Abstract

CCP (Center for Counter Plagiarism) is a tool designed to aid in the process of verifying digital documents in order to ensure their originality and appropriate usage of references. In this paper we describe the design and implementation of CCP. We compare our approach with other existing anti-plagiarism systems and their associated techniques. The main contribution of CCP is a simple yet meaningful graphical user interface that helps in the decision making process when drawing the line between fair use and plagiarism. We also analyze and discuss the preliminary results of performance and usability tests run on CCP.

1. Introduction

The wide availability of digital documents through the World Wide Web has opened a myriad of possibilities for students, instructors and researchers. Information is now just a few clicks away, which is of great help in scholarly activities, but raises new issues on the vulnerability of digital documents as they can be reused in ways that may not always be in the best interest of their original authors. Technology has reached a point where it can provide unlimited access to resources; however, it still does not properly safeguard intellectual property [1].

As the amount of accessible information rapidly increases, so do the difficulties for tracing the sources of new materials and for ensuring their originality. This issue becomes critical especially in academic environments where research reports and other documents are submitted and reviewed nearly on a daily basis. Students’ reports are often based on Internet sources, rather than their physically printed counterparts. When dealing with multiple reports simultaneously, the task of manually verifying each of them becomes unwieldy, due to the practically countless number of possible sources.

The main concern, however, should be not only the proliferation of verbatim copying or intentional paraphrasing of other authors’ work but also an even more significant issue: the lack of education on the subject. Currently, students have scarce knowledge, if any, regarding copyright issues, fair use and proper source citation when it comes to digital sources accessible on line.

Drawing the line between fair use and plagiarism for digital contents is not a simple endeavor. Thus, publication boards and universities are issuing plagiarism policies, as well as using detection tools and techniques, in order to deal with the escalating amount of plagiarism incidents [16]. Take the case of the ACM’s Publications Board, which was compelled to issue its first policy on plagiarism on November 2005 to define the levels of offense, open an investigation process and set the inherent penalties [12]. Hence, great importance is given to the development of new tools for plagiarism detection, which will aid in the verification of digital documents.

In this paper we describe CCP, a tool we developed to explore various plagiarism detection techniques. We compare CCP with other existing anti-plagiarism systems and discuss the results of preliminary tests we have conducted.

2. Defining plagiarism

Although definitions of plagiarism do exist (typically involving the use of ideas of another author without permission or appropriately crediting the source), when it comes to addressing plagiarism issues in practice, the definition becomes blurry. Gotterbarn explains that after a birds-of-a-feather session on plagiarism and self-plagiarism, which took place at the SIGCSE 2006 Symposium, the main points of agreement were only that plagiarism is wrong, on the increase and hard to detect [5].

Boisvert and Irwin [12] define paper plagiarism as the verbatim copying, near verbatim copying or purposely paraphrasing portions of another author’s
paper. The underlying complexity is related to setting the boundaries upon the extent of the copying and the clarity of the case. There are still some gray areas that should be considered while deciding whether or not plagiarism exists, thus the significance of the above discussed policies.

2.1. Types of plagiarism

There are several possible approaches for the classification of plagiarism. Some authors think it is important to distinguish between plagiarism from other authors’ work and self-plagiarism, because it is widely accepted to reuse portions of previous work, as long as there is a reference to it, or it is a well-known definition or equation [6][12][13].

Another classification, which applies more specifically to cases where there is a collection of works, comprises intra-corpus plagiarism and extra-corpus plagiarism [8]. The collection of submitted works constitute the corpus, and we could search for plagiarism inside (intra-corpus) or outside (extra-corpus) the collection. For example, two students presenting the same academic report would be an instance of intra-corpus plagiarism, while a student submitting a report acquired from a paper mill would be a case of extra-corpus plagiarism.

Plagiarism could also be identified as intentional and unintentional. Although defining whether or not plagiarism is intentional could be a rather controversial issue, there are students who do not know how to properly document their sources, or that the original authors should be given credit for their work [14]. As Gotterbarn puts it, we are trying to explain that plagiarism is stealing intellectual property to a generation that is used to download anything they require from the Web [5]. Or as Boisvert [12] so clearly phrases it, cutting and pasting digital contents has become so natural that it is easy to forget whose work we are actually manipulating.

2.2. Plagiarism in the classroom

Some of the most common sources for students to plagiarize are available online, such as free research papers found in electronic databases, essays shared by other students or papers sold by commercial paper mills [14]. There are even “private tutors” who sell reports to the students, copy material from several sources on their behalf and send it electronically, like in an instance that caused the opening of an investigation at the RMIT University in Australia [9].

