Sub-topic Segmentation in Multi-document

Yun Xiaoyan, Teng Wei
School of software
Liaoning University of Science and Technology
Anshan, China
e-mail: yxy19800725@163.com

***Abstract*—The similar sentences in multi-document set are combined into one class, and each class is one sub-topic. Describing the sub-topics from the perspective of understanding makes the multi-document summarization become the one with greater coverage and less redundancy. This paper presents a sub-topic segmentation method based on maximum tree algorithm. And based on sentences similarity matrix, maximum tree is calculated, as well as the sub-topic segmentation is realized through the analysis of the different communities for the sub-topic. The experiment shows that the method achieves the desired result.**

Keywords-Multi-document Summarization; Sub-topic Segmentation;Maximum tree algorithm

I. INTRODUCTION

With the rapid development of the Internet, thousands of WebPages with the same theme make it difficult for people to get rapid access to information. So there is an urgent need for similar document processing and extraction tools to facilitate reading.Multi-document summarization is aimed to solve the problem.Multi-document summarization can be applied to the processing module of the new generation of search engine system (QA) when it returns to the answer, and can also be applied to the monitoring and tracking of the topics, etc..

Multi document summarization is a text-based natural language processing technology which can refine a plurality of text information into a single text according to the

compression ratio.Multi document summarization is the same subject under a plurality of text information according to the compression ratio of the main refining for a text-based natural language processing technology.

The multiple document summarization researches, according to the different methods adopted, can be roughly divided into the following categories: Multiple document summarization system based on a single document summarization [1, 2]; multi document summarization system based on information extraction technology [3,4];document collection features of bilingual document abstract system [5,6,7].

II. SUB TOPIC SEGMENTATION BASED ON MAXIMUM TREE METHOD

A collection of documents refers to the collection of different documents with the same subject, which is characterized with much common information between documents, but with its own emphasis. The same or similar sentences in a collection of documents are classified into one class through clustering method, and each class is the sub-topic of the collection of documents [8].

The sub-topics gained by means of sentence clustering, are expressing the original document collections' information with different meanings in a parallel way, while the sentences with the same meaning are covered by the same sub-topic as a whole. It actually strengthens the information with different meanings and weakens the information with the same meaning. The abstracts generated by means of the sub-topics, will cover more information with less redundancy.

This paper presents a method based on maximum tree method sub topic segmentation method. The first computing sentence similarity, constructs a fuzzy similarity matrix, and then using the maximum tree method for fuzzy clustering, finally the biggest tree generating partitions document child theme.

A. sentence similarity computing

Sentence similarity computing can reflect the local topic information fitting degree. In order to better characterize the sentence relevance, this paper adopts the methods based on semantic dictionary. By means of semantic lexicon to sentences in the vocabulary of deep understanding, to calculate the semantic distance between words is then transformed into a sentence similarity value formula.

The overall train of thought is as follows: the word is mapped to the 《Tongyici Cilin semantic space》 [9], obtain the corresponding semantic encoding, then the word similarity computation. Specific methods see Ref. [10].

The first document preprocessing, after stop words, clauses after filtering high-frequency words, low-frequency words, word segmentation. Is a combination of the words in linear sequence; then the semantic coding extraction; and then the word semantic coding similarity calculation leads to final sentence similarity value.

In Multiple document collection $D = \{d_i \mid i = 1, 2, \cdots n\}$, each document is represented as a collection of text unit. In this paper the text unit mentioned is the sentence, and text can be expressed as $d_i = \{x_i | k = 1, 2, \cdots m\}$, x_i expressed as a sentence. Matrix represents similarity relationship between sentence and sentence in a multiple document collection. R_{ii} Expression x_i and its correlation, recorded as 1 in the

matrix, R_{ij} expression the correlation of sentences x_i and x_i Due to $R_{ij} = R_{ji}$, only the left lower matrix calculation is taken into consideration.

$$Q = \begin{Bmatrix} 1 & & & & \\ R_{21} & 1 & & & \\ R_{31} & \cdots & 1 & & \\ R_{n-1,1} & \cdots & \cdots & 1 & \\ R_{n,1} & \cdots & \cdots & R_{n,n-1} & 1 \end{Bmatrix}$$

Figure 1. Sentence similarity between sentences in a collection of documents

Sentence similarity between sentences in a collection of documents is as shown in Figure 1 .

