Automatic Conceptual Analysis for Plagiarism Detection

Heinz Dreher
School of Information Systems, Curtin University of Technology
Perth, Western Australia

h.dreher@curtin.edu.au

Abstract

In order to detect plagiarism, comparisons must be made between a target document (the suspect) and reference documents. Numerous automated systems exist which check at the text-string level. If the scope is kept constrained, as for example in within-cohort plagiarism checking, then performance is very reasonable. On the other hand if one extends the focus to a very large corpus such as the WWW then performance can be reduced to an impracticable level. The three case studies presented in this paper give insight into the text-string comparators, whilst the third case study considers the very new and promising conceptual analysis approach to plagiarism detection which is now made achievable by the very computationally efficient Normalised Word Vector algorithm. The paper concludes with a caution on the use of high-tech in the absence of high-touch.

Keywords: academic malpractice, conceptual analysis, conceptual footprint, semantic footprint, Normalised Word Vector, NWV, plagiarism.

Introduction

Plagiarism is now acknowledged to pose a significant threat to academic integrity. There is a growing array of software packages to help address the problem. Most of these offer a string-of-text comparison. New to emerge are software packages and services to 'generate' assignments. Naturally there will be a cat and mouse game for a while and in the meantime academics need to be alert to the possibilities of academic malpractice via plagiarism and adopt appropriate and promising counter-measures, including the newly emerging algorithms to do fast conceptual analysis. One such emergent agent is the Normalised Word Vector (NWV) algorithm (Williams, 2006), which was originally developed for use in the Automated Essay Grading (AEG) domain.

AEG is a relatively new technology which aims to score or grade essays at the level of expert humans. This is achieved by creating a mathematical representation of the semantic information in addition to checking spelling, grammar, and other more usual parameters associated with essay assessment. The mathematical representation is computed for each student essay and compared with a mathematical representation computed for the model answer. If we can represent the semantic content of an essay we are able to compare it to some standard model- hence determine a grade or assign an authenticity parameter relative to any given

corpus; and create a persistent digital representation of the essay.

AEG technology can be used for plagiarism detection because it processes the semantic information of student essays and creates a *semantic footprint*. Once a mathematical representation for all or parts of an essay is created it can be efficiently compared to other similarly constructed representations and facilitate plagiarism checking through *semantic footprint* comparison.

The Plagiarism Problem

The extent of plagiarism is indeed significant. Maurer et al. (2006) provide a thorough analysis of the plagiarism problem and possible solutions as they pertain to academia. They divide the solution strategies into three main categories. The most common method is based on document comparison in which a word for word check is made with each target document in a selected which could be the source of the copied material. Clearly this is language independent as one is essentially comparing character strings; it will also match misspellings. The selected set of document is usually all documents comprising assignment or paper submissions for a specific purpose. A second category is an expansion of the document check but where the set of target documents is 'everything' that is reachable on the internet and the candidate to be checked for is a characteristic paragraph or sentence rather than the entire document. The emergence of tools such as Google has made this type of check feasible. The third category mentioned by Maurer et al. is the use of stylometry, in which a language analysis algorithm compares the style of successive paragraphs and reports if a style change has occurred. This can be extended to analyzing prior documents by the same author and comparing the stylistic parameters of a succession of documents.

However, the issue of plagiarism is not merely a matter for academics. Austrian journalist Josef Karner (2001) writes "Das Abschreiben ist der eigentliche Beruf des Dichters" ("Transcription is the virtual vocation of the poet"). Is then the poet essentially a professional plagiarist, taking others' ideas and presenting them in verse as his own and without attribution? This may be a rather extreme position to hold, but its consideration does point up interesting possibilities which the etymology of plagiarism may illuminate.

As yet there is a paucity of statistics available to help us understand the extent of plagiarism. However a recent Canadian study (Kloda & Nicholson, 2005) has reported that one in three students admit to turning to plagiarism prior to graduation - serious enough one may think. The truly shocking statistic is that one in 20 has actually paid for someone else to write or provide an assessment paper which they subsequently submitted as their own (**Figure 1**).

# of Canadian students who admit to plagiarizing at least once before graduating = 1/3

# of Canadian students who admit to submitting a paper they had purchased online as their own = 1/20

# of Canadian universities and colleges that subscribe to Turnitin™ = 28/90 (31%) or 28/130 (22%)

Turnitin™ is one of the well established plagiarism checking systems

Figure 1: Plagiarism statistics (Kloda & Nicholson, 2005)

Despite having plagiarism detection technology available, its effective implementation can be a challenge in itself. In one prominent case where technology was forced on students, the reaction led to a court ruling granting the student the right to bypass a university mandated plagiarism check prior to assignment submission (**Figure 2**).

A student at McGill University has won the right to have his assignments marked without first submitting them to an American, anti-plagiarism website.

Jesse Rosenfeld refused to submit three assignments for his second-year economics class to Turnitin.com, a website that compares submitted works to other student essays in its database, as well as to documents on the web and published research papers.

Last Updated: Friday, January 16, 2004 | 11:11 AM ET

Figure 2: McGill student wins fight over anti-cheating website

Source: http://www.cbc.ca/canada/story/2004/01/16/mcgill_turnitin030116.html

To help effective plagiarism detection implementation in educational institutions around the world, advice for students and staff is readily available on a growing number of plagiarism-dedicated web-based resources – a sample appears in **Figure 3**.

Australia	http://www.lc.unsw.edu.au/onlib/plag.html http://academichonesty.unimelb.edu.au/ http://startup.curtin.edu.au/study/plagiarism.html http://www.teachers.ash.org.au/aussieed/research_plagiarism.htm
Austria	http://www.iaik.tugraz.at/aboutus/people/poschkc/EinfuehrungInDieTelematik.htm http://ipaweb.imw.tuwien.ac.at/bt/index.php?id=aktuelles
Canada	http://library.acadiau.ca/tutorials/plagiarism/ http://www.library.ualberta.ca/guides/plagiarism/ http://www.ucalgary.ca/~hexham/study/plag.html
Germany	http://plagiat.fhtw-berlin.de/ http://www.frank-schaetzlein.de/biblio/plagiat.htm http://www.spiegel.de/unispiegel/studium/0,1518,227828,00.html
USA	http://plagiarism.phys.virginia.edu/ ; http://www.umuc.edu/distance/odell/cip/links_plagiarism.html http://www.ece.cmu.edu/~ee240/)

Figure 3: Sample plagiarism resources

Whilst the plagiarism problem is significant, it is not solvable only by applying plagiarism detection techniques. There needs to be a recognition that the students are not entirely to blame (WilliamsJ 2002). Quite obviously we need to agree on a working definition of plagiarism which is simple to understand and to check.

In a light-hearted vein, the entry for plagiarism in The Devil's Dictionary by Ambrose Bierce reads: PLAGIARISM, n.
A literary coincidence compounded of a discreditable priority and an honorable subsequence.

This might be the sort of definition which would be used to justify excusing a first or minor instance of plagiarism but it does not admit of the measures which may be needed to detect it. A more precise and practically applicable definition, that indicates the measures which may be needed to detect plagiarism, is found on the www.plagiarism.org site:

- copying words or ideas from someone else without giving credit
- 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 (see our section on "fair use" rules)

From the above we can see the essential elements: words; style or structure; and ideas. Therefore, checking systems must look for matching words, analyze style, and create a map of the ideas con-

tained in candidate plagiarism cases. The first of these is well catered for by the established systems, such as string-of-text matching.

