Intelligence Integration in Distributed Knowledge Management

Dariusz Król
Wroclaw University of Technology, Poland

Ngoc Thanh Nguyen
Wroclaw University of Technology, Poland

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE

Hershey · New York

Director of Editorial Content: Kristin Klinger
Managing Development Editor: Kristin M. Roth
Senior Managing Editor: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Copy Editor: Lanette Ehrhardt
Typesetter: Jeff Ash
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Intelligence integration in distributed knowledge management / Dariusz Krol and Ngoc Thanh Nguyen, editors.

p. cm.

Includes bibliographical references and index.

Summary: "This book covers a broad range of intelligence integration approaches in distributed knowledge systems, from Web-based systems through multi-agent and grid systems, ontology management to fuzzy approaches"--Provided by publisher.

ISBN 978-1-59904-576-4 (hardcover) -- ISBN 978-1-59904-578-8 (ebook)

1. Expert systems (Computer science) 2. Intelligent agents (Computer software) 3. Electronic data processing--Distributed processing. I. Krol, Dariusz. II. Nguyên, Ngoc Thanh.

QA76.76.E95I53475 2009

006.3--dc22

2008016377

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter XVII
Utilizing Past Web for Knowledge Discovery

Adam Jatowt
Kyoto University, Japan

Yukiko Kawai
Kyoto Sangyo University, Japan

Katsumi Tanaka
Kyoto University, Japan

ABSTRACT

The Web is a useful data source for knowledge extraction, as it provides diverse content virtually on any possible topic. Hence, a lot of research has been recently done for improving mining in the Web. However, relatively little research has been done taking directly into account the temporal aspects of the Web. In this chapter, we analyze data stored in Web archives, which preserve content of the Web, and investigate the methodology required for successful knowledge discovery from this data. We call the collection of such Web archives past Web; a temporal structure composed of the past copies of Web pages. First, we discuss the character of the data and explain some concepts related to utilizing the past Web, such as data collection, analysis and processing. Next, we introduce examples of two applications, temporal summarization and a browser for the past Web.

INTRODUCTION

As the Web changes continuously, it is necessary to preserve the past content of pages for a future reuse. The Internet Archive[1] is the best-known and largest public Web archive containing data crawled since 1996. Other Web archives exist, for example, ones containing Web pages from particular countries (e.g., Arvidson, Persson, & Mannerheim, 2000; Hallgrimsson & Bang,

2003). Besides, there are also numerous repositories of past copies of pages such as caches, site archives, personal page repositories or search engine caches.

Web archives provide a view on the history of the Web reflecting past societal states. Past content of pages can reveal the histories of underlying elements represented by these pages, such as institutions, companies, people or other entities. For example, one could approximately detect when a particular member left some laboratory by detecting the time point at which her or his name was removed from the list of laboratory's personnel. In general, the use of Web archives can greatly benefit researchers and practitioners in many areas, such as history, sociology or marketing.

Furthermore, analyzing information from the past can help not only in better understanding the history of our society but also understanding its present state. This is because Web archives can provide contextual information about Web pages and the objects or concepts discussed on them as well as their inter-relations. For example, we can analyze information from Web archives concerning a given company in order to use it as a context for better understanding the present information about this company. In general, mining past Web content has a potential to stimulate and improve the traditional Web mining process in the sense that it provides contextual information and sheds new light on present data.

Past Web is considered here as a part of the WWW space where pages no longer have any change potential; they are "frozen" past snapshots of pages. The live Web, on the other hand, is the present Web, containing pages that we can currently view online. These pages may be changed or updated and they usually provide full interaction capabilities.

In the past Web each page has its history and lifetime. Links between the old content of pages can be reactivated again. In this way, a temporal structure can be obtained reflecting connectivity between pages in the past. Another aspect of the past Web is missing data. A given content after its deletion from a page may never be reproduced if it has not been preserved in any repository. Besides, due to the rapid growth of the Web, selective type archiving often needs to be done.

In this chapter, we approach the problem of discovering knowledge from the past Web. First, we discuss the character of data that is used and methods for acquiring and processing it. We propose techniques for analyzing and selecting candidate Web pages for mining. This approach is based on analyzing long-term characteristics of pages with a special focus on their content changes as they are most interesting from the viewpoint of pages' evolution. Next, we introduce temporal summarization, which is an adaptation of a traditional text mining task into the past Web scenario. We propose summarizing histories of Web pages to generate abstraction of events and salient concepts described in selected portions of the past Web. We also discuss the possibility of discovering object histories in past content of Web documents. Finally, we describe an application for browsing and navigating the past Web. We show an implementation that is similar to those of traditional browsers for the live Web and of video players.

The rest of this chapter is organized as follows. In the next section, we discuss the related research and attempt to place this work in the wider context of text and Web mining. The following two sections describe the data accumulation, preparation and analysis. In the next section we discuss temporal summarization and investigate the possibility of object history detection from the past Web. The next section describes a browser for the past Web, while the last section concludes the chapter with a brief summary.

RELATED RESEARCH

Web Dynamics

The dynamics of the Web has been measured in many experiments (Brewington & Cybenko, 2000; Cho & Garcia-Molina, 2000; Fetterly, Manasse, Najork, & Wiener, 2003; Ntoulas, Cho, & Olston, 2004) which demonstrated that the content and link structure of the Web continuously change. Although many pages on the Web are short-lived, meaning they are deleted shortly after being created (Ntoulas et al., 2004), many important Web documents persist over time. Popular and main, or top-ranked, pages usually belong to this category as it often takes a long time for a page or site to gain popularity and accumulate a high number of in-links.

The results of Web dynamics research indicate the level of volatility of the Web as a whole. On the other hand, the study of update patterns of individual pages has been carried out for prediction of their future changes (Cho & Garcia-Molina, 2000, 2003; Ntoulas et al., 2004). The frequencies and degree of changes are the most often used measures to set up crawling schedules for maintaining fresh indexes of search engines. In practice, however, it is usually difficult to predict content changes in pages although some Web documents, for example, newswire sources, change in a more or less periodical fashion. In this research, we go beyond the simple analysis of change statistics as we focus on the distribution of content and its context over time.

Text Mining

Text mining is defined as a nontrivial extraction of implicit, previously unknown and potentially useful information from textual data. Text mining evolved from data mining and is a promising field as much information nowadays is stored in the form of electronic text. We consider our approach to be similar to temporal text mining, because, to a certain extent it resembles efforts that were taken in analyzing and mining streams of text data. Generally, mining news articles or other text streams along the time dimension has been studied well (Allan, Gupta, & Khandelwal, 2001; Allan, 2002; Kleinberg, 2003; Li, Wang, Li, & Ma, 2005; Mei & Zhai, 2005; Papka, 1999; Swan & Allan, 2000; Wang & McCallum, 2006). For example, the well-known TDT (Topic Detection and Tracking) research initiative (Allan, 2002) was aimed at detecting, classifying, and tracking events in news corpora. Recently, Wang and McCallum (2006) identified topics persisting over dynamic collections of documents. Another work showed the development of topic patterns in news articles over time (Mei & Zhai, 2005). Li et al. (2005) proposed a probabilistic model for retrospective event detection in news corpora. An approach toward temporal summarization of news events was proposed in Allan (2001) where novelty and usefulness of sentences retrieved from newswire streams were calculated for the construction of a final summary. Another related work called TimeMines (Swan & Allan, 2000) was proposed for finding and grouping significant features in historical document collections based on applying chi-square test.

