Test cases for plagiarism detection software

Debora Weber-Wulff

HTW Berlin
Treskowallee 8
10318 Berlin

weberwu@htw-berlin.de
http://www.f4.htw-berlin.de/~weberwu

Abstract

There are numerous plagiarism detection software systems that claim to discover plagiarism of all sorts, given a digital text. This paper first discusses a typology of plagiarism, which makes clear that plagiarism is more than just an exact copy. Then a collection of 42 test cases in German are presented that were developed at the HTW Berlin for testing plagiarism detection software. The test cases have been used in three tests, in 2004, 2007, and 2008 and are available online. The test suite will be extended to include English-language test cases in 2010.

Keywords

plagiarism, plagiarism detection systems, test cases, taxonomy

1. Introduction

There are many software systems that are called plagiarism detection systems (PDS), which claim to be effective in finding plagiarism in a text that is available in a digital form (Weber-Wulff, 2009). These systems are marketed to universities in order to help them discover students using unreferenced sources, to companies to help them find plagiarism of their work online, and to publishers to help them discover plagiarism before they are in print.

Many institutions in the market for such software need some way of testing the effectiveness of the systems. In particular, they want to measure how much actual plagiarism can be detected and whether or not systems can deal with the minor changes made, for example, when paraphrasing is done. Texts with clearly defined plagiarism portions can be used to assess how effective the systems are. Original matter is also needed, so that it can be determined how the systems respond to this situation. Some institutions just take a few pages found on the web as test material, but there have been a number of surveys that used systematically prepared test material.

In 2001, the Joint Information Systems Committee in the UK (JISC) tested five plagiarism detection systems using eleven documents from six academic disciplines grouped into four types of plagiarized material (Bull, Collins, Coughlin & Sharp, 2001). In 2007 JISC commissioned Scaife (2007) to repeat the test. Eleven systems were tested with eleven test cases: Three essays were based on a PDF web page, an extract from a book, an extract from an online course, two essays taken from popular web pages, an original document, an extract from a book which is also available online, a purchased existing essay, and a purchased made-to-order essay[1]. An Austrian survey of systems (Maurer, Kappe, & Zaka, 2006) used one English paragraph that was heavily paraphrased to evaluate popular systems. In Sweden, two systems were tested (Nordström & Sjöberg, 2006a, 2006b) using 20 popular publications from four different departments in Swedish and in English.

At the HTW Berlin[2] we have tested plagiarism detection systems three times since 2004. We have developed a collection of test cases in German that we use to test plagiarism detection systems, most recently in 2008. Our corpus now includes test cases for collusion as well, and is available online as exercises for an eLearning unit for teachers about plagiarism called "*Fremde Federn Finden*"[3].

---

[1] These test cases are available by CD from JISC.

[2] The school changed its name from *FHTW Berlin* to *HTW Berlin* in April 2009.

[3] Literally: "finding false feathers", http://plagiat.htw-berlin.de/ff/uebung/start (in German)

2. Plagiarism and Collusion Detection

The exact algorithms that are used by PDS are often closely guarded industrial secrets, although many mechanisms have become visible during our system tests and are described elsewhere (Maurer, Kappe, & Zaka 2006; Schleimer, Wilkerson, & Aiken, 2003). Maurer et al. discuss document source comparison, manual search of characteristic phrases, and stylometry as an attempt to identify authors. Schleimer et al. present the fingerprinting and searching method called winnowing for detecting partial digital copies that is used in their MOSS plagiarism detection system.

Some companies use their own databases, others use the programming interfaces offered by the various search machines to look for possible matches. For example, the product Turnitin[4] advertises that is stores material such as term papers, which have been submitted to it for a plagiarism check, so that it can flag future copies of this paper. This has, of course, legal implications[5] and requires extremely large databases.

