Generating Personalized Summaries Using Publicly Available Web Documents

Chandan Kumar, Prasad Pingali, Vasudeva Varma
Language Technologies Research Centre
International Institute of Information Technology
Hyderabad, India
chandan_kumar@research.iiit.ac.in
pvvpr,vv@iiit.ac.in

Abstract

Many Knowledge workers are increasingly using online resources to find out latest developments in their specialty and articles of interest. To extract relevant information from such multiple online information sources summarization is being used. Current summarization systems produce a uniform version of summary for all users. However summaries which are generic in nature do not cater to the user's background and interests. In this paper we propose to make the summarization process user specific and present a design for generating personalized summaries of online articles that are tailored to each person's interest. The user's data available on web is used for model their background and interest. A controlled user-centered qualitative evaluation carried out on news articles of science and technology domain, indicates better user satisfaction with personalized summaries compared to generic summaries.

1 Introduction

With huge amount of online data available, knowledge workers find it increasingly difficult to extract information relevant to them. Therefore it is of great help to present the content of several articles in a condensed way using summarization. Automatic summarization is the process through which the relevant information from one or several sources is identified in order to produce a briefer version for the user [6].

While different professionals may have different perspectives on the same text, based on their field of expertise and interest, present summarization systems produce one uniform summary for all users without considering the user's personal interest. Thus there is a great need for summaries to cater to the user's personal background and interests. An effective summarization, thus, should not only be a function of the input text but also of who the reader is and what his prior knowledge is. So a good summary should change in accordance to preferences of its reader. In this paper we propose to extract this kind of user specific personalized summary for knowledge workers. We extract the personal information of the user using information available on the web. Web is a huge source of information and it contains a lot of personal information of web users. These pages contain enough personal data to model the users. As these documents are available publicly, there is no privacy concern in collecting user data, and user can be modeled anonymously without any effort from user.

In this work we concentrate on process of summarization that preserves the specific information that is relevant for a particular user, rather than information that simply summarizes the content of the document set.

This paper is organized as follows. In Section 2 we discuss the motivation for including user's background in summarization process. Section 3 describes the summarization model of sentence extraction. In section 4 we describe our user specific scoring mechanism and summary generation algorithm. In Section 5 we analyze system performance in a user-centered evaluation.

2 Motivation

One of the issues studied ever since the inception of automatic summarization in the 1960s was that of human agreement[4]: different people can choose different content for their summaries. Marcu-1997[3] found percent agreement of 13 judges over 5 texts from scientific America is 71 percent. Rath-1961[4] found that extracts selected by four different human judges had only 25 percent overlap. Salton-1997[5] found that most important 20 paragraphs extracted by 2 subjects have only 46 percent overlap. These results show that each person has different perspective on the same text and when persons of different background and expertise summarize the same articles, they include different content from each other, reflecting their personal interest and background knowledge. Thus there is a need to incorporate user knowledge in the automatic summarization process to provide them specific summaries. But at present most summarizers generate summaries using a generic notion of salience. In other words, what is important to summarize is determined by features of the text only, not by who the reader is, or what his background knowledge is. So in this work we explore the possibility of adding user personal knowledge into automatic summarization process. We treat Summarization process as not only a function of the input text but also of its reader.

3 Automatic Summarization

Text summarization is the process of selecting the most salient information in one or more textual documents. Recently there have been attempts to generate abstracts[6] as a summary. Extractive based approach [8, 7] is still most successful and useable where the primary aim is to extract highly informative snippets, i.e. keywords, sentences and paragraphs, which can be read in lieu of the original documents. Our approach also comprises of the same sentence extractive mechanism. Here we score sentences and extract the top ranking sentences and put them up verbatim, possibly after some re-ordering, as a summary. While there has been substantial effort in text summarization area for generic sentence extraction [8, 7], it is found that the term frequency based approach is very satisfying [1, 2], studies shows that frequency is indeed a powerful predictor in content selection with very good performance. In this work we use the same term frequency based approach to score sentences based on its importance with respect to the input documents.

