Plagiarism Detection in Open Access Publications

Jens Brandt*, Martin Gutbrod†, Oliver Wellnitz† and Lars Wolf*
*IBR,Technische Universität Braunschweig, Germany, {brandt|wolf}@ibr.cs.tu-bs.de
†Physikalisch-Technische Bundesanstalt (PTB), Braunschweig, Germany, {martin.gutbrod|oliver.wellnitz}@ptb.de

***Abstract*—The idea of Open Access publication is to provide free access to research articles on the Internet. In recent years, many research institutions as well as libraries worldwide started to provide all or parts of their publications following the paradigm of Open Access. Free access, however, bears the risk that the content may be misused without acknowledging the work of the original authors which should be avoided in any case. Therefore we started the Open Access Plagiarism Search (OAPS) project that exclusively uses OA documents for plagiarism checks. By offering this service free of charge to OA data providers, our project supports the OA community and at the same time acts as an incentive for closed access providers. In this work, we present the aims of OAPS and show how this project can help to ensure the integrity of OA publications.**

I. Introduction

In the last two decades a new paradigm for the publication of research results has grown in the research communities. Similar to the evolution on the Internet in the direction of people giving free access to all sorts of contents, like source codes, video streams or personal blogs, several researchers started to give free access to their scientific publications. This was the beginning of the Open Access (OA) movement which is still evolving. Today, OA publications are getting more and more popular in different research communities all over the world. This kind of publication increases the visibility of the research work and may also increase the quality of the publication compared to traditional publications. Different studies substantiate that an OA publication increases the number of citations of the publication which may in turn increase the standing of the author in the community. Lawrence (2001), for instance, found out that freely available conference articles in the area of computer science are cited more often than not freely available papers. Similar results were also reported for other scientific areas such as philosophy, political science, electrical and electronic engineering as well as for mathematics by Antelman (2004), for the area of astrophysics by Kurtz et al. (2005), or for 1.3 million cross-disciplinary papers of the Institute for Scientific Information (ISI) by Hajjem et al. (2005).

Freely available documents, however, bear the risk that they may easily be used by third persons without paying attention to the copyright of the original authors. There are several recent examples of this kind of copyright violation such as students copying contents from Wikipedia, PhD students copying text passages from the Internet, book authors using contents from Internet blogs or researchers using results and text from already published papers without acknowledging the original sources. Nevertheless, the unrestricted accessibility of OA publications is their main advantage, especially with regard to copyright protection. Due to their free availability, OA documents are also well-suited for automatic plagiarism search services.

This paper focuses on the potential of OA publications for automatic plagiarism search as well as on the mutual benefits of both aspects. The quality of pure OA publications can be increased by improving the efficiency of the review process as well as by the opportunity to continuously monitor potential copyright infringements for a large number of documents. The quality of plagiarism detection systems on the other hand may be increased by including OA documents. In contrast to documents which are freely accessible on the Internet, for instance on the author's personal website, OA documents are typically hidden from traditional web crawlers in so-called OA repositories. As presented by McCown et al. (2006), generic search engines like Google, Yahoo or Bing do not cover all documents that are available from OA repositories on the Internet. About 21% of the documents provided by OA repositories are not covered by major Internet search engines. Further, the main objectives of such generic search engines do not cover finding exact matches but more or less related matches according to the search request. Thus, the usage of existing OA repositories is beneficial for any plagiarism detection process.

In our current research project that is called Open Access Plagiarism Search (OAPS) we are developing an

online plagiarism search service for the OA community that exclusively uses OA publications for plagiarism detection. In this paper we present the aims and ideas of this project. The rest of this paper is organized as follows: Section II provides some background information about OA publishing and OA repositories. Afterwards, section III describes the type of plagiarism we are focusing on and gives some information about the existing plagiarism search service Docoloc, that we are using in OAPS. Section IV describes the aims and goals of our research project OAPS and presents the benefits that plagiarism detection may get from using OA documents as well as the benefits for the OA community from OAPS. Finally, section V concludes this paper.

II. Open Access

In 1991 Paul Ginsparg set up an online archive for preprints of scientific articles in his research area, a specific area of physics, at the Los Alamos National Laboratory (LAN-L). This archive was used by several researchers to share their results and can be seen as the first OA repository for scientific publications, although the notion of OA had not yet been introduced. Due to a rapid growth of the archive, as well as a growing acceptance in further areas of physics, mathematics, computer science, and biology, this repository evolved from a preprint archive for high energy physics to one of the largest OA repositories today: *arXiv.org*. Today *arXiv.org* contains about 594,500 free research articles which are freely accessible.

