Journal of Universal Computer Science, vol. 12, no. 8 (2006), 1050-1084
submitted: 10/8/06, accepted: 25/8/06, appeared: 28/8/06 © J.UCS

Plagiarism - A Survey

Hermann Maurer
(Institute for Information Systems and Computer Media
Graz University of Technology, Austria
hmaurer@iicm.edu)

Frank Kappe
(Institute for Information Systems and Computer Media
Graz University of Technology, Austria
frank.kappe@iicm.edu)

Bilal Zaka
(Institute for Information Systems and Computer Media
Graz University of Technology, Austria
bzaka@iicm.edu)

Abstract: Plagiarism in the sense of "theft of intellectual property" has been around for as long as humans have produced work of art and research. However, easy access to the Web, large databases, and telecommunication in general, has turned plagiarism into a serious problem for publishers, researchers and educational institutions. In this paper, we concentrate on textual plagiarism (as opposed to plagiarism in music, paintings, pictures, maps, technical drawings, etc.). We first discuss the complex general setting, then report on some results of plagiarism detection software and finally draw attention to the fact that any serious investigation in plagiarism turns up rather unexpected side-effects. We believe that this paper is of value to all researchers, educators and students and should be considered as seminal work that hopefully will encourage many still deeper investigations.

Keywords: Plagiarism, cheating, similarity detection, IPR
Categories: K.0, K.3, K.4, K.5

1 Introduction

1.1 Defining Plagiarism

There are many definitions of what constitutes plagiarism, and we will look at some of them in more detail below. However, according to research resources at plagiarism.org, the things that immediately come to mind as description of plagiarism are:

- turning in someone else's work as your own
- copying words or ideas from someone else without giving credit
- failing to put a quotation in quotation marks
- giving incorrect information about the source of a quotation
- changing words but copying the sentence structure of a source without giving credit

- copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not [Plagiarism.org 2006]

The border-line between plagiarism and research is surprisingly murky. After all, advanced research is only possible by “standing on the shoulders” of others, as it is often said. In some areas (such as e.g. literature or law) a scholarly paper may well consist of a conjecture followed by hundreds of quotes from other sources to verify or falsify the thesis. In such case, any attempt to classify something as plagiarized vs. not-plagiarized just based on a count of lines of words that are taken literally from other sources is bound to fail. In other areas (like in a paper in mathematics) it may be necessary to quote standard literature just to make sure that readers have enough background to understand the important part, the proof of a new result whose length may well be below one third of the paper! In other disciplines like engineering or computer science the real value of a contribution may be in the device or algorithm developed (that may not even be explicitly included in the paper) rather than the description of why the device or algorithm is important that may well be spelled out in a number of text books. In summary, we believe that there is no valid definition of even textual plagiarism that is not somewhat domain dependent, complicating the issue tremendously.

A good survey of further ideas about how to define plagiarism, and famous examples of suspected or perpetrated plagiarisms can be found in the Wikipedia[1]. Let us now turn, however, to an attempt to classify various types of plagiarism:

Plagiarism is derived form the Latin word “plagiarius” which means kidnapper. It is defined as “the passing off of another person's work as if it were one's own, by claiming credit for something that was actually done by someone else” [Wikipedia:Plagiarism 2006]. Plagiarism is not always intentional or stealing some things from some one else; it can be unintentional or accidental and may comprise of self stealing. The broader categories of plagiarism include:

- Accidental: due to lack of plagiarism knowledge, and understanding of citation or referencing style being practiced at an institute
- Unintentional: the vastness of available information influences thoughts and the same ideas may come out via spoken or written expressions as one's own
- Intentional: a deliberate act of copying complete or part of some one else’s work without giving proper credit to original creator
- Self plagiarism: using self published work in some other form without referring to original one [Wikipedia:Plagiarism 2006] [Beasley 2006].

There is a long list of plagiarism methods commonly in practise. Some of these methodologies include

- copy-paste: copying word to word textual contents.
- idea plagiarism: using similar concept or opinion which is not common knowledge.

[1] www.wikipedia.com/wiki/plagiarism

- paraphrasing: changing grammar, similar meaning words, re-ordering sentences in original work. Or restating same contents in different words.
- artistic plagiarism: presenting some one else's work using different media, such as text, images, voice or video.
- code plagiarism: using program code, algorithms, classes, or functions without permission or reference.
- forgotten or expired links to resources: addition of quotations or reference marks but failing to provide information or up-to-date links to sources.
- no proper use of quotation marks: failing to identify exact parts of borrowed contents.
- misinformation of references: adding references to incorrect or non existing original sources.
- translated plagiarism: cross language content translation and use without reference to original work.

1.2 Impact

A survey (released in June, 2005) conducted as part of Center of Academic Integrity's Assessment project reveals that 40% of students admitted to engaging in plagiarism as compared to 10% reported in 1999 [CAI 2005]. Another mass survey conducted by a Rutgers University professor in 2003 reports 38% of students involved in online plagiarism [Rutgers 2003]. These alarming figures show a gradual increase. The new generation is more aware of technology than ever before. Plagiarism now is not confined to mere cut and paste; synonymising and translation technologies are giving a new dimension to plagiarism.

Plagiarism is considered to be a most serious scholastic misconduct; academia everywhere is undertaking efforts to educate the students and teachers, by offering guides and tutorials to explain types of plagiarism and how to avoid it.

This growing awareness is forcing universities and institutes all around to help students and faculty understand the meaning of academic integrity, plagiarism and its consequences. Since plagiarism is often connected with the failure to reference or quote properly, many institutions suggest following one of the recognized writing styles as proposed by major publishing companies like Springer, or by using well defined citation styles like: Modern Language Association (MLA) style[2], Chicago Manual of style[3], or American Psychological Association (APA) style[4].

2 Response of academic institutions

Although plagiarism is reasonably well defined and explained in many forums, the penalty for cases detected varies from case to case and institution to institution,

Many universities in the United States have well defined policies to classify and deal with academic misconduct. Rules and information regarding it are made available to students during the enrolment process, via information brochures and the

[2] http://www.mla.org/style

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

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

university web sites. Academic dishonesty can be dealt with at teacher-student level or institute-student level. The penalties that can be imposed by teachers include written or verbal warning, failing or lower grades and extra assignments. The institutional case handling involves hearing and investigation by an appropriate committee, with the accused aware and part of whole process. The institutional level punishments may include official censure, academic integrity training exercises, social work, transcript notation, suspension, expulsion, revocation of degree or certificate and possibly even referral of the case to legal authorities. To be specific, we have collected a number of examples:

Stanford University: Stanford University provides its students with a well defined academic misconduct policy (Honor Code, in force since 1921) and a good collection of copyright and fair use resources [Stanford Copyright 2006]. According to an article in the Stanford daily, the Stanford's office of judicial affairs saw 126 percent increase in honor code violation from 1998 to 2001. This precipitated the increasing usage of anti plagiarism software among instructors at individual levels [Stanford Daily 2003]. As per the Stanford Honor Code "The standard penalty for a first offence includes a one-quarter suspension from the University and 40 hours of community service. In addition, most faculty members issue a "No Pass" or "No Credit" for the course in which the violation occurred. The standard penalty for multiple violations (e.g. cheating more than once in the same course) is a three-quarter suspension and 40 or more hours of community service" [Stanford Honorcode 1921]. Some sample cases and sanctions are available at, University's Judicial Affairs website[5].

Yale University: Yale College Executive Committee Yearly Chair Reports [Yale 2005] indicate that the committee had to deal with a sizeable number of plagiarism cases every year. They show great concern about increase in web plagiarism. There are discussions about its causes and possible preventive measures mentioned in the reports. Punishments vary from case to case starting from reprimands, probations and extending to suspension. Despite clear academic misconduct policies there were cases of accidental or mistaken plagiarism, which suggests that there is a need of more effective ways of communicating details to students. Teachers are encouraged to explain plagiarism, citation rules and writing styles to students.

U.C. Berkeley: This university also has clear policies and preventive procedures against academic dishonesty. Instructors are encouraged to resolve the matter personally and issue academic sanctions; in case an accused person does not agree with allegations or sanctions, the matter is handed over to student judicial affairs for further investigations and resolution. Sanctions at U.C. Berkeley for plagiarism are warning/censure, community service, letters of apology, counselling, additional coursework, disciplinary probation, suspension, dismissal, and restitution [Berkeley 2006].

Massachusetts Institute of Technology: MIT has well defined policies and procedures for handling academic misconduct [MIT policies 2006]. Teachers are encouraged to educate students about permissible academic conduct. MIT's online writing and communication center [MIT Writing 2006] provides a platform to improve writing abilities and explains various aspects of plagiarism. According to a report available at MIT News Office portal, usually the discipline committee has to

[5] http://www.stanford.edu/dept/vpsa/judicialaffairs/students/pdf/plagiarism.cases.pdf

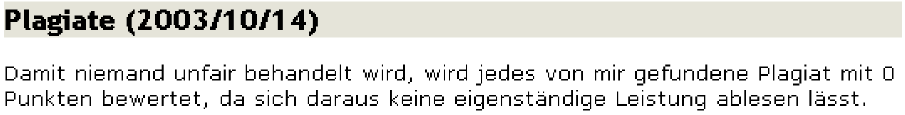
handle 12 to 15 cases annually with a tendency of increase in number of cases in recent years [MIT News 2003]. The penalties follow a similar trend as in other universities, starting from reduced grades, warning letters, redo of exam or assignment and in extreme cases with recommendation of the discipline committee, suspension or expulsion.

In Europe, UK is probably ahead of the other countries by taking collective measures against plagiarism. Most of the universities have online guides and tutorials available for students and researchers, helping them to understand academic integrity and improving writing skills. The higher education community in UK took a collective measure by forming a plagiarism advisory service [JISC 2006] giving all UK institutes access to an online plagiarism detection service.

University of Cambridge: At Cambridge, suspected plagiarism cases involve separate academic and disciplinary elements. Examiners are asked to evaluate and make recommendations about suspected work but they can not impose any penalty. The proctors, university advocates and courts decide about the sanctions in light of recommendations by examiners and investigations [Cambridge 2006].

Oxford University: According to the University Gazette March 2005, six plagiarism related cases were dealt with during the previous term. “Three cases were dealt with by the Court of Summary Jurisdiction; in each, the examiners were instructed to disregard the plagiarised work and the candidates were permitted to resubmit (with a marks penalty in one case). The Disciplinary Court dealt with 2 plagiarism cases; in one case the examiners were instructed to disregard the plagiarised work. The candidate was failed in a previously completed M.St. examination but permitted to retake the examination, and if the examiners are satisfied, permitted to re-enter the degree for M.Phil. In the second case, a candidate had previously been convicted of plagiarism by the Court of Summary Jurisdiction. He/she was permitted to submit new work and some of this was subsequently found to contain plagiarised material. A charge of attempting to cheat or act dishonestly was dismissed, but the candidate was nevertheless failed in the BCL examination. Following a proctorial investigation, and taking into consideration certain mitigating factors, the Examiners were instructed to disregard a candidate's original M.Phil submission. He/she was given permission to submit replacement work to be determined by the Examiners” [Oxford Gazette 2005].

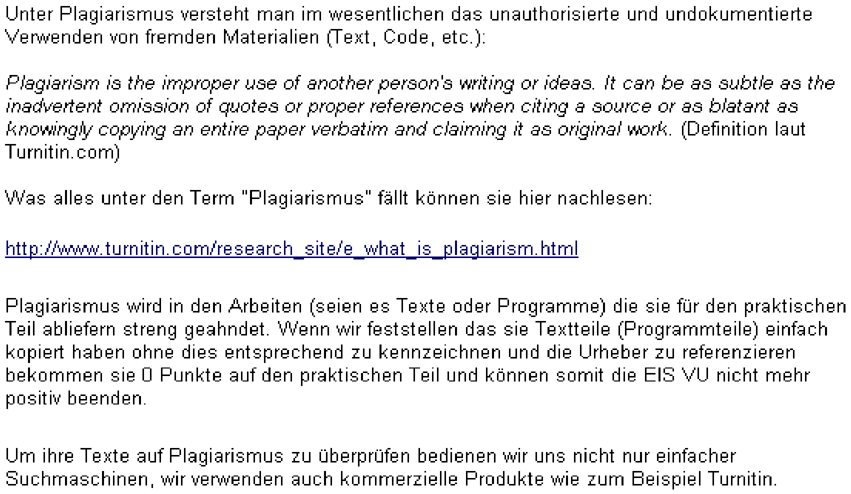
Elsewhere in Europe, there is also a growing concern and individual efforts have been started by teachers at departmental levels to educate researchers and students about plagiarism. At Graz University of Technology, Austria, a Commission for Scientific Integrity and Ethics defines guiding principles to deal with cases of plagiarism. A catalogue of possible academic, civil and criminal consequences will be ready by end of 2006. Instructors at various institutes of the university started adding information and warnings about plagiarism some time ago, e.g. figure 1, 2 & 3 show responses to plagiarism cases on course websites at various institutes of Technical University Graz.



Plagiate (2003/10/14)

Damit niemand unfair behandelt wird, wird jedes von mir gefundene Plagiat mit 0 Punkten bewertet, da sich daraus keine eigenständige Leistung ablesen lässt.

Figure 1: Taken from course information page by Harald Krottmaier, Institute of Computer Graphics and Knowledge Visualization, TU Graz



Unter Plagiarismus versteht man im wesentlichen das unauthorisierte und undokumentierte Verwenden von fremden Materialien (Text, Code, etc.):

Plagiarism is the improper use of another person's writing or ideas. It can be as subtle as the inadvertent omission of quotes or proper references when citing a source or as blatant as knowingly copying an entire paper verbatim and claiming it as original work. (Definition laut Turnitin.com)

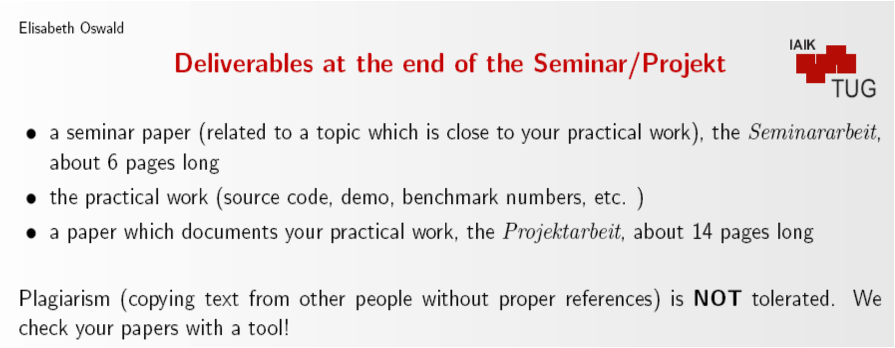
Was alles unter den Term "Plagiarismus" fällt können sie hier nachlesen:

http://www.turnitin.com/research_site/e_what_is_plagiarism.html

Plagiarismus wird in den Arbeiten (seien es Texte oder Programme) die sie für den praktischen Teil abliefern streng geahndet. Wenn wir feststellen das sie Textteile (Programmteile) einfach kopiert haben ohne dies entsprechend zu kennzeichnen und die Urheber zu referenzieren bekommen sie 0 Punkte auf den praktischen Teil und können somit die EIS VU nicht mehr positiv beenden.

Um ihre Texte auf Plagiarismus zu überprüfen bedienen wir uns nicht nur einfacher Suchmaschinen, wir verwenden auch kommerzielle Produkte wie zum Beispiel Turnitin.

