Web Page Summarization Using Dynamic Content

Adam Jatowt
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, Japan 113-8656
+81-3-58416755
jatowt@miv.t.u-tokyo.ac.jp

Mitsuru Ishizuka
University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, Japan 113-8656
+81-3-58416755
ishizuka@miv.t.u-tokyo.ac.jp

ABSTRACT
Summarizing web pages have recently gained much attention from researchers. Until now two main types of approaches have been proposed for this task: content- and context-based methods. Both of them assume fixed content and characteristics of web documents without considering their dynamic nature. However the volatility of information published on the Internet argue for the implementation of more time-aware techniques. This paper proposes a new approach towards automatic web page description, which extends the concept of a web page by the temporal dimension. Our method provides a broader view on web document summarization and can complement the existing techniques.

Categories and Subject Descriptors
H.4.0 [**Information systems Applications**]: General; I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis.*

General Terms
Algorithms

Keywords
web page summarization, change detection, web document

1. INTRODUCTION
The web is abundant in information, which often changes or evolves as time passes. Thus summarizing web pages basing only on their momentary contents faces risk of producing incomplete or skewed results. Considering several versions of the same web document obtained over specified time period offers the possibility for grasping long-term topic of the web page. In order to gain high confidence that a particular web page is about a given topic one should consider possibly many versions of this page. This confidence is related to the number of document samples and the time lag between their retrieval.

Current methods of web page summarizing can be basically divided into two groups: content- and context-based ones. The first type utilizes textual content of the web documents in question [2][3]. This concept does not differ much from the traditional document summarization approaches. However some new possibilities or obstacles arise here due to such characteristics of HTML documents as: markup language, flexibility of structure or existence of graphical elements. The disadvantage of this method becomes evident when a particular web page contains little textual content relying mostly on visual language communication.

Since a web page alone might not always provide any ready clues about its meaning an alternative approach was conceived. Context-based methods, which are making use of the hypertext structure of the web, have been recently proposed [1][4]. Usually the most common procedure is to exploit the textual content retrieved from the documents which link to the web page in question. Paragraphs or other text units that are close to the links pointing to the particular document are used to create the summary. By contrast to the content-based methods the final summary can be constructed using already created abstracts or other contextual data. Hence it is not limited to the expressions and terms extracted only from the summarized web page. However utilizing user-created abstracts or hints poses a risk of subjectivity. Different people have different opinions, knowledge, relationships or aims towards particular topics. Therefore, it can happen that the contextual information found on neighboring pages may be influenced by a variety of factors. Additionally, there can be a shortage of linking documents or available information for new or unknown web pages. In such cases it is necessary to combine several methods for producing a correct summary.

Our approach is an extension of content-based techniques but differently to them it considers a web page as a dynamic object. It means that we focus on the changing content retrieved from given number of versions of a web document. A similar method was proposed for summarization of changes in topical web collections [5]. We retrieve the consecutive versions of the same web page and compare them for different parts. In this way the content of a page is not restricted to only one moment but is extended in time. The increase in the volume of available data for summarization should result in better estimation of the document topic and characteristics. One example where the proposed technique can be used are search engines, which periodically crawl web pages and can store past versions of web documents.

2. METHODOLOGY
First, consecutive versions of a web page need to be fetched and stored. Let t be the time between the retrieval of sequential page versions and let T denote the whole period of summarization. The shorter is t, the higher will be the precision of summary for interval T due to the increased possibility of retrieving all short-life content. Changes in the web document can be found by comparing its consecutive versions. This comparison is carried out on the sentence level. It means that any sentence appearing in the new version of the document while absent from the previous one is considered as a change or an insertion. The total change for two consecutive web page versions is the union of extracted

WWW 2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

insertions and will be called a dynamic part of the document version. In this way the content of a web page is divided into two groups: static and a dynamic one.

After having fetched all versions of a web page during period *T* the standard text processing steps are taken, such as stop-word elimination and stemming. Next, terms are extracted from each page version as single words and bi-grams.