Ten years after the first version of *arXiv.org* had gone online, in 2001, the *Budapest Open Access Initiative (BOAI)* was founded by several European and American scientists to promote the idea of free access to scientific knowledge. They formulated the first defining statement about the new paradigm which they called ”open access” to scientific and scholarly literature: ”By open access to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself” *Budapest Open Access Initiative (BOAI)* (2002).

A. Different Ways to Open Access

In general there are two different possibilities for OA publishing: the green and the golden way to Open Access (Guedon; 2004). The golden way is the name for publishing original articles in OA journals that use a peer reviewing process similar to traditional journals. Instead of requesting subscription fees from the readers, OA journals request publishing fees from the authors. After publication, the articles are freely accessible on the Internet. A list of examples of such journals can be found in the *Directory of Open Access Journals (DOAJ)*. The green way to OA is the self-archiving of documents that are published with a traditional, non-OA publisher. Most publishers today allow the authors to publish preprints of submitted documents as well as often also postprints of reviewed and accepted articles on their personal or institutional website or repository. An overview of the rights that different publishers grant to authors is provided by *The RoMEO Project (Rights MEtadata for Open archiving)*.

B. Open Access Repositories

As the intention of OA publishing is the free and open accessibility of research works and results, the accessibility of OA articles is an important aspect. Similar to traditional publishers that provide their content via different interfaces including web portals that allow for searching and accessing the published articles, OA publications also need to be accessible and searchable in a convenient manner. Therefore, OA articles and documents are stored and provided in OA repositories. Such repositories may contain documents from one institution, i.e., so-called institutional repositories, or articles from a certain research area, i.e., disciplinary repositories. These OA repositories provide free access and may support some functionality of searching within the metadata as well as within the content of provided texts.

In the context of OA, data providers and service providers need to be distinguished. Data providers provide access to the relevant data of provided documents or items and service providers use data providers to provide different services upon the data and possibly upon different data providers. An institutional OA repository, for instance, is a data provider that provides access to OA documents and items of a certain institution. An example of a service provider is a search service that provides a full-text search interface to all documents provided by several different data providers. For interoperability between data and service providers, the *Open Archives Initiative (OAI)* developed the Protocol for Metadata Harvesting (OAI-PMH) (Lagoze et al.; 2002). This HTTP-based protocol allows well defined access to metadata records of all documents that are

provided by an OAI-PMH compliant repository. Such repositories are typically registered at meta registries like the *Registry of Open Access Repositories (ROAR)*, the *Directory of Open Access Repositories (OpenDOAR)*, or the *Open Archives Initiative List of Registered Data Providers*. A service provider may use this protocol to regularly harvest all metadata records from different data providers to build a new service upon this data.

III. Plagiarism Search with Docoloc

To detect text plagiarism, several different tools are available. Docoloc is an online plagiarism search service that started in 2005 for plagiarism detection in student work at Technische Universität Braunschweig, Germany (2005). It was born out of the necessity that a lasting use of eLearning can only occur if not only pure learning, but also all processes around learning are optimized and made more efficient and more effective. Today, Docoloc is used by several universities, schools and research labs all over the world with a main focus on institutions in Germany, Austria and Switzerland.

Documents are uploaded by the use of the Docoloc web application or by using a web service interface. The web-based interface is designed and optimized for fast and easy human usability. Less mouse-clicks, mouse-movements, page requests and user-interaction are aspired. The renouncement of graphic and smart web pages assures high speed in usage and also makes the use with smart-phones comfortable. Programmers can use a Simple Object Access Protocol (SOAP) interface to implement automated plagiarism checks in document management systems (DMS) or conference management systems. The *EDAS Conference Service* is such a system using Docoloc to search other published papers for possible plagiarism in the review process for conferences, workshops and journals.

The generated report is an interactive HTML document which can be archived off-line in repositories or databases and used later on whenever needed. The examination is running in background. Extensive reviews use the Internet and also private intranets. Docoloc is designed not only to find plagiarisms, but also to detect copyright infringements, quotations or other sources of the document on the Web.

