**Finding what is missing from a digital library: A case study in the Computer Science field**

This article proposes a process to retrieve the URL of a document for which metadata records exist in a digital library catalog but a pointer to the full text of the document is not available. The process uses results from queries submitted to Web search engines for finding the URL of the corresponding full text or any related material. We present a com-prehensive study of this process in different situations by investigating different query strategies applied to three general purpose search engines (Google, Yahoo!, MSN) and two specialized ones (Scholar and CiteSeer), considering five user scenarios. Specifically, we have conducted experiments with metadata records taken from the Brazilian Digital Library of Computing (BDBComp) and The DBLP Computer Science Bibliography (DBLP). We found that Scholar was the most effective search engine for this task in all considered scenarios and that simple strategies for combining and re-ranking results from Scholar and Google significantly improve the retrieval quality. Moreover, we study the influence of the number of query results on the effectiveness of finding missing information as well as the coverage of the proposed scenarios.

**1. Introduction**

On-line access to the full text of cataloged documents is an important requirement for satisfying the needs and expecta-tions of the users of a digital library (DL) of scientific articles (Laender et al., 2008). However, in many of such DLs, mainly those built by aggregating metadata from heterogeneous sources, not all metadata records have a direct pointer (e.g., a URL) to the corresponding full text. This situation is also common when the DL makes available information about citations and references but there is no direct access to the referenced items.

Even the presence of a direct pointer may not be useful for the user, for instance, in the cases in which the access to the full text requires the payment of a fee or the pointer became invalid due to the Web dynamics. In these cases, a service that retrieves the respective missing full texts from other Web sources would be of great value to the DL’s users.

In this article, we propose a process to provide such a service. It explores general purpose and specialized Web search engines to retrieve the URLs of full-text documents for which metadata records exist in a DL catalog but a full-text pointer is not available. The idea is to explore the potentiality of the existing search engines and to study how they behave in this specific task. For this, we experimented with records of documents from DLs in the Computer Science field. We also study the potential of the proposed process in finding other related documents and information that may be useful for the user, such asother publications (e.g., a thesis or dissertation) authored by one of the authors of the searched article and additional meta-data that complements the information already in the DL (e.g., references). In our experiments, we consider content freely available on the Web (e.g., in one of the authors’ homepage) as well as that provided by restricted sources (e.g., a publisher’s Web site).

We investigate how to specify, for each considered search engine, the most effective queries for the task at hand and which search engine is the best given distinct user requirement levels. We also explore search engine combination strategies to improve the overall effectiveness of our process. Finally, we analyze the capability of the proposed process to find the de-sired content in the considered scenarios.

In our evaluation, we use metadata records from the BDBComp – Brazilian Digital Library of Computing1 metadata catalog (Laender, Gonçalves, & Roberto, 2004) complemented with a set of records extracted from DBLP – The DBLP Computer Science Bibliography2 corresponding to conference papers authored by Brazilian researchers but not present in the first collection. We randomly sampled a set of records from these two collections and used them to build queries which we submitted to three general purpose search engines (Google, Yahoo!, and MSN) and two specialized ones (Scholar and CiteSeer) aiming at retrieving the corresponding full texts or other relevant but missing information. Our experimental results demonstrate that our proposed process is effective and provides a very simple strategy for finding the full text of documents cataloged in a DL for which a cor-responding URL is missing. They have also shown that, among the five tested search engines, Scholar is the most effective one for this task and that, when combined with Google, significant gains are achieved for all considered scenarios.

It is important to note that, despite the study presented in this article has been carried out with metadata records of doc-uments from Computer Science, a field very well represented on the Web, recent studies show that search engines, such as Scholar, also cover in a reasonable manner the content of other fields (Walters, 2007). Since the proposed process does not rely on intrinsic characteristics of any academic field, such as publication patterns, standards or preferences, but depends only on the metadata cataloged in DLs, we believe that it can be applied to other fields with similar results.

In summary, the main contributions of this article are:

(1) The proposal of a process for finding the URL of the corresponding full text, or of any relevant related material, for those documents cataloged in a DL but for which this information is missing.

(2) A comprehensive study of this process using different query strategies applied to different search engines and consid-ering different user needs and profiles.

(3) The proposal of a strategy for combining the results coming from specific search engines and re-ranking them, which improves the overall quality of the retrieved URLs.

(4) An analysis of the existent trade-off between the efficiency and the effectiveness of the proposed process.

The remainder of this article is organized as follows. Section 2 addresses related work. Section 3 describes the process proposed for retrieving the missing URLs. Section 4 discusses our experimental environment. Section 5 describes the meth-odology and the metrics used to compile the experimental results. Section 6 presents and discusses these results. Section 7 analyzes the impact of several factors in the likelihood of finding the missing URLs. Section 8 presents the conclusions and future work.

