International Review of Research in Open and Distributed Learning

Protecting Students' Intellectual Property in the Web Plagiarism Detection Process

Sergey Butakov, Vadim Dyagilev et Alexander Tskhay

Volume 13, numéro 5, décembre 2012

Special Issue: Technology-Enhanced Information Retrieval for Online Learning

URI : https://id.erudit.org/iderudit/1066984ar
DOI : https://doi.org/10.19173/irrodl.v13i5.1239

Aller au sommaire du numéro

Éditeur(s)

Athabasca University Press (AU Press)

ISSN

1492-3831 (numérique)

Découvrir la revue

Citer cet article

Butakov, S., Dyagilev, V. & Tskhay, A. (2012). Protecting Students' Intellectual Property in the Web Plagiarism Detection Process. *International Review of Research in Open and Distributed Learning*, *13*(5), 1–19. https://doi.org/10.19173/irrodl.v13i5.1239

Résumé de l'article

Learning management systems (LMS) play a central role in communications in online and distance education. In the digital era, with all the information now accessible at students' fingertips, plagiarism detection services (PDS) have become a must-have part of LMS. Such integration provides a seamless experience for users, allowing PDS to check submitted digital artifacts without any noticeable effort by either professor or student. In most such systems, to compare a submitted work with possible sources on the Internet, the university transfers the student's submission to a third-party service. Such an approach is often criticized by students, who regard this process as a violation of copyright law. To address this issue, 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 is to move document preprocessing and search result clarification from the third-party system back to the university system. The proposed architecture changes would allow schools to submit only limited information to the third party and avoid criticism about intellectual property violation.

IRRODL
THE INTERNATIONAL
REVIEW OF RESEARCH IN
OPEN AND DISTANCE LEARNING

Protecting Students' Intellectual Property in the Web Plagiarism Detection Process



Sergey Butakov
Concordia University College of Alberta, Canada
Vadim Dyagilev
Altai State Technical University, Russia
Alexander Tskhay
Altai Academy of Economics & Law, Russia

Abstract

Learning management systems (LMS) play a central role in communications in online and distance education. In the digital era, with all the information now accessible at students' fingertips, plagiarism detection services (PDS) have become a must-have part of LMS. Such integration provides a seamless experience for users, allowing PDS to check submitted digital artifacts without any noticeable effort by either professor or student. In most such systems, to compare a submitted work with possible sources on the Internet, the university transfers the student's submission to a third-party service. Such an approach is often criticized by students, who regard this process as a violation of copyright law. To address this issue, 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 is to move document preprocessing and search result clarification from the third-party system back to the university system. The proposed architecture changes would allow schools to submit only limited information to the third party and avoid criticism about intellectual property violation.

Keywords: Intellectual property protection; plagiarism detection; system architecture; social issues; learning management systems

Introduction

The rapid development of the Internet along with increasing computer literacy has made it easy and tempting for digital natives to copy someone else's work and paste it into their own. Plagiarism is now a burning issue in the education, industry, and research community (Spafford, 2011). For example, one group of researchers estimated up to 90% of students in high schools are involved in different kinds of plagiarism (Jensen, Arnett, Feldman, & Cauffman, 2002). Other research has shown that students with extensive exposure to the Internet are more inclined to be engaged in copy-paste practices (Underwood & Szabo, 2003).

Distance and online education can be even more vulnerable to plagiarism because of its remote and asynchronous nature. Concern about such vulnerability is growing along with the increasing number of online programs (Heberling, 2002; Marais, Minnaar, & Argles, 2006). In online course settings, PDS should be used as a tool to combat plagiarism and educate students on proper research and writing practices. As an example, Jocoy and DiBiase (2006) showed how automated plagiarism detection tools could be of great assistance in helping educators to raise student awareness about plagiarism and improve course outcomes. Due to its increasing importance, there have been a number of research projects concentrated on plagiarism recently. In this study, we concentrate on the plagiarism detection process, particularly focusing on the architecture of software tools used for detecting possible sources of plagiarism.

A major problem that arises for anyone searching for the sources of a suspicious paper is the degree of effort necessary to perform a document search to compare the suspected plagiarism with all possible sources. The comparison base can be relatively small if the search has to be performed in the scope of a single learning community, such as one college. At the other extreme, if the same search has to be done against all publicly available web pages, then the searcher must consider billions of documents. To perform such a search at the individual school level, each institution would require its own web crawler. This option is prohibitively expensive for most of the universities worldwide. The other option—outsourcing the search—may lead to intellectual property (IP) violation charges from the students (Bennett, 2009).

