THE FUTURE OF COPY DETECTION TECHNIQUES

Zdenek Ceska[1]

Abstract: Internet is one of the richest encyclopaedias in the world. Students can easily download various free documents and then plagiarize their content. This paper describes the current state of copy detection methods and proposes some new trends. New approaches, closer to nature language processing, can essentially improve identification of hardly-detectable cases of plagiarism, i.e. single-word changes and sentence structure changes. Synonyms and Latent Semantic Analysis are discussed in detail for better understanding of the semantics within documents.

Keywords: Plagiarism, Copy Detection, N-grams, Nature Language Processing, Synonyms, Singular Value Decomposition, Latent Semantic Analysis.

1 INTRODUCTION

With the growing popularity of the Internet, many various documents are available free. People can easily search for the required documents and make their copy instead of writing the documents themselves. These practices have an enormous impact on the education system. In addition, the problem is also supported by many servers, such as www.schoolsucks.com, www.cheathouse.com, and www.seminarky.cz, which offer a wide range of various topics. As the defence against widely spread plagiarism, effective copy detection methods have to be employed.

Document protection techniques, which disable copy&paste and printing, are insufficient. Some vulnerabilities are often exploited and the protection is bypassed. Moreover, many of the specialized protections require some additional tools, which prevent reading the protected documents on the computers where administrative access is not available.

A large database of existing documents is a better solution. The main idea of this protection rests in psychology because every plagiarized document can be easily identified when compared to the database. To reach successful identification of the plagiarized documents, two conditions should be met: the database covers up all common documents on the required topics and an effective detection method is employed.

Most of the plagiarists only copy a part of a document and do not try to hide this activity. This is an evident case of plagiarism that can be easily identified because a large continuous text is copied. The consistent plagiarists copy some parts of sentences and sometimes exchange several words to cause confusion. Clough (2000, 2003) and Maurer (2006) describe other methods of plagiarism and even paraphrasing in detail.

2 RELATED WORK

Many older systems are built on various modifications of Vector Space Model. One of the most famous modifications is system SCAM described by Shivakumar (1995, 1996). Another system COPS, presented by Brin (1995), compares sentences instead of word frequency used in SCAM. Many other modifications exist there using words, sentences, paragraphs, or whole

[1] Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Univerzitni 22, 306 14 Pilsen, Czech Republic, e-mail: zceska@kiv.zcu.cz

documents. Smaller granularity improves detection of partial copying. However, too small chunks, such as words, interrupt sentence pattern and can produce some noise.

Modern copy detection methods employ document chunking and compare N-grams instead of term frequency, presented by Monostori (2002), Pataki (2003), and Burrows (2004). The document chunking is usually performed by a fixed size, e.g. 5 words, on all documents in a database. Then N-grams of the same size are extracted from a particular tested document and compared to the stored chunks. N-grams also effectively solve the text-shifting issue caused by word insertion and removal from the plagiarized documents. Ceska (2007) describes suitable chunking strategy and tools for N-gram extraction in detail.

3 CLASSIFICATION OF COPY DETECTION METHODS

The most general classification of copy detection methods is to **free text** or **source code**. Source code copy detection methods are widely covered and the current implementations work well. Therefore, the main objective of this paper is to propose some new methods to identify plagiarized students' works. The following classification is intended just for free text copy detection methods.

Lancaster (2005) introduced the classification depending on the complexity and the number of processed submissions, which can be seen in Table 1.

Type of classification		**Description**
Complexity of the used method	Superficial	The metrics is computed without any knowledge of the linguistic rules or a document structure.
	Structural	The metrics is computed with a partial understanding of documents, e.g. words are converted into their linguistic root, or replaced by a synonym.
Number of processed by the used method	Singular	A single document is processed to compute the metrics. Several Singular metrics can be employed to calculate how similar the documents are.
	Paired	Two documents are processed together to compute the metrics.
	Multi-dimensional	*N* documents from a corpus are processed together to compute the metrics.
	Corpal	All documents contained in a corpus are processed together to compute the metrics.

