Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008

PERSONALIZED MULTI-DOCUMENT SUMMARIZATION IN INFORMATION RETRIEVAL

XIAO-PENG YANG[1], XIAO-RONG LIU[2]

[1]Jiujiang University, Jiujiang, 332005, Jiangxi, China,
[2] Network and Information Center, East China Institute of Technology, Fuzhou, 344000, Jiangxi, China
E-MAIL: yxp_plus@163.com

Abstract:

This research is directed towards automating open-domain Multi-Document Summarization in the framework of Web search. We present a novel approach to achieve this object. Given an unrestricted user query, our system retrieves documents related to and summarizes them. In the process of summarization, the sentences in a given document are scored based on the relevant value and the informativeness value, which are realized by using word overlap and semantic graph. Then, the sentences with highest scores are incorporated into the output summary together with their structural context. Experimental results show that our query-topic focused summary could return a topically relevant extractive summary. And the summarization quality is relatively competitive.

Keywords:

Multi-Document Summarization; The Relevant Value; Web Page Summarization; The Informativeness Value

1. Introduction

The continuous and rapid growth of Web information requires efficient assistance to Web users for information filtering and interpretation. 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. However, most search engines interact with user in a "one size fits all" fashion and ignore the user's preferences [1], 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, so the user needs to scroll down the huge number of results. This has several disadvantages. First, loading the page takes time. Second, if the target page is long and complex, the determination of the relevancy may be both difficult and time consuming. Also, there are usually a big number of returned results and it is not feasible for the user to open each particular link to find out its relevance. To overcome these problems and improve the effectiveness of Web search, better approaches regarding the summarization are needed.

It would be very helpful if an effective search engine could provide more contextual and summary information to help these users explore the retrieval set more efficiently. So we can say that automatic summarization for the search lists will make a great significance of research and bright prospect of vast application, who can help concisely describe the information and facilitate the users to understand the document cluster, especially provide personalized services.

As the summary should be represented the web document cluster and satisfy the user's need. In this paper, we present a novel approach to summarization of Web documents using a feature fusion based sentence scoring strategy. After retrieved documents are summarized according to the user query, the sentences in a given document are scored based on the relevant value and the informative-ness value. Then, the sentences with highest scores are incorporated into the output summary together with their structural context.

The rest of the paper is organized as follows. First, a brief overview of related work is given. This is followed by the description of the proposed system. Then, Experiments and Results are presented. At last, we give the conclusion and future work.

2. Related work

In generally speaking, web page summarization derives from text summarization techniques, while it is a great challenge to summarize Web pages automatically and effectively [2], because Web pages differ from traditional text documents in both structure and content. Instead of coherent text with a well-defined discourse structure, Web pages often have diverse contents such as bullets and images.

Currently there is no effective way to produce personalized, coherent and informative summaries of Web

978-1-4244-2096-4/08/$25.00 ©2008 IEEE

pages automatically. Amitay et al [3] propose a unique approach, which relies on the hypertext structure. This approach is applied to "generate short coherent textual snippets presented to the user with search engine results". And all major Web search engines use short summaries of document contents in displaying their results. Such as google [4], who displays short extracts that usually two lines of text fragments under the search results. The summaries can be very useful in determining the relevance of each result with respect to the query. However, in practice, such short summaries are usually inadequate for the user to determine the relevance of the documents.

Many competitive methods and systems have been developed recently also. For example, The systems designed by McKeown et al. [5] and Radev.D [6] et al provide summaries of popular recent events, which are discussed in some chosen news sources. However, there is a need for an application that could summarize information from any types of web pages. In other words it should be a system that could produce summaries of collections of web pages, which are not limited to newswire extracts.

With the message from the structure of documents, H. Alam et al. [7] realize a "table of content"-like hierarchy of sections and subsections for each document using some heuristics on HMTL tags present in the documents and incorporates this structural information in the output summaries. However, they method only creates general-purpose summaries, not tailored for particular user queries or Web search task.

