

Overview and Comparison of Plagiarism Detection Tools

Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, and Václav Snášel

Department of Computer Science, VŠB-Technical University of Ostrava,
17. listopadu 15, Ostrava - Poruba, Czech Republic
asim070@yahoo.com, hussamdahwa@hotmail.com, vaclav.snasel@vsb.cz

Abstract. In this paper we have done an overview of effective plagiarism detection methods that have been used for natural language text plagiarism detection, external plagiarism detection, clustering-base plagiarism detection and some methods used in code source plagiarism detection, also we have done a comparison between five of software used for textual plagiarism detection: (PlagAware, PlagScan, Check for Plagiarism, iThenticate and PlagiarismDetection.org), software are compared with respect of their features and performance.

1 Introduction

"Plagiarism, 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." Encyclopedia Britannica [5].

Plagiarism can be considered as one of the electronic crimes, like (computer hacking, computer viruses, spamming, phishing, copyrights violation and others crimes). Plagiarism defined as the act of taking or attempting to take or to use (whole or parts) of another person's works, without referencing or citation him as the owner of this work. It may include direct copy and paste, modification or changing some words of the original information from the internet books, magazine, newspaper, research, journal, personal information or ideas. According to the Merriam-Webster Online Dictionary, to "plagiarize" means:

- To steal and pass off (the ideas or words of another) as one's own.
- To use (another's production) without crediting the source.
- To commit literary theft.
- To present as new and original an idea or product derived from an existing source.

Also according to Turnitin.com, plagiarism.org and Research Resources this are considered plagiarism:

- Turning in someone else's work as your own.
- Copying words or ideas from someone else without giving credit.
- Failing to put a quotation in quotation marks.

- Giving incorrect information about the source of a quotation.
- Changing words but copying the sentence structure of a source without giving credit.
- Copying so many words or ideas from a source that it makes up the majority of your work, whether you give credit or not (see our section on "fair use" rules).

Plagiarism can be classified into five categories:

1. Copy & Paste Plagiarism.
2. Word Switch Plagiarism.
3. Style Plagiarism.
4. Metaphor Plagiarism.
5. Idea Plagiarism.

There are two types of plagiarism are more occurs:

1. Textual plagiarisms: this type of plagiarism usually done by students or researchers in academic enterprises, where documents are identical or typical to the original documents, reports, essays scientific papers and art design.
2. A source code plagiarism: also done by students in universities, where the students trying or copying the whole or the parts of source code written by someone else as one's own, this types of plagiarism it is difficult to detect.

2 Why Plagiarism Detection is Important

In some of the academic enterprises like universities, schools and institutions, plagiarism detection and prevention became one of the educational challenges, because most of the students or researchers are cheating when they do the assigned tasks and projects. This is because a lot of resources can be found on the internet. It is so easy to them to use one of the search engines to search for any topic and to cheat from it without citing the owner of the document. So it is better and must all academic fields they should have to use plagiarism detection soft-wares to stop or to eliminate students cheating, copying and modifying documents when they know that they will be found.

Some types of plagiarism acts can be detected easily by using some of the recent plagiarism detection soft-wares available on the market or over the internet. However for some of the expert plagiarism who is using some of the anti-plagiarism soft-wares which are available over the internet, it needs more efforts to detect the plagiarism or cannot be detected at all.

Plagiarism is practiced not only by student but also there are some staff members who like to publish papers in which some parts are directly copied or partially modified to be one of the famous people.

There is a big number of plagiarism soft-wares used for plagiarism detection and many of detection tools have been developed by researchers but still they have some limitations as they cannot prove or they show evidence that the documents has been plagiarized from another document or sources it only shows

the similarity and give hints to some other documents. This is if the paper has been published globally in some international journal, but some of universities and some of the research centers still do not taking any action against plagiarism detection which help people to cheat more and more.

So still now by using the recent detection software, plagiarism can not 100% be detected?

Copyrights and legal aspects for use of published documents also can be covered by using plagiarism software, so it can show whether this person has legally or illegally copied the documents or not and it also show the whether this person has permission from the owner to use this document or not.

Plagiarism detection is also one of the most important issues to journals, research center and conferences; they are using advanced plagiarism detection tools to ensure that all the documents have not been plagiarized, and to save the copyrights from violation for the publishers.

