Query-Based Summarization for Search Lists

Xinghuo Ye, Hai Wei
Huazhong Normal University Hankou Branch, 430212, Wuhan
E-mail:yxhez@tom.com

Abstract

In this paper we describe a Query-based summarization system. We propose a practical approach for this task by identifying the sentences with high query-relevant and high information density. This paper introduces a statistical model for query-relevant *summarization: adopt Word overlap feature to capture the power of correlation with the query and mine the relations between the sentences. While the first is executed by computing semantic similarity between the sentence and the query, and the other is executed by using semantic graph. Then these two kinds of features are blessed to score each sentence. At last with the help of MMR for reducing redundancy, we get the summary. Experimental results indicate that this method is encouraging for both those retrieved documents that correspondingly concentrating to one subject and retrieved documents who have many sub-topics and comparatively being related to the query.*

1. Introduction

The goal of a good document summary is to provided user with a presentation of the substance of a body of material in a coherent and concise form. Ideally, a summary would contain only the "right" amount of the interesting information and it would omit all the redundant and "uninteresting" material. In order to help alleviate the information overload problem and help users to find the information they need, many researchers turn to IR for help, who can help reduce the information overload problem by allowing a user to do a centralized search, but at the same time too many web pages are returned for a single query.

However, most search engines interact with user in a "one size fits all" fashion and ignore the user's preferences, search context or the task context. The burden is then placed on the user to scan, navigate, and read the retrieved documents to identify what s/he wants. As a result, users frequently have to refer to the full text of the document, making the process of relevance judgement time-consuming.

As query-based summarization should be both a "compressed version" of the document cluster and satisfy the user's need. This suggests that the selected sentences should be highly relevant to the query and representative of the documents at the same time. In order to judge whether a sentence should be appropriately included in the summary, we acquire two kinds of features for each sentences: the power of correlation with the query and the power of global connectivity, while the first is executed by computing the semantic similarity between the sentence and the query, and the other is mined from semantic graph. These two factors can be combined to measure the importance of each sentence. At last with the help of MMR for reducing redundancy, we get the summary.

In this paper, work in related areas is discussed In section 2; a description of the system is in Section 3;The method of query-based summarization is in Section 4; we present the experiments and the evaluation of experiment results in section 5; Finally, Session 6 gives a summary and talks about future work.

2. Related work

Query-Based summarization has synthesized all the technological merit available such as MDS□IR and QA, while avoiding their deficiency to a certain extent. Therefore, it will make a great significance of research and bright prospect of vast application, when put in use in many fields such as acquisition of massive information, recommendation of personalized information, the digital library, the figure analysis of commerce intelligence, the electron's administration and the calculation moving. During the past few years, a number of exploratory researches have been undertaken by some of the domestic and international staff.

Saggion et al. [1], a simple query-based scorer by computing the similarity value between each sentence

0-7695-3090-7/08 $25.00 © 2008 IEEE
DOI 10.1109/WKDD.2008.14

and the query is incorporated into a generic summarizer to produce the query-based summary .White et al. [2] and Goldstein et al. [3] create query-dependent summaries with a sentence extraction model. After cutting the documents into their component sentences, they order all the sentences according to features such as Word overlap feature between them and the query, their originality position in the passages and relevant paragraphs, and the core words they contain. At last they choose a number of the high-scoring sentences to form the summary. And Hovy et al. [4] select the important sentences based on the scores of basic elements (BE). However, all of the above methods ignore possible semantic connections between the sentences. CATS is a topic-oriented multi-document summarizer which first performs a thematic analysis of the documents, and then matches these themes with the ones identified in the topic [5].

In order to general the summary with the most query-relevant and the most content it covers, our approach naturally extracts the sentences from relevant documents by exploiting both the power of correlation with the query and the power of global connectivity. The first guarantees the relevance to the query set, and the other reflects the importance of the textual units to the whole.

3. System description

We present our work in the framework of a summary extraction. Important sentences are extracted and re-organized to form a summary with the most query-relevant and the least redundancy. First we give the system an overview, and then describe the important steps in detail.

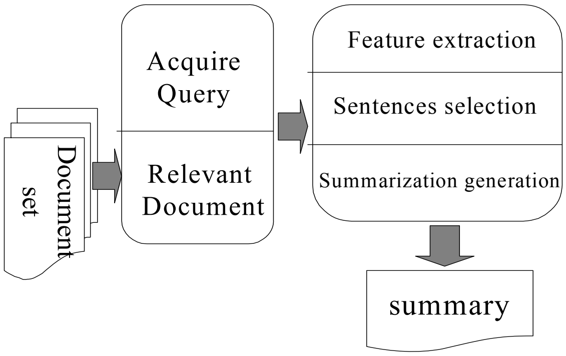

Figure 1.System Overview

Thus, the total has four steps: 1. acquire relevant documents; 2. feature extraction; 3. sentences selection; 4.summarization generation.

