PLAGIARISM AT UNIVERSITIES — HOW TO FIGHT IT? CASE OF THE CZECH REPUBLIC

Jiří Přibil, Ondřej Lešetický, and Hana Karásková
Faculty of Management
University of Economics, Prague
Czech Republic

Abstract

There has been an increasingly urgent need to prevent a very specific type of unethical behavior — plagiarism — at Czech universities/colleges in the last few years. With more and more increasing availability of electronic sources for professional scientific texts the level of its "violation" by plagiarism raises proportionally. The chance for intercepting such a plagiarized content is very low. The paper will present a basic comparison of the current situation on the plagiarism in the Czech Republic with the situation in the United States. At the conclusion there will be introduced unique system developed for extremely fast analysis of the plagiarized content.

Introduction

The Encyclopædia Britannica defines plagiarism as "the act of taking the writings of another person and passing them off as one's own. The fraudulence is closely related to forgery and piracy practices generally in violation of copyright laws. If only thoughts are duplicated, expressed in different words, there is no breach of contract. Also, there is no breach if it can be proved that the duplicated wordage was arrived at independently." The Compact Oxford English Dictionary simply states, that "Plagiarism is taking the words or ideas of someone else and pass it off as one's own." The Cambridge Advanced Learner's Dictionary uses a very similar definition: "Plagiarize: to use another person's idea or a part of their work and pretend that it is your own."

These definitions deal not only with questions of text copying but most also with the problem of ideas copying. That's a big problem in the real world: we suppose, that there is no effective way how to automatically detect other people's ideas or thoughts — there must always be some kind of human judge (and this judge must be an expert in his subject field) who decides whether the suspicious document is a plagiarism or not. Although there have been some ontological tools developed capable of detecting the semantic similarity between documents, none of them is able to state: "This idea is not original, it's plagiarism" for sure.

But the reuse of "other people words" is quite common and correct under one basic circumstance: the credit must be given to the original author. This situation becomes again more difficult because of the fact that there are many ways of using the regular citations (e.g. APA Style is most often used in social sciences; the IEEE Style is particularly used in computer science; the Harvard Style is recommended by the British Standards Institution; ISO 690 and ISO 690-2 are both used, for example, in the Czech Republic, etc.) and these citation styles are quite varying.

Thus it is possible to very precisely measure the percentage of plagiarized text in document, but the decision about the guilt and punishment is again up on the human factor. The plagiarism measurement should work as an "early warning system" and it should detect suspicious texts — and that should be the primary purpose of every plagiarism detection service.

In this paper, we would like to introduce our own plagiarism detection system DIANA (Document Identity Analysis) developed in the Faculty of Management, University of Economics, Prague. We introduce a quite new concept of "unordered *n*-grams" in the process of plagiarism detection and show some interesting results of DIANA achieved during the test phase based on out previous work (Přibil, Kubalová, & Kincl, 2007).

It's important to state that our system currently covers documents written in Czech language, but the usability in other languages is only a question of other language tools (vocabularies, stemmers, stop-words collection, etc.) and doesn't affect the generality of problem solved. The examples as provided in this text are in English to facilitate a better insight into the problems involved.

Current Situation

Types of Detection Services

There are basically four types of plagiarism detection: a) commercial online detection services, b) free online detection services, c) locally installed commercial applications, and d) self-developed applications used by one (or few) school or college.

All these detection methods have their advantages and disadvantages and we'll try to define them, because the decision between "to use ready-made solution", "to buy application for our institution" or "to develop internal solution" is very important in the beginning of the process of plagiarism detection in praxis.

It's also important to say that: a) some services allow detection of plagiarism between documents in corpus (database, file system — it's very useful in the case of schools); b) some services only compare plagiarized parts of documents against documents on the Internet; and c) some are able to combine both of these methods.

Commercial online detection services. Internet detection services like MyDropBox Suite (mydropbox.com) and Turn It In (turnitin.com) are quite favourite solutions for many institutions in the English-speaking countries. These services are able to detect wide spectrum of plagiarized text in student's assessments because of huge database of source documents commonly written in English. The biggest disadvantage of these services is very low support for foreign (non-English) languages. A quite important aspect of these services is their price. For example, Turn It In offers a single campus licence for £1000 GBP plus an extra £0.51 GBP per student over a 12-month period. This cost allows for an unlimited number of submissions and tutor enrolments per year.

