Computer Assisted Methods in Engineering and Science
31(2): 139–155, 2024, doi: 10.24423/cames.2024.493

This article belongs to the *Special Issue on Scientific Computing and Learning Analytics for Smart Healthcare Systems* edited by Dr. C. Chakraborty, Dr. S. Barbosa and Dr. L. Garg

Semantic Web Techniques for Clinical Topic Detection in Health Care

R. RAMAN[1], Kishore Anthuvan SAHAYARAJ[2], Mukesh SONI[3]*, Nihar Ranjan NAYAK[4], Ramya GOVINDARAJ[5], Nikhil Kumar SINGH[6]

[1] *Department of Electronics and Communication Engineering, Aditya College of Engineering, Surampalem, Andhra Pradesh, India; e-mail: ramanphdr@gmail.com*

[2] *Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Tamil Nadu, India; e-mail: kishorea1@srmist.edu.in*

[3] *Dr. D. Y. Patil Vidyapeeth, Pune, Dr. D. Y. Patil School of Science & Technology, Tathawade, Pune, India; e-mail: mukesh.research24@gmail.com*

[4] *Department of Computer Science & Engineering and Information Science, Presidency University, Bangaluru, India; e-mail: nayak.niharranjan0@gmail.com*

[5] *School of Information Technology and Engineering, Vellore Institute of Technology, India; e-mail: ramya.g@vit.ac.in*

[6] *Department of Computer Science Engineering, Maulana Azad National Institute of Technology, Bhopal, India; e-mail: nikhilsinghmanit@gmail.com*

* *Corresponding Author e-mail: mukesh.research24@gmail.com*

The scope of this paper is that it investigates and proposes a new clustering method that takes into account the timing characteristics of frequently used feature words and the semantic similarity of microblog short texts as well as designing and implementing microblog topic detection and detection based on clustering results. The aim of the proposed research is to provide a new cluster overlap reduction method based on the divisions of semantic memberships to solve limited semantic expression and diversify short microblog contents. First, by defining the time-series frequent word set of the microblog text, a feature word selection method for hot topics is given; then, for the existence of initial clusters, according to the time-series recurring feature word set, to obtain the initial clustering of the microblog.

Keywords: clinical text, frequent word set, feature selection, clustering, topic detection, time sequence, semantics.

1. Introduction

Topic detection and tracking (TDT) refers to the automatic identification of the topic of the news data stream without manual intervention to deal with the increasingly severe problem of information overload [1]. Its main task is to extract important features from the media information stream, monitor different news reports, and detect and organize previously unknown topics in the information flow without prior knowledge of issues [2]. Since the research on topic detection was jointly initiated by the Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST) in 1996, it has received a wide range of responses [3], and some scholars have used single-pass incremental, Smart Cities measures for data mining [4]. Some techniques provide for digital watermarking over wireless communication using the Internet of Things (IoT) application. During the detection, the topic detection effect has been improved to a certain extent [5, 6]. Semantic web strategies are crucial in healthcare field. They are extremely beneficial for the secure transmission and confidentiality of healthcare data transferred between a user and a server through a wireless connection. To address the issues posed by microblog short text clustering, this study employs the frequent itemset hierarchical clustering (FIHC) technique, the concept of "first clustering, then deleting duplication, and finally reducing", and provides a novel technique integrating time-series data FIHC (TS-FIHC).

The limitation of the TDT technique is that it is linked to the preceding issue in a roundabout way, creating the object tracking latency. Topic recognition still confronts several issues deserving additional exploration due to the uniqueness of hot topics in clinical texts [7]. It has a restricted vocabulary, and the degree of similarity across languages does not alter dynamically when themes change.

Researchers also propose secure data transfer scheme for mobile nodes in smart city applications in the wireless network to secure data transfer by implementing an energy-efficient and secure communication scheme [8].

Due to the short, fast, and variable topics of clinical texts, as well as the high dimensionality and sparsity of vectors brought about by its unstructured format, clinical texts differ from conventional texts in several ways [9]. A clinical text is enlarged on the time axis and in sequential order, according to its time-series properties. It can be scattered to various times if a time window is defined on the time axis.

Traditional news topic detection research also faces new challenges:

1) The classic text representation model – vector space model (VSM) constructed with bag of words will lead to the "high-dimensional curse" problem, and new feature selection methods for microblog texts need to be studied;
2) The feature sparsity of short texts will lead to traditional similarity based on spatial distance. The calculation method cannot effectively measure the similarity between two short texts, and new text similarity calculation methods need to be studied;

3) Traditional clustering algorithms usually need to preset the number of initial clusters or termination clusters, but this prior knowledge in the topic recognition of text clustering is often unknown. If inappropriate number parameters are set, it will lead to unsatisfactory clustering results.

Therefore, it is necessary to reasonably assess the initial clusters and terminations of topics according to the internal nature of the microblog text collection – the number of groups. As a result, visualization techniques might be investigated further in order to demonstrate the interrelationships between topic clusters and intuitively facilitate topic recognition.