3 Plagiarism Detection Methods

In both the textual document plagiarism and source code plagiarism, detection can be either: Manual detection or automatic detection.

- Manual detection: done manually by human, its suitable for lectures and teachers in checking student's assignments but it is not effective and impractical for a large number of documents and not economical also need highly effort and wasting time.
- Automatic detection (Computer assisted detection): there are many software and tools used in automatic plagiarism detection, like PlagAware, PlagScan, Check for Plagiarism, iThenticate, PlagiarismDetection.org, Academic Plagiarism, The Plagiarism Checker, Urkund, Docoloc and more.

3.1 Textual Plagiarism

Many of researchers are developed a set of tools used in textual automatic detection like:

Grammar-based method The grammar-based method is important tool to detect plagiarism. It focuses on the grammatical structure of documents, and this method uses a string-based matching approach to detect and to measure similarity between the documents. The grammar-based methods is suitable for detecting exact copy without any modification, but it is not suitable for detecting modified copied text by rewriting or switching some words that has the same meaning. This is considered as one of this method limitations [4].

Semantics-based method The semantics-based method, also considered as one of the important method for plagiarism detection, focuses on detecting the similarities between documents by using the vector space model. It also can calculate and count the redundancy of the word in the document, and then they use the fingerprints for each document for matching it with fingerprints in other documents and find out the similarity. The semantic-based method is suitable for non partial plagiarism as mentioned before use the whole document and use vector space to match between the documents, but if the document has been partially plagiarized it cannot achieve good results, and this is considered as one of the limitations of this method, because it is difficult to fix the place of copied text in the original document [4, 1].

Grammar semantics hybrid method Grammar semantic hybrid method is considered as the most important method in plagiarism detecting for the natural languages. This method, so effective in achieving better and improving plagiarism detection result, is suitable for the copied text including modified text by rewriting or switching some words that have the same meaning, which cannot be detected by grammar-based method. It also solves the limitation of semantic-based method. Grammar semantic hybrid method can detect and determine the location of plagiarized parts of the document, which cannot be detected by semantic-based method, and calculating the similarity between documents [4, 1].

External plagiarism detection method The external plagiarism detection relies on a reference corpus composed of documents from which passages might have been plagiarized A passage could be made up of paragraphs, a fixed size block of words, a block of sentences and so on. A suspicious document is checked for plagiarism by searching for passages that are duplicates or near duplicates of passages in documents within the reference corpus. An external plagiarism system then reports these findings to a human controller who decides whether the detected passages are plagiarized or not. A naive solution to this problem is to compare each passage in a suspicious document to every passage of each document in the reference corpus. This is obviously prohibitive. The reference corpus has to be large in order to find as many plagiarized passages as possible [20].

This fact directly translates to very high runtimes when using the naive approach. External plagiarism detection is similar to textual information retrieval (IR) [3]. Given a set of query terms an IR system returns a ranked set of documents from a corpus that best matches the query terms. The most common structure for answering such queries is an inverted index. An external plagiarism detection system using an inverted index indexes passage of the reference corpus' documents.

Such a system was presented in [7] for finding duplicate or near duplicate documents.

Another method for finding duplicates and near duplicates is based on hashing or fingerprinting. Such methods produce one or more fingerprints that de-

scribe the content of a document or passage. A suspicious document's passages are compared to the reference corpus based on their hashes or fingerprints. Duplicate and near duplicate passages are assumed to have similar fingerprints. One of the first systems for plagiarism detection using this schema was presented in [2]. External plagiarism detection can also be viewed as nearest neighbor problem in a vector space R^d .

Clustering in plagiarism detection Document clustering is one of the important techniques used by information retrieval in many purposes; it has been used in summarization of the documents to improve the retrieval of data by reducing the searching time in locating the document. It is also used for result presentation. Document clustering is used in plagiarism detection to reduce the searching time. But still now in clustering there are some limitations and difficulties with time and space [8].

Most of the above methods have been used by textual documents plagiarism detection.

