Topical Crawling for Business Intelligence

Gautam Pant and Filippo Menczer*

Department of Management Sciences
The University of Iowa, Iowa City IA 52242, USA
{gautam-pant,filippo-menczer}@uiowa.edu

Abstract. The Web provides us with a vast resource for business intelligence. However, the large size of the Web and its dynamic nature make the task of foraging appropriate information challenging. General-purpose search engines and business portals may be used to gather some basic intelligence. Topical crawlers, driven by richer contexts, can then leverage on the basic intelligence to facilitate in-depth and up-to-date research. In this paper we investigate the use of topical crawlers in creating a small document collection that helps locate relevant business entities. The problem of locating business entities is encountered when an organization looks for competitors, partners or acquisitions. We formalize the problem, create a test bed, introduce metrics to measure the performance of crawlers, and compare the results of four different crawlers. Our results underscore the importance of identifying good hubs and exploiting link contexts based on tag trees for accelerating the crawl and improving the overall results.

1 Introduction

A large number of business entities — start-up companies and established corporations — have a Web presence. This makes the Web a lucrative source for locating business information of interest. A company that is planning to diversify or invest in a start-up would want to locate a number of players in the area of business. The intelligence gathering process may involve manual efforts using search engines, business portals or personal contacts. Topical crawlers can help in extracting a small but focused document collection from the Web that can then be thoroughly mined for appropriate information using off-the-shelf text mining, indexing and ranking tools.

Topical crawlers, also called focused crawlers, have been studied extensively in the past [6,7,3,10,2]. In our previous evaluation studies of topical crawlers, we found a similarity based Naive Best-First crawler to be quite effective [11,12, 18]. However, the Naive Best-First crawler made no use of the inherent structure available in an HTML document. We also studied algorithms that attempt to identify the context of a link using a sliding window and a distance measure based on number of links separating a word from a given link. However, such approaches

* Current affiliation: School of Informatics and Computer Science Department, Indiana University. Email: fil@indiana.edu

T. Koch and I.T. Sølvberg (Eds.): ECDL 2003, LNCS 2769, pp. 233–244, 2003.

often performed no better than variants of the Naive Best-First approach [12, 18]. In this paper we consider a crawler that identifies the context of a link using the HTML page's tag tree or Document Object Model (DOM) representation.

The problem of locating business entities on the Web maps onto the general problem of Web resource discovery. However, it has some features that distinguish it from the general case. In particular, business communities are highly competitive. Hence, it is unlikely that a company's Web page would link to a Web page of its competitor. A topical crawler that is aware of such domain level characteristic may utilize it to its advantage.

We evaluate our crawlers over millions of pages across 159 topics. Based on the number of topics and the number of pages crawled per topic, our evaluations are the most extensive in the currently available topical crawling literature.

2 The Problem

Searching for URLs of related business entities is a type of business intelligence problem. The entities could be related through the area of competence, research thrust, comparable nature (like start-ups) or a combination of such features. We start by assuming that a short list of URLs of related business entities is already available. However, the list needs to be further expanded. The short list may have been generated manually with the help of search engines, business portals or Web directories. An analyst may face some hurdles in expanding the list of relevant URLs. Such hurdles could be due to lack of appropriate content in relevant pages, inadequate user queries, staleness of search engines' collections, or bias in search engines' ranking. Similar problems plague information discovery using Web directories or portals. The staleness of a search engine's collection is highlighted by the dynamic nature of the Web [9]. Cho and Garcia-Molina [4] have shown, in a study with 720,000 Web pages and spanning over 4 months, that 40% of the pages in the `.com` domain changed every day. Hence, it is reasonable to complement traditional techniques with topical crawlers to discover up-to-date business information.

Since our focus is on studying the effect of different crawling techniques, we do not investigate the issue of ranking. We believe that ranking is a separate task best left until a collection has been created. In fact, many different indexing, ranking or text mining tools may be applied to retrieve information from the collection. Our goal is to find ways of crawling and building a small but effective collection for the purpose of finding related business entities. We measure the quality of the collection at various points during the crawl through precision-like and recall-like metrics that are described next.

3 The Test Bed

For our test bed, we need a number of *topics* and corresponding lists of related business entities. The hierarchical categories of the Open Directory Project