Retrieving metadata for your local scholarly papers

David Aumueller

University of Leipzig
david@informatik.uni-leipzig.de

Abstract: We present a novel approach to retrieve metadata to scholarly papers stored locally as PDF files. A fingerprint is produced from the PDF fulltext to query an online metadata repository. The returned results are matched back to identify the correct metadata entry. These metadata can then be stored in the PDF itself, indexed for a desktop search engine, and collected in a user's or community's bibliography. We think this hitherto missing link but with our tool now available data eases the organization of scholarly papers, and increases accessibility to one's collected academic content.

1 Motivation

Desktop search engines and the so-called semantic desktop depend on the availability of according metadata. When downloading PDFs from the web the according metadata attributes are lost unless sophisticated download managers are in place picking up the metadata from the repository the PDF is catalogued in. [Lu08] reviews a current tool that at least helps to collect the metadata of papers found on such scholarly repositories. **The missing link:** Papers locally stored as PDF often lack correct metadata and thus are hardly accessible even via current desktop search engines. E.g. the stored title in a PDF document often resembles the filename of the source document the PDF got produced from. The author/creator often is set to the login name of the user having created the PDF, or worse, the initial creator of the template the document was started on (cf. fig. 1).

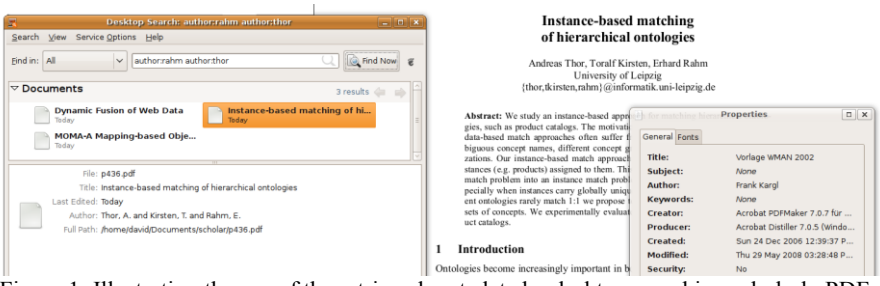

Figure 1: Illustrating the use of the retrieved metadata by desktop searching scholarly PDFs

For scholarly papers important metadata are authors, title, year, and reference (venue such as conference or journal), as available in metadata repositories such as the ACM digital library or Google Scholar (GS). Extracting such metadata directly from the PDF is difficult, cf. [Ha03], nevertheless GS, Citeseer etc. enlarge their repositories using extraction techniques. Generally, in the scholarly domain the mapping between the

metadata of an article and its fulltext as PDF file is indirectly given by following a download link on a webpage describing the article. This mapping though is lost when storing the file without reference to the source. Thus, local collections of articles grow without corresponding metadata. **Metadata to local files via web service:** We establish the mapping by matching the contents of local PDF files to a metadata repository. Analogue examples in other domains include the establishment of mappings between locally stored movie files and its corresponding film descriptions on movie web sites. Considering audio files, e.g. MP3-files, mappings to the corresponding metadata entries can be established using a variety of already existing tools that e.g. calculate a musical fingerprint and compare that to a user-contributed database on the web, cf. [HG04].