3.1. Query Description

The input given to the summarization system is a user's query, and it has to produce a fluent, well-organized summary. Here we assume that there must exist documents that contain some sentences which can summarize the topic briefly in our document corpus. As there are thousands of documents and only some relevant to the query, it first has to pick out relevant documents for the next analyses. Considering the limitation and one-sidedness of the message that the user provides, we choose term senses to act as the features of the query. For example if a query is: $Q=(q_1,q_2,q_3...q_n)$.

After inquiry expansion, it is defined as: $Q=(c_1,c_2,c_3...c_n)$.

Where $c_i=(w_1,w_2,w_3...w_h)$ is the senses of term q_i. Here we use the mapping algorithm for extracting term senses proposed by MengWang [7]. From the retrieved result, we choose the top n documents as the mine where the summary sentences are from.

4.Query-Based summarization

4.1. The Specificity power

Here, we adopt Word overlap feature to capture the power of correlation with the query. And the weight of each common word is computed by the Okapi. The power of correlation with the query is defined as followed:

$$P_{i1}=\sum_{j\in Q\cap S_i} In\frac{N-df+0.5}{df+0.5}\times\frac{(k_1+1)tf}{(k_1(1-b)+b\bullet\frac{dl}{avdl})+tf}\times\frac{(k_3+1)\bullet qtf}{k_3+qtf} \quad (1)$$

where tf is the candidate's frequency in document, qtf is the candidate's frequency in query, N is the total number of documents in the collection, df is the number of documents that contain the candidate, dl is the document length (in words), avdl is the average document length and k1 (between 1.0–2.0), b (usually 0.75), and k3 (between 0–1000) are constants.the score obtained by all sentences is normalized by the maximum score, so that the new maximum feature value corresponds to 1. This normalization will facilitate an easier combination of different feature scores for a sentence.

$$P_{i1} = \frac{P_{i1}}{\max\limits_{j\hat{I}\{1,2\ldots n\}} (p_{j1}, Q)} \qquad (2)$$

4.2 The informativeness power

In order to mine the relations between the sentences, and extract the most representative ones, we apply a diffusion process on the graph to obtain a more appropriate undirected graph.

If sentences are considered as nodes, the sentence collection can be modeled as an undirected graph by generating the link between two sentences if their weight exceeds 0, i.e. an undirected link between s_i and s_j ($i \neq j$) with suitable weight SUI(s_i,s_j) is constructed if SUI(si,sj)>0; otherwise no link is constructed. Thus, we construct an undirected graph G reflecting the semantic relationship between sentences by their content similarity. Note that given a link between a sentence pair of s_i and s_j, if s_i and s_j comes from the same document, the link is an intra-document link; and if s_i and s_j comes from different documents, the link is an inter-document link. We believe that inter-document links are more important than intra-document links for information richness computation.

The power of global connectivity is defined as:

$$P_{i2} = \frac{d_{s_i}}{d_{\max}} \qquad (3)$$

While d_{si} is the degree of the node S_i , d_{max} is the degree of the node that has the maximum edges using for normalization.

4.3. Sentences selection

The score of a sentence is used to measure how important it is to be included in the summary, for each sentence, the score is calculated as the weighted linear combination of the above two features.

$$P_i = a P_{i1} + b P_{i2} \qquad (4)$$

a , b are the experience weights assigned by human, which can be viewed as adjusting parameter for the query-independent and dependent parts of the scoring function respectively. With the sentence scores, we sort all sentences from high to low. At last with the help of MMR (Maximal Marginal Relevance) technique (Carbonell and Goldstein, 1998) for reducing the redundancy, we get the most representative ones.

4.4. Summarization generation

Here, after obtaining all summary sentences, we choose the document that contains the largest number of summary sentences as a frame of reference. Then summary sentences are inserted into the document according the similarity between them and the sentences in it, and then re-order them, we get the final summary.

5. Experiments and evaluation

5.1. Data

The data sets used in this survey are from the news corpus built by Monitor and Research Center for National Language Resource (Network Multimedia Sub-branch Center), which contains about 900,000 news articles in the 2005 and 2006. At the same time, correlative text sets are created for several news events. In our research, we randomly selected fifty kinds of topic sets and blended them together as our test sets, while each set contains about 1000 Chinese news articles.

