Source Retrieval for Plagiarism Detection

Šimon Suchomel and Michal Brandejs
Faculty of Informatics, Masaryk University, Brno, Czech Republic
Email: {suchomel, brandejs}@fi.muni.cz

***Abstract*—Plagiarism has become a serious problem mainly because of the electronically available documents. An online document retrieval is a weighty part of a modern anti-plagiarism tool. This paper describes an architecture and concepts of a real-world document retrieval system, which is a part of a general anti-plagiarism software. Up to date systems for plagiarism detection are discussed from the source retrieval perspective. The key approaches of source retrieval are compared. The system recommendations stem from design, implementation, and several years of operation experience of a nationwide plagiarism solution at Masaryk University in the Czech Republic. The design can be adapted to many situations. Proper usage of such systems contributes to the gradual improvement of the quality of student theses.**

***Index Terms*—plagiarism detection, plagiarism, source document retrieval, candidate document retrieval, system design, system architecture**

I. Introduction

Plagiarism is usually defined as passing off someone else's work as one's own. It is a moral offence which can appear in many forms. It is well known especially in the form of text plagiarism, for example, in journalism, which breaches moral ethics and in academia is referred to as an academic dishonesty. If such an offence is proven it, discredits the person who plagiarized and sometimes leads to resignations or expulsions.

Later appearance of plagiarism in academia also discredits the institution where it originated from because it passed unnoticed and should have been detected and dealt with accordingly at the time of submission. Higher educational institutions are not usually fond of making such cases public, on the contrary, they try to conceal it and resolve the issue internally as much as it is possible. Generally, acquiring an academic degree by deception has a bad influence on contemporary society. In some countries it is also legislatively impossible to revoke an academic degree if serious problems for the thesis defence are proven later on. Prevention and early detection are the best ways of solving plagiarism issues.

These are some of the reasons why the issue of plagiarism is not only complicated, but also very sensitive. There have been many publicly well known cases of plagiarism among high ranking politicians, journalists, artists or professors in the last decade which have proven that plagiarism cannot be taken lightly.

Manuscript received December 1, 2014; revised April 11, 2015.

Not all plagiarism is actual cheating. Much of it arises from a lack of text-using skills. Students sometimes do not know how to cite correctly, how to work with other text sources, or which actions lead to breaching the honor code. Therefore, they should be taught such skills as early as possible.

All plagiarism should ideally by detecting and handled accordingly. If it is found to be undertaken in purpose, i.e. a cheating, it should be dealt with without delay, according to the honour codes or law. However, detecting plagiarism may be quite a difficult, tedious and time consuming process and so, when marking papers, such as theses or seminar works, an automated computer system facilitating the task of checking for plagiarism, may prove to be very helpful. However, such systems never detect actual plagiarism, they cannot decide about what is right and what is not, the issue of plagiarism is very complex for a computer to decide. The automated systems can detect similarities among documents, mostly textual similarities, but it can be any kind of similarity which is somehow calculable. In the real-world today's computers cannot discover all forms of plagiarism, simply by its definition. For example, it would hardly detect that someone stole another person's idea. It is always up to a human specialist, a reviewer, a supervisor, or an authorized person to decide this.

The main goal of existence of plagiarism detection systems is to improve the quality of textual works. The mere existence of such systems puts pressure on students to have their texts correct knowing it has to pass some control. Also a formative feedback with assistance from an automated text reuse detection system has a positive impact on students' final submissions [1].

The prevention of plagiarism is the best success which the education about the problematic and also the tools for helping revealing plagiarism can achieve. If someone wants to cheat, they will do it anyway. Attackers are trying to deceive the system witch various techniques (some of them are discussed later), of which detection are after discovery implemented and the process continues with other techniques all over again. The plagiarism detection cannot be perfect, and the cheaters will always be one step ahead of the automated systems. However, a good trade-off of modern plagiarism detect system is that the cheater must put more effort into deceiving the system than to write an original text.

Standard systems for plagiarism detection, which operate on the basis of detailed document comparison, cannot detect similarities unless they possess both the source and the plagiarized document. An algorithm to

doi: 10.12720/jait.6.1.18-26

evaluate document similarity must build inner indices for a detailed document comparison. A modern plagiarism detection process can be divided into two main tasks. The source retrieval and the detailed document comparison, which can be reduced to pairwise document comparison. The pairwise document comparison is very computationally demanding, especially for real-time plagiarism solutions which must evaluate millions of documents [2]. Fig. 1 shows the approach of a modern plagiarism detection. For an input suspicious document, the outputs of the plagiarism revealing software are annotated passages of that document, which may have been plagiarized. This paper discusses the source retrieval as part of the whole anti-plagiarism detection process in detail.