Table 1: Classification of free text copy detection methods

Almost all current free text or source code copy detection systems are Paired (COPS, Ferret, JPlag, MOSS, SCAM) or Singular (Jones). This means that every document must be compared with any other possible documents to analyze the whole corpus. The time complexity for this case takes $O(n^2)$, where n is the number of documents in the corpus. Therefore, Paired and Singular methods are suitable for seeking some possibly plagiarized documents, which are related to the concrete tested document. However, none of both methods is able to perform criss-cross comparison at a time.

The older systems, such as COPS or SCAM working on the term frequency, are purely Superficial. The current systems, which employ N-grams, are also rather Superficial than Structural. The reason is too time-consuming analysis of sentences whose grammar includes

many linguistic rules. Fortunately, some modern approaches from the other fields of nature language processing give us new possibilities to modification and improving the current plagiarism detection methods.

4 FUTURE DIRECTIONS

Widely spread plagiarism dictates us to develop new and better detection approaches. The current ones are able to identify a partial copy of a paragraph or sentence, but it is difficult to identify more granular sentence changes or paraphrasing. The advanced techniques of nature language processing must be employed for better understanding of the document topics.

The following subsections describe some familiar approaches, well-known from the other fields of nature language processing, which can essentially improve the precision of the current copy detection methods.

4.1 Lemmatization & Stop-Word Removal

Lemmatization uses a dictionary or an algorithm to convert words into their base root. Although lemmatization is considered useless for text classification due to its lesser precision, copy detection can exploit the word-space reduction by lemmatizing the words.

Stop-Word Removal is one of the basic methods for data reduction. It removes all common words, e.g. articles or some prepositions, to avoid the noise caused by very frequent words. Systems based on Vector Space Model usually retain 70% the most frequent words and the rest 30% is removed as inconvenient. On the other hand, some systems, which match sentence overlap, do not use Stop-Word Removal at all. In modern systems, where the semantics must be preserved, words should be removed very carefully. This can be reached applying a dictionary method using WordNet, which includes a word-feature-based description for many well-known languages.

4.2 Synonyms

Consistent people usually try to hide their activity and replace words by suitable synonyms in order to break up the continuous copied sentences. When it is done with a sufficient granularity, most of the current systems fail matching sentence overlap with the source document. Converting all words to their most common synonyms might help to solve this problem. However, we should be careful because not all synonyms correspond with every meaning. We propose using a deep word analysis in accordance with WordNet.

4.3 Latent Semantic Analysis

In linear algebra, the Singular Value Decomposition (SVD) represents factorization of a rectangular matrix. This process is widely used for data reduction. Latent Semantic Analysis (LSA), which is a nature language processing technique, incorporates SVD to analyze relationships between a set of documents and their terms. You can see more detailed information about SVD and LSA in Landauer (1998).

Let matrix M be composed of several vectors $[M_1, M_2, \ldots, M_n]$, where the vector M_i represents terms occurred in document i. For each term, its frequency is placed into vector M_i. The decomposition of matrix M into three matrices U, Σ and V is expressed by the following equation

$$M = U \times \Sigma \times V^T . \quad (1)$$

U is an m-by-n unitary matrix denoting left singular vectors, i.e. phrase singular vectors. Σ is an n-by-n diagonal matrix without negative and zero numbers, which represents singular values. V is an n-by-n unitary transpose matrix denoting right singular vectors, i.e. document singular vectors.

For our purpose, we need to analyze phrases or, in other words, chunks of sentences. This is represented by the topic of a document and gives us the possibility to identify paraphrased documents. The conventional LSA, employed in Latent Semantic Indexing (LSI), uses a single word as a term. To convey more semantic content, we propose to employ phrases instead of single words. Therefore, we have refined matrix M to compose of vectors M_i representing phrases occurred in document i. Figure 1 presents the decomposition of matrix M in a more detailed fashion.