Established Plagiarism Checkers

As can be seen from Maurer et al. (2006), there are many systems doing the string-of-text matching. Here we briefly consider the performance of two of them which were readily available to the author – WCopyfind and EVE2.

Case 1: WCopyfind

The University of Virginia's freely available WCopyfind software (http://plagiarism.phys.virginia.edu) is a delightful example of the power of the computer to help in addressing the plagiarism problem. It makes text-string comparisons and can be instructed to find sub-string matches of given length and similarity characteristics. Such fine tuning permits the exclusion of obvious non-plagiarism cases despite text-string matches.

To determine the efficacy of WCopyfind the author devised a trial. Some 600 student assignments from a course on Societal Impacts of Information Technology were checked for within-cohort plagiarism. The assignments were between 500 and 2000 words and were either in the English or German language. The system is computationally very efficient and took only seconds to highlight five cases requiring closer scrutiny.

Figure 4, **Figure 5**, and **Figure 6** show WCopyfind – system interface, WCopyfind – report, and WCopyfind – document comparison, respectively.

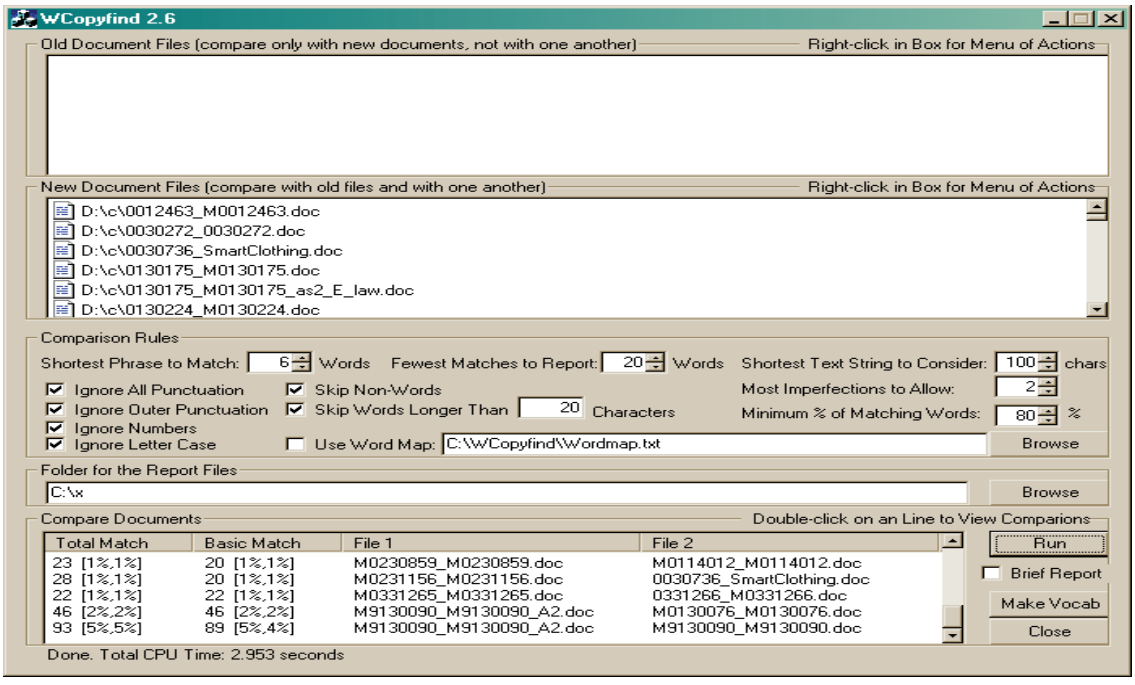
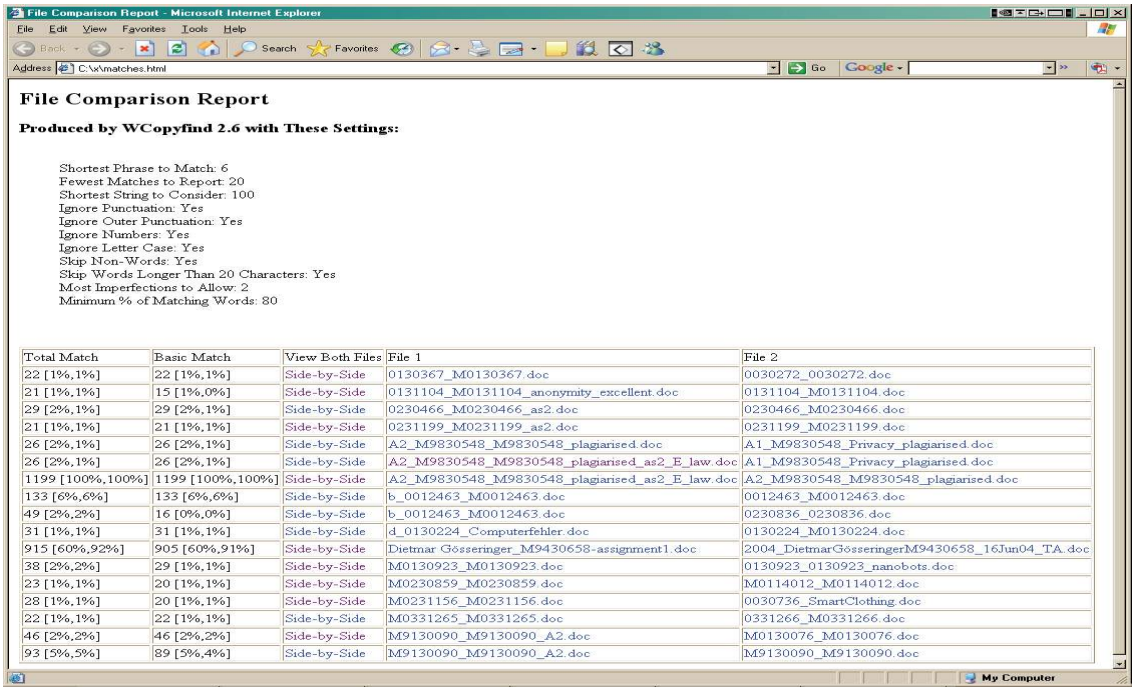


Figure 4: WCopyfind – system interface

Since the WCopyfind works at the string-of-text level, language is unimportant and matches are readily identified from the candidate documents submitted for analysis. Note that such a procedure cannot find plagiarism based on documents not submitted, for example Web resident documents. Of course, further analysis of a small subset can be submitted for Web-based document comparison with Google for example. In this case a sample of the identified within-cohort plagia-

rized text was submitted for a Google search and immediately revealed a source on the Web containing the same text (**Figure 7**).



