Seminar on Providing Access to Grey Literature 2013: The 6th year of the seminar focused on storage and providing access to the grey literature, 23th October 2013 [online]. Praha: National Library of Technogy, 2013. Available at WWW: http://nusl.techlib.cz/Sborniky. ISSN 1803-6015.

A COMPARISON OF ANTI-PLAGIARISM SYSTEMS FOR THESES AND DISSERTATIONS

JAN MACH

machj@vse.cz

Centre of Information and Library Services, University of Economics, Prague, Czech Republic

Institute of Information Studies and Librarianship, Faculty of Arts, Charles University in Prague, Czech Republic

Abstract
This paper focuses on a test and a comparative analysis of systems for detecting duplicates (so-called anti-plagiarism systems) used for the repositories of higher education theses and dissertations in the CR. A text corpus containing the most frequent sources of plagiarism was created for the needs of the test, and the modifications made by plagiarists were simulated. The success of duplicity detection by the most important anti-plagiarism systems was verified experimentally, and a comparative analysis and verification of stipulated hypotheses were performed. The evaluation was also performed on the author's own prototype application using the Google search engine.

Keyword
plagiarism, theses and dissertations, repository, Turnitin, Theses.cz, Ephorus, Google

INTRODUCTION

The operators of repositories of theses and dissertations, repositories of the scientific/research publication activities of employees, magazine publishers etc. may require a check on the originality of stored texts. The objective of this work is to compare important anti-plagiarism applications used primarily during checks of students' theses and dissertations, and to offer help to the administrators of grey literature repositories in the selection of the optimum tool.

This work describes the results of an analysis of 300 samples of text from 50 documents, randomly selected from ten differing types of source used by students when writing theses and dissertations. The test compared the Turnitin, Ephorus and Theses.cz systems, and the author's own GooglePlagiarism application, which searches online for duplicates using the Google search engine.

The analysis was performed in 2013 as part of doctorate studies at the Institute of Information Studies and Librarianship at the Faculty of Arts of Charles University, in which the author focused on designing and operating repositories of theses and dissertations in the CR. The

complete procedure for the preparation of the text corpus, the detailed results of the test, and the evaluation of the hypotheses form part of the author's dissertation under preparation.

A DEFINITION OF PLAGIARISM

We can find a basic definition of plagiarism in ČSN ISO 5127:2003 Information and Documentation – Dictionary (1), which stipulates the terms for facilitating international communication in the area of information and documentation, and defines selected terms. In this standard plagiarism is labelled as "presenting the intellectual work of another author borrowed or imitated in whole or in part, as a person's own" (1), meaning using an idea without indicating its author. The Copyright Act (2) does not actually define plagiarism itself, but defines a Work (Section 2) and the related personal rights (Section 11), economic rights (Section 12 - Section 27), and permitted exceptions, including citations (Section 30).

On the basis of a worldwide questionnaire-based survey of 900 secondary school and higher education students, ten different manifestations of plagiarism were stipulated (3). The most common sins, according to this survey, included cloning (presenting somebody else's work, word for word, as your own[1]), CTRL-C (presenting somebody else's work as your own with a minimum amount of modifications) and mixing sources (combining more than one source into a single text without attribution), when there is no significant modification of the original text. The individual software tools mentioned in this work help to uncover precisely these types of plagiarism. An application that identifies only duplicate passages between an analysed text and other indexed documents cannot assess whether this is a correct citation of the source or, for example, a regular phrase for which a citation of the source is not needed. The reviewer must decide whether plagiarism really occurred through an evaluation of the duplicates found.

CREATION OF THE TEXT CORPUS

For the purposes of this study focusing on plagiarism in theses and dissertations, it was not possible to use any of the normally available corpuses, as they are primarily intended for English texts or cover only specialised sources. A special text corpus was therefore prepared for the test based on Czech and English articles containing documents from the types of sources from which students most frequently copy indicated below:

- - the Czech and English versions of Wikipedia,
- - theses and dissertations in the CR written in Czech and English,
- - theses and dissertations from abroad,

[1] The word-for-word translation of a source without the proper citation can also be understood as this type of plagiarism. Translating from a foreign language without citation is more common with Czech texts than with English ones, and one objective of the test will be to verify the ability of the anti-plagiarism tools to also detect translations

- - Czech media monitoring,
- - foreign expert documents under the Open Access system,
- - Czech and English websites,
- - paid electronic information sources.