Reasons why students copy another author’s ideas or words are diverse, as well as the strategies they have devised to overcome the rules of what could be called appropriate academic behavior. Some of these strategies include [13]:

- Copying or downloading an entire paper.
- Cutting and pasting from several sources to create a new report.
- Intentionally paraphrasing another author’s work, without crediting the source.
- Quoting less than what has been copied, or entirely neglecting the quotation marks even for verbatim copies.
- Faking a reference, especially when the source is hard to find.

One of the most common kinds of plagiarism is the so called “cut and paste” plagiarism, thus making it the more broadly discussed. Luckily, Internet seems to be both the problem and the solution, as online digital contents may be used to automate the tasks of plagiarism detection and source tracing.

2.3. Assuring the fair use of digitally available contents

Fair use of another author’s work implies crediting the source, appropriately referencing the use of ideas and using quotation marks to enclose textual citations to their full extent. How can we assure the “good” use of digitally available contents? We believe that instructors should proactively try to prevent plagiarism from happening in the first place, by educating students on the issue: defining what is considered plagiarism, discussing the reasons why it is important to reference sources, establishing policies and enforcing them by making penalties clear before reports or essays are even assigned.

Once students fully understand the problem and are aware of the consequences of plagiarizing, detection techniques can be applied and tools can be used to ensure that they are actually complying with the good practice of fair use, and what is considered to be appropriate academic behavior.

Techniques for detecting plagiarism could be divided into manual and automated. The manual approach is labor intensive, time consuming and requires the expertise of the reader to accurately detect plagiarism [16]. The reader must carefully look for anomalies such as: mixed citation styles, lack of references, unusual formatting, tone or voice differences, unfamiliar terminology or verbatim copying of a renowned source [13]. Search engines such as Google’s or Yahoo’s can be used to assist in
manually finding matching words, sentences or paragraphs.

Automated tools apply multiple techniques to identify possible plagiarism instances, some of which are listed below:

- In the case of intra-corpus search, adding a watermark to each work at submission time, and comparing subsequent works with it for detecting possible duplicates [2].
- Finding similar documents by comparing their “fingerprints” (small set of character sequences that summarize a document, obtained using an algorithm such as winnowing) [6].
- Comparing structural characteristics (a merge between the keywords and the parsed tree for sections and paragraphs layout) of the documents [1].
- Using stylometry as a statistical approach to determine the authorship of a document [15].
- Using the sliding window technique and average length per word metric to analyze documents [16].

### Table 1. Comparison of plagiarism detection systems

<table>
<thead>
<tr>
<th>System</th>
<th>Interface</th>
<th>Process</th>
<th>Corpus search</th>
<th>Availability</th>
<th>Display</th>
<th>Search</th>
<th>Detection technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>arXiv</td>
<td>-</td>
<td>Local</td>
<td>Intra-corpus</td>
<td>-</td>
<td>Graphic</td>
<td>No</td>
<td>Fingerprints</td>
</tr>
<tr>
<td>CHECK</td>
<td>Desktop</td>
<td>Local</td>
<td>Intra-corpus</td>
<td>-</td>
<td>Text</td>
<td>No</td>
<td>Structural comparison</td>
</tr>
<tr>
<td>CopyCatch Gold</td>
<td>Desktop</td>
<td>Local</td>
<td>Intra-corpus</td>
<td>Commercial</td>
<td>Text</td>
<td>No</td>
<td>Collision detection</td>
</tr>
<tr>
<td>Swansea</td>
<td>Desktop</td>
<td>Local</td>
<td>2 files</td>
<td>Free</td>
<td>Text</td>
<td>No</td>
<td>Stylometry statistics</td>
</tr>
<tr>
<td>WORD Check</td>
<td>Desktop</td>
<td>Local</td>
<td>2 files</td>
<td>-</td>
<td>Text</td>
<td>-</td>
<td>Text comparison</td>
</tr>
<tr>
<td>ROBOPROF</td>
<td>Applet</td>
<td>Remote</td>
<td>Intra-corpus</td>
<td>-</td>
<td>-</td>
<td>No</td>
<td>Watermark</td>
</tr>
<tr>
<td>SNITCH</td>
<td>Applet / Web</td>
<td>Remote</td>
<td>Extra-corpus</td>
<td>Free</td>
<td>Text</td>
<td>Yes</td>
<td>Sliding windows</td>
</tr>
<tr>
<td>Paperbin</td>
<td>Web</td>
<td>Remote</td>
<td>Extra-corpus</td>
<td>Commercial</td>
<td>Text</td>
<td>Yes</td>
<td>Web metasearch</td>
</tr>
<tr>
<td>EVE2</td>
<td>Web</td>
<td>Remote</td>
<td>Extra-corpus</td>
<td>Commercial</td>
<td>Text</td>
<td>-</td>
<td>Web search</td>
</tr>
<tr>
<td>GLATT</td>
<td>Web</td>
<td>Remote</td>
<td>-</td>
<td>-</td>
<td>Text</td>
<td>No</td>
<td>Screening program</td>
</tr>
<tr>
<td>My Drop Box</td>
<td>Web</td>
<td>Remote</td>
<td>Extra-corpus</td>
<td>Commercial</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Turn-it-in</td>
<td>Web</td>
<td>Remote</td>
<td>Extra-corpus</td>
<td>Commercial</td>
<td>Text</td>
<td>No</td>
<td>Fingerprints</td>
</tr>
</tbody>
</table>