File Comparison Report - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address C:\x\matches.html

File Comparison Report

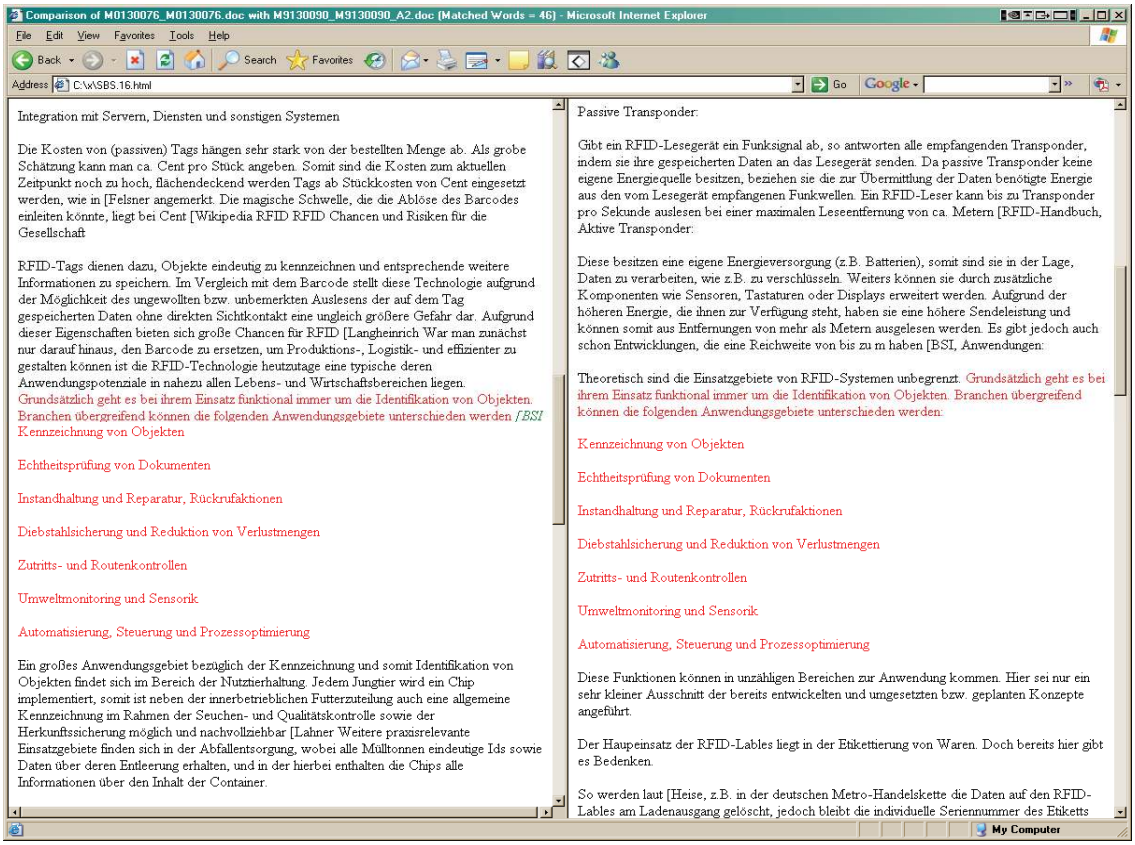
Produced by WCopyfind 2.6 with These Settings:

Shortest Phrase to Match: 6
Fewest Matches to Report: 20
Shortest String to Consider: 100
Ignore Punctuation: Yes
Ignore Outer Punctuation: Yes
Ignore Numbers: Yes
Ignore Letter Case: Yes
Skip Non-Words: Yes
Skip Words Longer Than 20 Characters: Yes
Most Imperfections to Allow: 2
Minimum % of Matching Words: 80

Total Match	Basic Match	View Both Files	File 1	File 2
22 [1%,1%]	22 [1%,1%]	Side-by-Side	0130367_M0130367.doc	0030272_0030272.doc
21 [1%,1%]	15 [1%,0%]	Side-by-Side	0131104_M0131104_anonymity_excellent.doc	0131104_M0131104.doc
29 [2%,1%]	29 [2%,1%]	Side-by-Side	0230466_M0230466_as2.doc	0230466_M0230466.doc
21 [1%,1%]	21 [1%,1%]	Side-by-Side	0231199_M0231199_as2.doc	0231199_M0231199.doc
26 [2%,1%]	26 [2%,1%]	Side-by-Side	A2_M9830548_M9830548_plagiarised.doc	A1_M9830548_Privacy_plagiarised.doc
26 [2%,1%]	26 [2%,1%]	Side-by-Side	A2_M9830548_M9830548_plagiarised_as2_E_law.doc	A1_M9830548_Privacy_plagiarised.doc
1199 [100%,100%]	1199 [100%,100%]	Side-by-Side	A2_M9830548_M9830548_plagiarised_as2_E_law.doc	A2_M9830548_M9830548_plagiarised.doc
133 [6%,6%]	133 [6%,6%]	Side-by-Side	b_0012463_M0012463.doc	0012463_M0012463.doc
49 [2%,2%]	16 [0%,0%]	Side-by-Side	b_0012463_M0012463.doc	0230836_0230836.doc
31 [1%,1%]	31 [1%,1%]	Side-by-Side	d_0130224_Computerfehler.doc	0130224_M0130224.doc
915 [60%,92%]	905 [60%,91%]	Side-by-Side	Dietmar Gösseringer_M9430658-assignment1.doc	2004_DietmarGösseringerM9430658_16Jun04_TA.doc
38 [2%,2%]	29 [1%,1%]	Side-by-Side	M0130923_M0130923.doc	0130923_0130923_nanobots.doc
23 [1%,1%]	20 [1%,1%]	Side-by-Side	M0230859_M0230859.doc	M0114012_M0114012.doc
28 [1%,1%]	20 [1%,1%]	Side-by-Side	M0231156_M0231156.doc	0030736_SmartClothing.doc
22 [1%,1%]	22 [1%,1%]	Side-by-Side	M0331265_M0331265.doc	0331266_M0331266.doc
46 [2%,2%]	46 [2%,2%]	Side-by-Side	M9130090_M9130090_A2.doc	M0130076_M0130076.doc
93 [5%,5%]	89 [5%,4%]	Side-by-Side	M9130090_M9130090_A2.doc	M9130090_M9130090.doc

My Computer

Figure 5: WCopyfind – report



Comparison of M0130076_M0130076.doc with M9130090_M9130090_A2.doc (Matched Words = 46) - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address C:\x\SBS.16.html

Integration mit Servern, Diensten und sonstigen Systemen

Die Kosten von (passiven) Tags hängen sehr stark von der bestellten Menge ab. Als grobe Schätzung kann man ca. Cent pro Stück angeben. Somit sind die Kosten zum aktuellen Zeitpunkt noch zu hoch, flächendeckend werden Tags ab Stückkosten von Cent eingesetzt werden, wie in [Felsner angemerkt. Die magische Schwelle, die die Ablöse des Barcodes einleiten könnte, liegt bei Cent [Wikipedia RFID RFID Chancen und Risiken für die Gesellschaft

