INTER-UNIVERSITY COOPERATION ON PLAGIARISM DETECTION SYSTEMS IN CZECH REPUBLIC

Ľuboš Lunter, Daniel Jakubík, Šimon Suchomel, Michal Brandejs

Abstract: Plagiarism detection systems can help to discover plagiarised material by finding similar documents. The field of plagiarism detection has been dealt with by Masaryk University since 2006 when the users of the Information System of Masaryk University (IS MU) started using e-learning tools including electronic study materials. In 2008, the joint initiative of most public universities in the Czech Republic resulted in development of the National Register of Theses and plagiarism-tracking system Theses.cz. This was followed by the start of the system tracing plagiarisms in seminar papers Odevzdej.cz. Repozitar.cz continues the inter-university control of plagiarisms in the field of scientific and professional publications. The paper shows how these systems work and discuss the benefits of inter-university cooperation. Plagiarism analysis of the document can detect the suspicious similarities with identified source document.

Introduction

In the past few years there is an obvious trend of keeping and processing documents electronically. The main source of electronically available documents is the Internet. It contains not only Web pages, but also vast collection of other text documents such as scanned books or archived theses. There are lots of benefits in having documents electronically available and easily accessible, but it also increases the risk of misuses for plagiarism.

The area of plagiarism is very huge and it can cover many creative human activities in lots of different forms. However, it is very easy to plagiarise an electronic document on contrary to detection of corresponding falsification. Therefore it is important to develop and provide some automated systems which can help to detect potentially plagiarized documents. As reaction for this increasing need, several anti-plagiarism systems have been developed (Brandejs, M. et al., 2009). For purposes of this paper, the anti-plagiarism system will be understood as a service focused on discovering potential plagiarism in a given document.

In this paper we provide an insight into the family of systems for plagiarism detection developed by Masaryk University. Furthermore we introduce well-established procedures for preventing and dealing with plagiarism, which are used at Masaryk University and which are also recommended to other institutions using these systems. Providing anti-plagiarism services (for more than 40 institutions) offers the wide field of experiences. Last but not least the benefits of joint effort in anti-plagiarism activities will be presented.

Text bellow is organized as follows. In section 2, the issue of plagiarism in academic world and reaction of Masaryk University will be discussed. That also include brief introduction of systems Theses.cz, Odevzdej.cz and Repozitar.cz. Section 3 focuses on detecting similarities against Internet sources. Section 4 summarizes benefits of presented approach and finally, in section 5, we give our conclusion.

Background

The issue of plagiarism has recently been intensively discussed in academic and scientific circles. The increasing availability of online data such as learning materials, theses or scientific articles tempts students, and sometimes even teachers, to submit someone else's results as theirs. This situation calls for tools capable of revealing instances of plagiarism, which would, at the same time, function as a deterrent to the would-be copyright violators.

As a reaction for this trend, developer team of Information System of Masaryk University have created a service for plagiarism detection which was initially intended only for university purposes. Consequently, the service has become the basis of the platform on which the anti-plagiarism systems Theses.cz, Odevzdej.cz and Repozitar.cz were built. The core of this uses proprietary distributed chunk-based algorithm. From the text form of each document, there are extracted overlapping chunks of consecutive five words. The algorithm evaluates portion of the same chunks found among documents. More information about used algorithm can be found in (Kasprzak et al., 2008)

For successful fight again plagiarism it is very important to hold huge base of data. Therefore for all three systems their mutual interconnection is characteristic, thus similarity searches can be done across all files they contain. Thanks to the quality of modern search engines and the huge amount of online data it was also desirable to include other sources inclusive of the Internet, detailed description can be found in next section "Detecting similarities against Internet sources".