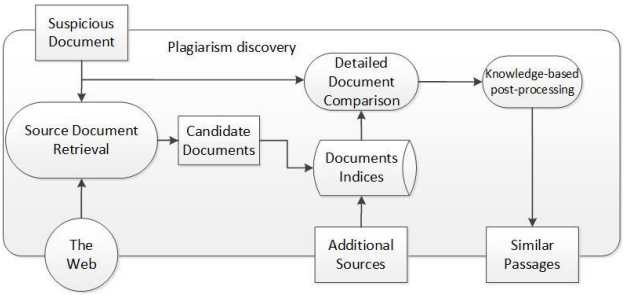

Figure 1. A global view of a modern anti-plagiarism software.

A source retrieval is a process for anti-plagiarism software to be performed for each suspicious document before it computes pairwise document similarities to find potential sources of plagiarism. The goal is to enlarge the document database of the anti-plagiarism system of relevant documents only. More relevant document means better opportunity to discover specific similarity. On the other hand, since the similarity computation is very demanding it is not wise to maintain a uselessly vast document database, for instance by crawling the Web, unless you possess really high computational capabilities like modern web search engines.

A. Source Retrieval

Having a suspicious document d_{susp} and a very large document collection D of potential source documents, the source retrieval task is to select a small subset $D_{ret} \subseteq D$ of documents $d_{ret} \in D_{ret}$ which probably served as a source of plagiarism. For example, documents which have a sufficient probability of $sim(d_{susp}, d_{ret}) > 0$, where SIM can represent any inter-documents measurable similarity. In a realistic scenario the D would contain all documents on the Web and the access method would be through the standard interface of a modern search engine, where we do not have direct access to its internal index.

The retrieved candidate documents are subsequently indexed for the purpose of the detailed document comparison.

II. State of the Art

There are also several types of plagiarism such as submitting another's work, which may also be bought; copying a text without citing the source, paraphrasing other texts, copying document structure; reusing own texts; or translations.

There are many tools that deal with uncovering plagiarism in one way or the other. Different approaches are also applied for detecting different types of plagiarism. Nonetheless, a system in order to be successful, must be aware of the original document, therefore the general usage of tools that compare documents only with a local corpus is limited, and this results in a sophisticated extension of the document base being an essential part of a successful tool for detecting plagiarism.

The most straightforward document source is the Web. In this section we further mention selected tools for plagiarism detection that are somehow extending their document bases from the Web.

A. CheckForPlagiarism.net

It is the name and the address of a commercial web-based service for scanning documents for reused text and showing potential plagiarism. The system assesses sentence structure in the suspicious document and makes a "digital snapshot" of each paper. The structure of paragraphs and sentences is cross-referenced to database of collected publications and it is "simultaneously sent to crawlers who scour the World Wide Web for possible matches" [1]. However, it is unclear what structure it is evaluating and how the online sources are looked up and retrieved. They do not publish the developed technology in more details.

B. DOC Cop [2]

It is a commercial web application for detecting plagiarism. It also offers web check for form-submitted suspicious text. It uses simple exhausted online search. The string length $n = [6, 12]$ is selected by the user prior to the text check. The submitted text is divided into strings of the selected length n – created sequentially from the input text. Each string is shifted by one word from the previous one, thus each neighbouring strings contain $n-2$ overlapping words. The first string is created from n words from the beginning of the submitted text. These strings are passed to Bing search engine and results are compared to the suspicious text. One submission to DOC Cop is limited to 1100 words of input text.

C. Masaryk University's Anti Plagiarism Solution

The Information System [3] of Masaryk University (IS MU) provides study administration and supports university e-learning. It also provides plagiarism detection among its documents. It is mainly designed for checking university theses prior to their defence. The document database and the plagiarism detection is interconnected also with papers of other schools [3], where the IS MU is being outsourced; next with the

[1] http://www.checkforplagiarism.net/service-features/sentence-structure
[2] https://www.doccop.com
[3] http://is.muni.cz

Czech National Archive of Graduate Theses (Theses.cz)[4]; the project for seminar works and papers (Odevzdej.cz)[5]; and the project for storing academic publications (Repozitar.cz) [4]. All documents respect a rich variety of access permissions.

The candidate documents are retrieved based on each document entering the database for plagiarism detection, from online sources according to principles discussed in this paper.

D. PlagScan[6]

It is a commercial software for plagiarism detection. It can be accessed via a web browser and is also offered to be installed on-site, on a dedicated server into one's own data processing center. PlagScan claims to be searching for thematically related documents. It offers to include its document base and also other local databases into similarity calculations with respect to document permissions. As a detailed document comparison it uses similarities based on chunks from consecutive three words of the texts. It utilizes Yahoo search as a means of obtaining relevant online sources. How the queries are constructed is not explained. The results are however displayed in a sentence-based manner.

E. turnitin; iThenticate; Ephorus[7]

Turnitin and iThenticate are commercial tools developed for originality check and plagiarism detection. Turnitin is designed for teachers and educational institutions, it helps them to organize and control the process and the quality of student papers. iThenticate is designed for individuals such as academic workers and writers, it allows them to check submitted document for unintentional plagiarism and to verify its originality.