Neto et al. [8]describes a text mining tool that performs document clustering and text summarization. They used the Autoclass algorithm to perform document clustering and used TF*ISF (an adaptation of TF*IDF) to realize sentence ranking and generate the summarization output. Our work is different from theirs because we perform a personalized Multi-Document Summarization based on the retrieval result. A more complicated sentence ranking function is employed to perfore the ranking performance.

More related work can be found in [9]. They use Meta-Crawler to perform web-based search and automatically generate summaries for each URLs retrieved. They only support single document summarization in their engine and the compression rate of the summarizer is also not customizable.

Compared with these work, our method supports multiple-document summarization for the search lists. In order to general the summary with the most query-relevant and the most content it covers, we mine the word overlap feature and use the semantic graph to discover the relation between sentences. With these two kind features, the sentences for a Web Pages Summary could be discovered. The experiments show that the proposed approach is encouraging.

3. System description

With the given query, our system first craws the relevant Web page，important sentences are extracted and re-organized to form a summary with the least redundancy and the most salient within topically relevant Web pages. We present our work in the framework of a summary extraction. First we give the system an overview, and then describe the important steps in detail.

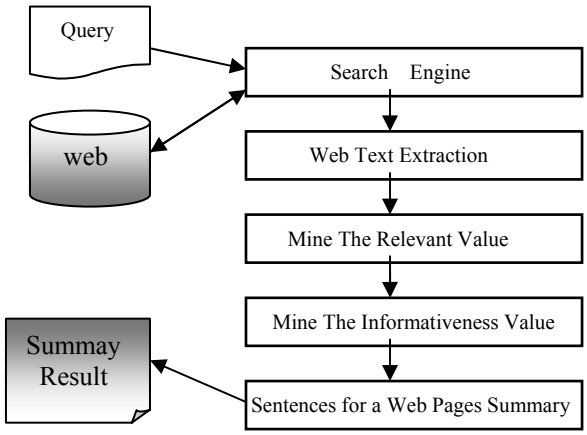

Figure 1.System Overview

The total has four steps: 1. Web Text Extraction; 2. The Relevant Value; 3. The Informativeness Value; 4. Sentences for a Web Pages Summary.

3.1. Web Text Extraction

The user enters an unrestricted query in our search engine, The result is a set of related documents. Spiders are used to fetch URLs from the Internet. After a URL is downloaded, the following steps are applied to normalize the text:

Table 1. Steps For Normalizing The Web Text

1. Parse the HTML file.
•Remove all links, images, and meta-information from the web pages;
•Remove pages containing adult-oriented contents ;
•Remove HTML markup information from the pages;
• Remove pairs whose pages contained frames;

- Remove pairs whose pages that had been moved since they had been included in the list; in other words, pages which were just "Page not found errors;
- Remove duplicate web pages;

2. Apply Porter's stemming algorithms to each keyword.
3. Index each keyword into the database along with its frequency and position information.

3.2. The Relevant Value

The words in a sentence can be of two cases: stop words and non-stopwords. From anther way to say, the non-stopwords overlap between the sentence and the query somewhat reflects the degree of their association. But it is not the result for the stop words, who have a negative effect on the information density of a sentence if its length is fixed.

We design a mode to measure the relevant feature for each candidate based on word overlap, while a negative value is assigned according to the number of stop words in the sentence, and a positive value is assigned according to the non-stopwords overlap. It is difficult to have a uniform measure among sentences of different lengths. That is, there often exists a long sentence bias for the non-normalized values. The longer the sentence is, the higher the values usually are. At last, the values mentioned above obtained by all sentences is normalized by the total word number of the sentence owns.

$$Rel\_Score_i = (\alpha * N_{stopwords} + \beta * N_{non-stopwords}) / N_{all}$$

$N_{\text{stopwords}}$ and $N_{\text{not-stopwords}}$ are the word number of the two cases respectively and N_{All} as the total word number of the sentence owns.