Figure 2: Taken from teaching information page, Institute for Applied Information Processing and Communications (IAIK) TU Graz



Elisabeth Oswald

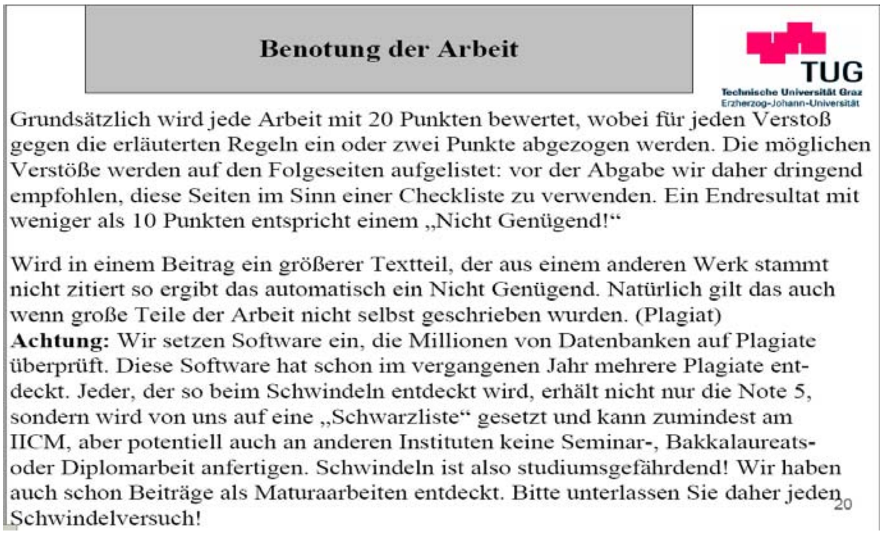
Deliverables at the end of the Seminar/Projekt

- a seminar paper (related to a topic which is close to your practical work), the *Seminararbeit*, about 6 pages long
- the practical work (source code, demo, benchmark numbers, etc.)
- a paper which documents your practical work, the *Projektarbeit*, about 14 pages long

Plagiarism (copying text from other people without proper references) is **NOT** tolerated. We check your papers with a tool!

Figure 3: Taken from Bachelor seminar project contents by Elisabeth Oswald, Institute for applied information processing and communications (IAIK), TU Graz

In the information on a course (shown in Figure 4) on how to write scientific contributions the first author of this paper states explicitly, that plagiarism will result in expulsion and therefore failing the course.



Benotung der Arbeit

TUG
Technische Universität Graz
Erzherzog-Johann-Universität

Grundsätzlich wird jede Arbeit mit 20 Punkten bewertet, wobei für jeden Verstoß gegen die erläuterten Regeln ein oder zwei Punkte abgezogen werden. Die möglichen Verstöße werden auf den Folgeseiten aufgelistet: vor der Abgabe wir daher dringend empfohlen, diese Seiten im Sinn einer Checkliste zu verwenden. Ein Endresultat mit weniger als 10 Punkten entspricht einem „Nicht Genügend!"

Wird in einem Beitrag ein größerer Textteil, der aus einem anderen Werk stammt nicht zitiert so ergibt das automatisch ein Nicht Genügend. Natürlich gilt das auch wenn große Teile der Arbeit nicht selbst geschrieben wurden. (Plagiat)
Achtung: Wir setzen Software ein, die Millionen von Datenbanken auf Plagiate überprüft. Diese Software hat schon im vergangenen Jahr mehrere Plagiate entdeckt. Jeder, der so beim Schwindeln entdeckt wird, erhält nicht nur die Note 5, sondern wird von uns auf eine „Schwarzliste" gesetzt und kann zumindest am IICM, aber potentiell auch an anderen Instituten keine Seminar-, Bakkalaureats- oder Diplomarbeit anfertigen. Schwindeln ist also studiumsgefährdend! Wir haben auch schon Beiträge als Maturaarbeiten entdeckt. Bitte unterlassen Sie daher jeden Schwindelversuch!

Figure 4: Taken from a course presentation of author, Institute for Information Systems and Computer Media (IICM) TU Graz

The problem of academic misconduct and plagiarism also exists in universities of developing countries. The situation there has different dimensions where language problems and lack of guidance create further complications. The concept of plagiarism is generally less known and very little institutional efforts are made to educate students and staff about the plagiarism. However, this is changing rapidly, because of high profile incidents causing an alarming situation and introduction of strict measures to address the problem. The Higher Education Commission of Pakistan issued detailed guidelines and zero tolerance policy against plagiarism to all universities of the country [HEC Press 2006]. This was initiated due to the discovery of high profile plagiarism cases at Pakistani universities which lead to the resignation of involved faculty members and expulsion of students.

At some places the fight against plagiarism is more about grooming the writers with organized guidelines, tutorials and honor codes; in other cases it is more about detection and punishment. However, a well balanced combination of both is the most effective approach.

3 Detecting plagiarism

Plagiarism detection methods can be broadly categorized into three main categories; the most common approach is by comparing the document against a body of documents, basically on a word by word basis where documents may reside locally or not. The other two approaches are not exploited as much, yet can also be surprisingly successful. One is by taking a characteristic paragraph and just doing a search with a good search engine like Google. And the other is by trying to do style analysis; in this case either just within the document at issue or performing writing style comparison with documents previously written by the same author. This is usually called stylometry.

Let us look at the three approaches in more detail:

3.1 Document source comparison:

This approach can be further divided into two categories; one that operates locally on the client computer and does analysis on local databases of documents or performs internet searches, the other is server based technology where the user uploads the document and the detection processes take place remotely. The most commonly used techniques in current document source comparison involve word stemming or fingerprinting. This is an approach introduced by Manber [Manber 1994] where moderately sized strings (Fingerprints) from a document are compared for similarities with preprocessed indexes from other documents. The result gives a similarity approximation among documents being checked. Figure 5 shows a generic structure of document source comparison based plagiarism detection system.

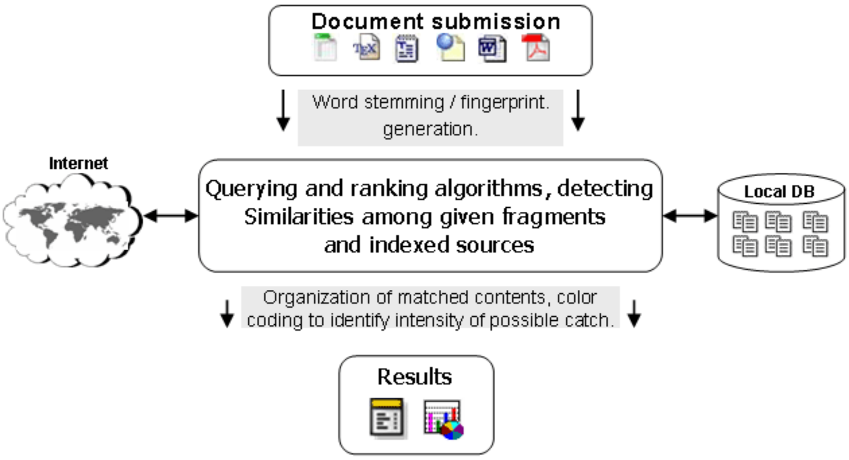

Document submission
Word stemming / fingerprint. generation.
Internet
Querying and ranking algorithms, detecting Similarities among given fragments and indexed sources
Local DB
Organization of matched contents, color coding to identify intensity of possible catch.
Results

Figure 5: Plagiarism detection with document source comparison

The core finger printing idea has been modified and enhanced by various researchers to improve similarity detection. Many current commercial plagiarism detection service providers claim to have proprietary fingerprinting and comparison mechanisms. The comparison can be local or it can be across the internet. Some services utilize the potentials of available search engines. Many such tools use Google Search API[6] providing querying capabilities to billions of web resources. Recent steps

[6] http://www.google.com/apis/

taken by Google to index the full text of some of the world's leading research libraries [Band 2006], and its well known searching and ranking algorithm makes it an ideal choice not only for open source and free tools but is also used by many commercial service providers and applications. The more popular commercial and server based approaches claim to use their own search and querying techniques over more extensively indexed internet documents, proprietary databases, password protected document archives and paper mills. (We will mention more on those in the next paragraph.). The detection services or tools usually represent the similarity findings in a report format, by identifying matches and their sources. The findings are then utilized by users of the service to determine whether the writing under question is actually plagiarized or whether there are other reasons for match detection. We come back to this later in the paper.

Returning to the issue of paper mills, this term refers to "website where students can download essays, either free or for a service charge. Online paper mills usually contain a large, searchable database of essays. Most paper mills today offer customized writing services, usually charging by the page. Some sites now even offer ready-made college application essays from applicants who have been accepted" [Wikipedia:papermill 2006].

There are a number of web sites that even list paper mills![7]

3.2 Manual search of characteristic phrases

Using this approach the instructor or examiner selects some phrases or sentences representing core concepts of a paper. These phrases are then searched across the internet using single or multiple search engines. Let us explain this by means of an example.

Suppose we detect the following sentence in a student's essay

"Let us call them eAssistants. They will be not much bigger than a credit card, with a fast processor, gigabytes of internal memory, a combination of mobile-phone, computer, camera"

Since eAssistant is an uncommon term, it makes sense to input the term into a Google query. Indeed if this done the query produces:

"(Maurer H., Oliver R.) The Future of PCs and Implications on Society -
Let us call them eAssistants. They will be not much bigger than a credit card, with a fast processor, gigabytes of internal memory, a combination of ...
www.jucs.org/jucs_9_4/the_future_of_pcs/Maurer_H_2.html - 34k -"

This proves that without further tools the student has used part of a paper published in the Journal of Universal Computer Science[8]. It is clear that this approach is labor intensive; hence it is obvious that some automation will make sense, as is done in SNITCH [Niezgoda & Way 2006].

[7] see http://www.coastal.edu/library/presentations/mills2.html

[8] see http://www.jucs.org

3.3 Stylometry

Stylometric analysis is based on individual and unique writing styles of various persons. The disputed writing can be evaluated using different factors within the same writing. Or it can be cross compared with previous writings by the same author. The detection of plagiarism within the document domain or without any external reference is well described as “intrinsic plagiarism detection” by Eissen and Stein [Eissen & Stein 2006]. This approach requires well defined quantification of linguistic features which can be used to determine inconsistencies within a document. According to Eissen and Stein “Most stylometric features fall in one of the following five categories: (i) text statistics, which operate at the character level, (ii) syntactic features, which measure writing style at the sentence-level, (iii) part-of-speech features to quantify the use of word classes, (iv) closed-class word sets to count special words, and (v) structural features, which reflect text organization.” [Eissen & Stein 2006] The paper quoted, adds a new quantification statistic “the averaged word frequency class” and presents experiments showing its effectiveness. As an example of simple generic intrinsic plagiarism analysis let us take the following paragraph.

*“**Our** goal is to identify files that came from **the same source** or contain parts that came from **the same source**. **We** say that two files are similar if they contain a significant number of common substrings that are not too small. **We** would like to find enough common substrings to rule out chance, without requiring too many so that **we** can detect similarity even if significant parts of the files are different. However, **my** interest in plagiarism lies within academic institutions, so the document domain will be local research articles. The limited scope of domain will make it easier to determine if it is **same source** or not.”*

A careful reading reveals the following inconsistencies:

- There is a change in pronoun from “our/we” to “my”
- The writer used the article “the” with “same source” in two sentences and missed the article in another.

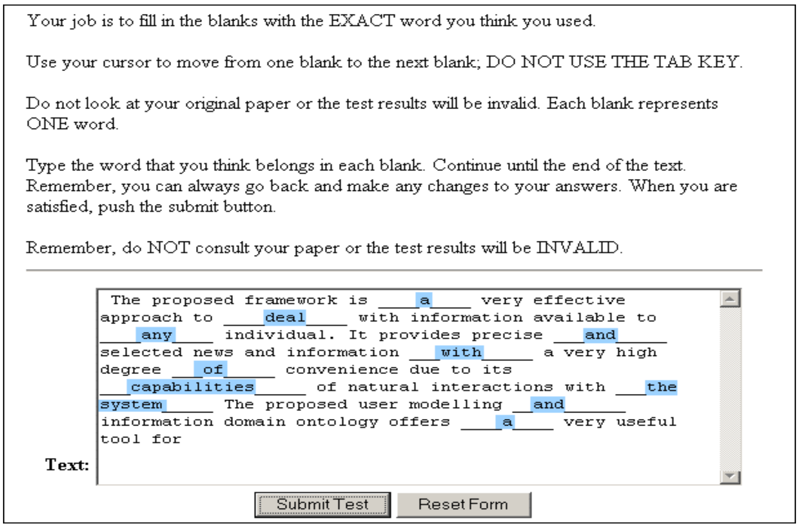
The bold words show the inconsistency and thus exhibit the possibility of plagiarism, where the writer took text from some source not matching the overall writing style. This approach can be hard to use in case of collaboratively written text where multiple writers are contributing to a single source.

Cross comparisons include a check on change of vocabulary, common spelling mistakes, the use of punctuation and common structural features such as word counts, sentence length distributions etc. (see example of using structural features to detect similarity in “Advanced Techniques” section). In order to further explain stylometry and another approach, we look at a service by Glatt [Glatt 2006], which uses Wilson Taylor's (1953) cloze procedure. In this approach every fifth word in a suspected document is removed and the writer is asked to fill the missing spaces. The number of correct responses and answering time is used to calculate plagiarism probability. For example the examiner suspects that the following paragraph is plagiarized.

“The proposed framework is a very effective approach to deal with information available to any individual. It provides precise and selected news and information

with a very high degree of convenience due to its capabilities of natural interactions with users. The proposed user modelling and information domain ontology offers a very useful tool for browsing the information repository, keeping the private and public aspects of information retrieval separate. Work is underway to develop and integrate seed resource knowledge structures forming basis of news ontology and user models using.....”

The writer is asked to take a test and fill in periodic blank spaces in text to verify the claim of authorship. A sample test based on above paragraph is shown in figure 6.



Your job is to fill in the blanks with the EXACT word you think you used.

Use your cursor to move from one blank to the next blank; DO NOT USE THE TAB KEY.

Do not look at your original paper or the test results will be invalid. Each blank represents ONE word.

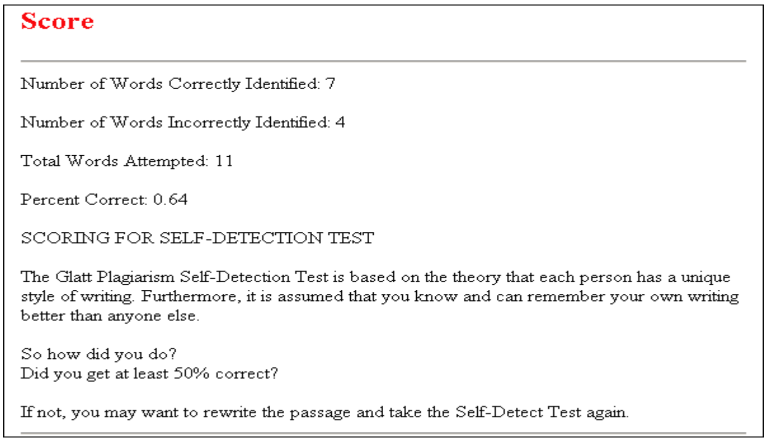
Type the word that you think belongs in each blank. Continue until the end of the text. Remember, you can always go back and make any changes to your answers. When you are satisfied, push the submit button.

Remember, do NOT consult your paper or the test results will be INVALID.

Text:

```
 The proposed framework is ____a____ very effective
approach to ____deal____ with information available to
____any____ individual. It provides precise ___and_____
selected news and information ___with_____ a very high
degree ___of_____ convenience due to its
___capabilities_____ of natural interactions with ___the
system_____ The proposed user modelling __and______
information domain ontology offers ____a____ very useful
tool for
```

Figure 6: Stylometric test, Glatt Plagiarism Self-Detection Program



Score

Number of Words Correctly Identified: 7

Number of Words Incorrectly Identified: 4

Total Words Attempted: 11

Percent Correct: 0.64

SCORING FOR SELF-DETECTION TEST

The Glatt Plagiarism Self-Detection Test is based on the theory that each person has a unique style of writing. Furthermore, it is assumed that you know and can remember your own writing better than anyone else.

So how did you do?
Did you get at least 50% correct?

