Used Hierarchical Topic to Generate multi-document Automatic Summarization

XU Yong-Dong, QUAN Guang-Ri, ZHANG Ting-Bin, WANG Ya-Dong

School of Computer Science of Technology
Harbin Institute of Technology at WeiHai
Wei Hai China
ydxu@insun.hit.edu.cn

***Abstract*—A concept of hierarchical topic is proposed for multi-document automatic summarization task, which used multi-layer topic tree structure to represent the text set. Each node in the topic tree represent specific topic and contains multiple similar sentences in the text set. The hierarchical topic structure may describe accurately the similarity between sentences at different levels of granularity. Therefore it can reflect the real content of the text set than single layer topic set. And can be used to find the important sentences in the important topic which can compose the summary of the text set. Concretely, a series of algorithms including building hierarchical topic tree, key sentences extraction based on hierarchical topic tree and summarization generation are proposed. The capability of summarization system is testified by sets of experiments and shows good result.**

Keywords-Natural lauguage processing; Multiple document automatic summarization; Hierarchical topic;

I. INTRODUCTION

Automatic text summarization researching can be traced back to Luhu's work in 1952. Researchers mainly focus on extracting and presenting the most important content from single text. However in recent, with the rapid growth of the Internet, the massive amounts of information make it more difficult to efficiently access the usable information. Thus, the ability to automatically compress the information covering multiple documents and present the summary to the users would help to solve this problem and in fact, has received a great deal of attention in recent research.

Because the multi-document automatic summarization system should have the ability to identify the general topic of the entire set rather than the topic of each text in the set [1], the technique that combines such heterogeneity information in multiple documents into structured data for further processing has attracted researchers' interest. Some researchers used text units clustering technique, including paragraphs-level clustering [2][3], sentences-level clustering [4][5], to achieve information fusion. The idea of method is built on the hypothesis that because these articles describe the same subject, the similar information from different articles can be regarded as the important content of set of texts. Thus topic identification is an important technology in multi-document automatic summarization. One important idea of the topic region identification is to gather text units into a cluster by clustering method[6]. Determining the range of the topic is a crucial issue of topic identification task. If the range is wide, two units with lower similarity could belong to the same topic, otherwise, the units with high similarity might belong to the different topics. In multi-document environment, a topic might include several sub-topics. So the traditional single layer topic partition method could hardly obtain appropriate topics set even if analysed manually by human experts. Thus we proposed the concept of hierarchical topic: used multi-layer topic structure to take place of the traditional single layer topic structure, and proposes a sentences selection algorithm based on hierarchical topic tree. The contents of this paper are as follows: section 2 and 3 introduced the concept of hierarchic topic and the algorithms of hierarchical topic identification. Section 4 introduced experimental analysis and section 5 showed the conclusion.

II. HIERARCHICAL TOPIC

A document set *D* is comprised of a series of topics. The units in each topic are similar to each other and the units from the different sub-topic should be dissimilar.

In traditional methods, the text set was always divided into single layer topic set. In fact, most text sets, especially the text set which has incoherence content, can hardly be divided into appropriate sub-topics by a proper threshold in which each sub-topic was divided "big" enough to contain all sentences that describe a whole event and at the same time "small" enough not to contain any other information.

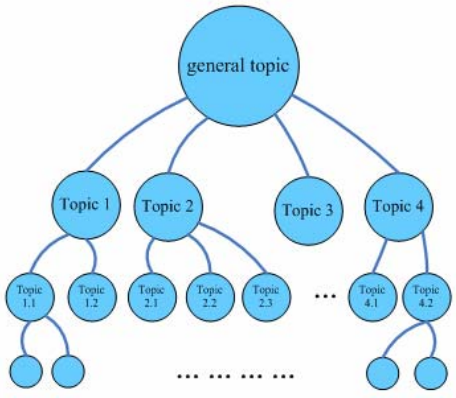

Figure 1. Hierarchical Topic Tree

Thus, the topic structure of a text set should be a hierarchical tree structure. As shown in Figure 1, each node of the hierarchical topic tree represented a topic. The root node was the general topic of the text set. The leaf nodes represented the fine grain sub-topics which can no longer be divided. The solid lines represented the inheritance relationship between the upper and the lower nodes.