Generic Sentence Scoring: Given the document set D to summarize, we compute the probability distribution over the words w appearing in the input D, $p(w|D)$

$$p(w|D) = \frac{tf(w,D)}{|D|} \tag{1}$$

where $tf(w,D)$ is the frequency of word w in the document D and $|D| = \sum_w D(w)$ is total number of content word tokens in the input document set D , it is essentially the length of the document set D.

Now For each sentence S in the input, assign a weight equal to the average probability of the words in the sentence, i.e.

$$score(S) = \sum_{w \in S} \frac{p(w|D)}{|\{w|w \in S\}|} \tag{2}$$

This score is used to select the most relevant sentences in general, which will be used to form an generic extract of the document set later used in the summary.

4 User Specific Summarization

4.1 Estimating User Background model

Our aim is to provide specific summaries to user based on their field of expertise and personal interest. To achieve this the system should know about the person's background knowledge, i.e. it should have an analysis of the goodness of his field. To model the user we propose to make use of his personal data available on web. For example for a research professional his information can be in an affiliation page, a project page,a conference page, an online paper, or even in a blog written by himself or others about him. These pages contain enough personal data to model the users. so we propose to extract the personal information of the user using information available on the web. Major benefit here is that there are no privacy issues in user modeling, since the personal data can be obtained anonymously and without any effort from user.

We used search engine to acquire these Web pages. It is reasonable to use a search engine because it can search the whole world wide web and also tracks the temporal variance of the information available on the Web. For profile creation the first step is to put the person's full name to a search engine (name is quoted with double quotation such as "Albert Einstein") and retrieve documents related to the person. From the search results 'n' top documents are taken and retrieved from their corresponding source websites to define that person's profile. These documents are parsed to extract text content. After performing the removal of stop words and stemming, a unigram language model is learned on the extracted text content. This model can be interpreted as the probability of a word w being related to the person's profile U.

$$p(w|U) = \frac{tf(w,U)}{|U|} \tag{3}$$

4.2 User Specific Sentence Scoring

To calculate sentence relevance for a particular person we consider its generic importance as well as importance related to profile. While scoring the sentences, the term probability of the document set D $p(w|D)$, and the user profile U $p(w|U)$ have been merged using a linear weighted combination. The score of a sentence S for user u is given as

$$score_u(S) = \sum_{w \in S} \frac{\alpha.p(w|D) + \beta.p(w|U)}{|\{w|w \in S\}|} \tag{4}$$

where α and β are the weighing parameters.

4.3 Summary Generation

After sentence scoring, top ranking sentences are selected to produce summary after eliminating redundancy. For redundancy identification, we use the measure of number of terms overlapping between the already generated summary and the new sentence being considered. Once sufficient number of sentences are picked to make the required length of summary(250 words), they are arranged based on chronological ordering (between documents i.e. based on the time stamp) and order of occurrence (within the document). Thus, sentences coming from different document will be ordered based on their source documents date of publication and if two sentences originate from the same document their original order in the source document will be considered. Any additional words than the required length of summary are truncated. So for a given document cluster containing articles related to a topic, two types of summary can be generated: Generic summary using sentence scoring function using equation 2, and a user specific Personalize summary using equation 4. Algorithm 1 shows the operation flow:

Algorithm 1 Summary Generation Steps

```
Step 1: Identify sentence boundaries in the given set of documents to decompose the document set into individual sentences and form the candidate sentence set $S$=$\{s_i | i = 1, 2, .., n\}$.
Step 2: For each sentence $i \in S$ compute its relevance score using proposed scoring mechanism, then sort the sentences in descending order.
Step 3: Select sentence $i$ with highest relevance score, and move it to the summary set $F$ and remove it from $S$.
Step 4:
while $|F|$ <required summary length do
  Pick the top ranking sentence $k$ in set $S$
  if term overlap between $F$ and $k < w$ where $w$ is redundancy threshold then
    add $k$ to $F$, remove $k$ from $S$
  else
    remove $k$ from $S$
  end if
end while
Step 5: arrange the sentences in F in chronological order i.e. in the order found in the source documents.
```