Thanks to the common platform developed systems can share even other useful, during years well-approved, features. They provide several options how records, consisting of meta-data and set of attachments, can be inserted into the system. That includes both manual method for individual or bulk insert by users via system-specific web forms and different forms of automatic downloading from remote systems of participated universities and other institutions. Bulk uploads of records is using xml files with specific semantic and syntax. Each system has its own set of allowed elements given by its individual purposes. The most commonly used automatic methods are based on protocols OAI-PMH or transmission of the data using the program Curl. Once the record is stored, the embedded OCR system then automatically converts the attached files into the plain text and PDF format. Typically, the PDF file serves to public presentation and text format is used for further processing including full-text search system indexing, splitting into chunks of text for finding these chunks inside other documents, and other tasks of text analysis. For participants using automatic methods the plagiarism detection task is most important and they would like to have this information available even in their own information systems. Therefore as soon as the similarities are computed they are automatically exported back to the interconnected system.

The access interface of all the systems more or less uses common full-text search engine which indexes both meta-data as well as the full text of the thesis, seminar papers, articles or other publication records. Last but not least the common communication tools should be mentioned. They include discussion forums which enable users to

share their problems, advice or experiences with other users, and message boards for mass communication.

Bellow, the description of individual systems follows, focusing on their specific characteristic.

Theses.cz

In 2006, Masaryk University provided its students and teachers a unique service that helped them to identify potentially plagiarized parts in the thesis. At that time, demand for a similar service arose at other universities as well. This demand represented an impetus for the joint project Theses.cz—National Registry of Theses and Plagiarism-Tracing System. It has started in 2008 and involves Masaryk University as well as the 16 other universities. Later on, some other universities joined the system. Nowadays there are 38 (36 Czech and 2 Slovak) universities involved in the project altogether.

The main objective of this system is to provide plagiarism detection for schools involved. Theses.cz is designed to cover different requirements. Every institution uses the system in own way according to internal procedures. The most of customers collect the theses in electronic mode via local information system. It turned out that theses' supervisors necessarily need effective tools for exploring the originality of submitted works. The system also serves as an archive of theses, with the option to publish metadata and full texts in several modes according to individual configuration. Although the recommended configuration is to publish all theses in accordance with The Higher Education Act and Open-Access principles, many of the universities involved hide the full texts. However, there are not hidden from the plagiarism detection system. Thanks to the Theses.cz, the unpublished documents enrich the corpus, too.

Odevzdej.cz

Even before the emergence of Theses.cz we held discussions with representatives of other universities on the topic of plagiarism in seminar papers. Originally, we thought that it would be possible to use Theses.cz for this kind of students' works. Due to different copyright mode of seminar papers, this showed up later as not optimal solution. In the Theses.cz there are stored metadata for each thesis and system has to solve accessibility of metadata and full texts. On the contrary, generally there is no need to collect any metadata and publish seminar papers. Therefore we came with an idea of individual system for seminar papers shortly after the introduction of Theses.cz. At that time, 9 other universities in the Czech Republic decided to participate in the project of developing system for plagiarism detection in seminar papers—Odevzdej.cz. (Brandejs, M. et al., 2009)

Odevzdej.cz provides partial e-learning solution for collecting seminar papers and essays to schools involved into the project. Teachers create so-called Vaults—the folder where students upload their papers. This eliminates the administration associated with the collection of papers, such as personal communication by e-mail and individual downloading attachments or subsequent archiving. There is a list of actual submissions in the Vault which is always available for the teacher. The system automatically closes submitting to the Vault in accordance to the deadline, which can be set in the Vault

options. Odevzdej.cz provides to teachers tools for managing, plagiarism-detecting and evaluation of uploaded papers.

Odevzdej.cz is open to anybody. Everyone can check 3 papers a day without the need to register in the system. The developer team was inspired to this approach by experiences with Theses.cz. There was a high demand for inspection of papers by public users (without an authenticated access into any of the mentioned systems). The uniqueness of such free access to the system lies in the fact that everybody has the opportunity to inspect the document for plagiarism and thus contributes to greater copyright protection. User can upload the file he wishes to check via the uploading form at the title page of the system. After processing uploaded document, the system sends results to the user-specified e-mail address. The files are automatically deleted in 5 days and nobody else can see the similarity with these temporary files.