978-0-7695-4353-6/11 $26.00 © 2011 IEEE
DOI 10.1109/ICICTA.2011.84

III. HIERARCHICAL TOPIC IDENTIFICATION BASED ON THE TEXT UNITS CLUSTERING ALGORITHM

A. Main Idea

Given the text units set P, $P=\{p_1,p_2,...p_n\}$, the result of hierarchical topic identification method is a tree structure, like the Figure 1 shows. This paper used the hierarchical agglomerative clustering algorithm to establish the hierarchical topic tree. Firstly, each unit was an independent cluster. Then the most similar clusters were merged. This process continued until all the units were merged into one cluster, and establish a clustering tree finally. In general, at least one of intermediate results of this processing could correspond to the proper topic division. As for the k-layer hierarchical topic tree, we supposed that each layer can accord with one intermediate result in the clustering tree. In another word, through determining suitable threshold value Φ_1 to suspend the clustering process, an intermediate result in clustering tree or a layer in topic tree can be obtained. The algorithm restarts once more, it will obtain the upper layer of the hierarchical topic tree by another threshold value Φ_2, this process continue until there is only one topic in the topic set. In this algorithm, sentences similarity computation and determination of the multi-threshold value of clustering are the key questions. Section 3.2 and 3.3 will give out the detailed description of the solution of above two questions.

B. The Similarity Between sentences

The traditional full text similarity computation methods seems insufficiency to calculate the similarity between text units. So we used a multiple features fusion method to calculate sentences similarity [7]. The similarity of text units are determined by multiple features extracted from these units, including word frequency vectors, part of speech vectors and semantic vectors. These features are fused automatically by logistic regression analysis model:

$$Y=\frac{e^{(\alpha+\beta_1 x_1+...+\beta_k x_k)}}{1+e^{(\alpha+\beta_1 x_1+...+\beta_k x_k)}} \tag{1}$$

Where α, β are logistic regression coefficient which are fitted automatically by training corpus. The response variable Y is the probability that pair of units are similar, and the explanatory variables $X=(X_1,X_2,...X_k)$ are a group of values, where k is the number of features, $X_i=(X_{i1},X_{i2},...X_{in})$ is the partial similarity calculated by ith feature, x_{ij} represents the partial similarity of the jth pair of units. For each feature, x_{ij} are calculated by using the cosine angle between the two vectors ω , γ that correspond to the two vector features extracted from the jth pair of paragraphs as follows:

$$sim(\overline{\omega},\overline{\gamma})=\frac{\sum_{i=1}^{N}\omega_i\gamma_i}{\sqrt{\sum_{i=1}^{N}\varpi_i^2\cdot\sum_{i=1}^{N}\gamma_i^2}} \tag{2}$$

C. Multi-threshold Automatic Identification

The role of threshold Φ is to suspend iterative process of clustering algorithm. Φ must both sufficiently "small" enough to accurately partition all sub-topics, and sufficiently "big" enough to not partition a whole sub-topics. In previous researches, García proposed an automatic threshold identification method of the hierarchical clustering algorithm[8]. A lot of experiments had shown that the algorithm were very effective for non-convex clusters. So this paper adopted this idea and used the clustering entropy ε to determine the thresholds. ε is the square-error sum of the cluster set, and when the value of clustering entropy was minimum algorithm can get the most reasonable cluster set, the formula of the square-error sum is :

$$\varepsilon(\chi)=\Lambda+\Gamma \tag{3}$$

χ is the cluster set. Λ and Γ separately represent the square-error sum of the internal clusters and the square-error sum between clusters which are calculated as follow:

$$\Lambda=\sum_{j=1}^{k}\sum_{i=1}^{n_j}e(p_i^{(j)},p_0^{(j)})=\sum_{j=1}^{k}\sum_{i=1}^{n_j}d(p_i^{(j)},p_0^{(j)})^2 \tag{4}$$

$$\Gamma=\sum_{j=1}^{k}e(p_0^{(j)},c_0) \tag{5}$$

n_j is the size of cluster C_j, $d(p_i,p_j)$ is the distance between elements. $p_0^{(j)}$ is cluster centroid. c_0 is the overall centroid. k is the number of clusters.

Λ should meet the following conditions: 1. single cluster without any contribution to Λ, 2. In the initial stage the initial value of Λ was 0, because there was only one cluster, 3. The Λ has the biggest value when all elements gathered together, 4. In the clustering processing, Λ monotonically increasing.