### 3. Related work

Table 1 summarizes the characteristics we considered relevant for our research on multiple plagiarism detection tools. Some of the data was directly obtained from research publications about the tools; however, some was acquired from reviews by other authors.

Table 1 comprises information about the following systems: arXiv (detection in arXiv’s corpus) [6], CHECK [1] [6], CopyCatch Gold [3] [4] [15], EVE2 [13] [14] [15] [16], GLATT [14] [15], MyDropBox [16], Paperbin [14], ROBOPROF [2], SNITCH [16], Swansea Institute (proposed tool) [15], Turn-it-in [4] [14] [15] [16] and WORDCheck [14]. Whenever the characteristic did not apply to the system or was unknown, a dash has been inserted.

From all the reviewed tools, EVE2 and Turn-it-in are perhaps the more broadly discussed. CCP’s approach is similar to that of EVE2, but includes
several additional characteristics. Unlike most of the available plagiarism detection tools that are commercial and display the results through text, CCP is freely available and introduces a graphical interface to present reports. Searches are performed both intra-corpus and extra-corpus, giving the system yet another advantage over other tools, which usually choose only one of those, or compare only 2 documents at a time.

In short, CCP is Web-based, works both on internal and external collections, is publicly available, displays results graphically and takes advantage of available search engines via Web services.

4. CCP design and implementation

CCP is designed as a central server that groups the functionality of several remote and local search engines to provide homogeneous text or document search across diverse repositories. Communication depends on each engine’s proprietary languages and specifications, and the system is divided into modules that interact with each engine. These exchangeable, extensible and reusable modules are handled like Lego pieces, which can be easily added or removed as required (see Figure 1).

Our aim was to design an automated tool to facilitate the process of searching for text using web search engines and intranet services following a standardized scheme. CCP is available as a Web application, which allows users to upload a document or a piece of text for verifying its originality and detecting possible plagiarism instances. Each document is divided into sentences, so a quick fine grained search is possible by sending those sentences to several engines that retrieve similar matches. Afterwards, a graphical report is generated and presented on a Web-based interface. All the process is centralized on the server side of the system, thus the connecting clients are lightweight and require only limited resources from the client machine.

4.1. Implementation

Implementation of the system was based on the Java programming language and complies with the J2EE specifications, since both support the development of reusable components and data exchange using open standards and protocols. Further flexibility of the system was achieved by relying on widely accepted standards and open technologies, such as XML [17] for the data model, and SOAP [10] as the protocol for communicating system components.

Compliance with these standards fosters the seamless integration of the modules, created with different purposes, to operate as a single web service for a more specific application. Currently, the system is capable of retrieving data results from two public search engines [7] [18], and the connection to Intranet resources is implemented using Nutch-based search services [11]. Other functionality can be added as the project continues to grow by using public services and APIs as they become available.

4.2. User interface

It is important to clarify that CCP is a tool that supports the verification process, supplying elements that might point out plagiarism instances. The responsibility of analyzing the results, however, lies in the users of the system as CCP does not intend to provide a final verdict. Keeping in mind that the results...
should be as helpful as possible for easily determining whether further examination is required, we focused our efforts on producing a semantically meaningful graphical interface for presenting the results in a report that enables a quick interpretation.

Analyzed sentences are catalogued into one of three categories: copied, similar or not found. They are given a color representation resembling that of a traffic light (red, yellow and green, respectively). Reports with increasing levels of detail are generated, thus allowing users to analyze documents using different approaches, from an overall look at the percentages, to a fine-grained detail of every sentence in the document. Information is revealed gradually, so as to make it as simple as possible for the user to decide whether or not a more detailed investigation is required.