In this paper we propose an improved way to build a PDS with an architecture that allows a school to use the applicable IP policy and yet also allows the PDS to maintain acceptable search capability. The proposed framework is based on the idea that only a limited amount of information from the original submission is required to locate potentially similar documents on the Web. The scope of the proposed approach is to build a PDS based on a conventional web search engine. The goal is not to disclose the entire suspicious submission to the third party—the outsourcing company that runs the PDS. This third party may or may not have its own web crawler. In the latter case, the PDS has to employ a conventional search engine to look for potential sources. This approach restricts the usage of such functions as the hashing of search queries to hide content because conventional search engines such as Google, Yahoo, or Bing do not work with hashed queries. The approach assumes that the

information will be transmitted to the PDS in an unencrypted form, and thus to preserve students' IP we have to limit the amount of information that can be transmitted to the external part of the PDS.

The rest of the paper is organized as follows: The second section, Related Works and Existing Solutions, describes the major options for how a general purpose PDS can be built and outlines why current architectures may be considered inappropriate from the IP protection point of view. The third section, Proposed Architecture, discusses a few of the legal cases against one of the major commercial PDS available and proposes a solution that would allow educators to avoid such cases. The fourth section, Experiments, provides more technical details about the proposed solution, outlining the modified client-server architecture for PDS. It also shares the experimental results that show how much of the original submission should be transmitted to PDS to locate similar documents on the Web.

Related Works and Existing Solutions

Plagiarism detection services are in demand in many areas, including but not limited to education, publishing, and research proposal evaluations. Some of these areas may be affected by IP protection legislation. Most likely in education, not all student works are subject to IP protection, but senior projects and master's and PhD theses could fall in to this category. The same applies to research papers, proposals, reports, and patent applications. In this paper we will be talking about IP protection during the process of plagiarism detection in student papers, but most of the concepts are applicable to other areas.

There are a number of research studies that deal with detecting duplicated material available on the Internet. The applied side of this research topic has evolved from earlier projects examining plagiarism detection in source code (Donaldson, Lancaster, & Sposato, 1981; Krsul & Spafford, 1997) to copy-paste detection in essays and program codes (Burrows, Tahaghoghi, & Zobel, 2007). A number of related studies on system architecture have been done to examine web indexing (Zaka, 2009), spam protection (Urvoy, Chauveau, Filoche, & Thomas, 2008), and writing style detection to identify individuals on anonymous Web sites (Abbasi & Chen, 2008).

Technical intellectual property protection is a well-developed area in digital rights management (DRM) systems. There are many ways that DRM systems can be implemented and managed. One of the areas that could be related to plagiarism detection is digital watermarking, especially implementations that establish accountability for copying the protected digital object (Arnold, Schmucker, & Wolthusen, 2003). DRM and PDS are similar in the sense that they allow users to identify the source. For example, the embedded watermark in an Oscar Academy award movie sent out for prescreening can reference the person who was given that particular copy (Cox, Miller, Bloom, & Fridrich, 2009), and a side-by-side comparison of two documents can reveal similarities in the texts. But the ability to implement watermarking or steganography on plain text files is very limited as changes will be visible.

In contrast to the well-established research field of DRM, at the moment there are not very

many studies on IP protection during the plagiarism detection process. If we look at the PDS currently available on the market, we can see that the PDS architecture for papers submitted locally is very straightforward. The school maintains a database of all student works and compares each new document with the existing ones upon submission. In terms of IP protection, the school informs students that their submissions will remain as digital files in the school database and will be used solely for PDS.

Conversely, the Internet search assumes that the paper must be compared against all possible sources on the open Web. As mentioned above, such a search can theoretically be performed on the school side or outsourced to a company that specializes in plagiarism detection. Performing such a search would require schools to maintain a web crawler similar to the ones used by major search engines. Thus this option is cost-prohibitive for most schools. Outsourcing, the second option, can be done in two major ways: outsourcing the whole process or outsourcing the most data-intensive parts of it. An overview of these approaches is displayed in Figure 1. There are two important points that should be highlighted here: (1) the complete student submission is transmitted to the PDS and (2) the PDS retains a copy of the document to use for comparison to other submissions in the future.

Figure 2 illustrates how information flows in the case of complete outsourcing: It starts with a student submission on the left side and ends with two similarity reports after passing through the university side. Such an approach is used by major players in the PDS market, such as iParadigms LLC, with its well-known Turnitin® service (www.turinitin.com), and BlackBoard's SafeAssign (www.safeassign.com).

The second way is to outsource the most difficult part of the detection process: the global search for candidate sources—documents that may be similar to the suspicious submission. In other words, this part of the process should narrow down the scope of the search from the tens of billions of documents available on the Web to just dozens of documents. As Figure 1b shows, the university portal in this case has to perform the detailed comparison, but it is not as intensive as a complete Internet search. Many smaller scale companies utilize public search engines to perform such searches. Crot, one example, utilizes a global search API from Microsoft's Bing search engine to perform this kind of selection. It uses a sliding window *x* words in length, thus sending to the search engine all the phrases from the document that have been formed by this sliding window and in doing so performs a very exhaustive search (Butakov & Shcherbinin, 2009).

