Header Extraction from Scientific PDF Documents

Mr. DLib[1], which represents an acronym for "machine-readable digital library" [4], is an easy to use web service providing academic literature and corresponding bibliographic data in different machine-readable formats, e.g. XML or JSON. Mr. DLib is a joint research effort of the projects Bibliographic Knowledge Network[2] (University of California, Berkeley), Docear[3] (University Magdeburg, Germany) [1,2] and SciPlore[4] (University Magdeburg, Germany / University of California, Berkeley) [3].

For the summer of 2012 we are offering internships to students participating in the RISE worldwide program, and who are interested in enhancing Mr. DLib. The intern positions are offered in cooperation with the Department of Statistics[5] at the University of California, Berkeley[6]. The interns will be supervised by the PhD students Jöran Beel and Bela Gipp and work at the UC Berkeley campus.

Project Description

Most documents in Mr. DLib are available in PDF format. For adding metadata of documents as new entries or matching them to existing records in the database, tools for extracting such data from the PDF documents are needed. In addition, Mr. DLib offers services for PDF metadata extraction to third parties, such as the academic literature suite Docear.

Interns will work on evaluating and improving of new or existing methods for automated PDF header extraction. It will be your task to find the best way to identify and extract authors, titles, affiliations, journals, etc. from the fulltext of a PDF. Your results will be used by Mr. DLib and other tools, such as Docear (whose predecessor SciPlore MindMapping already has attracted several thousand users). Depending on the quality of your work, we will also consider publishing a paper on your results recognizing you as an author.

Requirements

Successful participation in the project requires at least intermediate skills in programming using JAVA. Knowledge in statistics, other programming languages (especially C/++ or Python) and/or MySQL, Hibernate, Jersey, REST Web Services is beneficial, but not required.

Research vs. Programming

All our projects have a strong focus on both research and programming. Depending on your personal interests, goals and skills, your tasks during the internship may vary. Please let us know in your application whether you would prefer to focus on research, such as designing and evaluating concepts and algorithms (although, a certain level of programming will still be required), or implementing the aforementioned (yet, some research would still be required). A balanced mixture of both is also possible.

[1] http://www.mr-dlib.org/

[2] http://www.bibkn.org/

[3] http://www.docear.org/

[4] http://www.sciplore.org/

[5] http://www.stat.berkeley.edu/

[6] http://www.berkeley.edu/

The University and Around

The University of California, Berkeley is one of the world's most reputable universities. 70 Nobel Prize winners worked at UC Berkeley and only the best students are accepted to study here. Student tuition fees for nonresident and international students exceed US $20,000 per semester.

We can guarantee that there won't be any boredom during your time in Berkeley. Berkeley itself is a fascinating city with a wide variety of recreational, cultural and nightlife activities. Vibrant San Francisco with endless opportunities for exploration and exciting nightlife can be easily reached by public transportation in under 20 minutes. Also, the famous Silicon Valley, home of Google's Headquarter and virtually every known IT company is just about one hour away. During the weekends you could visit Los Angeles, go skiing in Lake Tahoe or hiking in some of North America's most beautiful national parks.

Administration and Housing

Interns will receive the status of "visiting student researcher" at UC Berkeley. As such, you will incur administration fees that amount to approximately US $1,000 including the fees for obtaining a visa.

We are flexible regarding the length and starting point of your internship. The internship may last anywhere from 6 to 12 weeks. If desired, we are willing to combine your internship with the supervision of a Bachelor, Master or Diploma thesis. Please contact us beforehand if you are considering to write a thesis related to your internship.

The costs for renting a room in a shared apartment in Berkeley are approx. US $600 (+/- US $100 depending on the size of the room and location of the apartment). Rooms shared with other persons are usually about US $400. We will support you in finding a room that meets your preferences.

Contact

If you have any questions, please contact us at info@mr-dlib.org (you may write in German).

About the Projects

The **Bibliographic Knowledge Network** is a project to develop a suite of tools and services to encourage formation of virtual organizations in scientific communities of various sizes, including conference groups and departmental research groups. The Bibliographic Knowledge Network will allow such organizations to filter out relevant documents from various input streams, select and enhance the quality of bibliographic data associated with the organization, and attract students, teachers and researchers to contribute to the activity of the organization.

Docear is an "academic literature suite" that bundles several applications for scientists: academic search engine, PDF reader, reference manager, word processor, mind mapping module, and recommender system.

The **SciPlore** project focuses on research of novel approaches in citation analysis for identifying and quantifying similarities between scientific articles [5]. The similarity assessments allow for improved clustering of similar documents, as well as the recommendation of academic literature. Furthermore, they can be used to detect forms of plagiarism [6] that could not be identified automatically so far. This was proven by the project team e.g. by analyzing the plagiarized doctoral thesis of Karl-Theodor zu Guttenberg [7]. In the future, the innovative technologies are supposed to significantly improve the quality of recommendation and plagiarism detection systems.

References

[1] Joeran Beel, Bela Gipp, Stefan Langer, and Marcel Genzmehr. Docear: An academic literature suite for searching, organizing and creating academic literature. In *Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL'11)*, 2011.

[2] Joeran Beel, Bela Gipp, and Christoph Müller. 'SciPlore MindMapping' – A Tool for Creating Mind Maps Combined with PDF and Reference Management. *D-Lib Magazine*, 15(11), November 2009. Brief Online Article. doi: 10.1045/november2009-inbrief.

[3] Bela Gipp, Joeran Beel, and Christian Hentschel. Scienstein: A Research Paper Recommender System. In *Proceedings of the International Conference on Emerging Trends in Computing (ICETiC'09)*, pages 309–315, Virudhunagar (India), January 2009. Kamaraj College of Engineering and Technology India, IEEE.

[4] Joeran Beel, Bela Gipp, Stefan Langer, Marcel Genzmehr, Erik Wilde, Andreas Nürnberger, and Jim Pitman. Introducing Mr. DLib, a Machine-readable Digital Library. In *Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL'11)*, 2011.

[5] Bela Gipp and Joeran Beel. Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. In Birger Larsen and Jacqueline Leta, editors, *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 2, pages 571–575, Rio de Janeiro (Brazil), July 2009. International Society for Scientometrics and Informetrics. ISSN 2175-1935.

[6] Bela Gipp and Joeran Beel. Citation Based Plagiarism Detection – A New Approach to Identify Plagiarized Work Language Independently. In *Proceedings of the 21th ACM Conference on Hyptertext and Hypermedia*. ACM, June 2010.

[7] Bela Gipp, Norman Meuschke, and Joeran Beel. Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag. In *Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL`11)*, Ottawa, Canada, 2011.

Docear and SciPlore are funded by

Gefördert durch:
Bundesministerium
für Wirtschaft
und Technologie
aufgrund eines Beschlusses
des Deutschen Bundestages

ESF
Europäischer Sozialfonds
für Deutschland