The methodology of search behind the systems is common. Concerning the growth of document database for text comparison from online sources, they adopted the Internet crawling[8] methodology like contemporary search engines, meaning that the company has sufficient resources to maintain crawlers for downloading and indexing the Web for detailed comparison. They index as much content as possible no selection of relevant sources is made in the state of resource retrieval. The source retrieval task is shifted to the whole index of retrieved documents where standard methods of information retrieval can be utilized.

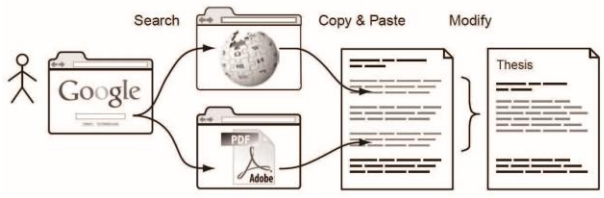

Figure 2. The generic steps of text reuse from the Web [5].

Ephorus is another major commercial software for preventing plagiarism. It administers also a large document database which stems from involved institutions. Online sources are also retrieved broadly via automated crawling. In autumn 2014, Ephorus was acquired by turnitin. Ephorus provides integrations with many various e-learning systems and others can be integrated via API.

F. Viper[9]

It is an anti-plagiarism scanner and free windows-based desktop application. It is intended for individual use and it offers a real-time Internet scan for plagiarism detection. To fulfil this task, it divides the input text into not overlapping chunks of about twenty consecutive words. It does not bother to remove any special characters or punctuation. Such text chunks are encoded into URL encoding and passed to the Yahoo search engine. Similarities between search results and the input text are, after evaluation, displayed in the application. Due to its inner encoding methods it does not support texts written in languages other than English.

III. Conceptual System Characteristics

The outer behaviour of the source retrieval system should be as much like the behaviour of a student who searches for documents on the Web and reuse a text from them. Martin Potthast depicts a standard process of text reuse from the Web as shown at Fig. 2.

Considering a standard plagiarism detection tool, we suppose that suspicious documents are single-themed. That is the most common situation. Such documents are, for example, theses or seminar papers. The majority of documents which are expected to be checked for plagiarism are single-themed. This assumption leads to the possibility of extracting keywords from the whole document without significantly lowering the performance of automated keywords extraction. Keywords extraction is one of the fundamental features of the source retrieval system (see the following section for more details). Under this assumption, an example of unsuitable use of an anti-plagiarism tool would be the checking of one diverse-themed document, like newspapers uploaded in a single file. Such a document should then be divided into separate documents according to the articles which the newspapers contain, which will result again in single-themed documents.

From a single document point of view, the system pre-processes the input document and creates appropriate queries which are submitted to a search engine interface. The search results must also be processed accordingly. The system should follow several considerations: i) maximizing precision and recall; ii) minimizing the overall cost; iii) be scalable and robust.

A. Retrieval Performance

The demand to maximize the recall and precision of retrieved documents is obvious. However, it is usually

[4] http://theses.cz
[5] http://odevzdej.cz
[6] http://www.plagscan.com
[7] http://turnitin.com
[8] Turnitin's web crawler is called TurnitinBot previously known as SlySearch https://turnitin.com/robot/crawlerinfo.html.
[9] http://gateway.scanmyessay.com

balanced with acceptable computational load of the system. It is also very difficult to measure precision and recall of a real-world web document retrieval system. Let D_{src} denote the set of documents that served as a source of plagiarism for document d_{susp}, and let D_{ret} denote the set of retrieved documents. Then the precision and recall can be defined as $prec = |D_{ret} \cap D_{src}| / |D_{ret}|$ and $rec = |D_{ret} \cap D_{src}| / |D_{src}|$ respectively. However, this standard information retrieval calculation is far from being applicable, namely because of the existence of so called near-duplicate documents on the Web [6]. The source document retrieval system can select a near duplicate document d_{ret} which certainly is true positive detection and it does not have to be the same source document d_{src} from which it was plagiarized. In order to measure a near duplicate, some characteristics must be defined. For an anti-plagiarism system, the positive value of similarity $sim(d_{susp}, d_{ret}) > 0$ can be sufficient to consider the retrieved document d_{ret} as a true positive. The similarity can be any kind of likeness between two documents which is computed by the detailed document comparison subsystem of the anti-plagiarism software.

Organizers of PAN [10], competition on plagiarism detection, determine whether a document d_{ret} is a near-duplicate to any document from the set of source document D_{src} by three characteristics: i) whether they are actually equal $d_{ret} = d_{src}$; ii) whether they are similar according to an empirically set threshold of Jaccard similarity $sim_{jac}(d_{ret}, d_{src}) > n$; iii) or whether the passages in a suspicious document d_{plag} that are known to be reused from d_{src} are also contained in d_{ret} [7]. We can now observe that one document can be a near-duplicate of more than one source document and one source document can have more than one near-duplicate. Next, they denote a set D_{dup} of all near-duplicates of all source documents D_{src} of d_{plag} and a subset $D'_{ret} \subseteq D_{src}$ containing documents having at least one positive detection in D_{ret}. Then precision and recall of set D_{ret} based on a suspicious document d_{plag} are defined as follows:

$$precision = \frac{|D_{ret} \cap D_{dup}|}{|D_{ret}|}, recall = \frac{|D'_{ret} \cap D_{src}|}{|D'_{ret}|} \quad (1)$$

This results in the fact that, retrieving more than one near-duplicate document to a single source document does not increase recall and it does not decrease precision either. Retrieving the first of the near-duplicate documents into a single source document increases both recall and precision.

