The Role of Social Connections in Plagiarism Detection

Aljaž Zrnec(✉) and Dejan Lavbič

Faculty of Computer and Information Science, University of Ljubljana,
Večna pot 113, 1000 Ljubljana, Slovenia
{Aljaz.Zrnec,Dejan.Lavbic}@fri.uni-lj.si

Abstract. Plagiarism is considered as an unethical act. Over the past few years its rate has increased considerably due to a widespread access to electronic documents on the Web. Existing tools for plagiarism detection are not efficient enough and if we want to successfully prevent these kind of acts we must improve today's plagiarism detection approaches. The paper proposes a framework for improved detection of plagiarism, where we focus on integration of information from social networks, information from the Web and semantically enriched visualization of information about authors and plagiates. Visualization enables exploring data and seeking of advanced patterns of plagiarism. We also developed a special tool to support the proposed framework. The results of evaluation confirmed our hypothesis that employment of social network analysis and advanced visualization techniques improves plagiarism detection process.

Keywords: Integration · Social network anaylsis · Plagiarism detection process · Plagiarism visualisation

1 Introduction

Widespread availability of computers, mobile devices and contents on the Web change the approaches to teaching and learning processes. At the same time, we have to improve current plagiarism detection process, to be able to cope with increasing rate of plagiarism in the way of becoming more efficient.

Plagiarism is defined as unethical act of copying someone else's work [23]. Culwin and Lancester defined Four-Stage Plagiarism Detection Process (FSPDP) [6] used to systematically search for plagiarism in a given set of documents. FSPDP consists of four stages: collection, analysis, confirmation and investigation. Usually all stages were performed by human investigator, but with the development of different plagiarism detection methods supported by computers some stages in this process can be fully automated and some can be automated just in parts [13]. The effectiveness of the detection depends on the similarity engine [1, 9]. For a submission to become a plagiate, confirmation stage must be completed, where the submission is examined and verified by a human investigator. This stage can also be automated, but false positive and false negative results may occur. Any similarity from the second stage, which is concerned as positive, is further considered in the investigation stage.

L. Uden et al. (Eds.): LTEC 2015, CCIS 533, pp. 54–63, 2015.
DOI: 10.1007/978-3-319-22629-3_5

Most of the today's approaches [1] for detecting plagiarism are focused on the first two stages, leaving the investigator to perform latter two stages manually. That was a motivation for us to propose an approach where we focus on social aspects of potential plagiarists, by taking into account their social network connections and information on the Web to support investigator's work in the third stage of FSPDP. We believe that plagiarism detection process can be improved by reducing the number of manual examinations of potential plagiarised work. This could be achieved by employment of new visualization techniques that enable semantically enriched view of relationships among potential plagiarists. In this way, the confirmation or rejection of plagiarism can be more efficient.

The paper is organized as follows. In Sect. 2, we present related work and give the proposal for solution. Section 3 describes our Social Plagiarism Detection Framework and supportive software tool. In Sect. 4, we present the evaluation method for assessing our approach and discuss the obtained results. Section 5 concludes the paper discussing the possibilities for future work.

2 Related Work

2.1 Approaches and Tools

With the expansion of various types of plagiarism many different approaches for plagiarism detection have also emerged, especially with the proliferation of digital documents on the Web and the advent of social networks. Several successful studies have been applied to traditional approaches [15, 16, 20] focusing on program code or plain text.

According to [15], there are five different types of plagiarism varying from verbatim copying to advanced types of plagiarism [22] such as copying of ideas and plagiarism of translated text. The increasing use of computers and Web 2.0 tools have mainly positive effect on learning but also increase the possibility of using different types of plagiarism [21]. Early approaches to plagiarism detection heavily relied on methods that were based on string matching while modern methods include document parsing and using a synonym thesaurus. However neither of these methods perform well when faced with complex types of plagiarism [15] such as stealing of ideas or translation of the text.

Several approaches have been implemented in various types of software tools varying from autonomous applications to web services. Applications are run locally and scan for plagiarisms within a given corpus of documents. On the other hand web services allow us to search for plagiarisms among several sources on the Internet.

Software tools for detecting plagiarism are used to detect similarities in program code, text or both. Some of the most commonly used tools for detecting plagiarisms in source code are Sherlock [12, 16], JPlag [17] and Moss [19]. Their basic functionality is very simple. Selected submissions are run through similarity engine, which provides pair-wise results with potential plagiarisms. Modern software for plagiarism detection in source code is based not only on methods for string matching but also include methods for searching lexical and structural modifications in programming code [2, 3, 8, 10, 12]. On the other hand WCopyFind [4], Ephorus [7] and TurnItIn [5, 14, 18] are tools for detecting plagiarisms in free text. They are used to find the amount of text shared between two or more documents on the basis of fingerprinting [11, 15].