Free online detection services. There are only few free Internet plagiarism detection services (e.g. DOC Cop) — but their future is always very unclear as they can stop working any day. Also the support for foreign languages other than English is weak.

Locally installed commercial detection systems. Some companies offer client plagiarism detection applications distributed as stand-alone applications running on customer's computers. A good example of a plagiarism system's possibilities is the Essay Verification Engine – EVE2. But again, there is one big disadvantage: this application can only "determine if students have plagiarized material from the World Wide Web" and thus it can't check plagiarism between documents in corpus.

Self-developed detection solutions. Some institutions try to go another way to fight plagiarism: they develop their own detection software and don't sell it or offer for public use. Open source solution Wcopyfind is one of few systems available for downloading; it searches plagiarized parts in files on a local file server and returns a comparative log of reused text segments.

Current Situation in Czech Republic

Table 1 shows results of two surveys realized at University of Virginia in 2005 (University of Virginia, 2006) and currently repeated at Faculty of Management, University of Economics, Prague, about a year ago.

Table 1: Unethical behaviour comparison — Czech Republic (CZ) and USA (US)

Specific Behavior (in %)	**Never**		**Once**		**> Once**		**Not Relevant**	
	US	CZ	US	CZ	US	CZ	US	CZ
Fabricating or falsifying research data.	79	68	3	17	1	5	17	11
Fabricating or falsifying a bibliography.	87	81	8	10	3	1	2	9
Copying sentences from written source w/o footnoting.	66	63	18	14	15	13	2	10
Copying from electronic source w/o footnoting.	64	58	20	17	15	16	1	10
Copying material, word for word, from written source.	96	68	2	15	1	7	1	12
Turning in paper obtained from term paper "mill" or site.	98	87	<1	3	<1	1	2	11
Turning in work done by someone else	97	89	2	5	1	1	1	7

Most rates are similar enough, surprisingly high number of Czech students use large parts of written sources (about 20 per cent compared to 3 per cent at University of Virginia). Our conclusion is simple: The situation is bad enough; twenty to thirty per cent of our students plagiarize their term papers, and what's even worse, almost nobody is punished — in fact, you can count them on the fingers of one hand every year.

Theoretical Background of Plagiarism Detection Process

First off, let's define the main options for pre-processing of the documents being parsed (input) the detection systems. As noted in the following chapters, some of these steps used in DIANA system are considered to be crucial.

Document Pre-processing

It's not very practical to measure plagiarism rate on the original documents — and it's very useful to do some basic "document pre-processing" actions before the measurement itself. This pre-processing's aim is an "information concentration" of the original documents. Whole process of the pre-processing consists of three main following) concepts — linearization, filtration and stemming (Garcia, 2005).

Document linearization. Document linearization is a process of a document reduction. There are usually two steps:

- Mark-up and format removal. During this phase, all mark-up tags and special formatting are removed from the document (all colors, headers, fonts, etc. are removed and the document is converted to the plain text).

- Tokenization. In this phase, all remaining text is lowercased and all punctuation is removed as well as the number sequences; thus, the document is represented as one very long "sentence."

Filtration (stop-words removing). There are words that can be marked as "content-unattractive" in every language (MySQL, 2009) — that means words, whose use in a language is so common that balancing and measuring of their presence in a concrete document is useless. We can say that including these words incomputation of documents' similarity measurement is ineffective as it: a) increases the computing complexity (the algorithm has to work with much more words than necessary); and b) distorts the final rate of documents similarity.

In the Czech language, many words can be supposed to be stop-words, i.e. all conjunctions, prepositions and pronouns, some adjectives, etc. In English, many different stop-words lists are used, but all of them contain common words as "a", "of", "the", "I", "it", "you", and "and", for example.

Stemming. Stemming is a process that transforms words to their base form. Thus the related words have the same "stem." The process of stemming is quite complex in most languages and uses different algorithms; the goal of this process is to reduce the count of words used and for many languages the problem of using different grammatical cases and other linguistic rarities.