Γ should meet the follow conditions: 1. c_0 must same as the cluster centroid that contained all the elements; 2. Γ had the biggest value at the initial stage of the clustering. 3. The value of Γ was the minimum when all elements gathered as a cluster. 4. Γ monotonically decreasing in the clustering processing.

We calculated the similarity of units with multi-feature fusion-based approach mentioned in section 3.2, it was difficult to get the cluster centroid because we do not use a single vector to represent unit. So when given category C_i, we optioned a unit from the initial two units in the cluster as the cluster centroid. If C_i and C_j merge into a new cluster C_{ij} in the clustering process, C_i was generated in step i, C_j was generated in step j, and $i<j$, then the cluster centroid of C_i is the new cluster centroid.

Our method also met the four conditions mentioned above. The conditions 1-3 were clearly true, and we prove the last one as follows:

Supposed two clusters C_i and C_j merged into a new cluster C_{ij} in the clustering processing, α and β separately represented the internal square-error sum before and after merging, if C_{ij} and C_i had the same cluster centroid, then α and β could be calculated as following:

$$\alpha=\sum_{k=1}^{n_i}d(p_k^{(i)},p_0^{(i)})^2+\sum_{k=1}^{n_j}d(p_k^{(j)},p_0^{(j)})^2 \tag{6}$$

$$\beta=\sum_{k=1}^{n_i}d(p_k^{(i)},p_0^{(i)})^2+\sum_{k=1}^{n_j}d(p_k^{(j)},p_0^{(j)})^2 \tag{7}$$

So if $\beta-\alpha>0$, condition 4 could be proved.

$$\begin{aligned}\beta-\alpha&=\sum_{k=1}^{n_j}d(p_k^{(j)},p_0^{(i)})^2-\sum_{k=1}^{n_j}d(p_k^{(j)},p_0^{(j)})^2\\&=\sum_{k=1}^{n_j}[d(p_k^{(j)},p_0^{(i)})^2-d(p_k^{(j)},p_0^{(j)})^2]\end{aligned}\tag{8}$$

As a element of the category C_j, the distance from $p_k^{(j)}$ to cluster centroid C_i was bigger than $p_k^{(j)}$ to cluster centroid C_j, (if not, $p_k^{(j)}$ would be a element of cluster C_i, not the element of cluster C_j), so $d(p_k^{(j)},p_0^{(i)})^2-d(p_k^{(j)},p_0^{(j)})^2>0$ and $\beta-\alpha>0$, prove complete.

Like the internal square-error sum, we can't get the central vector of the cluster, so we optioned a unit from the initial two units that initially formed, and calculate external square-error sum as follows:

$$\Gamma=\sum_{j=1}^{k}(\frac{1}{n_j}\sum_{i=1}^{n_j}d(p_i^{(j)},p_0^{(1)})^2)\tag{9}$$

$p_0^{(1)}$ was the cluster centroid at the beginning of the clustering, Γ also met 4 conditions we had mentioned above, The conditions 1-3 were clearly, and we prove the last one as follows:

Suppose two categories C_i and C_j merged into the new category C_{ij} in the clustering processing, α and β separately presented the external square-error sum before and after merging, α and β can be calculated as following:

$$\alpha=\frac{1}{n_i}\sum_{k=1}^{n_i}d(p_k^{(i)},p_0^{(1)})^2+\frac{1}{n_j}\sum_{k=1}^{n_j}d(p_k^{(j)},p_0^{(1)})^2\tag{10}$$

$$\begin{aligned}\beta&=\frac{1}{n_i+n_j}\left[\sum_{k=1}^{n_i}d(p_k^{(i)},p_0^{(i)})^2+\sum_{k=1}^{n_j}d(p_k^{(j)},p_0^{(i)})^2\right]\\&=\frac{1}{n_i+n_j}\sum_{k=1}^{n_i}d(p_k^{(i)},p_0^{(i)})^2+\frac{1}{n_i+n_j}\sum_{k=1}^{n_j}d(p_k^{(j)},p_0^{(i)})^2\end{aligned}$$

Comparing the previous one and the latter one in formula α and β, obviously $\beta<\alpha$, so Γ was monotonically decreasing, prove complete.