Despite many different approaches, researchers in [15] point out that currently available detection systems have numerous drawbacks which can be divided into two categories: (1) issues concerning the user friendliness of today' detection tools (implementation of the system) and (2) issues about limitations of the existing technologies for plagiarism detection.

2.2 Problem Definition and Proposal for Solution

The review of today's approaches pointed out that the research work is focused on the first two initial stages - collection (1st stage) and mainly analysis (2nd stage). Majority of existing approaches conclude the user support by providing pair-wise content similarity of documents and leaving the investigator to perform confirmation and investigation stage manually. In our approach we try to focus on social aspects of potential plagiarists, by taking into account their social network connections, activities and also information from the Web. This provides improved support for plagiarism detection in confirmation (3rd stage) and investigation (4th stage). We argue that our approach facilitates plagiarism detection by providing investigator better support. This results in the reduced number of potential plagiarised work that investigator has to examine manually and provision of new visualisation techniques that enable semantically enriched view of relationships. Therefore the confirmation or rejection of plagiarism can be more efficient. We also provide a tool to support the process that can visualize more corpora with additional information. That enables the investigator to have an overview of author's plagiarism in context of his previous work and work related to his colleagues not only by content similarity.

3 Social Plagarism Detection Framework

3.1 Descrition of the Proposed System

With introduction of Social Plagiarism Detection Framework (SPDF) we focus on latter stages of plagiarism detection process (confirmation and investigation) as depicted in Fig. 1.

Contributions of our approach are as follows: (1) **integration of social network information and information from the Web** that facilitates plagiarism detection process and (2) **advanced semantically enriched visualization** (semantic graph, co-occurrence matrix) of information about authors and documents that enables exploration of data in seeking of advanced patterns of plagiarism.

In confirmation stage, the system evaluates the content similarity report provided in analysis stage and performs additional evaluation of general search engine results and connectivity of authors on social networks. In case of ambiguity the investigator is provided with an option to review the social network analysis results. Based on all given information in the context (content similarity and user connections on social networks) he can confirm or reject pair-wise plagiarism. The main benefit of our approach is improved ranking of potential pair-wise plagiarisms where social information and also

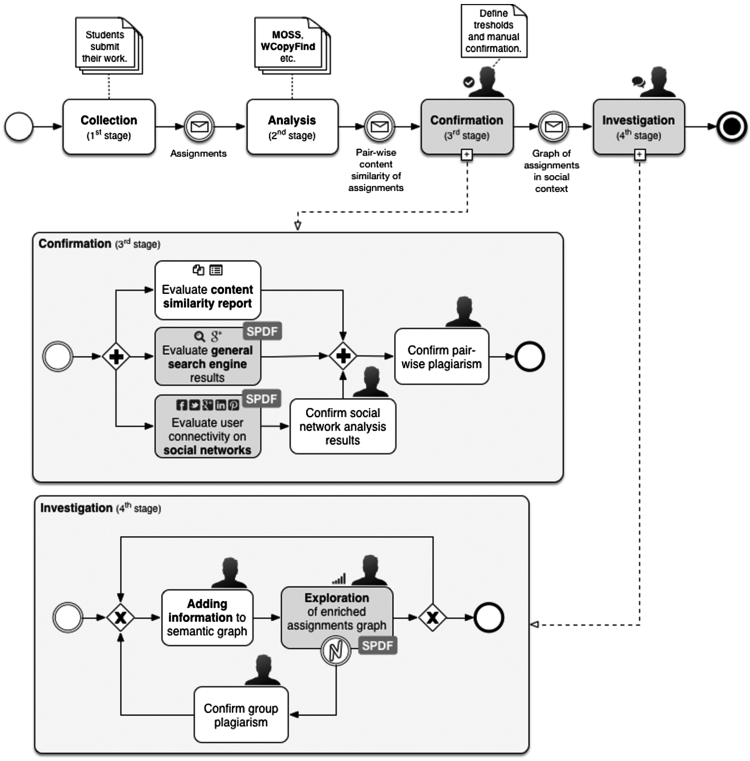

Fig. 1. Social Plagiarism Detection Framework (SPDF).

information from the Web is taken into account and therefore minimizing the effort required by investigator in confirmation stage. We argue and give comprehensive evaluation in Sect. 4 that impact of social information is statistically significant in performing plagiarism detection.

