Chapter 8.15
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.

DOI: 10.4018/978-1-59904-576-4.ch017

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