If not, you may want to rewrite the passage and take the Self-Detect Test again.

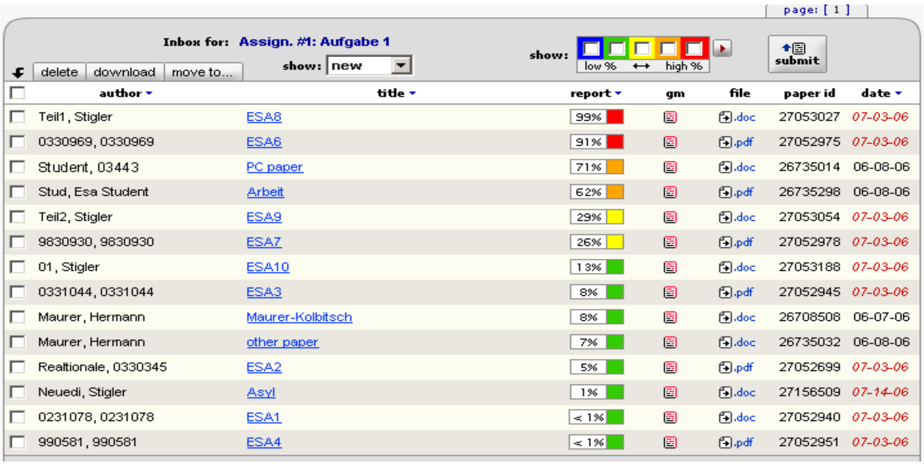
Figure 7: Stylometric test results

The percentage of correct answers can be used to determine if the writing is from the same person or not. The result of the mentioned test is shown in figure 7. This approach is not always feasible in academic environment where large numbers of documents are needed to be processed, but it provides a very effective secondary layer of detection to confirm and verify the results.

4 Available tools

Several applications and services exist to help academia detect intellectual dishonesty. We have selected some of these tools which are currently particularly popular and describe their main features in what follows.

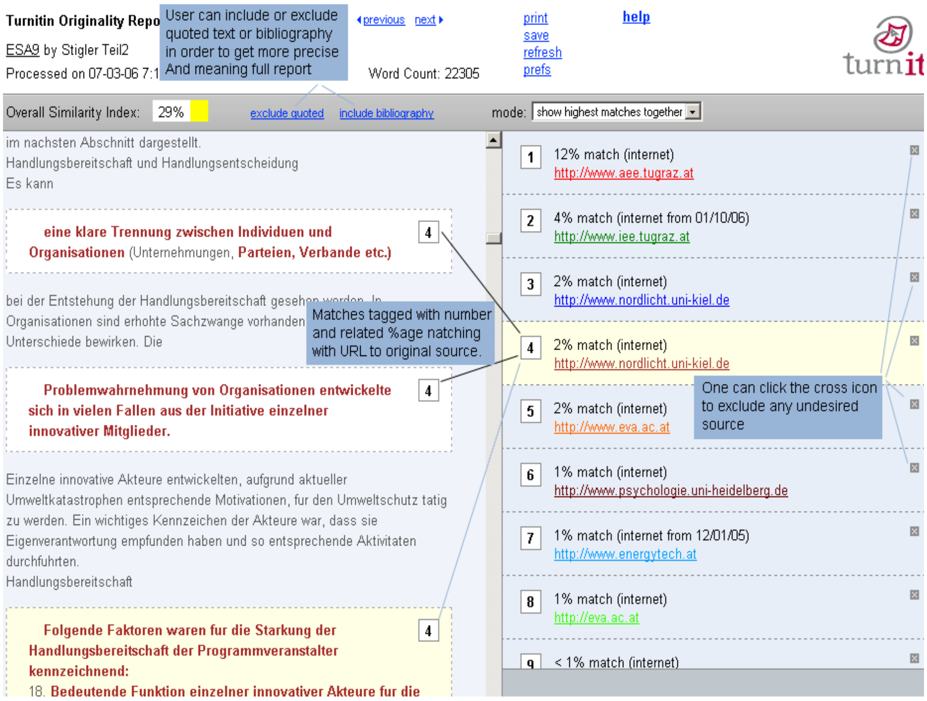
Turnitin: This is a product from iParadigms [iParadigm 2006]. It is a web based service. Detection and processing is done remotely. The user uploads the suspected document to the system database. The system creates a complete fingerprint of the document and stores it. Proprietary algorithms are used to query the three main sources: one is the current and extensively indexed archive of Internet with approximately 4.5 billion pages, books and journals in the ProQuest™ database; and 10 million documents already submitted to the Turnitin database.



□	author	title	report	gm	file	paper id	date
□	Teil1 , Stigler	ESA8	99% 🟥		.doc	27053027	07-03-06
□	0330969, 0330969	ESA6	91% 🟥		.pdf	27052975	07-03-06
□	Student, 03443	PC paper	71% 🟧		.doc	26735014	06-08-06
□	Stud, Esa Student	Arbeit	62% 🟧		.pdf	26735298	06-08-06
□	Teil2, Stigler	ESA9	29% 🟨		.doc	27053054	07-03-06
□	9830930, 9830930	ESA7	26% 🟨		.pdf	27052978	07-03-06
□	01 , Stigler	ESA10	13% 🟩		.doc	27053188	07-03-06
□	0331044, 0331044	ESA3	8% 🟩		.pdf	27052945	07-03-06
□	Maurer, Hermann	Maurer-Kolbitsch	8% 🟩		.doc	26708508	06-07-06
□	Maurer, Hermann	other paper	7% 🟩		.doc	26735032	06-08-06
□	Realtionale, 0330345	ESA2	5% 🟩		.pdf	27052699	07-03-06
□	Neuedi, Stigler	Asyl	1% 🟩		.doc	27156509	07-14-06
□	0231078, 0231078	ESA1	< 1% 🟩		.doc	27052940	07-03-06
□	990581 , 990581	ESA4	< 1% 🟩		.pdf	27052951	07-03-06

Figure 8: Turnitin, Instructor view of assignment inbox

Turnitin offers different account types. They include consortium, institute, department and individual instructor. The former account type can create later mentioned accounts and have management capabilities. At instructor account level, teachers can create classes and generate class enrolment passwords. Such passwords are distributed among students when joining the class and for the submission of assignments. Figure 8 and 9 gives an idea of the system's user-interface.



Turnitin Originality Repo

ESA9 by Stigler Teil2

Processed on 07-03-06 7:1

User can include or exclude quoted text or bibliography in order to get more precise And meaning full report

◂previous next▸

Word Count: 22305

print
save
refresh
prefs

help

Overall Similarity Index: 29%

exclude quoted include bibliography

mode: show highest matches together

im nachsten Abschnitt dargestellt.
Handlungsbereitschaft und Handlungsentscheidung
Es kann

eine klare Trennung zwischen Individuen und Organisationen (Unternehmungen, **Parteien, Verbande etc.)** 4

bei der Entstehung der Handlungsbereitschaft gesehen werden. In Organisationen sind erhohte Sachzwange vorhanden Unterschiede bewirken. Die

Matches tagged with number and related %age natching with URL to original source.

Problemwahrnehmung von Organisationen entwickelte sich in vielen Fallen aus der Initiative einzelner innovativer Mitglieder. 4

Einzelne innovative Akteure entwickelten, aufgrund aktueller Umweltkatastrophen entsprechende Motivationen, fur den Umweltschutz tatig zu werden. Ein wichtiges Kennzeichen der Akteure war, dass sie Eigenverantwortung empfunden haben und so entsprechende Aktivitaten durchfuhrten.
Handlungsbereitschaft

Folgende Faktoren waren fur die Starkung der Handlungsbereitschaft der Programmveranstalter kennzeichnend: 4
18. **Bedeutende Funktion einzelner innovativer Akteure fur die**

1 12% match (internet) http://www.aee.tugraz.at ☒

2 4% match (internet from 01/10/06) http://www.iee.tugraz.at ☒

3 2% match (internet) http://www.nordlicht.uni-kiel.de ☒

4 2% match (internet) http://www.nordlicht.uni-kiel.de ☒

5 2% match (internet) http://www.eva.ac.at ☒

One can click the cross icon to exclude any undesired source

6 1% match (internet) http://www.psychologie.uni-heidelberg.de ☒

7 1% match (internet from 12/01/05) http://www.energytech.at ☒

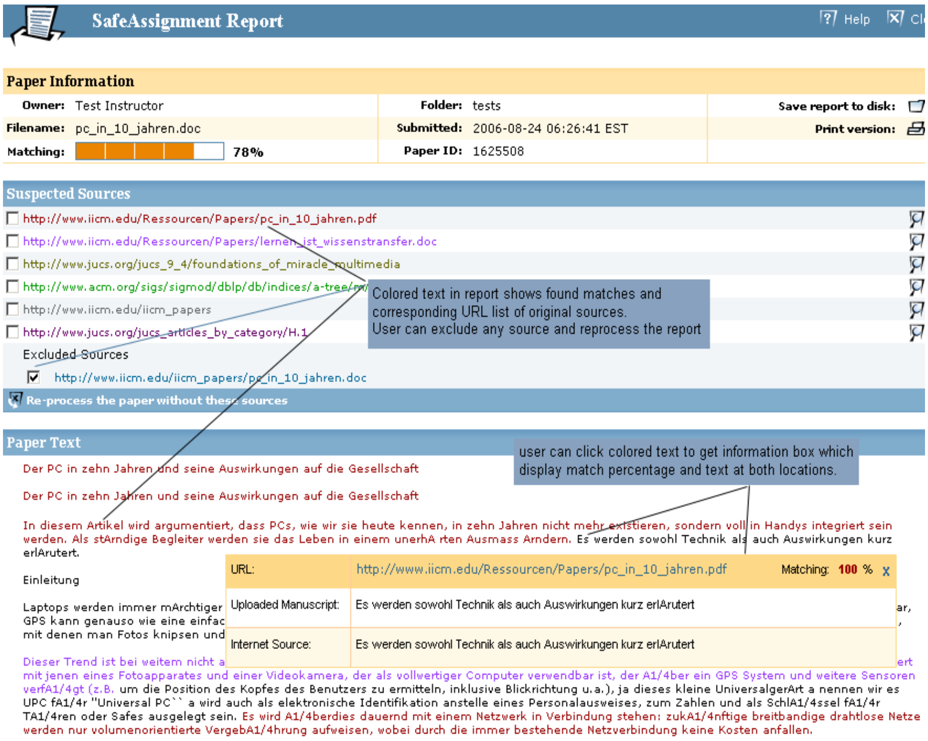
8 1% match (internet) http://eva.ac.at ☒

9 < 1% match (internet) ☒

Figure 9: Turnitin, originality report of a submission

The system generates the originality report within some minutes of submission. The report contains all the matches detected and links to original sources with color codes describing the intensity of plagiarism [Turnitin tour 2006]. It is however not a final statement of plagiarism. A higher percentage of similarities found do not necessarily mean that it actually is a case of plagiarism (for further explanation see Section 3). One has to interpret each identified match to deduce whether it is a false alarm or actually needs attention. This service is used by all UK institutes via the Joint Information Systems Committee (JISC) plagiarism Advisory Program [JISC 2006].

SafeAssignment: This web based service by Mydropbox, claims to search an index of 8 billion internet documents, ProQuest™, FindArticles™ database by LookSmart™ and other major scholastic databases. The system also searches 300,000 documents that are known to be offered by Paper Mills. SafeAssignment also utilizes proprietary archives of institutional partners. Password protected and zipped archives can be indexed on demand. This product keeps fingerprints of the submitted papers in separate databases belonging to the account owner institute in order to avoid any legal or copy right problems. The service uses proprietary searching and ranking algorithms for match detection of fingerprints with its resources. The plagiarism detection result is presented to the user after a couple of minutes of submission, i.e. is similar in this respect with previously mentioned products [Mydropbox 2006]. Figure 10 displays report of a processed paper.



SafeAssignment Report

Paper Information

Owner:	Test Instructor	Folder:	tests	Save report to disk:	
Filename:	pc_in_10_jahren.doc	Submitted:	2006-08-24 06:26:41 EST	Print version:	
Matching:	78%	Paper ID:	1625508		

Suspected Sources

- □ http://www.iicm.edu/Ressourcen/Papers/pc_in_10_jahren.pdf
- □ http://www.iicm.edu/Ressourcen/Papers/lernen_ist_wissenstransfer.doc
- □ http://www.jucs.org/jucs_9_4/foundations_of_miracle_multimedia
- □ http://www.acm.org/sigs/sigmod/dblp/db/indices/a-tree/m/
- □ http://www.iicm.edu/iicm_papers
- □ http://www.jucs.org/jucs_articles_by_category/H.1

Excluded Sources

☑ http://www.iicm.edu/iicm_papers/pc_in_10_jahren.doc

Re-process the paper without these sources

Colored text in report shows found matches and corresponding URL list of original sources.
User can exclude any source and reprocess the report

Paper Text

user can click colored text to get information box which display match percentage and text at both locations.

Der PC in zehn Jahren und seine Auswirkungen auf die Gesellschaft

Der PC in zehn Jahren und seine Auswirkungen auf die Gesellschaft

In diesem Artikel wird argumentiert, dass PCs, wie wir sie heute kennen, in zehn Jahren nicht mehr existieren, sondern voll in Handys integriert sein werden. Als stArndige Begleiter werden sie das Leben in einem unerhA rten Ausmass Arndern. Es werden sowohl Technik als auch Auswirkungen kurz erlArutert.

Einleitung

Laptops werden immer mArchtiger ar,
GPS kann genauso wie eine einfac ,
mit denen man Fotos knipsen und

URL:	http://www.iicm.edu/Ressourcen/Papers/pc_in_10_jahren.pdf Matching: 100 % x
Uploaded Manuscript:	Es werden sowohl Technik als auch Auswirkungen kurz erlArutert
Internet Source:	Es werden sowohl Technik als auch Auswirkungen kurz erlArutert

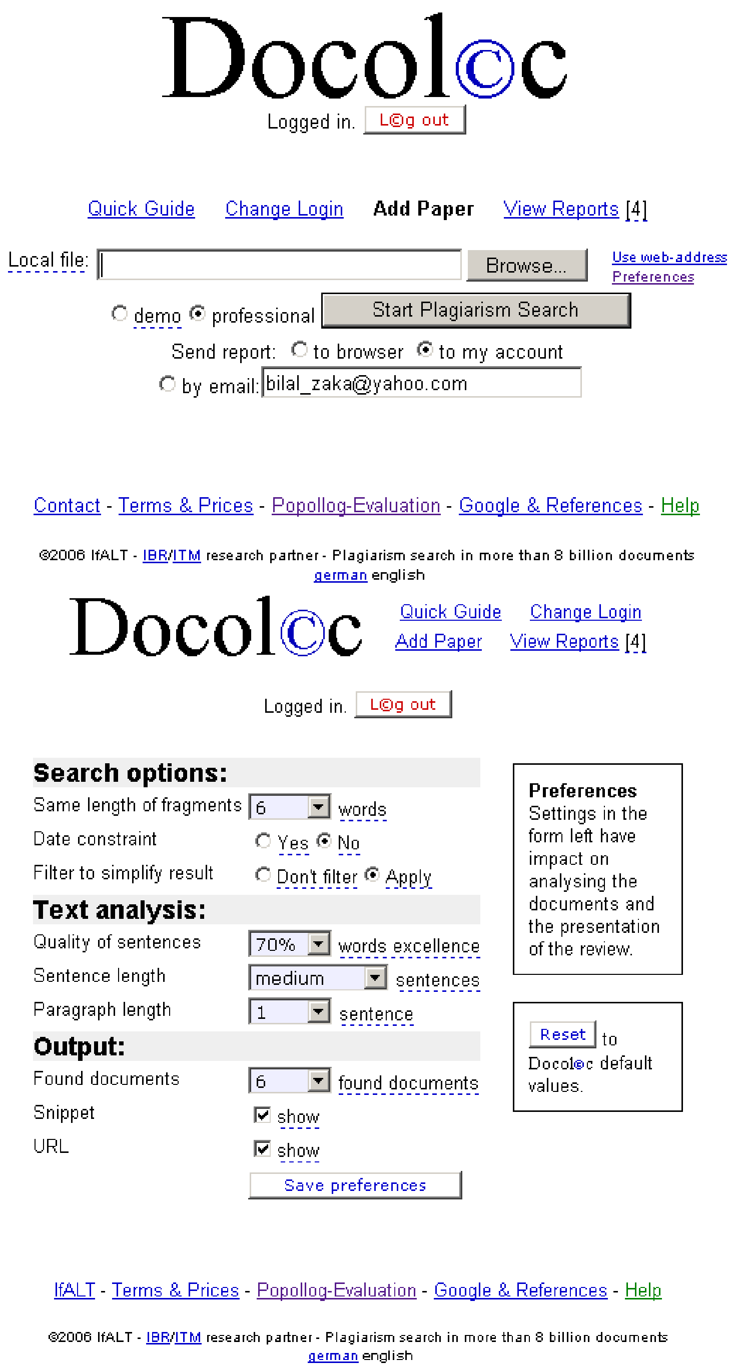
Dieser Trend ist bei weitem nicht a ert mit jenen eines Fotoapparates und einer Videokamera, der als vollwertiger Computer verwendbar ist, der A1/4ber ein GPS System und weitere Sensoren verfA1/4gt (z.B. um die Position des Kopfes des Benutzers zu ermitteln, inklusive Blickrichtung u.a.), ja dieses kleine UniversalgerArt a nennen wir es UPC fA1/4r "Universal PC`` a wird auch als elektronische Identifikation anstelle eines Personalausweises, zum Zahlen und als SchlA1/4ssel fA1/4r TA1/4ren oder Safes ausgelegt sein. Es wird A1/4berdies dauernd mit einem Netzwerk in Verbindung stehen: zukA1/4nftige breitbandige drahtlose Netze werden nur volumenorientierte VergebA1/4hrung aufweisen, wobei durch die immer bestehende Netzverbindung keine Kosten anfallen.