While news articles and, in general, any text streams are usually represented as transient text snapshots, the content of pages often persists over time. Duration of content has certain relation to its semantics and relative importance in a page. Thus, in contrast to typical text data streams, one has to consider three types of content in pages at every time point: static (persisting over time), deleted, and added. Additionally, pages have certain inherent topics that determine the context of their transitory content and that can enhance the mining process.

Web Mining

Web mining is often described as the application of data mining techniques for extracting knowledge

from the Web. It is traditionally divided into usage, structure and content mining. Web usage mining identifies the behavior patterns of users visiting Web pages for the purpose of optimizing Web sites. It is usually based on historical data, which is collected during certain time periods for its subsequent analysis (Cooley, Srivastava, & Mobasher, 1997; Kosala & Blockeel, 2000). Web usage mining can show how the users' access to Web sites changes over time. Web structure mining focuses on the link structure and graphical representation of the live Web. There have been, however, several approaches proposed to analyze the evolution of links over time (Amitay, Carmel, Herscovici, Lempel, & Soffer, 2004; Chi et al., 1998; Toyoda & Kitsuregawa, 2003). For example, temporal link analysis was used for detecting trends in page collections (Amitay et al., 2004) or for visualizing evolutions of Web communities (Chi et al., 1998; Toyoda & Kitsuregawa, 2003).

Web content mining uses the content of Web pages for knowledge extraction. Blog related research is probably the most prominent example of Web content mining in which the temporal aspect of pages is considered (Gruhl, Guha, Liben-Nowell, & Tompkins, 2004; Kumar, Novak, Raghavan, & Tomkins, 2003). Blogs help to detect and analyze social structures and social relations as well as provide information on society opinions, hot topics or recent trends. Blogs, however, are a unique media type as they usually contain complete versions of their past content with explicit timestamps provided as well as they are highly personalized and subjective. We believe that a general framework for mining any page types in the past Web is required.

Although most approaches to Web content mining generally neglected the temporal dimension of pages (Cooley et al., 1997; Kosala & Blockeel, 2000), there were, however, several works that investigated the usefulness of data on page histories for knowledge discovery (Arms et al., 2006; Aschenbrenner & Rauber, 2006; Jatowt & Tanaka, 2007; Rauber, Ascenbrenner, & Witvoet, 2002; Yamamoto, Tezuka, Jatowt, & Tanaka, 2007). Rauber et al. (2002) discussed the possibility of analyzing past Web data for identifying changes in Web-related technologies, particularly in the features and characteristics of Web pages, such as a file format, language, size, and so forth. The objective was to create statistics describing Web changes over time. Aschenbrenner and Rauber (2006) surveyed the work that has been done toward mining large portions of Web content with consideration of its temporal aspect. They also provided a general outlook on the potential of mining Web archives. Arms et al. (2006) have reported on building a research library for facilitating study of the Web evolution. This is an ongoing project aiming to build an infrastructure for analysis of massive portions of the data that is stored in Internet Archive. Practical usage of the past Web has been recently demonstrated by Yamamoto et al. (2007), who have proposed an application similar to question answering systems for extracting and combining knowledge from the Web and Web archives. It uses Web archive data for detecting changes in opinions and user knowledge over time.

Mining the content of the past Web is different from the usual Web content mining in several aspects. First, the temporal dimension of content and links in page histories poses new challenges and opportunities for understanding their roles and interrelations in contrast to traditional Web content mining. Second, pages and Web sites should be treated as dynamic objects having certain age, histories, trends, patterns, and so forth. Thus, the notions of a page and its content need to be separated in a way in which the latter one is considered as a transient component occurring in a higher level object, that is, a page. Content has then its own duration of occurrence while the page history is considered as the composition of different content occurring throughout the page's lifetime. Third, there is an issue of missing and incomplete data. In order to obtain satisfactory results, multiple snapshots of the past content of

pages have to be found and acquired as well as approximation methods need to be applied for an optimal page history reconstruction.

DATA ACQUISITION AND PREPARATION

Data acquisition and preparation are important steps in the knowledge discovery process. In the mining of the past content of the Web these steps mean the retrieval of data from Web archives and the reconstruction of Web document histories (Jatowt & Tanaka, 2007). The following issues are involved here. First, it is by definition an ex post facto process, as the data is the past content of pages. If one could predict beforehand which Web pages are going to be used, one could simply set up a crawler with a suitable crawling frequency so that page evolution would be captured with a desired precision. However, it is assumed that the user is unable to make such a prediction, and rather that she or he wishes to acquire knowledge in real time using the available, preserved data. Hence, past snapshots of Web pages are gathered in real time from available resources with the aim of reconstructing the past with the highest possible precision. Thus, when talking about crawling in the context of the past Web, we mean querying past Web repositories for the data they contain. Second, because data is scattered in different repositories, it has to be searched for and identified before being used. Therefore, it is necessary to use efficient search and download techniques to locate and gather multiple snapshots of past content with a minimal cost. Due to the large size of data, in practice, usually, only its small portion can be fetched and analyzed locally. Therefore, the focus of this research is on the analysis of the limited amount of data rather than on building a framework for examining the past Web from a macroscopic viewpoint. In addition, there is an issue of the trustworthiness of past content, which is directly related to the trustworthiness of past Web repositories. For example, data obtained from a personal Web repository would normally be less trustworthy than the data collected from a large Web archive containing millions of pages and having a professional maintenance and control. Finally, only fragmentary data can be obtained due to the unpredictable change pattern of the Web and limited resources of archival systems. This calls for employment of efficient techniques for estimation of actual content that pages had in the past.

Collecting Snapshots

***Definition 1*:** Past page snapshot is a copy of page content that was published in the Web at a given time point in the past. The timestamp of the snapshot indicates the date when it was captured.

As mentioned above, because of resource limitations, Web archives contain only fragmentary past data. As a general attempt to alleviate this problem a kind of meta-archive approach (Jatowt, Kawai, Nakamura, Kidawara, & Tanaka, 2006) can be used to maximize past Web coverage and consequently to increase the precision of history reconstruction. This approach presumes communication with several past Web repositories at the same time. An intermediary module is required between these repositories and the local system to translate queries into the format required for each repository. After receiving a request for a page history, the module queries the repositories about their data. The repositories should then send a list of stored page snapshots with their metadata so that a fetching policy can be determined.

The optimal strategy would be first to check the signatures (checksums) of snapshots, if they are provided, in order to detect the ones that actually contain content changes from among all data provided by the cooperating repositories. This would prevent downloading identical page snapshots from different repositories, thereby maximizing fetching efficiency[2]. However, currently, Web archives do not provide such infor-

mation. Instead, some repositories, such as the Internet Archive, provide lists of page snapshots that have any changed content when compared to the neighboring snapshots. By utilizing this information, only the snapshots with content changes inside archives would be fetched. In general, the efficiency of the data collection would depend on the metadata that is provided in past Web repositories.

Such a meta-archive approach would provide a unified interface to the history of the Web, making the data acquisition process less dependent on the resources of single Web archives. However, as Web archive interfaces are diverse, different data acquisition methods would have to be used. In addition, we make an assumption here that the URLs of pages remain the same over time, although, in practice, they may change even though the content of pages remains almost the same.