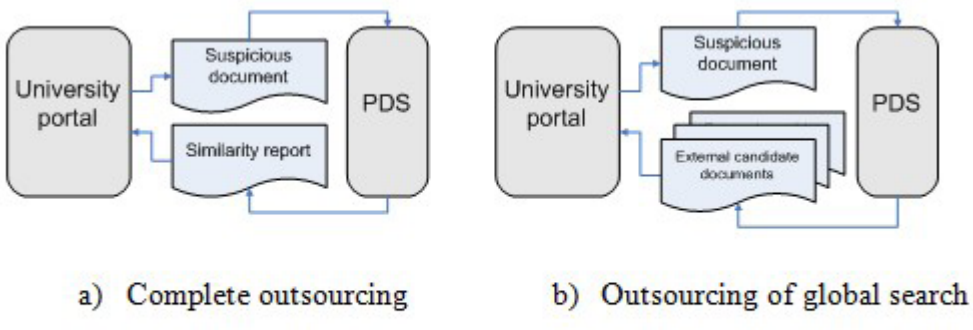


Figure 1. Two typical ways to outsource the plagiarism detection process.

Both of the plagiarism detection outsourcing approaches outlined in Figure 1 cause students to raise concerns that the PDS violates their IP rights. Up to now, four lawsuits have been filed against Turnitin by students. The company has won all of them, but in the latest lawsuit students were able to take their case to the point where the judge examined fair use policy tests outlined in US copyright legislation (Bennett, 2009). The students claimed that Turnitin PDS was making profits using their submissions, and therefore its owners should be liable for copyright violation. Although the service was not found guilty of this charge, such cases generate negative publicity for the schools involved. Also, similar cases could result in different decisions in other countries. For example, European Union law is known to be tougher concerning copyright protection.

Many universities are aware of these legal concerns. For example, the University of Maryland, University College suggests that faculty members consider the following questions when choosing PDS: "Will the [plagiarism detection] service archive all student submissions for further detection? Will all submissions be immediately destroyed after a report is generated?" (VAIL, 2012).

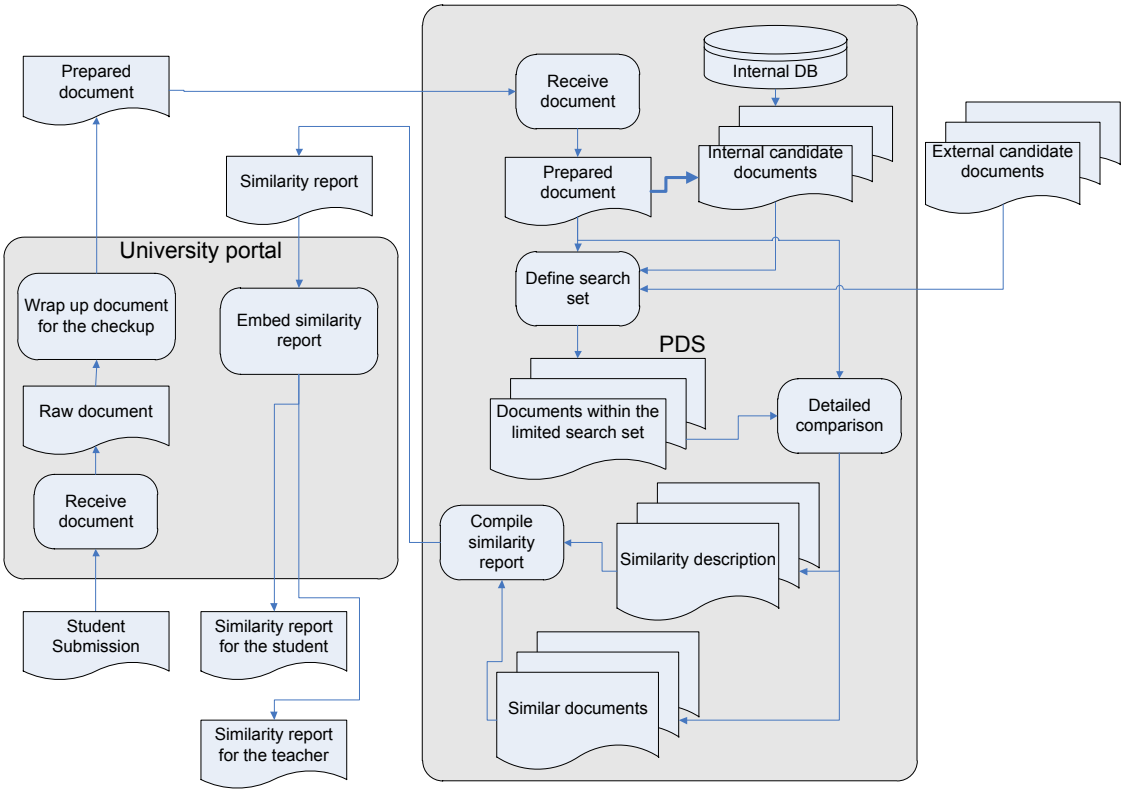


Figure 2. The complete outsourcing of the plagiarism detection process.