[10] http://pan.webis.de/

It is worth mentioning that in order to evaluate all near-duplicates we need to build an index of the whole corpus of all potential source documents, which could be searched via a given search engine. Therefore, such evaluation is infeasible in a real-world situation when the corpus of source documents is the Web.

In a real-world scenario, the recall is much more important than precision. If the precision is low it could affect time performance of the retrieved algorithm, since the system would process a lot of documents needlessly. It can also excessively extend the index for detailed document comparison, which is not a problem as long as the detailed document comparison is feasible according to user expectations. On the contrary, if the recall is low, the anti-plagiarism system may simply not be able to detect the plagiarized passage, since it may not have the source document retrieved and indexed in its database.

In addition to documents that contain similar passages with the suspicious document, we consider as a true positive retrieved result a document following the same theme as the source document. Themed documents are considered relevant. A theme can be detected by overlapping sets of equal keywords or keyphrases [8]. Existing themes are therefore defined by the characteristics of the suspicious documents within the database of the anti-plagiarism system.

B. Retrieval Cost

In a standard way the cost of the system consists of time and space requirements of all algorithms and data needed. Apart from that, the most costly component is the number of executed search queries, and secondly the number of Internet document downloads.

In any information retrieval system, there is always a correlation between the retrieval performance and the cost. Consider a system using an exhaustive search approach. For example, querying every sentence from a suspicious document would result in high recall, but it is simultaneously too cost demanding to be applicable elsewhere than in an experimental environment. On the contrary, in real-world systems, the number of search queries should be narrowed as much as possible, which can result in certain situations in executing only a single query per suspicious document.

It is crucial to reduce the number of queries since the real anti-plagiarism system must utilize modern search engines like Bing, Google or Yahoo. Furthermore, each search engine has strict rules about the amount of queries which can be submitted from one IP address, which prevents using the exhaustive search. The query execution is usually not particularly time consuming, yet the time consumption is not negligible. The automated querying can often be attended by additional fees. In the document retrieval system design, the querying represents the most expensive part.

The second significant part of the system cost is the number of document downloads. The download alone is in today's system, a cheap operation, but it can be very time and space consuming while considering a huge number of downloads. Also a post-processing of the downloaded documents is a very time consuming

operation. The number of downloads must be tuned according to the system computational possibilities and expectations.

C. System Scalability and Robustness

The purpose of the system determines its scalability. The modern anti-plagiarism systems maintain database of millions of documents and are able to process new documents within hours or even minutes. The complete processing of a new document means that all results of candidate document retrieval, together with the suspicious document, must already be indexed for detailed document comparison. Afterwards, the evaluation of similarities of that document is usually real-time (within seconds). The design of the source retrieval system, which is further recommended, can scale easily by adding more computational nodes.

A need for robustness stem mainly from a huge diversity of Internet documents. It is discussed together with the detailed design of the system in the following sections.

IV. System Architecture

The source retrieval should run as several independent tasks, in order to be highly parallelizable and scalable. The tasks can share data via a transactional relation database. The database represents a central point for process control. If it is accessible over a network, the computational power can be increased by adding more computer nodes. The database should be utilized in order to keep detailed information about the progress of document processing. The tasks could be divided according to the following functions into 4 main groups: 1) parsing an input document; 2) searching the Internet; 3) downloading the results; 4) the results post-processing.

A. Parsing an Input Document

Let us assume that an input of this stage is a textual representation of a suspicious document d_{plag}. Since the anti-plagiarism system needs to build data structures for the detailed document comparison, the plaintext format is needed anyway. Therefore, the input document conversion into plaintext must generally also be considered. The output of this stage would be queries prepared for their execution.

Textual processing and keywords extraction algorithms may become quite time consuming. A standard algorithm optimization should be considered when needed. This stage, however, does not represent the most time consuming part of the overall source retrieval process. Each suspicious document is processed independently, thus the system may scale by simultaneously processing suspicious documents.

The matters to consider at this stage include: i) document cleaning and preprocessing; ii) language detection; iii) chunking; iv) keywords extraction; v) query formulation; vi) permanent storage of extracted queries and the input document information into the database.

Cleaning of the document may comprise the removal of special characters, original document structure violation, existing citation detection, or in-text urls extraction for a direct download.

1) Language detection

A modern anti-plagiarism system should also be able to detect similarities among and across multiple languages. This must be borne in mind during the system implementation. Many of the shelf tools for lingual processing or keyphrase extraction would not be possible to utilize.