RFID-Tags dienen dazu, Objekte eindeutig zu kennzeichnen und entsprechende weitere Informationen zu speichern. Im Vergleich mit dem Barcode stellt diese Technologie aufgrund der Möglichkeit des ungewollten bzw. unbemerkten Auslesens der auf dem Tag gespeicherten Daten ohne direkten Sichtkontakt eine ungleich größere Gefahr dar. Aufgrund dieser Eigenschaften bieten sich große Chancen für RFID [Langheinrich War man zunächst nur darauf hinaus, den Barcode zu ersetzen, um Produktions-, Logistik- und effizienter zu gestalten können ist die RFID-Technologie heutzutage eine typische deren Anwendungspotenziale in nahezu allen Lebens- und Wirtschaftsbereichen liegen. Grundsätzlich geht es bei ihrem Einsatz funktional immer um die Identifikation von Objekten. Branchen übergreifend können die folgenden Anwendungsgebiete unterschieden werden *[BSI* Kennzeichnung von Objekten

Echtheitsprüfung von Dokumenten

Instandhaltung und Reparatur, Rückrufaktionen

Diebstahlsicherung und Reduktion von Verlustmengen

Zutritts- und Routenkontrollen

Umweltmonitoring und Sensorik

Automatisierung, Steuerung und Prozessoptimierung

Ein großes Anwendungsgebiet bezüglich der Kennzeichnung und somit Identifikation von Objekten findet sich im Bereich der Nutztierhaltung. Jedem Jungtier wird ein Chip implementiert, somit ist neben der innerbetrieblichen Futterzuteilung auch eine allgemeine Kennzeichnung im Rahmen der Seuchen- und Qualitätskontrolle sowie der Herkunftssicherung möglich und nachvollziehbar [Lahner Weitere praxisrelevante Einsatzgebiete finden sich in der Abfallentsorgung, wobei alle Mülltonnen eindeutige Ids sowie Daten über deren Entleerung erhalten, und in der hierbei enthalten die Chips alle Informationen über den Inhalt der Container.

Passive Transponder:

Gibt ein RFID-Lesegerät ein Funksignal ab, so antworten alle empfangenden Transponder, indem sie ihre gespeicherten Daten an das Lesegerät senden. Da passive Transponder keine eigene Energiequelle besitzen, beziehen sie die zur Übermittlung der Daten benötigte Energie aus den vom Lesegerät empfangenen Funkwellen. Ein RFID-Leser kann bis zu Transponder pro Sekunde auslesen bei einer maximalen Leseentfernung von ca. Metern [RFID-Handbuch, Aktive Transponder:

Diese besitzen eine eigene Energieversorgung (z.B. Batterien), somit sind sie in der Lage, Daten zu verarbeiten, wie z.B. zu verschlüsseln. Weiters können sie durch zusätzliche Komponenten wie Sensoren, Tastaturen oder Displays erweitert werden. Aufgrund der höheren Energie, die ihnen zur Verfügung steht, haben sie eine höhere Sendeleistung und können somit aus Entfernungen von mehr als Metern ausgelesen werden. Es gibt jedoch auch schon Entwicklungen, die eine Reichweite von bis zu m haben [BSI, Anwendungen:

Theoretisch sind die Einsatzgebiete von RFID-Systemen unbegrenzt. Grundsätzlich geht es bei ihrem Einsatz funktional immer um die Identifikation von Objekten. Branchen übergreifend können die folgenden Anwendungsgebiete unterschieden werden:

Kennzeichnung von Objekten

Echtheitsprüfung von Dokumenten

Instandhaltung und Reparatur, Rückrufaktionen

Diebstahlsicherung und Reduktion von Verlustmengen

Zutritts- und Routenkontrollen

Umweltmonitoring und Sensorik

Automatisierung, Steuerung und Prozessoptimierung

Diese Funktionen können in unzähligen Bereichen zur Anwendung kommen. Hier sei nur ein sehr kleiner Ausschnitt der bereits entwickelten und umgesetzten bzw. geplanten Konzepte angeführt.

Der Haupeinsatz der RFID-Lables liegt in der Etikettierung von Waren. Doch bereits hier gibt es Bedenken.

So werden laut [Heise, z.B. in der deutschen Metro-Handelskette die Daten auf den RFID-Lables am Ladenausgang gelöscht, jedoch bleibt die individuelle Seriennummer des Etiketts

My Computer

Figure 6: WCopyfind – document comparison

It is interesting but perhaps not surprising to note that those who plagiarize from fellow students will also copy from elsewhere (personal experience). The analysis thus far has not proven plagiarism but simply highlighted its possible existence and located the evidence. Simply because text-strings match does not permit one to conclude plagiarism, as the text may be properly referenced. The suspect text was found in document www.bsi.de/fachthem/rfid/RIKCHA.pdf (**Figure 7**) and can now be carefully matched with student text to determine the extent and accuracy of the copying. In short, WCopyfind is one text-string-matching approach to plagiarism detection that is useful for within-cohort applications, but is not amenable to large scale 'extra-cohort' plagiarism detection (i.e., searching the WWW). Case study 2 investigates one program that is designed for this purpose.

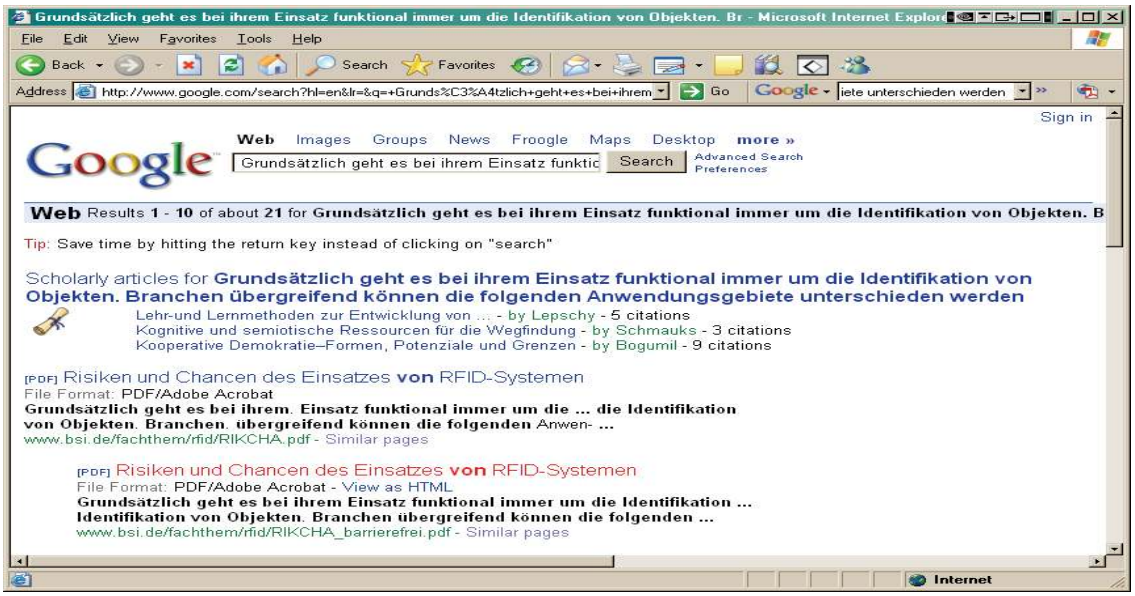


Figure 7: Google finds matching Web document

Case 2: EVE2

The Essay Verification Engine (EVE2) by CaNexus.com makes a reliable check of the Internet to track down possible instances of plagiarism. It examines the essays, and then quickly makes a large number of searches of the Internet to locate suspect sites. EVE2 visits each of these suspect sites to determine if they contain work that matches the essay in question. It all sounds rather simple and straightforward and in a testimonial one reads: "*...EVE aced the test, finding everything I had plagiarized.* EVE is faster, testing four papers in fifteen minutes, a fraction of the four hours it took Plagiarism.org to respond." (excerpted and adapted from the EVE2 website http://www.canexus.com/eve/).