The major contribution of this research is that it mainly focuses on the new problems faced by microblog topic detection research, proposes a new clustering method that considers the timing characteristics of frequent feature words and the semantic similarity of microblog short texts, and designs and implements microblog topic detection based on clustering results.

The paper is structured as follows. Section 1 presents several topics related to this study, Sec. 2 discusses the related work in this field followed by Sec. 3 describing the proposed methodology. Section 4 describes the experimental setup and result analysis, and Sec. 5 concludes the paper.

2. Related works

2.1. Research on clinical topic detection system method

As a new online media content, clinical texts have several characteristics, such as rapid information growth, real-time solid content, and random language. Twitter's hot topic recognition approach associated with social network assessment and pattern, which collects tweets as a collection of words acquires hot issues by building a basic database, extracting emergence terms via the term life cycle model, and identifying impact of specific users via the user's social media platform [10]. Some scholars have taken the lead in obtaining preliminary research results in clinical topic detection [11]. In some research, Twitter users are considered as network sensors, and the Bayesian decision-making method based on keyword evidence is adopted to design and develop a Twitter-based real-time earthquake monitoring prototype system, and achieves a detection rate of more than 80%. A method of collecting, grouping, sorting, and tracking breaking news on Twitter was proposed [12]. Tweets with high similarity were grouped into one group, treated as one news item, and then tracked based on the vocabulary of each news topic group [13]. These news topics were sorted by the degree of connection and popularity between them, and finally a hot topic was obtained. A Twitter hot topic detection method based on social system evaluation and time sequence, which extracts tweets as a collection of a series of words, mining emergent words through the word life cycle model, mining the influence of specific users through the user's social network, and calculating the importance of tweets based on this, and finally, obtaining hot topics by creating a basic topic table was proposed [14]. With the rapid popularity of mainstream platforms such as MedHelp, some scholars have also begun researching topic detection in clinical texts. A microblog news topic detection method based on the hidden topic model based on the characteristics of a large amount of microblog data and fragmented information was proposed. The method was based on the characteristics of clinical topic timing and short text semantic similarity, and topic detection and tracking system method based on clinical text clustering [15].

The present study takes the FIHC algorithm's principle of "first clustering, then removing duplicates, and finally condensing" to tackle the challenges of microblog short text clustering, and presents a novel technique integrating time series frequent and semantic clustering time and semantics (TS-FIHC).

2.2. Research on clustering algorithm for short microblog text

Text content clustering is still the core of microblog topic detection, but the unique attributes of microblog short texts make traditional clustering algorithms unable to obtain better application effects. Therefore, the research on microblog short text clustering methods is fundamental. The starting cluster is directly affected by the minimal cluster capability. The subject semantic membership calculation of the first clusters of the microblogs in the cluster centers of the sets is further influenced by the number of cluster features collected during the feature extraction stage, which finally influences the separation effect of the initial groups. The following mainly introduces two directions with excellent development potential [16]:

1) Based on extended semantic information, it was found that using external sources such as Wikipedia can expand the characteristics of short texts and improve the similarity between them. For example, the authors introduced WordNet to convert frequent word sets into standard concept sets and then proposed a text clustering algorithm based on standard concept sets [17]. Semantic web techniques play an important role in the healthcare industry [19]. They are very useful for providing security and privacy of healthcare information among users and servers over the wireless network [20]. For example, the wavelet based digital watermarking scheme for medical images for healthcare data was used against various attacks [18].
2) Based on co-occurrence between words or sequence the FTC (frequent term-based clustering) algorithm based on standard item sets was first proposed. The FTC algorithm uses standard word sets to represent clusters and adopts a greedy heuristic strategy [21]. The frequent word set selection order will affect the final clustering results. In [22], a text clustering

method called clustering based on frequent word sequences (CFWS) based on frequent word sequences was proposed. The algorithm constructs initial clusters by mining frequent word sequences and then uses the k-mismatch method to merge the initial groups and obtain the clustering results.

A regular itemset-based hierarchical clustering (FIHC, frequent itemset-based hierarchical clustering) algorithm is proposed for the shortcomings of the FTC algorithm. Various security protocols for security and privacy of healthcare information among users and servers over wireless networks are provided. This research offers a way to assess the complex data with hybrid neural network-based techniques based on ensemble methodology with extended gray wolf optimizing (E-GWO) feature extraction method. If we compare the proposed work with existing literature, we find that the proposed work investigates hot topic identification via clinical content, presents the challenge as a short text clustering problem, and provides a systematic approach.

The function of a semantic web has become more prevalent in several fields and the need for semantics in healthcare, online worlds, or knowledge extraction has increased significantly [23]. Moreover, we provide a detailed assessment of applying the semantic web to certain health industry internet forums and other data-gathering initiatives in this study.

The smart healthcare system is not only a network supported by smart technologies and devices. This is an overall advancement that has made all levels of healthcare institutions data-driven. For years, the conventional approach has been data-driven, but now it is time to create national health information environments, such as cities and neighborhoods, data-driven too.