3.3. The Informativeness Value

The informative-ness value reflects the coverage value of each content-element to all content-elements at the same granularity. Thus, it is computed based on the graph. Because in the structure of graph, the more neighbors a sentence has, the more informative it is; the more informative a sentences neighbors are, the more informative it is [10][11].

Given a sentence collection $S=\{s_i \mid 1<=i<=n\}$, the weight between a sentence pair of s_i and s_j is calculated using the cosine measure. If sentences are considered as nodes, the sentence collection an be modeled as an undirected graph by generating the link between two sentences if the connection between them exceeds 0. And the connection is defined as followed:

$$SIM(S_i, S_j) = \cos(S_i, S_j) = \frac{\sum_{k=1}^{n} w_{ki} \times w_{kj}}{\sqrt{(\sum_{k=1}^{n} w_{ki}^2)(\sum_{k=1}^{n} w_{kj}^2)}}$$

Thus, an undirected graph can be constructed to reflect the semantic relationship between sentences. And we use an adjacency matrix M to describe the graph with each entry corresponding to the weight of a link in the graph.

The informativeness value for a sentence si can be deduced from those of all other sentences linked with it and it can be formulated in a recursive form as follows:

$$Inf\_Score_i = \sum_{j \neq i} Inf\_Score_j * M_{(i,j)}$$

3.4. Sentences for a Web Pages Summary

The score of a sentence is used to measure how important a sentence is to be included in the summary. According to the method that we proposed in the previous stages for calculating the two features, all the sentences in the document cluster are calculated as the weighted linear combination of the above two features and are ranked in descending order.

$$Score_i = \lambda Rel\_Score_i + (1-\lambda) Inf\_Score_i$$

λ is the experience weight assigned by human, which can be viewed as adjusting parameter.

To extract sentences to compose a summary, our system iteratively ranks the candidate sentences. Since two sentences may have redundant information, it is not appropriate to extract both sentences into the summary. And the method we used is as follows. When sentence Si is extracted, the weight of the remaining sentence Sj is adjusted as:

$$Score_j = \sqrt{Score_j^2 - Score_j * Sim(S_i, S_j)}$$

In each iteration, we extract the sentence with the highest score, and then adjust scores of the remaining sentences using the above formula. Scores of sentences that are very similar with the extracted sentence are adjusted downwards in this way. This process is repeated until we reach the length restriction of the summary. And the total number of selected sentences for a summary is controlled by the requested percentages of top-ranked sentences in relation to the total number of sentences of the target Web pages.

4. Experiments and evaluation

It is important to measure systems in actual

information seeking situations, and real-world systems can only be meaningfully evaluated in real-world settings. In order to measure the overall performance of our approach, it was evaluated using the task-based extrinsic measure as suggested by Mani et al [12]. The experiment was set up as follows: Five sets of documents on different topics were selected prior to the experiment. While the number of documents and the variations of document length in each topic set are different. And this will help test the robustness of our summarization algorithms. Table 1 gives the topics and their corresponding document information.

Table 2.Topics and document set information

Topic	No.of cluster	Length
Data Miningand their application In Information retrieval	23	901k
Wildlife in Danger of Extinction	14	610k
Health-related problems related to working with computers	10	585k
Nobel Prize Winners in the Sciences and Economics	18	724k
Development of magnetic levitation rail dystems	8	423k

Four users were selected for evaluation of these summarization results. The standard is they need to analyse the relevance, importance, usefulness and complement of the summaries. For each guide line, each user was asked to read through the set of full articles for each topic first, followed by its summary. After that, they needed to assign a score (1-5). The higher the score is, the better is the summary.