Document Identity Indexes

We define two types of document identity indexes in our DIANA system:
- pair-wise identity,
- global identity.

Commonly, these characteristics are measured on short text segments. The elemental issue is how to define these segments. From the nature of the language these could be sentences (simple sentences or clauses) or whole paragraphs. Field observation shows that none of these approaches leads to any satisfactory results — plagiarists most frequently proceed by "compiling" short segments of stolen text and interlard it with texts of their own. With this kind of approach the whole sentences are not maintained; plagiarists also often change the words order, grammatical cases and so on.

Classical Approach: N-grams

As an appropriate method at the moment it seems to be the use of the so called *n*-grams (Cavnar & Trenkle, 1994), where *n*-gram is a sub-sequence of *n* items from a given sequence. In this specific case the items are words, thus each document is represented as a set of *n*-grams (substrings of *n* words length).

The “right” value of *n* is also a question for discussion — too low *n* (2 or 3) can reveal much more identical (“plagiarized”) substrings, but every language has many common phrases of this length and their use is really not a plagiarism (in English e.g. “and so on”, “there is” and many, many others). On the other side, high value of *n* shows another problem — it can’t reveal plagiarism of substrings with length (*n*-1) and shorter. For example, Zini, Fabbri, Moneglia, and Panunzi (2006) use 4-grams in their interesting multi-level comparison method.

In the case of “classic” *n*-grams the fixed-sequence of the individual words (expressions) is quite unsuitable for plagiarism detection. Since this factor disables completely (or in many cases) one of the classic technique of plagiarism, that is the alternation of the word sequence in the sentence. The only solution being left would be to decrease the value of *n* to very small figure/number ($n = 2$ or $n = 3$). That (as mentioned above) leads to a substantial increase of false positive plagiarisms (plagiarized parts of the document) detected.

Figure 1: Example of classic n-gram representation

source text: *Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc dapibus.*

2-grams (basic normalization accomplished): *(lorem, ipsum), (ipsum, dolor), (dolor, sit), (sit, amet), (amet, consectetur), (consectetur, adipiscing), (adipiscing, elit), (elit nunc), (nunc, dapibus)*

3-grams (basic normalization accomplished): *(lorem, ipsum, dolor), (ipsum, dolor, sit), (dolor, sit, amet), (sit, amet, consectetur), (amet, consectetur, adipiscing), (consectetur, adipiscing, elit), (adipiscing, elit, nunc), (elit, nunc, dapibus)*

Proposed Solution: Unordered *n*-grams

Our proposed solution uses so-called “unordered *n*-grams.” These *n*-grams have fixed length throughout the system (e.g. DIANA uses 5-grams), but the sequence of the individual words (expressions) has not been taken into account. This circumstance plays a fundamental role: not only in the case of so called “ideas plagiarism” but also in some specific cases of the “creative plagiarism”, because then the probability of disclosure of such an unethical behavior is much higher.

Unordered *n*-grams representation. At present we solve mainly the following question: how to store those unordered *n*-grams, how to tell that given *n*-gram is identical to other *n*-gram with only word sequence has been changed? The problem solution could be application of the hash function and representation of the unordered *n*-grams using the hash. We have considered many possible options, and as the most suitable we have established the following process:

1. The words (expressions) contained in given *n*-gram are sorted alphabetically in ascending order.

2. These words are concatenated into one string with specific length of the sum of individual words lengths in the *n*-gram.

3. For every string a simple hash is being calculated (DIANA uses very fast 128-bits RIPEMD-128 hash function) and saved into the database.

The key feature for detection of the plagiarized parts of the document is the use of hash value as an unordered *n*-grams representation, since this ensures that *n*-grams differing only in the word sequence are recognized automatically as plagiarism and also increases the plagiarism rates.

Plagiarism Rate Indexes

Now let’s define two of indexes determining the numeric value representing the rate of plagiarized text in the document, disregarding the chosen option of the source document processing being used: pair wise identity and global identity.