Some systems store hashes of documents gleaned from the Internet and then compare these with a hash fingerprint of the document to be examined. This mechanism has often proven to be very susceptible to slight changes in a document producing both false positives and false negatives. Just the use of German umlauts will often cause a plagiarism not to be found, as we have demonstrated in our tests. Also, such systems must constantly scour the Internet for new material.

The systems that use one or multiple search machine APIs are not able to check an entire document, so they will choose a sample string from the document, and send it to the search machine. In return, a list of candidate matches is returned. The candidates must each now be compared in more detail with the document in question in order to determine the degree of plagiarism – if any.

A number of the systems appear to use the free interfaces that are made available by the search machines, for example the Google SOAP API[6]. These interfaces offer only a fixed number of searches within 24 hours. This means that while the first plagiarism tests work relatively quickly, subsequent tests will take longer, so that a user may have to wait overnight for results.

---

[4] This product by iParadigms LLC uses the same database as other company products such as iThenticate. http://www.turnitin.com

[5] iParadigms LLC partially won an appeals suit brought against it by four high-school students in the US in April 2009 (http://pacer.ca4.uscourts.gov/opinion.pdf/081424.P.pdf) but the EU Intellectual Property Rights Center (http://plagiat.htw-berlin.de/ff/support/5_2/einreichungsdienste) finds that EU law does not permit such use of student papers without explicit permission.

[6] http://code.google.com/apis/soapsearch/, only personal use permitted, limited to 1000 searcher per day. The new AJAX API described in http://code.google.com/apis/ajaxsearch/ offers unlimited searching for commercial use, but the results must not be transformed.

Collusion is a special case of plagiarism that is much easier to find with help of software. Collusion happens when additional copies – perhaps slightly modified – are found in a group of documents to be checked. This might be the case in a large university class, for example, of 200 or more, in which a group of 3 or 4 students write on one paper that each submits as their own in a slightly modified form.

Collusion is much easier to detect than plagiarism from an external source, as the possible partners in collusion are all available. This restricts the search space so that a comparison of each paper with all others is feasible for reasonable numbers of documents (less than 100).

3. A Typology of Plagiarism

There are many different definitions of plagiarism to be found in the literature, many quite far-reaching and encompassing aspects of scientific misconduct such as ghostwriting or data falsification. The Modern Language Association defines the forms of plagiarism (Gibaldi, 2003) to

> *"include the failure to give appropriate acknowledgment when repeating another's wording or particularly apt phrase, paraphrasing another's argument, and presenting another's line of thinking."*

This definition misses one important aspect that has to be dealt with, especially in non-English speaking countries: translations. It also does not differentiate between what can be seen as two different forms of para–phrasing, shake & paste (a term we introduced) and clause quilts (sometimes called mosaics). We differentiate six different forms of plagiarism (Weber-Wulff & Wohnsdorf, 2006) that were used in the construction of the test cases.

3.1 Copy & Paste

These is the easiest kind of plagiarism to make – a portion of text (or the entire paper) is marked and with a keystroke combination a copy is made that can be simply inserted into another text. This is the kind of plagiarism that is assumed to be easy to find, as the passages are identical: The sequence of characters, including spaces and punctuation marks, is the same in the original and in the copy.

3.2 Translations

For a translation, the plagiarist chooses a text portion in a language different from the target language and either uses an online translation tool such as Babelfish or Google Translator to produce a rough draft, or produces a hand translation. Many students are not aware that the quality of a software-produced translation is not the best. A native speaker will quickly see that the text is in some way odd. Some will pick up on word order, others on strangely wrong word selections, or incorrect grammar.

Some students feel that through the work done in preparing a hand translation, effort has been expended and should be rewarded. Indeed, such a translation is a lot of work, but it is not original work, as it is still entirely based on the work of someone else.

3.3 Shake & Paste Collections

Another technique used in the preparation of a plagiarized paper is to take a number of sources and to copy the material paragraph-wise from the various sources. This tends to give the reader the impression that the paragraphs were put into a bag, shaken well, and taken out for pasting in a rather random order.