Because in clustering process the internal square-error sum is monotonically decreasing and the external square-error sum is monotonically increasing, thus the clustering entropy is minimum value when the two trends met, and algorithm can get a reasonable clustering result. Let Φ_0 is the minimum entropy, all threshold Φ that met condition $\Phi/\Phi_0<1.2$ can be added into the final threshold vector.

D. Hierarchical clustering algorithm

We used complete-link hierarchical clustering algorithm to build hierarchical topic tree. Let Φ is the threshold vector, $\Phi=\{\Phi_1, \Phi_2,\ldots \Phi_K\}$, k is the depth of the topic tree, $\Phi_1,\ldots \Phi_K$sorted by decreasing value. Let S is the initial cluster collection, $S=\{s_1,s_2,\ldots,s_n\}$, each element in S is a 2-tuple $s(c,T)$, c is the content of unit, T is the cluster vector, which contains k elements. In each iteration process, the algorithm traversed the cluster set and merged the most similar clusters. When the similarities among all clusters were less than the threshold value Φ_1, the first element of the cluster vector T was updated. Clustering algorithm continued until S contained only one element. The formal description of Multi-threshold clustering algorithm is given as below:

```
begin
l = 1.
while l ≤k begin
   while begin
      traverse the cluster set S, calculate the similarity between
      any two clusters basing formula 3-36,
      let D=max{sim(s_i,s_j)|s_i,s_j∈S}, if D> Φ_l, then merge s_i and
      s_j to one cluster, and update the type collection S. else break
   end while
l = l + 1
end while
output hierarchical topic tree.
```

IV. Summary Sentences Extraction

We proposed two-step sentences selection algorithm, respectively calculate the importance of one sentence in the document and the whole set.

A. Initializing Weight of Each Sentence in Text Set

We firstly assigned an initial weight of each sentence in text set by using several types of metrics, including sentence position (SP), sentence length (SL) and term weight (TW). In above metrics, the score of the sentence is:

$$Score_{SP}(s_i^{(j)}) = 1/(i+j)\tag{11}$$

$$\begin{aligned}Score_{SL}(s) &= 0\ (if\ L_s>L_{min})\\&=(L_s-L_{min})\ /\ L_{min}\ (otherwise)\end{aligned}\tag{12}$$

$$Score_{TW}(s_i)=\frac{1}{|s_i|}\sum_{w\in s_i}tf\cdot idf(w)\tag{13}$$

In formula 11, $s_i^{(j)}$is the ith sentence in the jth paragraph, in formula 12, L_{min} *is* a penalty threshold value which we set 10 in this paper. In formula 13, $tf \bullet idf(w)$ is the $tf \bullet idf$ score of word w in the sentence s which can be obtained by training process:

$$tf\cdot idf(w)=\frac{tf(w)-1}{tf(w)}\log\frac{DN}{df(w)}\tag{14}$$

DN is the number of documents in all sets. The total score for a sentence is determined using a scoring function:

$$Total-Score(s)=\sum\alpha_i Score_i(s)\quad(\sum\alpha_i=1)\tag{15}$$

Each parameter α_i was set manually with α_{sp} =0.507, α_{SL} =0.005 and α_{TW} =0.488.

B. Top-down Reciprocal Sentences Weighted Algorithm

Input: initial weight list of sentences, Hierarchical topic tree with depth k, Output: weight list of sentence nodes

1. make the weights of all nodes zero, variable l=1
2. while l<k, do
3. calculate the final weight of all sentences as follow:

$$W(i)=a_i+\frac{1}{n}\sum_{j=1}^{n}(a_j+f_j)$$

Where f_j is the similarity between sentence i and j calculated by hierarchical topic tree: f_j= $(k-m+1)/k$, m is the depth of first layer that sentence i and j are not belonged to the same topic. a_i and a_j are initial values of sentences i and j.

4. l=l+1
5. end while

V. SYSTEM EVALUATION

An ideal summarization system must possess at least three fundamental properties: 1) *Accuracy*: The ability to find and extract the important information across documents. 2) *Concision*: The ability to minimize redundancy between candidate sentences. 3) *Cohesion:* The summary should be readable. So, evaluation method should have the ability of evaluating above properties: evaluate how similar automatic summaries are to the "gold standard". A problem of this method is that there may exist multiple same or similar sentences to standard summaries and each of these can be regarded as an acceptable summary sentence. Thus, it will underestimate the quality of summarization by only matching it with the standard summary.

