Architecture of Plagiarism Detection Service that Does Not Violate Intellectual Property of the Student

Sergey Butakov[1], Craig Barber[1], Vadim Diagilev[2] and Alexey Mikhailov[3]

[1] SolBridge International School of Business, Woosong University, Daejeon, South Korea

[2] Department of Math and Simulation, Altai State Technical University, Barnaul, Russia

[3] Department of Public Safety, Altai State Technical University, Barnaul, Russia

Abstract. Plagiarism detection services (PDS) have become a vital part of Learning Management Systems (LMS). Commercial or non-commercial PDS can be easily attached to the most popular LMS these days. In most such systems, to compare a submitted work with all possible sources on the Internet a university has to transfer the student submission to the third party. Such an approach is often criticized by students who may see a violation of copyright law in this process. This paper outlines an improved approach for PDS development that should allow universities to avoid such criticism. The major proposed alteration of the mainstream architecture of the improved PDS is a move of document preprocessing and search result clarification from the server side to the client side. Such a split allows users to submit only limited information to the third party, and to do so in a way that will not make it possible to fully recover the submitted work but will allow the PDS to maintain the same search quality.

1 Introduction

Digital plagiarism is not a new phenomenon any more. The rapid development of the Internet along with increasing computer literacy made it easy and tempting for digital natives to "borrow" someone's work. Plagiarism is now a burning issue in education, industry and the research community. For example, in education one paper estimated the number of students in American high schools involved in different kinds of plagiarism to be as high as 90% [12]. Worse, there have been a number of cases of research communities in which people misrepresent other's work as their own [13]. There are a number of research areas concentrated around plagiarism. In this study we concentrate on plagiarism detection with particular focus on the technical and legal sides of it.

One of the problems that arises before anyone searching for the sources of the suspected paper is the degree of effort necessary to perform a "one-to-many" comparison where the "many" part of the relationship represents all possible sources. This part can be relatively small if the search has to be performed in the scope of a single learning

Butakov S., Barber C., Diagilev V. and Mikhailov A..
Architecture of Plagiarism Detection Service that Does Not Violate Intellectual Property of the Student .
DOI: 10.5220/0003568701230131
In *Proceedings of the 8th International Workshop on Security in Information Systems* (WOSIS-2011), pages 123-131
ISBN: 978-989-8425-61-4

community. Such a case occurs if for example, someone wants to check a submitted paper across all the papers submitted in the same college in the last five years. In this case the scope of search will be on the order of tens of thousands of documents. At the other extreme, if the same searcher wants to check the paper against all the publicly available web pages then the searcher must consider billions of documents. Global comparison can be done either by tools available at the school or it can be outsourced. Such a search performed at school would require the school to have its own web crawler and therefore to own infrastructure comparable to the ones employed by major Internet search engines. This option looks prohibitively expensive for most of the universities worldwide. The second option – outsourcing the search - may lead to Intellectual Property (IP) protection charges from the students. This paper proposes a much improved way to build a PDS, an architecture that would allow the school to avoid such concerns and yet allow the PDS to maintain an acceptable search quality. The rest of the paper is organized as following. The second section describes the major options how a general purpose PDS can be built, it also outlines why current architectures may be considered inappropriate from the IP protection point of view. Third section discusses a few legal cases against one of the major commercial PDS available and proposes a way that would allow educators to avoid such cases. The fourth section provides more technical details on the proposed solution, outlining the modified client server architecture for PDS.

2 Typical Architecture for Plagiarism Detection Tools

The PDS architecture for local (or in-house) search is very straight forward. The school maintains a database of all student works and compares each new document with the existing ones upon submission. IP-protection-wise the school can inform students that their submissions will remain in digital form in the school database and will be used solely for PDS. Global search on the other hand assumes that the paper has to be compared against all possible sources in on the open web. Under the open web paradigm we assume the publicly accessible web (rather than the "deep" web of company portals and resources available to subscribers only) is the proper field of search. As mentioned above, such a search theoretically can be performed on the school side or outsourced to a company that specializes in plagiarism detection. Practically the first option looks impossible for the vast majority of educational institutions around the world. Outsourcing, the second option, can be done in two major ways: outsourcing the whole process or outsourcing the most difficult parts of it.