Following is the example showcasing our technique. The Topic of summary generation is ”Microsoft to open research lab in India”, 8 articles published in different new sources forms the news cluster. A generic summary, and User specific summaries for all user were generated from the news cluster, In the example we are showing the condensed summary(100 words) for two users. User A is from NLP domain and User B from network security domain. The italic text in user specific summary shows the differnce compare to generic summary.

Generic summary: The New Lab, Called Microsoft Research India, Goes Online In January, And Will Be Part Of A Network Of Five Research Labs That Microsoft Runs Worldwide, Said Padmanabhan Anandan, Managing Director Of Microsoft Research India. Microsoft's Mission India, Formally Inaugurated Jan. 12, 2005, Is Microsoft's Third Basic Research Facility Established Outside The United States.In Line With Microsoft's Research Strategy Worldwide,The Bangalore Lab Will Collaborate With And Fund Research At Key Educational Institutions In India, Such As The Indian Institutes Of Technology, Anandan Said.Although Microsoft Research Doesn't Engage In Product Development Itself, Technologies Researchers Create Can Make Their Way Into The Products The Company

User A Specific summary: The New Lab, Called Microsoft Research India, Goes Online In January, And Will Be Part Of A Network Of Five Research Labs That Microsoft Runs Worldwide, Said Padmanabhan Anandan, Managing Director Of Microsoft Research India.Microsoft's Mission India, Formally Inaugurated Jan. 12, 2005, Is Microsoft's Third Basic Research Facility Established Outside The United States. ***Microsoft Will Collaborate With The Government Of India And The Indian Scientific Community To Conduct Research In Indic Language Computing Technologies, This Will Include Areas Such As Machine Translation Between Indian Languages And English, Search And Browsing And Character Recognition.*** In Line With Microsoft's Research Strategy Worldwide,The Bangalore Lab

User B Specific summary: The New Lab, Called Microsoft Research India, Goes Online In January, And Will Be Part Of A Network Of Five Research Labs That Microsoft Runs Worldwide, Said Padmanabhan Anandan, Managing Director Of Microsoft Research India. ***The Newly Announced India Research Group Focuses On Cryptography, Security, Algorithms And Multimedia Security, Ramarathnam Venkatesan, A Leading Cryptographer At Microsoft Research In Redmond, Washington, In The US, Will Head The New Group. Microsoft Research India will conduct a four-week summer school featuring lectures by leading experts in the fields of cryptography, algorithms and security. The program is aimed at senior undergraduate students, graduate students and faculty***

5 Evaluation

The evaluation of this technique was carried out on five different research scholars working in different fields of computer science. Web based profile has been built for each of the researchers. News articles of science and technology domain were considered for summarization. Twenty five different topics were chosen with each topic having 5-10 articles. For each topic a generic and user specific summary was generated for each person. Each researcher was asked to judge the relevance of both versions of summaries for all 25 topics. They have been asked to evaluate the informative ness of summaries on 5 point scale, so each user provide manual scores to both generic and personalize summaries based on how relevant the summaries are for them.

The average scores for both types of summaries of all topics for each researcher is shown in Figure 1. It shows that the users prefer profile based personalized summaries compared to a generic summary given by general automatic summarization system. This means that personalization can benefit the automatic summarization process to improve user satisfaction towards summaries.