Due to the short text characteristics of microblogs, if the FIHC algorithm is directly applied to microblog clustering, it will face the following two problems: ordinary frequent item sets only indicate that the co-occurrence of some feature items is regular, and the co-occurrence cannot be guaranteed. Researchers can represent the hidden topics of the text collection and affect the correct structure and division of topic clusters. Because the content of clinical texts is short and the features are scarce, some clinical texts on the same topic may have similar semantics but different expressions and thus be classified incorrectly. Eventually, this affects the clustering results.

3. Proposed method

3.1. Proposed framework

To solve the problems faced by microblog short text clustering, this paper adopts the FIHC algorithm, the idea of “first clustering, then eliminating duplicates and then condensing”, and proposes a new method combining time series frequent and semantic clustering time and semantics (TS-FIHC) [25]. Topics usually have time attributes, and hot topics in clinical tests are more sequential; they surge at a certain point in time, and their development trends are highly unbalanced [26]. Therefore, firstly, the time-series trend of frequent word sets is defined according to the time sliding window, and a microblog text feature selection based on the time-series word frequency is proposed, and the selected time-series trend frequent word sets are used to divide the initial topic clusters of clinical texts; to be more precise, we eliminate the text overlap between the initial groups, adopt HowNet's semantic similarity model, and separate the initial topic clusters according to the principle of full semantic membership. The microblog texts might be associated with several topics at the same time. As a result, it is worth looking at certain soft clustering algorithms to improve microblog topic detection. The subject groups that result from clustering typically include more relevant details; however, this implied data is difficult to be detected explicitly.

The proposed method plays an important role in the healthcare industry. It is very useful for providing security and privacy of healthcare information among users and servers over the wireless network. Finally, by defining the semantic similarity matrix between clusters, the aggregation of clinical topic clusters hierarchical clustering, according to the reference optimization, is conducted to obtain the final topic cluster and realize topic detection and tracking.

3.2. Clinical text feature selection based on time series word frequency

Clinical text is different from ordinary text. Its time-series characteristics indicate that a clinical text is expanded along the time axis and sequential order. If a time window is set on the time axis, the clinical text can be distributed at different times. A collection of microblogs based on a sliding window of time is obtained in the window. Therefore, the feature selection of clinical vocabulary can be considered from the timing.

Definition 1. Trend base. Put a specific feature word U_j in the i-th time sliding window. The trend base CU_{ji} is defined as the average of the frequency UG of the word in the previous k consecutive time sliding windows (here k is defined as the time window parameter, and when $i \leq k$, take $k = i - 1$):

$$CU_{ji} = \frac{\sum\limits_{s=1}^{k} UG_{(i-s)}}{k}. \tag{1}$$

Definition 2. Trend growth rate. Let the trended base of a particular characteristic word U_j be CU_{ji}, the frequency that appears in the i-th time sliding window is UG_i, and the trend growth rate of U_j in the i-th time sliding window is defined as:

$$HU_{ji} = \frac{UG_i}{CU_{ji}} = \frac{kUG_i}{\sum\limits_{s=1}^{k} UG_{(i=s)}}. \tag{2}$$

Definition 3. Time series trend degree. The topic trend of a feature word is directly proportional to its trend growth rate and trend base. Therefore, the calculation formula for defining the time series trend degree of the frequent word sets of clinical topics is:

$$UU_{ji} = HU_{ji} lbCU_{ji}. \tag{3}$$

Next, let us verify the effect of topic trend feature vocabulary selection and standard frequent feature vocabulary selection. Given that the current general clinical topic popularity is measured in days, and the news popularity of social networks generally does not exceed one week, this article collects statistics on MedHelp, randomly selecting according to the number of days set in the sliding window of time (no more than 7). The difference ratio curve of the feature vocabulary selection results based on the time series frequent word set and the general frequent word set is given. Figure 1 depicts differences in universal frequent characteristic word groups and time series prominent word collections of various sliding window sizes. The x-axis denotes the window size and the y-axis refers to the feature word sets. Figure 1 shows that the results of clinical feature selection based on time-series word frequency show differences compared to the available method based on frequent word sets when the number of features is small. For the applicability of the topic, a sliding window of 3 days is appropriate.

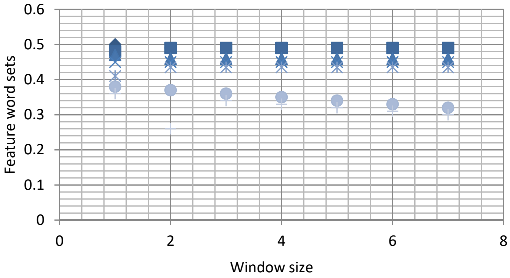

FIG. 1. Time-series frequent feature word sets of different sliding window sizes and differences in general frequent feature word sets.

3.3. Clinical initial clustering based on time series frequent word sets

Definition 4. For an item set Y in the text set E, if the number of occurrences of the item set Y in E is more significant than a preset ratio, then Y is the frequent item set in the text set E, and this preset ratio is called the minimum support s.