The corresponding web applications (full-text search engines) were stipulated for the indicated types of sources, specifically www.wikipedia.cz, www.wikipedia.org, www.Theses.cz, SCIRUS ETD Search (4), Anopress, arXiv.org, Google[2] and the paid electronic information sources EBSCO, ProQuest Central, eBrary, JSTOR and OECD iLibrary. Five documents were selected in each of the total of ten search engines according to a previously specified methodology for the stipulation of key words[3].

From these documents one sentence was randomly selected from the start of an article, containing if possible continuous text with the key word in question without parenthesis and without upper indexes for footnotes, etc. In addition, a paragraph of text containing around 5 sentences was selected at random. Any eventual links to notes, sources etc. were removed from this paragraph. In addition to the sentences and the paragraphs, the following transformations of the selected sentence were incorporated to the test corpus:

- - a sentence with two words transposed,
- - a sentence with diacritics removed,
- - a sentence with a single word replaced with another with a similar meaning (paraphrasing a word),
- - a sentence with several words replaced with others with similar meanings (paraphrasing a sentence),
- - a sentence machine-translated into Czech/English.

The resulting text corpus also contained a total of 300 records with the indicated transformations – samples of text obtained from 50 documents from 10 types of source. To support the preparation of the corpus a special application was programmed to record source

[2] According to a survey by Effectix [5], Czech users more often select Google, and therefore this search engine was given priority in the test in the Czech and English versions over the Seznam search engine.

[3] The key words used for the selection of documents were from the field of mobile communications, namely the abbreviations LTE, UMTS, EDGE, WiMax and Wi-Fi. The selected words only minimally influence the result of the test, as the test corpus contains whole sentences from found documents and the sources used are generally focused.

documents (texts and metadata in XML format), the automated transformation of texts[4], the generation of test files in HTML and statistics in Excel. The application prepared HTML files with texts that were either directly – or after conversion into Word format – uploaded into the individual tested anti-plagiarism systems.

EVALUATING THE ANTI-PLAGIARISM SYSTEMS

Within the framework of the test the functionality was compared of the most important systems used to check for plagiarism in theses and dissertations in the Czech Republic – the applications Turnitin, Ephorus and the Masaryk University systems (Theses.cz, Odevzdej.cz, Repozitar.cz). As part of the comparative analysis these systems were compared with the GooglePlagiarism application developed by the author of this study.

Turnitin (www.turnitin.com)

The Turnitin application enables the administration of documents via a web interface (in 15 language versions, not including Czech) or connection to an external system (e.g., Moodle). According to information from the company, a database of over 24 billion web pages, 300 million archived theses and dissertations and 120 million articles from more than 110 000 magazines and books is used for comparison purposes.

A licence is provided based on the number of full-time students registered at the institution in question. The price of an annual licence for a university without discounts is £1,430 plus a surcharge of £1.16 per student attending the institution (if integrated with Moodle or another of the offered online search tools the charge is £0.23 per student).

The actual searching for duplicates takes around 30 seconds. For each mail box the educator has available a clear table with a list of uploaded documents with a percentage-based, graphical and colour-coded scale of the number of duplicates found. When browsing files the original document with the original formatting is displayed, with colour-coded similarities and a summary of the discovered duplicates and an indication of the sources and the percentage similarity. For each source it is possible to display the corresponding indexed original text, or to display the source online.

Ephorus (www.ephorus.com)

According to the operator, the Ephorus database contains billions of web pages, documents sent by involved schools and other sources such as magazines, reference materials etc. However, the operator does not provide precise data and it is not possible to confidently evaluate the comprehensiveness of its index merely from the description of the application. The web environment for data management in the Ephorus application contains a Czech interface in addition to other languages.

[4] For paraphrasing and machine translations was used the programming interface of the Microsoft Translator application[6], specifically the Translate and Paraphrase API methods.

The results of the check are sent to the educator by e-mail in the form of PDF attachments, or are available in the HTML format after logging in to the website. The formatting of the reports is the same in PDF and on the web – regarding the original formatting, paragraph breaks are retained, but not fonts, unlike with the Turnitin system. In the case of the test corpus, the delay between entering the document into the system and the sending of the results by e-mail was 1 hour 45 minutes.