As illustrated by this example, from the IP protection point of view, using an external service to only narrow down the search scope is better than outsourcing the whole detection process because student submissions will not be stored and reused for profit. But there is still a concern because the suspicious submission goes to the third party. From a legal standpoint, some suggested using only a limited amount of information from the suspicious text to obtain potential sources from the Web (Butakov & Barber, 2012), but this left an open question about how such selection would affect the detection quality. Referring back to Figure 1, this approach means that the external PDS should not keep the submitted document and should not get it in its original form. The next section outlines the details of the architecture based on these principles and can serve as a blueprint for this sort of PDS development.

Proposed Architecture

Figure 3 shows the main concepts of the proposed architecture for PDS. The service itself is divided into an internal part running on university infrastructure and an external part running on a third-party system. 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 data processing is done on university infrastructure. The only step in the process performed on third-party infrastructure is the preselection of can-

didate documents or locating the probable sources of the plagiarized paper on the Web. The major difference between the general approach presented in Figure 1b and Figure 3 is the form and amount of information that is transferred to the third-party infrastructure. The proposed architecture sends out essential queries, instead of sending the complete student's submission.

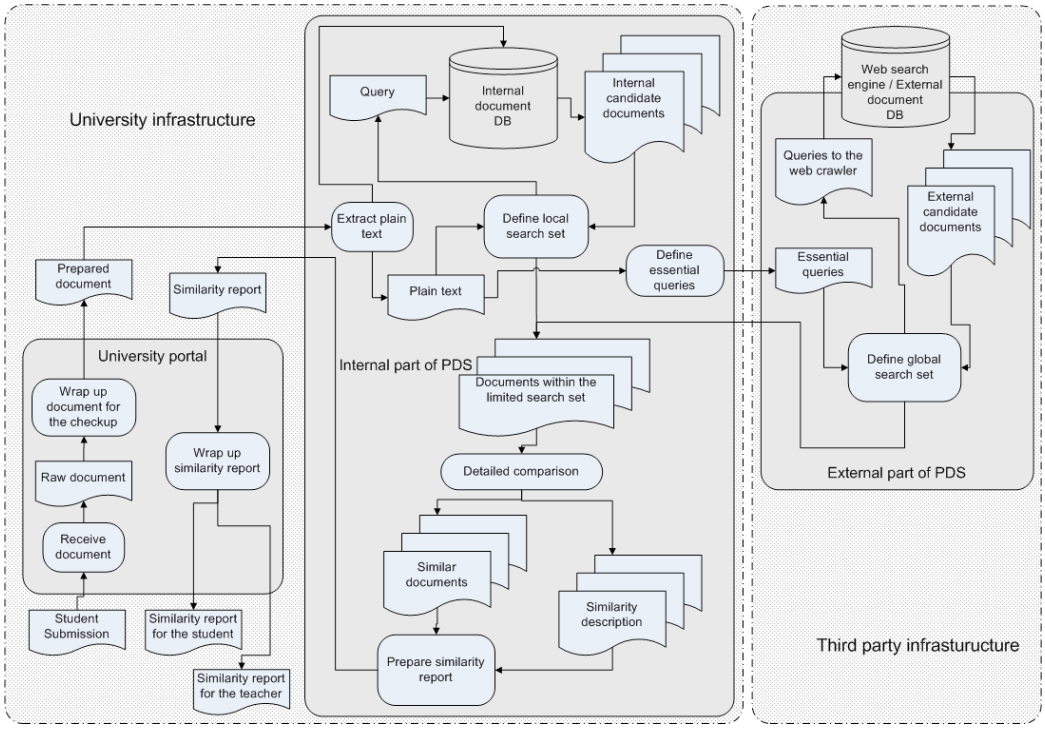


Figure 3. Outline of the proposed architecture for PDS.

To use the proposed architecture, we assume the whole suspicious submission is not required to select candidate sources. This assumption is based on the results of two reports that indicated an exhaustive search is not required to detect plagiarism. As Culwin and Child (2010) have illustrated, plugging exact phrases into a public search engine can be an effective way to locate the source of a suspicious paper, but the question of how to locate such an indicative phrase in a document remains. Butakov and Shcherbinin (2009) also indicated that if a significant part of the paper was plagiarized from the Internet, there was no need to send all possible queries to the search engine: Even as few as 10% of these can help to locate the source.