B. sub topic segmentation based on maximum tree method

Relationship chart can clearly show the content distribution between cases. If the graph is a connected graph, it shows that the content of the document is relatively concentrated, with all the sentences around only one theme. If the graph is not a connected graph, it shows that the document content involves several aspects. And this kind of graph can be described in the form of trees with every tree representative of a sub topic.

C. maximal tree clustering method

The Maximal tree makes use of the related theory in the graph theory. After obtaining fuzzy similar matrix, it shows fuzzy similarity relation through the graph, and then it concludes with the maximum fuzzy support tree [11] (referred to as the maximum tree).

The points in the graph represent the sentences. The connection between the nodes and nodes represents the two sentences are related, and no connection means the sentences are not related. The weights show the degree of the correlation between the 2 sentences.

The algorithm is described below:

Step1: calculate the sentence similarity, and a fuzzy similarity matrix $Q = Sim(R_{ij})$.

Step 2: $Sim(R_{ij})$ arranged the weights from big to small recorded as $lw = b_1 > b_2 > \cdots > b_m$

Step3: take all sentences as vertices, connect the vertices whose similarity are b_1 ,and in the corresponding segment mark them as the weights of the edges (not intersecting line). If a loop appears when the vertices are connected, do not connect the edge.

Step4: to $b_2 > b_2 > \cdots > b_\kappa (k \le m)$ repeat Step3 until the whole sentence is connected so as to obtain the maximum tree. So the tree, the edge weights, is called the maximum tree $SubPN$.

The method prefers sentence weights and connects the most similar sentences together. Unlike other general hierarchical clustering methods in which objects are treated equally, it is more conducive to summarize the abstract from the perspective of sub-topic.

D. *Partitioning based on Maximum Tree*

The sub-topic is made up with the sentences that are close related with each other internally but seemed loosely related externally, and the internal sentence similarity is greater than the external sentence similarity. Each sub topic is a uniting structure. The size and the number of potential uniting structures depend on the maximum tree $SubPN$ topology structure and the edge weights distribution.

The partitioning of maximum tree makes use of depth-first law division to divide the similarities of the connecting edges, that is, to search the edge of the maximum weights which is directly connected to the current node.

Algorithm description:

Step 1: generate the maximum tree $SubPN$, $SubPN(PN = (v, e, b))$, e edge, b weights. Node lw , values in descending order, preserved in the list lw .

Step 2: initialize uniting structure set $CSubset = \{s_j\}$ (j=1)

Step 3: according lw , from $SubPN$ put the current maximum b_i edge e into the uniting structure s_j as a seed edge lw, and remove the value of b.

Step 4: examine s_j which is directly connected with b_j , if $b_j < b_i$, add e to s_j , $b_i = b_j$, e removed from lw the value b_j , jump to jump to the Step 4, otherwise Step 5.

Step 5: Separate $SubPN$ out of s_j as the j uniting structure and put it into $CSubset$ make j=j+1, and then jump to steps 3.

Step 6: after dividing all the uniting units, stop dividing process, and output results of division.

III. EXPERIMENTS AND ANALYSIS

In general, two indexes are used to evaluate clustering results, clustering precision and clustering recall rate. Clustering precision reflects the level of the class merged by similar text unit and the different text unit, The higher the clustering precision is , the more concentrated the contents are. Cluster recall rate reflects the level of the class merged by similar text units. The higher the clustering recall rate is, the more concentrated are the similar text units and the less chance is there for them to be divided into other classes. Clustering precision reflects the ability to distinguish between different subjects, and clustering recall rate reflects in the same the ability to recognize the same subject.

