Summarization of Dynamic Content in Web Collections

Adam Jatowt and Mitsuru Ishizuka

University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan
{jatowt,ishizuka}@miv.t.u-tokyo.ac.jp

Abstract. This paper describes a new research proposal of multi-document summarization of dynamic content in web pages. Much information is lost in the Web due to the temporal character of web documents. Therefore adapting summarization techniques to the web genre is a promising task. The aim of our research is to provide methods for summarizing volatile content retrieved from collections of topically related web pages over defined time periods. The resulting summary ideally would reflect the most popular topics and concepts found in retrospective web collections. Because of the content and time diversities of web changes, it is necessary to apply different techniques than standard methods used for static documents. In this paper we propose an initial solution to this summarization problem. Our approach exploits temporal similarities between web pages by utilizing sliding window concept over dynamic parts of the collection.

1 Introduction

In document summarization research summaries are usually built from newspaper articles or some static documents. However in the age of the growing importance of the Web, it is becoming necessary to focus more on the summarization of web pages. Until now, few methods have been proposed that are especially designed for summarization in web genre (e.g., [3], [4]). The Web is a dynamic and heterogeneous environment. These characteristics cause difficulties for adapting traditional text analysis techniques into the web space. One of the most important differences between web pages and other document formats is the capability of the latter ones to change their content and structure in time. Many popular web pages continuously change, evolve and provide new information. Thus one should regard a web document as a dynamic object or as a kind of slot assigned to the URL address. This approach enables to consider volatile content, which is inserted or deleted from web documents for summary creation. Such summarization task differs from the standard multi-document summarization in the sense that it focuses on the changed contents of web pages (Figure 1).

There are several cases where summarization of changes in web documents could be beneficial. For example a user may be interested in knowing what was popular in his favorite web collection during given period of time. It can be too difficult for him to manually access each web document for discovering important changes. By carefully choosing information sources, one can construct a web collection which is informative about a particular topic. Such collection would be considered as a single, complex information source about the user's area of interest. Then main events and popular changes concerning user-defined topic could be acquired to the extent, which depends on the quality and characteristics of the input collection.

J.-F. Boulicaut et al. (Eds.): PKDD 2004, LNAI 3202, pp. 245–254, 2004.

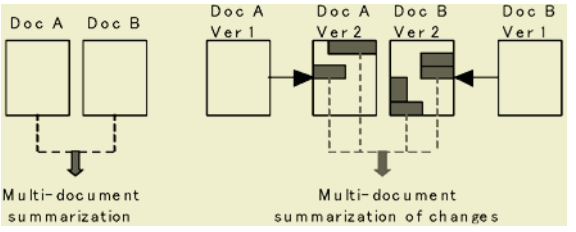

Fig. 1. Difference between traditional and new summarization. In the new one temporal versions of two or more documents are compared to reveal their changes, which are later summarized.

Another motivation for our research comes from the observation that current search engines cannot retrieve all changing data of web pages. Thus much information is lost because the web content changes too fast for any system to crawl and store every modified version of documents.

The method presented in this paper can be generally applied to any types of web pages. However, perhaps some modified approach could be more efficient for particular kinds of web documents like for example newswires, company web pages or mailing lists. Anyway, due to the large number of different page types and the difficulty of their classification we have attempted to provide generic summarization solutions, which are not tailored for any specific kinds of documents. Another concern is that different types of web pages have different frequencies and sizes of changes. Our approach works well for dynamic web pages, which have enough changing content so that meaningful summaries can be created. Therefore for rather static web documents the output may not be satisfactory enough and, in such a case, some existing document summarization methods (e.g., [8]) could work better. The speed and the size of changes of a web page can be approximated as the average change frequency and the average size of changes over the whole summarization period. Additionally to obtain a meaningful summary there should be a continuity of topics in temporal versions of a web page. Therefore we make an assumption here that the topical domain and the main characteristics of a document do not change rapidly so that a short-term summarization could be feasible. In other words, we assume semantical and structural continuity of different versions of the same web page.