Like the typical architecture outlined earlier, the process starts with a student submission. The university portal prepares the document for the checkup, wrapping it with information about the course, assignment, type of required checkup, and so on. The internal part of the PDS checks the document against the local database and prepares queries for the external part. The distinctive feature of this proposed architecture is the way these queries are prepared. According to the legal requirements that protect IP, these queries should not contain enough information to recover the submission in its original form, and they should appear

significantly altered from the original document. Such a transformation would give the PDS protection from accusations of IP violation.

One of the more appealing ways to hide the content of the submitted document from the third party is to encrypt the queries and perform the search in an encrypted index (Song, Wagner, & Perrig, 2000; Goh, 2003). To maintain privacy during the search, these techniques require the search index to also be encrypted; therefore, it should be managed by the PDS. In many cases, private PDS outsource search index support. For example, Crot and SafeAssign employ Microsoft's Bing index to perform the selection of candidate documents. Such outsourcing makes it impossible for schools to use encryption mechanisms to hide a submission from the third party.

Experiments indicate that even limited numbers of properly selected search queries can help to locate plagiarism sources on the Web (Culwin & Child, 2010; Butakov & Shcherbinin, 2009). Essentially this means that the part of the PDS located on school infrastructure can prepare some queries from key parts of the text. These essential queries should be enough to locate the candidate sources. This technique is very similar to the approach used by many language professors when they see a grammatically perfect sentence written by a non-native student. They put the suspicious sentence in quotation marks and use conventional search engines to perform the search. There are a number of benefits to this approach. Besides offering a limited selection of results, these queries can be randomly shuffled as systems accumulate the results of search queries before selecting the ones with multiple appearances in the search results. Randomly shuffled, selected key phrases from a suspicious document can be considered a significant alternation from the initial submission. Such a query will not likely be subject to attack on IP infringement grounds because the complete student submission is not transmitted to the third party and cannot be completely restored from the information transferred for the search.

Compilation of the essential queries should consider the amount of information to be transferred to the third party. The sliding window algorithm that is implemented in the Crot PDS 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 a 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," and "that is the question." Obviously if that 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 the 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 the inequality (1), then we can guarantee it will be impossible to fully recover the initial student submission from the queries that will be sent to the external part of the PDS.

$$y_1 < y/x \qquad (1)$$

Of course, inequality (1) does not guarantee that parts of the document cannot be restored, but such a recovery may not be considered a violation of students' IP rights. The submitted

text has been significantly altered, and its recovery on the third-party side would be considered a deliberate attempt to gain access to the copyrighted material.

At the moment, related research has made no noticeable effort to protect student IP while performing the plagiarism detection process. The architecture we have proposed here is focused on this issue and leaves room for schools to be flexible in IP protection management. When it is necessary to establish a policy on how much of a student submission can go to the external PDS, the school may decide to accept the tradeoff between the granularity of search, that is, the probability of catching small-scale plagiarism (one sentence to a few paragraphs) versus sending copyrighted material to the third party. Inequality (1) can provide a boundary to help school officials make that decision.

The potential flaw of our proposed architecture is that theoretically it will not be able to catch a paper plagiarized from one submitted at a different school if this paper never appeared online and was not accessed by a conventional Internet search engine. The scale of this peer-to-peer copying can be assessed by finding out the actual percentage of such transfers between schools. Scanlon and Neumann (2002) indicated that the Internet is indeed the main source of plagiarized texts. Another study also indicated that about half of the surveyed students knew someone who had plagiarized from the Internet (Jones, Johnson-Yale, Millermaier, & Pérez, 2008). It also indicates that in many cases, the "deep" Web could be a source for plagiarized papers as the majority of students feel that academic papers on library databases are a reliable source of information. Even so, major search engines such as Bing and Google have access to subscription-based databases of research journals, and therefore outsourcing the search can help to locate these possible sources. Based on these two factors, we feel that the scale of peer-to-peer copying is not critical compared to the advantages provided by the proposed architecture for PDS. Moreover, local plagiarism within the school is taken care of by the database of local submissions outlined on Figure 3. Such a database is located on the university infrastructure and will not be subject to copyright claims from the students.

The following factors should be considered to assess the scalability of the proposed approach.

- The architecture increases the requirements of the available computational power and storage capacity of university infrastructure. Additional storage is required to keep digital fingerprints of the submitted documents. If typical fingerprinting algorithms (Schleimer, Wilkerson, & Aiken, 2003) or T9-like algorithms are used to compile fingerprints, then we one can expect an additional 20 to 100% increase in required storage comparing to the space required for the plain texts. However, such an increase is insignificant, as plain text does not take up much space, and memory prices have followed a declining trend for years.

- Additional computational power is required to calculate a single document fingerprint, which is necessary for the fast comparison of documents. Since there are many algorithms available that are linear to the document size, such an increase can be also considered insignificant.

