

Plagiatsuche für Open Access

Jens Brandt

Institut für Betriebssysteme und Rechnerverbund (IBR)
Technische Universität Braunschweig

Fachtagung ViFa Recht - AjBD 2010
25./26. Februar 2010

Plagiat



”Plagiat ist geistiger Diebstahl. Schuldig macht sich, wer sich als Autor eines fremden Textes ausgibt und fremde Gedankengänge oder Argumente als die eigenen verkauft.”

(ZEIT Campus 01/2010)

Verschiedene Plagiate



- Wörtliche Übernahme fremder Texte (Copy & Paste)
- Übersetzung fremder Texte
- Paraphrasierung fremder Texte
- Übernehmen fremder Ideen oder Argumente
- Plagiiere eigener Texte (Eigenplagiat)
- Übernehmen von Ergebnissen, Daten, Bildern

⇒ Jeweils ohne Kenntlichmachung und Angabe der Quellen

Textplagiate in Wissenschaft und Lehre



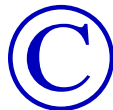
- arXiv.org: 67 Dokumente wurden 2007 wegen Plagiarismus entfernt (Nature, 06.09.2007)
- Die IEEE überprüft ab 2010 sämtliche Einreichungen zu 24 Magazinen und 30 Konferenzen. (IEEE, 05.02.2010)
- Die Universität Klagenfurt prüft seit 2006 sämtliche Abschlussarbeiten auf Plagiate; zwei Doktorinnen wurde der Dokortitel aberkannt. (Kleine Zeitung, 16.02.2010)
- Ein Professor der HU Berlin hat beim Verfassen eines juristischen Lehrbuches plagiiert. (Spiegel, 12.05.2007)

Gründe für Textplagiate



- Digitalisierung erleichtert das Kopieren fremder Inhalte
- Überwindung sprachlicher Hürden bei Publikationen
- Kulturell anderer Umgang mit geistigem Eigentum
- Publikationsdruck (z.B. wegen Drittmittelwerbung)
- Recyclen eigener Arbeiten verlängert eigene Publikationsliste
- Zeitdruck, bspw. Studiengebühr und viele BA/MA Module
- Mangels besseren Wissens

Sind Plagiate ein Problem?



- Erschleichung von Leistungen
- Integritätsverlust von Verlagen
- Missachtung der Urheberschaft

Vermeidung von Plagiaten

- Intensive Kontrollen und Begutachtungen
⇒ Zeitintensiv, kaum realisierbar
- Automatisierte Plagiatsuche erleichtert das Auffinden
- Aufklärung zu guter wissenschaftlicher Praxis
- Sensibilisierung für Plagiate

Plagiate und Open Access



Frei zugängliche Inhalte erleichtern ein direktes Kopieren

- Schüler und Studenten kopieren Inhalte aus Wikipedia
- Autorin Helene Hegemann schreibt aus einem Blog ab
- Doktoranden kopieren Texte aus dem Internet

Frei zugängliche Texte erleichtern die Erkennung von Plagiaten

- Internet Suchmaschinen fördern die Quellen zu Tage
- Automatisierte Plagiatsuche möglich
- Aufwertung der Ergebnisse durch Metadaten
- Vermeidung von Eigenplagiaten



Open Access Plagiarism Search



Ziele

- Neuartiger OAI-Serviceprovider zur Plagiatsuche
- Vermeidung von Textplagiaten in OA Repositorien
- Stärkung der Qualität von OA Veröffentlichungen
- Sensibilisierung für Textplagiate in der Wissenschaft
- Unterstützung in Begutachtungsprozessen

Umsetzung

- Aufbau eines Volltextindex aller verfügbaren OA Dokumente
- Volltext-Suchmaschine für OA Repositorien



Mehrwert durch Open Access



- Auswertung der Metadaten gefundener Quellen
- Kennzeichnung von Quellen aus speziellen Repositorien (bspw. mit DINI-Zertifikat)
- Anzeige erweiterter Informationen zu den gefundenen Quellen
- Bewertung der Wichtigkeit gefundener Referenzdokumente
- Nutzung von Zitationsanalysen für Dokumentenvergleiche

Projektpartner

Docoloc



TECHNISCHE UNIVERSITÄT
CAROLO-WILHELMINA
ZU BRAUNSCHWEIG

PTB

gefördert durch

DFG Deutsche
Forschungsgemeinschaft



Docoloc Plagiatsuche

Docoloc

- Webbasierter Dienst zur automatisierten Plagiatsuche
- Ursprung: Aufdeckung von Plagiaten studentischer Arbeiten
- Fokus auf Hochschullehre und den wissenschaftlichen Begutachtungsprozess
- Ausgründung des Instituts für Betriebssysteme und Rechnerverbund
- Verbreitung seit 2006:
 - über 65 Universitäten und Hochschulen (DE, A, CH)
 - über 70 Allgemein- und Berufsbildende Schulen (DE, CH)
 - Integration in das Konferenzmanagement System EDAS
 - Verlage, internationale Forschungseinrichtungen
 - 210.000 Dokumente mit 2,6 Mio. Seiten jährlich

Herkunftsreport

Digital signiert

Überprüftes Dokument: **testfragmente.txt-gsoap.txt**

Überprüft am: **Thu, 4.6.2009 11:41:45 CEST**

Es wurden insgesamt **47** Textstellen überprüft. Davon wurden **27** Textstellen (57,4%) in anderen Dokumenten gefunden. Die kritischen Textstellen wurden in der folgenden Dokumentenvorschau **gelb** markiert. Die Markierungen können angeklickt werden und zeigen daraufhin maximal 6 gefundene Quellen.