Definition 5. For a specific trend word set X in the text set E, if the support degree of X in E is $s(X) \geq \min\_s$; then the trend word set X is called the frequent trend word set on the text set E, and $\min\_s$ is the global minimum support. In this paper, a priori frequent set mining algorithm calculates periodic trend word sets. The execution is as follows. Steps:

1) Scan the text set E, use the word frequency trend to count the number of occurrences of the candidate item set, and collect the item set that meets the minimum support $\min\_s$ setting and record it as a frequent item set;
2) Use the generated frequent k-item sets to construct strong union rules, and use frequent k-item sets to create a candidate.

Choose $(k+1)$-item set, and repeatedly iterate until the candidate $(k+1)$-item set is empty. Frequent trend word sets can better describe the topic information hidden in clinical texts. This article uses standard trend word sets to construct initial clusters, that is, to divide clinical texts containing a specific frequent trend word set into one group, and get words based on periodic trend words in the initial collection of the set.

Definition 6. If the proportion of a time series vocabulary feature item set in the clinical set contained in the initial cluster exceeds a preset minimum ratio, the time series feature item set is called a cluster time series frequent item set.

This preset minimum ratio is recorded as the cluster minimum support θ. Since the frequent item sets of cluster time series refer to those core words that frequently appear in clusters, these core words are at a certain level. The degree represents the implicit semantics of the topic described by this cluster. The semantic topic information of this initial cluster can be obtained by extracting the cluster time series frequent item sets of each initial collection.

3.4. Inter-cluster overlap reduction algorithm based on semantic membership

As the goal of topic detection is to assign each microblog to a topic cluster, it is also necessary to design an algorithm to reduce the overlap between initial sets. It is an initial cluster overlap reduction algorithm based on the division of semantic membership to overcome the short semantic expression and diversification of short microblog texts. Combining the characteristics of the short text of microblogs, overlapping clinical text is finally allocated to the most suitable initial cluster.

Definition 7. If clinical doc_j is allocated to the initial cluster D_i, it is said that Clinical doc_j supports collection D_i.

Definition 8. Note that E_i and E_j are the set of all microblogs that support clusters D_i and D_j and $E_i \cap E_j \neq \emptyset$; that is, there are shared microblogs between groups D_i and D_j, so clusters D_i and D_j are said to be overlapping. Advance step by step, and remember that the set of microblogs overlapping between clusters is D^E, where $E = \{D_i, D_j\}$, $D\hat{}E = E_i \cap E_j$.

Definition 9. The degree of semantic membership of clinical topics. This paper defines the topic semantic membership function of clinical doc_j for the initial cluster D_i as:

$$\text{Score}\left(doc_j - D_i\right) = \frac{\sum\limits_{l=1}^{n} \max\limits_{k=1,2,\ldots,m} \{\text{sim}(g_{ik}, u_{jl})\}}{n}, \tag{4}$$

where frequent cluster 1-items set $\{g_{i1}, g_{i2}, ..., g_{im}\}$ represents the topic feature item of the initial cluster D_i, $\{u_{j1}, u_{j2}, ..., u_{jn}\}$ represents the microblog text j in the initial cluster D_i, doc_j feature item; $\text{sim}(g_{ik}, u_{jl})$ cluster feature item g_{ik} and text feature item u_{jl} defined in “HowNet” semantic similarity [23], n is the number of feature items of microblog text doc_j, m is the number of cluster feature items. Of course, the number of topic features in the initial cluster can be controlled by setting a reasonable minimum cluster support θ.

Algorithm 1 presents a detailed description of the first cluster overlap reduction approach based on clinical semantic membership.

Algorithm 1. Cluster overlap reduction approach.

```
Input  : Initialize cluster with overlap $D_1, D_2, ..., D_n$
Output : the initial set of overlap reduction $D'_1, D'_2, ..., D'_n$
: Record the initial cluster set of $m$ overlapping Clinical texts as
      $D^{E_1}, D^{E_2}, ..., D^{E_n}, \quad E_i \subset \{D_1, D_2, ..., D_n\}.$
: Initialize a two-dimensional array vector cluster Score hash :={,}
: for each $i$ from 1 to $n$
    : for each Clinical $doc_j$ in $D^{E_i}$
        : for each cluster $D_k \subset D^{E_i}$
            : CurrScore = Score $(doc_j \to D_k)$
            : if $doc_j \notin$ clusterScoreHash
                : add <$doc_j$, currScore> to cluster Scorehash
            : elseif currScore $\geq$ currScore of $doc_j \in$ clusterScoreHash
                : Update <$doc_j$, currScore> to clusterScorehash
            : else
                : del $doc_j$ from $D_k$
            : endif
            : $D'_k = D_k$
        : endfor
    : endfor
: endfor
```

The algorithm's complexity is $O(n)$, that is, only one scan of the clinical text in all overlapping initial clusters can reduce the overlap between all initial groups. Finally, the empty collections with a size of 0, after the initial sets are separated, are deleted. Then, a non-empty candidate topic cluster can be obtained.