- Additional requirements will arise to perform one-to-many detailed comparisons in the internal PDS. These requirements will impose a high load on the database management system (DBMS), and therefore the cost of the DBMS licensing and maintenance on the university side will be the main factor that will affect the price of maintaining the proposed architecture. In most cases, universities already own and maintain some DBMS, and therefore licensing cost increases may not be significant. Hardware investments along with regular maintenance will have a major impact.

- The scalability of external search capacities could be an issue with the exponential growth of Internet content. But if the search is outsourced to one of the major search engines, the proposed architecture gets the whole power of this engine. Moreover, an OCLC report (2011) indicated that 90% of students begin their search for information using a search engine. Thus the probability that a particular source of plagiarism was indexed by a general purpose search engine is very high. If such a search engine is included in the architecture, the scalability and reliability of the PDS will be sufficient to detect copy-paste types of plagiarism.

- Increased requirements on the university infrastructure will ease heavy requirements for the DBMS on the third-party infrastructure. This factor can decrease the cost of third-party service and reduce the market entrance cost for start-ups, thus promoting competition among PDS.

The following section shows the results of our prototype of the four-component architecture for PDS: Learning Management System – Internal PDS – External PDS – Conventional Search Engine.

Experiment

The experiment was run on a set of papers that simulated different plagiarism levels from the Web. The main goal of the experiment was to check the applicability and practical implications of the proposed approach. The experiment was conducted to answer the following questions: (a) How does the amount of information used in the search affect the search quality? (b) What is the minimal amount of information required for guaranteed detection of different levels of plagiarism?

The experiment was conducted using a set of documents that were tailored to simulate different levels of plagiarism from Wikipedia pages. All 500 documents in the set were roughly the same length, 3,000 words. We used 50 different pages from Wikipedia as sources. The sources were grouped into the following categories: countries, objects from a specialized domain, social concepts and actions, and general objects. Wikipedia text was placed as one consecutive piece in a random place in the original document. Random placement of plagiarized passages was done to ensure the validity of simulated plagiarism since in real assignments, plagiarism can appear anywhere in the text. We used 10 different levels of plagiarism that took varying amounts (from 5 to 50%) of text from Wikipedia. In total, 500 documents were generated according to these standards. A 5% increase in the amount of

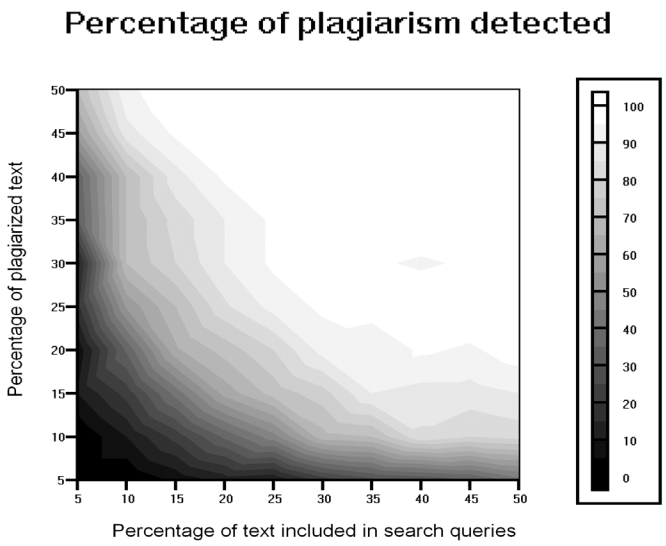
plagiarized text was used to maintain the balance between the granularity of the search results and the time required to run the experiment. The total amount of 500 documents and the coverage of four categories supported experiment result reliability.

The scan was scheduled to randomly select anywhere from 5 up to 50% of all possible phrases made up of six words from each document. For example, for a 5% selection of a 3,000-word document, the total number of queries sent to the search engine was 150, making the maximum possible value of y_1 equal to 900 words. The selection of six words is based on two previous studies that have indicated a six-word running window is an effective search pattern for English language texts (Butakov & Shcherbinin, 2009; Culwin & Child, 2010). The Appendix contains tables showing the detailed information of plagiarism detection results for the different source categories mentioned above.

The minor inconsistency can be seen in the results when a higher percentage of search queries sent to the search engine resulted in actually lower detection level. This inconsistency could have been caused by network timeouts and the unavailability of certain resources on the Web at the time the experiment was conducted.

Figure 4 summarizes the results for all the documents. It displays the surface of the detection reliability on two axes: the percentage of text plagiarized and the percentage of queries sent to the search engine. Bright areas indicate better results. As can be seen from Figure 4, in the experiment's setting, reliable detection could be achieved for low y_1 only if a significant portion of the suspicious text was plagiarized from the Web.