In this paper the accurate rate of each category P_i in a document collection see Formula 1, recall rate R_i see formula 2

$$P_i = \frac{SetA_i \cap SetB_i}{SetB_i} \quad \text{(formula 1)}$$

$$R_i = \frac{SetA_i \cap SetB_i}{SetA_i} \quad \text{(formula 2)}$$

Among them $SetA_i = \max_j (SetA_i \cap SetA_j)$, $SetB_i$ belong to the categories generated by machine, $SetA_i$ belongs to the standard category, At this moment, the standard category represents the machine type I and the category that shares the most same sentences with the standard category $SetA_i = \max_j (SetA_i \cap SetA_j)$, i.e.

With the method above, more than 10 document collections, each of which contains 5-6 documents are experimented and the experimental results are shown in the following table:

TABLE I. EXPERIMENTAL RESULTS

The document collection number	Hierarchical clustering method	Based on the maximum tree sub topic segmentation method
1	70%	70%
2	70%	73%
3	57%	61%
4	65%	68%
5	70%	73%
6	72%	75%
7	70%	72%
8	65%	56%
9	68%	70%
10	69%	69%

Through the experiment, the following conclusions can be arrived at:

The hierarchical clustering algorithm can not make a very good effect because it is not a backtracking algorithm, which means if a sentence is divided into a class, it can not be changed again; besides, the hierarchical clustering results need a threshold set by people, which may make a great difference to the results, and that means the users are required to have basic knowledge about the objects to be clustered, which makes the practical operation inconvenient.

When adopting the maximum tree fuzzy clustering method, there is no need to set any parameters, and the clustering process is more automatic and universally adaptive, while clustering results are stable.

IV. CONCLUSIONS

This paper presents a sub topic segmentation method based on maximum tree method. The feature of the method lies in its comprehensiveness and conciseness of the abstract whose content can simultaneously gain the balance between the coverage and the redundancy.

REFERENCES

[1] R.Radev, Hongyan Jing, Malgorzata Budzikowska,Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In ANLP/NAACL 2000 Workshop, April 2000:21-29

[2] Chin-Yew Lin,Eduarad Hovy .From Single to Multi-document Summarization: A Prototype system and its Evaluation. In proceeding of the 40tth anniversary meeting of the association for computational linguistics(ACL-02), Philadephia,USA,2002:25-34

[3] Dragomir R.radev,Kathleen R.Mckeovwn. Generating Natural Languages Summaries from multiple on-line Sources. Computational Linguistics.1998,24(3):21-29

[4] Sanda Harabagiu, Steven Maiorano. Multi-document summarization with GISTexter Proceedings of the third LREC Conference 2002(LREC 2002),June 2002,Canary Islands,Spain

[5] E.Filatova,V.Hatzivassiloglou,Event-based Extractive Summarization.In the proceedings of ACL Workshop on Summarization,Barcelona,Span,July 2004

[6] Endre Boros,Paul B.Kantor,David J.Neu.A Clustering Based Approach to Creating Multi-Document Summaries.In Proceedings of the 24th Annual International ACM SIGIR Conference on Reseach and Development in information retrieval ,New Orleans,LA,2001

[7] Pascale Fumg,Grace Ngai.Combining Optimal Clustering and Hidden Markov Model for Extractive Summarization.Proceedings of the ACL 2003 workshop on multilingual summarization and question answering .2003:21-28

[8] Qin bing,liu ting,gao ye Identification of logical topic of multi-document set

[9] Mei jiaju Corpus annotation[M] Shanghai Lexicographical Publishing Bureau, 1983

[10] Tang ming zhu,Zhang yuan ping,Yang jia, Method of text fuzzy clustering based on concept similarity Science technology and engineering Vol 7,no5,