3.2 Source code plagiarism

Source code plagiarism or it called programming plagiarisms usually done by students in universities and schools can be defined act or trial to use, reuse, convert and modify or copy the whole or the part of the source code written by someone else and used in your programming without citation to the owners. Source code detection mainly requires human intervention if they use Manual or automatic source code plagiarism detection to decide or to determine whether the similarity due to the plagiarism or not. Manual detection of source code in a big number of student homework's or project it is so difficult and needs highly effort and stronger memory, it seems that impossible for a big number of sources.

Plagiarism detection system or algorithms used in source-code similarity detection can be classifies according to Roy and Cordy [9] can be classified as based on either:

- 'Strings - look for exact textual matches of segments, for instance five-word runs. Fast, but can be confused by renaming identifiers'.
- "Tokens - as with strings, but using a lexer to convert the program into tokens first. This discards whitespace, comments, and identifier names, making the system more robust to simple text replacements. Most academic plagiarism detection systems work at this level, using different algorithms to measure the similarity between token sequences".
- "Parse Trees - build and compare parse trees. This allows higher-level similarities to be detected. For instance, tree comparison can normalize conditional statements, and detect equivalent constructs as similar to each other".
- "Program Dependency Graphs (PDGs) - a PDG captures the actual flow of control in a program, and allows much higher-level equivalences to be located, at a greater expense in complexity and calculation time".

- "Metrics - metrics capture 'scores' of code segments according to certain criteria; for instance, "the number of loops and conditionals", or "the number of different variables used". Metrics are simple to calculate and can be compared quickly, but can also lead to false positives: two fragments with the same scores on a set of metrics may do entirely different things".
- "Hybrid approaches - for instance, parse trees + suffix trees can combine the detection capability of parse trees with the speed afforded by suffix trees, a type of string-matching data structure".

There are many methods developed by researcher for source code plagiarism detection like:

- Cynthia Kustanto and Inggriani Liem: they developed a tool for automatic source code detection call Deimos, used in source plagiarism detection, to provide a clear readable form and to erase the displayed result. It was develop to be used with LISP and Pascal programming languages. The time consumed by this tool for section a number of 100 LISP was efficient [11].
- Boris Lesner, Romain Brixtel, Cyril Bazin and Guillaume Bagan: they introduced a new frame work named A Novel Framework to Detect Source Code Plagiarism, mainly used in detection of four type of code source plagiarisms which are change the code name, rebuilt or recoded again, move, add, change and remove the code and replace some text from place to place with the code. A bottom-up approach has been implemented to six steps which are: 1- first the Pre-flattering the source code: they use common method in filtering a source code that by indicating and rename each alphanumerical string in the code. 2- Second they segment the source code to segmentation and measure the similarity on it 3- thirdly they matched each segment and reposted it for filtering. 4-5: Fourthly the use matrix M that have been used in filtering in evaluation of the document 6- In this stage start to analysis the original document according to the evaluation done by document wise distance. This method has been applied to copra languages and shows a great result [12].
- Ameera Jadalla and Ashraf El Nagar: They develop Plagiarism Detection Engine used for detection of source code plagiarism for Java (PDE4Java). The proposed search engine divided in to three steps 1- step one is the process of the tokenization for the Java code 2- second step is to find and measure the similarity between the original code and the tokenized code 3- lastly is to cluster the Java code in order to be used in plagiarism detection as reference. This search engine can be used with all programming language due to its flexibility. Report can show for each cluster code besides the textual [10].

4 Comparisons

We compare the plagiarism software used in textual and source code plagiarism into two categories: first according to features and secondly according to performance [6]. Qualitative comparison used in comparing the features of software, where we are looking for properties of the tools. Quantitative comparisons used

in comparing the performance of software, where it depends on the result. Here comparison of some textual soft-wares:

4.1 PlagAware

Is an online-service used for textual plagiarism detection, which allows and offers some services for the user for example can search, find, analyze and trace plagiarism in the specified topic similar to the topics, PlagAware is a search engine, which is considered as the main element, which is strong in detecting typical contents of given texts. It uses the classical search engine for detecting and scanning plagiarism, and provide different types of report that help the user or the document owner to decide that is his document has been plagiarized or not. The two primary fields for PlagAware plagiarism search engine is webpage monitoring for theft contents and transmitted text assessment. In [13], there are three application fields of PlagAware [14]:

- Tracing content theft: Webmaster can use PlagAware for detecting and tracing plagiarisms of websites, in order to find out the plagiarized or the copied contents. PlagAware is considered as strong total solution software systems, which allow the operators of websites to do an automatic observation of their own pages against possible content theft.
- PlagAware is used in search for plagiarisms of student’s academic documents and analyze them. Also it is used to assess plagiarism, also to follow and prove the origin of the works including all of academic documents. It generates report that helps them to fast detection of plagiarism.
- Proof of authorship is also provided by PlagAware: it became more important to the authors to ensure that the authorization have been granted to their publication including all types of publication this gives them additional competitive advantage and increase the value of your work.