The first way is to completely outsource the whole process, asking the external PDS to provide a detailed report on plagiarism if it was detected in the submitted document. There are two important points that should be highlighted here:

1. The complete student submission is transmitted to the PDS.
2. The PDS retains a copy of the document to use for the comparison with other submissions in the future.

Such an approach is used by one of the major players on PDS market - iParadigms LLC – in its well known Turnitin(tm) service (www.turinitin.com). This approach

raises concerns from students that the PDS is making profits using their submissions and therefore violates their IP rights. The next section of this paper discusses this issue further.

The second way is to outsource the most difficult part of the detection process – the global search for the documents that may be similar to the submission. In another words this part of the process should narrow down the scope of search from tens of billions to just tens of documents. Such an approach is actively used by language teachers: when they see a perfectly written sentence from a non-native speaker they often put it in the quotes and use a search engine to look for exact phrase on the web. Of course such manual search is slow and not very granular. A similar and very simple technique which employs Google Alerts has been proposed to monitor inappropriate copying from blogs [7]. Crot, one of the free PDS, uses similar brute force approach for search. It uses a sliding window of X words length, thus sending to the search engine all the phrases from the document that have been formed by this sliding window and so performing a very exhaustive search [6].

Two reports indicated that actually exhaustive search is not required to detect plagiarism. As Culwin & Child have indicated, the exact phrase search technique with use of public search engine can be effective to locate the source [10]. Crot authors also indicated that if a significant part of the paper was plagiarized from the Internet there is no need to send all possible queries to the search engine: even as low as 10% of these queries can help to locate the source [6]. This reduction can be very helpful for IP protection and will be discussed further in section 4.

Let's take a closer look at outsourcing global search to an external service. The internet search experiment with the Crot software indicated a linear dependence of the search time on the number of words in a document. The experiment has been performed on a set of 60 documents with length ranging from 350 to 3500 words. It was done on a dedicated server with a 100Mbs internet channel. Experiment indicated that the search time was about 5 minutes per 1000 words of the document. This time consists of the following elements: (1) time required for querying the search engine, (2) time required for downloading the suspected sources; (3) time for detailed comparison and (4) document hashing time. The latter two can be ignored because the documents were relatively small (average size of ~1200 words) and there were only a few hundred documents in the local database, which makes the detailed comparison very fast. Thus we can state that querying the search engine and downloading the suspected sources consume most of the time required for the search. Further investigation indicated that on average 26% of the search time can be attributed to the first component – querying the search engine. Therefore a significant reduction of total detection time can be achieved if either one or both of these steps is performed more quickly: sending queries and downloading documents.

From the IP protection point of view the approach of using an external service only to narrow down the search scope is better than the outsourcing of the whole PDS process. The external service is not getting the student submission in the same form as it was submitted and the original submission can barely be restored by the third party (search engine) from the search queries.

The solution to the problem of slow search would be to change the mode of submitting the search queries. Crot and similar systems submit the search queries one by one to the search engine and receive and process replies after that. They may use a multi-

thread approach at this point to improve the speed at this stage but multi-thread imposes certain requirements on Internet bandwidth. As the small experiment above indicated, querying the search engine causes a significant delay in the search time. Improvement could be made if the download and preliminary comparison could be outsourced to the search engine, e.g. if there is a service, complimentary to the conventional internet search engine, that would work as a proxy to the search engine facilitating the plagiarism detection process. Such a labor division where the on-premises part of PDS performs only detailed comparison and out-of-premises part does the global search could help to improve the speed. Such architecture has to be justified whether it can be subject to IP violation concerns. The next section analyzes a few cases on IP protection and provides insights on how the information flow could be organized to avoid collisions with IP protection mechanisms.

3 Legal Issues Related to Plagiarism Detection

There has been considerable controversy about the anti-plagiarism service by iParadigms LLC called "Turnitin®". Turnitin® requires the teacher or student to submit the student paper to them whole. Turnitin® then creates a "digital fingerprint" of the paper and archives the paper. This had generated considerable outrage on the part of students, and interestingly, in certain parts of the academic world [11]. The complaints have gone as far as legal action including law suits against iParadigms LLC, discussed in some studies [18, 4]. The threat of lawsuit is a severe disincentive for schools to using an anti-plagiarism system. The major problem here is that the university sends the student work to the third party (iParadigms LLC) and this third party stores the submission, using it to generate profit.