Figure 2 shows the scores given by a particular user

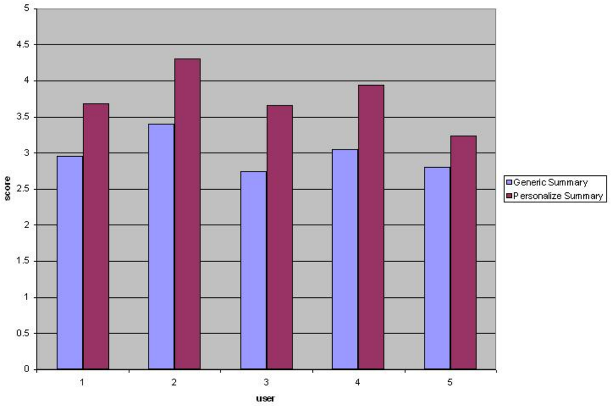

Figure 1. Average Scores for different Users

across different topics. We see that for most of the topics user find personalized summaries relevant for him. Also the personalized summaries for the topics strongly related to the user's domain are more relevant to him. For topics which are not closely related to user's field, the personalized and generic summaries are quite similar. These topics are the ones which got least influenced with personalization. For a few rare topics the user did not find personalized summary better, which may be because of presence of noisy data in their profile.

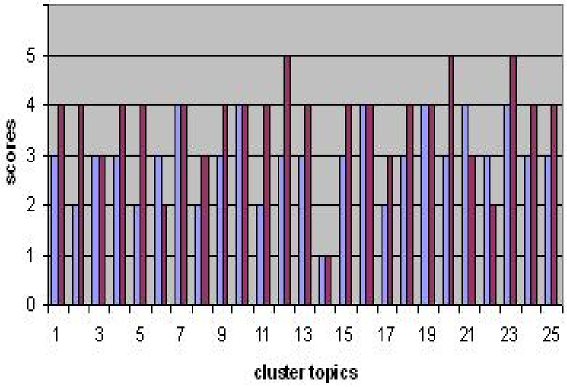

Figure 2. Score of Different topics for a User

6 Conclusion and future work

This paper introduces a document summarization approach for knowledge workers based on their online information available. Web based profile is used to generate user specific summary from a set of news articles. Two types of summaries are produced: generic and user specific. Their performance has been compared by a group of users which shows that the profile based summary is more user satisfactory than a generic summary. The proposed framework of generating user specific summaries is not restricted to web based profile creation for a user. Current results with the proposed user model are encouraging and this motivates us to carry out these experiments with other richer ways of building user background models to benefit more users and community in future. Presently we experimented only with online news articles. This technique can be adapted for summarization of technical papers and journals and this will be a part of our future work.

Acknowledgment

We acknowledge the financial support given by The Nokia Global University Relations Grant in conducting the research reported in this paper.

References

[1] A. Nenkova and L. Vanderwende.:The impact of frequency on summarization. Technical report, MSR-TR-2005-101, 2005.

[2] A. Nenkova, L. Vanderwende, and K. McKeown.: A compositional context sensitive multidocument summarizer. In Proc. of SIGIR, 2006.

[3] Daniel Marcu. From Discourse Structures to Text Summaries. Proceedings of the 14th National Conference on Artificial Intelligence AAAI-97

[4] GJ Rath, A Resnick, TR Savage. The formation of abstracts by the selection of sentences. American Documentation,12(2): 139143, April 1961.

[5] G Salton, A Singhal, M Mitra, C Buckley. Automatic text structuring and summarization. Information Processing and Management,33(2): 193-207, 1997.

[6] I. Mani and M. Maybury: Advances in Automatic Text Summarization.The MIT Press, 1999.

[7] Jade Goldstein, Mark Kantrowitz, Vibhu O. Mittal, and Jaime Carbonell. Summarizing Text Documents:Sentence Selection and Evaluation Metrics. In Proceedings of SIGIR-99, Berkeley, CA, August 1999.

[8] M. Jaoua and A. Ben Hamadou.: Automatic text summarization of scientific articles based on classification of extract's population. In Proceedings of Computational Linguistics and Intelligent Text Processing, 2003.