Naturally such claims are encouraging but will it really work so well in my case? That's the real question. Given the 'speed' claim above the author decided to submit 16 pages comprising some 7,300 words of a Master's thesis which was in preparation. It was a chapter on "previous work and literature review" so one would expect to find some matches.

The first observation was that EVE2 took circa 20 minutes to complete the task. This is not fast at all; rather, it is so slow that one could not check but a few carefully selected items at this rate. **Figure 8** shows the computer's CPU usage during the analysis.

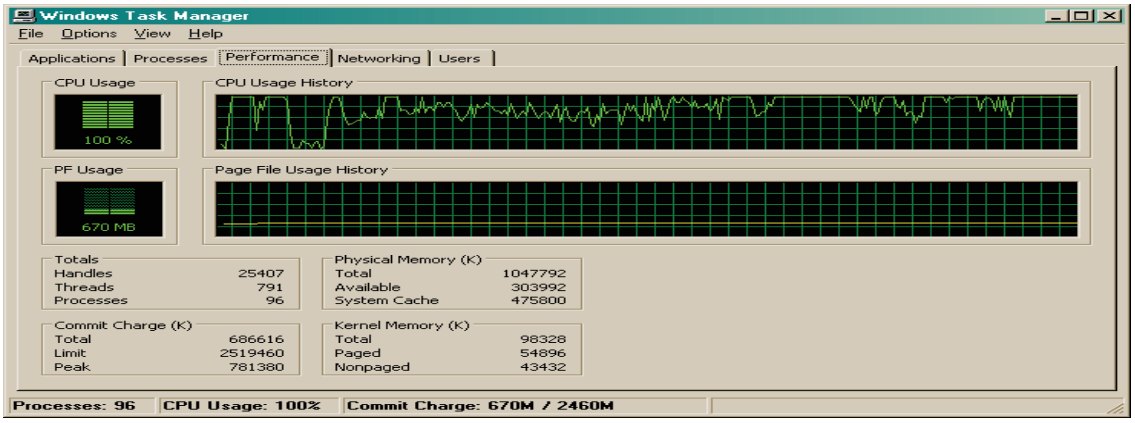


Figure 8: EVE2 computational demand

The result was 'disappointing' too – EVE2 only flagged a low level of potential plagiarism most of which was due to legitimate referencing and flagged two websites (**Figure 9**). On the other hand one is delighted that one's research students are creating their own work!

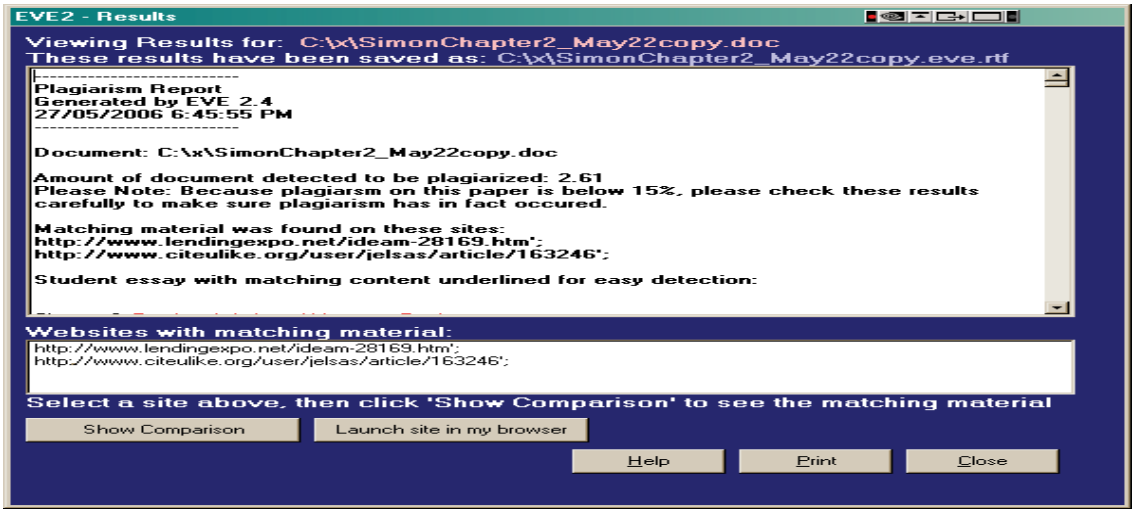


Figure 9: EVE2 result for Master's thesis chapter

The Limitation of String-Of-Text Plagiarism Checking

Finding exact matches of text is, in principle, a simple and straightforward matter, and having computer support certainly makes it possible to undertake checking on a large scale. However one usually does not want everything checked against everything else since there will always be some legitimate text matches. In the two case analyses above the author has found that checking at the level of text-strings can be useful, especially as the process is language independent (although translations of plagiarized text will not be revealed), and very fast if the scope is restricted (Case 1).

In Case 2 the processing required to check thousands of words in the source with billions+ of documents on the web is obviously impractical in the usual situation and would have to be reserved for special cases where the outcome is critical.

Typically however, further and usually laborious analysis and detection will be required and human intervention and attention cannot be spared. There are some cases which are rather obvious

to humans but not so simply detected automatically. Consider the assignment fragment in **Figure 10**. These words appeared in an assignment submitted by a student doing a capstone course in Information Systems & Technology.

> Web sites involves a mixture development between print publishing and software development, between marketing and computing, between internal communications and external relations, and between art and technology. Software engineering provides processes that are useful in developing the web sites and web site engineering priniciples can be used to help bring web projects under control and minimize the risk of a project being delivered late or over budget.

Figure 10: Excerpt from student essay (coursework degree)

As one reads the words one senses a rather unsophisticated level of expression up until the first comma, then the reader is suddenly confronted with a rather engaging triplet of comparisons to 'complete' the sentence. The second and final sentence reverts to the banality with which the paragraph began.

Looking at the context of the triplet of comparisons ("between marketing and computing, between internal communications and external relations, and between art and technology") from **Figure 10** it can be seen that these words do not really integrate into the sentence naturally – somehow it seems contrived. Such an instance can spark further investigation and in this case led to the identification of a published article which used these very words. Interestingly, and not at all surprisingly, this article was the source of considerable chunks of un-attributed replicated text in the student assignment.

Does automated support for more sophisticated processing of essays, perhaps at the semantic level exist and which may complement the string-of-text analyses considered above?

Before turning our attention to this, one should mention here that of the three approaches to the plagiarism problem identified at the outset – namely: words; style or structure; and ideas – the author has covered the first, and merely makes a reference to the possibility of style analysis, leaving its treatment to a future paper.

Turning our attention now to the type of processing done by humans, i.e. idea processing, we consider applying the new and very computationally efficient Normalised Word Vector (NWV) algorithm (WilliamsR, 2006) to the task of plagiarism detection.

