The Icecite Research Paper Management System

Hannah Bast and Claudius Korzen

Department of Computer Science, University of Freiburg, Germany
{bast,korzen}@informatik.uni-freiburg.de

Abstract. We present Icecite, a new fully web-based research paper management system (RPMS). Icecite facilitates the following otherwise laborious and time-consuming steps typically involved in literature research: automatic metadata and reference extraction, on-click reference downloading, shared annotations, offline availability, and full-featured search in metadata, full texts, and annotations. None of the many existing RPMSs provides this feature set. For the metadata and reference extraction, we use a rule-based approach combined with an index-based approximate search on a given reference database. An extensive quality evaluation, using DBLP and PubMed as reference databases, shows extraction accuracies of above 95%. We also provide a small user study, comparing Icecite to the state-of-the-art RPMS Mendeley as well as to an RPMS-free baseline.

1 Introduction

This paper is about *Icecite*, a new research paper management system (RPMS) that provides the following unique set of features:

(1) Automatic Metadata AND Reference Extraction: Icecite automatically extracts, with accuracies over 95%, bibliographic metadata (title, authors, year, conference, etc.) as well as references from academic research papers uploaded to the system.

(2) On-Click Download of New Papers: When reading a paper, other papers cited or listed in the reference section can be downloaded with a single click. Using the metadata from the reference extraction from (1), Icecite automatically searches the web for the correct PDF and uploads it to the system.

(3) Collaborative Annotation: Research papers can be annotated in the browser using the PDF standard. This ensures, that annotations remain modifiable in all standard (annotation-enabled) PDF viewers. Internally, annotations are kept separately from the PDF files. This enables collaborative annotation with other users in both online and offline mode (when annotating offline, annotations will be synchronized the next time the user goes online).

(4) Offline Availability: Icecite is web-based (no software download required), but papers can be read and annotated also when offline.

(5) Full-Featured Search: With Icecite, all the metadata, references, annotations, full texts as well as the underlying reference databases can be searched interactively (search as you type).

X. Lin et al. (Eds.): WISE 2013, Part II, LNCS 8181, pp. 396–409, 2013.

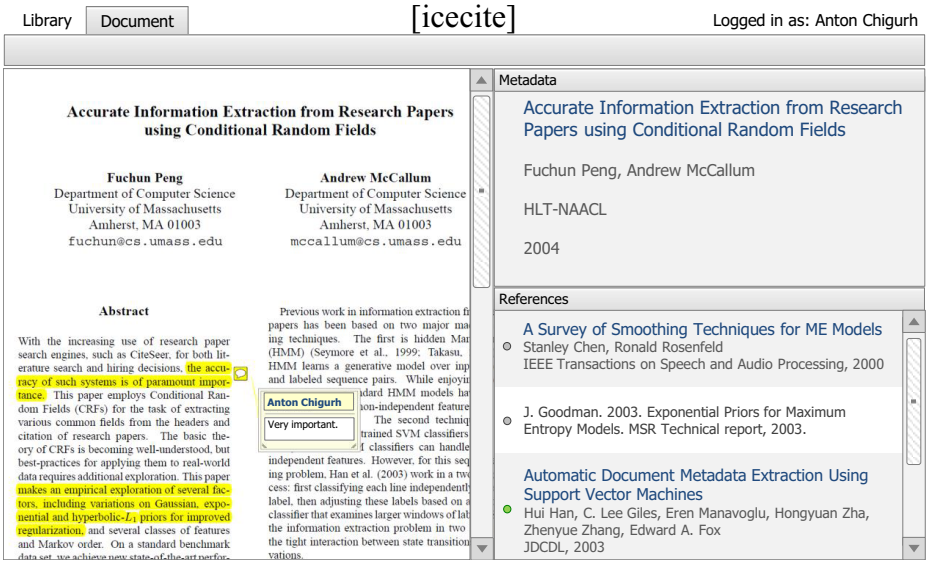

Fig. 1. A screenshot of the *Document View*. The left panel displays the PDF file, the right panels display the metadata (upper right) and the extracted references (lower right). The PDF file can be annotated in the browser using standard PDF annotations. The metadata and references panel can be arbitrarily resized, or hidden to display the PDF file in full screen mode. The references are listed with their full metadata. If no metadata record was found in the reference database, only the extract is displayed (as for the 2nd reference). The documents of the user are organized in a personal library (accessible by clicking the tab "Library" in the header). The colored bullet besides each reference indicates its availability in the user's library. A green bullet means: The document is already stored in the library and can be called by clicking it. A gray bullet means: The reference is not available in the library and can be clicked to import it.

The feature set described above looks quite natural and straightforward for a RPMS. However, *none* of the many existing RPMSs provides this combination of features. In fact, not one of these systems is able to provide even automatic metadata AND reference extraction (with acceptable accuracy). We provide an overview and comparison of fifteen RPMSs in Section 2.

Technically, Icecite combines known techniques in a (more or less) clever way to do what it does. The main idea behind the high-accuracy metadata and reference extraction is a combination of a rule-based recognition (of the passages in the text referring to metadata) with a fast index-based approximate search on a reference database. This is described in more detail in Sections 3 (metadata) and 4 (references). The results of our experimental evaluation, as well as a description of our reference databases are provided in Section 6.

The annotation and offline features are realized using the capabilities of the new HTML5 standard, namely its *Filesystem API* and the *Application Cache*. Annotations are merged using a standard text-based concurrent versioning