A human reader will often pick up on this type of plagiarism because of the changes in style, diction, or even formatting. Researchers in the area of intrinsic plagiarism analysis (Stein & Meyer zu Eissen, 2007) attempt to identify these kinds of plagiarism, as this avoids the complexities of having to address matches found in unspecified locations throughout the Internet.

3.4 Clause Quilts or Mosaics

There seems to be a belief amongst students[7] that changing a specific number of words from the source somehow makes a paper not be a plagiarism. Howard (1999) calls this kind of paraphrasing "patchwriting". Students take bits and pieces of text from different authors and edit them, changing an adjective here or switching the word order there. Some even go to the trouble of lifting only a clauses and particularly apt terms, and then stitching what ever else they find together to produce a text that reminds of a crazy quilt of causes or a mosaic.

3.5 Structural Plagiarism

In structural plagiarism, the plagiarist will paraphrase another author without giving credit. This can include using the argumentative structure, the sources (sometimes even in the exact order as in the original work), the experimental setup, or even the research goal.

This kind of plagiarism is quite hard to detect or even to prove, given the two sources in question, but is often the basis of heated accusations amongst scientists that others have "stolen" their ideas.

3.6 Collusion

Collusion, as briefly discussed above, is the term used to describe the situation in which either a number of authors cooperate on writing a paper

---

[7] lol (2008). How many words do you have to change in a document for it not to be plagiarised? http://answers.yahoo.com/question/index?qid=20080504003338AAegQeG Accessed 14 May 2010

and each turns it in as her own, or when each make a few small changes to the text before turning it in.

Collusion occurs both in natural language papers as well as in programming language assignments, in which one student solves the problem and the others make slight changes (especially in the variable naming scheme or the commenting structure) before handing in the result.

3.7 Plagiarism is more than copying

As can be seen from the typology given above, exact copies of sources are not found in two of the six kinds of plagiarism (translations, structural plagiarism). With clause quilts, there are only small portions of text that are exact copies, although a quick read will show the similarity of the texts.

On the other hand, original papers often contain copies of text from other works – properly quoted and sourced. There are numerous ways of setting off quoted content, using single or double quotation marks (Danish, English, French, and German use different symbols, and there are also the so-called straight quotation marks), or by indenting larger portions of text. So if a system purports to detect plagiarism, it must not count proper quotations as plagiarism.

Thus, plagiarism detection and copy detection are two different things. And as we have often shown in our tests, PDS find copies, not plagiarism.

4. Test cases for plagiarism detection systems

This section discusses the development of test cases for plagiarism detection systems. These test cases were designed to test how well the different plagiarism detection systems detect the different types of plagiarism. Short original papers were also included, in order to see if any false positives would be flagged. Since there is no known data on exactly how students plagiarize and to what extent they combine techniques, in general only one type of plagiarism was used in each plagiarism test case, although a few with multiple plagiarism types have been included.

Permission was obtained from each copyright owner for their material to be used in the test and published online as exercise material in the eLearning unit. The test cases are guarded with commands to search machines to disregard them upon indexing, so that they are not included in any search results. But although there is a *robots.txt* file and a

```
<meta name='robots' content='noindex,nofollow' />
```

in the header of all files, both Bing and Google have recently started indexing the files. This poses a problem both for testing and for using the test cases in teaching teachers how to search for plagiarism using search machines, as the file itself is marked as a possible source.

4.1 Testing Plagiarism Detection Systems

In 2004, the author was preparing an eLearning unit on plagiarism detection for teachers. She was often asked what software was recommended for catching plagiarists. It seemed that a good way to test the software would be to measure how effective the systems were in differentiating original papers from plagiarized ones. Ten papers were written (numbered 0-9 in the collection), three were original and seven were various types of plagiarism. In this test, only a yes/no evaluation of the results was undertaken: did the software "correctly" judge the test case to be a plagiarism or not. But it quickly became clear during the test that PDS are not litmus tests, cleanly marking the presence or absence of plagiarism. In particular, complicated plagiarism involving multiple sources were not well identified by the software.