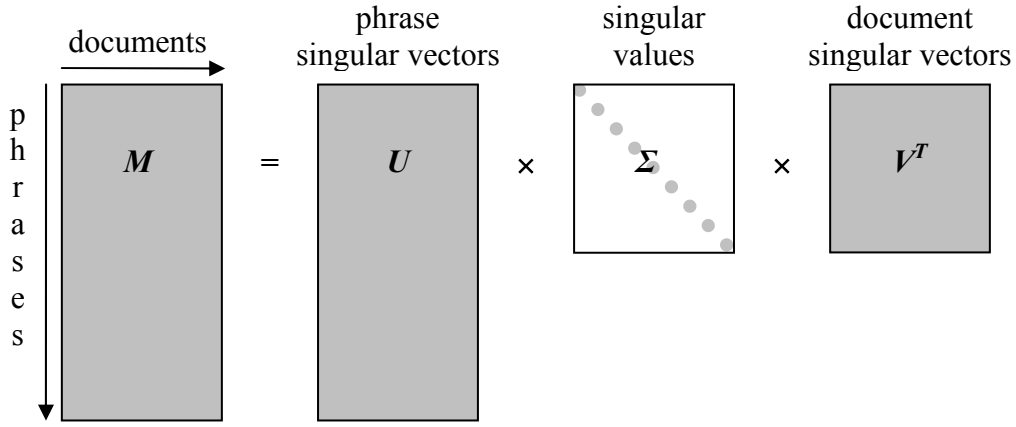

Figure 1: Latent Semantic Analysis

Because every document contains only a subset of all phrases, matrix M is highly sparse. As a result, the computational time depends on the number of non-zero numbers in matrix M. Despite of this fact, if the phrase-space becomes too large, the decomposition will be very time-consuming.

5 ADVANCED PLAGIARISM DETECTION SYSTEM

Modern plagiarism detection methods extract N-grams from documents and then search for an overlap in a database. This is done for each pair of documents in the database to reach successful identification of all plagiarists. We can classify these methods as Paired and Superficial because no semantics in sentences is used. The similarity between two documents R and S is expressible as

$$sim(R,S) = \frac{|gram(R) \cap gram(S)|}{|gram(R) \cup gram(S)|}, \tag{2}$$

where the output of *gram(R)* is a set of produced N-grams for document R.

We are proposing an advanced system, which is built on LSA and might reveal hardly-detectable cases of plagiarism, i.e. paraphrasing. The schema of this system is shown in Figure 2. Our system is designed as Corpal/Multi-dimensional rather than Paired or Singular. This means that we need only one computation to process all documents at a time. Moreover, our system is designed as Structural and utilizes some linguistic rules for more accurate document comparison.

Students' works are stored in a local database in the plain text format. To save memory resources, some parts of the texts can be shared. If wider surroundings need to be included, external data sources such as the Internet can employed. Only one related issue might be a distinct stored mechanism, which requires some necessary data conversion.

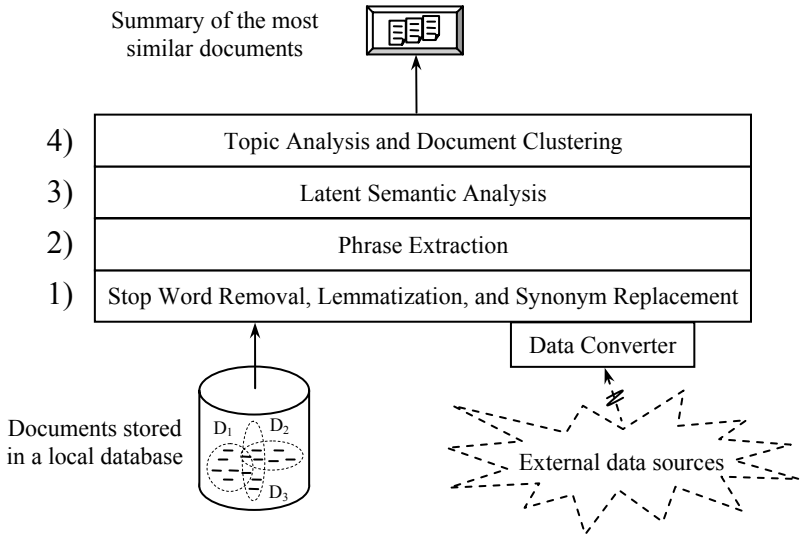

Figure 2: Advanced plagiarism detection system

First, all documents or a set of documents for comparison is extracted from the sources and filters, such as Stop Word Removal, Lemmatization and Synonym Replacement, are applied. These filters have two utilizations. They convert text into a structure with simplified grammar and even reduce the phrase-space, which is necessary for efficient LSA.