Conceptual Analysis for Essay Grading

Considerable knowledge and intelligence is needed to detect plagiarism at the concept or idea level irrespective of which actual words are used to express the idea. Humans are able to do this very well of course, naturally one might say, but are limited severely by capacity. Computers can find text-string matches but cannot readily prioritize the cases for manual checking (this is a well known problem in searching – one receives millions of hits within fractions of a second), which clearly leaves a gap to be filled by some more sophisticated technology.

The NWV algorithm was devised by Bob Williams for the Automated Essay Grading system MarkIT developed by Bob Williams and Heinz Dreher at Curtin University of Technology in Perth, Western Australia (www.essaygrading.com). In essence the approach is to use vector algebra to represent similarities in content between documents. A thesaurus is used to *normalise* words by reducing selected words from essay to a thesaurus root word, and deriving occurrence frequency measures to create a vector representation. This mathematical representation is a *conceptual footprint* of the essay and can be used for comparison purposes including plagiarism checking at the semantic level.

The next section of the paper presents some case analyses using a promising new technology to aid in plagiarism detection – the use of the Normalised Word Vector (NWV) algorithm to create a *conceptual footprint* of student assignments.

Case 3: Conceptual Analysis Using NWV

In this case study the author has selected the 51 English language essays from the set used in Case 1. Since the thesaurus currently being used by the NWV is an English language thesaurus it follows that only English language essays can be processed.

Firstly, note that the time taken to process these 51 essays, which is to do a conceptual analysis, was three minutes (**Figure 11**).

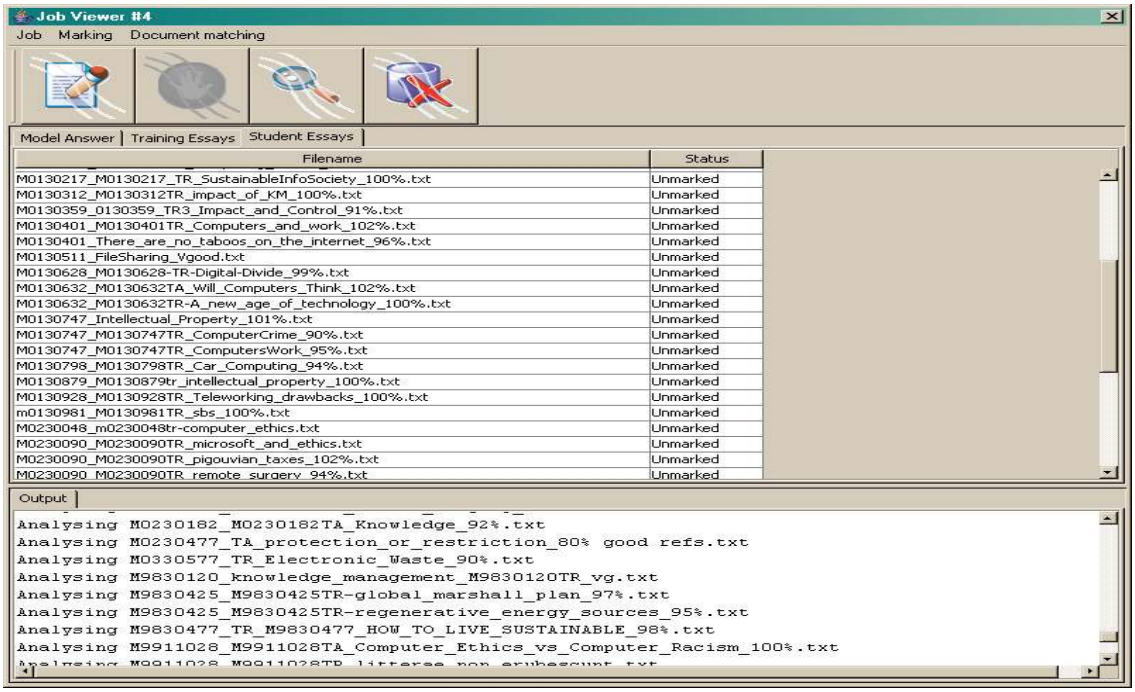


Figure 11: Concept analysis of 51 essays

Next, look at the *conceptual footprint* matching, by considering a comparison of an assignment entity X with itself. In this case X = assignment M0130097tr written on a topic in computer security. Inspection of **Figure 12** reveals the desired, and expected, perfect match on a concept by concept basis. Note how the un-normalised words highlighted in the text (right panel, upper compared with middle) match and how the bar graphs match in the lower panel). The left side of the screen provides a map of the objects being conceptually compared – document files containing the assignments, and paragraphs within documents, with matching *%closeness*. The object document under consideration obviously has 100% closeness.

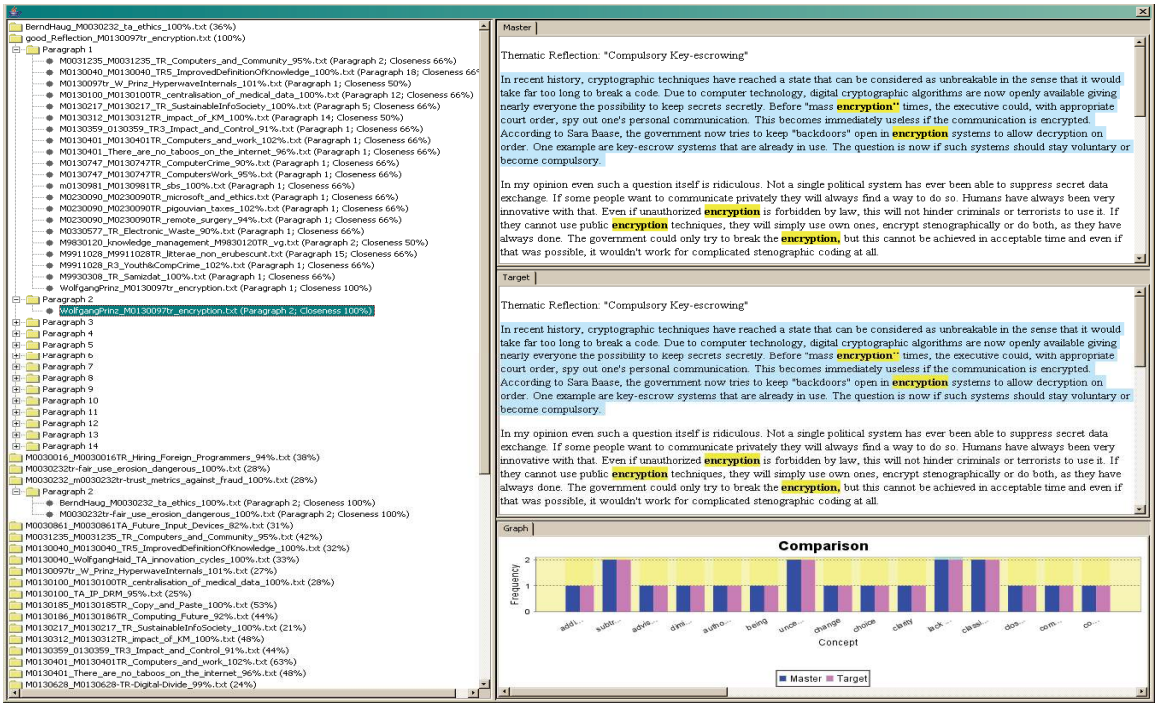


Figure 12: Conceptual analysis of X with X – 100% closeness

The user, the plagiarist detective or plagiarism analyst, can focus on particular concepts (with matching thesaurus root definitions). Observe that using this visual feedback one could engage in quite a meaningful and persuasive discussion with a 'plagiarist' for the purpose of establishing the presence and level of copying.