During the summer of 2007 the test was repeated re-using the test cases from 2004 and adding twelve new ones. One test case was just a copy of one of the old test cases, but with a German umlaut in the name of the file. It had been determined in the first test that the German umlauts, which are not part of the portion of the ASCII code traditionally used in English-language computing, cause an enormous amount of trouble for some systems.

Another test case was a copy of one that had included many footnotes (a structural plagiarism), but without the footnotes. The footnotes were causing many systems to flag the test case as a plagiarism for the wrong reason, that is, the references fit exactly to references in other papers. This kind of false positive is very confusing for a teacher using the system, as the text in question was clearly not identical to the "source" identified by the software, although a high plagiarism index was indeed indicated.

Since many software companies announced new versions of their systems shortly after the test results for 2007 were published, the test was repeated in the summer of 2008[8]. We used 20 additional test cases: Eleven new papers were produced, one duplicate test case was made replaced all the letters "e" in one paragraph with the Greek letter epsilon to see if it confused the systems (for some it does), and eight collusions of other test cases were prepared.

The scoring for this test ended up having to be readjusted during the test and all of the previous tests repeated, because some of the test material was suddenly pirated on other sites and findable with Google, two sources disappeared in the middle of the test, and the Wikipedia article used in one plagiarism had been highly edited.

This is a problem that is often encountered when using plagiarism detection software in real life. Between the time that a teacher discovers the plagiarism with the help of software, and the time that the case is dealt with by an honor board, the Internet sources may change. As an administrative procedure, the

[8] The test is planned for repeating in 2010, this time extending the test to English-language constructed plagiarism.

conservation of copies of sources is absolutely necessary. Even if the sources for our test cases are no longer available online, they are still plagiarism, although no longer easily detectable by software. The results for all three test series, including the English-language summary of the tests[9] are available online.

4.2 The Test Cases

There are now 42 test cases available in the corpus, all in German and about a page or a page and a half in size. Although this is not enough to thoroughly exercise all aspects of the various plagiarism detection systems, since the plagiarism amount is known and there are a class-sized number of test cases, this allows for the simulation of a real-life use case for such systems.

The test cases were all hand-produced in a manner simulating what we suppose to be the method that a student would use to produce a paper. For many of the cases, topics of general interest were posed and then a search was made to see what was available on the internet.

We used the change tracking mechanism from Word to document what was copied and what was changed. This helps produce a good visualization of the plagiarism. We then measured the amount of plagiarism in terms of words and characters for those test cases which were not 100% plagiarisms.

The test cases are discussed in detail in the appendix.

4.3. Collusion Test Material

For the two papers 29 and 30 (one original, one a partial plagiarism), eight additional versions of the test cases were prepared. There were four collusions each made of each papers. Version *a* had the first and last sentences different, version *b* had the first paragraph completely different, version *c* substituted synonyms throughout the text, and version *d* used a different font for the text.

For testing purposes we mixed these collusions in with the other test cases in order to create a class-sized collection of papers. Three systems were able to flag the collusions completely (Weber-Wulff, 2008). The complete results of this test are also available online[10].

---

[9] Results from 2007 were presented at the 3rd International Plagiarism Conference in 2008, a paper with the 2008 results was submitted to *Plagiary* for review in February of 2009. Although sent for review in March 2009, numerous inquiries have gone unanswered, so the English version has been published on the web site as of May 2010. The results are available in German at http://plagiat.htw-berlin.de/software/2008/.

[10] http://plagiat.htw-berlin.de/software/2008/collusion/.

4. Summary

Over a period of five years, a collection of short test cases in German for plagiarism detection software systems has been developed. There are 31 different basic test cases, and an additional eight cases that are collusions of two of the test cases. There are also copies of one test case without the references, one with a character replaced by a similar character, and one with an umlaut in the file name, making a total of 42 test cases. The majority of the test cases are 100% plagiarized in the sense of the definition of plagiarism, even if many are not copy and paste equivalents but have been edited in some manner using the editing operations defined in this paper.