Figure 10: Mydropbox, paper information report

Mydropbox products integrates with other learning management systems (Blackboard ®, WebCT) to extend plagiarism detection capabilities in existing systems running at institutes.

Docol©c: A web based service offered by Institut für Angewandte Lerntechnologien(IFALT)[9]. This service utilizes the searching and ranking capabilities of the Google API. The user of the service uploads the document that needs to be evaluated to a server. The software provides a simple console to set fingerprint (search fragments) size, date constraints, filtering and other report related options. The analysis report is sent to the browser or user's email identifying the matched fragments and internet sources. Figures 11 and 12 show different consoles and detection report by service.

[9] http://www.ifalt.com/



Docol©c

Logged in. L©g out

Quick Guide Change Login **Add Paper** View Reports [4]

Local file: [ ] Browse... Use web-address Preferences

○ demo ● professional Start Plagiarism Search

Send report: ○ to browser ● to my account

○ by email: bilal_zaka@yahoo.com

Contact - Terms & Prices - Popollog-Evaluation - Google & References - Help

©2006 IfALT - IBR/ITM research partner - Plagiarism search in more than 8 billion documents
german english

Docol©c Quick Guide Change Login Add Paper View Reports [4]

Logged in. L©g out

Search options:

Same length of fragments	6 words
Date constraint	○ Yes ● No
Filter to simplify result	○ Don't filter ● Apply

Text analysis:

Quality of sentences	70% words excellence
Sentence length	medium sentences
Paragraph length	1 sentence

Output:

Found documents	6 found documents
Snippet	☑ show
URL	☑ show

Save preferences

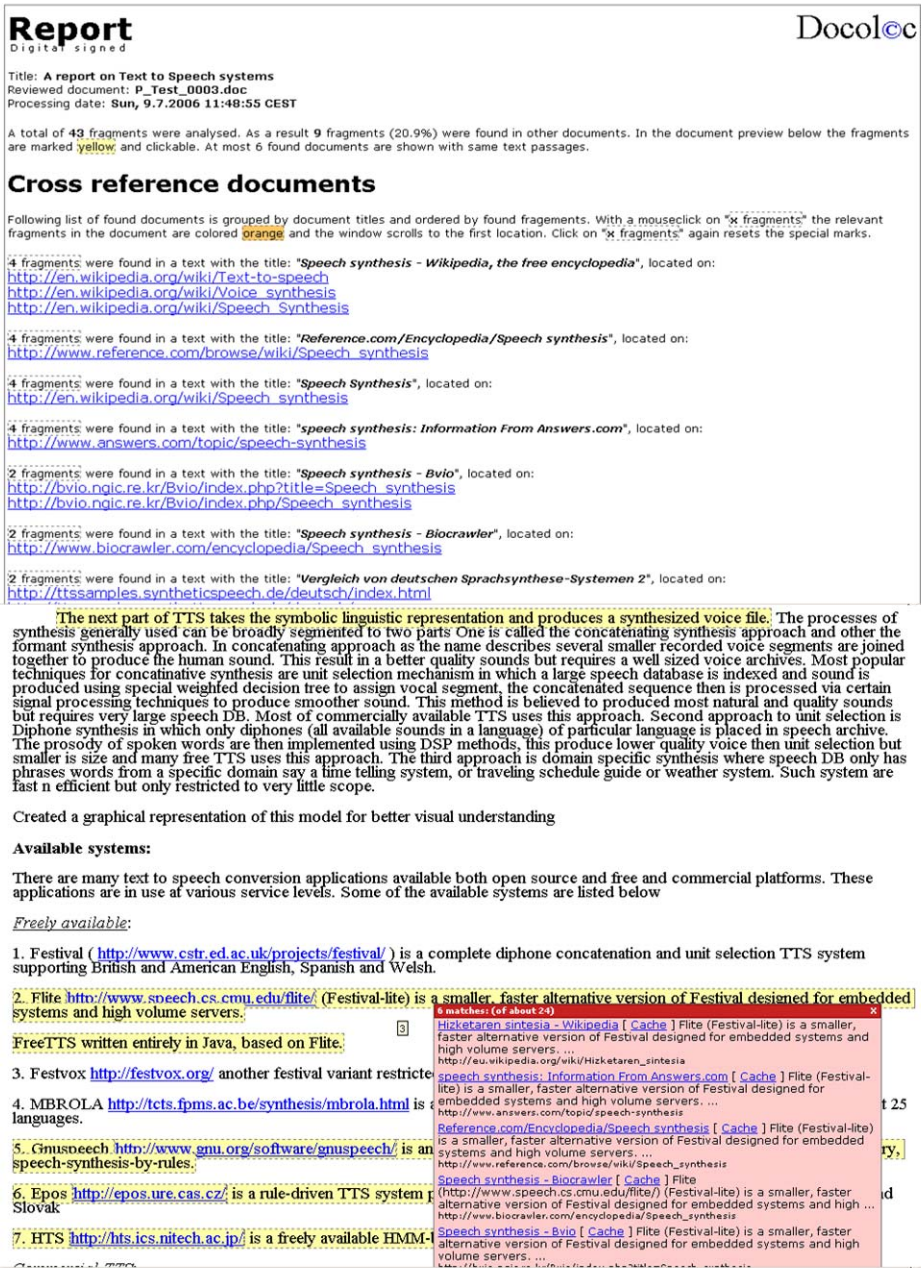
Preferences Settings in the form left have impact on analysing the documents and the presentation of the review.

Reset to Docol©c default values.

IfALT - Terms & Prices - Popollog-Evaluation - Google & References - Help

©2006 IfALT - IBR/ITM research partner - Plagiarism search in more than 8 billion documents
german english

Figure 11: Docoloc, Start page and detection preference settings



Report
Digital signed

Docoloc

Title: **A report on Text to Speech systems**
Reviewed document: **P_Test_0003.doc**
Processing date: **Sun, 9.7.2006 11:48:55 CEST**

A total of **43** fragments were analysed. As a result **9** fragments (20.9%) 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 fragements. 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.

4 fragments were found in a text with the title: "***Speech synthesis - Wikipedia, the free encyclopedia***", located on:
http://en.wikipedia.org/wiki/Text-to-speech
http://en.wikipedia.org/wiki/Voice_synthesis
http://en.wikipedia.org/wiki/Speech_Synthesis

4 fragments were found in a text with the title: "***Reference.com/Encyclopedia/Speech synthesis***", located on:
http://www.reference.com/browse/wiki/Speech_synthesis

4 fragments were found in a text with the title: "***Speech Synthesis***", located on:
http://en.wikipedia.org/wiki/Speech_synthesis

4 fragments were found in a text with the title: "***speech synthesis: Information From Answers.com***", located on:
http://www.answers.com/topic/speech-synthesis

2 fragments were found in a text with the title: "***Speech synthesis - Bvio***", located on:
http://bvio.ngic.re.kr/Bvio/index.php?title=Speech_synthesis
http://bvio.ngic.re.kr/Bvio/index.php/Speech_synthesis

2 fragments were found in a text with the title: "***Speech synthesis - Biocrawler***", located on:
http://www.biocrawler.com/encyclopedia/Speech_synthesis

2 fragments were found in a text with the title: "***Vergleich von deutschen Sprachsynthese-Systemen 2***", located on:
http://ttssamples.syntheticspeech.de/deutsch/index.html

The next part of TTS takes the symbolic linguistic representation and produces a synthesized voice file. The processes of synthesis generally used can be broadly segmented to two parts One is called the concatenating synthesis approach and other the formant synthesis approach. In concatenating approach as the name describes several smaller recorded voice segments are joined together to produce the human sound. This result in a better quality sounds but requires a well sized voice archives. Most popular techniques for concatinative synthesis are unit selection mechanism in which a large speech database is indexed and sound is produced using special weighted decision tree to assign vocal segment, the concatenated sequence then is processed via certain signal processing techniques to produce smoother sound. This method is believed to produced most natural and quality sounds but requires very large speech DB. Most of commercially available TTS uses this approach. Second approach to unit selection is Diphone synthesis in which only diphones (all available sounds in a language) of particular language is placed in speech archive. The prosody of spoken words are then implemented using DSP methods, this produce lower quality voice then unit selection but smaller is size and many free TTS uses this approach. The third approach is domain specific synthesis where speech DB only has phrases words from a specific domain say a time telling system, or traveling schedule guide or weather system. Such system are fast n efficient but only restricted to very little scope.

Created a graphical representation of this model for better visual understanding

Available systems:

There are many text to speech conversion applications available both open source and free and commercial platforms. These applications are in use at various service levels. Some of the available systems are listed below

Freely available:

1. Festival (http://www.cstr.ed.ac.uk/projects/festival/) is a complete diphone concatenation and unit selection TTS system supporting British and American English, Spanish and Welsh.

2. Flite http://www.speech.cs.cmu.edu/flite/ (Festival-lite) is a smaller, faster alternative version of Festival designed for embedded systems and high volume servers.

3

FreeTTS written entirely in Java, based on Flite.

3. Festvox http://festvox.org/ another festival variant restricte

4. MBROLA http://tcts.fpms.ac.be/synthesis/mbrola.html is t 25 languages.

5. Gnuspeech http://www.gnu.org/software/gnuspeech/ is an ry, speech-synthesis-by-rules.

6. Epos http://epos.ure.cas.cz/ is a rule-driven TTS system p d Slovak

7. HTS http://hts.ics.nitech.ac.jp/ is a freely available HMM-

6 matches: (of about 24) ×

Hizketaren sintesia - Wikipedia [Cache] Flite (Festival-lite) is a smaller, faster alternative version of Festival designed for embedded systems and high volume servers. ...
http://eu.wikipedia.org/wiki/Hizketaren_sintesia

speech synthesis: Information From Answers.com [Cache] Flite (Festival-lite) is a smaller, faster alternative version of Festival designed for embedded systems and high volume servers. ...
http://www.answers.com/topic/speech-synthesis

Reference.com/Encyclopedia/Speech synthesis [Cache] Flite (Festival-lite) is a smaller, faster alternative version of Festival designed for embedded systems and high volume servers. ...
http://www.reference.com/browse/wiki/Speech_synthesis

Speech synthesis - Biocrawler [Cache] Flite (http://www.speech.cs.cmu.edu/flite/) (Festival-lite) is a smaller, faster alternative version of Festival designed for embedded systems and high ...
http://www.biocrawler.com/encyclopedia/Speech_synthesis

Speech synthesis - Bvio [Cache] Flite (Festival-lite) is a smaller, faster alternative version of Festival designed for embedded systems and high volume servers. ...

Figure 12: Docoloc, Sections of test report

This service is totally dependent on the Google API and might become unavailable or change at any point. Service availability is NOT guaranteed by the providers.

Urkund: Another server based plagiarism detection web service which offers an integrated and automated solution for plagiarism detection. It utilizes standard email systems for submission of documents and viewing results. This tool also claims to search through all available online sources giving priority to educational and scandinavian origin. This system claims to process 300 different types of document submissions [Urkund 2006].

Copycatch: A client based tool used to compare locally available databases of documents. It offers 'gold' and 'campus versions' [CopyCatch 2006], giving comparison capabilities for large number of local resources. It also offers a web version which extends the capabilities of plagiarism detection across the internet using the Goggle API. Users are required to signup for personal Google API licences.

WCopyfind: An open source tool for detecting words or phrases of defined length within a local repository of documents [Wcopyfind 2006]. The product is being modified to extend searching capabilities across the internet using the Google API at ACT labs[10]. The resultant product SNITCH [Niezgoda & Way 2006] is expected to be an application version of Docol©c web service.

Eve2 (Essay Verification Engine): This tool works at the client side and uses it own internet search mechanism to find out about plagiarized contents in a suspected document [EVE 2006]. It presents the user with a report identifying matches found in the World Wide Web.

GPSP - Glatt Plagiarism Screening Program: This software works locally and uses an approach to plagiarism detection that differs from previously mentioned services. GPSP detection is based on writing styles and patterns. The author of a suspected submission has to go through a test of filling blank spaces in the writing. The number of correctly filled spaces and the time taken for completion of the test provides the hypothesis of plagiarism guilt or innocence [Glatt 2006]. This has already been discussed in some detail in Section 3.3.

MOSS - a Measure of Software Similarity: MOSS Internet service [MOSS 2006] "accepts batches of documents and returns a set of HTML pages showing where significant sections of a pair of documents are very similar" [Schleimer et al. 2003]. The service specializes in detecting plagiarism in C, C++, Java, Pascal, Ada, ML, Lisp, or Scheme programs.

JPlag: Another internet based service [JPlag 2006] which is used to detect similarities among program source codes. Users upload the files to be compared and the system presents a report identifying matches. JPlag does programming language syntax and structure aware analysis to find results.

When using server based applications to evaluate student's work it is advisable to inform students about the online submission of authenticity checks. Such services

[10] http://actlab.csc.villanova.edu/

keep a fingerprint version of student work in their database which is in turn used for further checking processes. This may be considered a violation of student's intellectual property copyrights [IPR overview 2006]. There are examples of students filing legal cases to prevent their work being submitted to such systems [CNN 2004] and threatening to sue for negligence when the institution was unable to provide clear policy statements about their prohibitions and treatment of plagiarism [Wikipedia:Kent 2006]. All this makes it very important for universities to have a well defined policy and guidance system when students enrol at a university that uses such services.

5 Unexpected Results

The broad scope of plagiarism makes one wonder about the potential of available services. Some of the test cases worth mentioning are listed in this section.

"Paraphrasing" means using some one else's ideas but rewriting it with different words. This is certainly also plagiarism. Plagiarists who want to avoid even the work of coming up with words of their own can use a thesaurus or some "synonymizer" to do the job for them. A proof of concept of such an obvious cheat is a limited dictionary tool the Anti-Anti Plagiarism System11. The library of words in such tools can be enhanced to fit individual requirements. A paraphrased portion of writing using this approach was tested with two of the more often used plagiarism detection services.