In our approach we have focused generally on singular web pages. Thus any links or neighboring pages have been neglected. However the algorithm can be extended to work with the collection of web sites or any groups of linked web documents. In these cases a given depth of penetration can be defined for tracking such networks of web pages. For example, we can examine any pages, which are linked from the company home page that is pages such as: company products, staff, vacancies etc. An intuitive solution is to combine all these pages together into one single document representing the selected part of the web site. The content of each joined page could have lower scores assigned depending on the distance from the starting web page. In this way all web sites or other sub-groups of pages in the collection would be treated as single web documents where the connectivity-based weighting scheme is applied to the content of every single page.

The rest of the paper is organized as follows. In the next section we discuss related research work. Sections 3 and 4 present dynamic characteristics of web collections and our methodology for summarization of changes. In Section 5 the results of the

experiments are demonstrated and discussed. Finally, the last section contains conclusions and future research plans.

2 Related Work

Topic Detection and Tracking (TDT) (e.g., [2]) is the most advanced research area which focuses on automatic processing of information from news articles. TDT attempts to recognize and classify events from online news streams or from retrospective news corpora. In our case we want to use collections of arbitrary kinds of web pages rather then news articles only. Thus we aim at detecting not only events reported by newswires but any popular concepts in a given topical domain representing user's interest.

Additionally, TDT or other automatic news mining applications like for example Google News [5] concentrate more on tracking and detecting particular events than on generating their topical summaries. The part of research, which centers on temporal summarization of news articles is represented by: [1], [9], [11], [14]. In [1] novelty and usefulness measures are applied for sentences extracted from newswire resources in order to generate temporal summaries of news topics. Newsblaster [9] or NewsInEssence [11] are other examples of applications developed for producing automatic summaries of popular events. The authors use some pre-selected resources of the newswire type for input data. Finally TimeMines [14] is a system for finding and grouping significant features in documents based on chi-square test.

There is a need for an application that could summarize new information from any, decided by users, kinds of resources. WebInEssence [12] is a web-based multi-document summarization and recommendation system that meets the above requirement. However, our approach is different in the sense that we attempt to do temporal summarization of documents, that is, summarization of their "changes" or dynamic content, instead of considering web pages as static objects. Temporal single-document summarization of web documents has been recently proposed in [7]. Multi-document summarization of common changes in online web collections has been shown in ChangeSummarizer system [6], which uses web page ranking and static contexts of dynamic document parts. Nevertheless, despite of the popularity of Web, there is still a lack of applications for retrospective summarization of changes in web documents.

3 Changes in Web Collections

There are two simple methods for obtaining topical collections of web pages. In the first case one may use any available web directory like for example ODP [10]. However, there is quite a limited number of topical choices in existing web directories, which additionally may have outdated contents. It means that a user cannot choose any arbitrary topic that he or she requires but is rather restricted to the pre-defined, general hierarchy of domains. Another straightforward way to obtain a web collection is to use search engine. In this case any combination of terms can be issued providing more freedom of choice. However the responding set of web pages may not always be completely relevant to the user's interest. Therefore an additional examination of search results is often necessary. Additionally, one should also filter collected documents to reject any duplicate web pages since they could considerably degenerate the final summary output.

In the next step, web page versions are automatically downloaded with some defined frequency. The interval *t* between the retrieval of each next version of a single web page should be chosen depending on the temporal characteristics of the collection. The longer period *t*, the lower the recall of changes is due to the existence of short-life content as it often happens in the case of newswire or popular pages. Some parts of web pages may change more than one time during interval *t* what poses a risk that the information can be lost. On the other hand, high frequency of page sampling should result in the increased recall but naturally also in the higher usage of network resources. Let $C_a = \{C_1, C_2, ..., C_n\}$ be a set of all changes occurring in a single web page *a* during given interval and $F_a = \{F_1, F_2, ..., F_u\}$ a set of discovered changes. If we assume that the page changes with a constant frequency t_a then the recall of changes can be approximated as:

$$R_a = \frac{|F_a|}{|C_a|} \approx \frac{t_a}{t} \quad \text{if} \quad t_a \le t \,. \qquad R_a = \frac{|F_a|}{|C_a|} = 1 \quad \text{if} \quad t_a > t \,. \tag{1}$$

Let *T* denote the whole time interval for which a summary will be created. Assuming short period *T*, which embraces only a few intervals *t*, we obtain a small number of pages that have any changes. In this case the influence of these web pages on the final summary will be relatively high. Thus probably the final summary could have lower quality with regards to the real changes in the topic of collection, since only few web pages are determining the summary. In case of a choice of long *T* containing many intervals *t*, we expect more changes to be detected, which cause a low influence of a single web page on the final summary.

