Temporal Web Page Summarization

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. In the recent years the Web has become an important medium for communication and information storage. As this trend is predicted to continue, it is necessary to provide efficient solutions for retrieving and processing information found in WWW. In this paper we present a new method for temporal web page summarization based on trend and variance analysis. In the temporal summarization web documents are treated as dynamic objects that have changing contents and characteristics. The sequential versions of a single web page are retrieved during predefined time interval for which the summary is to be constructed. The resulting summary should represent the most popular, evolving concepts which are found in web document versions. The proposed method can be also used for summarization of dynamic collections of topically related web pages.

1 Introduction

Web pages can change their contents any number of times. It is tempting to analyze the changing content retrieved from temporally distributed versions of web documents. The amount of data available for summarization could be increased by considering dynamic and static content of a web document. In this paper we provide methods for summarization of multiple, spread in time versions of a single web document.

Our aim is to provide a summary of main concepts and topics discussed in the web page over given time period. There is an important difference between multi-document summarization and the summarization of multiple versions of a single web page. Usually for the former one there are several documents discussing the same event or topic. However temporal versions of a single web page do not always contain the same concepts. The stress is put here more on the time dimension and the evolution of the content due to the temporal character of data. Therefore for temporal summarization one should use text mining techniques that can help with extracting common, related information which is distributed in some defined time interval. We assume here that the scope of topics and the structure of a document in question do not change rapidly so that a short-term summarization can be performed.

There are several situations when single or multi-document temporal summaries can be of some use. For example a user can be interested in main changes or popular topics in his favorite web page, web site or the collection of pages during given period. It can happen that for some reasons he may not be able to track all changes in the

chosen web page or the group of pages. Moreover, due to the large size of new content it may be impossible to manually compare each document version. Thus automatic methods for the summarization of temporal contents of web pages could be helpful. Such summaries could be also beneficial from the commercial point of view. For example companies may want to gather and summarize important information from their competitors' websites.

The summarization approach proposed here is suitable for dynamic or "active" type of web pages, which have enough changing content so that successful summaries can be constructed. The applicability of a web document for such a summarization can be simply estimated by calculating the average change frequency and the average volume of changes of the document during the desired period.

Additionally, apart from analyzing single web documents, our algorithm can be applied for multi-document temporal summarization. This task assumes summarization of topically related collections of web pages over some time intervals [7], [9].

This paper is organized as follows. In the next section we discuss the related research work. Section 3 explains the method used for term scoring. In Section 4 we introduce the sentence selection and the sentence ordering algorithms. In the next section the results of our experiment are presented. We conclude and discuss future research in the last section.

2 Related Research

Summarization of web pages is particularly useful for handheld devices whose size limitations require converting web documents to more compact forms. The straightforward approach utilizes textual content of web documents in question [4], [5] together with additional structural or presentational information like HTML tags. This method works well for web documents which contain enough textual content and few multimedia. On the other hand, there are also context-based methods which are making use of hypertext structure of the Web [3], [6]. They typically exploit parts of content like anchor texts which are found in documents linking to the chosen web page. Usually anchor texts contain useful information about target web documents or their summaries.

Recently, summarization of temporal versions of web documents has been proposed as a way of describing distributed in time content and topics of web pages [8]. This approach can be used to summarize web documents which have frequently updated content. Consecutive web page versions are combined together as an input for summarization instead of analyzing only momentary snapshots of web documents' content.

On the other hand, temporal summarization in web collections of related pages has been proposed in [9] and [7]. ChangeSummarizer [9] is an online summarization system for detecting and abstracting popular changes in the groups of web documents which uses temporal ranking of web pages. In [7] a sliding window is applied on retrospective collection of web documents whose content changes are separated into deletion and insertion type.

Topic Detection and Tracking (TDT) [2] is another research area that is close to our work. TDT focuses on detection, clustering and classification of news articles from online news streams or retrospective corpora. Temporal summarization of news events was presented in [1] where novelty and usefulness of sentences retrieved from newswire streams are calculated for the construction of a summary.

Statistical methods have been used by some researchers for detecting trends and mining textual collections. In [12] comparison of probability distributions was done for trend detection in news collections. On the other hand, regression method was used in [11] to enhance IR efficiency by considering the changing meaning of words in long time spans. Lent et al. [10] proposed a query searching of shapes of trends by using Shape Definition Language (SDL) in a patent database.

