A Dynamic Access Planning Method for Information Mediator*

Yasuhiko Kitamura, Tomoya Noda, and Shoji Tatsumi

Department of Information and Communication Engineering
Faculty of Engineering, Osaka City University
3-3-138 Sugimoto, Sumiyoshi-ku, Osaka 558-8585, Japan
{kitamura, tnodai, tatsumi}@kdel.info.eng.osaka-cu.ac.jp
http://www.kdel.info.eng.osaka-cu.ac.jp/~kitamura/

Abstract. The Internet is spreading into our society rapidly and deeply and is becoming one of our social infrastructures. Especially, WWW technologies are widely used for doing business and research, creating communities, disseminating personal information, and so on. We usually access WWW pages one by one through a browser, but we can add more value to them by integrating information collected from various WWW sites. However, to realize such WWW information integration, we face obstacles such as distributed information sources, access cost, and frequently and asynchronously updates of information. We here adopt mediator, which integrates information from distributed information sources, with cache mechanism to reduce access cost. We also propose a dynamic access planning method to cope with frequently updating information sources. In a limited time period, it can construct an appropriate answer by accessing information sources effectively considering reliability and quality of cached data. We show its performance through a real-world flight information service comparing with a conventional access strategy.

1 Introduction

The Internet is spreading into our society rapidly and deeply and is becoming one of our social infrastructures which are indispensable for our daily life. Especially, WWW technologies are widely used for doing business such as electronic commerce, doing research, creating special interest communities, disseminating personal information, and so on. We usually use the WWW by accessing WWW pages one by one through a browser, but we can add more value to it by integrating information collected from various WWW sites. Of course, we can make a link page or a portal site where related WWW pages are hyperlinked. Search engine is another elaborated approach for information integration, which dynamically generates a list of WWW pages for specified keywords. However, these approaches provide only a link set of related WWW pages and leave users collecting the pages, extracting data from them, and integrating the data to

* This work is partly supported by NTT Communication Science Laboratories.

© Springer-Verlag Berlin Heidelberg 2000
obtain desired information. To reduce this bothering work, WWW information integration [7] aims at providing a way to collect, extract, and integrate information from various WWW sites flexibly and automatically.

A typical example of WWW information integration is flight information service. In Japan, each of major airline companies provides a flight information service which we can consult about flight schedule and availability through the WWW. To our data input about departing date, origin, and destination, it returns a list of flight number, schedule, and availability. However, in case several airline companies operate on common routes, we need to access each of the WWW sites respectively to find connecting flights over different airline companies. WWW information integration service can provide a comprehensive and unified view of flight information over multiple information sites by collecting and integrating the information on behalf of users.

However, we have three major issues to tackle for achieving WWW information integration as follows.

1) **Distributed Autonomous Information Sources:** WWW information sources are distributed on the Internet and maintained individually. Hence, we need to collect related information from distributed information sources.

2) **Access Cost:** It takes time and cost to access information sources through the Internet. Moreover, WWW information integration often needs to collect a large amount of data, so it takes much time and cost.

3) **Frequently and Asynchronous Updates:** Some information sources are frequently and asynchronously updated. The timing of update depends on the source, so we cannot know when it is updated until we actually access it.

To cope with (1), we adopt mediator [15] which integrates distributed information as shown in Fig. 1. When a mediator receives a query from a user, it accesses multiple WWW sites to collect WWW pages. It then extracts data from the collected pages to construct an answer to the user's query. To cope with (2), we can cache collected data from WWW sites, so we can improve the response time to users because we can reduce the number of accessing WWW sites.

However, as shown in (3), WWW sites are frequently updated, so cached data may well be obsolete shortly and may lead to construct an incorrect answer. On the other hand, if we do not use a cache mechanism, we need to take a long time to collect a large amount of data. Hence, how to collect data properly in a limited time from frequently updated WWW sites becomes an important research issue. In this paper, we propose a dynamic access planning method for information mediator which collects and caches data to construct an answer to user's query. It makes an access plan dynamically considering reliability and quality of cached data to construct a proper answer in a limited time.

In Section 2, we propose a dynamic access planning method, and show its performance by applying it to a real-world flight information service in Section 3. We discuss related work in Section 4 and conclude our discussion in Section 5.