We can define P and D as non-empty sets of people and documents respectively,

$$P = \{p | p \textit{ is a person}\},$$

$$D = \{d | d \textit{ is a document, written by } p \in P\},$$

where W is a set of pairs $\langle p, d \rangle$ with p as an author of a document d

$$W = \{\langle p, d \rangle \, | \exists p \in P \wedge d \in D : p \textit{ is an author of } d\}.$$

We can further define content similarity

$$CS = \left\{ \left\langle d_i, d_j, s_{ij} \right\rangle \, | \exists d_i, d_j \in D \wedge s_{ij} \in (0, 1] : s_{ij} \textit{ is a share of content from } d_i \textit{ in } d_j \right\}$$

as a set of pairs of documents d_i and d_j, with directed content similarity s_{ij} between documents. With integration of social information to plagiarism detection we also introduce

$$SN = SN_{direct} \cup SN_{undirect},$$

which defines a set of social network connections SN between people and consists of directed SN_{direct} and undirected $SN_{undirect}$ social network connections. We can further define

$$SN_{direct} = TW \cup GP \text{ and } SN_{undirect} = FB \cup LN,$$

where TW and GP are sets of Twitter and Google + followers respectivelly and FB and LN are sets of Facebook and LinkedIn connections. Directed social networks are defined as

$$SN_{direct} = \left\{ \left\langle p_i, p_j \right\rangle | p_i, p_j \in P : p_i \, follows \, p_j \right\}$$

and undirected as follows

$$SN_{undirect} = \left\{ \left\langle p_i, p_j \right\rangle | p_i, p_j \in P : p_i \, follows \, p_j \wedge p_j \, follows \, p_i \right\}.$$

When determining pairs $\left\langle p_i, p_j \right\rangle$ of connected people the fuzzy search with employment of Levenshtein distance is performed that requires further actions by investigator in case of ambiguity with multiple account and/or people matching.

We also introduce a set of related items from general search engine SE between person p_i and p_j as

$$SE = \left\{ \left\langle p_i, p_j, n \right\rangle | \exists p_i, p_j \in P, n \in \mathbb{N} : n \, is \, number \, of \, related \, items \right\}.$$

There are multiple search queries performed using the following keywords KW_{ij} between pairs $\left\langle p_i, p_j \right\rangle$ of connected people and are defined as follows:

$$KW_{ij} = KW_{p_i} \cup KW_{p_j} \cup KW_{assignment},$$

where KW_{p_i} and KW_{p_j} are keywords related to person information (e.g. name, surname etc.) and $KW_{assignment}$ is a set of assignment related keywords to narrow down the result set.

When performing plagiarism detection, the goal is to define the set of pairs of documents DP, where plagiarism has been confirmed:

$$DP = \left\{ \left\langle d_i, d_j \right\rangle | \exists d_i, d_j \in D : d_i \, is \, a \, plagiat \, of \, d_j \right\}.$$

When using existing approaches, the investigator, who performs plagiarism detection, tries to find elements of DP, while considering CS and some tacit knowledge TK by investigation. The result of confirmation and investigation stage can be defined as a function $check_{woSocio}$, performed by the investigator

$$check_{woSocio} : CS \times TK \rightarrow DP.$$

When our approach is utilized, the following function $check_{\text{wSocio}}$ is defined

$$check_{wSocio} : CS \times TK \times SN \times SE \rightarrow DP$$

that takes into account social information of document authors. We furthermore argue that employment of $check_{wSocio}$ is more straightforward than $check_{woSocio}$ and enables investigator to perform the confirmation stage more efficiently. This results in total number of documents suspected of plagiarism that investigator has to manually review and confirm or reject plagiarism. For evaluation purposes the supporting tool has been developed to test and compare the aforementioned scenarios.

3.2 Plagiarism Detection Assistant

To support the proposed process, we developed the Plagiarism Detection Assistant (PDA) tool which supports the following functionalities: (1) creating and managing projects, (2) integration of existing plagiarism detection tools, (3) automatic acquisition of social network information and general search engine results, (4) confirming/rejecting assignments, (5) advanced visualization.

The initial action in the process is creating a project by an investigator and collecting the submissions. Then the following steps include preparation of data for confirmation stage: (1) **performing content analysis** by selected existing plagiarism detection tool, where pair-wise content similarity report is retrieved and (2) **acquisition of social network information** and general search engine results for investigated students.

After data is prepared, investigator enters the confirmation stage. The goal of this step is to assign one of the following status to the pair-wise assignments by two people that are considered to perform plagiarism: (1) “not checked” - assignments similarity has not been considered yet, (2) “rejected” - the assignments are not plagiarisms or (3) “confirmed” - the assignments are plagiarisms.

When making the decision the PDA tool assists investigator by providing extensive report of matches found on assignments submitted by different authors. There is a history of all assignments and their corresponding content similarity enriched with social component. The colours used depict the severity of warnings. When investigator reviews all provided information he can make a decision and confirms or rejects plagiarism. By performing these steps the confirmation stage of plagiarism detection is concluded (see Fig. 1) and investigation can start.