McCown and Nelson (2006) and McCown Smith, and Nelson (2006) have recently measured the persistence and availability of page copies inside the repositories of major search engines and the Internet Archive. The objective was to estimate the possibility and to provide methodology for reproducing the latest versions of Web sites in case of the loss of Web data.

Reconstruction of Page Histories

***Definition 2*:** Page history reconstruction is the process of reproducing the past content of a page using available snapshots for obtaining the continuous representation of page history.

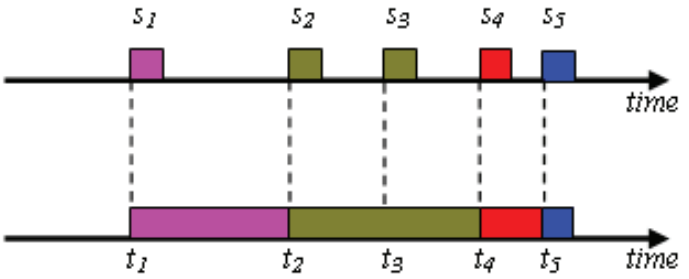
***Definition 3*:** Optimal page history reconstruction is a reconstruction which accurately reproduces page history; that is, the errors resulting from such a reconstruction are equal to zero. Having determined an optimal page history, it is possible to recreate page content for any time point in the past that shows the actual content the page had at that time.

However, unless the page was unchanging, it has been crawled continuously or the implicit information about its past changes is provided, there will be usually some error involved in the history reconstruction. Only for certain types of pages, for example wikis, complete past data is available as the preservation of their versions is usually automatically done. In case of such pages, the reconstruction error would be equal to zero as all past changes can be derived from available page versions. In addition, some pages may contain temporal annotations in their present content that can be used to enhance the history reconstruction. For example, blogs often provide timestamps of content insertion. Nevertheless, for the majority of hypertexts, usually, neither implicit version management nor temporal annotations are provided.

We propose a simple approach for the page history reconstruction (Jatowt & Tanaka, 2007). First, collected snapshots are chronologically ordered according to their timestamps. If past snapshots are not associated with any temporal metadata then they cannot be directly included in the ordered sequence of past snapshots without a prior determination of their timestamps. For example, Yahoo! search engine provides cached snapshots of Web pages but it does not attach any timestamps to them. Estimating a timestamp of a snapshot could be possibly done by comparing similarities between its content and the content of other snapshots with known timestamps.

Second, every previous page snapshot is considered to represent the actual state of page content for the time period until the next page snapshot in the sequence. For example, suppose that five snapshots have been collected, s_1, s_2, s_3, s_4 and s_5, with timestamps, t_1, t_2, t_3, t_4 and t_5, where $t_1<t_2<t_3<t_4<t_5$ (Figure 1). Let us also suppose that snapshots s_2 and s_3 are exactly same. After the simple approximation, the page content is assumed to be the same as that in s_1 during the period $[t_1, t_2)$, the same as that in s_2 during $[t_2, t_4)$ and equal to s_4 in $[t_4, t_5)$. The reconstructed page history is then represented as a minimal sequence of 2-tuples containing different page versions and

Figure 1. Example of page history reconstruction



their starting dates ($\{(s_1, t_1), (s_2, t_2), (s_4, t_4)\}$ in the above case).

Page history reconstruction could be improved by considering additional information, for example, by analyzing changes in other pages belonging to the same site. Also, using the results of the temporal analysis of pages, especially their updating patterns, could make the reconstruction more accurate. Finally, historical snapshots of mirror pages, if there are any, could be utilized.

History Reconstruction Error

Usually, it is difficult to determine an accurate page history that would reflect the actual page content as it was at any arbitrary time point in the past unless the complete set of actual page versions is provided, for example, by a page author. Hence, mining the content of the past Web will typically be carried out using incomplete data with varying levels of precision and trust. It is thus necessary to consider the issue of missing data.

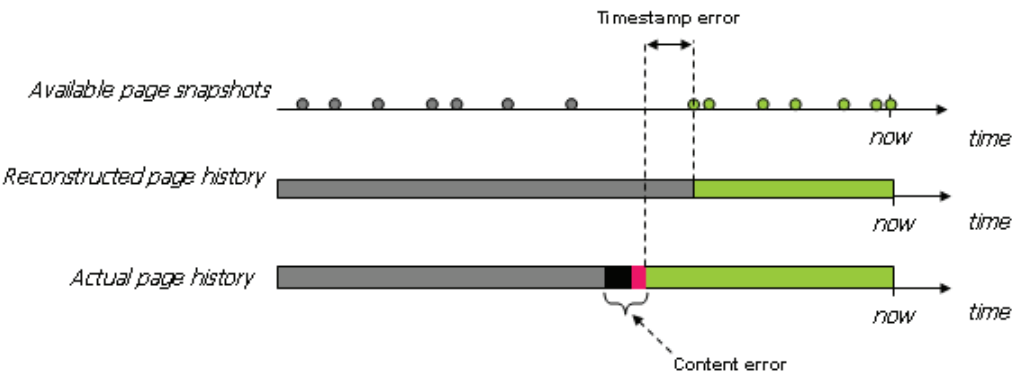
We can distinguish two types of errors in the page history reconstruction assuming that the page crawling was independent from the page update pattern (Jatowt & Tanaka, 2007). The first one, which we call a content error, is caused by uncertainty related to the content that appeared on a page in the past. Consider two retrieved past versions of the page (v_L and v_R) captured at time points t_L and t_R ($t_L<t_R$). The probability, $P(v_i)$, that there is any version v_i satisfying $t_L<t_i<t_R$ and containing any content different from that in v_L and v_R depends on many factors such as the length of the period from t_L to t_R, page type, content difference between v_L and v_R or the average change degree of the page. Intuitively, the shorter the time distance between the page snapshots and the more even their distribution over time are, the lower is the average probability of any transient, undetected content occurring in the page.

The second error type, which we call a timestamp error, is due to uncertainty in estimating the dates of content changes. The timestamp error, like the content one, depends also on the number of acquired past snapshots and their distribution over time. Figure 2 illustrates both error types. The top timeline shows available past snapshots of a page. For simplicity, let us assume that the page snapshots are empty (i.e., blank page) or they contain only one content element, be it a picture or a text snippet. Those snapshots that contain the element are marked by a green color, while the empty snapshots are marked by a grey color. After reconstructing the history of the page (the middle timeline) and comparing it with the bottom timeline that shows the actual page history, we can see that the reconstructed history contains both content and timestamp errors.

Site History Reconstruction

Pages usually belong to larger information units, or Web sites. Reconstructing histories of a Web site requires detecting the changes in site's topology over time and retrieving past content of pages

Figure 2. Content and timestamp errors in the reconstructed page history



that belonged to the site. As an input the starting page (e.g., the top page of a site), time frame *T*, and the depth *D* (i.e., the number of hops from the starting page) need to be specified.