The second step deals with phrase extraction from the simplified text. Phrases are extracted as N-grams with one fixed size or several different sizes. Regarding previous experiments on copy detection methods, e.g. Pataki (2003), we can estimate that this would be between 3 and 10 words. To choose the right size, some experiments have to be performed because we must be very careful choosing higher N-gram sizes. If the phrase-space becomes too large, SVD simply takes too much time. The suitable method for N-gram extraction is a sliding window, which can be easily modified so that it could accept punctuation and would prevent creating phrases across two different sentences. This requires a bit deeper analysis because not all dots and commas separate sentences, e.g. decimal point.

The next step is LSA. For each document, the occurrence of the phrases created in the previous step is placed into matrix M. Then it is decomposed into three matrices U, Σ and V.

The last step analyses the decomposed matrices in order to receive an evaluation of the most similar documents. Matrix V^T contains the document singular vectors involving some document cohesion. Along with matrix Σ, where singular values of factorized matrix M are presented, it is possible to compute a document correlation. Then, we propose using document clustering based on this score. As a result, we obtain several clusters where documents from the same cluster are considered plagiarized. Then the shared phrases are identified according to the values in matrix U and their overlap is computed.

6 CONCLUSION

In this paper, we denoted the future directions of copy detection techniques. SVD is one of the most interesting approaches, which has not been employed in this field yet. This is because of

too large phrase- and document-space of matrix *M*. Consequently, we are proposing employing Stop Word Removal, Lemmatization, and Synonym Replacement to reduce the phrase-space. The advantage of these techniques is the simplification of the grammar and better understanding of the semantics within documents. We argue that Synonym Replacement together with LSA can identify paraphrased documents that are hardly-detectable cases of plagiarism.

Our proposed solution is designed as Corpal/Multi-dimensional meaning that a set of documents can be processed at a time. This is another essential advantage in comparison to the other Paired or Singular approaches. We are going to implement this solution and compare our results with the other existing tools. A detail description of the achieved results will be reported in a following paper.

Acknowledgement: This research was supported in part by National Research Programme II, project 2C06009 (COT-SEWing).

REFERENCES

Clough, P., 2000. Plagiarism in natural and programming languages: An overview of current tools and technologies. Internal Report CS-00-05, Department of Computer Science, University of Sheffield.

Clough, P., 2003. Old and new challenges in automatic plagiarism detection. Plagiarism Advisory Service, vol. 10, Department of Computer Science, University of Sheffield.

Maurer, H., Kappe, F., Zaka, B., 2006. Plagiarism – A Survey. Journal of Universal Computer Science, vol. 12, no. 8, pp. 1050-1084.

Shivakumar, N., Garcia-Molina, H., 1995. SCAM: A copy detection mechanism for digital documents. In Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries, Austin.

Shivakumar, N., Garcia-Molina, H., 1996. Building a Scalable and Accurate Copy Detection Mechanism. Proceedings of 1st ACM DL International Conference, Besheda, Maryland.

Brin, S., Davis, J. Garcia-Molina, H., 1995. Copy Detection Mechanisms for Digital Documents. In Proceedings of ACM SIGMOD Annual Conference, San Jose, Canada.

Monostori, K., Finkel, R., Zaslavsky, A., Hodász, G., Pataki, M., 2002. Comparison of Overlap Detection Techniques. Proceedings of the International Conference on Computational Science, Amsterdam, The Netherlands.

Pataki, M., 2003. Plagiarism Detection and Document Chunking Methods. The Twelfth International Word Wide Web Conference, Budapest, Hungary.

Burrows, S., Tahaghoghi, S., Zobel, J., 2004. Efficient and Effective Plagiarism Detection for Large Code Repositories. Proceedings of the Second Australian Undergraduate Students' Computing Conference, pp. 8-15, Australia.

Ceska, Z., 2007. Využití N-gramů pro odhalování plagiátů. Proceedings of ITAT 2007, pp. 63-66, Polana, Slovakia. ISBN 978-80-969184-6-1.

Lancaster, T., Culwin, F., 2005. Classification of Plagiarism Detection Engines. E-journal ITALICS, vol. 4 issue 2, ISSN 1473-7507.

Landauer, T., Foltz, P., Laham, D., 1998. An Introduction to Latent Semantic Analysis. Discourse Processes 25: 259-284.