3.5. Agglomerated topic clustering algorithm

The agglomerated topic clustering algorithm based on semantic similarity can obtain candidate topic clusters for microblog clustering topic detection through the overlap reduction between initial sets. Still, sometimes these topic clusters can be attributed to a specific topic, so it is necessary to carry out agglomerative hierarchical clustering of candidate topic clusters, merge topic clusters to reduce the number of significant issues, and provide users with more focused hot microblog topics. To merge candidate topic clusters, first measure the similarity between two candidate topic clusters. Since the candidate topic cluster is composed of many microblog texts, to ensure the efficiency of agglomerative hierarchical clustering, all microblog texts in the candidate topic cluster should be avoided from participating in the calculation of the similarity measure. Therefore, this article selects the main frequent ones in the candidate topic cluster. The feature word set constitutes the feature vector of the group, and the feature vector is used to represent the candidate topic cluster.

Definition 10. Cluster feature vectors. The feature map clustering capability aggregates input classification model into bunches. The input parameters within the same grouping are coupled collectively based on similarity. Feature set findings identify relevant features utilized as information during creating groupings. For the candidate topic cluster DU_i, the frequent 1-item sets of the DU_i cluster are mined, that is, the cluster feature vector constituting the cluster, which is recorded as $\overrightarrow{DU_i} = (u_{i1}, u_{i2}, ..., u_{in})$.

Definition 11. Cluster similarity matrix. Remember that the cluster feature vectors of two different candidate topic clusters DU_i and DU_j are $\overrightarrow{DU_i} = (u_{i1}, u_{i2}, ..., u_{in})$ and $\overrightarrow{DU_j} = (u_{j1}, u_{j2}, ..., u_{jm})$. The cluster semantic similarity matrix formed by the feature items of DU_i and DU_j is defined in Table 1.

Table 1. The cluster semantic similarity matrix of topic clusters DU_i and DU_j.

	u_{i1}	u_{i2}	...	u_{in-1}	u_n
u_{j1}	$\text{sim}(u_{j1}, u_{i1})$	$\text{sim}(u_{j1}, u_{i2})$	...	$\text{sim}(u_{j1}, u_{in-1})$	$\text{sim}(u_{j1}, u_{in})$
u_{j2}	$\text{sim}(u_{j2}, u_{i1})$	$\text{sim}(u_{j2}, u_{i2})$	...	$\text{sim}(u_{j2}, u_{in-1})$	$\text{sim}(u_{j2}, u_{in})$
...	...	...	...	...	...
u_{jm-1}	$\text{sim}(u_{jm-1}, u_{i1})$	$\text{sim}(u_{jm-1}, u_{i2})$	...	$\text{sim}(u_{jm-1}, u_{in-1})$	$\text{sim}(u_{jm-1}, u_{in})$
u_{jm}	$\text{sim}(u_{jm}, u_{i1})$	$\text{sim}(u_{jm}, u_{i2})$	...	$\text{sim}(u_{jm}, u_{in-1})$	$\text{sim}(u_{jm}, u_{in})$

Definition 12. Topic cluster semantic similarity. To avoid the noise of the cluster semantic similarity words, only the similarity of the K group feature of the semantic similarity in the similarity matrix is selected to calculate the similarity between the candidate topic clusters and record $\{\mathrm{sim}\,(u_i, u_j)_1\,, \mathrm{sim}\,(u_i, u_j)_2\,, ...,$ $\mathrm{sim}(u_i, u_j)_k\}$. The semantic similarity between candidate topic clusters is defined as:

$$\mathrm{sim}\,(DU_i, DU_j) = \frac{\sum\limits_{l=1}^{k} \mathrm{sim}(u_i,\, u_j)_l}{k}. \tag{5}$$

Based on the semantic similarity of candidate topic clusters, λ is used to represent the minimum similarity threshold between candidate topic clusters.

The minimum threshold of cluster semantic similarity is set when two groups are merged. μ represents the minimum number of collections expected to be obtained after topic clusters agglomerate. The topic cluster aggregation hierarchical clustering method's operation steps are as follows:

1) Extract each candidate topic cluster's feature vectors and calculate the semantic similarity of the candidate topic clusters.
2) Construct the semantic similarity matrix of candidate topic clusters. From the definition of cluster similarity, we can see $\mathrm{sim}\,(DU_i, DU_j) = \mathrm{sim}(DU_j, DU_i)$, that is, the similarity matrix is symmetric.
3) Select the largest inter-cluster similarity from the similarity matrix and record it as $\max\{\mathrm{sim}\,(DU_i, DU_j)\}$, if $\max\{\mathrm{sim}\,(DU_i, DU_j)\}\leq \lambda$, execute 6; otherwise, execute 4.
4) Since $\max\{\mathrm{sim}\,(DU_i, DU_j)\}>\lambda$, the similarity between DU_i and DU_j is relatively large, so the two clusters DU_i and DU_j are merged to form a new collection DU_i', and the original DU_j is deleted, and recalculate the cluster feature vector and update the semantic similarity matrix.
5) If the number of rows or columns of the semantic similarity matrix between clusters is less than or equal to the preset minimum number of collections μ, execute 6; otherwise, the clustering has not ended, and return to 3.
6) The condensed hierarchical clustering ends and the final topic clusters are obtained.