To see how to make PDS architecture viable from the IP protection point of view we now analyze the legal cases related to Turnitin®.

In legal cases against the Turnitin® service, iParadigms LLC was using fair use doctrine to protect its way of doing business. "Fair Use" is defined in the Copyright Act § 107, 17 USC § 107 [19] as follows:

> " ...the fair use of a copyrighted work... for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright.
> In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include —
> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
> (2) the nature of the copyrighted work;
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
> (4) the effect of the use upon the potential market for or value of the copyrighted work" [19].

Before examining the four factor test for the proposed architecture, we consider the small number of cases decided to date in regard to Turnitin ®.

The situation of the proposed architecture legally is somewhat dependent upon the situation of Turnitin®, not just because Turnitin® is being used to establish the legal standards but also because Turnitin® problems are likely to impact the proposed architecture commercially, either for benefit or detriment. The standing of the proposed architecture in relationship to Turnitin® and similar systems becomes important.
There have been four US Turnitin® cases, and Turnitin® won all four [1, 9, 2, 15]. The important case [1] is one in which the lawyer for the students managed to get the court to consider the issue of copyright infringement, in contrast to the other cases, which were determined on secondary grounds. In this case, iParadigms LLC and the court had to resort to the "Fair Use" doctrine to defend the Turnitin® software. The cases show the very high level of resistance to Turnitin®, and the persistence of the opposition. As can be seen from the cases, the court is well aware that Turnitin® is copying copyright materials without the permission of the copyright owners. However, other courts, especially courts in countries other than the USA, may not be as helpful in the future.
The brief analysis provided above shows that the architecture for a PDS will be less subject to legal actions if it is based on the following principles:

- Do not transfer the student work to third party computers in a form that can be considered a copy of the work.
- Do not store the copy of the paper on the third party computers for later use.

4 Proposed Plagiarism Detection Service Architecture

Figure 1 shows the main concepts of the proposed architecture for plagiarism detection service. The service itself is divided into an internal part running on university computers and an external part running on a third party computer. The internal part plays a service role when it communicates with the university portal and a client role when it submits search requests to the external part of the service. Most of the work is happening on university computers. The only outsourced part performed on third party software is pre-selection of probable sources of plagiarized paper from the web.
Like in the typical architecture outlined earlier, the process starts with student submission. The university portal prepares the document for the checkup, wrapping it with information on course, assignment, type of required checkup, etc. The internal part of PDS checks the document against the local database and prepares queries for external part. The distinctive feature of the proposed architecture is the way these queries are prepared. According to the legal requirements provided in the previous section these queries should not contain enough information to recover the submission but from an information retrieval point of view these queries should be enough to find similar documents on the web. Experiments indicate that even limited numbers of properly selected search queries can help to locate plagiarism sources on the web [6]. Essentially this means that the part of the PDS located on school infrastructure can prepare some queries from the key parts of the text, randomly shuffle them and send them as one large query to the external part of the PDS. Such a submission will not be subject to attack on IP grounds.

The sliding window algorithm that is implemented in Crot uses all possible queries that can be generated from the text. For example, for the Shakespearean quote «to be, or not to be: that is the question» with window length X=4 the algorithm will prepare seven queries: «to be or not», «be or not to», «or not to be», «not to be that», «to be that is », «be that is the», «that is the question». Obviously that if text has Y words the total number of queries can be defined as $N=Y-X+1$. Since we know that Y will be much larger than X we can say that sliding window algorithm will form almost Y queries. The initial student submission may be easily required from these queries because two neighboring queries q_i and q_i+1 have $X-1$ common words. The total number of words that will be sent to the search engine will be about $Y*X$. If we decide to select only Y_1 queries and Y_1 satisfies inequality (1) than there is a guarantee that it will be impossible to fully recover the initial student submission from the queries that will be sent to the external part of PDS.

$$Y_1 < Y/X \tag{1}$$

The essential queries that will go out to the third party must be sent in the plain text form. Any encryption / alternation of these queries will make impossible the use of web crawlers to narrow down the search scope. Of course inequality (1) does not guarantee that parts of the document cannot be restored but such partial recovery may not be considered as significant violation of student IP rights.