The data accumulation system collects all available snapshots of the starting page that have timestamps within *T*. It then searches their content for any links to other pages on the site (i.e., pages having the same domain name). For each such a link, it collects available, previous snapshots of the linked page that have timestamps within *T*. These snapshots are then searched in the same way for links to other pages on the same site. The entire process is repeated until the specified depth *D*. In general, a page is considered to belong to the site's history if, during the time frame *T*, it was linked from another page belonging to the site at that time and if it was located a smaller number of hops from the starting page than the specified depth *D*. Intuitively, the number of page snapshots collected at the initial steps of the crawl (few hops from the starting page) has an influence on detecting pages at later steps. This is because pages may remain undiscovered if the links pointing to them occurred only in the undetected, transient content of other pages in the site. We call the error caused by the missing links a page error.

A site history is represented as a set of reconstructed page histories that belonged to the site in the past. The precision of the site reconstruction can be enhanced by utilizing topological information preserved in the past content of site-map pages if they existed. Many Web sites include site-map pages designed to help users navigate sites. Utilizing the site-map page history could help to detect transient pages that have not been discovered by the above crawling approach and thus minimize the page error, as well as it could help to more precisely determine the actual time points of page creation and deletion within sites (timestamp error for whole pages).

PAGE TEMPORAL ANALYSIS

Page temporal analysis is the study of page content over time. Its results should be particularly useful if pages are associated with specific objects such as companies, institutions, persons or other entities. Understanding the temporal characteristics of a page over a long time frame can shed light on the associated objects or on other information appearing on the page. For example, if certain content occurred for a long time on a page which was updated frequently and regularly, then we can treat the content in a different way or with a different level of trust than if it occurred on a page that was generally static or even obsolete. A similar idea applies to a page devoted to a specific topic vs. a page that deals with many varying top-

ics throughout its history. In other words, page temporal analysis can be used to find temporal context for information from the past. Having determined the context, it is possible to better understand the connection between Web pages and their transient content as well as to identify pages most relevant to target objects.

When mining the histories of Web pages for real-world information, we must distinguish between the valid time and transaction time of events, both of which are often used in the database research. The valid time of an event is considered as the time at which the event occurred in the real world. The transaction time is the time at which the information about the event was stored in a database or, in our case, added to a certain Web page. It can be estimated by searching the page history for the earliest occurrence of the content related to the event (Jatowt, Kawai, & Tanaka, 2007). The valid time, on the other hand, can be detected from temporal expressions appearing in the content of past page versions. This would require using special taggers and resolvers of temporal expressions in text. In addition, techniques such as the one described by Bar-Yossef, Broder, Kumar, and Tomkins (2004) could be applied for classifying page content as current or obsolete.

Next, we present a simple framework for analyzing page histories. After page history reconstruction, HTML tags, scripting code, and multimedia objects are removed from available page versions. Vector representation is then created for textual content of the past versions using a weighting method such as a term frequency. Let $\boldsymbol{V}=(\boldsymbol{v}_1,\boldsymbol{v}_2,..,\boldsymbol{v}_n)$ denote the sequence of vectors of the consecutive page versions, where $\boldsymbol{v}_j$ is the vector of a page version at time point t_j ($t_1 \leq t_j \leq t_n$). Next, the contents of the neighboring versions are compared with each other using a change detection algorithm such as *diff*. Added content appearing in the page's history is thereby found. All changes in each version are then grouped together and represented as a change vector. Consequently, a sequence of change vectors is obtained, $\boldsymbol{C}=(\boldsymbol{c}_{(1,2)},\boldsymbol{c}_{(2,3)},..,\boldsymbol{c}_{(n-1,n)})$, where $\boldsymbol{c}_{(j,j+1)}$ is a vector for an added-type change obtained by comparing page versions $\boldsymbol{v}_j$ and $\boldsymbol{v}_{j+1}$.

The content of past versions can be compared against any query containing terms describing given topic of interest. In order to do so, at each selected time point, a query vector, $\boldsymbol{q}_j$, is constructed by assigning uniform weights to all query terms. The sequence of query vectors is denoted as $\boldsymbol{Q}=(\boldsymbol{q}_1,\boldsymbol{q}_2,..,\boldsymbol{q}_n)$. Different values can be assigned to $\boldsymbol{Q}$ at different time points to reflect changes in the chosen topic of interest. Otherwise, the query vector is made static by having the same content at all times. To measure the relationship of past page content to the query topic, the similarity between $\boldsymbol{V}$ and $\boldsymbol{Q}$ is calculated using a cosine similarity measure. In result, the sequence of similarities is obtained: $\boldsymbol{sim}(\boldsymbol{V},\boldsymbol{Q})=(sim(\boldsymbol{v}_1,\boldsymbol{q}_1),sim(\boldsymbol{v}_2,\boldsymbol{q}_2),..,sim(\boldsymbol{v}_n,\boldsymbol{q}_n))$,

Figure 3. Similarity calculation between the sequence of version vectors and the sequence of query vectors

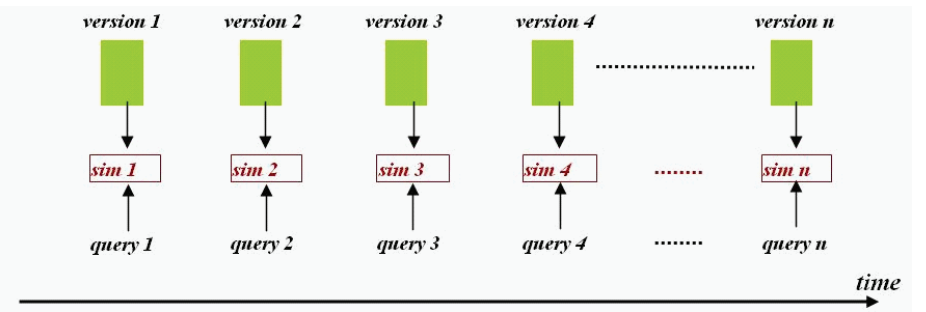
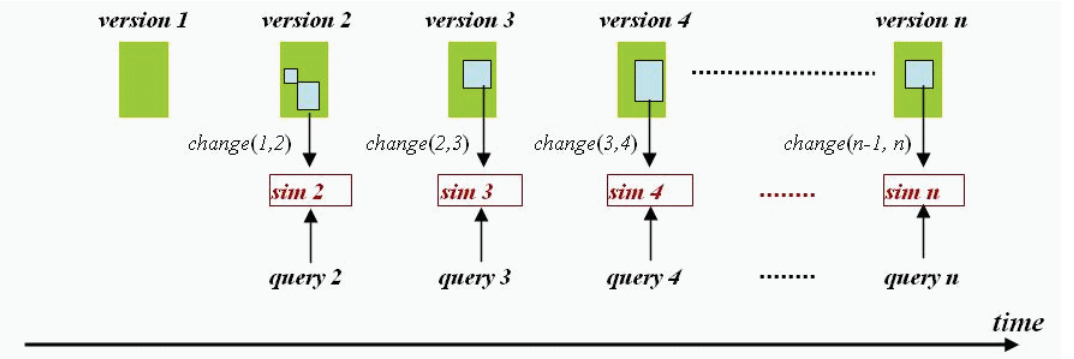


Figure 4. Similarity calculation between the sequences of change and query vectors; changes are depicted as small rectangles inside page versions



where $sim(\boldsymbol{v}_j,\boldsymbol{q}_j)$ is the cosine similarity between the vector of past version $\boldsymbol{v}_j$ and query vector $\boldsymbol{q}_j$ (Figure 3). Similarly, the sequence of similarities between the vectors of the changes and the query vectors is calculated, $\boldsymbol{sim}(\boldsymbol{C},\boldsymbol{Q})=(sim(\boldsymbol{c}_{(1,2)},\boldsymbol{q}_2),sim(\boldsymbol{c}_{(2,3)},\boldsymbol{q}_3),..,sim(\boldsymbol{c}_{(n-1,n)},\boldsymbol{q}_n))$, where $sim(\boldsymbol{c}_{(j,j+1)},\boldsymbol{q}_{j+1})$ is the cosine similarity between change vector $\boldsymbol{c}_{(j,j+1)}$ and $\boldsymbol{q}_{j+1}$ (Figure 4).