In other words, if only up to 15% of the queries are used from the suspicious document then this approach is applicable to the texts with plagiarism level of 45% and above. Or to rephrase, if more than 50% of the text is plagiarized then only 15% of the queries will be enough to reliably detect the plagiarism. Table 2 in the Appendix indicates that if the topic of the paper is very specific, then reliable detection can be achieved with 15% of the queries if about 25% of text is plagiarized. This can be explained by the fact that queries with more specific keywords draw better results from conventional search engines.



Note. Dark areas represent undetected plagiarism.

Figure 4. Summary of the plagiarism detection results.

The less reliable results for documents with lower levels of plagiarism could have been caused by the fact that these fragments were not covered by the random selection of the search queries. This means that better results for documents with less plagiarism can be achieved if, instead of randomly selecting queries, we can prescan a suspicious document and define segments that must be covered by the search queries. One possible way to do that is to implement an appropriate technique from the authorship detection and stylometry domain (Stamatatos, 2009). This issue will be addressed in subsequent research on the proposed architecture.

Conclusion

In this study, we concentrated on architecture for a PDS. The proposed solution contributes to a number of 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 profit from student work. Limited and scrambled portions of student work that departs from the school's IT infrastructure cannot be used to fully recover the material.

A second distinctive feature is the decision to outsource the most time-consuming part of the plagiarism checkup to a third party, thereby reducing the workload on the university IT infrastructure. Such outsourcing removes the necessity for the PDS in each school to have its own private web crawler and allows different schools to rely on a common search engine for PDS. Such major search engines improve the probability of plagiarism detection because they have very high indexing capacities.

In future research projects, we are planning to work on improving the details of the proposed architecture. One possible direction to take might be to include stylometry in the

external part of the PDS to do preliminary checkups. This feature could lead to better scalability of the service, allowing the external part of the PDS to download suspicious sources of plagiarism with a higher probability rate and filter them before submitting the results to the internal part of the PDS.

References

Abbasi, A., & Chen H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, *26*(2). doi:10.1145/1344411.1344413. 7:2 – 7:29

Arnold, M., Schmucker, M., & Wolthusen, S. D. (2003). *Techniques and applications of digital watermarking and content protection*. Norwood, MA: Artech House.

Bennett, M. G. (2009). A nexus of law & technology: Analysis and postsecondary complications of A.V. et al. v. iParadigms. *LLC Journal of Student Conduct Administration*, *2*(1), 40–45.

Burrows, S., Tahaghoghi, S. M., & Zobel, J. (2007). Efficient plagiarism detection for large code repositories. *Software: Practice and Experience*, *37*(2), 151–175. doi: 10.1002/spe.750

Butakov, S., & Barber, C. (2012). Protecting student intellectual property in plagiarism detection process. *British Journal of Educational Technology, 3*(4), E101–E103. doi:10.1111/j.1467-8535.2012.01290.x

Butakov, S., & 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* (pp. 482–483). Washington, DC: ICALT, IEEE Computer Society.

Cox, I., Miller, M., Bloom, J., & Friedrich, J. (2007). *Digital watermarking and steganography*. San Mateo, CA: Morgan Kaufmann.

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*. Retrieved from http://www.plagiarismadvice.org/

Donaldson, J. L., Lancaster, A.-M., & Sposato, P. H. (1981). A plagiarism detection system. In *Proceedings of the Twelfth SIGCSE Technical Symposium on Computer Science Education* (pp. 21–25). New York, NY: ACM. doi: 10.1145/800037.800955

Goh, E. J. (2003). *Secure indexes* (Report 2003/216). Retrieved from Cryptology ePrint Archive website: http://eprint.iacr.org/2003/216/

Heberling, M. (2002). Maintaining academic integrity in online education. *Online Journal of Distance Learning Administration*, *5*(1). Retrieved from http://www.westga.edu/~distance/ojdla/spring51/heberling51.html

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.

Jocoy, C., & DiBiase, D. (2006). Plagiarism by adult learners online: A case study in detection and remediation. *The International Review of Research in Open and Distance Learning*, *7*(1). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/242/466

Jones, S., Johnson-Yale, C., Millermaier, S., & Pérez, F.S. (2008). Academic work, the Internet and US college students. *The Internet and Higher Education*, *11*(3–4), 165–177.

Krsul, I., & Spafford, E. H. (1997). Authorship analysis: Identifying the author of a program. *Computers & Security*, *16*(3), 233–257.

Marais, E., Minnaar, U., & Argles D. (2006). Plagiarism in e-Learning systems: Identifying and solving the problem for practical assignments. In *Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies (ICALT '06)* (pp. 822–824). Washington, DC: IEEE Computer Society.

OCLC (2011). *Perceptions of libraries and information resources: A report to the OCLC membership*. Dublin, OH: OCLC. Retrieved from http://www.oclc.org/reports/2005perceptions.htm

Scanlon, P. M., & Neumann, D. R. (2002). Internet plagiarism among college students. *Journal of College Student Development*, *43*(3), 374–385.