Docoloc is the strategic partner for plagiarism detection in any kind of publications and provides good experiences in plagiarism search algorithms and the generation of detection reports. For each document that is checked by Docoloc a result report is generated that includes all sources from which some portions were

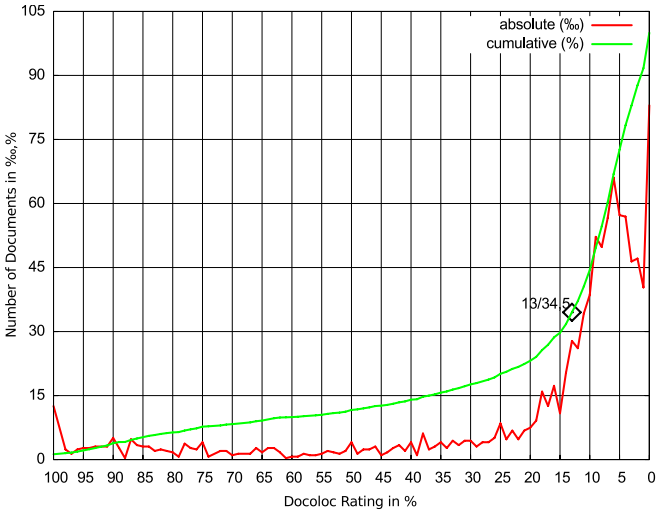

Fig. 1. Ratings of different documents checked by Docoloc

found in the analysed text. For the ease of use, these reports also include the Docoloc rating which is the percentage of fragments that were found in any other sources, compared to the total number of fragments that were inspected by Docoloc. Empiric studies showed that a Docoloc rating value greater than or equal to 13% clearly indicates plagiarism within the checked document. Figure 1 shows the results from the aforementioned investigation of the Docoloc ratings for all documents checked within one year. With regard to the cumulative values, 34.5% of all investigated writings have a rating of 13% or higher – what shows that around one third of all essays checked by Docoloc are not free of plagiarism, but are at least partially copied from one or several sources, depending on the document size. Thus, copying of text without acknowledging the sources really is a common problem.

IV. Open Access Plagiarism Search

In nearly all recent examples of copyright violations in scientific, academic and scholarly areas the original source of the plagiarised passages can be found on the Internet. This substantiates the suspicion that freely accessible content is more likely to be used without acknowledging the work of the original authors. One reason might be that a plagiarising author assumes that freely available content could be used without referencing the original work. Nevertheless, using text passages from any sources without mentioning these sources is a kind of plagiarism and needs to be avoided in any case.

Documents provided by OA repositories are very well suited to be included into existing plagiarism search

services because of their free accessibility from data providers. In the list of *Registered Service Providers of the OAI* several different metadata and full-text search services are available. However, none of them provides any plagiarism search or detection functionality. Therefore, we started the research project Open Access Plagiarism Search (OAPS) in 2009. The primary goal of OAPS is to fill this gap by building an OA service provider that provides a plagiarism search service to OA data providers. By providing this service free of charge, our project supports the OA community and helps to strengthen the quality of OA publications.

A. The OAPS Approach

To build a plagiarism search service, the contents of all documents that should be included in the search process needs to be searchable in an efficient manner. Although, OA publications are freely available on the Internet, they are not always covered by generic Internet search engines like Google, Yahoo or Bing. McCown et al. (2006), for instance, investigated that 21% of 3.3 million inspected OA documents were not covered by any major search engine. Additionally, generic search engines are typically optimized to find documents and websites that are somehow related to the given search term. For plagiarism search, however, one of the main objectives is to reliably find exact matches in authentic documents, rather than getting semantically related results on websites. Therefore, we are currently developing our own search index that contains the complete contents of all included texts, the origin of the documents as well as all available metadata. This search index provides an efficient and fast search interface that can fulfill the requirements of a plagiarism search service.

For building our search index, we are concentrating in the first phase of the project solely on documents that are provided by data providers via an OAI-PMH interface. This interface provides structured access to all metadata records of any OAI compliant repository. A very first run of our metadata harvesting process was able to harvest about 14.9 million metadata records from 1894 different OA repositories. Unfortunately, not all of these records contain an URL that links to the corresponding document. Moreover, when accessing those documents for which we were able to extract a URL from the metadata records we faced the problem: Most of these URLs do not link directly to the document itself, but link to a jump-off page containing the metadata in a human readable form. An URL usable for full-text harvesting is often missing. To solve these problems we are developing different algorithms and approaches to get access to the documents that belong to the provided metadata records.

To broaden the scope of our document index, we started to also include documents into the search index that are not provided by any OAI-PMH compatible repository but are provided on personal or institutional websites. Therefore, we are developing a specialized web-crawler that extracts information about documents from websites.

An optimized search index is an integral component for providing a reasonable plagiarism search service, but only if it is used by efficient plagiarism detection mechanisms. In the scope of OAPS, such mechanisms are provided by Docoloc as presented in the previous section. Each document that is submitted to OAPS to be checked is sent to the Docoloc servers by using a web service API. Docoloc in turn may send search requests to the OAPS search index and thereby can find candidates from which the inspected text might be copied from. The results are transferred back to the OAPS servers where they can be postprocessed and presented to the user of the OAPS service.

