



Technische
Universität
Braunschweig



Plagiarism Detection in Open Access Publications

Jens Brandt, Martin Gutbrod, Oliver Wellnitz, Lars Wolf

4th International Plagiarism Conference, 21-23 June 2010

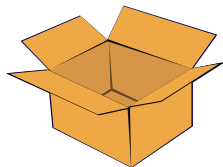
Outline

Introduction

Open Access

Open Access Plagiarism Search

Conclusion



Open Access

”[...] 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

[<http://www.soros.org/openaccess/read.shtml>]

Plagiarism and Open Access



Free access facilitates copying of third-party contents

- Students copy contents from Wikipedia
- PhD students copy contents from the Internet
- Book authors copy text from blogs

Free access facilitates plagiarism detection

- Internet search engines can be used to find the sources
- Automatic plagiarism search
- Avoidance of self-plagiarism

History of Open Access



1991

- Paul Ginsparg set up an online archive for preprints
- Provides access to articles in the area of high energy physics
- Today arXiv.org contains more than 600,000 documents

2001

- Budapest Open Access Initiative (BOAI)
- Founded by European and American scientists
- Formulated the first defining statement about Open Access

History of Open Access (cont.)

2003

- Bethesda statement on Open Access publishing
- Berlin declaration on Open Access to knowledge in the sciences and humanities



2004

- Organisation for Economic Cooperation and Development (OECD) statement on access to research data
- International Federation of Library Associations and Institutions (IFLA) statement on Open Access to scholarly literature and research documentation

Different Ways to Open Access



The green way to Open Access

- Open Access self-archiving
- Preprints or postprints
- Personal or institutional website

*RoMEO Project
(Rights METadata for
Open archiving)*

The golden way to Open Access

- Open Access publishing
- Peer reviewing process
- Publishing fees

*Directory of Open
Access Journals
(DOAJ)*

Open Access Repositories



- OA documents are stored and provided by OA repositories
- Institutional and disciplinary repositories
- Data providers provide access to relevant data
 - The metadata of the document
 - The document itself
- Service providers use existing data providers to build services
 - Services based on the data of several data providers
 - Examples: search engines, citation indexing

OAI-Protocol for Metadata Harvesting (PMH)



- Defined by the Open Archives Initiative (OAI)
- Interoperability between data and service providers
- Uses *Hypertext Transfer Protocol (HTTP)*
- Exchange of *XML*-Messages
- Provides access to metadata records
- Request information about the repository
- Different metadata standards
 - Dublin Core (mandatory)
 - Several different formats

Open Access Plagiarism Search (OAPS)



Goals

- Plagiarism search service for OA data providers
- Avoid text plagiarism in OA repositories
- Support the OA community
- Strengthen the quality of OA publications

Approach

- Development of a full-text index of available OA documents
- Implementation of a search engine for plagiarism checks
- Act as an OA service provider

The OAPS Approach

- Make OA documents available for plagiarism checks
- Google, Yahoo and Bing do not cover all available OA documents
 - 21% or 3.3 million inspected OA document were not covered (McCown et al., 2006)
- Internet search engines are not optimized for plagiarism checks

OAPS Approach

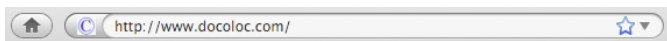
- Harvesting of available OA documents
- Specialized search index
 - Covers all available OA documents
 - Optimized for plagiarism checks
- Plagiarism detection service is provided by Docoloc

Plagiarism Detection with Docoloc

Docoloc

- Online plagiarism search service
- Started in 2005 at University of Braunschweig
- Main objective: plagiarism detection in student work
- Widely used in Germany, Austria and Switzerland
- Web service interface with SOAP
- Easy integration into existing systems
- Integrated into the EDAS Conference Service

Docoloc Web-Interface



Docoloc

ID: Password: [Log in](#)

[Quick Guide](#) [Create Account](#) [Add Paper](#) [View Reports](#)

Local file: [Use web-address](#)

[demo](#) professional

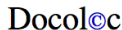
Send report: to browser to my account

by email:

[Contact](#) - [Terms & Prices](#) - [Policy](#) - [References](#) - [Help](#)

©2010 Docoloc KG - [IBR/ITM](#) research partner - Plagiarism search in billion documents
[german](#) [english](#)

Docoloc Report



Report

Digital signed

Reviewed document: **testfragments.txt**

Processing date: **Tue, 15.6.2010 23:42:46 CEST**

A total of **51** fragments were analysed. As a result **22** fragments (43.1%) were found in other documents. In the document preview below the fragments are marked **yellow** and clickable. At most 6 found documents are shown with same text passages.

Cross reference documents

Following list of found documents is grouped by document titles and ordered by found fragments. With a mouseclick on "**x** fragments" the relevant fragments in the document are colored **orange** and the window scrolls to the first location. Click on "**x** fragments" again resets the special marks.