A focus on concept by concept reveals the lowest level of granularity available with our system. In the example shown in **Figure 13** the concept with name "hiding" is being scrutinized. The thesaurus entry is shown on the left and any words belonging to the thesaurus entry for "hiding", which appear in either of the documents of interest, are shown in the upper right panels.

Obviously such detailed scrutiny even with automated support is time consuming and would be reserved for special and few cases of intense interest or where the stakes are high.

Dynamically generated maps showing a visual or graphical representation of the *conceptual footprint* and facilitates plagiarism analysis and discussion.

At this juncture our development work has produced a report showing the results of the "Extract from master essay" components juxtaposed with conceptually similar "Matched extract(s)". Such reports are amenable for use in checking large numbers of essays for possible similarity at the semantic level. An extract from a sample report is shown in **Figure 14**. For a large number of essays such a report would contain potentially hundreds of pages if the conceptual similarity among the essays is significant. With regard to the entire process described, the report production and reading consumes the greatest resource for computer and human respectively.

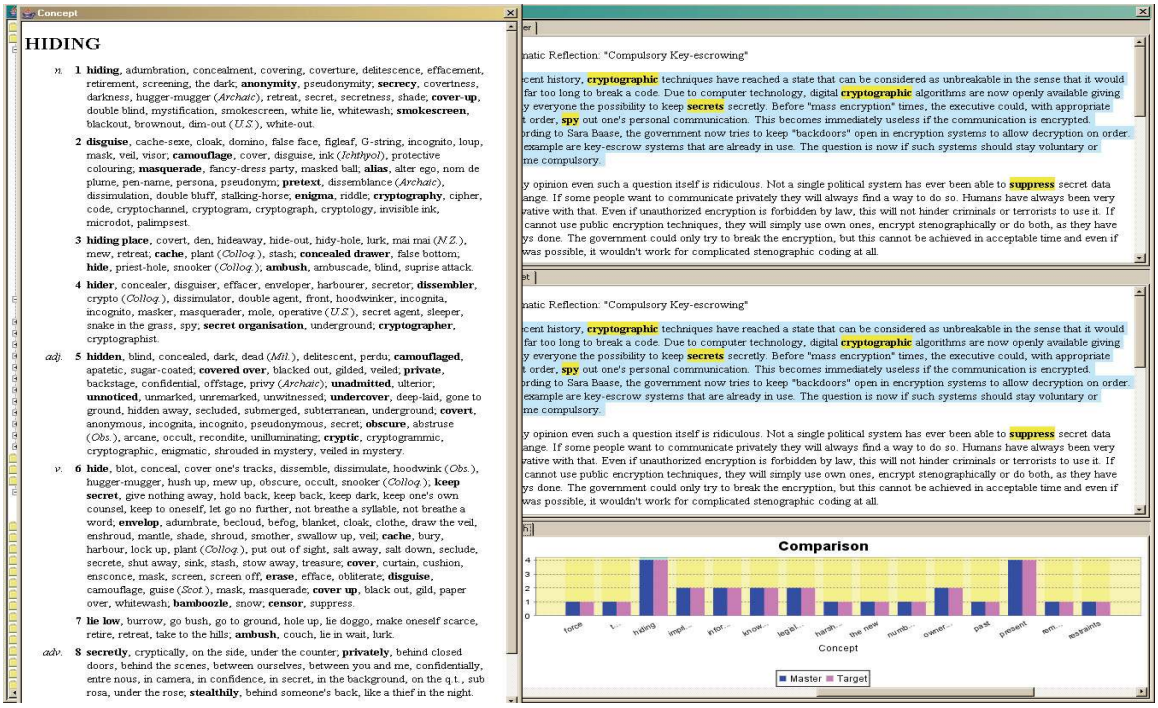


Figure 13: Analysis of concept with name "hiding"

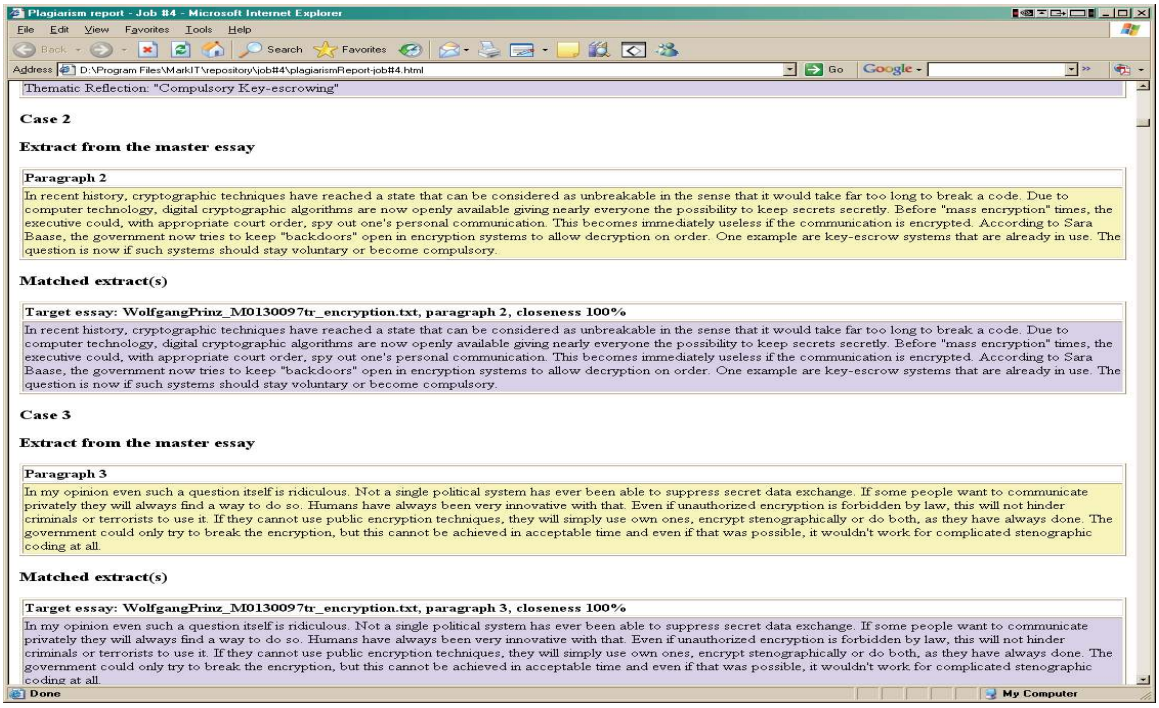


Figure 14: Extract from a sample conceptual analysis report

Since conceptual analysis is the main topic being addressed in this paper we consider a second example – this time with assignments of 300 to 500 words written by year 10 students on the topic of School Leaving Age" (**Figure 15**).

