Discovering and Analyzing World Wide Web Collections

Sougata Mukherjea
Verity Inc., Sunnyvale, CA, USA

Abstract. With the explosive growth of the World Wide Web, it is becoming increasingly difficult for users to discover Web pages that are relevant to a topic. To address this problem we are developing a system that allows the collection and analysis of Web pages related to a particular topic. In this paper we present the system’s overall architecture and introduce the focused crawler used by the system. We also discuss the various techniques we use to allow the user to analyze and gain useful insights about a collection. Finally, we present some statistics on the collections.

Keywords: Authorities; Focused crawling; Graph algorithms; Hubs; Site graph analysis

1. Introduction

The World Wide Web (WWW) is undoubtedly the best source for getting information on any topic. Therefore, more and more people use the Web for topic management (Amento et al., 1999): the task of gathering, evaluating and organizing information resources on the Web. Users may investigate topics both for professional or personal interests.

Generally the popular portals or search engines like Yahoo and Google are used for gathering information on the WWW. However, they are not suitable for users who use the Web to gather detailed information on a topic of interest, for a number of reasons:

- The explosive growth of the Web poses basic limits of scale for today’s generic search engines. Thus much relevant information may not have been gathered and some information may not be up to date. For serious understanding of a topic having the latest information is essential.
- Most queries generate a large number of results and the search engines show these results as pages of scrolled lists. Going through these pages to retrieve the relevant

1 Some sections of this paper have appeared in the Proceedings of the WWW9 Conference (in Amsterdam) and the Hypertext 2000 Conference (in San Antonio).

Received 8 Mar 2002
Revised 5 Aug 2002
Accepted 10 Sep 2002
information is tedious. More sophisticated techniques for organizing and analyzing the information are needed.

- The search engines retrieve documents containing the user-specified keywords. Some relevant documents may not be retrieved because they may not contain that specific word. For example, if a user is interested in Wireless Technology, searching using those words won’t retrieve pages that have information on WAP or Bluetooth even though they are very relevant. Therefore, a more intelligent way of gathering the Web pages of interest is needed.

- Most corporate intranets have a large amount of information that is of interest. These cannot be collected by the search engines because they are under the firewall. It will be of great help to corporate users if both the intranet and Internet information can be gathered, organized and presented in a single unified view.

Because of these problems, there has been much awareness recently that for serious Web users focused portholes are more useful than generic portals (Chakrabarti et al, 1998). Therefore, systems that allow the user to collect and organize the information related to a particular topic and allows easy navigation through this information space is becoming essential.

We have built a specialized search engine (SSE) for the collection and organization of information on the Web related to a particular topic. The system uses a focused crawler to gather relevant pages. It also allows the user various ways to analyze the collection. This paper discusses the various features of the system. The next section cites related work and Section 3 describes the architecture of the system. Section 4 explains our focused crawler, which collects information from the WWW about a particular topic. Section 5 discusses the various analysis techniques that are present in the system to allow the user to gain useful insights about the information space. In Section 6 we present some statistics on the collections. Finally, Section 7 concludes the paper.

2. Related Work

2.1. World Wide Web Topic Management

In recent times there has been much interest in collecting Web pages related to a particular topic. Focused crawlers for collecting topic-specific Web pages are presented in Hersovici et al (1998), Chakrabarti et al (1999) and Aggarwal et al (2001). Our focused crawler is similar to these systems. However, we have incorporated some heuristics to improve performance. These are explained in Section 4.

Mapuccino (formerly WebCutter) (Maarek and Shaul, 1997; Hersovici et al, 1998; Ben-Shaul et al, 1999) and TopicShop (Terveen and Will, 1998; Amento et al, 1999) are two systems that have been developed for WWW topic management. Both systems use a crawler for collecting Web pages related to a topic and use various types of visualization to allow the user to navigate through the resultant information space. While Mapuccino presents the information as a collection of Web pages, TopicShop presents the information as a collection of Web sites. We believe that it is more effective to present the information at various levels of abstraction depending on the user’s focus.

2.2. Analyzing Web Page Collections

Modifications of the traditional information retrieval clustering techniques are being applied to the WWW. For example, Pirolli et al (1996) and Pitkow and Pirolli (1997)