First, a change frequency can be defined (Equation 1).

$$CF = \frac{fc}{n}$$

Here, *fc* is the number of non-zero elements in $\boldsymbol{C}$. Another measure called a change degree indicates the average change size of a page (*size(a)* denotes the size of element *a*).

$$CD = \frac{\sum_{j=1}^{n} \frac{size\left(c_{(j,j+1)}\right)}{size\left(v_j\right)}}{n}$$

Besides these simple measures, the long-term relevance of a page to the query topic can be calculated. It is expressed as the weighted average of the elements of $\boldsymbol{sim}(\boldsymbol{V},\boldsymbol{Q})$ by taking into account the duration of page content over time (Equation 3).

$$TR = \frac{1}{\sum_{j=1}^{n}\left(\alpha_{j+1} * \left(t_{j+1} - t_j\right)\right)} \sum_{j=1}^{n}\left[\alpha_{j+1} * sim\left(v_j, q_j\right) * \left(t_{j+1} - t_j\right)\right]$$

A page is considered relevant if its content overlaps with the sequence of query vectors during a large portion of a chosen time period. Using this approach, we can estimate the degree of page relevance to any topic within a given time frame. As the recent content is often likely to be more important, Equation 3 is adjusted by applying a weighting scheme depending on the age of page versions.

$$\alpha_j = e^{-\lambda\left(t_{now} - t_j\right)}$$

In addition, the long-term topic stability of a page can be computed by detecting the average similarity between consecutive past versions over time (Equation 5).

$$TS = \frac{1}{\sum_{j=1}^{n}\left(\alpha_{j+1} * \left(t_{j+1} - t_j\right)\right)} \sum_{j=1}^{n}\left[\alpha_{j+1} * sim\left(v_j, v_{j+1}\right) * \left(t_{j+1} - t_j\right)\right]$$

The long-term relevance and long-term topic stability are calculated considering the whole page content in the past, including the static content (the content that did not change between consecu-

tive page versions). In contrast, we can compute a measure showing the degree of page updating based on the amount of changed content over time that is related to the query (Equation 6).

$$TRC = \frac{1}{\sum_{j=1}^{n}\left(\alpha_{j+1} * \left(t_{j+1} - t_j\right)\right)} \sum_{j=1}^{n}\left[\alpha_{j+1} * \frac{sim\left(c_{(j,j+1)}, q_{j+1}\right)}{t_{j+1} - t_j}\right]$$

A combination of different measures can also be used. For example, the measure of the temporal quality of a page is based both on the relevance of the changed content over time to the query topic and on the size of the changes:

$$TA = \frac{1}{\sum_{j=1}^{n}\alpha_{j+1}} \sum_{j=1}^{n}\left[\alpha_{j+1} * \frac{sim\left(c_{(j,j+1)}, q_{j+1}\right)}{\left(t_{j+1} - t_j\right)} * \frac{size\left(c_{(j,j+1)}\right)}{size\left(v_j\right)}\right]$$

According to this measure, a page is more attractive from the viewpoint of the query topic if its changes were relevant to that topic and if they were relatively large. Small changes are usually less likely to be attractive than large ones. Additionally, the temporal quality of the page is higher if the page was modified often in the past. In general, the greater the number and the larger the size of related changes that occurred within short time periods, the higher is the attractiveness of the page. The page temporal quality can be used to identify candidate pages for mining. Naturally, the precision of results depends on the amount and characteristics of the input data that is on the size of errors resulting from the history reconstruction process.

Finally, the trend of page relevance to query can be measured by fitting a regression line to the historical plot of the similarity between page content and query vectors. This allows for estimating the long-term change direction of the page relevance. A rising trend would mean that the page content becomes closer to the query topic.

TEMPORAL SUMMARIZATION

Document summarization is a well-known text mining task. Automatic summarization of Web pages aims at creating compact versions of Web documents that would contain only the most important content. Traditionally, summaries were constructed from static snapshots of Web pages (Berger & Mittal, 2000; Buyukkokten, Garcia-Molina, & Paepcke, 2001; Delort, Bouchon-Meunier, & Rifqi, 2003). However, as pages are dynamic, their content is often changed. In this section, we briefly describe the concept of temporal summarization which is the extension of the traditional summarization task into the time dimension (Jatowt & Ishizuka, 2004a, 2004b; Jatowt & Ishizuka, 2006). It is used to summarize temporal versions of Web documents in order to provide information on important content, hot topics or popular events described in pages over time. Web users are often overloaded with large amounts of data. Automatic temporal summarization would help them in discovering salient information from parts of the past Web such as histories of pages or their collections.

Following the classical division of document summarization research, two types of temporal summarization can be distinguished: single- and multi-page temporal summarization. Single-page temporal summarization attempts at capturing salient content that occurred on a page over a certain time period. The summary should thus reveal main page topics during a predefined time frame. On the other hand, in multi-page temporal summarization, multiple snapshots of a topical collection of pages are analyzed for changes over time. The summary should reveal important events or concepts that occurred in a given topical area over time. The key issue in this type of summarization is gathering pages which are up-to-date and related to the target topic so that a reliable and consistent topical collection can be synthesized. Below we discuss the multi-page temporal summarization in more detail.

Multi-Page Temporal Summarization

Web collection for multi-page temporal summarization can be obtained in several ways; for example, it can be created from a user-provided set of related Web documents that she or he usually revisits for fresh information or it could be downloaded from existing Web directories. While Web directories group topically related Web documents, they provide only a limited number of categories. In a more flexible way, the collection could be synthesized by filtering search engine results based on the analysis of their temporal characteristics such as long-term relevance or temporal quality. Naturally, duplicate pages should be discarded in this process. After the initial set of topically related pages is ready, it is extended in time by reconstructing page histories for a chosen time period.

In the following step, textual data is extracted from the accumulated past versions. Then, an extractive type summarization algorithm is used to detect useful sentences for constructing a summary. First, so-called long-term scores are calculated for all terms by comparing terms' distributions in documents over time. These scores are later used to identify important sentences to be included in the summary. We propose two approaches for the long-term score calculation. One uses a sliding window that is sequentially moving through the temporal collection to search for bursts of terms in added or deleted content in the collection (Jatowt & Ishizuka, 2004a). Any terms that were added to or deleted from many pages in the collection at around the same time have high values of the long-term scores. Another approach to the calculation of long-term scores is based on the analysis of term frequency plots. The parameters of term frequency plots such as variance, slope of a regression line and intercept are calculated and compared for identification of salient terms (Jatowt & Ishizuka, 2004b). The terms with outstanding features, such as the ones with upward trends or high variance would be then scored highly. More details on the both term scoring methods can be found in Jatowt and Ishizuka (2004a, 2004b) and Jatowt and Ishizuka (2006).