Schleimer, S., Wilkerson, D., & Aiken, A. (2003). Winnowing: Local algorithms for document fingerprinting. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 76–85). New York, NY: ACM.

Song, D. X., Wagner, D., & Perrig, A. (2000). Practical techniques for searches on encrypted data. In *Proceedings of the 2000 IEEE Symposium on Security and Privacy* (pp. 44–55). doi: 10.1109/SECPRI.2000.848445

Spafford, E.H. (2011). Security, technology, publishing, and ethics (part II). *Computers & Security*, *30*(1), 2–3.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science*, *60*, 538–556. doi: 10.1002/asi.21001

Underwood, J., & Szabo, A. (2003). Academic offences and e-learning: Individual propensities in cheating. *British Journal of Educational Technology*, *34*(4), 467–477.

Urvoy, T., Chauveau, E., Filoche, P., & Thomas, L. (2008). Tracking web spam with HTML style similarities. *ACM Transactions on the Web*, *2*(1)., 3:2-3:28 doi: 10.1145/1326561.1326564

VAIL (Virtual Academic Integrity Laboratory) (2012). Detection tools and methods: Stu-

dent copyright and detection services [Web page]. Retrieved from http://www.umuc.edu/cip/vail/faculty/detection_tools/services.html

Zaka, B. (2009). Empowering plagiarism detection with a web services enabled collaborative network. *Journal of Information Science and Engineering*, *25*(5), 1391–1403.

Appendix: Plagarism Detection Results

Table 1

Plagiarism Detection Results on all Simulated Documents

		Percentage of queries sent to the search engine									
		5	10	15	20	25	30	35	40	45	50
Percentage of plagiarized text	5	0	0	0	6	11	17	22	11	28	17
	10	0	6	11	28	33	56	50	72	67	72
	15	6	17	17	39	56	61	83	78	78	78
	20	6	22	50	61	72	89	89	94	94	94
	25	0	33	56	78	94	94	100	100	100	100
	30	6	61	61	78	94	100	94	89	94	94
	35	22	56	72	78	94	100	100	100	100	100
	40	39	67	78	94	94	100	94	100	100	100
	45	33	72	89	94	94	94	100	94	100	100
	50	50	83	94	100	100	100	100	100	100	100

Note. 500 generated assignments taken from 50 source documents.

Table 2

Plagiarism Detection Results for the Objects from a Specialized Domain Category

		Percentage of queries sent to the search engine									
		5	10	15	20	25	30	35	40	45	50
Percentage of plagiarized text	5	0	0	0	25	0	25	50	50	25	50
	10	0	0	25	75	75	100	100	100	100	100
	15	25	50	50	100	100	100	100	100	100	100
	20	25	25	75	100	100	100	100	100	100	100
	25	0	100	75	100	100	100	100	100	100	100
	30	0	50	100	100	100	100	100	100	100	100
	35	25	75	100	100	100	100	100	100	100	100
	40	0	100	75	100	100	100	100	100	100	100
	45	75	100	100	100	100	100	100	100	100	100
	50	100	100	100	100	100	100	100	100	100	100

Note. 40 generated assignments taken from 4 source documents.

Table 3

Plagiarism Detection Results in the Social Concepts and Actions Category

		Percentage of queries sent to the search engine									
		5	10	15	20	25	30	35	40	45	50
Percentage of plagiarized text	5	0	0	10	14	10	19	24	43	43	62
	10	0	10	29	29	48	67	81	90	81	86
	15	14	19	52	67	76	81	90	90	90	95
	20	0	48	67	67	81	90	86	95	90	100
	25	29	57	71	81	86	95	95	100	100	100
	30	19	71	86	95	95	90	95	95	95	95
	35	38	76	86	95	95	100	100	100	95	100
	40	29	57	95	95	95	100	100	100	100	100
	45	67	95	100	100	100	100	100	100	100	100
	50	76	100	100	100	100	100	100	100	100	100

Note. 210 generated assignments taken from 21 source documents.

Table 4

Plagiarism Detection Results in the General Objects Category

		Percentage of queries sent to the search engine									
		5	10	15	20	25	30	35	40	45	50
Percentage of plagiarized text	5	0	0	14	14	14	43	43	29	29	43
	10	0	29	29	57	71	86	71	86	100	86
	15	0	29	71	86	86	100	100	100	100	100
	20	14	43	86	100	86	100	100	100	100	100
	25	43	86	100	100	100	100	100	100	100	100
	30	14	100	100	100	100	100	100	100	100	100
	35	57	86	86	100	100	100	100	100	100	100
	40	71	100	100	100	100	100	100	100	100	100
	45	86	100	100	100	100	100	100	100	100	100
	50	86	100	100	100	100	100	100	100	100	100

Note. 70 generated assignments taken from 7 source documents.

Athabasca University