentralization with clustering analysis. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 107247–107258. doi:10.1109/ACCESS.2019.2932334

Mingyang Li received the Master's degree from the Jilin University of Finance and Economics in 2019. She is currently pursuing the Ph.D. degree with the School of Management, Jilin University, China. Her research interests include computer science, big data analysis, machine learning, data mining, and text mining.

Xinhua Bi, professor, corresponding author of this paper, has published more than 100 academic papers on Social Science Citation Index (SSCI), Chinese Social Science Citation Index (CSSCI), Chinese core journals of natural science and important international conferences (EI, ISTP retrieval), and 5 academic monographs and textbooks.

Limin Wang, professor, received the Master's and Ph.D. degrees in computer science and technology from Jilin University, in 2004 and 2007, respectively. Her current research interests include big data analysis, evolutionary algorithm, and intelligent decision optimization.

Xuming Han, professor, received the Master's and Ph.D. degrees in computer application technology from Tianjin University and Jilin University, in 2006 and 2010, respectively. His main research interests include data mining, machine learning, simulation modeling, intelligent computing, swarm intelligence optimization.

Lin Wang is currently pursuing the Ph.D. degree with the School of Management, Jilin University, China. Her main research direction is machine learning.

Wei Zhou received the Master's degree from the Jilin University of Finance and Economics, in 2020. He is currently pursuing the Ph.D. degree with School of Computer Science and Technology, Changchun University of Science and Technology, China. His research interests include computer science, big data analysis, data mining, machine learning, and intelligent decision optimization.