After the long-term scores of terms are computed, the summarization system searches for sentences suitable for constructing the summary. Sentence selection is based on analyzing plots of the terms that have the highest long-term scores. The plots are examined to identify intervals with the closest match to the shape of an ideal plot. For example, the system may search for a time period where the frequency plot of a term has a shape that most resembles the ideal shape in which the plot suddenly increases and remains at a relatively high level over a long time. Such a plot shape may indicate the onset of an important event represented by the term. Thus, sentences containing the term are extracted from the collection within the selected time period. The system tries here also to maximize the number of different terms with top long-term scores in the selected sentences. Lastly, after the predefined number of sentences is extracted, the system orders them based on their timestamps and relative locations in their original page versions. Each sentence may also have a link to its page version added to be used in case users wish to obtain more details. Furthermore, a number of additional heuristics may be used to increase the coherence and readability of the final summary, for example, by inserting explanatory content or by modifying or reordering the selected sentences.

Discovering Object Histories

Related to temporal summarization is object history reconstruction. Objects are defined here as higher level concepts and abstractions that represent persons, institutions, ideas, organizations, and so forth. Objects can be represented by groups of related words or n-grams. Thus, object histories could be modeled using the histories of the representative terms and their inter-relation-

ships. Time points of changes and the durations of terms' occurrences on pages would provide clues about the timing of events related to objects represented by these terms.

Object's history should be most accurate if it has been derived from a source that directly represents the object (e.g., company homepage, personal blog). The relationship between the page and objects discussed on this page can help in understanding the content related to the objects. In general, contextual information about objects can be derived from the characteristics and topical scopes of analysed pages. Furthermore, the co-occurrence of similar information among different resources increases its trustworthiness as well as helps to better determine the starting and ending points of events. The larger is the number of different data sources devoted to an object, the more reliable and accurate the discovered knowledge should be.

A possible example of object history reconstruction is an automatic creation of personal bibliographies or their parts. There is much personal data published on the Web. For example, employment data is sometimes reported on company or personal Web pages (e.g., on blogs), and other personal information can be found. This information could be collected and processed to construct biography parts.

By analyzing semantic and temporal clues derived from past Web content it could be possible to improve the detection process by employment of various heuristics. For example, the temporal information derived from the chronological ordering of events reported on past pages might help in understanding the events and may provide hints for a further search. One such possible heuristic is the detection of person's employment dates. Suppose that at some time point a person's name was removed from the page of some laboratory. Then, the system could search for the page of another institution that reported hiring the person at around that time. Note, however, that there might be certain latency between the actual events and their reports in the Web (i.e., valid and transaction times).

BROWSING PAST WEB

Apart from mining the content of the past Web, it is important to have a tool that allows for viewing data in detail, for example, in order to manually inspect the data from the viewpoint of discovered results. Such a tool should be intuitive, easy to use and possibly resemble similar applications used for the current Web. In this section, we describe the framework for a past Web browser (Jatowt et al., 2006) that supports browsing and navigation in the past Web. A browser built using this framework would be a client-side system that downloads, in a real time, past page snapshots from Web archives for their customized presentation. Such a browser would enable viewing the evolution of pages and browsing the past structures of the Web.

The proposed browser integrates histories of Web pages with their present versions and has a standard functionality of a traditional browser for the live Web. Consequently, browsing the live and past Web can be done almost at the same time. Thanks to this, users browsing the live Web can access the histories of viewed pages in case they need to find some content from the past, observe the page evolution or, simply, to access the latest page snapshot if the present page cannot be properly viewed due to any reasons such as a server failure.

Browsing

Two basic types of browsing are distinguished here: vertical and horizontal. The former means browsing different pages around a certain point of time by following links, while the latter means viewing past snapshots of a single page along the time direction, that is, browsing the past Web in a horizontal direction. A mixture of both kinds of browsing enables users to traverse the past Web both in time and space dimensions.

To start the horizontal browsing, the URL of a page and a point of time have to be provided. The browser fetches a page snapshot whose time-stamp is closest to the user-provided time point. Next, the browser automatically downloads the following page snapshots and displays them in a passive manner. This type of viewing results in a minimum user interaction, because page snapshots are presented to the user one by one, like in a slideshow, with a certain delay predefined for each snapshot. As when watching a video, the user can pause or stop the motion, enabling the detailed examination of the currently presented snapshot or following its links. Besides, the user may enter a new date or a different URL to make a jump to another snapshot. In addition, a timeline is automatically constructed and displayed above the page content (Figure 5). It shows the distribution of page snapshots indicating the points of time for which snapshots are available. The currently viewed snapshot is indicated in the timeline by a blue rectangle. The information provided by the timeline prevents users from being lost in the hyperspace of the past Web by informing them about the current time point of browsing and the overall distribution of snapshots. At the same time, it is also a navigation tool thanks to which users can make a jump to any page snapshot simply by clicking on any point on the timeline. The timeline can be also zoomed to provide the more detailed view. Besides the timeline, the clickable list of all page snapshots together with their timestamps is also displayed (Figure 5).

Horizontal browsing is enhanced by a page presentation in which content changes are detected and emphasized. Keeping in mind the large size of the past Web, with lots of static, redundant data, the most effective method for horizontal browsing seems to be the one using change visualization. We think that changed data is the most important in page histories and that enhancing horizontal browsing with the change indication can portray page evolution and, in addition, help reduce the amount of browsing needed, especially in the case of static (unchanging) pages. Both content additions and deletions between neighboring page snapshots are then detected using a change detection algorithm and emphasized to indicate the content variance in pages. This enables users to spot not only the added content in consecutive page snapshots but also to identify the removed one. However, effectively showing both change types in a combined view on a single page would be difficult, especially in the case of large and overlapping changes. Thus, we propose using animation effects in order to efficiently show both change types. The change presentation algorithm displays the changes gradually, in the form of animation. Content that was deleted in the page history first blinks for a certain time period and then disappears, followed by the inserted content that first appears on the page and then blinks for a short time. Page snapshots are processed in this way line by line from left to right and from top to bottom. Content that was static between consecutive snapshots remains displayed on the page. After the page transition between two consecutive page snapshots is completed, the browser waits a predefined time period with the latter page snapshot displayed and then it proceeds to analyze the following page snapshot. The user can control the speed of the presentation using a slider provided in the top-right corner of the browser (Figure 5). Besides, as sometimes page snapshots may be too large to be shown at once, a user can choose between the automatic scrolling option and the option of displaying only the top part of page content.

Animation of changed content results in a smooth transition between sequential page snapshots. By animating changes user's attention is drawn to the changed content. In addition, changes are also highlighted by different colors to increase their visibility. However, for simplicity, in the case when the amount of change in a page snapshot is higher than the predefined threshold, no animation is done and changes are emphasized using only different background colors.

The user can stop the horizontal browsing at any time by pressing stop or pause buttons in a similar way to video players. Next, she or he can view the currently displayed page snapshot in detail or follow any link. Upon clicking on a link, the browser loads the snapshot of the linked page that is closest in time to the one being currently viewed and, after a short time period, it automatically starts the horizontal browsing on the new page.