Current effective automated language identification methods are based on frequency analysis, such as utilizing the principle of language-characteristic sequences of n-grams. For the usage of such methods one needs to construct a referential vector language model for each supported language. Other beneficial, less computational demanding method, can be language detection by stop-words matching. Only lists of language specific stop words are kept and the language with the highest number of matches is selected. This method works reliably for longer textual parts. Next, a method based on word relevancy can be utilized for shorter texts. It is also applicable for the web page language identification [9]. With supporting of multiple languages the automated language identification must also be applied on every retrieved web document.

Please consider that in many theses, there are usually small parts of text written in multiple languages, such as the abstract or summary. The detection method should detect the major language of the text or identify those language-different parts.

2) Chunking and keywords extraction

The purpose of chunking is to distribute focus of text processing algorithms evenly across the document and thus lower the possibilities of influencing the efficiency of that algorithms by unexpected characteristics of the text. Chunking is also applied in order to detect textual differences, where one cannot pre-set the exact boundary in a document, where the textual characteristics are changed. For this purpose, the principle of a sliding window is usually used, where two primary parameters must be determined. The first stands for the size of the window and the second represents the size of the overlap between two neighbour windows during sliding. The size of the overlapping part also influences the detected characteristic differences between two chunks. If the overlapping size is too big the difference would probably pass unnoticed. On the other hand, using small size of that interval sharpens algorithm detection edges, but poses a risk of placing the window's centre on the textual characteristic boundary, resulting accidentally in no detection. It may also be considerable to process more than one pass of the algorithm with different sliding windows settings—a type of cross validation.

Other considerable chunking approaches are no chunking, paragraph based chunking [10], chapter based chunking, or sentence chunking. Some approaches use also chunks of pre-set size [11], which is applicable in all situations, but is little correlated with the structure of the document.

The Keywords or keyphrase extraction is the most straightforward process for subsequent query formulation. The high quality keywords extraction is crucial for the proper query formulation. Modern keywords extraction methods are based on the word repetition allied to a statistical estimate of likelihood. Also the most widely used method in all PAN competitions on plagiarism detection (since 2012) in the source retrieval subtask was keywords scoring by $tf \cdot idf$ (term frequency—inverse document frequency) [12], [7], [13].

Keyphrases can also be extracted from the selected chunks. However, in the real-world scenario it appears to be more beneficial to extract global keywords from the whole document. Such keywords are fully related to the document theme and should suitably describe the individual document. From a longer textual part, there can also be obtained more descriptive keywords than from the shorter part.

3) Query formulation

The query formulation is the most important part of the source retrieval system, since it has the highest impact on the overall performance and costs. In order to control the cost, a maximum number of queries submitted per document should be set. The total number of executed queries influences directly, not only the cost, but also the time demands of the input document process and the number of Internet documents to be processed.

Suchomel *et al.* [14] propose a methodology based on the combination of three different types of queries. The first type of queries is constructed from keywords or keyphrases extracted from the whole document. They suggest using use 5 word long keywords based queries. The query length is important since it directly influences the number and the relevance of retrieved results. If the query is too long, it could be too specific, which will probably lead to no retrieved results. On the other hand, if the query is too short, it will be too general resulting in retrieval of many irrelevant documents. The purpose of the keywords based queries is to retrieve theme bounded documents.

Other types of the proposed queries are extracted from different chunks of the suspicious document. It deepens the search for those specific parts and aims for retrieval of more text-related documents. They also suggest to detect suspicious passages of the document by evaluating textual characteristics with intrinsic plagiarism detection methods and deepen the search in those passages. Queries constructed from small text parts of the source document can be characterized as phrasal queries and are usually longer (up to 10 words).

The proposed methodology is applicable in a real-world document retrieval system and it also performed best in terms of the total system workload, while maintaining a good retrieval performance in PAN 2012 competition on plagiarism detection [12]. In following runs, this methodology was improved with enhanced download control and the third type of queries was changed from header based to paragraph based queries. However, the main idea remains the same [15], [10].

This methodology is also expected to perform better in real-world scenarios while utilizing modern search engines, than in PAN competition environment. It is because the search engine used during the competition did not support phrasal search, which influenced a significant part of queries of the proposed methodology.

It also scales up to a single query per document. The first query is constructed from the keywords which obtained the highest score. In the next step, the keywords based queries are formulated from the consecutive extracted keywords sorted by their score up to the score threshold or a up to the pre-set maximum number of queries of this type. After that, the search can be deepened by phrasal types of prepared queries.

A multilingual search can be accomplished by a query translation, which is generally easier than the full sentence translation. It is sufficient to translate all the query terms consecutively, especially if the query is constructed from keywords only. Translation can be done by the dictionary associations. Still, a quality disambiguation may pose a problem for successful translations. It is therefore, better to use words in their canonical forms in keywords based queries, since the search engine will not distinguish between different forms of one word. A different situation is at phrasal queries, they should remain unchanged, since the modern search engines will attempt to appraise the meaning of the search, like for example, the new Google's searching algorithm called Hummingbird.