We chose the following paragraph:

*"**According** to many **observers**, the **coming decade** will be the **decade** of **speech** technologies. Computer systems, whether **stationary** or mobile, **wired** or **wireless**, will **increasingly** offer users the **opportunity** to **interact** with **information** and **people through** speech. This has been made **possible** by the **arrival** of **relatively robust**, speaker-**independent**, **spontaneous** (or **continuous**) **spoken dialogue** systems in the late 1990s as well as through the **constantly falling costs** of computer speed, bandwidth, storage, and component **miniaturisation**. The **presence** of a speech recogniser in most **appliances combined** with distributed speech processing technologies will **enable** users to speak their **native tongue** when **interacting** with computer systems for a **very large** number of **purposes**. "*

[Bryan Duggan, Mark Deegan, "Considerations in the usage of text to speech (TTS) in the creation of natural sounding voice enabled web systems", ACM International Conference Proceeding Series; Vol. 49, 2003]

Paraphrasing it, using a simple automatic word replacement tool we obtain:

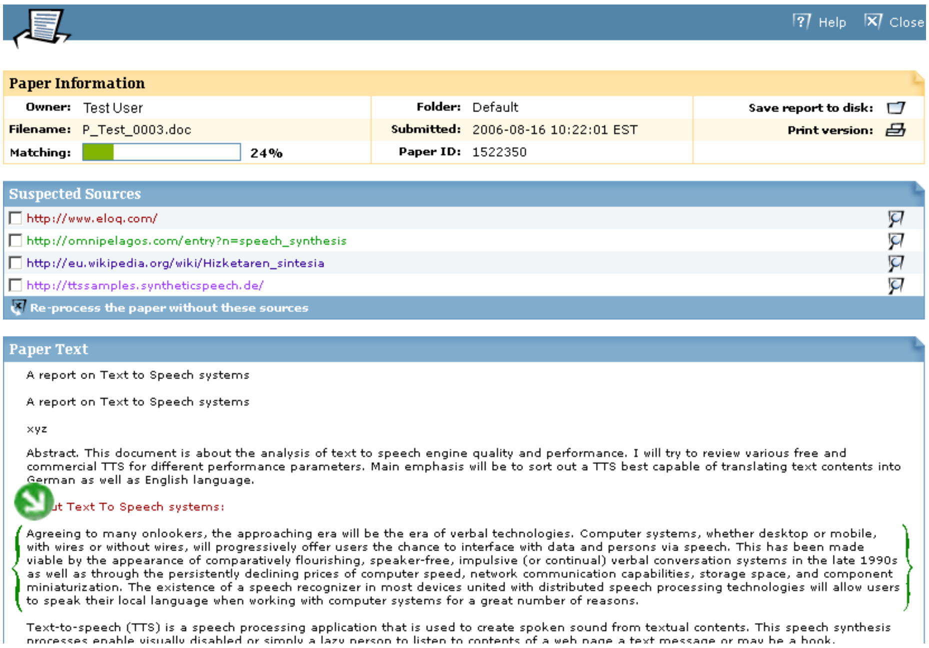
*"**Agreeing** to many **onlookers**, the **approaching era** will be the **era** of **verbal** technologies. Computer systems, whether **desktop** or mobile, **with wires** or **without wires**, will **progressively** offer users the **chance** to **interface** with **data** and **persons via***

[11] http://sourceforge.net/projects/aaps

*speech. This has been made **viable** by the **appearance** of **comparatively flourishing**, speaker-**free**, **impulsive** (or **continual**) **verbal conversation** systems in the late 1990s as well as through the **persistently declining prices** of computer speed, network communication capabilities, **storage space**, and component **miniaturization**. The **existence** of a speech recognizer in most **devices united** with distributed speech processing technologies will **allow** users to speak their **local language** when **working** with computer systems for a **great** number of **reasons**. "*

Note in passing that such simple automatic paraphrasing results in fairly poor English. To really use such an anti-anti/plagiarism tool more sophisticated linguistic techniques are essential.

The originality reports from two service providers in figure 13 and 14 show failure of detection.



Paper Information

Owner:	Test User	Folder:	Default	Save report to disk:	
Filename:	P_Test_0003.doc	Submitted:	2006-08-16 10:22:01 EST	Print version:	
Matching:	24%	Paper ID:	1522350		

Suspected Sources

- □ http://www.eloq.com/
- □ http://omnipelagos.com/entry?n=speech_synthesis
- □ http://eu.wikipedia.org/wiki/Hizketaren_sintesia
- □ http://ttssamples.syntheticspeech.de/

Re-process the paper without these sources

Paper Text

A report on Text to Speech systems

A report on Text to Speech systems

xyz

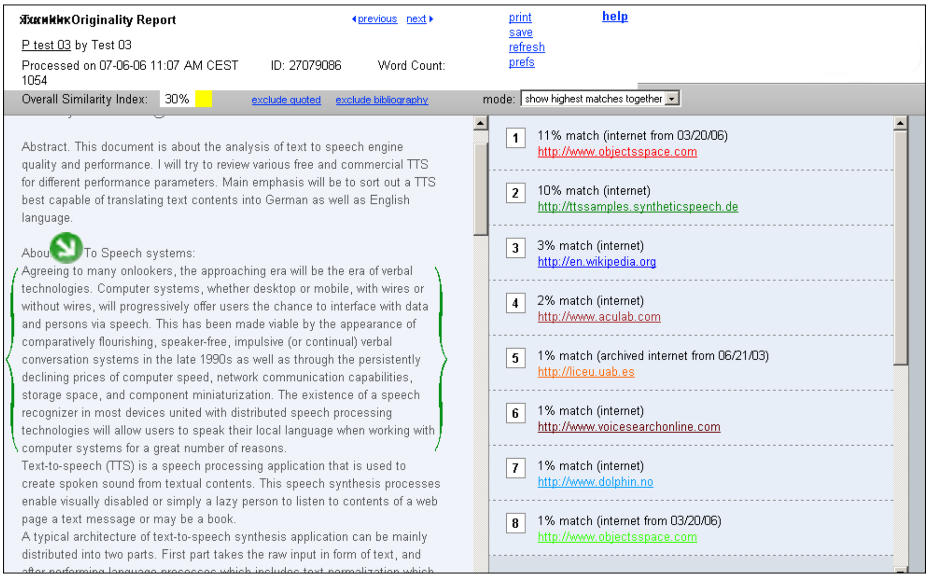
Abstract. This document is about the analysis of text to speech engine quality and performance. I will try to review various free and commercial TTS for different performance parameters. Main emphasis will be to sort out a TTS best capable of translating text contents into German as well as English language.

ut Text To Speech systems:

Agreeing to many onlookers, the approaching era will be the era of verbal technologies. Computer systems, whether desktop or mobile, with wires or without wires, will progressively offer users the chance to interface with data and persons via speech. This has been made viable by the appearance of comparatively flourishing, speaker-free, impulsive (or continual) verbal conversation systems in the late 1990s as well as through the persistently declining prices of computer speed, network communication capabilities, storage space, and component miniaturization. The existence of a speech recognizer in most devices united with distributed speech processing technologies will allow users to speak their local language when working with computer systems for a great number of reasons.

Text-to-speech (TTS) is a speech processing application that is used to create spoken sound from textual contents. This speech synthesis processes enable visually disabled or simply a lazy person to listen to contents of a web page a text message or may be a book.

Figure 13: Originality report by first service



Originality Report

previous next

print
save
refresh
prefs

help

P test 03 by Test 03

Processed on 07-06-06 11:07 AM CEST ID: 27079086 Word Count: 1054

Overall Similarity Index: 30% exclude quoted exclude bibliography mode: show highest matches together

Abstract. This document is about the analysis of text to speech engine quality and performance. I will try to review various free and commercial TTS for different performance parameters. Main emphasis will be to sort out a TTS best capable of translating text contents into German as well as English language.

Abou To Speech systems:
Agreeing to many onlookers, the approaching era will be the era of verbal technologies. Computer systems, whether desktop or mobile, with wires or without wires, will progressively offer users the chance to interface with data and persons via speech. This has been made viable by the appearance of comparatively flourishing, speaker-free, impulsive (or continual) verbal conversation systems in the late 1990s as well as through the persistently declining prices of computer speed, network communication capabilities, storage space, and component miniaturization. The existence of a speech recognizer in most devices united with distributed speech processing technologies will allow users to speak their local language when working with computer systems for a great number of reasons.
Text-to-speech (TTS) is a speech processing application that is used to create spoken sound from textual contents. This speech synthesis processes enable visually disabled or simply a lazy person to listen to contents of a web page a text message or may be a book.
A typical architecture of text-to-speech synthesis application can be mainly distributed into two parts. First part takes the raw input in form of text, and

1 11% match (internet from 03/20/06)
http://www.objectsspace.com

2 10% match (internet)
http://ttssamples.syntheticspeech.de

3 3% match (internet)
http://en.wikipedia.org

4 2% match (internet)
http://www.aculab.com

5 1% match (archived internet from 06/21/03)
http://liceu.uab.es

6 1% match (internet)
http://www.voicesearchonline.com

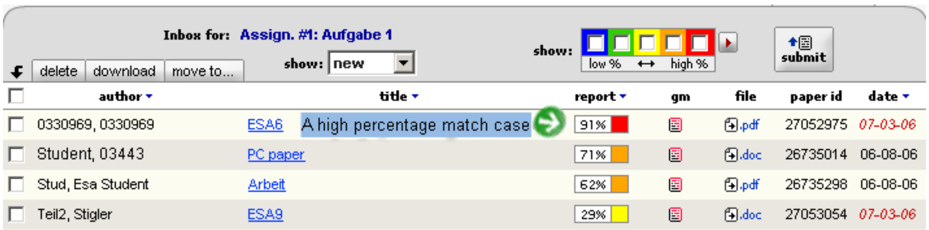
7 1% match (internet)
http://www.dolphin.no

8 1% match (internet from 03/20/06)
http://www.objectsspace.com

Figure 14: Originality report by second service

The above example shows the weakness of word by word comparison or using fingerprints just involving the exact words occurring in a text. We will come back to this issue later in section 6 where we will discuss possible solutions for this problem.

At times, various systems show a very high percentage of matches; this does not necessarily mean that the document is plagiarized. Rather, it can be due to the fact that we are checking some paper that has already been put on some server, hence the match is made with exactly the same contribution by the same author. In such a case, one can use the facility to exclude the high percentage matching original source and regenerate the report showing other matches detected by the system. Figures (15 - 17) show such a case and two versions of originality report.

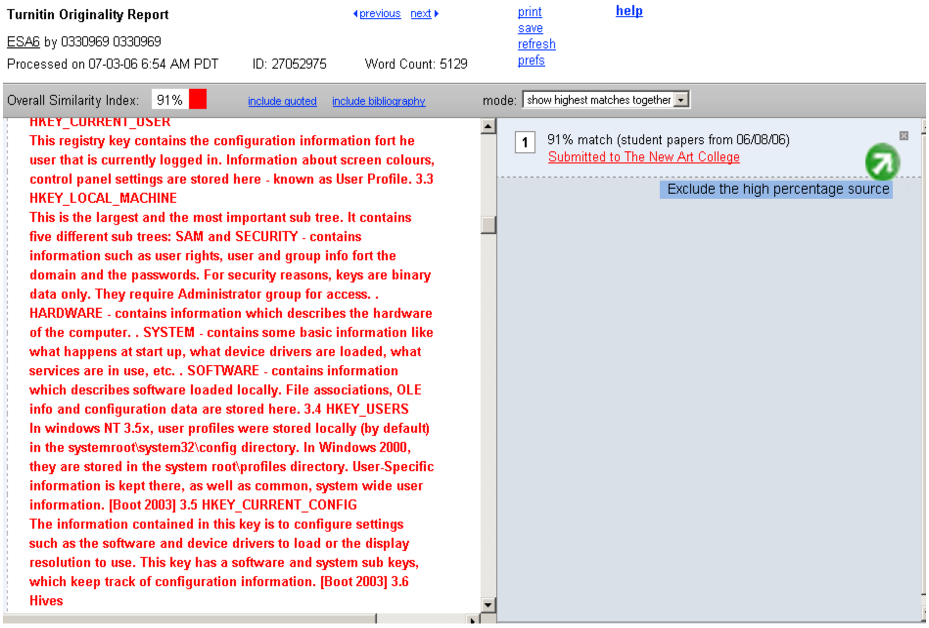


Inbox for: **Assign. #1: Aufgabe 1**

delete download move to... **show:** new **show:** low % ↔ high % **submit**

□	author	title	report	gm	file	paper id	date
□	0330969, 0330969	ESA6 A high percentage match case	91% 🟥	▤	.pdf	27052975	07-03-06
□	Student, 03443	PC paper	71% 🟧	▤	.doc	26735014	06-08-06
□	Stud, Esa Student	Arbeit	62% 🟧	▤	.pdf	26735298	06-08-06
□	Teil2, Stigler	ESA9	29% 🟨	▤	.doc	27053054	07-03-06

Figure 15: System showing 91% match for a particular paper



Turnitin Originality Report

ESA6 by 0330969 0330969

Processed on 07-03-06 6:54 AM PDT ID: 27052975 Word Count: 5129

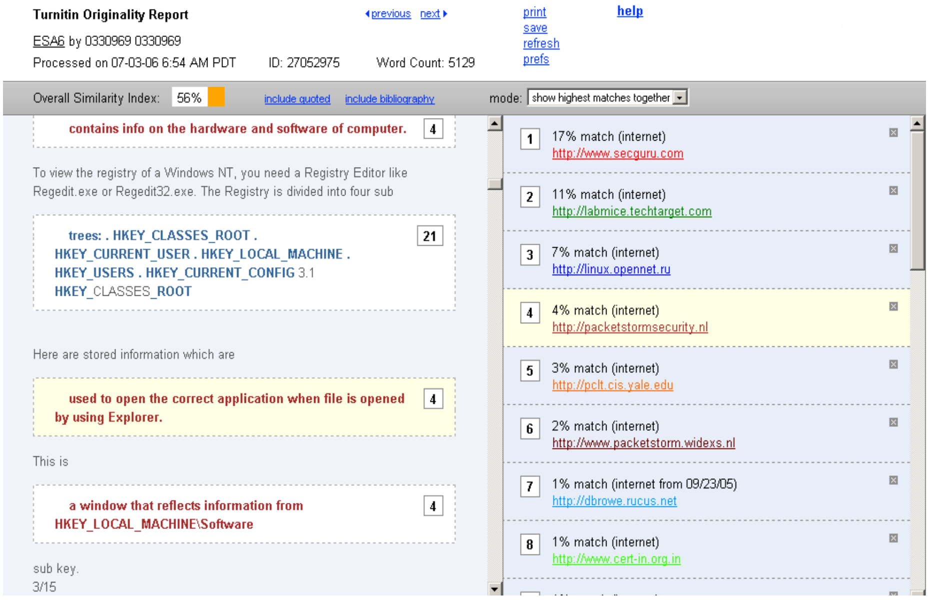
Overall Similarity Index: 91%

mode: show highest matches together

HKEY_CURRENT_USER
This registry key contains the configuration information fort he user that is currently logged in. Information about screen colours, control panel settings are stored here - known as User Profile. 3.3 HKEY_LOCAL_MACHINE
This is the largest and the most important sub tree. It contains five different sub trees: SAM and SECURITY - contains information such as user rights, user and group info fort the domain and the passwords. For security reasons, keys are binary data only. They require Administrator group for access. . HARDWARE - contains information which describes the hardware of the computer. . SYSTEM - contains some basic information like what happens at start up, what device drivers are loaded, what services are in use, etc. . SOFTWARE - contains information which describes software loaded locally. File associations, OLE info and configuration data are stored here. 3.4 HKEY_USERS
In windows NT 3.5x, user profiles were stored locally (by default) in the systemroot\system32\config directory. In Windows 2000, they are stored in the system root\profiles directory. User-Specific information is kept there, as well as common, system wide user information. [Boot 2003] 3.5 HKEY_CURRENT_CONFIG
The information contained in this key is to configure settings such as the software and device drivers to load or the display resolution to use. This key has a software and system sub keys, which keep track of configuration information. [Boot 2003] 3.6 Hives