For two similar web pages the delay of the reaction to an event occurring at a particular point of time can be different. We assume that these web pages always report the most important and popular events concerning user's area of interest. It is usually expected that a newswire source would mention the particular information in a matter of hours or days. However it may take longer time in the case of other type of web pages which are more static. We call this difference a "time diversity" of web pages in order to distinguish it from the "content diversity". The choice of too short *T* may result in poor precision of the final summary because the reactions to a particular event could be spread in time in different web pages. However, on the other hand, longer *T* increases the probability that many unrelated and off-topic changes from the collection are taken into consideration what may cause the reduced quality of an output.

4 Methodology

To extract the changing content, two consecutive versions of every web page are compared with each other. The comparison is done on the sentence level. Sentences from proximate versions of a document are compared so that inserted and deleted ones can be detected. We have decided to focus only on a textual content of web documents. Thus pictures and other multimedia are discarded. There are two types of

textual changes that can occur in a page: an insertion and a deletion. If a particular sentence appears only in the later version of a web page then it is regarded as an insertion. In case it can be found only in the previous version we define such sentence as a deletion.

Next, standard text preprocessing steps are conducted such as stemming and stopwords removal. We consider words and bi-grams extracted from the changes in the collection as a selected pool of features. Each such a term is scored depending on its distribution in the dynamic parts of collection during interval *T*. The term scoring method assumes that popular concepts are found in the same type of changes in high number of web page versions which are in close proximity to each other. Therefore terms appearing in changed parts of many documents will have higher scores assigned than the ones that are found in changed sections of only a few web pages (Equation 2). Moreover, a term that appears frequently inside the changes of many versions of a single web page should also have its score increased. However, in the concept of the "popularity", document frequency of the term is more important than its term frequency therefore the equation part concerning document frequency has an exponential character. Document frequency *DF* is the number of document versions that contain given term. Term frequency TF_j is the frequency of the term inside the dynamic part of a single document version *j*. In Equation 2 term frequency is divided by the number of all terms inside each change, that is the size of a change S_j, and averaged over all web page samples *N*n* where *N* is the number of different web documents and *n* the number of versions of each web page. In general, the basic scoring scheme is similar to the well-known *TFIDF* weighting [13].

$$S_{term} = \frac{\sum_{j=1}^{N*n} \frac{TF_j}{S_j}}{N*n} * \exp(DF) \tag{2}$$

As it has been mentioned before there are two possible types of changes: the deleted and the inserted change. Intuitively, deletions should be considered as a kind of out-dated content, which is no longer important and thus can be replaced by a new text. However if many web documents have deleted similar content in the same time then one may expect that some important event, expressed by this content, has been completed. In this case terms occurring in such deletions should have high value of importance assigned. On the other hand terms which are found in many proximate insertions will also have high scores. Finally the overall score of a term will be a combination of both partial scores calculated over deleted and inserted textual contents.

Let d_x be a deletion and i_x an insertion of a single web page, where *x* is the number of the web document version (Figure 2). Specifically d_x, i_x indicate the content that was deleted from the *x-1* version and the content that was added to the *x* version of the web page.

The total amount of deletions *D* of a single web page over period *T*, called a "negative change" of the page, is the union of the deleted text for all page versions:

$$D = \bigcup_{x=1}^{x=n} d_x \,. \tag{3}$$

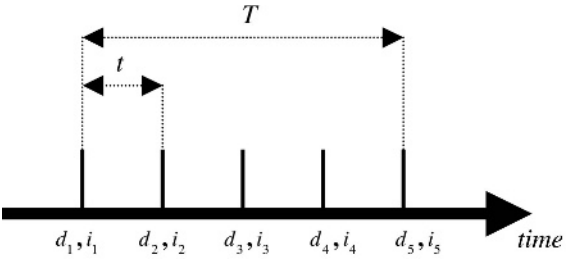

Fig. 2. Temporal representation of changes in a web page.