The browser is also equipped with two back and two forward buttons to enable navigation in the space as well as in time dimensions. Besides, there is an additional navigation mechanism provided (automatic jumping facility). It enables the browser to skip periods in the page history during which the content did not change or did not change much. When this functionality is switched on, the browser displays only those page snapshots that contain more than a certain amount of change. This enables faster viewing of page evolution by omitting changeless periods.

Finally, a search option enables users to specify queries for filtering changes. If a query is issued, only the changes that contain the query terms are animated. Other changes are treated as static content and thus are not animated. This browsing style results in the filtered view of page history. Users can thus observe page histories from the viewpoint of topics that they are interested in. For example, a newswire page history could be browsed for information about "Iraq" or "presidential election" over selected time periods.

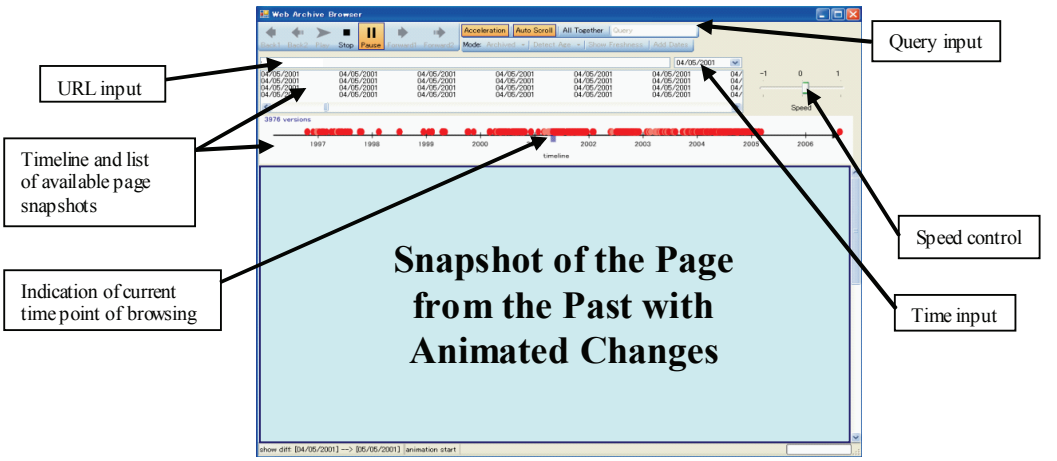
Related Research and Future Work

Visual Knowledge Builder (VKB) (Shipman & Hsieh, 2000) was an early proposal of an application that provides a mechanism for enabling history navigation in private hypertexts. The objective was to allow users to play back the history of a hypertext for witnessing the authoring of hypertexts, understanding the context of their creation and authors' writing styles. The browser interface had some similarity to VCR players.

WERA[3] (Web ARchive Access) and Wayback Machine[4] are applications for accessing Web archives. WERA supports time and URL input for specifying a particular page snapshot. There is a timeline provided showing the available page snapshots and indicating the currently browsed one. Users can view the consecutive page snapshots by clicking arrows in the timeline.

Wayback Machine is a Web-based interface to the Internet Archive. After a user inputs a URL, optionally with a time period specified,

Figure 5. Past Web browser



links to the available page snapshots are listed on the "directory" page. The user can then click on any snapshot to view its content or follow its links if the linked snapshots are also stored in the archive. The directory page indicates also page snapshots that contain changes by marking them with asterisks. Horizontal browsing using Wayback Machine is difficult, as users need to access the directory page each time if they wish to view other snapshots.

Both the Wayback Machine and WERA are server-side applications designed for single Web archives. Our proposed browser is a higher-level, client-side application that allows for the usage of multiple past Web repositories at the same time, thus, enabling browsing of the past Web rather than browsing single archives. Browsing the past Web is also facilitated by combining passive, automatic page viewing together with a change presentation. The framework has also functions that minimize the user effort and time required to find specific information in the past snapshots of pages. In addition, navigation mechanisms are provided to enable traversal of the link structure of the past Web. Testing the browser built on the proposed framework demonstrated its usefulness (Jatowt et al., 2006). Users were able to move freely in the past Web, find desired information and relatively easily obtain an overall view of pages' evolution.

In a multi-authoring area, an interesting application has been recently proposed for effective visualization of histories of wiki pages (Viégas, Wattenbeg, & Dave, 2004). It allows viewing contributions of different authors and their persistence over time as demonstrated on the example of Wikipedia pages[5].

There are several possible directions for expanding the proposed framework. For example, location-based browsing would allow a user to select a certain area on a page and then view its evolution over time, provided that the structure of the page did not change substantially. This would limit the presentation to only those changes that occurred in the selected area, for example, in the sports section of a newswire page. Next, links on visited snapshots could be annotated with timestamps of page snapshots that will be accessed when following these links. Thanks to it, a user would know how much time jump she or he is going to experience upon clicking on a certain link. Lastly, a comparative past web browser could enable comparison of histories of two or more pages highlighting their common or similar parts.

CONCLUSION

The Web has become nowadays a major means of communication and an important information repository. Due to its dynamic, ever evolving character, much of the content regularly disappears from the live Web and can only be accessed through Web archival repositories. Knowledge discovery from past Web is a challenging and promising research direction. Mining the content of the past Web differs from traditional Web content mining and thus requires a novel approach. In this chapter, we have described several issues related to mining data in Web archives. First, we provided the outlook on the data collection and preparation steps and emphasized their importance. Next, we demonstrated the methodology for determining page temporal characteristics as a source of contextual information for describing pages and their transient content. Then, data summarization and object history detection were described as examples of mining tasks on the past Web. Finally, we proposed the application for browsing and navigation in the past Web.

ACKNOWLEDGMENT

This research was supported by the MEXT Grant-in-Aid for Scientific Research in Priority Areas entitled: Content Fusion and Seamless

Search for Information Explosion (#18049041, Representative Katsumi Tanaka), and by the Informatics Education and Research Center for Knowledge-Circulating Society (Project Leader: Katsumi Tanaka, MEXT Global COE Program, Kyoto University) as well as by the MEXT Grant-in-Aid for Young Scientists B entitled: Information Retrieval and Mining in Web Archives (#18700111).

REFERENCES

Allan, J. (Ed.). (2002). *Topic detection and tracking: Event-based information organization*. Norwell, MA, USA: Kluwer Academic.

Allan, J., Gupta, R., & Khandelwal, V. (2001). Temporal summaries of news topics. In *Proceedings of the 24th Annual Conference on Research and Development in Information Retrieval*, (pp. 10-18). New Orleans, LA, USA: ACM Press.

Amitay, E., Carmel, D., Herscovici, M., Lempel, R., & Soffer A. (2004). Trend detection through temporal link analysis. *Journal of the American Society for Information Science and Technology, 55*(14), 1270-1281.

Arms, W.Y., Aya, S., Dmitriev, P., Kot, B. J., Mitchell, R., & Walle, L. (2006). Building a research library for the history of the Web. In *Proceedings of the Joint Conference on Digital Libraries*, (pp. 95-102). Chapel Hill, NC, USA: ACM Press.

Arvidson, A., Persson, K., & Mannerheim, J. (2000). The Kulturarw3 project—the Royal Swedish Web Archive—an example of "complete" collection of Web pages. In *Proceedings of the 66th IFLA Council and General Conference,* Jerusalem, Israel.

Aschenbrenner, A., & Rauber, A. (2006). Mining Web collections. In J. Masanes (Ed.), *Web archiving* (pp. 153-174). Berlin, Heidelberg, Germany: Springer-Verlag.