1 91% match (student papers from 06/08/06)
Submitted to The New Art College

Figure 16: Report showing high percentage of match from a single source



Turnitin Originality Report

ESA6 by 0330969 0330969

Processed on 07-03-06 6:54 AM PDT ID: 27052975 Word Count: 5129

Overall Similarity Index: 56%

mode: show highest matches together

contains info on the hardware and software of computer. 4

To view the registry of a Windows NT, you need a Registry Editor like Regedit.exe or Regedit32.exe. The Registry is divided into four sub

trees: . HKEY_CLASSES_ROOT . HKEY_CURRENT_USER . HKEY_LOCAL_MACHINE . HKEY_USERS . HKEY_CURRENT_CONFIG 3.1 HKEY_CLASSES_ROOT 21

Here are stored information which are

used to open the correct application when file is opened by using Explorer. 4

This is

a window that reflects information from HKEY_LOCAL_MACHINE\Software 4

sub key.
3/15

1 17% match (internet)
http://www.secguru.com

2 11% match (internet)
http://labmice.techtarget.com

3 7% match (internet)
http://linux.opennet.ru

4 4% match (internet)
http://packetstormsecurity.nl

5 3% match (internet)
http://pclt.cis.yale.edu

6 2% match (internet)
http://www.packetstorm.widexs.nl

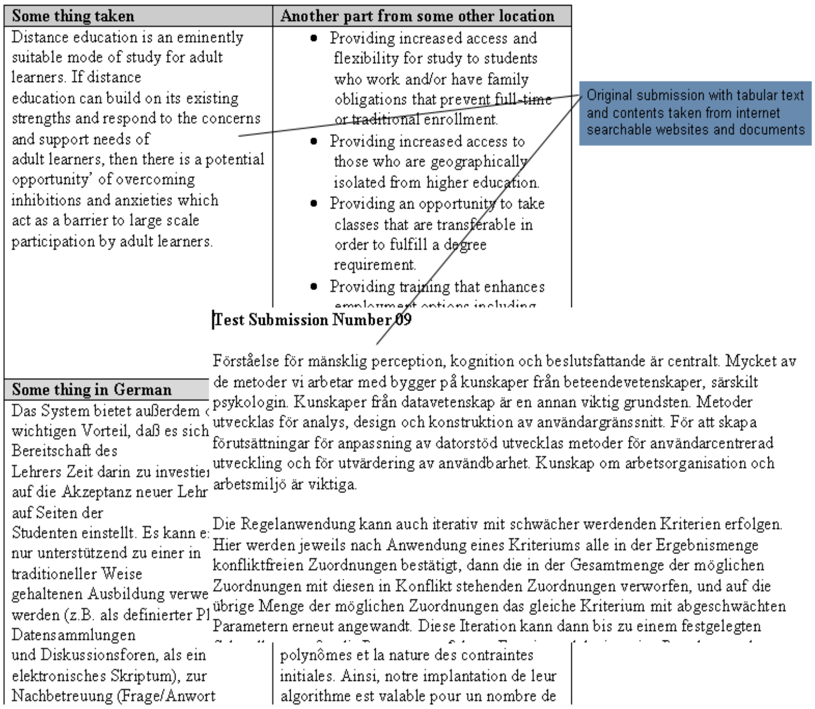
7 1% match (internet from 09/23/05)
http://dbrowe.rucus.net

8 1% match (internet)
http://www.cert-in.org.in

Figure 17: More meaning full report after excluding the high percentage source

Hence if a system finds a very high percentage match it can mean that the uploading was done in the wrong order!

Testing with tabular information and text in languages with special characters (German, Swedish, French etc.) showed that some of available systems are unable to correctly process data in table cells. Figure 18 shows few portions of test documents submitted to different systems. The collected text in test comes from internet available documents and websites.



Some thing taken	Another part from some other location
Distance education is an eminently suitable mode of study for adult learners. If distance education can build on its existing strengths and respond to the concerns and support needs of adult learners, then there is a potential opportunity' of overcoming inhibitions and anxieties which act as a barrier to large scale participation by adult learners.	• Providing increased access and flexibility for study to students who work and/or have family obligations that prevent full-time or traditional enrollment. • Providing increased access to those who are geographically isolated from higher education. • Providing an opportunity to take classes that are transferable in order to fulfill a degree requirement. • Providing training that enhances employment options including

Original submission with tabular text and contents taken from internet searchable websites and documents

Test Submission Number 09

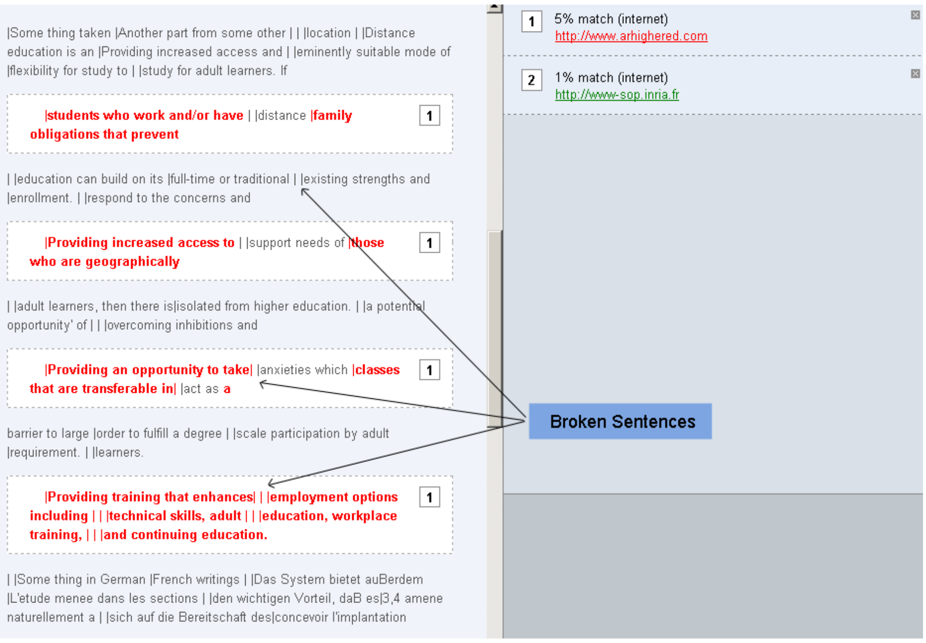
Förståelse för mänsklig perception, kognition och beslutsfattande är centralt. Mycket av de metoder vi arbetar med bygger på kunskaper från beteendevetenskaper, särskilt psykologin. Kunskaper från datavetenskap är en annan viktig grundsten. Metoder utvecklas för analys, design och konstruktion av användargränssnitt. För att skapa förutsättningar för anpassning av datorstöd utvecklas metoder för användarcentrerad utveckling och för utvärdering av användbarhet. Kunskap om arbetsorganisation och arbetsmiljö är viktiga.

Die Regelanwendung kann auch iterativ mit schwächer werdenden Kriterien erfolgen. Hier werden jeweils nach Anwendung eines Kriteriums alle in der Ergebnismenge konfliktfreien Zuordnungen bestätigt, dann die in der Gesamtmenge der möglichen Zuordnungen mit diesen in Konflikt stehenden Zuordnungen verworfen, und auf die übrige Menge der möglichen Zuordnungen das gleiche Kriterium mit abgeschwächten Parametern erneut angewandt. Diese Iteration kann dann bis zu einem festgelegten

Some thing in German	
Das System bietet außerdem wichtigen Vorteil, daß es sich Bereitschaft des Lehrers Zeit darin zu investie auf die Akzeptanz neuer Lehr auf Seiten der Studenten einstellt. Es kann e nur unterstützend zu einer in traditioneller Weise gehaltenen Ausbildung verwe werden (z.B. als definierter Pl Datensammlungen und Diskussionsforen, als ein elektronisches Skriptum), zur Nachbetreuung (Frage/Anwort	polynômes et la nature des contraintes initiales. Ainsi, notre implantation de leur algorithme est valable pour un nombre de

Figure 18: Original tabular data with text containing special characters

Processing of testing documents through different detection services showed that in some cases the sentences are broken irregularly making a wrong fingerprint which might lead to false or no match detection. Some systems are also unable to properly process special characters; this might be the cause of no or lesser percentage of match detection in few test cases. Figures (19 - 20) show few portions of resulting reports.


|Some thing taken |Another part from some other | | |location | |Distance education is an |Providing increased access and | |eminently suitable mode of |flexibility for study to | |study for adult learners. If

|students who work and/or have | |distance |family obligations that prevent 1

| |education can build on its |full-time or traditional | |existing strengths and |enrollment. | |respond to the concerns and

|Providing increased access to | |support needs of |those who are geographically 1

| |adult learners, then there is|isolated from higher education. | |a potential opportunity' of | | |overcoming inhibitions and

|Providing an opportunity to take| |anxieties which |classes that are transferable in| |act as a 1

barrier to large |order to fulfill a degree | |scale participation by adult |requirement. | |learners.

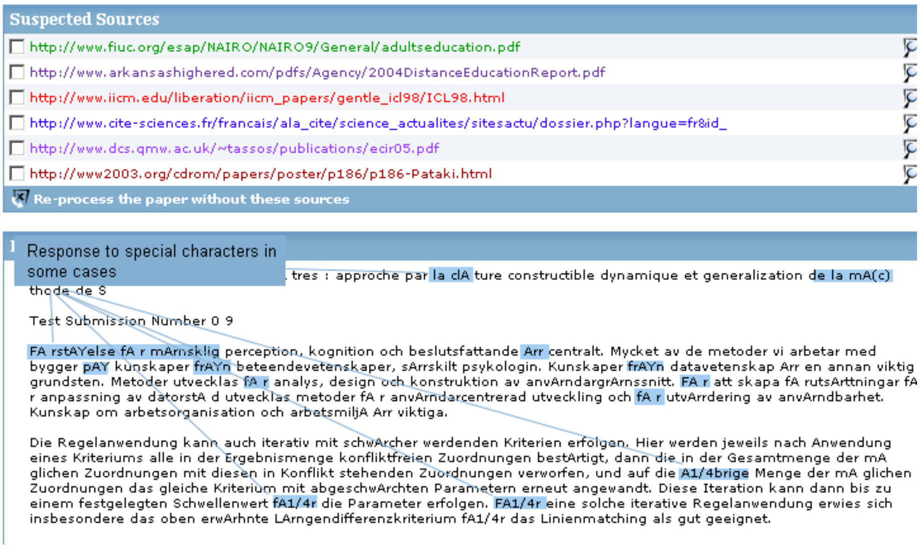
|Providing training that enhances| | |employment options including | | |technical skills, adult | | |education, workplace training, | | |and continuing education. 1

| |Some thing in German |French writings | |Das System bietet auBerdem |L'etude menee dans les sections | |den wichtigen Vorteil, daB es|3,4 amene naturellement a | |sich auf die Bereitschaft des|concevoir l'implantation

1 5% match (internet) http://www.arhighered.com

2 1% match (internet) http://www-sop.inria.fr

Figure 19: Report with broken table cell text


Suspected Sources

- □ http://www.fiuc.org/esap/NAIRO/NAIRO9/General/adultseducation.pdf
- □ http://www.arkansashighered.com/pdfs/Agency/2004DistanceEducationReport.pdf
- □ http://www.iicm.edu/liberation/iicm_papers/gentle_icl98/ICL98.html
- □ http://www.cite-sciences.fr/francais/ala_cite/science_actualites/sitesactu/dossier.php?langue=fr&id_
- □ http://www.dcs.qmw.ac.uk/~tassos/publications/ecir05.pdf
- □ http://www2003.org/cdrom/papers/poster/p186/p186-Pataki.html

. tres : approche par la clA ture constructible dynamique et generalization de la mA(c) thode de S

Test Submission Number 0 9

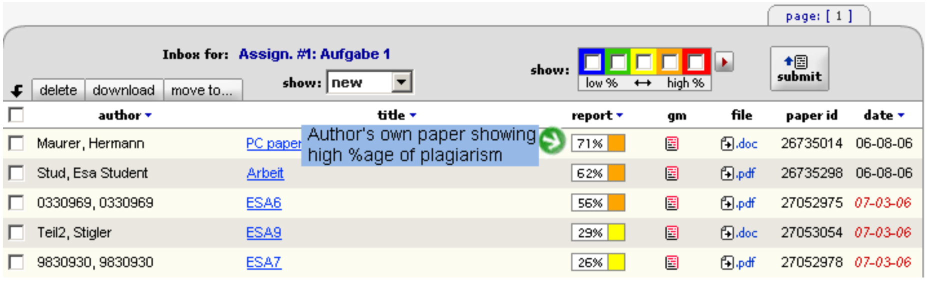
FA rstAYelse fA r mArnsklig perception, kognition och beslutsfattande Arr centralt. Mycket av de metoder vi arbetar med bygger pAY kunskaper frAYn beteendevetenskaper, sArrskilt psykologin. Kunskaper frAYn datavetenskap Arr en annan viktig grundsten. Metoder utvecklas fA r analys, design och konstruktion av anvArndargrArnssnitt. FA r att skapa fA rutsArttningar fA r anpassning av datorstA d utvecklas metoder fA r anvArndarcentrerad utveckling och fA r utvArrdering av anvArndbarhet. Kunskap om arbetsorganisation och arbetsmiljA Arr viktiga.

Die Regelanwendung kann auch iterativ mit schwArcher werdenden Kriterien erfolgen. Hier werden jeweils nach Anwendung eines Kriteriums alle in der Ergebnismenge konfliktfreien Zuordnungen bestArtigt, dann die in der Gesamtmenge der mA glichen Zuordnungen mit diesen in Konflikt stehenden Zuordnungen verworfen, und auf die A1/4brige Menge der mA glichen Zuordnungen das gleiche Kriterium mit abgeschwArchten Parametern erneut angewandt. Diese Iteration kann dann bis zu einem festgelegten Schwellenwert fA1/4r die Parameter erfolgen. FA1/4r eine solche iterative Regelanwendung erwies sich insbesondere das oben erwArhnte LArngendifferenzkriterium fA1/4r das Linienmatching als gut geeignet.

Figure 20: Document report with special characters

One interesting fact about the use of plagiarism detection services is that they can be also employed to discover illegal copies of our own writing as well. One such example is shown below: A paper produced by the first author of this paper showed a

71% match using one of the plagiarism detection services. A more detailed analysis of the report revealed the fact that various portions of the paper were used illegally at different places. Figures (21-22) show the relevant reports.