On the other hand, the whole pool of insertions expressed by *I* is described as a "positive change" of the document (Equation 4).

$$I = \bigcup_{x=1}^{x=n} i_x \,. \tag{4}$$

We want to assign maximum weights to terms, which are inserted or deleted from high number of documents in proximate time. In this way we use temporal similarity of terms in a similar fashion as TDT considers bursts of event-type data. The range of this similarity is determined by the user in the form of a sliding window of length *L*. The window moves through the sequentially ordered collection so that only *L/t* versions of each web document are considered in the same time (Figure 3). Terms are extracted from the positive and negative types of changes inside every window and are scored according to the weighting scheme from Equation 1. However, now the differences between document and term frequencies in both kinds of changes are considered. The score of a term in each window position is denoted by S_{term}^{win} and expressed as:

$$S_{term}^{win} = \frac{\sum_{j=1}^{N*n} \left| \frac{TF_j^I}{S_j^I} - \frac{TF_j^D}{S_j^D} \right| * \exp\left| DF^I - DF^D \right|}{N * n * L}. \tag{5}$$

In this equation the superscripts *I* and *D* denote the respecting types of changes inside one window position. Thus the term and document frequencies of each term are calculated only for the area restricted by the window. The overall term score (Equation 6) is the average distribution of the term in changes inside all window positions *Nw*.

$$S_{term}^{overall} = \frac{\sum_{win=1}^{Nw} S_{term}^{win}}{Nw} \,. \tag{6}$$

If inside many window positions a term was occurring generally in one type of changes then its overall score will be quite high. On the other hand, terms, which exhibit almost equal distributions in positive and negative types of changes inside majority of window positions, will have assigned low scores. In other words we favor terms that occur in bursts of deletions or bursts of insertions in the substantial number of window positions. This is implemented by considering the absolute values of differences in term and document frequencies of both types of changes (Equation 5). The

length of the window is chosen by the user depending on whether short- or long-term concepts are to be discovered.

In the last step, sentences containing popular concepts are extracted and presented to the user as a summary for period *T*. To select such sentences we calculate the average term score for each sentence in the changes of the collection and retrieve the ones with the highest scores. The length of the summary is specified by the user. We also implement a user-defined limit of sentences, which can be extracted from a single document version. This restriction is put in order to avoid situations where one or only a few documents will dominate the final summary. To increase the summary understandability we add preceding and following sentences surrounding selected, top-scored sentences. Additionally to minimize summary redundancy we calculate cosine similarities between all sentences and reject the redundant ones. Lastly, sentences are arranged in the temporal order and are provided with links to their original web documents to enable users the access to the remaining parts of pages.

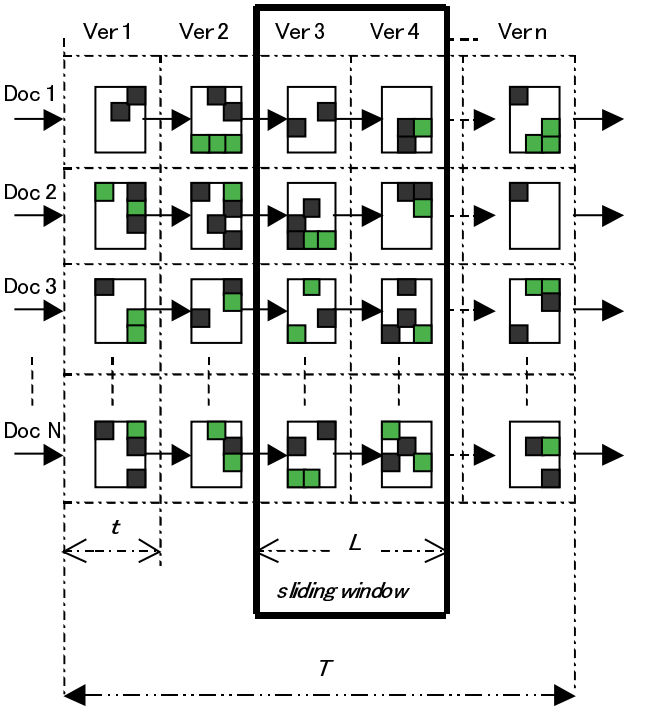

Fig. 3. Sliding window in the web collection of *N* documents with *n* versions. Dark areas symbolize insertions and deletions in web page versions.