3 Term Scoring

First, a user has to specify the length of an interval *T* for which he requires a summary of the selected web page (Figure 1). Additionally, the sampling frequency for fetching web page versions should be defined. This frequency determines the number of web document samples and hence, the amount of available content to be analyzed. High sampling frequency results in higher probability of detecting all changes, which have occurred in the web page during the period *T*. It is especially critical for fast-changing web documents. Therefore the user should choose a tracking scheme which is appropriate for his web page. Document versions will be automatically downloaded during the interval *T* and their content will be extracted by discarding HTML tags. We have limited our focus to the textual contents of web documents. Thus pictures and other multimedia are rejected. In the next step frequent words are eliminated by using stop-word list. Then, each word is subjected to stemming. Except for single words we also use bi-grams as basic features for the summarization. Thus the final pool of terms contains single words and bi-grams which have been filtered by stop-word list and stemmed.

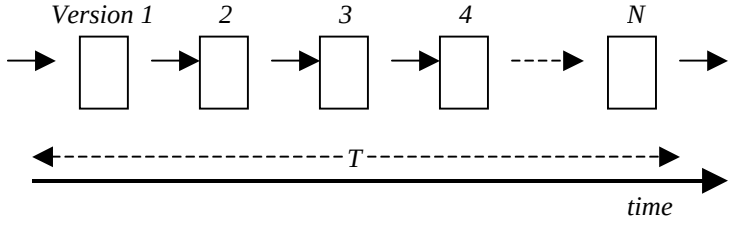

Fig. 1. Temporal versions of a web page

For each term its frequencies are calculated for every web page version. Because the document size may change in time term frequencies are divided by the total number of terms in the given web page sample. Consequently, term weight will reflect the relative importance of the term in the page version. Thus the weight of a term, which has fixed frequency in the collection, can change if the document size varies in different versions of the web page. For example, when a large part of the document

ent versions of the web page. For example, when a large part of the document has been deleted then the relative importance of a given term will increase even if the frequency of this term has not changed.

As it was mentioned before, temporal summary of the collection of multiple, related web pages can also be created using the proposed method. In this case instead of calculating the frequency of a term in each web page version one should use the document frequency of the term. Document frequency is a ratio of the number of web pages that contain the term to the whole number of different web documents in a given time point of tracking process. After calculating document frequencies of terms for all points during *T* we may observe how their importance was changing in time.

In the next step terms are scored according to their distributions in web page versions. We are looking for a scoring scheme, which would favor popular and active terms in the collection. If some concept or topic is becoming popular then it is usually often discussed throughout the collection of web page versions. However, just only the level of popularity estimated as the average term frequency does not reflect properly the significance of the term. Terms, which occur frequently in many web page versions, do not have to be important from the point of view of the temporal summarization. For example high but equal frequency of a word in the large number of consecutive versions may indicate that no change has occurred in concepts or events associated with this term.

To represent long-term changes in the frequency of a term during the interval *T* we apply simple regression analysis. Regression is often used to summarize relationships between two variables, which in our case are: time and the frequency of a term. We have restricted our analysis to simple regression because of its simplicity and the sparseness of our data. However some better fitting regression would be necessary to more correctly approximate changes in the importance of a term. The slope *S* and the intercept *I* are calculated for every term so that scatter plots of terms can be fit with regression lines. Slope value informs whether term frequency is stable, rising or decreasing during the specified period. Intercept point, on the other hand, can be used to check how popular a given term was at the beginning of the web page monitoring. Additionally for every term, its variance *V* is calculated. Variance conveys information about the average magnitude of changes in the term importance for the interval *T*. To determine which terms are significant for the summary we need to compare their statistical parameters. The slope of a term different from zero is an indicator of an emerging or disappearing event or concept. A term that has a clearly rising trend whose value of the slope is much higher than zero should have a high score assigned. If additionally such a term has a low intercept point we can conclude that there is a high probability for the term to discuss a new, emerging topic. A falling trend, on the other hand, may indicate a term, which is associated with an already accomplished event or with a concept that is no longer important. The score of such a term should have a low value assigned providing that a user is interested only in emerging or ongoing events. However if the user wants to construct a general summary of any popular concepts in the web page then terms with evidently falling trend lines could also be useful. On the other hand, a term with nearly horizontal slope is probably not an active word in the sense that its average importance remains on the same level throughout the summarization period. To test this hypothesis the term variance should also be exam-

ined. The low value of variance together with the slope value being close to zero indicates a term, which probably was static in the document versions throughout the interval *T*. Its importance remained on the same level because it was occurring mostly in the static parts of many web page versions. Finally, the intercept plays a role of a novelty indicator showing how popular given terms were before the beginning of the interval *T*. Thus the intercept allows us to choose terms that are not only popular but also novel during tracking process.