The main features of PlagAware are [15]:

- Database Checking: PlagAware is a search engine that allows the user to submit his document and Plagaware start searching over the internet. So mainly it does not have local database but it offers checking other database that are available over the internet.
- Internet Checking: PlagAware is an online application and it considered as one of search engine, allows the student or webmaster to upload and check their academic documents, homework, manuscript and articles to be searched against plagiarism over world wide web.ans also provides a webmaster to have capability to do automatic observation of their own page against possible contents theft.
- Publications Checking: PlagAware: support mainly used in academic filed so it provides checking of most types of submitted publication like homework, manuscript, documents, including, books, articles, magazines, journals, editorial and PDFs etc.

- Synonym and Sentence Structure Checking: PlagAware does not support synonym and sentence structure checking.
- Multiple Document Comparison: PlagAware offers comparison of multiple documents.
- Supported Languages: PlagAware supports German as primary language, English and Japanese as secondary languages.

4.2 PlagScan

PlagScan is online software used for textual plagiarism checker. PlagScan is often used by school and provides different types of account with different features. PlagScan use complex algorithms for checking and analyzing uploaded document for plagiarism detection, based on up-to-date linguistic research. Unique signature extracted from the document's structure that is then compared with PlagScan database and millions of online documents. So PlagScan is able to detect most of plagiarism types either directs copy and paste or words switching, which provides an accurate measurement of the level of plagiarized content in any given documents [16]. The Main features of PlagScan are [15]:

- Database Checking: PlagScan it has own database that include millions documents like (paper, articles and assignments), and articles over World Wide Web. So it offers database checking whether locally or others database over the internet.
- Internet checking: PlagScan is an online checker so it provides internet checking to all submitted documents. Whether that the document available on the internet or available in the local database or cached.
- Publications Checking: PlagScan: is mainly used in academic filed so it provides checking most types of submitted publication like documents, including, books, articles, magazines, journals, newspapers, PDFs etc. online only.
- Synonym and Sentence Structure Checking: PlagScan does not support synonym and sentence structure checking but provides Integration via application programming interface in your existing content management system or learning management system possible.
- Multiple Document Comparison: CheckForPlagiarism.net offers comparison of multiple documents in parallel.
- Supported Languages: PlagScan supports all the language that use the international UTF-8 encoding and all language with Latin or Arabic characters can be checked for plagiarism.

4.3 CheckForPlagiarism.net

CheckForPlagiarism.net was developed by a team of professional academic people and became one of the best online plagiarism checkers that used to stop or prevention of online plagiarism and minimizes its effects on academic integrity. In order to maximize the accuracy CheckForPlagiarism.net has used the some

methods like document fingerprint and document source analysis to protect document against plagiarism. The fingerprint-based approach used to analyze and summarize collection of document and create a kind of fingerprint for it. Some of numerical attributes can be used by fingerprint that somehow reflects in the structure of the document. So by creating fingerprint for each document with some of numerical attributes for each document in the collection, we can easily find the matching or the similarity between documents across billions of articles. Using this feature by CheckForPlagiarism.net increased the efficiency in detecting most types of plagiarisms [17]. The main features of CheckForPlagiarism.net are [15]:

- Database Checking: CheckForPlagiarism.net uses its own database that include millions documents like (paper, articles and assignments), and articles over World Wide Web. So it offers fast and reliable depth database checking, also provides checking through all other databases in different fields like medical database, law- related database and other specialty and generalized databases.
- Internet Checking: CheckForPlagiarism.net: live(online) and cached links to websites used for extensive internet checking to all submitted documents. One more advantage is that it can still check your documents against if a website that is no longer online, this include all contents of website like forums, message boards, bulletin boards, blogs, and PDFs etc., all this check is done automatically and in (almost) real-time.
- Publications Checking: CheckForPlagiarism.net offers detailed and deep checking of most types of submitted publication documents, including, books, articles, magazines, journals, newspapers, PDFs etc. this is done whether the publications is available online (active on the internet) or not available on the internet offline (store paper based).
- Synonym & Sentence Structure Checking: CheckForPlagiarism.net is said to have a sole advantage, that other soft-wares do not support, which is the fact that it uses a "patented" plagiarism checking approach. In which the sentence structure of a document is checked to ensure improper paragraphing and thus is susceptible to plagiarism. Also a synonym check is done to words and phrases to identify any attempt of plagiarism.
- Multiple Document Comparison: CheckForPlagiarism.net can compare a set of different documents simultaneously with other documents and can diagnose different type of plagiarisms at the sometimes [15].
- Supported Languages: CheckForPlagiarism.net supports English languages, Spanish, German, Portuguese, French, Italian, Arabic, Korean, and Chinese languages [15].

4.4 iThenticate

iThenticate one of the application or services designed especially for the researchers, authors' publisher and other. It provided by iParadigms that have introduced Turnitin in 1996 to become the online plagiarism detection. It is designed to be used by institutions rather than personal, but lastly they provided

a limit service for single plagiarism detection user like master and doctoral students and this allows them to check a single document of up to 25,000 words. So they can use this service to insure or to check their draft thesis whether containing correct citation and content originality [18]. The main features of iThenticate are [15]:

- Database Checking: iThenticate used its own database that contain millions of documents like (books, paper, essays, articles and assignments), with a large number of this documents that have been stored in iThenticate database locally, allowing the users who have account to do either online and offline comparison of submitted documents against it and to identify plagiarized content.
- Internet Checking : iThenticate, is considered as the first online plagiarism checker that provides live and cached links to websites and database to have extensive internet checking to all submitted documents. This Provides deep internet checking. One more advantage is that it can still check your documents even if a website is no longer online, this include all contents of website like forums, message boards, bulletin boards, blogs, and PDFs etc., all this check is done automatically and in (almost) real-time.
- Publications Checking: iThenticate offers an online and offline detailed and depth checking most types of publication like documents, including, books, articles, magazines, journals, newspapers, website and PDFs etc.
- Synonym & Sentence Structure Checking: Not supported by iThenticate.
- Multiple Document Comparison: iThenticate offers two types of document comparison document to document and multiple documents checking against database and also direct source comparison word to word also.
- Supported Languages: iThenticate supports more than 30 languages, it mean that it supports most of languages likes "English, Arabic, Chinese, Japanese, Thai, Korean, Catalan, Croatian, Czech, Danish, Dutch, Finnish, French, German, Hungarian, Italian, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovak, Slovenian, Spanish, Swedish, Greek, Hebrew, Farsi, Russian, and Turkish." [18].

4.5 PlagiarismDetection.org

PlagiarismDetection.org: an online service provides high level of accuracy result in plagiarism detection. Mainly designed to help the teachers and student to maintain and to ensure or prevent and detect plagiarism against their academic documents. It provides quickly detect plagiarism with high level of accuracy [19]. The main features of PlagiarismDetection.org [15]:

- Database Checking: PlagiarismDetection.org used it own database that contains millions of documents like (books, paper, essays, articles and assignments).
- Internet Checking: PlagiarismDetection.org is an online plagiarism detector, so it is mainly based on the internet checking and is faster in plagiarism detection, it does not support offline detection.

- Publications Checking: PlagiarismDetection.org offers the students and teachers to check their publication against the published document and support most types of publication.
- Synonym & Sentence Structure Checking: PlagiarismDetection.org not supports Synonym & Sentence Structure Checking.
- mMultiple Document Comparison: PlagiarismDetection.org does not support multiple document comparison but it takes long time to return the result.
- Supported Languages: PlagiarismDetection.org supports English languages and all languages that using Latin characters.

Table1: Summarization of the comparison according to the software features: Key of the table figure1: The following expressions denote that: ***** Excellent, **** Very good, *** Good, ** Acceptable and * Poor.