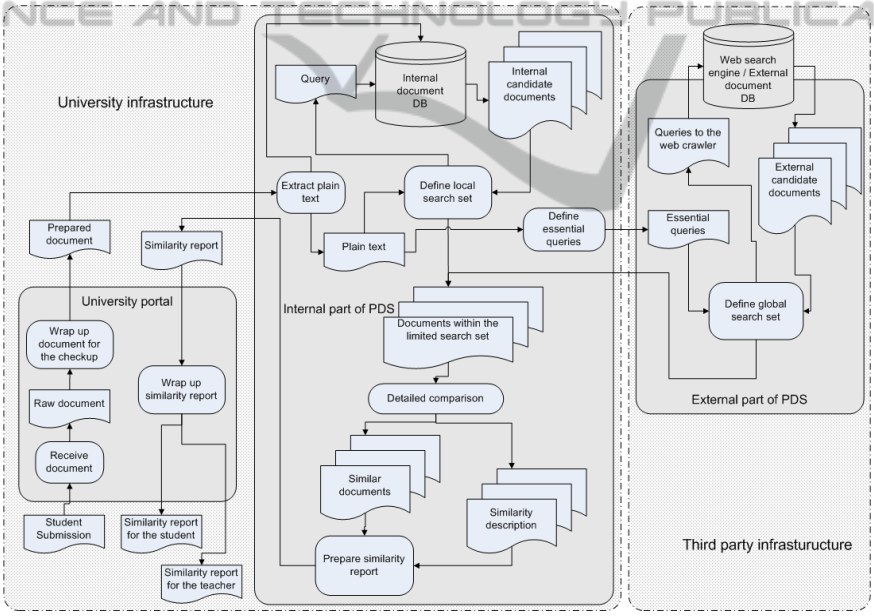

Fig. 1. Outline of the proposed architecture for PDS.

There are a number of researches that deal with detection of duplicated material available on the Internet. They vary from straightforward plagiarism detection in texts and program source codes [5] to web indexing [8, 14] and writing style detection for identification of individuals on anonymous web sites [3]. However we have not found any information on architectures for PDS that would enforce student IP protection

while performing the plagiarism detection process. The proposed architecture is focused on this issue and leaves room for a school to be flexible in IP protection management. When it is necessary to decide how much of student submission can go to the external PDS the school may decide on the tradeoff between the probability of catching small scale plagiarism (one sentence to a few paragraphs) versus sending of the copyrighted material to the third party. Inequality (1) can provide a boundary for that decision.

The proposed architecture increases the requirements to available computational power and storage capacity of university infrastructure. Additional storage is required to keep digital fingerprints of submitted documents. If such algorithms as Winnowing [17] or T9 will be used to compile fingerprints then we can expect additional 20-100% of plain text size increase in terms of required storage. Such an increase can be considered as insignificant as plain text does not consume much of the space and storage prices have followed a declining trend for years.

Additional computational power is required for calculation of a single document fingerprint, which is necessary for fast comparison of documents. Since there are many algorithms available that are linear to the document size such an increase can be also considered as insignificant. Additional requirements will arise to perform one-to-many detailed comparison on the internal part of PDS. These requirements impose high load on DBMS and therefore cost of DBMS licensing and maintenance on the university side will be the main factor that can affect the price of service in the proposed architecture. In most of the cases universities already own and maintain DBMS and therefore licensing cost increase may not be significant. Hardware investments should be also considered on the university side in this project. On the other hand there are no such heavy requirements to the DBMS on the third party infrastructure. This factor can decrease the cost of third party service and improve the possibility for start-ups to enter the PDS market, thus promoting competition.

5 Conclusions

In this study we concentrated on an architecture for a plagiarism detection service. The proposed solution contributes to many aspects of service architecture development. First of all this novel architecture makes student copyright protection a main goal and guarantees that no third party directly or indirectly makes any profit out of student work. Those small and scrambled portions of the student work that depart from the school IT infrastructure cannot be even used to fully recover the student work. A second distinctive feature is the outsourcing of the most time consuming part of the plagiarism checkup to the third party, thereby reducing the workload on the university IT infrastructure. Such outsourcing removes the necessity for the PDS instance in each school to have its own private web crawler and allows reliance on a common search engine for PDS in different schools.