B. Searching the Internet

During this task, the prepared queries are submitted to the search engine. A search control should be implemented in order to minimize the total number of submitted queries. As a consequence of the limited query budget, the queries should be processed stepwise and search results should be evaluated, in terms of a basic feedback for the search controller, after each query. During the searching the search controller can reschedule the submission of prepared queries, which may include query skipping or reformulation. The basic search control represents submission of queries according to their priority up to a specific number of submissions. Haggag and El-Beltagy [16] check subsequent queries against all previously downloaded documents, which were downloaded based on an analysis of one suspicious document, through a simple token matching. They skip queries which show 60% or higher tokens match. Suchomel *et al.* [15] submit document global keywords based queries at first. Next, according to obtained similarities between the suspicious and retrieved document, they skip queries covering by their position the portions of the suspicious document, which were already mapped to the source.

Issues to consider during this task include: i) the search control; ii) the feedback from retrieved documents; iii) the storage of query records and retrieved results with results filtering. The storage of the query record prevents

executing of the same queries if prepared for different input documents. Also the date and time of each query execution can be decisive for the eventual query resubmission.

C. Downloading the Results

Real Internet searches include many types of documents, such as textual rich formats or multimedia formats. A plaintext needs to be extracted from retrieved documents, therefore only plaintext convertible documents should be downloaded. Downloads can pose huge bandwidth and disk storage demands. In the real-world, there is little information known prior to the download, which influences download decision making. The type and size of the document can usually be determined from headers of Internet documents prior to the full document download, which still pose a header request to a web server. This leads to having to leave the decision making about the quality of the search results to the post-processing phase.

The Web is a very wild and volatile environment, thus more emphasis must be put on the robustness of the downloading subsystem. Addition to standard timeouts, other techniques should also be considered. For example, more attempts to download a document should be carried out if the previous one was unsuccessful. A maximum file size limit should be set for *html* files, otherwise the downloader can get stuck in endless web files. On the other hand, certain types of documents (like *pdf* files) have to be downloaded completely in order to extract the text from them properly. The request for header can help to set the maximum file size of such types of documents. Unfortunately, Internet files headers do not always provide veracious information.

Various metadata of downloaded documents need to be stored permanently. The date and time of the last download and the original internet document are needed also to be able to ascertain plagiarism.

Assuming database driven data exchange, the time demands of the hypothetical simple downloader consist of database operations; establishing connections to the target web server; downloading the actual data; and saving the data. Our tests show that beneficial speedup of download is favourable by the process parallelization only. The connection establishment can be sped up, for example, by the DNS record caching. Not any crawler-like optimization can be performed, since the search results generally contain various web sites. All downloads can be sorted according to the hosting site in order to use cached DNS records.

TABLE I. PERFORMANCE OF VARIOUS WEB DOCUMENT DOWNLOAD TECHNIQUES.

Δt [min.]	domain ordering	no domain ordering
one process	1:49	1:44
db dedicated thread	1:56	1:48
4 download dedicated threads	0:31	0:34

Table I presents averaged times of 2 passes of downloads of 137 different Internet documents obtained from searches based on different queries. 91 of those 137 documents were downloadable at the time of the tests. Others ended with various HTTP errors among which the HTTP 404 (Not Found) was the most abundant. The tests ran in homogenous network and hardware conditions. The *domain ordering* column shows times, when the downloader tried to optimize Internet requests by accessing the same sites consecutively. The times shown in the second column were obtained while accessing Internet documents in the order as they were added to the database for download by the search algorithm. The second data line of the table shows times of the changed downloader differing from the first line of the table in a threading approach. A dedicated thread was used for downloading the documents, and database operations together with the other logic remained in the main thread of the programme. The third line represents multi-threaded download process—4 threads were used for the documents download and the main thread remained unchanged. The results show that the database operations are negligible when compared to time demands for downloads. Also, the site ordering will probably not pay off, since it burdens the algorithm with the additional sorting. Here the downloads are certainly the most time consuming operations, but they are also easily parallelizable. The third row of the table shows that n additional threads will almost linearly n times decrease the total time for download.

D. The Results Post-Processing

In the post-processing phase, the task is to evaluate the quality of the downloaded document and if the quality is sufficient, to pass the document to the indexer. Only among the indexed documents the similarities can be calculated, which represents the subsequent stage of the plagiarism detection process.

A plaintext needs to be extracted from every downloaded document in order to evaluate similarities. Information must be obtained before the actual text conversion, which includes: file type identification—the file type given by a *HTTP* response header cannot be trusted, therefore the file type must be identified by MIME detection tools.

Sometimes the file type must be identified according to the file extensions, if any, or according to other heuristics. The encoding of the file needs to be determined in pursuance of the correct tokenization and indexing. The language of the document should also be known supposing appropriate lingual classification.

A modern source retrieval system should be able to convert the most common web file formats which include: *html* and other markup languages document formats; Microsoft Office family file formats; Open Office file formats; and probably the most common *pdf* files.