5.2. Evaluation metrics

Here we used the task-based extrinsic measure (Mani et al.1999). Two professors of linguistics and three graduate students were invited to participate in the experiment as the users. All of them were first asked to read all the correlative documents for each query, and then all the users assigned a readability score that arranged from 1 to 5 for each summary that generated by our method.

5.3. Experimental result

In this section, we provide the results that our method performed. In our experiments, the parameters are empirically set as: The compression rate of sentences extraction to form a summary is 10% and 20%. These rates yield the number of extracts in the summary comparable to the number of actual extracts in a given test document set. In the process of acquiring relevant documents, the parameter n is 7-20. The threshold a of the cosine similarity is 0.2.

The parameter l for combining the power of correlation with the query and the power of global connectivity is 0.2.For each topic, our system generated both 10% and 20% summaries. The detailed evaluation results based on readability score are shown in Table 1.

Table 1: Evaluation Topics and their corresponding experiment results

Item	Compression	User 1	User 2	User 3	User 4	User 5
□□□□□	10%	3	3	3.5	2.5	3
	20%	3.5	3.5	4	3.5	3.5
□□□—□□□□□□□□□□□□□	10%	3	2.5	3	3	2.5
	20%	3.5	4	4.5	4.5	4
□□“□□”□□□□□	10%	2.5	3	2	3.5	3
	20%	3	4	3	3.5	4
□□□□□□	10%	2	2	2.5	2	2
	20%	2.5	3	3	2	2.5
□□□□□□□□□□□□□□□□□	10%	3.5	3.5	4	4	3
	20%	2	3	2	2.5	2

By analyzing correlative document set, we discover that for those retrieved documents who have many sub-topics and comparatively being related to the query, the result performs good when the compressibility is 20%. For example, the retrieved documents contain many sub-topics in the result of "query 2" such as "the interrelated reports of three rescued deeds", "Shunyou Wang joins the moved china 2005", "the survey of Shunyou Wang's family", "response from society about Shunyou Wang's story", "the experience of work that Shunyou Wang met". These sub-topics are related to the query closely, so all of them should be included in summary.

Secondly, regarding the other kind of documents whose subject is correspondingly concentrated, it can get satisfying result when compression rate is 10%. For example, all the documents expound viewpoints of six countries about how to solve Korean Peninsula nuclear peacefully in the result of "query 5". Here, if compression ratio is enhanced, it would get much more redundant information instead.

6. Conclusion and future work

In this paper, we propose a method for query-based summarization by sentences scoring, which integrates natural language processing and information retrieval techniques to perform automatic customized summarization. The result of evaluation shows that our method is encouraging. And there are still lots of rooms for us to improve. In the future, we will go on our work from the following aspects: Named entities are the important part of the sentences, whose degree is more significant than other element. Generally speaking, the more named entities a sentence has, more important the sentence is. In our method, we just er terms, and we will r sentences scoring.

Second, further experiments on larger text corpora are needed to evaluate the performance of our method. Specially, the threshold is just the average value of the training process, which sometimes may perform not steadily. At ast, we need to do more research on improving the linguistic quality of summaries.

References

[1] H. Saggion, K. Bontcheva and H.Cunningham. Robust generic and query-based summarization. In Proceedings of EACL'2003.

[2] R. W. White, I. Ruthven and J. M. Jose: Finding Relevant Documents using Top RankingSentences: An Evaluation of Two Alternative Schemes, SIGIR, 2002

[3] J.Goldstein,M. Kantrowitz,V,Mittal,J.Carbonell: Summarizing text documents:Sentenceselection and evaluation metrics.ACM SIGIR,1999

[4] E. Hovy,C.-Y.Lin and L.Zhou.2005.A BE-based multi-document summarizer with query interpretation. In Proceedings of DUC 2005.

[5] Atefeh Farzindar, Frederik Rozon and Guy Lapalme: CATS a topic-oriented multi-document summarization system at DUC 2005.

[6] Lin Zhao,Xuanjing Huang,Lide Wu.2005.Fudan University at DUC 2005. In Proceedings of DUC 2005.

[7] □□,□□□,□□□,□□□,□□ HowNet □□□□□□□□□□□□□□2005 □□□□□□□□ 3 □

[8] Qin Bing, Liu Ting, Li Sheng. Summarization Based on Physical Features and Logical Structure of Multi Documents. High Technology Letters, 2005