2 Approach – matching fulltext to metadata

Our novel approach to establish a mapping between locally stored scholarly articles in PDF file format and corresponding metadata entries consists of the following steps: converting the PDF to text, determining a suitable fingerprint, i.e. the query terms that are likely to locate the paper in a metadata repository, retrieving and parsing the resulting metadata entries, matching these with the document in question, and storing the according metadata. **PDF fingerprint:** As Google Scholar indexes the fulltext and not only the metadata of papers we can query this service using any identifying text fragment from the fulltext. Our approach takes a fragment from the beginning of the document which usually contains the content of identifying attributes, such as title and abstract excerpts. As we are using Linux' pdftotext we still have to replace some special characters such as ligatures with their counterparts in plain ASCII. Also, often author indices, such as [1, 2], need to be set off from the name itself. Thus, to be more robust, we restrict our search terms to phrases from the beginning of the document containing only words in the [a-z] character class. We experimented with the whole document head including title, authors, affiliations and email addresses but experienced issues regarding diverting textual representations locally and in the remote repository. If no match is found using the first bunch of phrases as query, it might be due to some extra information placed on the top of the document. In this case we skip the first phrase for a second query. Furthermore, at least the title of a document can more often also be acquired from general web search engines, as e.g. Yahoo also indexes PDF fulltexts and successfully extracts titles thereto. These titles could then be used to query the more specific, scholarly metadata repositories. **Result matching:** As querying a search engine will usually result in multiple different entries, the correct entry, if available, has to be matched to the document in question. We match back the result set to the fulltext of the local document by checking whether the title string is contained in the fulltext, preferably in the head, i.e. the fragment above the first mention of the keyword 'abstract'. For this match task we normalize both the title retrieved from the metadata service and the local fulltext alike. **Evaluation:** We evaluated our approach with local PDF files of the VLDB 2007 proceedings (91 papers in research track). For this set we yielded 100% accuracy, i.e. for each paper the corresponding metadata entry was found in Google Scholar. Generally, not every scholarly PDF file is indexed in Google Scholar, e.g. many articles from the BTW series are still missing. Surplus, PDF encryption or restrictions on reading may be in place hindering the extraction of text to create a query.

3 Application – using the scholarly metadata to local PDFs

Bibliographies: Metadata to a collection of scholarly PDFs can be maintained in local or remote databases, e.g. in the form of BibTex entries. As with our tool collected BibTex entries link back to the local PDF file, these fulltexts can be opened instantaneously from within the bibliography manager, e.g. JabRef. Community based bibliographies, such as the ones driven e.g. by Caravela [AR07], could be augmented with papers and their corresponding metadata including extracted abstract simply by uploading or pointing to PDF files. **Desktop search:** Generally, the more metadata attributes available in search engines the more expressive queries the users can pose. The retrieved metadata to local scholarly articles in PDF documents can be integrated in desktop search engines. Searching for locally stored papers e.g. via author name, title, year, and/or venue becomes possible. As proof-of concept we implemented a 'filter' for 'beagle', a popular desktop search engine for Linux. Figure 1 illustrates this by showing a rendered PDF in the background, its improper PDF metadata in the property window (right), and the article found by desktop searching for its real metadata (left).

4 Demonstration

On-site demonstration attendees may download scholarly PDFs off the net or supply otherwise and use our tool to retrieve according metadata. We showcase results within a bibliography manager as well as a desktop search engine (querying for some metadata to see whether the supplied PDFs are correctly indexed). The approach can be scrutinized step-by-step, i.e. converting the PDF to text, determining the query terms, inspecting the search results, matching the result entries to the document, and storing the according metadata. For a first experience we set up a demo on labs.dbs.uni-leipzig.de, incorporating our "PDF MEtadata Acquisition Tool" (pdfmeat), currently coded in Perl.

Querying web services to retrieve metadata for local files is quite new and applicable to other domains as well, e.g. to index movie details to locally stored films by querying an online movie database via cleaned filenames. Here, we presented our approach to establishing a mapping between local scholarly PDFs and its corresponding metadata entries and described how this data may increase accessibility to one's academic content.

References

[AR07] Aumueller, D., Rahm, E., Caravela: Semantic Content Management with Automatic Information Integration and Categorization. ESWC, 2007

[Ha03] Han, H. et al. Automatic document metadata extraction using support vector machines. Digital Libraries, 2003

[HG04] Howison, J., Goodrum, A. Why can't I manage academic papers like MP3s? The evolution and intent of Metadata standards. Colleges, Code and Intellectual Property Conference, 2004

[Lu08] Lucas, Daniel V. A product review of Zotero. Master's thesis, University of North Carolina at Chapel Hill, School of Information and Library Science, 2008