Table 1. Comparison of the software

Features	PlagAware	PlagScan	iThenticate	CheckForPlagiarism.net	Plagiarismdetecting.org
Database Checking (online and offline)	*****	*****	*****	****	*****
Internet Checking	*****	*****	*****	*****	*****
Publication Checking	*****	*****	*****	***	***
Multiple document comparison	*****	*****	*****	*****	*****
Multiple languages support	*****	*****	*****	*****	*****
Sentence structure and synonym checking	****	**	****	*****	**

5 Conclusions

The comparison of the software shown that still now their no software that can detect or to prove that the document has been plagiarize 100%, because each software and tool has advantages and limitation, according to the following features and performance, (Database checking whether locally or online, internet checking whether is online or offline, publications documents checking and supported types, capabilities of multiple document comparison, supported languages and synonym and sentence structure checking) we ranked them as follows, starting from the best, PlagAware, iThenticate, PlagScan, CheckForPlagiarism.net and lastly PlagiarismDetection.org. Academic enterprises can use one of the above software in their detecting of plagiarism but, due the limitation of most of the software and importance of plagiarism detection to the academic fields we suggest some rules that can be used to limit or to reduce student plagiarism teacher should educated student about plagiarism and its impact, copy right, citation and ownership.

In future work we may want to extend our comparison to larger and more varied set of real-life data and to extent our comparison to include more textual plagiarism software like Academic Plagiarism, The Plagiarism Checker, Urkund

and Docoloc, also to extended to include code source plagiarism detection software according to Supported languages, Extendibility, Presentation of results, Usability, Exclusion of template code, Exclusion of small files, Historical comparisons, Submission or file- based rating, Local or web-based and Open source

References

1. Alzahrani, S. Salim, N. Abraham, A. Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods, preprint 2011.
2. Brin, S. and Davis, J. and Garcia-Molina, H. Copy Detection Mechanisms for Digital Documents. In: ACM International Conference on Management of Data (SIGMOD 1995), May 22-25, 1995
3. Baeza-Yates, R. & Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley, 1999.
4. Bao Jun-Peng, Shen Jun-Yi, Liu Xiao-Dong, Song Qin-Bao, A Survey on Natural Language Text Copy Detection, Journal of Software, 2003, vol.14, No.10, pp. 1753-1760.
5. Encyclopedia Britannica, <http://www.britannica.com/EBchecked/topic/462640/plagiarism> (last access February 7, 2011)
6. Hage, J. Rademaker, P. Vugt, N. A comparison of plagiarism detection tools, Technical Report UU-CS-2010-015, June 2010, Department of Information and Computing Sciences Utrecht University, Utrecht, The Netherlands.
7. Hoad, T. C., Zobel, J. Methods for Identifying Versioned and Plagiarized Documents. JASIST 54(3): 203-215 (2003)
8. Jain, K. Murty, M. N., Flynn, P. J. Data clustering: a review. ACM Computing Surveys, 31(3):264–323, 1999.
9. <http://www.cs.queensu.ca/queensu.ca/TechReports/Reports/2007-541.pdf>. (last access February 7, 2011)
10. Jadalla, A. Elnagar, A. PDE4Java: Plagiarism Detection Engine for Java source code: a clustering approach. IJBIDM 3(2):121-135 (2008)
11. Kustanto, C. Liem, I. Automatic Source Code Plagiarism Detection. SNPD 2009:481-486.
12. Lesner, B. Brixtel, R. Bazin, C. Bagan, G. A novel framework to detect source code plagiarism: now, students have to work for real! SAC 2010:57-58.
13. http://www.plagaware.com/about_plagaware (last access February 7, 2011)
14. http://www.plagaware.com/about_plagaware/application (last access February 7, 2011)
15. <http://plagiarism-checker-review.toptenreviews.com/index.html> (last access February 7, 2011)
16. <http://www.plagscan.com>. (last access February 7, 2011)
17. <http://www.checkforplagiarism.net> (last access February 7, 2011)
18. <http://www.ithenticate.com/index.html> (last access February 7, 2011)
19. <http://www.plagiarismdetection.org> (last access February 7, 2011)
20. Zechner, M., Muhr, M., Kern, R., Michael, G. External and intrinsic plagiarism detection using vector space models. In: Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. pp. 4755 (2009).