In general the score of a term is defined by the following formula.

$$W_i = \alpha * \frac{S_i^r}{N_s^r} + \beta * \frac{V_i^r}{N_v^r} + \delta * \frac{I_i^r}{N_I^r} . \quad (1)$$

The score W_i of a term *i* is expressed as a linear combination of the ranks of this term S_i^r, V_i^r, I_i^r in the respective ranking lists of slopes, variances and intercepts. Terms are ordered in an increasing way in the first two ranking lists according to their values of slopes and variances. However the list of intercept ranks is sorted decreasingly so that the terms with low intercepts have high values of ranks and thus have an increased influence on the final score. N_s^r, N_v^r and N_i^r denote the numbers of different ranks for all three lists while weights α, β and δ specify the strength of each variable. The choice of the weights depends on what kind of terms should have the highest scores assigned. In general any combination of weighs can be used for obtaining a desired type of a summary. According to the weighting scheme discussed above the weights have following values.

$$\begin{aligned} \alpha &= \frac{\left|2 * S_i^r - N_s^r\right|}{N_s^r} \\ \beta &= 1 - \alpha \\ \delta &= 0 \end{aligned} \quad . \quad (2)$$

In this case the score of a term is dependent on the value of the slope if the rank of the slope is close to the maximum or minimum term slope in the web page. This is the situation of a rising or a falling trend. If the value of the slope is close to the average slope value for all terms, then the term score becomes more influenced by the variance of the term. Therefore "active" terms with horizontal regression lines will have higher scores than low variance terms with the same slope values.

It is important to note that the above statistical parameters are calculated for the whole contents of the web page versions in each time point rather than for new, inserted parts only. Some web documents have only minor changes. Therefore if we had analyzed only dynamic parts occurring in page versions then the importance of a static text would be neglected. However in some cases the longer a given content stays in the web page the more important it is. Additionally, it is also difficult to distinguish whether sentences have informational or other purposes like for example a structural one. Therefore we adopt a simplified and general approach where equal significance is given to the static and the changing data in web documents. In general we consider overall frequencies of terms in consecutive time points without distinction for unchanged, inserted or deleted type of content.

In this section we have described long-term scores of terms considering their statistics over the whole tracking interval. However for each web page version terms can have also so-called momentary scores computed. They are calculated using the local and neighboring frequencies of a term. After the term scoring has been completed important sentences can be selected for the inclusion into the summary. The algorithm for selecting and ordering sentences and the momentary scoring scheme are discussed in the next chapter.

4 Sentence Selection and Ordering

The algorithm for the sentence selection presumes the identification of points in time when particular terms have their highest importance. Every term will have its momentary score assigned for each page version depending on the local values of its frequency. The peak momentary score of a term would point to the web page version where the biggest frequency change has occurred. We believe that in such a page version there is a higher probability of the onset of an event or emergence of a concept associated with the term. Therefore it is advisable to select sentences whose terms have the highest local scores. The momentary weighting function has the following form.

$$M_i^j = \left(1 + \frac{\left(F_i^j - F_i^{j-1}\right)}{F_i^{\max}}\right) * \frac{F_i^j}{F_i^{\max}} . \quad \textbf{(3)}$$

Terms are weighted maximally in the web page versions where the frequency of the term *i* expressed as F_i^j has increased significantly in comparison to the frequency F_i^{j-1} in the previous version of the document. The momentary term score M_i^j in the version *j* will also increase when the term frequency has a high value in this version compared to the maximum term frequency $F_i^{\max}$ for all the page versions.

For constructing a summary we will select informative sentences, which have the highest sentence scores. The easiest way to estimate the score of a sentence is to compute its average term score. The score of contributing terms in this case could be calculated by multiplying their momentary scores by the long-term scores introduced in Equation 1. In this way the momentary score would be regarded as an indicator of the relative strength of a term in a particular web page version. Thus, when estimating scores of sentences we consider not only the words that they contain but also the web page versions in which these sentences appear.