The materials were prepared with the consent of the copyright owners and are available online as HTML documents. A ZIP-file with the texts in various formats is available from the author upon request. The test cases will be expanded to include English-language test cases for the 2010 test.

Acknowledgements

I am indebted to my research students Gabriele Wohnsdorf, Martin Pomerenke, Katrin Köhler, and Martin Hauffe for their help in conducting this research. Matrin Pomerenke and Katrin Köhler prepared many of the test cases. I also wish to thank my university for graciously granting me research time and research students in order to support my plagiarism work and David Zellhöfer for a critical reading of the text.

Appendix

This section describes the test cases in some detail.

0. **Leap year**: This is an original paper about the history of leap years, but with a properly quoted table that is often flagged as plagiarism by software.

1. **Djembe**: This is a translation of an English-language essay about the Djembe drum that was done using the online translation tool Babelfish. Numerous plagiarism of the English original essay exist, and many teachers participating in a course on plagiarism detection using this material have easily found either the original or another plagiarism. However, no software system has ever even come close to finding the source, although many “untranslatable” words are kept from the original. The picture included is also taken from the original work.

2. **Atwood**: This book report copies some paragraphs from the official review at the Amazon site with two paragraphs (932 characters / 141 words) from an anonymous review cut in. A typo (capital “I” in the middle of a sentence) is one of the key markers that something is wrong. Some sites have a plagiarism of the site.

3. **IETF:** This paper is taken from a technical report about the structure of the Internet. Four pages were copied, the quotations removed, two

spelling errors fixed, and the lead sentence rewritten. There exists at least one plagiarism in an online exhibition catalogue that was discovered during the first test. Although the original authors had requested takedown, the material was still to be found in the 2008 test.

4. **Döner**: This essay about the popular Turkish street food, is a carefully crafted shake & paste from three sources, a scientific one, a popular one, and the German Wikipedia. A second version of this paper was stored with a German umlaut in the file name.

5. **Telnet:** This paper is one that was actually submitted by a student to a colleague. It plagiarizes a bootleg PDF copy of a hacker's book that circulates on the Internet. The PDF was scanned and has typical character recognition errors, such as "~" instead of "-". The student was aware that the plaintext dates in the telnet protocol looked odd (i. e. they were far too old), so these values were changed. The time stamps on the telnet commands, however, were not changed, which struck my colleague as being quite odd. A search on a timestamp is quickly successful.

6. **Friðrik Þór Friðriksson***:* This is an original biography written about the Icelandic film director containing 21 Icelandic characters and 2 Danish characters in the names. After the 2004 test the paper was included in the German Wikipedia with the correct author named in the history tab. Some students do this – put their reports online before they are graded – and it can cause a false positive, especially if a teacher does not check the authorship of the Wikipedia article given as the source by the PDS.

7. **Maple Syrup**: This report is a clause quilt from a children's TV show script available online and an article from the Wikipedia.

8. **Reinhard Lettau**: This original biography was placed by the author in both the English and the German Wikipedia and noted as such. There exist an enormous number of legal and illegal copies of the Wikipedia online, making test cases 6 and 8 appear to have very many sources to some systems.

9. **Grass frogs**: This essay was purchased from a paper mill and is used by permission. Human searchers (biology teachers) have found the schoolbook from which this paper was cribbed, there is now a PDF version of the book available online. It can be seen by teachers to be a plagiarism, as it uses German spellings from before the last reform[11]. The unreferenced picture is from a public domain animal picture database.

10. **Fraktur**: This paper about a German type family is taken from a PDF that is itself written in a *Fraktur* typeface. That means that all ligatures are encoded, and since every second or third word includes a ligature, it is highly unlikely for this to be found by a software match (i.e. a hash (#) encodes ff, sz encodes ß, and so on). There are also 6 Scandinavian

[11] Official German spelling was reformed in 1996 with a goal of simplifying spelling rules, with major revisions occurring in 2004 and 2006. http://rechtschreibrat.ids-mannheim.de/

ligatures in the text. Paragraphs from the PDF are mixed as a shake & paste collection with paragraphs from a book about *Fraktur*, including 10 pictures of words with ligatures.