In tests for finding duplicates the Ephorus system produced the worst results. It only found similarities in 10% of the sources, and these were mainly short, regular phrases such as transcription of the abbreviation WIMAX. The findings were rarely longer passages of text that would be direct proof of plagiarism.

MUNI systems (www.theses.cz)

The Theses.cz, Odevzdej.cz and Repozitar.cz applications were developed within the framework of centralised development projects of the Ministry of Education of the CR, and are operated by the Faculty of Informatics at Masaryk University. Until 2012 these applications were free of charge for the public higher education institutions participating in the projects, while since 2013 plagiarism searches for higher education institutions cost tens of thousands of Czech crowns per year, depending on the number of students. In spite of focusing on differing types of texts the applications use the same algorithm and database in the background, against which the search for duplicates is carried out.

With the MUNI systems the result of a check is not available until several hours after uploading (the test file was processed by the Odevzdej.cz application for more than 12 hours). This means that documents cannot be checked immediately after uploading, but the results are usually available the next day.

The MUNI anti-plagiarism systems work on the principle of comparing a pair of documents against each other. The application cannot stipulate the overall percentage of similarity of an analysed document against all other documents (it only knows the similarity between individual pairs). Only similarities of over 5% of one or the other document in the pair being compared are displayed, meaning that if, for example, a document under analysis has 50 pages, this system would not even report a case in which 2 pages of the text were copied from another similarly large document[5].

The similarities discovered between a tested and original document are available to educators in a text report, similar to that from the Ephorus system, in which the system retains only paragraph breaks. The size of the letters, headings and other formatting is ignored, meaning that navigation within a text is more difficult than in the case of the Turnitin system.

[5] On 2 September 2013 a new version for the display of similarities was put into operation in the MUNI systems, without a second list in the default state. The display of the overall percentage of similarities is planned for the future.

GooglePlagiarism

The GooglePlagiarism desktop application for the Windows operating system was programmed by the author of this document in March 2010, and later, depending on need, modified to its current form used in this test. The application was used by the author, for example, for the preparation of analysis materials for the TV shows Reportéři ČT and Události, komentáře (5), (6), (7).

The application splits (parses) an input Microsoft Word document into sentences, and then searches for exact matches (as phrases) using the Google Web Search API. Due to the limits on the free version of the Web Search API regarding numbers of queries over a specific time, the analysis of the test corpus took 3 hours. The results of the analysis are available as an HTML report without retention of formatting, with colour-coded duplicates and an overview of sources.

The GooglePlagiarism application found matches (at least partial) for 58% of the tested documents. It thus achieved the best score in terms of the total quantity of documents found.

TEST EVALUATION

The text file generated from the test corpus was evaluated using the individual testing systems for the presence of duplicates. The findings were then assessed and manually evaluated as a finding directly proving plagiarism, a finding only resulting in the suspicion of plagiarism, or a random finding that did not constitute suspicion of plagiarism.

Table 1 contains the author's subjective evaluation of the control and available functions in the individual tested systems on the basis of the tests performed.

Criteria	Thesis	Turnitin	Ephorus	GooglePl.
time to process	✗	✓	!	✗
readability of results	✗	✓	!	!
total sum of the similarity	✗	✓	!	!
minimal similarity	✗	✓	✓	!
price	✓	✗	!	✓
integration in to IR's	✓	✗	!	✗
deduplications of results	✗	✓	✗	✓

Table 1 Evaluation of system control and functions

The comparison of the applications shows a fundamental difference between the Theses.cz and Turnitin systems, where the Theses.cz system stands out because of its low price and

level of integration with a school information system. On the other hand, Theses.cz very much lagged behind in terms of the long time needed to process documents, the lack of clarity of the results, the absence of de-duplication of sources, and its inability to calculate the overall level of similarity. In terms of functionality the Ephorus system takes the middle road, however its usefulness is compromised by its poor results in the duplicate search test.