**2. Related work**

Current approaches to find documents missing from DLs rely mostly on focused crawlers (Chakrabarti, van den Berg, & Dom, 1999). For instance, in Zhuang, Wagle, and Giles (2005), the authors investigate the feasibility of using publication metadata to guide the crawler towards author’s homepages to harvest documents that are missing from a DL collection. However, relying on focused crawlers to maintain collections of scientific documents requires the construction of a complex software infrastructure. Therefore, in this work we advocate taking advantage of the current content already indexed by existing search engines but having just a small effort of formulating appropriate queries to such search engines.

The use of the infrastructure provided by search engines has been beneficial in many situations. In Qin, Zhou, and Chau (2004), for example, the authors discuss limitations of traditional focused crawling algorithms and argue that the use of meta-search can help overcome such deficiencies. They also propose that answers of queries submitted to search engines can be used to make more diverse the search space of such algorithms, which are normally limited to the content located close to the seeds selected as initial points for the crawling process. In Harrison and Nelson (2006), the authors describe strategies for finding information related to pages missing from Web sites. Cached versions of the missing pages retrieved from search engines are used for generating a lexical signature of a set of terms that captures the essential information pre-sented in the page which is then used to find similar documents or alternative copies of the original document. This strategy is the basis of a framework that aims at preserving the information available on the Web.

Some systems that provide searching or crawling services for scientific articles have been reported in the literature, such as HPSearch and Mops (Hoff & Mundhenk, 2001), and Paper Search Engine (PaSE) (On & Dongwon, 2004). However, these works focus on searching for scientific articles in general. In our work, we restrict our investigation to articles for which metadata records exist in a DL but a full-text pointer is not available. Thus, we evaluate the effectiveness of existing search engines to accomplish this task.

Finally, comparative studies evaluating the effectiveness of generic search engines to satisfy general information needs are very common (Bharat & Broder, 1998; Chu & Rosenthal, 1996; Gordon & Pathak, 1999; Lawrence & Giles, 1998, 1999). However, we have been unable to find any work comparing the use of generic and specialized search engines for the specific task described here.

**3. Proposed process**

We envision a service aimed at helping users to find DL missing content on the Web. The proposed process is depicted in Fig. 1. Interacting with the Digital Library Interface, if the user notices that a document d has no full text in the DL, she may request the service to search for the missing information. By using the Article Metadata record md, the Query Interface auto-matically generates and submits queries to one or more search engines requesting the missing information. Candidate URLs are extracted from the resulting pages as an ordered list Cd. The results in Cd follow a ranking that prioritizes answers coming from vertical search engines3 since they index collections of scientific articles, potentially minimizing noise. Results from a same search engine have their relative positions preserved.

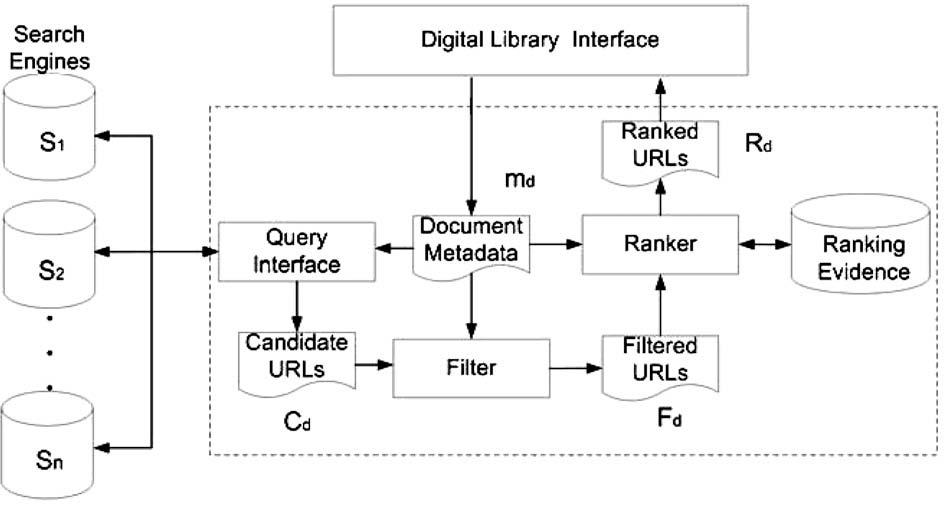


Fig. 1. Service architecture.

Next, the Ranker ranks the Fd list and generates a new ranked list Rd possibly using additional evidence such as the title of the returned document and the source of the result. The Ranker works by trying to put, on top of the ranked list, those documents with higher chance of satisfying the user’s needs, as we shall see later. The Rd list is then returned to the Digital Library Inter-face, which shows it to the user.

**4. Experimental environment**

In order to find the best configuration for the proposed service when searching for the full text of documents whose metadata records have been taken from the two Computer Science collections described next, we first investigated the effectiveness of individual search engines for this specific task. We tested five popular search engines available on the Web: Google,4 Yahoo!,5 MSN Search,6 Google Scholar,7 and CiteSeer.8 The first three are general purpose search engines and are among the ones with the largest audience on the Web. The last two are specialized search engines that index scientific publi-cations, being CiteSeer focused on the Computer Science field.