Repozitar.cz

System Repozitar.cz has arisen in order to provide services for long-term storage and presentation of academic papers and other scientific outputs (Jakubík et al., 2011). It serves the purpose of digital library and includes technical solution form the necessary organizational, social and legal environment. According to the Open Access idea there is an effort to make the most of records open to the public but for various reasons it is not possible to fully achieve the goal. Therefore the system handles a wide range of access rights through which the authors can restrict the access to their publication outputs.

The main access interface of the system is based on a full-text search which indexes both meta-data as well as the full text of the articles and other publication records. This full-text search system is also used for searching of similar documents or generating various lists through its ability to add any additional information in the form of "virtual tokens" to the index. The advanced search options extend the application by adding the possibility of progressive refinements to the query. It allows search not only by publication own meta-data but also by departments, R&D projects or other data required for transmission to the RIV. To each one of found records, the seeker can display abstract, list of citations, attached files, similar documents and other detail information. Additionally, a final list of matching records can be entered into a user box which enables the transfer of a selection into other applications, work with it and subsequently process them en bloc.

Detecting similarities against Internet sources

The easiest way for plagiarizing is to use favorite search engine to obtain an online available document and use this document as a source of plagiarism. The task of candidate document retrieval is to identify for a certain input document relatively small set of source documents which may be plagiarized from. Those candidate documents are usually further processed using detail document comparison methods in order to discover potential plagiarism. It is up to the detail comparison method what kind of plagiarism it is able to discover. Nevertheless the standard methods use algorithms

based on document similarities which may by infeasible to compute among large collection of documents which is in case of online plagiarism the whole Web.

Since the detailed document comparison is based on computing similarities among documents which the system indexes we need to provide the system with the source documents which may have served for plagiarism. The computational complexity and storage demands of vast amount of documents prevent us from crawling and indexing the whole Web. The most straightforward method of candidate document retrieval is utilizing ideally the same search engine as the plagiarists did. Latter-day commercial search engines like Google, Yahoo or Bing are indexing the World Wide Web continuously. Despite the fact, that they cannot provide fully up to date data, we can simply rely on them, since the plagiarist could not also find a source which has not yet been indexed by that search engine. Commercial search engines do not provide access to their internal index, nor do they usually publish used retrieving methods in detail. Therefore we are using standard query interface the very same as the regular users do. The problem of candidate document retrieval is than converted into the problem of constructing befitting queries for the selected search engine. Accordingly it is wise to use more than one search engine. Our research shows that generally more search engines provide slightly different results on the basis of the same query.

We must apply several considerations on constructing queries from suspicious documents. The first consideration is to minimize the total number of executed queries, because the number of executed queries is usually limited by the search engine or not free of charge. Also every query takes not insignificant amount of time to process. Second consideration is to decide how many and which of the resulting documents to download and process which influence the performance measure. We must also maximize the precision and recall of the downloaded documents regarding actual potential plagiarism sources of the suspicious document.

There is a need to obtain only relevant documents. We define a document relevant if it is similar to the input document in the way of document similarity computed by our document similarity evaluation algorithm or if it discusses the same theme.

Two documents following the same theme usually share a set of overlapping words. Consequently querying quality keywords extracted from a given text would result in obtaining relevant documents following the same theme. Assume that an input document does not cover more than one coherent theme. That is the case of for example theses and seminar works which we mainly focus on. On the contrary for example one edition of newspapers usually contains many different articles. Based on this assumption we extract key-words from the whole document. This means that they cover the document as a whole and serve for obtaining theme-bounded documents. We developed an in-house key-word extraction algorithm currently supporting 5 European languages, which is easily extensible. The algorithm is based on term frequency analysis combined with the TF-IDF statistical measure.