The following hypotheses stipulated at the start of the testing were evaluated on the basis of the statistics processed in Excel:

1. The application is able to detect a single sentence copied from a source document.
2. The application is able to detect a single paragraph copied from a source document. The application is not affected by potential line breaks, indexes etc. in the source or tested document.
3. Successful detection is not impacted if the plagiarist adds/removes a word in the copied sentence.
4. The application can detect Czech texts irrespective of the use of diacritics.
5. Successful detection is not impacted if the plagiarist paraphrases a single word in a sentence.
6. Successful detection is not impacted if the plagiarist paraphrases a whole sentence.
7. Successful detection is not impacted if the plagiarist translates text from/to Czech.
8. The Theses.cz system should achieve the best results in the detection of plagiarism in Czech theses and dissertations.
9. A low percentage of the total number of similarities will be detected from the Anopress source compared to sources freely available on the internet.
10. Better results with EIZ and Open Access sources are achieved by foreign tools rather than Czech ones.
11. Very good results for web sources will be achieved by systems using web search services.

The envisaged validity of hypotheses 1 and 2 is calculated as the ratio of documents detected by the system in question to the number of documents in the whole text corpus. In the case of hypotheses 3 to 7 the validity of the hypotheses for the system in question was stipulated as the ratio of the detected number of concrete modifications to the number of detected duplicate copied sentences evaluated by the system (meaning that if the system detects a sentence without modification, it should also detect a modified sentence). The detailed results of the test are given in a separate annex.

Table 2 shows the evaluation of the individual hypotheses shown above in the tested systems (value <33% - hypothesis rejected, ≥ 67% - hypothesis confirmed, ≥ 33 and < 67 – hypothesis could not be confirmed or rejected).

Hypothesis	Thesis	Turnitin	Ephorus	GooglePl.	avg
1	12%	40%	2%	56%	28%
2	14%	42%	6%	46%	27%
3	100%	100%	0%	0%	50%
4	100%	100%	0%	80%	70%
5	67%	100%	0%	4%	43%
6	0%	88%	na	0%	29%
7	0%	0%	0%	0%	0%
8	10%	50%	10%	30%	25%
9	0%	0%	0%	0%	0%
10	0%	40%	0%	70%	28%
11	20%	50%	0%	80%	38%

Table 2 Evaluation of the hypotheses in the tested systems

CONCLUSION

None of the tested systems is able to perfectly detect the source used. The Turnitin and GooglePlagiarism systems, however, achieve significantly better results in terms of the detection of copied sentences or paragraphs (hypotheses 1 and 2). The Theses.cz application achieved worse results probably because of the effect of the minimum necessary similarity of 5% of the content of the document. The Ephorus system fared worst in the test, finding a minimum of documents indicating plagiarism.

Hypotheses 3 and 5 were confirmed with the MUNI and Turnitin systems. Successful detection is not impacted if the plagiarist adds/removes/paraphrases a word in a copied sentence. If a whole sentence in a foreign-language document is paraphrased (hypothesis 6), only the Turnitin system achieves good detection results. Except for the Ephorus system, the other applications are not impacted if diacritics are removed from a copied text (hypothesis 4).

None of the applications is able to detect a text translated from a foreign language. Not even the support for translations in the Turnitin system (beta version) was able to help with the detection. Hypothesis 7 was not confirmed in any of the tested systems.

In view of the 5% limit for detection in the Theses.cz system, hypothesis 8 (that this system would be the best for searching Czech theses and dissertations) was not confirmed (success rate 10%). For these purposes the Turnitin (success rate 50%) and GooglePlagiarism (success rate 30%) systems produced the best results.

It was confirmed that the systems will have problems detecting duplicates from the Anopress source (hypothesis 9). In searching Open Access materials and EIZ the GooglePlagiarism application in particular posted very good results (success rate 70%) followed by the Turnitin system (success rate 40%). In view of the low success rate of the search by the Ephorus application, however, it is not possible to unequivocally confirm hypothesis 10 – that all the foreign applications would achieve better results from Open Access and web sources.

Hypothesis 11 was confirmed, namely that an application using web search services will achieve very good results for web sources (success rate 80% compared to the second-place Turnitin with a 50% success rate).

The conclusions given above are valid for the analysis prepared from 300 samples of text from 50 documents. More precise results may be achieved through incorporating a significantly higher number of documents into the corpus, when at least partial automation of the evaluation of findings would already be appropriate.