Master essay: B18 E4129 50G.txt

The age at which students are legally allowed to leave school should not be raised from 15 years of age to 17 years of age.

If a student is kept in school against their will they are less likely to do well in their studies. As a majority of them wont even try to do well and learn what is being tort. They will also disrupt other student that are willing to learn.

Also at the age of 15 these people are becoming adults. They are beginning to form their own ideas of what they want to do in the future. Some students are good at school and it would benefit them to stay others are not good at school and are better off in the work force. These people know there strengths and weaknesses and are there for better equipt to make the decision for them selves.

People have to be allowed to make their own mistakes. There are always options if the person is unhappy with their decision. Such as seeking higher education in the field they wish to go into at places like TAFE or they could even go back to school if they decide they have made the wrong choice. The point is it should be their responsibility to make the choice that is going to effect the rest of there lives.

Plagiarism detected

Extract from the master essay

Paragraph 1

The age at which students are legally allowed to leave school should not be raised from 15 years of age to 17 years of age.

Matched extract(s)

Target essay: B14 E4137 117R.txt, paragraph 1, closeness 50%

I belivev that the legal age to leave school should be raise from the age of 15 to the age of 17.

Target essay: B14 E4137 182G.txt, paragraph 1, closeness 63%

I agree with the Minister of Education, that the legal age that students should be allowed to leave school, is at the age of 17 years old not 15 years old.

Target essay: B20 E4129 179G.txt, paragraph 3, closeness 50%

The arguement that is stated in this essay is should students be allowed to leave school at 15 years of age or should be change to a later age.

Target essay: B21 E4129 155G.txt, paragraph 1, closeness 57%

Whether you student have to atend school at 15 to 17 years of age.

Target essay: B3 E4126 268G.txt, paragraph 5, closeness 77%

* According to the Minister of Education, the legal age for students to leave school will be changed from 15 years of age to 17 in 2002.

Figure 15: Concept analysis – "School Leaving Age" example

The "Master essay" is presented in full at the beginning of the report. Then for each paragraph of the "Master essay" (labeled **Paragraph 1** in the yellow/light panel) the "Matched extracts" from other essays (five instances of **Target essay** in the turquoise/darker panels) are presented with an indication of the conceptual closeness as a percentage agreement with the "Master essay". In the example, five essays contain conceptually similar content. Of course this does not necessarily

indicate plagiarism, however the reader may contemplate how well the computer, the NWV algorithm actually, has determined semantic proximity, which may be an indicator of plagiarism.

Conclusion

Through three case studies the author has illustrated how text-string comparison can be effective in detecting within-cohort plagiarism (Case 1), but can be inefficient for plagiarism detection on a lar-ger scale such as the WWW (Case 2). Furthermore it has been shown that while text-string comparisons are effective they may not flag the replication of others' ideas using semantically similar words. To detect such forms of copying one needs to use a conceptual analysis. We have applied the NWV algorithm because it is the fastest method known to extract semantic content from essays of arbitrary length, the efficacy of which was shown in Case 3.

Whilst the results achieved with this 'hi-tech' approach are promising one should stress that a 'hi-touch' approach is not to be ruled out and may be used in a complimentary manner for increased efficacy in detecting and addressing plagiarism (**Figure 16**).

In Fig 16 the 'hi-tech' approach can be seen used in step 2) whereas the 'hi-touch' approach is relied upon for the remainder of the steps. The term 'hi-tech/hi-touch' comes from Naisbitt (1982). As in all cases where humans rely on technology to help solve problems, in this situation there is a very large degree of reliance on human (6 out of 7 steps), as opposed to artificial, intelligence.

1) select some text fragment which is 'unlikely' to come from the nominated source and search for <selected text>

2) compare search results with original & highlight matching text

3) professor invites student for interview – bring paper copy of assignment

4) ask student to highlight all words which have been copied

5) compare student's highlighting with professor's highlighting and you can guess the student's reaction: DISBELIEF

6) professor listens patiently student's explanation, protestation, justification .

7) professor explains:
if we are HONEST in the assessment process and with each other,
then we can TRUST that the system is FAIR to everyone
and society will RESPECT the worth of your degree from this university:
for this reason we both have the RESPONSIBILITY to uphold academic **INTEGRITY**

Figure 16: *Low-Tech* & *High-Touch* Plagiarism Detection Method

References

Bierce, A. (1911). *The Devil's dictionary*. Retrieved from http://www.thedevilsdictionary.com - text by Ambrose Bierce, 1911; copyright expired.

Kloda, L.A. & Nicholson, K. (2005). Plagiarism detection software and academic integrity: The Canadian perspective. In *Proceedings Librarians' Information Literacy Annual Conference (LILAC)*, London (UK). Retrieved from http://eprints.rclis.org/archive/00005409/

Karner, J. (2001). *Der Plagiator* Retrieved from http://old.onlinejournalismus.de/meinung/plagiator.html

Maurer, H., Kappe, F. & Zaka, B. (2006). Plagiarism – A survey. *Journal of Universal Computer Science*, *12*(8), 1050-1084.

Naisbitt, J. (1982). *Megatrends. Ten new directions transforming our lives*. Warner Books.

Turnitin. (2007). http://www.turnitin.com/static/home.html

Williams, J.B. (2002). "The plagiarism problem: Are students entirely to blame?" In *Proceedings of AS-CILITE 2002*. Retrieved from http://www.ascilite.org.au/conferences/auckland02/proceedings/papers/189.pdf

Williams, R. (2006). The power of normalised word vectors for automatically grading essays. *The Journal of Issues in Informing Science and Information Technology, 3*, 721-730. Retrieved from http://informingscience.org/proceedings/InSITE2006/IISITWill155.pdf

Biography

Heinz Dreher is Associate Professor in Information Systems at the Curtin Business School, Curtin University, Perth, Western Australia. He has published in the educational technology and information systems domain through conferences, journals, invited talks and seminars; is currently the holder of Australian National Competitive Grant funding for a 4-year E-Learning project and a 4-year project on Automated Essay Grading technology development, trial usage and evaluation; has received numerous industry grants for investigating hypertext based systems in training and business scenarios; and is an experienced and accomplished teacher, receiving awards for his work in cross-cultural awareness and course design. In 2004 he was appointed Adjunct Professor for Computer Science at TU Graz, and continues to collaborate in teaching & learning and research projects with European partners.

Dr Dreher's research and development in the hypertext domain has centred on the empowering aspects of text & document technology since 1988. The systems he has developed provide support for educators and teachers, and document creators and users from business and government. 'DriveSafe', 'Active Writing', 'The Effectiveness of Hypertext to Support Quality Improvement', 'Water Bill 1990 Hypertext Project', 'A Prototype Hypertext Operating Manual for LNG Plant Dehydration Unit', 'Hypertextual Tender Submission - Telecom Training Programme', were all hypertext construction and evaluation projects in industry or education. The Hypertext Research Laboratory, whose aim was to facilitate the application of hypertext-based technology in academe, business and in the wider community, was founded by him in December 1989

Acknowledgements

The author would like to acknowledge the InSITE reviewers for their helpful comments and in particular thank Carl Dreher for his extensive and critical appraisal of early drafts of the paper.