Another way for selecting important sentences is to take *n* terms with the highest average long-term scores and use them as a base for the sentence extraction. Let *Vi* denote a page version when a given top-scored term *i* has its peak momentary score. From all the sentences containing the term *i* in *Vi* the one with highest average long-term score will be chosen for the summary. The same procedure is followed for the next top terms if they are not present in any of already selected sentences. In case they have already been included, the algorithm checks if the differences of their maximal momentary scores $M_i^{\max}$ and the momentary scores in those sentences are lower then

some predefined threshold. If the threshold condition is met then the next term is examined and the value of *n* is increased by one. The sentence selection algorithm is summarized below.

1. Define an empty set *Z* of sentences which will be included into summary
2. For each term *i* from *1* to *n*:
    a. For each sentence *Sj* from *Z*:
        i. if term *i* exists in sentence *Sj* then
            1. if ($M_i^{\max} - M_i^j <$ Threshold) then increase *n* by one and go to (2.)
    b. Find page version *Vi* where the term *i* has the highest M_i^j
    c. Choose a sentence from *Vi* which contains term *i* and has the highest average long-term score, and insert it into *Z* if it has not been already included
3. Return *Z*

The above algorithm ensures that a specified number *n* of the top terms will be represented in the summary. On the other hand if the sentence selection is based on the average term score then there is a possibility that some top-terms will not occur in the final summary. Thus the presented sentence selection algorithm could be used in the case when a user requires all the top-scored terms to be included into summary.

For redundancy elimination, the comparison of vectors of candidate sentences can be done in order to eliminate similar sentences. Because sentences are selected from different web page versions their local contexts can be different. To increase the coherence of the summary we also extract and include the preceding and the following sentences for all the sentences that have been selected by our algorithm. Additionally we employ a sentence ordering method, which is a combination of a temporal and a content ordering (Figure 2). The first one orders sentences according to the sequence of web page versions in which they appear. The latter one arranges sentences according to their relative positions in the document. Each sentence has its insertion and deletion time points that restrict the period during which the sentence can be found in the document. Thus, for temporal ordering one needs to check in which versions the sentence occurs. The combined ordering algorithm works in the following way. First the selected sentences are ordered according to the timestamps of page versions from which they have been extracted. In order to detect if it is necessary to reverse the order of two neighboring sentences from different versions we must make sure if the latter sentence occurs also in the same web page version as the former one. If they occur

then the sentences are ordered according to the sequence in which they appear in this web page version. In other words, sentences are sorted by two keys. The first key is the timestamp of a sentence in its earliest web page version and the second one is the position of the sentence in this document sample.

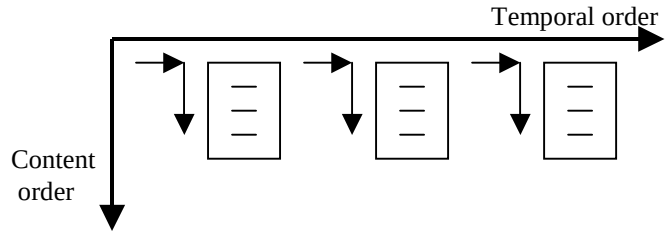

Fig. 2. Temporal and content ordering

5 Results

We show the example of a summary of "USATODAY.com – Travel - Today in the Sky" web page in Table 2. This page contains news about airline industry. We have collected 20 web page versions for the period of 3 months starting from 15th April 2003. In Table 1 the ranking lists of the top terms for different values of α, β and δ are displayed.

The summary contains information about financial problems of major US airlines, the reasons that caused their poor earnings and the measurements that the airlines take to improve their situation. The output is presented to a user in such a way that every sentence contains a link to its original web page version. Thanks to it, it is possible to see the background context of each sentence. Additionally each web page may contain colored text for changing contents to be distinguished.