For constructing queries, we are using similar approach as we used during the international competition on plagiarism detection PAN 2012 published in (Suchomel, Kasprzak & Brandejs, 2012). Our approach resulted in the achievement of the best results in overall performance of the system. More information about the course of the competition can be found in (Potthast et al., 2012). The team of IS MU has competed

in the international competition on plagiarism detection PAN three times. Firstly in 2009 with overall second place in the External Plagiarism Detection task. Secondly in 2010 with overall first place in the Plagiarism Detection task. Thirdly in 2012 with second place in the Detailed Document Comparison task and with the best performing approach in the Candidate Document Retrieval task.

As opposed to the published method, we are not using the queries based on extracted headers from suspicious document since they provide the least performance benefit. We are using two types of queries: the key-words based queries; and the intrinsic plagiarism based queries.

Since December 2010 our system checked about 300 thousand documents for plagiarism against the Internet. It is enriched by the candidate document retrieval process of more than 6 million relevant documents from aimed Internet search and about 7 million Web pages from the Wikipedia. The Wikipedia is crawled and stored on regular basis, since it is often used as a quality information source and may easily serve as a source of plagiarism.

A question we wanted to answer by means of utilizing the system is whether the Internet is in real time "inexhaustible" source of documents. In other words whether there still will be loads of new URLs found as relevant according to our search method for similar documents. Our research shows that the number of unique URLs found is significantly lowering in time of processing due to the fact, that we have downloaded lots of URLs which are repeated among searches. For example after two months of system running there were added during the next month 130 thousand new unique URLs for download and another 250 thousand were also found but rejected, because they have been already downloaded.

Plagiarism evaluation

There are several aspects of potential plagiarism that need to be considered. Percentage of similarity is only one of them. The outline of the plagiarism detection is a list of the similarity rates (percentage of similarity) linked with the source documents. Described systems provide the ability to explore each similarity by clicking the link which opens your document with highlighted suspicious similar passages. The user has to inspect the highlighted text and evaluate compliance with citation ethics and the range of citations whether they correspond to the type of paper/thesis. The problem is that some users expect from the system decision whether the paper/thesis contains plagiarism or not. After the detailed examination of all aspects (of all similarities from the list), the final evaluation could be formulated.

In theses context, there is very important the ongoing cooperation between students and their advisor. The quality of the thesis can be controlled better through incremental paragraphs than final text. The advisor can affect the content, form of the thesis and also citations. The final decision about an examined text should be the collaborative decision of Thesis Defense Committee or Disciplinary Board. Any system cannot be responsible for evaluation and following sentence whether a thesis is in accordance with requirements on the specific type of work. The advisor, opponent and Thesis Defense Committee are guarantors of the quality. They also have the option to refuse

the thesis. After the thesis has been defended, additional objections can hardly be taken into account. Therefore, it is strongly recommended to check and evaluate plagiarism by advisor (or other responsible person) before the thesis defense.

Checking of scientific articles and other works without student/advisor relation must follow different approach. Therefore on-line availability of plagiarism detection module for regular users (not only administrators, teachers or persons with "high" access rights) is characteristic for all three systems. This ensures that every document can be controlled at any time by anyone.

Even if very well similarity detection is available to users, the issue of plagiarism can never be decided by a computer system. The final verdict about plagiarism content must be done by a human; the system tries to facilitate the decision process and point out suspicious documents or suspicious document parts. (Suchomel, 2012)

Reflection of inter-university cooperation

The area of the plagiarism is very large and anti-plagiarism systems can be utilized in many different cases. The system Theses.cz can help the teacher to identify suspicious paragraphs in theses and seminar papers. Contrary, the students can benefit from public access to the Odevzdej.cz and inspect their own theses before submission. They can easily check the correctness of citations or consult the range of passages taken over from another author. Electronically available scientific papers could be abused similarly as freely accessible theses. Therefore, system Repozitar.cz combines functionality of digital library with plagiarism detection.