4. Result analysis and discussions

We evaluated the large transition in the discussion in Sec. 2 to assess the influence and feasibility of the clustering algorithm. This study screened the crawling bloggers and labeled data issues, resulting in 10 manually annotated topic groupings and a total of 13 356 microblogs. Table 2 depicts the topic tagging situation.

Table 2. Distribution of 10 topic categories manually labeled.

Serial number	Tag cluster	Cluster size
1	{Protecting the Diaoyu Islands, Diaoyu Islands}	1438
2	{Jingdong, e-commerce}	1427
3	{sea anemone}	2463
4	{Meteor Shower}	157
5	{Headshot brother, Zhou Kehua}	680
6	{London, Olympics}	2000
7	{Millet}	528
8	{Constellation, Horoscope}	3115
9	{Ye Shiwen}	799
10	{badminton}	749

Without loss of generality, this article uses two indicators, purity and F-measure value, to objectively evaluate the clustering effect. Generally speaking, the greater the purity of the clustering result, the better the clustering effect; the more significant the F-measure value of the clustering result, the better the clustering effect.

4.1. Parameter analysis of the clustering algorithm

The minimum cluster support θ directly affects the initial cluster, and the number of cluster features obtained in the feature extraction stage further affects the topic semantic membership calculation of the initial clusters of the microblogs in the overlapping parts of the groups and ultimately affects the separation effect of the initial sets. The experiments validate the efficiency of this strategy by analyzing actual microblog data. Topic recognition still confronts several issues deserving of additional exploration due to the uniqueness of hot subjects in clinical texts. To analyze the selection effect of the parameter θ, firstly, the clusters of the 10 manually labeled categories were randomly divided into 2 groups, each containing 5 manually labelled collections, which were marked as "#PartI" and "#PartII" (10 different groups were obtained randomly. #PartI and #PartII); Different cluster minimum support θ was chosen, and the influence of varying θ on the average value of the clustering result F-measure is shown in Fig. 2. This figure shows the distribution of the minimum support of different clusters on the F value.

According to the test results, when the minimum cluster support θ is 0.5 to 0.6, a better clustering effect can be obtained.

In agglomerative hierarchical clustering, λ is the minimum similarity threshold between candidate topic clusters. When the similarity between all groups is

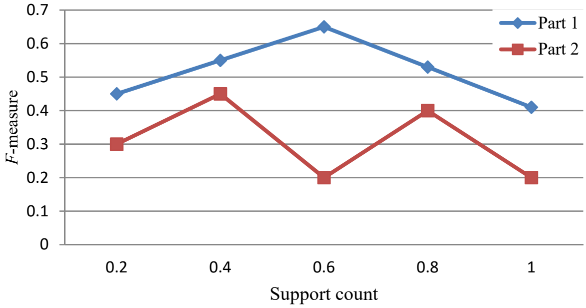

FIG. 2. Influence of the minimum support θ of different clusters on the F value.

smaller than λ, topic merging is terminated; in the experimental results, this can be obtained when λ is 0.6~0.7. The best clustering effect is shown in Fig. 3. This figure shows the influence of the minimum threshold of topic cluster similarity on the F value using cluster similarity.

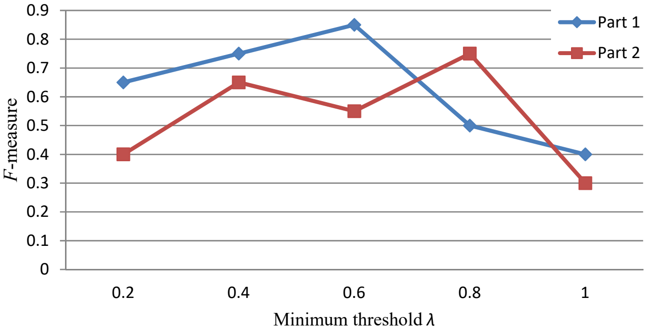

FIG. 3. Influence of the minimum threshold λ of topic cluster similarity on the F value.

4.2. Comparison of the effects of TS-FIHC and FIHC

Five groups of topic data, including 2, 4, 6, 8, and 10 annotated topics in artificially annotated cases, were extracted as test benchmarks and compared the purity and F-measure values. Comparisons of cluster F-scores and cluster purity of different topics are shown in Figs. 4 and 5, respectively. This study takes the FIHC algorithm's principle of "first clustering, then removing duplicates, and finally condensing" to tackle the challenges of microblog short text clustering, and presents a novel technique integrating time series frequent and semantic clustering time and semantics (TS-FIHC).

The TS-FIHC algorithm uses TS-FIHC-k to represent frequent k-item sets $(k = 1, 2, 3, 4)$.