5 Experiment

The results of our experiment are presented for the web collection which was built after issuing the query "EU enlargement" to the search engine and downloading 200 top-ranked web pages. We have manually filtered duplicate pages and documents which were topically distant from the meaning of the query. Change tracking was

performed with the time delay t of 3 days during interval from 12th March to 12th May 2004. Table 1 displays the top scored sentences for this query.

Unfortunately, to the best of our knowledge there is no annotated corpus available, which could be used for the evaluation of change summarization in web collections. Since the creation of such corpus is not a straightforward task we have restricted the evaluation here to presenting an example of a summary and to discussing a simplified, experimental scenario, which shows the influence of different window lengths on the term score. Given the diversity of the types of web pages and discussed topics for this quite general query, it should not be surprising that results may not constitute coherent and high quality summary. Intuitively, it is very important to construct the appropriate web collection with closely related pages. We have noticed that results are better for narrow, topics, where documents tend to be more related topically with each other.

Table 1. Summary for "EU enlargement" query.

Europe reunited means a stronger, democratic and more stable continent, with a single market providing economic benefits for all its 450 million citizens. The European Union has come a long way since the original six member states joined forces to create the European Coal and Steel Community in 1951 and the European Economic Community in 1957, calling upon the peoples of Europe "who share their ideas to join their efforts." The six became nine in 1973, and had grown to 15 by 1995.
An enlarged EU: an opportunity for health? 1st May 2004 represents a both historical and symbolic landmark in the process of European integration. Initiated more than 50 years ago, the concept of the European Union now includes some of the countries that once belonged to the former Communist block. While undoubtedly a milestone in the "ever wider and closer" this new enlargement also raises a whole range of serious challenges particularly in terms of health and health policies.
Bulgaria has made economic progress but still has a long way to go and must close down nuclear power plants. The state of the Romanian economy is far from where it needs to be for EU accession. It must make more progress and improve its child-care institutions.
Never before has the European Union invited such a large group of countries which has had such a remarkably different social and economic system. This challenge will be especially large in the field of Structural Funds, the mechanism of the European Union that aims to achieve economic and social cohesion across the European territory.

Let us imagine a situation when only two instances of the same term are present in the changes of a collection. The term occurs once as a deletion and once as an insertion of a single document during period T. The rest of the changes of the collection can be empty. Different relative locations of both instances of the term should result in adequate term scores. In Figure 4 we plot the score of the term against the relative positions of its both instances. One instance of the term, for example an insertion, is fixed in the middle of the horizontal axis (value 9 in Figure 4) representing sequential versions of documents. In other words the term is occurring in the inserted part of 9th version of the document. The other instance (deletion) can be placed in any position in the collection. We will put it in every consecutive version of the document and compute term scores for all such positions. Thus starting from the beginning of the horizontal axis until the end of T we move the deletion instance and calculate the score of the term. In result the term score is plotted against all possible distances between both term instances. The score depends also on the window length. In Figure 4 three different lengths of the sliding window are used for the score calculation. Thus

in general, the graph shows the influence of the size of the sliding window and the temporal closeness of both changes (temporal distance between the term instances) on the final term score.

From Figure 4 we see that in the middle of the horizontal axis the overall score is equal to zero since both instances of the term nullify each other. The relative term score is decreasing when the distance between both term instances becomes smaller. This decline starts later for windows with smaller length *L*, which embrace only few web page versions. Thus for a short length of the window the algorithm works in a short-term mode and assigns high scores to terms which can change often in close time throughout the summarization interval. On the other hand, for the high value of *L* long-term changes are favored, so in this case, the temporal granularity of an event discovery is diminished. Therefore for wider windows some short-life events can be overlooked. However, unlike in the former case there is the advantage of the reduced effect of temporal diversity between different web pages.

