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

Jiří Přibil, Ondřej Lešetický, and Vladislav Bína  
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 and colleges in the last few years. With increasing availability of electronic sources for professional scientific texts the level of “violation” by plagiarism rises proportionally. The chance for intercepting such plagiarized content is very low. The paper will present a comparison of the current situation on plagiarism in the Czech Republic with the situation in the world. The conclusion will introduce a unique system developed for extremely fast analysis of content for plagiarism.

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 passing them off as one’s own.” The Cambridge Advanced Learner’s Dictionary uses 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 idea copying. That’s a big problem in the real world: we suppose that there is no effective way how to detect automatically other people’s ideas or thoughts purely — there must always be some kind of human judge (and this judge has to be an expert in his subject field) who decides whether the suspicious document is really plagiarized or not. Although some ontological tools capable of detecting the semantic similarity between documents have been developed, none of them is able to state for sure: “This idea is not original, it’s plagiarism.”

But the reuse of “other people words” is quite common and correct under one typical circumstance: when the credit is given to the original author. This situation becomes more complex in conjunction with the process of automatic plagiarism detection because there are many ways of using the regular citations (e. g., APA style is most often used in social sciences, IEEE style is used in computer science, Harvard Style recommended by
the British Standards Institution, ISO 690 and ISO 690-2 both used for example in Czech Republic) and these citation styles are quite varied.

While measuring very precisely the percentage of plagiarized text in a document is possible, the decision about the guilt and punishment is part of the human factor. The plagiarism measurement should work as “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 at 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 DIANA achieved during the test phase based on our previous work (Příbıl, Kubalová, & Kincl, 2007).

Our system currently covers documents written in Czech language, but its use in other languages is only a question of other language tools (vocabularies, stemmers, stop-words collection, etc.) and doesn’t generally affect the findings. 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

Basically there are 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 a few) school or college.

All these detection methods have their advantages and disadvantages and we’ll try to define them, because the decision whether to use a ready-made solution, buy an application for an institution, or to develop an 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 — 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 Turn It In (turnitin.com) are quite favourite solutions for many institutions in English-speaking countries. These services are able to detect a wide spectrum of plagiarized text in students’ assessments because of their 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 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 customers’ 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 it 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 surveys done at the University of Virginia in 2005 (University of Virginia, 2006) and repeated at the Faculty of Management, University of Economics, Prague in 2009.

<table>
<thead>
<tr>
<th>Specific Behavior (in %)</th>
<th>Never</th>
<th>Once</th>
<th>&gt; Once</th>
<th>Not Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>US</td>
<td>CZ</td>
<td>US</td>
<td>CZ</td>
</tr>
<tr>
<td>Fabricating or falsifying research data</td>
<td>79</td>
<td>68</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>Fabricating or falsifying a bibliography</td>
<td>87</td>
<td>81</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Copying sentences from written source w/o footnoting</td>
<td>66</td>
<td>63</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td>Copying from electronic source w/o footnoting</td>
<td>64</td>
<td>58</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>Copying material, word for word, from written source</td>
<td>96</td>
<td>68</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>Turning in paper obtained from term paper &quot;mill&quot; or site</td>
<td>98</td>
<td>87</td>
<td>&lt;1</td>
<td>3</td>
</tr>
<tr>
<td>Turning in work done by someone else</td>
<td>97</td>
<td>89</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>
Many of the rates between the two groups are similar, although surprisingly a 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. What is 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

Let’s define the main ideas of document pre-processing; some of these steps used in the DIANA system are considered to be crucial.

Document Pre-processing

It’s not very practical to measure plagiarism rates on the original documents as they contain considerable information extraneous to measuring plagiarism — 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 and consists of three main concepts — linearization, filtration and stemming (Garcia, 2005).

Document linearization. Document linearization is a process of a document reduction and usually has two steps:

- Markup and Format Removal. During this phase, all markup 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 any language (MySQL, 2009). These are whose use in a language is so common that balancing and measuring of their presence in a concrete document is useless. Including these words to computation of documents similarity measurement is ineffective as it: a.) increases the computing complexity (the algorithm has to work with many more words than necessary); and b.) distorts the final rate of documents’ similarity.

In the Czech language, many words are treated as 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 such 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, and
- 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 interlaying them with texts of their own. With this kind of approach the whole sentences are not maintained; plagiarists also often change word order, grammatical case, and so on.

**Classical Approach: N-grams**

An appropriate method at the moment 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 word lengths).

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” and “there is” among many, many others). On the other side, a 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 in many cases disables completely one of the classic technique of plagiarism — the alternation of the word sequence in the sentence — the only solution left is to decrease the value of n to a very small figure/number (n = 2 or n = 3). This (as mentioned above) leads to a substantial increase of false positive plagiarisms (plagiarized parts of the document) detected.