It can be seen in the experimental results that the improved TS-FIHC method takes into account the text semantics, making the separation of the initial clus-

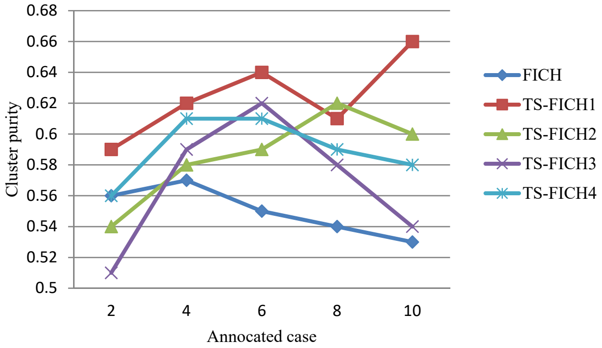

FIG. 4. Comparison of cluster purity of different topics.

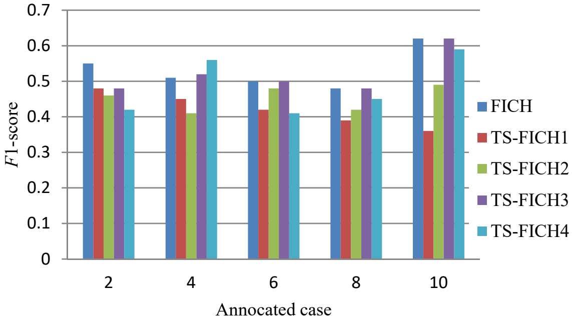

FIG. 5. Comparison of cluster F-scores of different topics.

ters and the merging of the candidate topic clusters more reasonable, and thus obtains a better purity and F-measure value than FIHC. On the other hand, the clustering effect of all TS-FIHC_1 is also better than that of TS-FIHC_k. At the same time, the TS-FIHC_1 algorithm can avoid the mining of frequent k-item sets, which significantly reduces the algorithm overhead.

5. Conclusion

This paper conducted research on hot topic detection based on clinical content, summarized the topic detection problem as a short text clustering problem, and proposed a systematic solution:

1) Using the timing characteristics of clinical topics, frequent feature words for clinical topics were proposed;
2) Aiming at the problem of microblog overlap between initial clusters, this paper proposes an initial cluster overlap reduction algorithm based on the division of semantic membership to overcome the short semantic expression and diversification of short microblog texts and topic ambiguity;

3) By defining the semantic similarity between initial clusters, a condensed hierarchical clustering method for microblog topics was presented, which can realize topic detection and tracking;
4) Through actual microblog data analysis, experiments verify the effectiveness of this method.

Due to the uniqueness of hot topics in clinical texts, topic detection still faces many problems worthy of further investigation:

1) The "HowNet" semantic database has a limited vocabulary, so the similarity between languages cannot change dynamically with different topics. Therefore, it can be further explored. For example, by studying the co-occurrence of clinical vocabulary and researching dynamic semantic similarity calculation methods based on vocabulary co-occurrence;
2) Not all microblogs discuss only one topic. Some microblogs may connect multiple cases in a series, that is, a microblog. In addition, blogs may belong to various issues at the same time. Therefore, it is worthwhile to study further some soft clustering methods to extend microblog topic detection;
3) The topic clusters obtained through clustering usually contain more potential information, but this implicit information is not easy to be discovered directly.

The potential of the research for the future is that it will be extremely useful in future investigations. As a result, visualizations may be studied further in order to highlight the links between subject clusters and intuitively aid topic detection.

References

1. A. Ejnioui, C.E. Otero, A.A. Qureshi, Formal semantics of interactions in sequence diagrams for embedded software, [in:] *2013 IEEE Conference on Open Systems (ICOS)*, 2–4 December, Kuching, Malaysia, 2013, pp. 106–111, doi: 10.1109/ICOS.2013.6735057.

2. J. Mansouri, B. Seddik, S. Gazzah, T. Chateau, Coarse localization using space-time and semantic-context representations of geo-referenced video sequences, [in:] *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 10–13 November, Orleans, France, 2015, pp. 355–359, doi: 10.1109/IPTA.2015.7367165.

3. L. Yao *et al.*, HSD: Hybrid MARTE sequence diagram, [in:] *2015 IEEE International Conference on Software Quality, Reliability and Security*, 3–5 August, Vancouver, BC, Canada, 2015, pp. 189–194, doi: 10.1109/QRS.2015.35.

4. Y. Fang *et al.*, Salient object detection by spatiotemporal and semantic features in real-time video processing systems, *IEEE Transactions on Industrial Electronics*, **67**(11): 9893–9903, 2020, doi: 10.1109/TIE.2019.2956418.

5. A. Shobanadevi *et al.*, Internet of things-based data hiding scheme for wireless communication, *Wireless Communications and Mobile Computing*, **2022**: Article ID 6997190, 8 pages, 2022, doi: 10.1155/2022/6997190.