Pair-wise Identity

Pair-wise Identity — noted as $ide_p(D_2, D_1)$ — is defined as “document-to-document identity”:

$$ide_p\left(D_2, D_1\right) = \frac{\sum n\text{-}grams\ used\ both\ in\ D_2\ and\ D_1}{\sum n\text{-}grams\ in\ D_2} \qquad (1)$$

Pair-wise identity calculates the rate of plagiarized *n*-grams in document D_2 compared to (single) document D_1.

Its value must be between 0 and 1. Value of 1 shows that all *n*-grams of document D_2 were plagiarized from document D_1. Value of 0 shows that in document D_2 there are no substrings plagiarized from document D_1.

Global Identity

Global Identity — noted as $ide_g(D_x, E)$ — is defined as “document-to-corpus identity”:

$$ide_g\left(D_x, E\right) = \frac{\sum n\text{-}grams\ used\ both\ in\ D_x\ and\ E}{\sum n\text{-}grams\ in\ D_x} \qquad (2)$$

$E=\{D_1, D_2, \ldots D_{x-1}\}$ is a set of (x-1) documents in the corpus.

Global identity calculates the rate of plagiarized substrings of document D_x compared to (many) documents from the set E. Its value is defined similar to pair-wise identity: Value of 1 shows that all n-grams form document D_x are plagiarized from one or more documents from the set E. Value of 0 shows that in the document D_x there has been nothing plagiarized from any documents from set E. This index doesn't penalize "multiple sources occurrence" — it doesn't matter if the plagiarized substring originates from one or more source documents simultaneously; the global identity index will always be the same. Therefore always stated as true:

$$ide_p\left(D_3, D_1\right) + ide_p\left(D_3, D_2\right) \geq ide_g\left(D_3, \left\{D_1, D_2\right\}\right) \qquad (3)$$

Comparison of the Plagiarism Detection Results Using *N*-Grams versus Unordered *N*-Grams

See Table 2 below for results of comparison performed on reduced (randomly selected) corpus of 100 real-life term papers submitted by students at Faculty of Management, University of Economics last year. It shows global identity rates for 5 different (randomly selected) documents (D_{13}, D_{41}, D_{62}, D_{87}, D_{93}) from this corpus compared with other 99 documents (E set). Three different approaches in plagiarism detection are used:

1. detection based on classic 3-grams, without any filtration or stemming (App A),
2. detection based on unordered 5-grams, documents are filtrated and stemmed (App B),
3. detection based on unordered 6-grams, documents are filtrated and stemmed (App C).

Table 2: ide_g values comparing detection systems based on classic n-grams and unordered n-grams

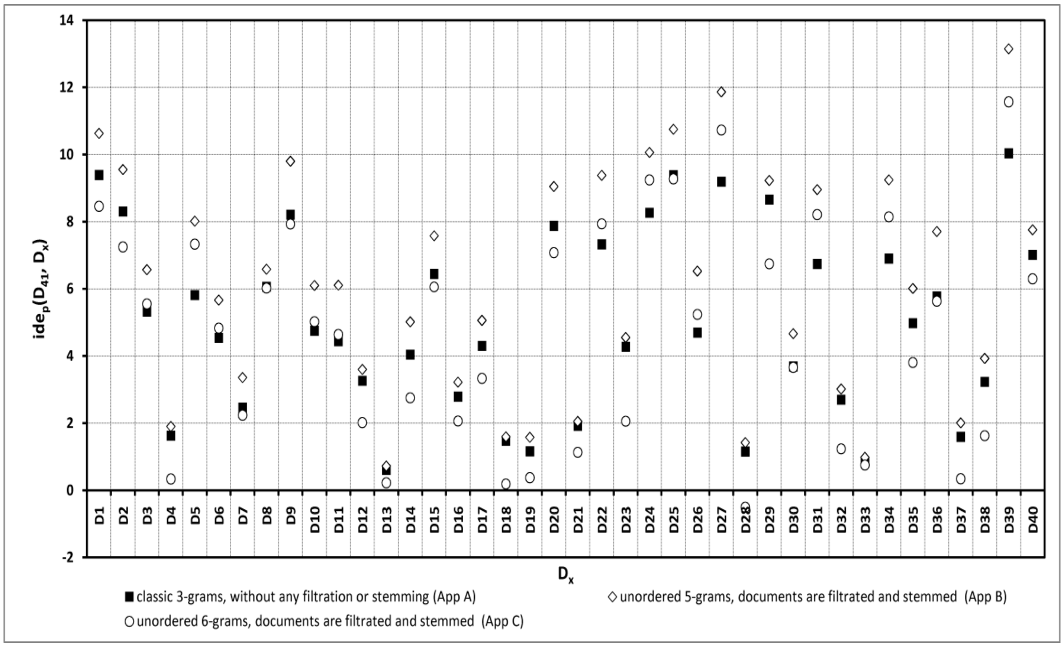
	Doc 13	**Doc41**	**Doc62**	**Do 87**	**Doc 93**
App A	7,14 %	22,09 %	4,98 %	11,81 %	17,30 %
App B	10,84 %	32,66 %	7,62 %	14,98 %	23,93 %
App C	9,96 %	30,35 %	5,32 %	12,70 %	23,26 %