5 fragments were found in a text with the title: "**KidsWingTsun und soziales Kompetenztraining für Kids im ...**", located on:

http://www.kidswingtsun-schule-koblenz.de/wtkids_gewalt.html

4 fragments were found in a text with the title: "**wingsun.de | Gewaltprävention | Warum brauchen Kinder und ...**", located on:

<http://www.wingsun.de/gewaltpraevention/warum-gewaltpraevention.html>

4 fragments were found in a text with the title: "**Skript: Qualitätsmanagement: Amazon.de: Rolf Mohr: Bücher**", located on:

<http://www.amazon.de/Skript-Qualit%C3%A4tsmanagement-Rolf-Mohr/dp/3638703967>

25% with fuzzy search (1 fragment)

We employed a Microstrip-to-CPW-to-Microstrip transition and via holes to transfer the current from the top to the bottom substrate layer and vice versa. The presented phase shifter is operating in a wide bandwidth between 5.5 and 17.2 GHz, with low insertion loss and reflection coefficients. Because the input and output microstrip lines are on the same layer, the presented phase shifter is suitable for a modified class of feeding networks for phased antenna arrays.

Various studies of throughput and channel utilization for wireless asymmetric channels, multi-hop interference, high traffic demand, and contention mechanisms for contention have been introduced to notify the network.

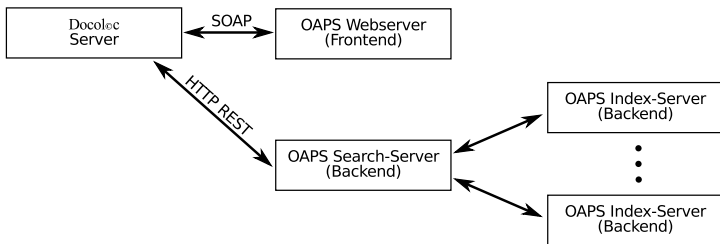
Given a mesh network, our goal is therefore to identify a set of maximal concurrent transmission sets whose union contains all the nodes in the network. By assigning one time slot to each of these sets, a node schedule can be created, maximizing throughput and ensuring that each node gets to transmit. This schedule is then repeated.

1 match:

PIER B Online - Wideband 180 Degree Phase Shifter Using ... Because the input and output microstrip lines are on the same layer, the presented phase shifter is suitable for a modified class of feeding networks ...
<http://ceta.mit.edu/pierb/pier.php?paper=07111507>

Interaction between OAPS and Docoloc

- Distinct user accounts
- OAPS uses the web service API of Docoloc
- Docoloc uses the OAPS search API



Full-Text Harvesting

Metadata Harvesting

- Protocol for Metadata Harvesting (OAI-PMH)
- Periodical harvesting of known repositories
- Use of meta-repositories
- Data provider may register repositories at OAPS

Data Extraction

- Extract full-text link from metadata records
- Extract text from document
- Support of different file types
- Harmonisation of metadata records

Benefits from Open Access



- Free and structured accessibility of OA documents
- Internet search engines does not cover all OA documents
- Use of metadata to increase the value of reports
 - Author information
 - Type of document
 - Date of publication
 - ...
- Build optimized search indexes

Integration

How can OAPS be used?

- Online service (web-based, API)
- Free of charge for OA data providers
- Integration into existing OA platforms
- Repositories may check every newly included document
- Integration into peer reviewing processes of OA publishers

Current Status



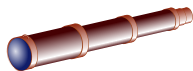
- Server infrastructure with 5 servers
- OAI-PMH Metadata harvesting
 - 3052 different OAI-PMH repositories
 - 14.2 million metadata records
 - 12.9 million records contain a link
- Development of different algorithms for full-text harvesting
- Harvesting of documents not available via OAI-PMH

Summary



- Plagiarism search service for the OA community
- OAPS is an OA service provider
- Harvesting of available OA documents
- Full-text search index, optimized for plagiarism checks
- Automatic plagiarism checks
- Strengthen the quality of OA publications
- Substantiate the integrity of OA repositories

Future Work



- Preview of the OAPS search index in July 2010
- First usable version of OAPS by the end of 2010
- Stable version in the mid of 2011
- Free of charge for OA data providers
- Business model for non-OA users
- Harvesting of further OA documents
- Integration of closed access contents

Questions?

Jens Brandt
brandt@oaps.eu

Open Access Plagiarism Search (OAPS)
<http://oaps.eu>

IBR, Technische Universität Braunschweig
<http://www.ibr.cs.tu-bs.de>

funded by



Deutsche
Forschungsgemeinschaft



Projekt Partners



funded by

Deutsche
Forschungsgemeinschaft

Plagiarism in Research and Education



- arXiv.org: 67 documents were deleted in 2007 due to plagiarism
(Nature, 2007-09-06)
- The IEEE starts using automatic plagiarism checks for all submissions to 24 journals and 30 conferences in 2010.
(IEEE, The Institute, 2010-02-05)
- Since 2006, the University of Klagenfurt checks all theses for plagiarism; two doctoral degrees were revoked.
(Kleine Zeitung, 2010-02-16)
- A professor from a university in Berlin plagiarised some portions of a judicial textbook.
(Spiegel, 2007-05-12)