Figure 2 illustrates an example of the graphical report, which the user visualizes by default. The upper section of the report uses a color bar to represent the percentage of sentences that belong in each category. Just by looking at the color bar, users should get an idea of whether further analysis of the document is needed. If only a few sentences were copied (i.e., only a small percentage of the bar is red), then it is very likely that the verbatim copies found are titles, names, dates, short sentences, etc. and there is no need for further analysis. On the other hand, if the red or yellow percentage of the bar is considerable, users should continue to examine other available reports.

The lower section of the report displays the total percentage of copied or similar sentences by Web page, and could be used as a second hint of the document’s origin. If only a few pages have high percentages of such sentences, some parts of the document might have been copied from there. In this section only the top three pages, ranked by percentage, are displayed. For a complete review of all the Web pages and a graphic display of the position of the copied sentences in the original document, users may open the report by Web pages (Fig. 3).

A more detailed report is depicted in Figure 4, in which the sentences are represented according to their position in the original document. The color dispersion along the bar is also a good indicator for detecting plagiarism; if there are several red or yellow sentences in a row, a sizable chunk of text might have been copied from the same source. As for the second section of the detailed report, all the sentences are enumerated in the same order in which they appeared in the document, for the users to verify that it is indeed a possible plagiarism instance. This information allows the users to promptly discard short sentences, appropriately referenced quotes and chapter titles.

Since documents of a considerable size may comprise a large number of sentences, we customized the interface to easily navigate through them. By clicking on a sentence’s corresponding mark on the upper section color bar, the interface automatically looks for the sentence in the lower section, which indicates the color representation and the original text that was initially searched for.
An even more detailed description about each sentence search can be displayed by opening the sentence details report (Fig. 5). In this report, the users can compare the original sentence with the snippet of the sentence found online, where all the matching words are displayed in boldface for easily identifying the degree of similarity between both sentences. There is also a link to the Web page where the sentence was found, the total percentage of sentences found in the same source and the graphical representation of their position in the document.

CCP processes and searches each sentence asynchronously, thus users can upload a document or a text segment and check how reports progress at a later time. It was necessary to use this approach because of the extent of some documents, which may involve lengthy, time-consuming operations. Response time varies depending on the size of the document, the number of repositories to search from, and the dynamic response time of the search engines being queried. A batch mode interface supports uploading several documents simultaneously; currently, it allows users to request the verification of up to 5 files at a time. Processing multiple documents will, however, increase the response time.

5. Evaluation

We have concentrated our initial assessment of the potential of CCP in the context of our home institution, the Universidad de las Américas Puebla (UDLA). As in many other institutions of higher education, plagiarism at UDLA has become a serious and increasingly common problem among students.

Faculty members are faced with the task of manually verifying multiple documents in order to detect possible instances of plagiarism; which is both laborious and time consuming. Even in those cases when the suspicion of plagiarism arises, it is not easy to exactly pinpoint the original source, and gathering evidence to justify a claim is still a delicate matter.

CCP was presented to our faculty as a tool to aid in this verification process, and an evaluation version was provided so they could test its functionality in actual scenarios with their courses. Subsequently, we conducted interviews and applied questionnaires to have a better grasp of their stance on the subject and the results attained by using the system.

In our test setting, assignments are submitted digitally; students have individual electronic portfolios (e-portfolios) where their work is stored for each course enrolled at the university. Thus, we start from the premise that instructors already have a digital version of the documents, which can readily be uploaded to CCP.

5.1. Usability tests

Usability tests were conducted in three different modalities: heuristic evaluation, the think aloud approach and inspection of recordings of the user interaction. Users’ tasks consisted of registering into the system and uploading documents for verification. The heuristic evaluation considered guidelines such as: status awareness, user control, consistency and standards, errors prevention, aesthetics, flexibility, efficiency, help and documentation. In the think aloud approach users verbally described their actions along with any idea that came to their mind while using the system. Sessions were video recorded. Examining session logs was useful for the analysis of the duration of each activity and the effectiveness of the tasks.

5.2. Functionality tests

A workshop with 15 faculty members was arranged for the functionality tests. The main goal of the session was to test the system in a real scenario with several users’ registration and multiple documents’ searches occurring simultaneously. Performance of the system was satisfactory, and the web services based architecture proved its reliability.