Referenzdokumente

Die folgende Übersicht ist gegliedert nach den Titeln der gefundenen Dokumente. Durch einen Klick auf „**x** Stellen“ werden die speziellen Stellen im Dokument in der Farbe **orange** hervorgehoben und direkt zur ersten Stelle gescrollt. Ein erneuter Klick auf „**x** Stellen“ setzt die Markierungen wieder zurück.

5 Stellen wurden gefunden in einer Textvorlage mit dem Titel: „**QUALITÄTSMANAGEMENT /v32706.pdf**“, zu finden unter:

<http://content.grin.com/document/v32706.pdf>

4 Stellen wurden gefunden in einer Textvorlage mit dem Titel: „**Hausarbeiten.de - Skript: Qualitätsmanagement - Skript**“, zu finden unter:

<http://www.hausarbeiten.de/faecher/vorschau/32706.html>

4 Stellen wurden gefunden unter

<http://buchwe>

Früher beschränkte sich Abschreiben auf das Kopieren von Hausübungen in der Schule und das Einsagen, Schwindelzettel oder Spickzettel schreiben oder ähnliches. Vereinzelt kam es auch vor, daß Studenten bei Ihrer Diplomarbeit ganze Auszüge aus Fachbüchern übernahmen, ohne den entsprechenden Quellenverweis anzugeben.

Ziel der Qualitätsverbesserung ist das "Null-Fehler-Prinzip". Damit wird nicht gesagt, dass ab sofort keine Fehler mehr auftreten dürfen, sondern dass wir uns dies beginnt schon im Kopf aller Beteiligten. Grundprinzip ist eines "Schuldigen". Je mehr über die Fehlerursache bekannt wird, desto besser kann diese beseitigt und zukünftig vorgebeugt werden. Dazu kommt erst nach der Fertigung erfolgt. Je später jedoch ein Fehler entdeckt wird, desto teurer wird seine Behebung (Abb.).

1 Treffer:

QUALITÄTSMANAGEMENT ab sofort keine Fehler mehr auftreten dürfen, sondern dass wir uns dieser Grenze annähern wollen. (Abb.). Abbildung 1-3: Grundsteine zum Null-Fehler-Prinzip ...
<http://content.grin.com/document/v32706.pdf>

Docoloc

Aufbau einer Plagiatsuche für Open Access

- Aufbau eines Volltextindex aller OA Dokumente
 - Aufbereitung verfügbarer OA Dokumente für die Plagiatsuche
 - Volltext Harvester
- Nutzung der Docoloc Plagiatsuche
- Zugriff auf den Volltextindex durch Docoloc
- Verteilte Architektur
- Integration der Plagiatsuche in OA Repositorien
- Entwicklung eines nachhaltigen Betreibermodells

Aufbau eines Volltext Harvesters

- Volltext Harvesting
 - Protocol for Metadata Harvesting (OAI-PMH)
 - Regelmäßiges Abfragen bekannter Repositorien
 - Nutzung von Meta-Repositorien
 - Data-Provider können Repositorien anmelden
- Fehlertolerante Datenextraktion
 - Unterstützung verschiedener Dokumentenformate
 - Harmonisierung von Metadaten
- Serverinfrastruktur
 - Lastverteilung
 - Redundanz

Warum eine weitere OA Suchmaschine?



Vorhandene Suchmaschinen mit Volltextindex

- Scirus (Elsevier Verlag, 370 Mio. Dokumente)
- Scientific Commons (Universität St. Gallen, 34 Mio. Dokumente)
- BASE (Universität Bielefeld, 20 Mio. Dokumente)
- Google, Microsoft, Yahoo, etc.

Mehrwert

- Schnittstellen und Abhängigkeiten
- Optimierung für Plagiatanalyse
- Metadaten als Ergänzung zum Volltext



Probleme



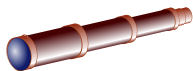
- Uneinheitliche Metadaten
- Fehlende Volltexte
- Rechtliche Einschränkungen durch Betreiber von Repositorien
- Beispiel: Analyse von 34 Repositorien mit DINI-Zertifikat:
 - 29 Repositorien bieten Volltexte nur über Umwege an
 - 5 Repositorien bieten direkte Links auf die Volltexte an
 - 2 Repositorien verbieten ein Volltext-Harvesting
 - Nur 3 Repositorien können uneingeschränkt genutzt werden
- Situation bei anderen Repositorien ähnlich

Projektstatus



- Förderung durch die DFG seit 2009 für 24 Monate:
- 25% der Projektlaufzeit vergangen
- Serverinfrastruktur mit 5 Servern
- Verteilte Softwarearchitektur für Volltextindex
 - ca. 2100 Repositorien
 - ca. 11 Mio. Datensätze
 - ca. 3,4 Mio Volltexte

Ausblick



- Plagiatsuche für Open Access
- Stärkung von Open Access Veröffentlichungen
- Volltextindex verfügbarer Open Access Dokumente
- Open Access Service Provider
- Erste nutzbare Version in 2010
- Frei zugänglich für Open Access Anbieter
- Integration von Closed Access Inhalten

Fragen?

Jens Brandt

brandt@ibr.cs.tu-bs.de

Institut für Betriebssysteme und Rechnerverbund
Technische Universität Braunschweig

<http://www.ibr.cs.tu-bs.de>

Open Access Plagiarism Search (OAPS)

<http://www.oaps.eu>