Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

A New Method of Automatic Summarization based on Semantic Rhythm

Fan Chen

Information Science & Technology Department, Tianjin University of Finance and Economics,

Tianjin 300200, China

Abstract: Language Rhythm is an important characteristic in language. Research and analysis the Language Rhythm, an important and potential rhythm: language semantic rhythm reflects the language characteristic in distribution of semantic. Based on it, a new method of automatic summarization is researched. How to get the language semantic rhythms and how to used it in automatic summarization are expounded. The simulation results show that it is significant to apply in automatic summarization.
Keywords: Automatic Summarization, Language Semantic Rhythm, Natural Language Processing.

---

1. INTRODUCTION

Today, massive information we absorbed every day from Internet. The 38thChina Internet development Statistical report [1] shows that there are 4,540,000 websites in China. So we need to face the mass information from the large scale websites and the other type of information from Internet. And how to improve the efficiency of understanding the documents, web page, news and so on that we can received. How to getting the most important information that we need. This is a difficult problem. A high quality abstract can be helped to understand the whole document efficiently. How to get a good and correctly abstract of a document and other text resource in Internet quickly, can improve the efficiency that accepting the information in Internet. Lots of studies in how to get the perfect summary of a document quickly and correctly have been developed. Automatic Summarization is studying on how to get the summary of a document and other style text resource automatically without human intervention. Automatic Summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax [2].There are four main methods in automatic summarization: automatic extraction, automatic summarization based on understanding, information extraction and automatic summarization based on structure [3].To get the summarization considering the understanding is limited by the domain knowledge. Information extraction is needed the knowledge structure about the information, so it limited by domain, too. Using the structure characters in summary, there is a very complex connect network in sentences, and it will need

more time and space resource. A new method considering the semantic character in language is be used in automatic summarization, and it can get the summary of the document very quickly without any limit, such as the length of document, the domain of document and the complex of document structure.

2. AUTOMATIC SUMMARIZATION SYSTEM PROCESS

Usually, the Automatic Summarization system process [4] is showed as figure 1:

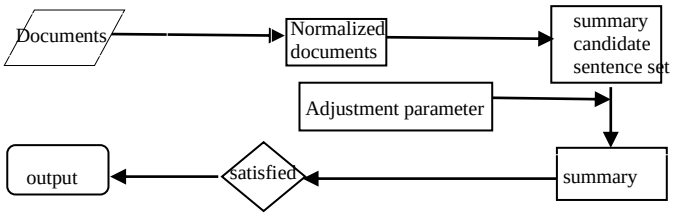

Fig. 1 The process of automatic summarization

The process of automatic summarization need to normalize the document first, the others processes are based it; the efficiency of the method in automatic summarization is influenced by the result of this step. Next, to choose the candidate sentence set of summary, these sentences in this set are the base in automatic summarization, so to choose the good sentences are so important in the task. Then to generate the summarization, and evaluate it, if it can be satisfied by the users then a good summary is generated, otherwise it need to adjust the parameters in summary generation until to get the satisfied summary.

In Automatic Summarization, it is important and difficult to get the appropriate sentences before generating the summary. Methods based on probability statistics [5], correlation analysis [6], graph based methods [7], and machine learning [8] can get the sentences without the information of semantic, they can get advantage in speed but lack in integrality. How to generate the summarization automatically and efficiently and considering the semantic information, then the summarization can keep some semantic characters. Generating summarization by using the semantic information easily and quickly without affecting the efficiency of automatic summarization, it is an important problem in automatic summarization.

3. LANGUAGE SEMANTIC RHYTHM

Language Rhythm is a nature feature and phenomenon in human language. Each document has the unique rhythm, and the rhythm of language can reflect some language characters in documents. There are five language rhythms, nature rhythm, grammar rhythm, logic rhythm, emotion rhythm and semantic rhythm. The semantic rhythm of one document reflects the semantic character of one document. And it can reflect the distribution of semantic in one document as the same time. One document is composed with words, in the other words; words are the basic element of one document. The words in the documents are classed as two classes, notional words and function words. The notional words are contained the semantic information,

nouns, adjectives, verbs and etc. are belong to it. They contain all kinds of semantic information of one document. The function words reflect the structure, the emotion and other kinds characters of one document without semantic. One document is composed with the notional words crossing the function words. One sentence has both notional words and function words. The more notional words in the sentence, the more semantic information contained. The more function words in one sentence, the other characters contained without semantic information. Some sentence in one document has much semantic information and some sentence has less semantic information. The semantic information is distributed in the document with some rhythm characters. So, it is needed to study that how to obtain the semantic rhythm of one document and how to use the semantic rhythm in automatic summarization. Then the semantic rhythm is defined and the method that to get it in one document is discussed next.

4. OBTAIN LANGUAGE SEMANTIC RHYTHM

The character of distribution of the notional words and function words can be catched. Semantic Density (SD) is reflecting the relation between notional words and function words in one sentence. It is calculated by formula 1

$$SD=\text{nums (notional words)}/\text{nums (words)} \quad (1)$$

nums(X): The function is to used to get the amount of X.

The SD of each sentence maybe is very different in one document. In one document, the sentence contains some notional words and some function words, has the unique SD. For the sentence uses the different notional words explain the different semantic information, and the other kind words –function words help to explain the semantic information in logic, structure and emotion. So the SD in each sentence of document is unique. And it can reflect the semantic character of the sentence, and all the SDs reflect the semantic character of one document. Some sentences can be chosen in the task of automatic summarization. The good quality sentences need to be chosen. It is a problem that how to find the good quality sentence? The higher SD, the better quality, maybe it is corrected. The sentences with the highest SD are chosen as the candidate sentence set in automatic summarization. But some sentences such as title (without function words mostly) are be chosen as shown in Figure2 and this result is not wanted.