Our experimental environment is a simplified version of the service architecture described in Fig. 1, where the Sample Catalog contains metadata records of Computer Science conference papers for which a URL is missing from the DL and each query qd, generated for a metadata record md, is submitted to a single search engine in order to find a relevant URL for md. For each search engine, except CiteSeer, we developed a specific Query Interface that submits the queries directly to the respec-tive query processor and extracts from Pd, the set of returned pages, the title and the URL of the retrieved documents in order to create the list Cd of candidate URLs. For CiteSeer, the queries were submitted to Google, using its filter option, restricted to the CiteSeer domain. This was due to the fact that CiteSeer was constantly unavailable at the time of our experiments. We have adopted such an alternative based on the results of a previous experiment in which we randomly selected 1060 records from the CiteSeer metadata catalog9 and then submitted queries to Google requesting for related content. For 98 1% of these records, using a 95% confidence interval, it was possible to retrieve at least one document with a title similar to the one in the metadata record.

To create the Sample Catalog we carried out a stratified random sampling of two collections: (a) a set of 3969 metadata records obtained from the complete catalog of BDBComp, a collection of papers published in proceedings of major Brazilian Computer Science conferences, and (b) a set of 3181 records extracted from DBLP, which we call DBLP-Br, corresponding to papers published by Brazilian researchers in proceedings of international conferences. Note that no records from the second collection belong to the first one. The percentage of articles without a full-text URL is about 66% in BDBComp and 36% in DBLP-Br. The resulting Sample Catalog comprises 200 metadata records with missing URLs.

**5. Evaluation**

After submitting the queries to the five search engines, we combined the obtained result lists into a single one. This pro-cedure resulted in 3676 pairs ðmd; uÞ, where u is a plausible URL for accessing the full text (or related information) of a spe-cific paper a. To classify these pairs according to their usefulness, we first defined the different user scenarios we are interested.

Different users of a DL may have different interests and needs in different circumstances. For instance, a user looking for material on a certain topic may be interested in having access only to the abstracts of the related articles. Having found these abstracts, the interest of this user may be shifted towards their full-text content. However, if this user is not prepared to pay for this content, she might prefer to have access only to those articles whose full texts are freely accessible. Later on, when complementing some bibliographic references, this same user may become interested in missing metadata such as page numbers. Therefore, different sort of material may be useful in different situations. Thus, in this context, bibliographic mate-rial can be classified according to its content and accessibility, as described below.

With respect to content, we have considered the items in our result list as belonging to the following six categories:

(1) Full text: the URL u points to the full text (or to a document containing a pointer to it) of the article described by md.

(2) Similar full text: the URL u points to the full text (or to a document containing a pointer to it) of a document d similar to a, such as another related article or a thesis or dissertation authored by one of the authors of a.

(3) Useful metadata: the URL u does not belong to any of the above categories and points to a document containing meta-data information about a not present in md.

(4) Similar metadata: the URL u does not belong to any of the above categories and points to a document containing meta-data that describe a document d similar to a, such as a related article or a thesis or dissertation authored by one of the authors of a.

(5) Redundant metadata: the URL u does not belong to any of the above categories and points to a document describing only metadata information already present in md.

(6) Others: the URL u does not belong to any of the above categories.

Regarding accessibility, we have considered the items in our result list as belonging to the following two categories:

(1) Restricted: the URL u provides access to the full text (or related document) by means of some sort of payment or subscription.

(2) Free: the URL u provides free access to the full text (or related document).

Based on the previously described categories, we derived eight scenarios that model users with different interests, as shown in Table 1. In this table, each scenario is derived based on the type of content a user is interested in and the required accessibility, i.e., free (f) or restricted (r). In the Strict scenario, users are interested only in the full text of an article, no matter how it can be accessed, while in the Strict & Free scenario, the full text is required to be freely accessible. In the flexible sce-narios, besides the full text, related documents and unknown useful metadata may also satisfy the users. The scenario Strict & Restricted covers, for example, users interested only in ‘‘official” documents coming from trusted sources. In the At Least Metadata scenario, additionally to all relevant documents considered in the Highly Flexible scenario, users are also interested in documents that contain either redundant metadata describing a searched article or metadata describing another docu-ment similar to the desired one, no matter how they can be accessed. Finally, in the No Requirements scenario, users have no requirements. Any kind of content related to a searched article is considered relevant, no matter also how it can be ac-cessed. The last three scenarios are defined only for the purpose of analyzing coverage, as discussed in Section 7.

We then asked 27 subjects, members of our research group, to classify the resulting 3676 pairs as useful or not according to the previously described usage scenarios. For instance, a page containing additional metadata related about a paper but not its full text could be considered useful in the At Least Metadata scenario, but not useful in the Strict scenario.

To evaluate our results, we use three metrics: average precision at seen relevant documents (Pq), mean average precision (MAP), and mean reciprocal rank (MRR).