Table 1. Top terms for different combinations of the values of weights α, β and δ

$\alpha=1\ \beta=0\ \delta=0$	$\alpha=0\ \beta=1\ \delta=0$	$\alpha=0\ \beta=0\ \delta=1$	α, β, δ as in Eq. 2.
washington	airlin	chicago	airport
chicago	airport	bag	secur
servic	secur	detroit	flight
fare	carrier	hare	travel
airport	associ	code	fare
offer	flight	impact	airlin
washington_po	associ_press	chicago_hare	servic
price	press	staff	press

profit	travel	access	carrier
increas	blue	lake	associ
detroit	war	dozen	associ_press
work	industri	spirit	long
bag	fare	tower	blue
fort	journal	public	atlanta
northwest	atlanta	missil	war
program	servic	hartsfield	industri
end	sar	kuwait	journal
access	cut	arm	profit
flight	long	trend	sar

Table 2. Summary for the weight combination from Equation 2

Bankrupt United announced today that it lost $1.3 billion during the first quarter, topping American's $1 billion loss for the same period. Not surprisingly, the airline blamed the war in Iraq and SARS fears for driving down traffic.
The proposed cuts simply "move Delta pilots from industry-leading pay to, well, industry-leading pay," J.P. Morgan analyst Jamie Baker told The Atlanta Journal-Constitution. He said that even if Delta's pilots agree to the cuts — as expected — their wages will still be about 12.5% higher than at United and up to 22% more than at American.
While the carrier still faces hurdles — witness today's announcement that United lost $1.3 billion — some say the airline has a decent shot to survive. "I'm much more optimistic now about United than I was a few weeks ago," said Phil Roberts, a consultant with Unisys R2A. "If it all comes together the way they'd like, they would emerge a very powerful carrier."
Airlines are slowly beginning to restore flights to their schedules after weeks of cutbacks due to war, a viral outbreak and the economic downturn, USA TODAY reports. U.S. carriers made some of the deepest cuts during the war on international routes, according to flight-schedule data provided by OAG.
"To convince passengers that the national carrier is free from SARS, we would pay $100,000 to any passenger who can prove he or she got SARS from flying Thai." — Thai Airways chairman Thanong Bidaya to The Associated Press.

6 Conclusions

In this paper we have presented a sentence extraction method for temporal summarization of web pages based on comparison of statistical parameters of text features. Temporal summarization of web documents attempts to summarize the contents of web pages spread in time. We have applied trend analysis and variance measure to reveal the main concepts of a web page during specified time period. Additionally we have

proposed the sentence selection algorithm which identifies sentences where given terms have the highest local importance. The algorithms presented here can be extended to summarizing entire web sites or the collections of related web documents.

There are some limitations of the proposed method that we would like to consider in the future research. A better fitting regression could help to approximate more correctly changes in the evolution of the importance of terms. Additionally, it would be beneficial to employ some more sophisticated concept representation. Lastly, we would like to make experiments with different kinds of web pages and to provide solutions, which are more specialized to the diverse types of documents.

References

1. Allan, J., Gupta, R. and 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. Amitay, E., and Paris, C.: Automatically Summarizing Web Sites: Is There Any Way Around It? Proceedings of the 9th International Conference on Information and Knowledge Management, McLean, Virginia USA (2000) 173-179
4. Berger, A. L., and Mittal, V. O.: Ocelot: a System for Summarizing Web Pages. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece (2000) 144-151
5. Buyukkokten, O., Garcia-Molina, H., and Paepcke, A.: Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. Proceedings of the 10th International World Wide Web Conference, Hong Kong (2001) 652-662
6. Glover, E. J., Tsioutsiouliklis, K., Lawrance, S., Pennock, D. M., and Flake, G. W.: Using Web Structure for Classifying and Describing Web Pages. Proceedings of the 11th International World Wide Web Conference, Honolulu USA (2002) 562-569
7. Jatowt, A., and Ishizuka, M.: Summarization of Dynamic Content in Web Collections. Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy (2004)
8. Jatowt, A., and Ishizuka, M.: Web Page Summarization Using Dynamic Content. Proceedings of the 13th International World Wide Web Conference, New York, USA (2004) 344-345
9. Jatowt, A., Khoo, K. B., and 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
10. Lent, B., Agrawal, R., and Srikant, R.: Discovering Trends in Text Databases. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Newport Beach, California, USA (1997) 227-230
11. Liebscher, R., and Belew, R.,: Lexical Dynamics and Conceptual Change: Analyses and Implications for Information Retrieval. Cognitive Science Online, Vol. 1, http://cogsci-online.ucsd.edu (2003) 46-57
12. Mendez-Torreblanca, A., Montes-y-Gomez, M., and Lopez-Lopez, A.: A Trend Discovery System for Dynamic Web Content Mining. Proceedings of the 11th International Conference on Computing, Mexico City, Mexico (2002)