5.3. Results

In addition to the supervised tests described above, CCP has been open for evaluation during the current semester (see http://ict.udlap.mx:8088/plagio). Initial results show a total of over 400 documents during three months, which represents an average of five documents per day. Each document was divided into an average of 33 sentences, which where then verified for literal occurrences and similar phrases. Since the project’s evaluation started it has searched for nearly 13,150 sentences, and from those 17 percent where detected as copied, 54 percent were suggested as similar and 29 percent where not found.
An interesting initial finding was that detecting a plagiarized document is not as uncommon as instructors expected, and that plagiarism is more frequent in online courses. Our faculty noted that CCP’s interface facilitated the process of detecting plagiarism. When that occurred, documents were returned to students, the situation was explained and they were requested to resubmit an improved version of their reports.

Results of the heuristic evaluation showed that instructions and informative messages could be improved but did not have a significant impact on the cases being observed. Using the think aloud approach, we identified common pitfalls in the interaction with the system and the users’ perception about the system’s interface. Initial impressions were favorable, as users reported that the graphical reports were useful, and that the use of colors was easy to understand. Users did notice that feedback was necessary to determine how much progress the system had made in verifying each document. The problems identified during the usability testing were addressed and a progress bar indicating the status of the verification was included to keep users aware of the process’ status.

Although CCP was initially devised as an aid for plagiarism detection, during the tests we realized it could also have a more positive side. One of the papers we used as an example for testing was a research paper written by some of our lab members, and was an instance of non plagiarized document. We were pleased to review the report and find out that several sentences, marked as similar, were actually references to related work and pointed to related ongoing research.

6. Ongoing and future work

Considering the results of usability tests, the system interface has been modified to improve its usability and make it more understandable. More changes, however, will be required as we gather the results of usage tests in courses from different areas at the university.

We also plan to explore the effectiveness of new search modules, both with regard to quality of the results and response time improvements. The modules we plan to add include intra-corpus search on repositories available within the university, and extra-corpus search using publicly available search engines.

As an ongoing work we are refining the algorithms for parsing the text into sentences. For example, we will exclude very short sentences, as well as titles, dates, names, and table and figure captions. This will not only reduce search overload but also the unnecessary red signals for copied sentences, making the graphic report more understandable for the user.

Based upon the interviews conducted in the evaluation process, instructors requested to have a bar included in the interface to display the percentage of completion for each report, and a mechanism to cancel an ongoing search. Since the information on the Web is not static and is constantly changing and expanding, we plan to implement a dynamic search. Two search requests for the same document can generate different results when performed separately; therefore the report could become a richer source of information if the search is performed periodically to check for updates on the search engine indexes, or additions to the intra-corpus repositories.

Currently, the integration of a translation module is considered only from Spanish to English and vice versa, but more modules will be added as required. Translation modules will also be included as Web services, and are expected to help in the detection of plagiarism instances where students directly translated texts found online. The possibility of refining and personalizing the search is also contemplated.

Further testing of the system is scheduled for the Fall semester, when the new courses start. Based on the results of those tests, CCP will be continuously improved to better suit the users’ needs.

7. Conclusions

Through this work we found that plagiarism in higher education is a rapidly increasing problem that has been intensified by the wide availability of online digital documents. In order to deter plagiarism among the students, we developed CCP, a system with basic searching capabilities both on the Web and Intranet repositories. CCP looks for verbatim copies and similar phrases from a digital document, to detect possible hints of plagiarism, and generates reports to help the professor explain to the student where the potentially misused information is located and how to appropriately credit other authors’ work.

Other advantages of the system are that it can be used asynchronously and the reports are permanently stored on a data base for later use. The persistency of the reports allows users to have an historic trace to analyze and evaluate how students are progressing over the time.

Users considered that the graphical interface for presenting the reports was clearly understandable and supported both an overall and a more detailed analysis
of the documents. Response time, however, was still considered a drawback; which is why we are considering the testing of more search engines and the addition of a mechanism to simplify the text parsing into sentences and to reduce the processing time.

As for the detection rate, there are many more indexed pages in English than in Spanish on the Web, and that represents a disadvantage when looking for documents written in Spanish (the mother tongue of most of the students enrolled at the University). We expect to overcome this by adding the translation module to CCP, which will be capable of searching in both languages simultaneously.

The basic characteristics, functionalities and overall integration of the project have already been tested. Integration of more services and usability tests are required in order to release a complete version of the system. However, initial results show that the Web Services approach is ideal for the integration of distributed search processes.

The system proved to be efficient in automating the process of verifying a document, avoiding the manual approach that was both laborious and time consuming for instructors. Hopefully, CCP will be used as a tool not only to detect plagiarism but also to educate students on the issue and making them aware of the importance of fair use of information.

8. References