As we can see, there are very interesting differences in these three approaches: from the global view, detection using unordered *n*-grams, even with higher N than classic *n*-grams, gives better results.

As stated earlier, it is not possible to determine the “quality of detection” (in terms of whether the declared parts of the documents parts are being plagiarized or not). On the other hand in case of the implementation of such a system as a tool for a “human judge,” the disclosed results are benefits.

The results of pair-wise identities calculations performed on the one selected document (D_{41}), in subsequent turn against the previous 40 other ones within the corpus: $ide_p(D_{41}, D_1)$, $ide_p(D_{41}, D_2)$, . . ., $ide_p(D_{41}, D_{100})$ — as shown in Figure 2 below.

Figure 2: Comparison of plagiarism detection systems based on classic n-grams and unordered n-grams



The results indicate that longer unordered 5-gram is able on testing (but real ones) data to detect more suspicious text sequences than the classic 3-grams. The crucial part of this process is of course pre-processing made in the case of unordered *n*-grams. The trend of decreasing number of the plagiarized *n*-grams with increasing *n* value is evident as well. The specific tests performed under real-world circumstances of the Czech language (validated by human rater/evaluator) suggest that the unordered 5-grams are the most fitting method for representation and plagiarism detection.

Real-life View

The DIANA plagiarism detection service is a server-side application written in Borland Delphi language and can be run on both 32-bit and 64-bit Windows operating system. The Firebird database has been used for saving documents; however any SQL based relational database could be a suitable option.

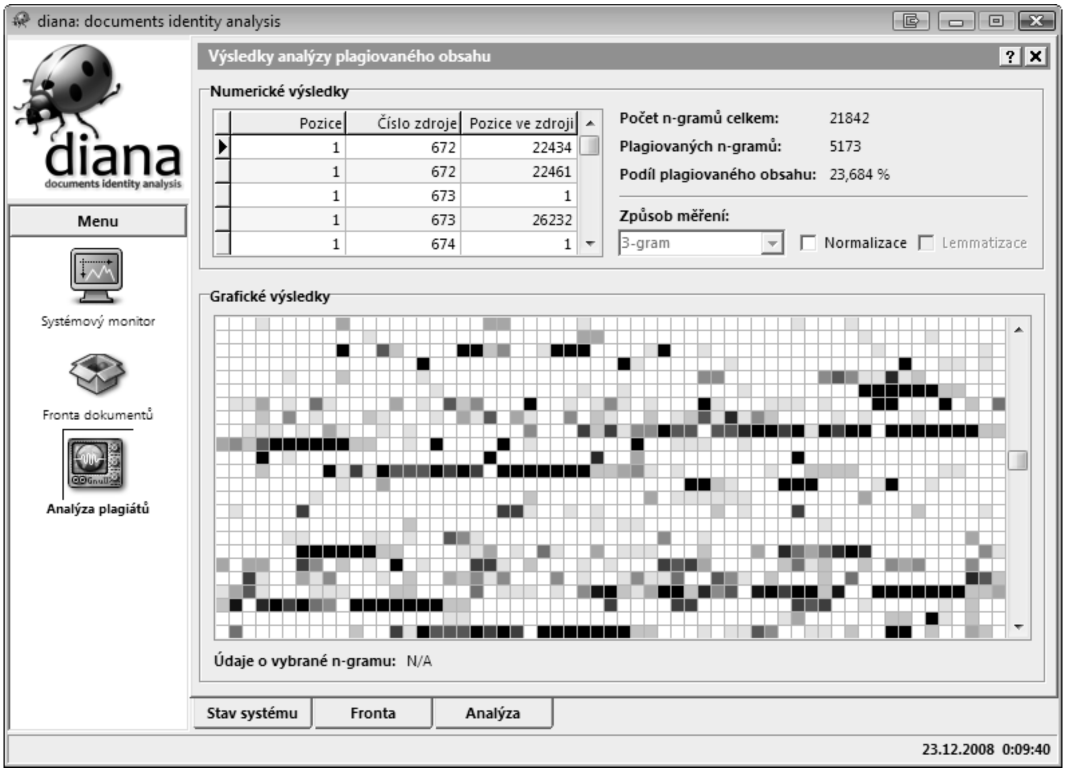
At the present time the DIANA users could upload/insert documents by two different ways:

1. By sending the file (named properly) in a supported format (plain text, MS Word, RTF) on a specific e-mail address.
2. By uploading file to a document server hosted at Faculty of Management, University of Economics.

The DIANA's results of detected plagiarisms are then sent to designated persons (supervising academics and authors) in a specific numerical and graphical representation — more advanced representation techniques are a subject of further research. A plagiarism rate index reaching circa 10 % is being perceived as a general threshold requiring educator's/rater's attention, perhaps using a more comprehensive graphic analytical tool.

Figure 3 below shows the "plagiarism rate heat map" — every cell sequentially represents one *n*-gram, the darkness of this cell shows the number of plagiarism sources.

Figure 3: DIANA system interface (Czech language)



Pozice	Číslo zdroje	Pozice ve zdroji
1	672	22434
1	672	22461
1	673	1
1	673	26232
1	674	1

Počet n-gramů celkem: 21842

Plagiovaných n-gramů: 5173

Podíl plagiovaného obsahu: 23,684 %

Způsob měření: 3-gram ☐ Normalizace ☐ Lemmatizace

Údaje o vybrané n-gramu: N/A

23.12.2008 0:09:40

Future Works

The developers would like to focus on the following areas in which the improved results could possibly be reached:

1. The interception of the inter-lingual plagiarism. Generally this idea should not be a problem with the superior translators available. Such procedure has already been tested with very promising results and only the obstacle seems to be a certain scarcity of the sufficient number of high-quality open-source translators.

2. The support for a much broader spectrum of files analyzable by the DIANA system, especially the OpenOffice and PDF formats.

Acknowledgements

This work has been supported by The Czech Republic National Research Programme II under project No. 2C06019 “(Medical) Knowledge Acquisition and Modelling” and by The Internal Grant Agency of Prague University of Economics under project No. 11/2010 “Advanced Methods of Text Plagiarism Detection and Visualisation.”

References

Cavnar, W. B., & Trenkle, J. M. (1994). Ngram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval* (pp. 161–175). Las Vegas, USA.

Garcia, E. (2005). *Document indexing tutorial.* Retrieved February 15, 2010, from http://www.miislita.com/information-retrieval-tutorial/indexing.html

MySQL. (2009). *Full-Text Stopwords*. Retrieved February 2, 2010, from http://dev.mysql.com/doc/refman/5.5/en/fulltext-stopwords.html

Přibil, J., Kubalová, K., & Kincl, T. (2007). Plagiarism detection in large text collections. In *Proceedings of the IASK International Conference — E-Activity and Leading Technologies* (pp. 72–80). Porto, Portugal.

Zini, M., Fabbri, M., Moneglia, M., & Panunzi, A. (2006). Plagiarism detection through multilevel text comparison. In *Proceedings of the Second International Conference on Automated Production of Cross Media Content For Multi-Channel Distribution* (pp. 181–185). Leeds, UK.