Ji suan ji chu li,ji suan ji bing du,wang luo hua tong
xin,ruan jian biao zhun hua,
fen bu shi shu zi chu li,ke suo wo zhan zheng zhong, gong ji
mu biao xi tong shi ji suan ji ge zhong xi tong
dang ran du wen jian yun xing shi,ji suan ji ping tai ti
gong le yi ge ping tai

zai hai wan zhan zheng zhong,rang bing du gan ran dui fang
dian zi xi tong ,qi gong ji mu biao shi ji suan ji ge zhong
xi tong

Fig. 2 Select the candidate sentences considering SD only

So it is not correct that the higher SD the sentences are better. The candidate sentences set are be chosen that considering the SD of each sentence only, remove the sentence with highest

SD, and the sentences are chosen from each part of one document. So the summary composed by these sentence s is cannot be understand smoothly. So it is not a good method to choose the candidate set sentences only considering the SD.

Semantic Cluster (SC) is proposed here to build the better candidate sentences set in the task of automatic summarization. Semantic Cluster is the set of some sentences with the appropriate SDs. The distribution characters of the sentences SDs is considered here. The distribution of the sentences is related with the semantic characters of one document. When the important thing is discussed, the sentences contain the higher SDs. So the distribution of the sentences SDs is about the semantic distribution. As a result, if the set of sentences which is discussing the same thing are chosen to be the set of the candidate sentences set, the summarization can be understand more easily. The SC is created as the Figue3 shown below:

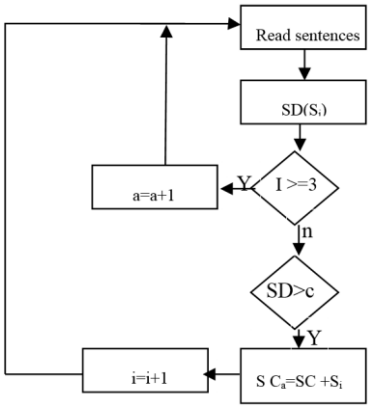

Fig 3 Process of creating SC

SC is set of the sentence with higher SD not the highest one. The sentence that can be absorbed in SC is considering both its own SD and the SD of its neighbors.

And it is an important problem that how to find the boundary of the SC. In a other word that how to build the SC. Formula 2 is used in SC creation:

$$SC = \{S1, S2, Sn\}, (SD(Si) > c, 0 < c < 1) \quad (2)$$

S_i is the sequence of the sentences, as Formula 2 show, the sentences that selected in SC is continuously. And c is the parameter to find the SC.

Then, calculate the average SD (ASD) of the SC by using the Formula 3 as shown below:

$$ASD(SCi) = \sum_{i=1}^{n} SD(Si)/Len(SCi) \quad (3)$$

The SC_i contains limit account sentences that is about 3 to 5, and the top 3 SCs ranked by the ASDs are selected as the summary in one document. The result is shown in Figure 4.

ji suan ji xi tong de cui ruo xing, ji suan ji bing du ru qin de he xin ji shu shi jie jue bing du de ru qin,wei ji suan ji bing du po huai ti gong le yi ge ping tai. ruan jian fang mian ye cun zai yin huan. rang bing du gan ran dui fang dian zi ti tong.zai xi tong qi dong pan de wen jian zhong jia ru yi ge bing du jian ce cheng xu.

Figure 4. Automatic Summarization with the appropriate ASD.

The summarization built with the SC is more readable than only considering the sentences' SD.

5. SUMMARY

The semantic information that considered in the automatic summarization is not based on understanding the knowledge in this method. It is only considering the distribution character of the semantic information. And select the candidate sentence set is not only using the isolated semantic information, but also the context sentences, too. As a result, the sentences in the summarization are explained the document main content. But the summarization needs to be smoother, so how to make sentences in the summarization more coherent is an important problem next.

REFERENCES

[1] The 38th China Internet development Statistical report . http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201608/t20160803_54392.htm. S.A. Lowe, "The Beta-Binomial Mixture Model for Word Frequencies in Documents with Applications to Information Retrieval," Proceedings of Euro speech'99,Budapest,September 1999.

[2] https://en.wikipedia.org/wiki/Automatic_summarization.

[3] Xu Jin, Bingru Yang, Zhigang Guan. Automatic abstracting method analysis [A],2004.

[4] Peter D. Turney, Patrick Pantel From Frequency to Meaning:Vector Space Models of Semantics Journal of Articial Intelligence Research [J] 2010(37) :141-188.

[5] Qianqian Cheng,Dagang Tian. Automatic Chinese Summarization Model Based on Basic Elements Method[J] New Technology of Library and Information Service 2010,26 (2).

[6] Hongling Wang,Minghui Zhang,GuodongZhou,Chinese multi-document summarization system basedon topic information[J]. Computer Engineering and Applications, 2012,48(25).

[7] Canhasi E,Kononenko I,Weighted archetypal analysis of the multielement graph for query focused multi-document summarization [J]Expert Systems with Applications,2014,41 (2)

[8] Yang Cao,Ying Cheng,Lei Pei. A Review on Machine Learning Oriented Automatic Summarization [J]. Library and Information Service,2014,58 (18).