From the nature of MS Office and Open Office formats it follows, that the plaintext conversion is possible, because those files carry a text source information. It is hidden in the proprietary structure of the files. Standard

tools for those format creation can extract the text; however, the extraction must be fully automated. Not all documents are generally convertible, since those formats allow to lock or create password protected text. There are also publicly available programme modules and extraction tools for Open Office [11] and MS Office [12] documents.

The plaintext conversion of *pdf* file is more complicated since the *pdf* is not an easily editable file format. There are many tools for creating various versions of *pdf* files, thus the issue of the *pdf* text conversion is far from being a smooth and errorless process. Firstly, the standard methods of text extraction from the *pdf* text layer together with the text correctness should be performed. If the text is not well-formed or if the extraction fails, other conversion methods should be applied. Further possibility of text extraction is to pass the document to an OCR recognition[13]. The check for text correctness is important even if the extraction from *pdf* layer was errorless. Typically non ASCI characters can be damaged and a profile of the text must align with any of the supported language. If a plagiarist obfuscates plagiarism by braking the textual layer and keeping the document to display correctly, the use of an OCR is also inevitable. For example a student creates a plagiarized text and replaces every space in the text with any letter in white colour, which will not be seen by a human reader. The text will have the character of a single huge meaningless word for the text extractor. Another cheater's known approach to confuse computers is replacing some types of characters with characters from a different alphabet using certain fonts, which look very similar and will pass unnoticed by the reader. For example the Roman character **o** can be replaced by the Greek omicron[14] if one uses a font in which they look the same. The use of OCR will recreate the text correctly, since it looks at the document in the same way as the user does.

The issue of text extraction from *html* family files is even more complicated. The majority of web pages include together with the main content also so-called boilerplate content. The boilerplate content is, for example, a navigation link, advertisement, header or footer. It is a meaningless content for document comparison. Having indexed all unchanged text from web pages, it would result in many false positives in evaluation of the document similarities. Internet documents would be spoiled with repeated parts of text, which do not carry needed information. An example of this would be the pages from Wikipedia, they all contain the same footer. Therefore, the text extraction from *html* files should be accompanied by a boilerplate removal, for example by means of context based approaches to identify and remove the boilerplate [17].

For example, Masaryk University runs proprietary servers for plaintext conversions inside the network document storage of the university's Information System. It includes dedicated client-server network hosted applications for MS Office, Open Office and *pdf*, including OCR, documents to plaintext conversion.

The plaintext conversion is generally very computationally demanding, it also takes a lot of tools and technologies to convert many document types. From all, the *pdf* files are possibly the most computationally demanding to convert and except for *html* files, the *pdf* files represent the most preferable file format to be published on the Internet.

Having extracted a plaintext from the downloaded documents allows for subsequent document evaluation. A decision whether to actually index the document for the plagiarism detection must be made. Straightforward evaluation is to compare the retrieved document d_{ret} with the suspicious document d_{plag} for a document similarity. However, considering the real-world plagiarism detection for many input documents, the source retrieval system can retrieve a theme bounded document based on a query created from a certain suspicious document, but the retrieved document could serve as the source for plagiarizing another document, which is also in the anti-plagiarism system database. Then the retrieved document is valuable, but evaluating it with only the document from which the query was constructed can result in no similarities. In such situations all retrieved good quality texts should also be indexed for all document similarities.

The text extraction can also be parallelized on the document level, like downloading, but especially the optical recognition can still be very time consuming. It may currently take tens of minutes to complete, based on used hardware.

V. Conclusion

This paper discussed the main points of the source retrieval system architecture. The candidate document retrieval is an unexpendable part of a modern anti-plagiarism detection system. The quality document base for detecting document similarities is for the anti-plagiarism system of critical importance. Such a system should retrieve potential sources of plagiarism from the Web based on each document entering the plagiarism detection and the main purpose is to retrieve a relatively small subset of similar documents, which may have been plagiarized from. Firstly, such methodology leads to retrieving textually similar documents, and secondly, which is particularly beneficial when done based on academic papers, it retrieves thematically related documents. Consequently, an anti-plagiarism system evaluates document similarities among all documents which the system operates with, together with the newly retrieved documents.

A user is usually provided with a percentual portion of document similarities between pairs of similar documents. The overall percentage can also be provided. However, the system does not decide about plagiarism, it only selects similar passages. The issue of plagiarism is far more complicated. It is always up to a user judge to

[11] http://freecode.com/projects/oo2txt
[12] http://www.adelton.com/perl/Docserver/
[13] http://en.wikipedia.org/wiki/Optical_character_recognition
[14] Omicron does not even have an LaTeX command.

decide, whether the given text is plagiarized or not. The system simplifies the tedious work of finding the sources of similar texts. For example, a page in a thesis can be copied from another text source, which would not be considered as a cheat if cited correctly.

There are also many types of plagiarism, for example paraphrasing, copying the structure of the document, copying the results or copying the texts. The overall quality of a plagiarism system can be evaluated using measurement based on what reused text obfuscation it can detect.

This paper summarized experience from the real-world operation of the anti-plagiarism system used at Masaryk University as part of a country wide plagiarism solution in the Czech Republic. It provided ideas, concrete methods, and concepts for a candidate document retrieval system construction. It also discussed concepts used on PAN competition on plagiarism detection. The research behind the competition provides additional topic-related information.