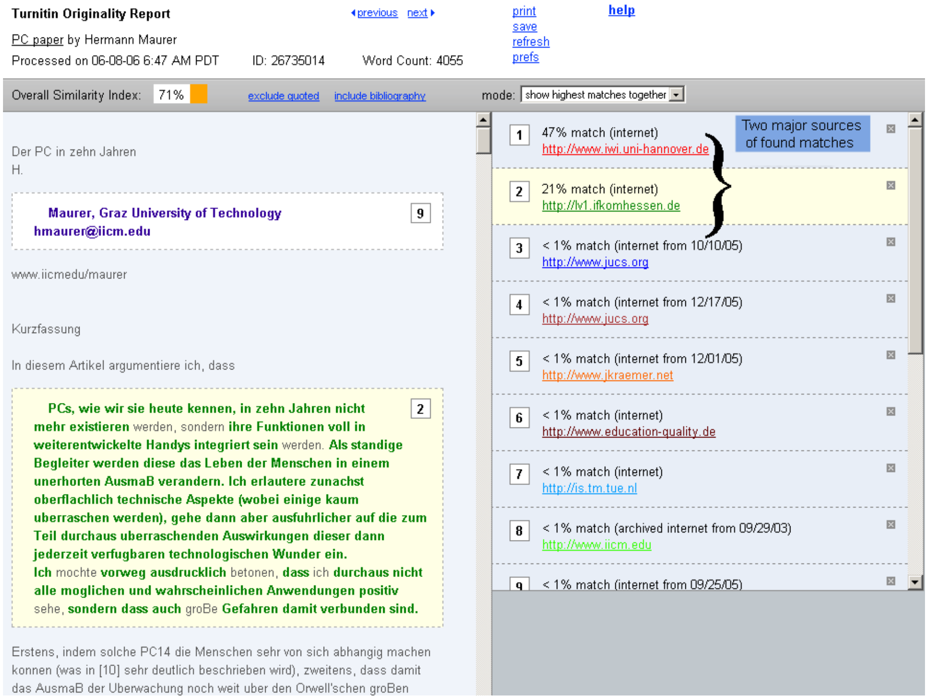
Inbox for: Assign. #1: Aufgabe 1

show: new

□	author	title	report	gm	file	paper id	date
□	Maurer, Hermann	PC paper	71% 🟧		.doc	26735014	06-08-06
□	Stud, Esa Student	Arbeit	62% 🟧		.pdf	26735298	06-08-06
□	0330969, 0330969	ESA6	56% 🟧		.pdf	27052975	07-03-06
□	Teil2, Stigler	ESA9	29% 🟨		.doc	27053054	07-03-06
□	9830930, 9830930	ESA7	26% 🟨		.pdf	27052978	07-03-06

Author's own paper showing high %age of plagiarism

Figure 21: Use of plagiarism detection tools to discover copies of own writings



Turnitin Originality Report

PC paper by Hermann Maurer

Processed on 06-08-06 6:47 AM PDT   ID: 26735014   Word Count: 4055

Overall Similarity Index: 71% 🟧

mode: show highest matches together

Der PC in zehn Jahren
H.

Maurer, Graz University of Technology
hmaurer@iicm.edu 9

www.iicmedu/maurer

Kurzfassung

In diesem Artikel argumentiere ich, dass

PCs, wie wir sie heute kennen, in zehn Jahren nicht mehr existieren werden, sondern ihre Funktionen voll in weiterentwickelte Handys integriert sein werden. Als standige Begleiter werden diese das Leben der Menschen in einem unerhorten AusmaB verandern. Ich erlautere zunachst oberflachlich technische Aspekte (wobei einige kaum uberraschen werden), gehe dann aber ausfuhrlicher auf die zum Teil durchaus uberraschenden Auswirkungen dieser dann jederzeit verfugbaren technologischen Wunder ein. Ich mochte vorweg ausdrucklich betonen, dass ich durchaus nicht alle moglichen und wahrscheinlichen Anwendungen positiv sehe, sondern dass auch groBe Gefahren damit verbunden sind. 2

Erstens, indem solche PC14 die Menschen sehr von sich abhangig machen konnen (was in [10] sehr deutlich beschrieben wird), zweitens, dass damit das AusmaB der Uberwachung noch weit uber den Orwell'schen groBen

1 47% match (internet)
http://www.iwi.uni-hannover.de

2 21% match (internet)
http://lv1.ifkomhessen.de

3 < 1% match (internet from 10/10/05)
http://www.jucs.org

4 < 1% match (internet from 12/17/05)
http://www.jucs.org

5 < 1% match (internet from 12/01/05)
http://www.jkraemer.net

6 < 1% match (internet)
http://www.education-quality.de

7 < 1% match (internet)
http://is.tm.tue.nl

8 < 1% match (archived internet from 09/29/03)
http://www.iicm.edu

9 < 1% match (internet from 09/25/05)

Two major sources of found matches

Figure 22: Report with links showing copied portion of text

The highlighted/plagiarised portions in the report are linked to a specific URL pointing to the source. Visiting these sources confirms that the text was illegally copied from the author's paper that had appeared in a journal previously.

	Turnitin	**Mydropbox**	**Docol©c**
Technology	Web based, server side processing, support internet and other external scholastic databases	Web based, server side processing, support internet and other external scholastic databases	Web based, server side processing, support internet searches via Google API
Supported file types	MS Word, WordPerfect, PostScript, PDF, HTML, RTF, and plain text	ZIP, DOC, TXT, PDF, RTF, HTML and Direct text paste in text box at site	PDF, DOC (Word®), RTF, HTML, PPT, (Power Point®), XLS (Excel®), and TXT
Verbatim/Cut-Paste check	Yes	Yes	Yes
Paraphrase check	No	No	No
Tabular information processing	Showed problem in some cases	Yes	Yes
Translation check	No	No	No
Image/multi-media checks	No	No	No
Reference validity check	No	No	No
Exclusion/selection of sources	Yes	Yes	No

Table 1: Comparison of plagiarism detection capabilities

We tested two commonly used commercial services (Turnitin and Mydropbox) with a selected set of submissions. The experiments showed generally similar results. We will return to a comparison of those tools and other techniques together with a discussion that shows how blurred the borders of plagiarism are in [Zaka & Maurer 2006], based on first observations in [Maurer et al. 2006].

6 Advanced techniques

Most services and tools described in earlier sections address verbatim plagiarism and utilize the document source comparison approach for detection. Thus, similarities that are not detectable by just comparison of word-based fingerprints usually escape those tools. However, more sophisticated similarity detection which is the core of source comparison is used to some extent already in many other areas such as data mining, indexing, knowledge management and automated essay grading.

Although we are not aware of concept-oriented or semantic similarity detection in existing plagiarism detection services we do find experimental research projects and other commercial products which utilize innovative similarity detection methodologies, often for simpler tasks e.g. just checking whether a question asked is similar to one in the list of available FAQs.

A research paper in this direction describing so-called Active Documents explains that the most satisfying approach for checking whether a similarity exists in the meaning of different pieces of text is of course to determine their semantic equivalence. "To actually prove that two pieces of text are semantically equivalent one would require a complete understanding of natural language, something still quite elusive. However, we can consider a compromise: rather than allowing a full natural language we restrict our attention to a simplified grammar and to a particular domain for which an ontology (semantic network) is developed. Clearly, sufficiently restricting syntactic possibilities and terms to be used will allow one to actually prove the equivalence of pieces of text." [Heinrich & Maurer 2000].

Before we further look at various experiments that use semantic information and find aspects that may limit their use in similarity analysis we first describe one mathematical approach generally used in similarity detection.

A popular approach to similarity detection or pattern recognition is the use of a vector space model to determine cosine (i.e. angular) similarity among vectors of keywords/function-words extracted from the text under inspection.

To elaborate more let us take an example of two sentences

Text A: "A rainy day with a cold wind"
Text B: "A sunny day with blue sky"
Each text is represented in a word frequency table as follows:

Text A:	**Text B:**	**Complete vocabulary:**
a: 2	a: 1	a
rainy: 1	blue: 1	blue
day: 1	day: 1	cold
with: 1	sunny: 1	day
cold: 1	sky: 1	rainy
wind: 1	with: 1	sky
		sunny
		wind
		with

Table 2: Word-frequency in text, and complete vocabulary

The representation of the two pieces of text as vectors based against the vocabulary is: Text A= {2,0,1,1,1,0,0,1,1} and Text B= {1,1,0,1,0,1,1,0,1}.

Now let us take some text for similarity detection e.g. C: "A cold day". The vector representation is C= {1,0,1,1,0,0,0,0,0}.

The cosine similarity measure between text A and C is calculated using formula $\frac{\text{Vector-A} \bullet \text{Vector-C}}{|\text{Vector-A}||\text{Vector-C}|}$

Calculations give us similarity measure of 0.769 between document A and C and 0.471 between B and C. Thus one can make assumption of similarity even if the two pieces of text are not completely identical. In real applications word vectors are made by the removal of stop words (frequently occurring words that can be ignored in a query, e.g. the, is, of, be, a etc.) and keyword vectors generally are made using tf-idf[12] weights. These are very common methods and their functionality and limitations are well known. One can imagine that using a semantic matrix of words and concepts for a large corpus of text and complete language information, the vector space can be easily too large for practical computation. Thus, we need ideas and methodologies to improve this analysis. Examples are limiting the domain (i.e. to the ontology of subject in question) as described earlier in this section or other techniques which we will discuss a bit later.

The plagiarists today are becoming aware of limitations of existing systems and avoid detection by using linguistic tools as demonstrated in one example above. They can replace functional words after small intervals by using synonyms, retaining the idea or concept behind the sentences, yet remain undetected.

However, semantic or syntactic elements of any language can be used to enhance similarity detection mechanism and anti plagiarism software as well. One such approach to empower document similarity detection using semantic analysis is discussed by Iyer and Singh. Their system extracts keywords (nouns, verbs, adjectives in this case, ignoring adverbs, pronouns, prepositions, conjunctions and interjections) representing structural characteristics of documents. Synonym clusters for keywords are looked up from WordNet[13] and each cluster is represented with a numeric value. All keywords that are present in the structural characteristic tree of the document also carry the numeric value of the synonym cluster they belong too. Thus, when comparing sources, the binary comparison of synonym cluster numbers tells whether two words are synonyms. The software runs the comparison algorithms initially on the structural characteristic tree of the complete document. If similarities are above a certain threshold, only then is sentence level comparison initiated. This makes the system capable of detecting similarity even with minor semantic modifications at sentence level [Iyer & Singh 2005].

Another approach of “Using Syntactic Information to Identify Plagiarism” shows the effectiveness of linguistic information to detect similarities among different words to express the same material. This experimental study goes beyond just using synonyms, it “presents a set of low-level syntactic structures that capture creative aspects of writing and show that information about linguistic similarities of works improves recognition of plagiarism” [Uzuner et al. 2005]. This research experiment identifies classes for different syntactic expressions for the same content, called “syntactic elements of expression”. These elements of expression include: different variations of initial and final phrases of a sentence, argument structures of verb

[12] http://en.wikipedia.org/wiki/Tf-idf

[13] http://wordnet.princeton.edu/

phrases and syntactic classes of verb phrases. All possible variations are considered to combat initial and final phrase structure alterations.

For example, a sentence may have following class of three different expressive alterations:
(a) Martha can finally put some money in the bank.
(b) Martha can put some money in the bank, finally.
(c) Finally, Martha can put some money in the bank." [Uzuner et al. 2005]

This research experiment also enriches its syntactic elements of expressions by employing Levin's classes [Levin 1993] of verbs. In Levin's classes verbs are classified using various syntactic alterations a verb is subject to, and the classes of verbs with similar meanings. These features are combined to create further elements of expression for testing data (including English translations of literary work by different translators). This data is then used for recognition of paraphrased writings with similar contents. Although this is a computationally expensive approach compared to conventional content recognition approaches such as comparing tf-idf weighted keywords, function words, distribution of word lengths and sentence lengths, the results presented show a significantly better average of similarity detection over baseline/conventional approaches [Uzuner et al. 2005].

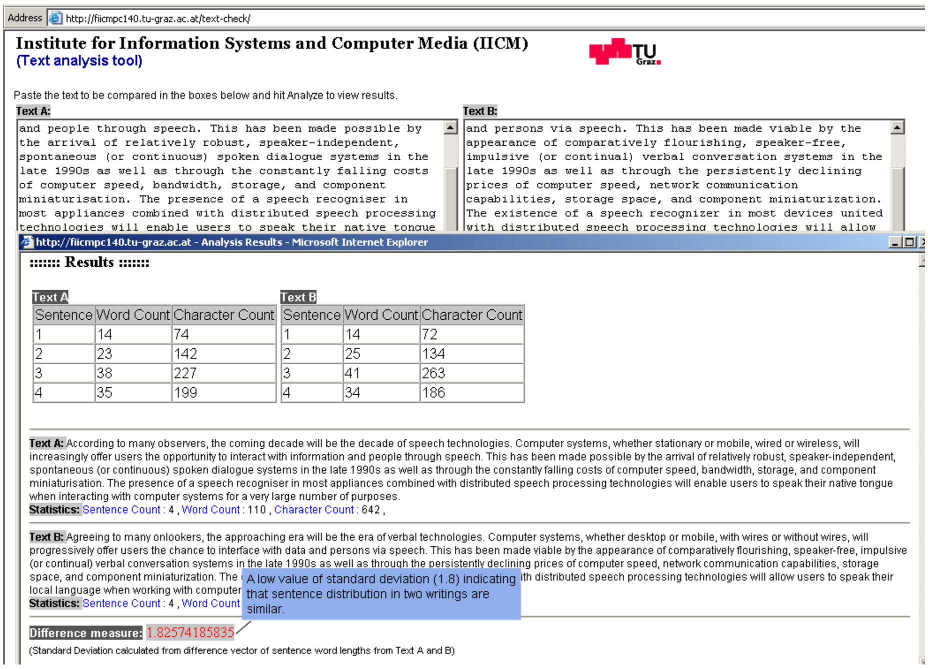
There are services available that evaluate the text contents on a conceptual level for automated essay grading. They compare semantic similarities among contents (written essay and domain knowledge) to calculate grades. A method used in such systems is "Latent Semantic Analysis" (LSA). This is a statistical technique for extracting and representing the similarity of meaning of words and passages by the analysis of large bodies of text" [LSA 2006]. A matrix of words and related segments is used to build a word to concept semantic domain space. The text needed to be checked for similarity with this domain space is also represented in document vector form. If the document vector is similar to the model answer vector (again the measure of angle between vectors defines closeness to each other) in this domain the document will have higher similarity grade. This kind of system which detects semantic similarities to grade some writing can also be used effectively for paraphrased plagiarism detection. But even with a singular value decomposition approach in LSA to reduce word and context matrix, the matrix dimensions are still large and the vector space analysis is computationally demanding.

As mentioned before, in the case of plagiarism detection we are usually dealing with a very large corpus of textual information making such analysis not as yet practical. This necessitates methodologies to enhance processing and making the methods mentioned feasible for practical environments.

Another approach utilizing the power of Normalized Word Vectors (NWV) is to further reduce the word-concept vector space by normalizing all words to a thesaurus root word. The convergence to a singular concept word reduces the domain space and document vectors significantly. The cosine similarity measure can then be used to find semantic relevance among answers [Williams 2006]. This in turn leads to a reduced computational load and can perhaps make such methodology practical for plagiarism detection.

A more generic technology of query formulation is being investigated which use NWV technology and dynamic ontological filtering to help extend the semantic similarity detection mechanism in various applications [Dreher & Williams 2006].

It is interesting to note that some times less computationally demanding simple text structure analysis techniques such as average word lengths, sentence counts, words per sentence etc. can be very useful in cases of suspected plagiarism in different documents. A simple example in Figure 23 show the use of sentence and word counts to determine style similarity between two paragraphs in the test case used before, where simple synonym replacement made similarity undetectable using conventional plagiarism detection services. We developed a simple program to calculate the standard deviation of the difference vector of the sentence lengths (calculated on the basis of number of words) of two suspected paragraphs. This can be a good indicator of text structure similarity, and can be used to identify potentially similar documents.



Institute for Information Systems and Computer Media (IICM)
(Text analysis tool)

Paste the text to be compared in the boxes below and hit Analyze to view results.