A significant benefit is the prevention effect of plagiarism detection systems. The risk of revelation connected with systematic usage of anti-plagiarism systems by tens of universities in the region further reduces the determination to misuse available electronic sources. To multiply the effect, it is important to raise awareness of citation ethics and copyrights. From our experience that leads to increasing quality of all types of academic works and possible improvement of the university prestige. Moreover, the obligation of publishing theses on the Internet naturally influences on the quality.

The systematic use of plagiarism detection tools requires integration to the common systems. A professional anti-plagiarism tool needs quality similarity detection algorithms which can scale up to the web. Also integration into internal guidelines and Study and Examination Regulations is needed. The students' theses, seminar papers, essays, etc. should be inspected not just randomly. Institutions should have clear rules and competencies to detect plagiarism.

Another interesting outcome on the quality of all types of works should have their publication on the Internet. People act naturally more responsibly knowing that their text will be obtainable.

Conclusion

In the age of information technologies the problem of plagiarism has become more frequent and turned into a serious issue. In this paper, we described its' complexity and the role of technology in combating it. We discussed the fact that plagiarism detection tools provide excellent service for detection of similar text between documents and they

can help users to easily identify suspicious works or paragraphs. But the final decision based on exploring the similarities and their evaluation are up to the users. They are responsible for the decision about plagiarism and they have to be prepared to submit substantiating arguments.

We provided an insight into the family of systems for revealing plagiarism which arises at Masaryk University. It consists of systems for plagiarism detection in theses (Theses.cz), in seminar papers (Odevzdej.cz), and in scientific papers (Repozitar.cz). All three systems are based on the same platform and share common database. Since the Internet can be considered as the best pool of resources for plagiarism, we have pinpointed principles and techniques which we use for detecting similarities against this vast amount of online data. Last but not least the enlightenment through seminars and workshops about plagiarism research and proper citation methodologies should be mentioned.

With such systematic approach to plagiarism, we hope that the institutional culture of the university can be transformed in a way, that the need and desire of students to plagiarise could be dramatically reduced. For the future, many tasks remain; nevertheless we can already see some great results that we have achieved so far. Together our team is preparing for the collaboration with other institutions in this area.

References

Brandejs, M. et al., 2009. Odhalovat plagiáty se daří díky spolupráci a odpovědnosti vysokých škol (Cooperation Among Conscientious Universities Reveals Plagiarism). In *UNINFOS 2009*. Nitra: Slovak University of Agriculture in Nitra

Jakubík, D. et al., 2011. A central repository of publication results, implemented as a part of systems for revealing plagiarism. *Seminář ke zpřístupňování šedé literatury 2011*, 4.

Kasprzak, J. et al., 2008. Distributed System for Discovering Similar Documents. In *ICEIS 2008*. Brno: Faculty of Informatics, Masaryk University, p. 14.

Suchomel, Š., 2012. *Systems for Online Plagiarism Detection*. Thesis. Brno: Masaryk University.

Suchomel, Š., Kasprzak, J. & Brandejs, M., 2012. Three way search engine queries with multi-feature document comparison or plagiarism detection. In P. Forner, J. Karlgren, & C. Womser-Hacker, (eds.) *CLEF (Online Working Notes/Labs/Workshop)*. Available at: http://dblp.uni-trier.de/db/confclef/clef2012w.html#SuchomelKB12.

Potthast, M. et al., 2012. Overview of the 4th international competition on plagiarism detection. In P. Forner, J. Karlgren, & C. Womser-Hacker, (eds.) *CLEF (Online Working Notes/Labs/Workshop)*. Available at: http://dblp.uni-trier.de/db/confclef/clef2012w.html#PotthastGHKMOTBGRS12.

Authors

Ľuboš Lunter, lunter@fi.muni.cz, Daniel Jakubík, jakubik@fi.muni.cz, Šimon Suchomel, suchomel@fi.muni.cz, Michal Brandejs, brandejs@fi.muni.cz, Faculty of Informatics, Masaryk University, Czech Republic