6. Z. Xiang, S. Zhi-qing, ASM semantic modeling and checking for sequence diagram, [in:] *2009 Fifth International Conference on Natural Computation*, 14–16 August, Tianjian, China, pp. 527–530, 2009, doi: 10.1109/ICNC.2009.218.
7. A. Kishor, C. Chakraborty, W. Jeberson, Reinforcement learning for medical information processing over heterogeneous networks, *Multimedia Tools and Applications*, **80**: 23983–24004, 2021, doi: 10.1007/s11042-021-10840-0.
8. M. Soni, G. Dhiman, B.S. Rajput, R. Patel, N.K. Tejra, Energy-effective and secure data transfer scheme for mobile nodes in smart city applications, *Wireless Personal Communications*, **127**: 2041–2061, 2021, doi: 10.1007/s11277-021-08767-8.
9. C. Jia, M.B. Carson, X. Wang, J. Yu, Concept decompositions for short text clustering by identifying word communities, *Pattern Recognition*, **76**: 691–703, 2018, doi: 10.1016/j.patcog.2017.09.045.
10. A. Cioppa, M.V. Droogenbroeck, M. Braham, Real-time semantic background subtraction, [in:] *2020 IEEE International Conference on Image Processing (ICIP)*, 25–28 October, Abu Dhabi, UAE, pp. 3214–3218, 2020, doi: 10.1109/ICIP40778.2020.9190838.
11. T. Wang, K. Jia, M. Yao, Sequence matching with discriminative binary features for robust and fast light-rail localization at high frame rate, [in:] *2020 IEEE International Conference on Big Data (Big Data)*, 10–13 December, Atlanta, GA, USA, pp. 1266–1272, 2020, doi: 10.1109/BigData50022.2020.9378494.
12. N.H. Kirk, K. Ramírez-Amaro, E. Dean-León, M. Saveriano, G. Cheng, Online prediction of activities with structure: Exploiting contextual associations and sequences, [in:] *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, 3–5 November, Seoul, South Korea, pp. 744–749, 2015, doi: 10.1109/HUMANOIDS.2015.7363453.
13. Q. Li, B. Tian, M. Zhang, An event sequence based method for audio scene analysis, [in:] *2011 4th IEEE International Conference on Broadband Network and Multimedia Technology*, 28–30 October, Shenzen, China, pp. 255–259, 2011, doi: 10.1109/ICB NMT.2011.6155936.
14. S. Han, Z. Xi, Dynamic scene semantics SLAM based on semantic segmentation, *IEEE Access*, **8**: 43563–43570, 2020, doi: 10.1109/ACCESS.2020.2977684.
15. C. Li, H. Xiao, K. Tateno, F. Tombari, N. Navab, G.D. Hager, Incremental scene understanding on dense SLAM, [in:] *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9–14 October, Daejeon, South Korea, pp. 574–581, 2016, doi: 10.1109/IROS.2016.7759111.
16. M. Siam, A. Kendall, M. Jagersand, Video class agnostic segmentation benchmark for autonomous driving, [in:] *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 20–25 June, on-line, pp. 2819–2828, 2021, doi: 10.1109/CVPRW53098.2021.00317.
17. J. Yang, F. Wang, J. Yang, Dynamic gesture recognition using LBRCN combined with attention mechanism, [in:] *2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT)*, 11–13 June, Changsha, China, pp. 466–469, 2021, doi: 10.1109/ISCIPT53667.2021.00100.
18. X. Li, Y. Tian, F. Zhang, S. Quan, Y. Xu, Object detection in the context of mobile augmented reality, [in:] *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 9–13 November, Porto de Galinhas, Brazil, pp. 156–163, 2020, doi: 10.1109/ISMAR50242.2020.00037.

19. C. Chakraborty, Performance analysis of compression techniques for chronic wound image transmission under smartphone-enabled tele-wound network, *International Journal of E-Health and Medical Communications,* **10**(2): 1–20, 2019, doi: 10.4018/ijehmc.2019040101.

20. K.N. Mishra, C. Chakraborty, A novel approach toward enhancing the quality of life in smart cities using clouds and IoT-based technologies, [in:] *Digital Twin Technologies and Smart Cities*, Springer, pp. 19–35, 2019, doi: 10.1007/978-3-030-18732-3_2.

21. S.K. Narayanasamy, K. Srinivasan, Y.-C. Hu, S.K. Masilamani, K.-Y. Huang, A contemporary review on utilizing semantic web technologies in healthcare, virtual communities, and ontology-based information processing systems, *Electronics*, **11**(3): 453, 2022, doi: 10.3390/electronics11030453.

22. A. Kishor, W. Jeberson, Diagnosis of heart disease using internet of things and machine learning algorithms, [in:] *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, Vol. 203, pp. 691–702, Springer Singapore, 2021, doi: 10.1007/978-981-16-0733-2_49.

23. C. Friedrich, A. Lechler, A. Verl, A planning system for generating manipulation sequences for the automation of maintenance tasks, [in:] *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, 21–25 August, Fort Worth, Texas, pp. 843–848, 2016, doi: 10.1109/COASE.2016.7743489.

Received February 2, 2022; revised version May 11, 2022;
accepted July 31, 2022; published online May 29, 2024.