Table 3. Semantic Value Assigned To The Summaries

Users	Relevant	Important	Useful	Complete
User 1	3.5	2.8	3.1	3.8
User 2	3.7	2.5	3.6	3.4
User 3	3.8	3.0	4.0	3.6
User 4	3.2	3.1	4.5	3.5
User 5	4.0	3.1	4.5	4.1

From the table, we can get the average readability score for summaries is 3.3 given by user 1; For user 2, the average readability score is 3.4 also; For user 3, user 4 and user 5, the average readability score for summaries is 3.6,3.57 and 3.9 respectively. So we can say the summarization quality is relatively competitive. The results also indicate the help users complete more tasks and complete their tasks more quickly. For the reason, our understanding is that the approach looks for terms in sentences which are very close to terms expressed in information need when computing the relevant feature for each sentences, and it is an important factor for sentences ranking. In this process we are not performing any deeper analysis such as, an understanding how these terms are related to each other or what is the semantic content of the sentence chosen. Therefore the summary holds a good chance of satisfying the information need, but this is not always the case.

By analyzing correlative approach proposed by us, a reasonably intelligent person would no doubt make use of information that our system ignores. For instance, the documents often appear in web pages in the form of images, but this information is lost without a front-end OCR module to extract this kind of text. At the same time, our approach does not exploit structural clues about what's important on the page. For instance, the text within the <title> ... </title> region is likely to be relatively important, while text within a <small> ... </small> is probably less important. This all what we will go on in future. The performance of the system could clearly benefit from better content selection and surface realization models.

5. Conclusion and future work

We have introduced a new research area of summarizing textual changes in web page collections and have presented a complete system. The system uses novel methodology for extracting and summarizing documents in web collections.

Our method has several limitations, which we want to focus on in the future. The research can be extended in several directions. First, the scoring method of sentences can be improved with query-biased methods. Second, the system can be refined considering different types of search tasks, such as searching for a particular fact or searching for background information about a subject, etc. Also, the heuristics used in the identification of structural information and incorporation in the output summary can be improved. Finally, natural language processing techniques is always open to improvement; e.g. incorporating verb phrases besides noun phrases to the system. As a future work, the system will be evaluated on a comprehensive task-based evaluation.

References

[1] Tag cloud Wikipedia, the free encyclopedia, http: //

en.wikipedia.org/wiki/Tag_cloud.

[2] E. Amitay and C. Paris. Automatically summarising web sites - is there a way around it? In ACM 9th International Conference on Information and Knowledge Management, 2000.

[3] E. Amitay and C. Paris. Automatically summarizing web sites - is there a way around it? In ACM 9th International Conference on Information and KnowledgeManagement, 2000.

[4] Google, 2006. http://www.google.com.

[5] McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B., Sigelman, S.: Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In Proceedings of Human Language Technology Conference. San Diego, USA (2002)

[6] Radev, D., Blair-Goldensohn, S., Zhang, Z., Raghavan, S.R.: NewsInEssence: A System for Domain-Independent, Real-Time News Clustering and Multi-Document Summarization. In Human Language Technology Conference. San Diego, USA (2001a)

[7] H. Alam, A. Kumar, M. Nakamura, A. F. R. Rahman, Y. Tarnikova and C. Wilcox, "Structured and Unstructured Document Summarization: Design of a Commercial Summarizer using Lexical Chains", In Proceedings of Seventh International Conference on Document Analysis and Recognition, IEEE Computer Society, 2003, pp. 1147-1150

[8] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, 2000) Readings in Information Retrieva

[9] http://extractor.iit.nrc.ca/

[10] V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, H. Karambelkar. Bidirectional Expansion For Keyword Search on Graph Databases. VLDB, 2005

[11] Salton, G., Singhal, A., Mitra, M., and Buckley, C. 1999. Automatic text structuring and summarization. In Mani, I. and Maybury, M. (Eds.), Advances in automatic text summarization.

[12] Mani, I. and Bloedorn, E. (1999). Summarizing similarities and differences among related documents. Information Retrieval 1(1): 35—67.