We have assumed the following hypothesis: the higher is the number of dynamic parts containing the particular term, the more important is this term. The frequency of a term inside every single document also plays a role, although a smaller one, in estimating the significance of the term. Therefore the term score (S_i) has the following form:

$$S_i = \frac{\sum_{j=1}^{N} \frac{n_{ij}}{n_{cj}}}{N} * \log\left(\frac{N_{di}}{N}\right) \quad (1)$$

Equation 1 is derived from the well-known *TFIDF* weighting scheme [6] and is composed of two parts. The first one estimates the average term frequency in the dynamic parts of web page versions. N and N_{di} are respectively: the number of all page versions and the number of versions with dynamic parts containing term *i.* The total amount of terms in a version *j* is denoted by n_{cj} while the number of instances of the term *i* in the dynamic part of this page sample is expressed as n_{ij} . Second section of the equation is a logarithm of the ratio of the documents with their dynamic parts containing at least one instance of the term. It gives preference to terms appearing often in insertions of different page versions. Additionally, the term score can be influenced by the number and the type of emphasizing or structural tags related to the term such as for example: <H1>, <b> or <title>.

The above method works well for web pages, which have rather dynamic character. By such pages we understand documents, whose substantial parts of the content change frequently. A different approach in needed for web pages, which have little changes throughout the summarization period. In such a case, one has to rely more on static part of a page. For a completely stationary web document we can only use static web page summarization methods. To evaluate the scope of the content changes we introduce a volatility parameter α , which is an average ratio of the size of insertions, which are denoted by n_{dj} , to the size of the total content for every version of a web page.

$$\alpha = \sum_{j=1}^{N} \frac{n_{dj}}{n_{cj}} \Big/ N \quad (2)$$

The value of the parameter α equal to *0* indicates a static document during interval *T* while value *1* characterizes a web page whose consecutive versions do not contain any common sentence. In a simplistic way we can decide that only for web pages with α value higher than *0.5* one should apply summarization techniques focused on dynamic parts of a document. However other approaches may use α for deciding the extent to what dynamic and static summarization methods could be used together for constructing the summary.

In the last step the scores of all sentences from the dynamic parts of document versions are calculated. Sentence weight is the average score of its terms. Sentences with scores above a defined threshold are arranged according to their temporal order and the relative sequence in the web page and presented to the user.

Web documents, which have high value of parameter α calculated for relatively low ratio of *t* to *T* often turn out to be newswire sources. Such type of web documents can discuss quite different concepts in short time and can have low topical continuity. In this case it might be difficult to create a coherent and meaningful summary especially for rather short intervals *T* due to the extended range of diverse topics covered by a document.

3. CONCLUSIONS

In the time of an increasing popularity of the Web, web documents should no longer be treated as static objects. The momentary content found on pages may be insufficient for the description of their long-term topics and characteristics. Context-based summarization techniques can be degraded by the subjectivity factor in other pages or simply by the quantity of contextual information. Therefore we propose a new approach, which advocates extracting and summarizing changes from consecutive web page versions over specified time intervals. By detecting similar concepts from instantaneous snapshots of the same page one can more correctly grasp the essence of the page.

However, the main obstacles for implementing the proposed solution are the lack of available web archives and a fairy static or scarce content of some web pages. Therefore we believe that our algorithm may be a sort of a complement to the existing methods of web page summarization rather then an alternative to them.

4. REFERENCES

[1] Amitay, E., and Paris, C. Automatically summarizing web sites: is there any way around it? Proc. of the 9th International Conference on Information and Knowledge Management (McLean, Virginia, November 2000), 173-179.

[2] Berger, A. L., and Mittal, V. O. OCELOT: a system for summarizing web pages. Proc. of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval (Athens, Greece, July 2000), 144-151.

[3] Buyukkokten, O., Garcia-Molina, H., and Paepcke, A. Seeing the whole in parts: text summarization for web browsing on handheld devices. Proc. of the 10th International WWW Conference (Hong Kong, May 2001), 652-662.

[4] Glover, E. J., Tsioutsiouliklis, K., Lawrance, S., Pennock, D. M., and Flake, G. W. Using web structure for classifying and describing web pages. Proc. of 11th International WWW Conference (Honolulu, Hawaii, May 2002), 562-569.

[5] Khoo, K. B., and Ishizuka, M. Emerging Topic Tracking System. Proc. of Web Intelligence Conference 2001. (Maebashi City, Japan, October 2001), 125-130.

[6] Salton, G. and Buckley, C., Term weighting approaches in automatic text retrieval. Information Processing and Management vol.24, no 5, (1988) 513-523.