ACKNOWLEDGMENT

The authors would like to thank to the Information System of Masaryk University for creating an opportunity to improve the plagiarism issue in Europe.

REFERENCES

[1] M. Davis and J. Carrol, "Formative feedback within plagiarism education: Is there a role for text-matching software?" *International Journal for Educational Integrity*, vol. 5, pp. 580-70, December 2009.

[2] J. Kasprzak, M. Brandejs, M. Křipač, and P. Šmerk, "Distributed system for discovering similar documents," *Brno: Faculty of Informatics, Masaryk University*, p. 4, 2008.

[3] L. Lunter, D. JakubÍK, Š. Suchomel, and M. Brandejs, "Inter-university cooperation on plagiarism detection systems in Czech republic," in *Jiří Rybička, Plagiarism across Europe and Beyond*. 1st. Brno: Mendel University in Brno, 2013, pp. 216-224.

[4] D. Jakub ÍK, L. Lunter, M. Brandejs, and J. BrandejsovÁ, "A central repository of publication results, implemented as a part of systems for revealing plagiarism," *Seminar on Providing Access to Grey Literature 2011*: *4th Year*, The National Library of Technology in Prague, 2011.

[5] M. Potthast, "Technologies for reusing text from the web," *Dissertation*, Bauhaus-Universit ät Weimar, December 2011.

[6] D. Fetterly, M. Manasse, and M. Najork, "On the evolution of clusters of near-duplicate web pages," in *Proc. the 1st Latin American Web Congress*, LA-WEB 2003. IEEE, 2003.

[7] M. Potthast, T. Gollub, M. Hagen, M. Tippmann, J. Kiesel, E. Stamatatos, P. Rosso, and B. Stein, "Overview of the 5th International competition on plagiarism detection," in *Proc. CLEF 2013 Evaluation Labs and Workshop*, 2013.

[8] Š. Suchomel. "Systems for online plagiarism detection," *Thesis, Faculty of Informatics, Masaryk University*, Brno 2012.

[9] R. Řehůřek and M. Kolkus, "Language identification on the web: Extending the dictionary method," in *Proc. 10th International Conference on Computational Linguistics and Intelligent Text Processing*, Mexico City, Mexico: Springer-Verlag, 2009.

[10] Š. Suchomel and M. Brandejs, "Heterogeneous queries for synoptic and phrasal search," *Notebook for PAN at CLEF 2014*, L. Cappellato, N. Ferro, M. Halvey, W. Kraaij, eds. *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, 15-18 September, Sheffield, UK. CEUR Workshop Proceedings, CEUR-WS.org (2014)

[11] V. Elizalde, "Using noun phrases and tf-idf for plagiarized document retrieval," *Notebook for PAN at CLEF 2014*, Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. eds. *CLEF 2014 Evaluation Labs and Workshop* – Working Notes Papers, 15-18 September, Sheffield, UK. CEUR Workshop Proceedings, CEUR-WS.org (2014)

[12] M. Potthast, T. Gollub, M. Hagen, J. Kiessel, M. Michel, A. Oberl ädner, M. Tippmann, A. Barr ón-Cedeño, P. Gupta, P. Rosso, and B. Stein, "Overview of the 4th International competition on plagiarism detection," *CLEF 2012, Evaluation Labs and Workshop*, 2012

[13] M. Potthast, M. Hagen, A. Beyer, M. Busse, M. Tippmann, P. Rosso, and B. Stein, "Overview of the 6th International competition on plagiarism detection," in *Proc. 2014 Conference Working Notes for {CLEF}*, Sheffield, UK, September 15-18, 2014.

[14] Š. Suchomel, J. Kasprzak, and M. Brandejs, "Three way search engine queries with multi-feature document comparison for plagiarism detection," *CLEF (Online Working Notes/Labs/Workshop)*, Rome, 2012.

[15] Š. Suchomel, J. Kasprzak, and M. Brandejs, "Diverse queries and feature type selection for plagiarism discovery," *CLEF 2013 Evaluation Labs and Workshop*, Valencia, 2013.

[16] O. Haggag and S. El-Beltagy, "Plagiarism candidate retrieval using selective query formulation and discriminative query scoring," *Notebook for PAN at CLEF 2013*, Valencia, 2013.

[17] J. PomikÁLEK. (2013). Justext: Heuristic based boilerplate removal tool. [Online]. Available: Google code, online http://code.google.com/p/justext/

Šimon Suchomel studied computer and information systems at Faculty of Informatics, Masaryk University in the Czech Republic. Currently, he studies PhD programme in the field of applied informatics in plagiarism detection. The author is a member of the Masaryk University's Information System development team, he designs and builds the online-source retrieval system for the Information System's plagiarism detection tool. He also works at the faculty as an administrator of windows-based networks. In his free time, he plays tennis and usually spends time with his family or sporting. He is also a devotee of dogs and south Moravian white wine.