In this paper, we resolve above problem by labeling each sentence a fuzzy value in training corpus. The sentence that can replace standard summary sentence in source text is called candidate summary sentence. According to the similarity between candidate summary sentence and standard summary sentence, each candidate summary sentence is labeled a value ranged from 0 to 1. Standard summary sentence and related all candidate summary sentences constitute a congeneric cluster. Two sentences in congeneric clustering are named congeneric summary sentences.

According to above-mentioned definitions, we can evaluate the quality of summarization system by following three criterions: precise, redundancy and synthesis quality

$$precise = (\sum_{i=1}^{k_1} \omega_i) / K$$

$$redundancy = (\sum_{i=1}^{k_1} (\sum_{j=i+1}^{k_1} \phi(s_i, s_j))) / K$$

$$Synthesis = precise - redundancy$$

Where K is number of all sentences in summarization. $k_1 = K \cap k_0$, k_0 is number of all sentences in standard summarization. w_i is weight of ith candidate summary sentence. $\Phi(s_i,s_j)$ is a binary discriminant function. $\Phi(s_i,s_j) = 1$ if s_i, s_j are congeneric summary sentences. $\Phi(s_i,s_j) = 0$ else.

TABLE I. THE RESULT OF OUR METHOD

	5 sentences	10 sentences	20 sentences
precise	70.58%	75.67%	67.5%
redundancy	0%	10%	11.33%
Synthesis quality	70.58%	65.67%	56.17%

TABLE II. THE RESULT OF BASELINE METHOD

	5 sentences	10 sentences	20 sentences
precise	54.3%	50.8%	65.85%
redundancy	0%	8%	19.33%
Synthesis quality	54.3%	42.8%	46.52%

We used a baseline system which likes our method, but adopt single layer topic structure to weight each sentence, to validate our system. For constructing training corpus, we download 120 newswire documents from Internet and cluster them into 15 topic-relative subsets manually. For each subset, we extracted a standard summary and congeneric cluster of each summary sentence. The experimental result by respectively comparing the output of two summarization methods with standard summary is showed in table I and II.

The result of experiment shows that the quality of summary produced by our method is better than the one produced by baseline method for all output length of summary.

CONCLUSION

In this paper, a hierarchical tree topic structure is proposed to address the problem of Chinese multi-document automatic summarization. This structure replace the traditional single-layered topics set and consequently, may describe accurately the text units similarity that are completely similar or incompletely similar to each other. Thus, the wrong division of topics brought by the units clustering may be most reduced. Furthermore, we propose summary sentences extracting and evaluating algorithms based on hierarchical topic tree. We validate our method by comparing result to baseline method. Experimental results show the better performance of our method in improving precise and reducing redundancy.

ACKNOWLEDGEMENTS

This work is supported by China National Natural Science Foundation (60803092), Promotive research fund for excellent young and middle-aged scientists of Shandong Province (2010BSA10014) and WeiHai City Science & Technology Fund Planning Project (2010-3-96)

REFERENCES

1. Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In ANLP/NAACLWorkshop on Summarization, Seattle,WA, April 2000.

2. McKeown, Kathleen R. and Dragomir R. Radev. Generating Summaries of Multiple News Articles. In Proceedings of 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, 74–82, 1995.

3. Hardy, Hilda. Cross-Document Summarization by Concept Classification. Workshop on Text Summarization (DUC 2001). New Orleans, 2001.

4. Endre Boros, Paul B. Kantora and David J. Neu. A Clustering Based Approach to Creating Multi-Document Summaries. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New. Orleans, LA, 2001.

5. Yi Guo and George Stylios A New Multi-document Summarization System. In Proceedings of the Document Understanding Conference, 2003.

6. H.Zha. Generic Summarization and Key Phrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Tampere. Finland. 2002 .

7. Yong-Dong Xu etc.. Using Multiple Features and Statistical Model To Calculate Text Units Similarity，Proceedings of the Fourth International Conference on Machine Learning and Cybernetics IEEE, Guangzhou, 19-21 August 2005

8. García J.A., Fdez-Valdivia J., Cortijo F. J., and Molina R. 1994. A Dynamic Approach for Clustering Data. Signal Processing, Vol. 44, No. 2, 1994, pp 181-196.