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 “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 questions: how to store those unordered n-grams and how to tell that a given n-gram is identical to other n-grams with only the word sequence 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 we have established the following process as the most suitable:

1. The words (expressions) contained in a 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 word lengths in the n-gram.

3. For every string a simple hash is calculated (DIANA uses a 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 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”:

\[
ide_p(D_2, D_1) = \frac{\sum \text{n-grams used both in } D_2 \text{ and } D_1}{\sum \text{n-grams in } D_2}
\]  

(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. A value of 1 shows that all n-grams of document \( D_2 \) were plagiarized from document \( D_1 \); a value of 0 shows that in document \( D_2 \) no substrings were plagiarized from document \( D_1 \).

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

\[
ide_g(D_x, E) = \frac{\sum \text{n-grams used both in } D_x \text{ and } E}{\sum \text{n-grams in } D_x}
\]  

(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: a value of 1 shows that all n-grams form document \( D_x \) are plagiarized from one or more documents from the set \( E \); a value of 0 shows that in the document \( D_x \) nothing has been 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 the following is always stated as true:

\[
ide_p(D_3, D_1) + ide_p(D_3, D_2) \geq ide_g(D_3, \{D_1, D_2\})
\]  

(3)
Comparison of Plagiarism Detection Results using n-grams versus Unordered n-grams

Table 1 below gives the results of comparison performed on a reduced (randomly selected) corpus of 100 actual term papers submitted last year by students of the Faculty of Management, University of Economics. It shows global identity rates for five different (randomly selected) documents ($D_{13}$, $D_{41}$, $D_{62}$, $D_{87}$, $D_{93}$) from this corpus compared with another 99 documents ($E$ set). Three different approaches in plagiarism detection are used:

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

Table 2: ide$_p$ Values Comparing Detection Systems based on Classic n-grams and Unordered n-grams

<table>
<thead>
<tr>
<th></th>
<th>Doc 13</th>
<th>Doc 41</th>
<th>Doc 62</th>
<th>Do 87</th>
<th>Doc 93</th>
</tr>
</thead>
<tbody>
<tr>
<td>App A</td>
<td>7.14 %</td>
<td>22.09 %</td>
<td>4.98 %</td>
<td>11.81%</td>
<td>17.30 %</td>
</tr>
<tr>
<td>App B</td>
<td>10.84 %</td>
<td>32.66 %</td>
<td>7.62 %</td>
<td>14.98%</td>
<td>23.93 %</td>
</tr>
<tr>
<td>App C</td>
<td>9.96 %</td>
<td>30.35 %</td>
<td>5.32 %</td>
<td>12.70%</td>
<td>23.26 %</td>
</tr>
</tbody>
</table>

We can see 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, determining the “quality of detection” in terms of whether the declared parts of the documents parts are being plagiarized or not is not possible. On the other hand in the case 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$_p(D_{41}, D_{100})$ — is shown in Figure 2.

The results indicate that longer unordered 5-gram is able on testing 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 with the Czech language (validated by
human rater/evaluator) suggest that the unordered 5-grams are the most fitting method for representation and plagiarism detection.

Figure 2: Comparison of Plagiarism Detection Systems based on Classic n-grams and Unordered n-grams

Real-life View

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

Presently DIANA users can upload/insert documents in two different ways:

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

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 perceived as a general threshold requiring the 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 the cell reflects the number of plagiarism sources.

Figure 3: DIANA System Interface (Czech language)

Future Work

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

- 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.

- 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
References


Jiří Příbil is Vice-Dean for External Affairs at the University of Economics in Prague, Faculty of Management and Senior Lecturer in Information Technologies and Management. He received his doctorate in Management in 2010 and he teaches courses in information science, project management, operation management, and multimedia. He has done research and published articles on plagiarism detection and e-learning technologies. His main recent research fields include data mining and text mining in business administration. He was a member or coordinator of many grant programs including a €2 million program (Medical) Knowledge Acquisition and Modeling funded by the Czech Republic Ministry of Education, Youth and Sports during 2006–2011.

Ondřej Lešetický is currently a Senior Lecturer and research fellow at the Institute of Healthcare Services Management, University of Economics, Prague. He received his degrees in Health Care Management and Public Services Administration. His research interests cover mainly the area of application of simulation techniques in the field of healthcare management, but as a member of multidisciplinary research groups he has also contributed to different health technology assessment studies.

Vladislav Bína is a post-doc lecturer and researcher at the University of Economics in Prague, Faculty of Management. He reached the PhD grade in 2011 with the thesis "Multidimensional Probability Distributions: Structure and Learning." Among his research interests rank the themes of probabilistic modeling, partially with stress on
combinatorial aspects and uninformed approach to text-mining. His teaching activities are focused on statistical courses and managerial decision making.