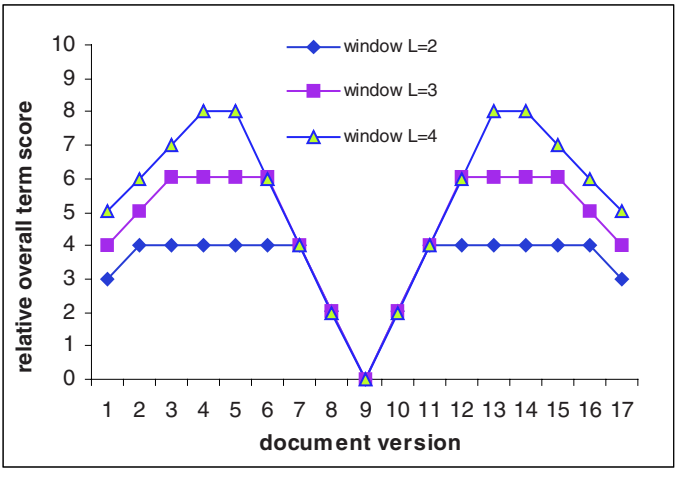

Fig. 4. Score for two opposite instances of the same term for different window lengths where the first instance occurs only in the middle of period *T* (the center of the horizontal axis) while the second one is placed in any document version.

6 Conclusions

We have introduced a new research area of summarization of dynamic content in retrospective web collections and have proposed an initial method, which employs sliding window over insertion and deletion types of changes. Our approach focuses on the temporal aspects of summarization in web genre. We have proposed to treat deletions as a part of dynamic data of web documents and invented a combined scoring approach for both types of changes. The advantages and challenges of summarization in dynamic web collections have been also discussed.

Currently we investigate evaluation methods which would enable us to compare different approaches to this summarization task. In the future we also would like to focus on the summarization of changes of the whole web sites and to make experiments with diverse kinds of web pages. Apart from that, we would like to take into

consideration more attributes of web documents. Besides textual content there are some other changeable elements of web pages that can be exploited for summarization purposes.

References

1. Allan, J., Gupta, R., Khandelwal, V.: Temporal Summaries of News Topics. Proceedings of the 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, USA (2001) 10-18
2. Allan, J. (ed.): Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Publishers. Norwell MA, USA (2002)
3. Berger, A. L., Mittal, V. O.: OCELOT: a System for Summarizing Web Pages. Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval. Athens, Greece (2000) 144-151
4. Buyukkokten, O., Garcia-Molina, H., Paepcke, A.: Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. Proceedings of the 10th International WWW Conference. Hong Kong, Hong Kong (2001) 652-662
5. Google News: http://news.google.com
6. Jatowt, A., Khoo, K. B., Ishizuka, M.: Change Summarization in Web Collections. Proceedings of the 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. Ottawa, Canada (2004) 653-662
7. Jatowt, A., Ishizuka, M.: Web Page Summarization Using Dynamic Content. Proceedings of the 13th International World Wide Web Conference. New York, USA (2004) 344-345
8. Mani, I., Maybury, M.T. (eds.): Advances in Automatic Text Summarization. MIT Press, Cambridge MA, USA (1999)
9. 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. Proceedings of Human Language Technology Conference. San Diego, USA (2002)
10. Open Directory Project (ODP): http://dmoz.org
11. 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 (2001)
12. Radev, D., Fan, W., Zhang, Z.: WebInEssence: A Personalized Web-Based Multi-Document Summarization and Recommendation System. In NAACL 2001 Workshop on Automatic Summarization. Pittsburgh, USA (2001) 79-88
13. Salton, G., Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, Vol. 24, No 5, (1988) 513-523
14. Swan, R., Jensen, D.: TimeMines: Constructing Timelines with Statistical Models of Word Usage. In ACM SIGKDD 2000 Workshop on Text Mining, Boston MA, USA (2000) 73-80