In future research we will work on improvements of the details of the proposed architecture. One of the possible directions will be to include stylometry on the external part of the PDS to do preliminary checkups. This improvement could lead to better scalability of the service allowing the external part of the PDS to download more

suspicious sources of plagiarism and filter them before submitting results to the internal part of the PDS.

Acknowledgements

The authors would like to acknowledge a great help from Ms. Marina Kim and Ms. Svetlana Kim for conducting the experiment with Crot PDS.In this study we concentrated on an architecture for a plagiarism detection service.

References

1. A. V. ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630, 2009 Copr.L.Dec. P 29,743, 90 U. S. P.Q.2d 1513 (4th Cir. 2009).
2. A. V. v. iParadigms, LLC, 544 F.Supp.2d 473, 232 Ed. Law Rep. 176 (4th Cir. 2008)
3. Abbasi A. & Chen H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. 26, 2, Article 7 (April 2008), 29 pages. DOI=10.1145/1344411.1344413
4. Bennett M. G.. (2009) A Nexus of Law & Technology: Analysis and Postsecondary Implications of A.V. et al. v. iParadigms, LLC Journal of Student Conduct Administration, 2009, Vol. 2, no.1. p. 40-45
5. Burrows S., Uitdenbogerd A.L., and Turpin A. (2009). Application of Information Retrieval Techniques for Source Code Authorship Attribution. In Proceedings of the 14th International Conference on Database Systems for Advanced Applications (DASFAA '09), Xiaofang Zhou, Haruo Yokota, Ke Deng, and Qing Liu (Eds.). Springer-Verlag, Berlin, Heidelberg, 699-713. DOI=10.1007/978-3-642-00887-0_61
6. Butakov, S. and Shcherbinin, V. (2009) On the Number of Search Queries Required for Internet Plagiarism Detection. In Proceedings of the 2009 Ninth IEEE international Conference on Advanced Learning Technologies - Volume 00 (July 15 - 17, 2009). ICALT. IEEE Computer Society, Washington, DC, 482-483
7. Carter M. (2008). How to Use Google Alerts to Detect Plagiarism. Retrieved on December 15, 2010 from http://www.suite101.com/content/how-to-use-google-alerts-for-web-writers-a86525
8. Chowdhury A., Frieder O., Grossman D., and McCabe M. C.. (2002). Collection statistics for fast duplicate document detection. ACM Trans. Inf. Syst. 20, 2 (April 2002), 171-191. DOI=10.1145/506309.506311
9. Christen v. Iparadigms, LLC, August 4, 2010, 2010 WL 3063137, Copr.L.Dec. P 29,960, (E. D. Va. 2010).
10. Culwin F., & Child M. (2010) Optimizing and Automating the Choice of Search Strings when Investigating Possible Plagiarism. In Proceedings of 4th International Plagiarism Conference, Newcastle, June 2010. Retrieved on November 20, 2010 from http://www.plagiarismadvice.org/
11. Horovitz, S. (2008). Two wrongs don't negate a copyright. Florida Law Review, 60, 229.
12. Jensen, L.A., Arnett, J.J., Feldman, S.S. & Cauffman, E. (2002), It's wrong, but everybody does it: academic dishonesty among high school students, Contemporary Educational Psychology, 27(2), 209-228

13. Kompas (2010) Saving Indonesia from Traps of Plagiarism. Retrieved on November 5, 2010 from http://english.kompas.com/read/2010/04/28/02563687/ Saving.Indonesia.from.Traps.of.Plagiarism
14. Manku G. S., Jain A., and Sarma A. D.. (2007). Detecting near-duplicates for web crawling. In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA, 141-150. DOI=10.1145/1242572.1242592
15. Mawle v. Texas A & M University Kingsville, 2010 WL 1782214 (S.D.Tex. 2010).
16. Scanlon P. M. and Neumann D. R. (2002), Internet Plagiarism Among College Students, 43 J. C. STUDENT DEV. 374, 379 (May–June 2002).
17. Schleimer S., Wilkerson D., and Aiken A. (2003). Winnowing: Local Algorithms for Document Fingerprinting. Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 76-85, June 2003.
18. Sharon, S. (2009) Do Students Turn Over Their Rights When They Turn in Their Papers? A Case Study of Turnitin.com (June 1, 2009). Available at SSRN: http://ssrn.com/abstract=1396725
19. The Copyright Act § 107, 17 USC § 107 (1976)