B. Benefits from Open Access

The most obvious benefit that a plagiarism detection service may gain from the OA community is the free accessibility of OA documents as well as the structured access via well defined interfaces. This provides the ability to build a dedicated document search index that contains all available OA documents. Compared to search indexes from traditional search engines on the Internet, this index can be optimized for plagiarism search mechanisms. Especially when trying to get results for exact match queries, traditional web search engines often produce lesser results than are truly available, which result from their optimization for typical web search queries. In contrast, a document index that is optimized for plagiarism search services can provide much better results.

Another benefit of OA is the fact that OA repositories not only provide access to the documents, but also provide metadata records for each contained document. These records typically include some set of meta information about the corresponding document or item, such as the names of the authors, the title of the document, the publication date, the type of document as well as an URL that links to the document itself. In case of plagiarism search, this metadata can be used to enhance the significance of the generated plagiarism reports. If,

for instance, some portions of an inspected document are found in a text of a law, this result might be less important than other portions that are found in an article of a conference proceedings or a scientific journal. Author information can also be used to distinguish between self-plagiarism and text that is plagiarised from other sources. The existence of this metadata records is a great benefit compared to the information provided by generic Internet search engines, as their quality typically is much better than meta information extracted from the documents itself in an automated way. Author and title information might be extracted automatically but further details like the information about the conference or journal an article originally belongs to cannot always be extracted automatically.

C. Integration of OAPS

As one of the goals of OAPS is to support the idea of OA publishing, we are also working on solutions to integrate an OAPS service into existing components in the OA community. One of these solutions, for instance, is the integration of an automated plagiarism check into popular repository implementations. This integration allows to check every document that should be included into the repository with OAPS and to detect any suspicious document.

For OA publishers we plan to provide an integration of the OAPS services into their peer reviewing process, which supports the reviewers to detect plagiarised portions of submitted documents.

V. Conclusions

The aim of OAPS is to strengthen the quality of Open Access publications and thereby also the acceptance of Open Access publications as well as the integrity of OA repositories, by providing an automatic plagiarism detection service for the OA community. Therefore, a specialized full-text search index of all available OA documents is currently developed. This index allows for efficient searching within the contained documents.

This work presented an overview about the goals and motivation of our research project, named OAPS, as well as a brief description of our approach. In the near future we will firstly provide the possibility to search within our document index on our web site, followed by a fully functional plagiarism search service for the OA community. By further extending our document index and by providing our services free of charge to OA service providers, OAPS will be a valuable service to strengthen the quality of OA publications.

VI. Acknowledgments

This work is funded by the German Research Foundation (DFG) within the framework of Scientific Library Services and Information Systems (LIS).

References

Antelman, K. (2004). Do open-access articles have a greater research impact?, *College and Research Libraries* **65**(5): 372 – 383.

arXiv.org http://arxiv.org

Budapest Open Access Initiative (BOAI) (2002). http://www.soros.org/openaccess/read.shtml

Directory of Open Access Journals (DOAJ) http://www.doaj.org

Directory of Open Access Repositories (OpenDOAR) http://www.opendoar.org

Docoloc Plagiarism Search (2005). http://www.docoloc.com

EDAS Conference Service http://www.edas.info

Guedon, J.-C. (2004). The "Green" and "Gold" Roads to Open Access: The Case for Mixing and Matching, *Serials Review* **30**(4): 315 – 328.

Hajjem, C., Harnad, S. and Gingras, Y. (2005). Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact, *IEEE Data Engineering Bulletin* **28**(4): 39–46.

Kurtz, M., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M. and Henneken, M. E. (2005). The Effect of Use and Access on Citation, *Information Processing and Management* **41**: 1395–1402.

Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S. (2002). The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0, http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.

Lawrence, S. (2001). Online or Invisible, *Nature* **411**(6837): 521–522.

McCown, F., Liu, X., Nelson, M. and Zubair, M. (2006). Search engine coverage of the OAI-PMH corpus, *IEEE Internet Computing Magazine* **10**(2): 66 – 73.

Open Archives Initiative List of Registered Data Providers (n.d.). http://www.openarchives.org/Register/BrowseSites.

Open Archives Initiative (OAI) http://www.openarchives.org

Registered Service Providers of the OAI http://www.openarchives.org/service/listproviders.html

Registry of Open Access Repositories (ROAR) http://roar.eprints.org

The RoMEO Project (Rights MEtadata for Open archiving) http://www.sherpa.ac.uk/romeo