11. **Henning Mankell**: This book report about a detective story by the Swedish author is an exact copy from the Internet (including typographical errors). During the 2007 test, an online student plagiarism was found. The author of the original book report was successful in having that site taken down.

12. **Microbreweries**: This test case is a hand translation from an article in the English Wikipedia about small breweries.

13. **Allspice**: This paper is a translation into German from an English translation of a Swedish original chapter in a book about spices. It is a shake & paste of paragraphs. It sticks out for a German-speaking teacher reading it because it discusses the Danish and Swedish names and uses of allspice, instead of German ones.

14. **Max Schmeling**: This biography of the German boxing legend is original, but the footnotes are made up. The information was found in a tourist brochure, so since this was not quotable, a scientific journal of local history (that does not exist) was invented as the source. This is not counted as plagiarism.

15. **Public toilets**: This report is taken from a DVD version of a German technology encyclopedia that was published in 1910 and is now in the public domain. Even though this is not a copyright problem, it is still plagiarism, as the source is not given. The dates in the footnotes have 100 years added to them to look more modern. The five pictures illustrating the work are the copperplates found in the encyclopedia – obvious to a teacher, but oblivious to software.

16. **Elfriede Jelinek**: This biography of the 2004 Nobel Prize laureate for literature is a shake & paste plagiarism from three sources. One was translated by hand from an English-language blog, one is an official book report, and the third a print newspaper article available online. The English blog is no longer available online.

17. **Square dancing**: This paper is almost original, except for one paragraph about the clothing that was taken verbatim from the home page of a club. There exists at least one plagiarism of the text on the pages of another club.

18. **Vikings**: This paper is a highly adapted clause quilt based on the online version of a scholarly journal article about the Vikings. Almost every sentence had some sort of change done to it – word order changed, synonyms used, etc. Only the quote of a rune stone text was left unchanged, which was useful for software-based detection.

19. **Blogs**: This is a structural plagiarism of a PDF about the so-called digital revolution. Sentences and paragraphs were used in ascending order, as

well as the footnotes. Each sentence was manipulated so that it was not identical to the source.

20. **Volleyball**: Two sentences of an otherwise original work about the sport were taken from a web page.

21. **Tibet**: Three sources were used for this shake & paste plagiarism, the Wikipedia, an article in a German daily newspaper, and an article from a weekly computer newspaper. There are a number of sources referenced, but the reference numbering scheme contains gaps – caused by sentences being removed in the middle and the references not being adjusted, something that is glaringly obvious to a human reader.

22. **Le Pont**: This test case was prepared from a French original using Google-Translate. The result was polished to make the German sentences read cleanly, because the sentence structure produced by the automatic translator was quite unintelligible.

23. **Wok**: This test case was prepared by using the Amazon "Search Inside" feature. A cookbook was found with an appropriate page describing a wok, the page was typed up by hand.

24. **Keyboard**: This article about the Dvorak keyboard was prepared as a shake & paste plagiarism from an online article.

25. **Surströmming**: This plagiarism was copied completely from an online article that itself plagiarized the Wikipedia heavily. Then additional, original sentences were added so that an originality quotient of about 20% was given. During the test, however, the source disappeared from the Internet without a trace. We adjusted the scoring to only score hits on the Wikipedia. A second version of this article was prepared in which in one paragraph all of the letters 'e' were replaced with an 'ε', a differently coded letter that looks similar to an 'e' on a quick read-through.

26. **Ajax**: An article from the online journal database of Springer was taken (with permission) as the basis for this copy & paste plagiarism. During the test we discovered that Ciando and Googlebooks also have the entire article indexed - although the link delivered by Google is just to a page for purchasing an electronic copy of the article.