Text A:

and people through speech. This has been made possible by
the arrival of relatively robust, speaker-independent,
spontaneous (or continuous) spoken dialogue systems in the
late 1990s as well as through the constantly falling costs
of computer speed, bandwidth, storage, and component
miniaturisation. The presence of a speech recogniser in
most appliances combined with distributed speech processing
technologies will enable users to speak their native tongue

Text B:

and persons via speech. This has been made viable by the
appearance of comparatively flourishing, speaker-free,
impulsive (or continual) verbal conversation systems in the
late 1990s as well as through the persistently declining
prices of computer speed, network communication
capabilities, storage space, and component miniaturization.
The existence of a speech recognizer in most devices united
with distributed speech processing technologies will allow

http://fiicmpc140.tu-graz.ac.at - Analysis Results - Microsoft Internet Explorer

:::::: Results ::::::

Text A

Sentence	Word Count	Character Count
1	14	74
2	23	142
3	38	227
4	35	199

Text B

Sentence	Word Count	Character Count
1	14	72
2	25	134
3	41	263
4	34	186

Text A: According to many observers, the coming decade will be the decade of speech technologies. Computer systems, whether stationary or mobile, wired or wireless, will increasingly offer users the opportunity to interact with information and people through speech. This has been made possible by the arrival of relatively robust, speaker-independent, spontaneous (or continuous) spoken dialogue systems in the late 1990s as well as through the constantly falling costs of computer speed, bandwidth, storage, and component miniaturisation. The presence of a speech recogniser in most appliances combined with distributed speech processing technologies will enable users to speak their native tongue when interacting with computer systems for a very large number of purposes.
Statistics: Sentence Count : 4 , Word Count : 110 , Character Count : 642 ,

Text B: Agreeing to many onlookers, the approaching era will be the era of verbal technologies. Computer systems, whether desktop or mobile, with wires or without wires, will progressively offer users the chance to interface with data and persons via speech. This has been made viable by the appearance of comparatively flourishing, speaker-free, impulsive (or continual) verbal conversation systems in the late 1990s as well as through the persistently declining prices of computer speed, network communication capabilities, storage space, and component miniaturization. The ith distributed speech processing technologies will allow users to speak their local language when working with computer
Statistics: Sentence Count : 4 , Word Count

A low value of standard deviation (1.8) indicating that sentence distribution in two writings are similar.

Difference measure: 1.82574185835
(Standard Deviation calculated from difference vector of sentence word lengths from Text A and B)

Figure 23: Statistical text structure analysis

Statistical analysis may determine a preliminary similarity measure. Suspected parts can then be put to further more advanced semantic or syntactic testing algorithms to confirm the detection.

7 Problems and Visions

Looking at the extent of the problem, it is quite obvious that academia requires tools and services to automate and enhance plagiarism detection. Our analysis of these tools revealed a number of areas which need attention.

Almost all tools and services produce results that can not be used as a final report without human interpretation. The problems pointed out by the system have to be analyzed by domain experts for verification and further investigation. This limitation suggests more work is required to adapt systems to provide an analysis layer that triggers further investigative matches and produces a more conclusive result. A viable solution will probably have to be interactive, with feedback from the examiner to confirm system assumptions before proceeding with additional analyses.

The results of research studies and experiments described in the previous section seem encouraging. However, to date, we found no evidence of any released tool or service which uses language information, syntactic and semantic aspects of writings to detect paraphrased or translated plagiarism. Current detection tools are lagging behind without having broad and generic ontology of linguistic or writing parameters which convert the search patterns to a certain level of abstraction.

Increased ease of access to global and multilingual contents makes detection of translated plagiarism a vital requirement for detection systems. The detection services can use translation tools to convert foreign language contents into a basic English form, apply normalization techniques to generate a generic index of document sources and apply semantic similarity checks for detection. To illustrate what we mean consider the following example

Synonym classes in German:
{Cabriolet, Cabrio, Zweisitzer, Automobil, Personenauto, PKW, Auto, …} → Auto
{tiefbblau, azurblau, türkisblau, blau, ...} → blau
{Klatsch, Plumps,…} → Lärm
{fallen, sinken, herunterfallen, hinunterfallen} →fallen
{laut, heftig, stark, groß,…} → groß
{Bach, Fluss, Teich, See, Wasser, …} → Wasser

Synonym classes English:
{cabriolet, car, limousine, automobile, …} → car
{deep blue, azul, azure, sky-blue, dark blue, …} → blue
{splash, splish, …} → noise
{fall, drop, …} → fall
{loud, strong, great, big, …} → big
{creek, brook, stream, river, pond, pool, lake, …} → water

Let us now see, how the two sentences: "Das azurblaue Cabriolet fiel mit lautem Klatschen in den Bach" (German) and "The deep-blue limousine dropped with a big splash into the river" (English) can be determined to be similar:

The sentence:
"Das azurblaue Cabriolet fiel mit lautem Klatschen in den Bach" is converted using grammatical rules (such as stemming, conjugation, etc.) and employing German synonym classes to:
“blau Auto fallen gross Lärm Wasser”

A machine translation of this will provide: "blue car fall big noise water".

The English sentence “The deep-blue limousine dropped with a big splash into the river” is converted using grammatical rules (like reducing to singular, nominative, infinitive, etc.) and synonym classes to: “blue car fall big noise water”

Bingo! The two sentences have been proven to be similar!

Another functionality lacking in existing systems is the ability to process textual images for similarity checks. Some times one has to deal with textual information in scanned format. Most of such images contain text in typed form which can be very accurately converted to text with the use of Optical Character Recognition (OCR) engines.

The missing components in existing systems also include better tabular information processing, proper support for foreign language characters, reference validity and relevance checks. It is likely that high quality services for plagiarism detection will have to combine a set of methods as described above.

8 Conclusion

It is fair to say, that current plagiarism detection tools work reasonably well on textual information that is available on the internet or in other electronic sources. They do break down:

(1) When systematic attempts are made to combat plagiarism tools by e.g. using extensive paraphrasing with the help of synonymising tools, syntactic variations or different expressions for same contents. (NOTE: most of the better systems are stable against the order in which paragraphs are arranged: fingerprinting is usually not done on a sequence but on a set of data, hence order does not matter)

(2) When plagiarism is based on documents that are not available electronically (Since they only are available in printed form, or in archives that are not accessible for the tool used)

(3) When plagiarism crosses language boundaries.

Of the three points mentioned above there is hope concerning item (2): more and more material is being digitized, and some tools have managed to get access to hidden material in paper mills and such. Item (3) will be challenge for some time to come. We believe that most headway can be achieved in connection with point (1) by using a multiphase approach:

Observe that we have mentioned that the similarity check of a small set of documents is possible using rather deep techniques that can determine conceptual equivalence even when heavy paraphrasing is used. However, those techniques break down if the volume of data becomes too large. Hence we think that the way to obtain a successful system that determines whether a particular document x is plagiarized will have to work as follows:

A fast algorithm scans the whole available docuverse (the set of all available documents) and eliminates all documents that 'clearly' have not been used for the document x at issue.

The remaining much smaller docuverse is now scanned by a better algorithm to again reduce the size of the set of still possible sources used for plagiarism. This continues, until a 'fairly small set' of documents remain for which it is feasible to use deep and computing intensive techniques.

Whether the number of 'passes' should be 2, 3 or more remains to be seen. Since all major plagiarism tools are proprietary it is not known to us how much this multi-pass technique is already in use. It is clear for us from the observations we have, however, that there is much room for further progress.

In closing we want to mention two further important points to which we return in [Zaka & Maurer 2006]:

First, plagiarism is not confined to academia. It is rampant and still not much recognized in schools, particularly in high schools where many assignments are of the general essay type, exactly the kind of stuff easily found on the internet. It also appears in a different form when government agencies or other organisations commission some 'study' or report to be compiled: in a number of cases they get what they want, pay quite some money for it, but what they get is just obtained by simply copying and pasting and minor changes or additions of existing material. In those cases it is not so much a question to detect plagiarism after the fact, but rather have some specialists spend a few hours searching on the net if the material requested it not available anyway before commissioning a report.

Second, plagiarism is getting lots of attention in academia right now. The reaction has been that many universities purchase tools for plagiarism detection. It is our belief that to detect plagiarism at a university you need more than a software tool: you need a set of them, specialists who know how to work with those tools, domain experts and also language experts if we ever want to go beyond the boundary of one language. This implies that a substantial group is necessary to do good work, and this cannot be achieved by any one university. It requires a joint effort i.e. a center for plagiarism detection that is run on a national or even supra-national (e.g. European) level.

References

[Band 2006] Jonathan Band, "The Google Library Project: Both Sides of the Story", Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification, 1 (2): 1-17. 2006

[Beasley 2006] James Douglas Beasley, "The Impact of Technology on Plagiarism Prevention and Detection" Plagiarism: Prevention, Practice and Policies 2004 Conference.

[Berkeley 2006] Berkeley University of California, Student Conduct, Sanctions, http://students.berkeley.edu/osl/sja.asp?id=1004, visited: 22 July 2006

[CAI 2005] The Center for Academic Integrity's Assessment Project Research survey by Don McCabe, http://www.academicintegrity.org/cai_research.asp, visited: 22 July 2006

[Cambridge 2006] University of Cambridge, Procedure for dealing with cases of suspected plagiarism , http://www.admin.cam.ac.uk/offices/gradstud/committees/plagiarism/procedure.html, visited: 22 July 2006

[CNN 2004] CNN COURTTV; Student wins battle against plagiarism detection requirement, January 2004, http://www.cnn.com/2004/LAW/01/21/ctv.plagiarism/, visited: 22 July 2006

[CopyCatch 2006] CopyCatch product website, http://www.copycatchgold.com/, visited: 22 July 2006

[Dreher & Williams 2006] Heinz Dreher and Robert Williams, "Assisted Query Formulation Using Normalised Word Vector and Dynamic Ontological Filtering" Flexible Query Answering Systems: 7th International Conference, FQAS 2006, Milan, Italy, June 7-10, 2006 pp. 282 – 294

[Eissen & Stein 2006] Sven Meyer zu Eissen and Benno Stein. Intrinsic Plagiarism Detection, To appear in the Proceedings of the European Conference on Information Retrieval (ECIR-06), Springer, 2006.

[EVE 2006] EVE Plagiarism Detection System website, http://www.canexus.com/eve/, visited: 22 July 2006

[Glatt 2006] Glatt Plagiarism Services website, http://www.plagiarism.com/, visited: 22 July 2006

[HEC Press 2006] Higher Education Commission Pakistan Press release (7 Feb. 2006), http://www.hec.gov.pk/htmls/press_release/2006/Feb/feb_6.htm and (10 May 2006), http://www.hec.gov.pk/htmls/press_release/May_06/May-10.htm, visited: 22 July 2006

[Heinrich & Maurer 2000] E. Heinrich, H. Maurer, "Active Documents: Concept, Implementation and Applications" J.UCS 6, 12 (2000), 1197-1202

[iParadigm 2006] iParadigms anti plagiarism product website, http://www.plagiarism.org/, visited: 22 July 2006

[IPR overview 2006] Intellectual Property Rights: Overview, March 2006, http://www.jisclegal.ac.uk/pdfs/IPROverview.pdf, visited: 22 July 2006

[Iyer & Singh 2005] Parvati Iyer and Abhipsita Singh, "Document Similarity Analysis for a Plagiarism Detection Systems" 2nd Indian International Conference on Artificial Intelligence (IICAI –2005), pp. 2534-2544

[JISC 2006] Joint Information Systems Committee (JISC) plagiarism Advisory Program website, http://www.jiscpas.ac.uk/, visited: 22 July 2006

[JPlag 2006] JPlag website, https://www.ipd.uni-karlsruhe.de/jplag/ visited: 22 July 2006

[Levin 1993] Levin. 1993. English Verb Classes and Alternations. A Preliminary Investigation. University of Chicago Press.

[LSA 2006] Latent Semantic Analysis, web site at University of Colorado, http://lsa.colorado.edu/, visited: 22 July 2006

[Manber 1994] Udi Manber. "Finding similar files in a large file system" Winter USENIX Technical Conference 1994, San Francisco, CA, USA

[Maurer et al. 2006] Hermann Maurer, Harald Krottmaier, Heinz Dreher, “Important Aspects of Digital Libraries“: International Conference of Digital Libraries, New Delhi, Dec.5-8, 2006, to appear

[MIT News 2003] MIT, (March 25, 2003), “Budget projections, student discipline report presented to faculty”, http://web.mit.edu/newsoffice/2003/facmeet.html, visited: 22 July 2006

[MIT policies 2006] Massachusetts Institute of Technology Policies and Procedures, http://web.mit.edu/policies/10.0.html, visited: 22 July 2006

[MIT Writing 2006] MIT Online Writing and Communication Center, http://web.mit.edu/writing/, visited: 22 July 2006

[MOSS 2006] MOSS, A System for Detecting Software Plagiarism website, http://www.cs.berkeley.edu/~aiken/moss.html, visited: 22 July 2006

[Mydropbox 2006] Mydropbox, SafeAssignment Product Brochure, http://www.mydropbox.com/info/SafeAssignment_Standalone.pdf, visited: 22 July 2006

[Niezgoda & Way 2006] Sebastian Niezgoda and Thomas P. Way. "SNITCH: a Software Tool for Detecting Cut and Paste Plagiarism." SIGCSE Technical Symposium (SIGCSE 2006), March 2006.

[Oxford Gazette 2005] Oxford University Gazette, (23 March 2005), http://www.ox.ac.uk/gazette/2004-5/supps/1_4728.htm, visited: 22 July 2006

[Plagiarism.org 2006] Research resources at plagiarism.org, http://www.plagiarism.org/research_site/e_what_is_plagiarism.html, visited: 22 July 2006

[Rutgers 2003] Study at Rutgers Confirms Internet Plagiarism Is Prevalent, http://ur.rutgers.edu/medrel/viewArticle.html?ArticleID=3408, visited: 22 July 2006

[Schleimer et al. 2003] S. Schleimer, D. S. Wilkerson, and A. Aiken. Winnowing: local algorithms for document fingerprinting. In SIGMOD: Proceedings of the 2003

[Stanford Copyright 2006] Copyright and Fair use portal at Stanford University, http://fairuse.stanford.edu/, visited: 22 July 2006

[Stanford Daily 2003] The Stanford Daily, Feb. 12, 2003 By Ali Alemozafar, http://daily.stanford.edu/article/2003/2/12/onlineSoftwareBattlesPlagiarismAtStanford, visited: 22 July 2006

[Stanford Honorcode 1921] Stanford Honor Code, http://www.stanford.edu/dept/vpsa/judicialaffairs/guiding/pdf/honorcode.pdf, visited: 22 July 2006

[Turnitin tour 2006] Plagiarism Tour at turnitin.com, http://www.turnitin.com/static/flash/tii.html, visited: 22 July 2006

[Urkund 2006] Urkund website, http://www.urkund.com/ visited: 22 July 2006

[Uzuner et al. 2005] Özlem Uzuner, Boris Katz, and Thade Nahnsen, "Using Syntactic Information to Identify Plagiarism" Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, pages 37–44, Ann Arbor, June 2005. Association for Computational Linguistics.

[Wcopyfind 2006] WCopyfind website.
http://plagiarism.phys.virginia.edu/Wsoftware.html, visited: 22 July 2006

[Wikipedia:Kent 2006] University of Kent. (2006, July 20). In Wikipedia,
http://en.wikipedia.org/w/index.php?title=University_of_Kent&oldid=64849655, visited: July 25, 2006

[Wikipedia:papermill 2006] Paper mill (essays). (2006, July 16). In Wikipedia, The Free Encyclopedia. Retrieved 09:13, July 25, 2006, from
http://en.wikipedia.org/w/index.php?title=Paper_mill_%28essays%29&oldid=64074352.

[Wikipedia:Plagiarism 2006] Plagiarism. In Wikipedia, The Free Encyclopedia. Retrieved 09:11, 22 July 2006, from
http://en.wikipedia.org/w/index.php?title=Plagiarism&oldid=65284248

[Williams 2006] Robert Williams, “The Power of Normalised Word Vectors for Automatically Grading Essays” The Journal of Issues in Informing Science and Information Technology Volume 3, 2006, pp. 721-730

[Yale 2005] Yale College Executive Committee Yearly Chair Reports,
http://www.yale.edu/yalecol/publications/executive/index.html, visited: 22 July 2006

[Zaka & Maurer 2006] Bilal Zaka and Hermann Maurer, “Plagiarism: Should and can we Fight against it?”, to appear