Bar-Yossef, Z., Broder, A. Z., Kumar, R., & Tomkins, A. (2004). Sic transit gloria telae: Towards an understanding of the Web's decay. In *Proceedings of the 13th International Conference on World Wide Web*, (pp. 328-337). New York: ACM Press.

Berger, A. L., & Mittal V. O. (2000). OCELOT: A system for summarizing Web pages. In *Proceedings of the 23rd Conference on Research and Development in Information Retrieval*, (pp. 144-151). Athens, Greece: ACM Press.

Brewington, E. B., & Cybenko, G. (2000). How dynamic is the Web? In *Proceedings of the 9th International World Wide Web Conference*, (pp. 257-276). Amsterdam, the Netherlands: ACM Press.

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001). Seeing the whole in parts: Text summarization for Web browsing on handheld devices. In *Proceedings of the 10th International World Wide Web Conference,* (pp. 652-662). Hong Kong, SAR, China: ACM Press.

Chi, E. H., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., & Card, S. K. (1998). Visualizing the evolution of Web ecologies. In *Proceedings of Conference on Human Factors in Computing Systems*, (pp. 400-407), Los Angeles: ACM Press.

Cho, J., & Garcia-Molina, H. (2000). The evolution of the Web and implications for an incremental crawler. In *Proceedings of the 26th International Conference on Very Large Databases*, (pp. 200-209). Cairo, Egypt: ACM Press.

Cho, J., & Garcia-Molina, H. (2003). Estimating frequency of change. *Transactions on Internet Technology, 3*(3), 256-290.

Cooley, R. Srivastava, J., & Mobasher, B. (1997). Web mining: Information and pattern discovery on the World Wide Web. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence,* (p. 558). IEEE Press.

Delort, J.-Y., Bouchon-Meunier B., & Rifqi, M. (2003). Enhanced Web document summarization using hyperlinks. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, (pp. 208-215). Nottingham, UK: ACM Press.

Fetterly, D., Manasse, M., Najork, M., & Wiener, J. (2003). A large-scale study of the evolution of Web pages. In *Proceedings of the 12th International World Wide Web Conference*, (pp. 669-678). Budapest, Hungary: ACM Press.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th International World Wide Web Conference,* (pp. 491-501). New York: ACM Press.

Hallgrimsson, Þ., & Bang S. (2003). Nordic Web Archive. In *Proceedings of the 3rd Workshop on Web Archives in conjunction with the 7th European Conference on Research and Advanced Technologies in Digital Archives*, Trondheim, Norway. Springer-Verlag.

Jatowt, A., & Ishizuka, M. (2004a). Summarization of dynamic content in Web collections. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, (pp. 245-254). Pisa, Italy: Springer-Verlag.

Jatowt, A., & Ishizuka, M. (2004b). Temporal Web page summarization. In *Proceedings of the 5th Web Information Systems Engineering Conference,* (pp. 303-312). Brisbane, Australia: Springer-Verlag.

Jatowt, A., & Ishizuka, M. (2006). Temporal multi-page summarization. *Web Intelligence and Agent Systems: An International Journal, 4*(2), 163-180. IOS Press.

Jatowt, A., Kawai, Y., Nakamura, S., Kidawara, Y., & Tanaka, K. (2006). Journey to the past: Proposal for a past Web browser. In *Proceedings of the 17th Conference on Hypertext and Hypermedia*, (pp. 134-144). Odense, Denmark: ACM Press.

Jatowt, A., Kawai, Y., & Tanaka, K. (2007). Detecting age of page content. In *Proceedings of the 9th ACM International Workshop on Web Information and Data Management*. Lisbon, Portugal: ACM Press.

Jatowt, A., & Tanaka, K. (2007). Towards mining past content of Web pages. *New Review of Hypermedia and Multimedia, Special Issue on Web Archiving*, *13*(1), 77-86. Taylor and Francis.

Kleinberg, J. M. (2003). Bursty and hierarchical structure in streams. *Data Mining Knowledge Discovery, 7*(4), 373-397.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations, 2*(1), 1-15.

Kumar, R., Novak, P., Raghavan, S., & Tomkins, A. (2003). On the bursty evolution of Blogspace. In *Proceedings of the 12th International World Wide Web Conference*. Budapest, Hungary: ACM Press.

Li, Z., Wang, B., Li, M., & Ma, W.-Y. (2005). A probabilistic model for retrospective news event detection. In *Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval,* (pp. 106-113). Salvador, Brazil: ACM Press.

McCown, F., & Nelson, M. (2006). Evaluation of crawling policies for a Web-repository crawler. In *Proceedings of the 17th Conference on Hypertext and Hypermedia*, (pp. 157-168). Odense, Denmark: ACM Press.

McCown, F., Smith, J.A., & Nelson, M.L. (2006). Lazy preservation: Reconstructing Web sites by crawling the crawlers. In *Proceedings of the 8th ACM International Workshop on Web Information and Data Management,* (pp. 67-74). Arlington, VA, USA: ACM Press.

Mei, Q., & Zhai, C-X. (2005). Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the*

11th International Conference on Knowledge Discovery and Data Mining, (pp. 198-207). New York: ACM Press.

Ntoulas, A., Cho, J., & Olston, C. (2004). What's new on the Web? The evolution of the Web from a search engine perspective. In *Proceedings of the 13th International World Wide Web Conference,* (pp. 1-12). New York: ACM Press.

Papka, R. (1999). *Online new event detection, clustering and tracking.* Unpublished doctoral dissertation, Department of Computer Science, University of Massachusetts, USA.

Rauber, A., Aschenbrenner, A., & Witvoet, O. (2002). Austrian online archive processing: Analyzing archives of the World Wide Web. In *Proceedings of the 6th European Conference on Digital Libraries*, (pp. 16-31). Rome, Italy: Springer-Verlag.

Shipman, F. M., & Hsieh, H. (2000). Navigable history: A reader's view of writer's time: Time-based hypermedia. *New Review of Hypermedia and Multimedia, 6*, 147-167. Taylor and Francis.

Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. In *Proceedings of the 23rd Conference on Research and Development in Information Retrieval*, (pp. 49-56). Athens, Greece: ACM Press.

Toyoda, M., & Kitsuregawa, M. (2003). Extracting evolution of Web communities from a series of Web archives. In *Proceedings the 14th Conference on Hypertext and Hypermedia,* (pp. 28-37). Nottingham, UK: ACM Press.

Viégas, F., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the CHI Conference*, (pp. 575-582). Vienna, Austria: ACM Press.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining,* (pp. 424-433), Philadelphia, PA, USA: ACM Press.

Yamamoto, Y., Tezuka, T., Jatowt, A., & Tanaka, K. (2007). Honto? Search: Estimating trustworthiness of Web information by search results aggregation and temporal analysis. In *Proceedings of the APWeb/WAIM 2007 Conference,* (pp. 253-264). Hunagshan, China: Springer-Verlag.

ENDNOTES

1. Internet Archive: http://www.archive.org
2. This efficiency is important in case when stream data is required.
3. WERA: http://archive-access.sourceforge.net/projects/wera
4. Wayback Machine: http://www.archive.org
5. Wikipedia, the free encyclopedia: http://en.wikipedia.org/wiki/Wiki