27. **Codfish**: This is a shake & paste plagiarism taken from a German weekly newspaper, an online special edition of another weekly newspaper, and the Wikipedia.

28. **Brantenberg**: This test case about the Norwegian author Gert Brantenberg was translated by hand from an online source in Norwegian. It contains many place names, so it should be discoverable.

29. **Facebook**: Half of this test case is taken as a copy & paste with permission from a student blog, the rest is original.

30. **Smoking ban**: This is an original paper about the smoking ban in public places recently introduced in Germany.

References

Bull, Joanna, Collins, Carol, Coughlin, Elizabeth and Sharp, Dale (2001). Technical Review of Plagiarism Detection Software Report. JISC. http://www.plagiarismadvice.org/documents/resources/Luton_TechnicalReviewofPDS.pdf Accessed 17 April 2009

Gibaldi, Joseph (2003). *MLA Handbook for Writers of Research Papers.* 6th ed. New York: Modern Language Association. Quoted from http://www.mla.org/repview_profethics#two

Howard, Rebecca Moore. (1999*). Standing in the shadow of giants: Plagiarists, authors, collaborators*. Stamford, Conn.: Ablex Pub.

Maurer, Hermann, Kappe, Frank and Zaka, Bilal (2006). *Plagiarism - A Survey.* In: Journal of Universal Computer Science, vol. 12, no. 8 (2006), 1050-1084.

Nordström, Anna and Sjöberg, Susanne (2006a): *Utvärdering av URKUND, ett verktyg für Plagiatkontroll* [Analysis of Urkund, a System for Plagiarism Detection]. Center for Regional Studies (CERUM) Report 16, Umeå University, Sweden http://www.cerum.umu.se/publikationer/pdfs/CRAP_16_06.pdf Accessed 9 March 2009

Nordström, Anna and Sjöberg, Susanne (2006b): *Utvärdering av GenuineText, ett verktyg für Plagiatkontroll* [Analysis of GenuineText, a System for Plagiarism Detection]. Center for Regional Studies (CERUM) Report 17, Umeå University, Sweden http://www.cerum.umu.se/publikationer/pdfs/CRAP_17_06.pdf Accessed 9 March 2009

Scaife, Bryan (2007). *IT Consultancy – Plagiarism Detection Software Report for JISC Plagiarism Advisory Service*. http://www.jiscpas.ac.uk/documents/resources/PDReview-Reportv1_5.pdf Accessed 17 April 2009

Schleimer, Saul, Wilkerson, Daniel S., Aiken, Alex: Winnowing: Local Algorithms for Document Fingerprinting. SIGMOD 2003, June 9-12, 2003, San Diego, CA. Available at http://theory.stanford.edu/~aiken/publications/papers/sigmod03.pdf

Stein, Benno and Meyer zu Eissen, Sven (2007). Intrinsic Plagiarism Analysis with Meta Learning. In: B. Stein, M. Koppel, and E. Stamatatos, Eds., *SIG-IR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07),* Amsterdam, Netherlands, 45–50

Weber-Wulff, Debora and Wohnsdorf, Gabriele (2006). Strategien der Plagiatsbekämpfung [Strategies for Combating Plagiarism]. *Information: Wissenschaft & Praxis*. 57(2), 90-98.

Weber-Wulff, Debora (2004, revised 2007). *Fremde Federn Finden* [Finding False Feathers]. An E-Learning unit. http://plagiat.htw-berlin.de/ff/ Accessed 30 March 2009

Weber-Wulff, Debora (2008). Kollusions-Erkennungs-Systeme [Collusion Detection Systems]. Online results of Plagiarism Detection Software Test 2008. http://plagiat.htw-berlin.de/software/2008/collusion/ Accessed 30 March 2009

Weber-Wulff, Debora (2009). Fremde Federn Finden – Plagiatssysteme im Vergleich. In: *Die Neue Hochschule*. 50(2-3), S. 40-47