One of the views in PDA tool in investigation stage is depicted in Fig. 2 where the support for advanced visualization is provided. Investigator can interactively explore the semantic graph and co-occurrence matrix equipped with information about content similarity, connectivity on social networks and general search engine results. The data from social networks and the Web are collected by means of social network’s public APIs and Web scrapping of publicly available data about authors. Because we do only pairwise analysis of data from the limited set of people, we don’t have any problems with processing. By visualising the context of entire group (e.g. class at University) the investigator can perform plagiarism detection by exploring group plagiarism (Fig. 3).

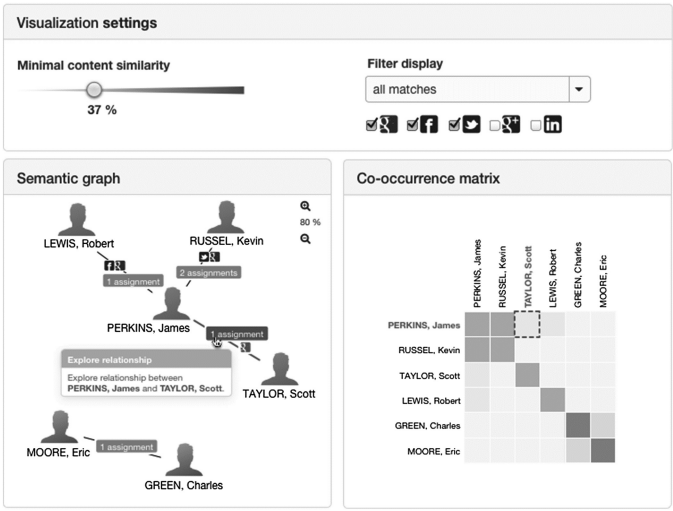

Fig. 2. Support for investigation in PDA tool.

1st person	2nd person	Content similarity	Status	Actions
WRIGHT, Richard	RUSSEL, Kevin	6 %	NOT CHECKED	View match · View result · Confirm match · Reject match
PERKINS, James	TAYLOR, Scott	7 %	REJECTED	View match · View result · Confirm match · Reject match
PARKER, Thomas	MOORE, Eric	8 %	CONFIRMED	View match · View result · Confirm match · Reject match
GREEN, Charles	MOORE, Eric	9 %	REJECTED	View match · View result · Confirm match · Reject match
GREEN, Charles	MILLER, George	9 %	REJECTED	View match · View result · Confirm match · Reject match

Fig. 3. Pair-wise review.

4 Evaluation of the Approach

4.1 Method

Our approach was evaluated on a case study of 76 students taking one of the lectures from Computer Science at undergraduate level. Each student had to submit 5 programming homework assignments during the semester that were later checked for plagiarism. There were 2 experiments performed with 2 groups of evaluators that followed different approach on the same data set (76 student submitted 5 assignments, where 22 assignments were missing so in total 358 assignments). Both groups of evaluators had the common goal - to identify plagiarisms in student work. In the first approach $check_{woSocio}$ evaluators employed MOSS [19] and performed manual investigation on pair-wise content similarity, while in the second approach $check_{wSocio}$ our method with additional social network analysis results was used. The information from social networks employed in the second approach was extracted from public profiles of students. In our

case study 54 students had publicly available information on Facebook and 43 students were active on Twitter. Students were informed about the use of all available public information in the process of plagiarism detection throughout the course.

The method used for evaluation of aforementioned approaches is generalized linear model with logistic regression where link function is defined as follows

$$g(Y) = log_e\left(\frac{n}{1-n}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j.$$

The logistic regression is applied to a situation where the response variable $Y = cheat\_confirmed$ is dichotomous (0, 1). The model assumes that Y follows a binomial distribution and it can be a fit to a linear model $g(Y)$. The conditional mean of Y is the probability $\pi = \mu_Y$ that cheat is confirmed, given a set of X values. The odds that cheat is confirmed are $\frac{n}{1-n}$ and $log\left(\frac{n}{1-n}\right)$ is the log odds or *logit*.

We've defined two models:

$$check_{woSocio}{:}cheat_{confirmed} \sim match\_cs \text{ and}$$

$$check_{wSocio}{:}cheat_{confirmed} \sim match_{cs} + match_{fb} + match_{tw} + se\_hits,$$

where $check_{woSocio}$ is nested model within $check_{wSocio}$ with the same response variable $Y = cheat\_confirmed$ and different predictors $X_{woSocio}$ and X_{wSocio}, where $X_{woSocio} \subseteq X_{wSocio}$. The predictor variables are as follows: $cheat\_confirmed$ is $\{true, false\}$ factor with information about confirmed plagiarism from investigator; $match\_cs$ is content similarity s_{ij} between documents d_i and d_j, where $\langle d_i, d_j, s_{ij} \rangle \in CS$; $match_{fb}$ is a S$\{true, false\}$ factor, based on existence of $\langle p_i, p_j \rangle \in SN_{direct}$; $match_{tw}$ is a $\{true, false\}$ factor, based on existance of $\langle p_i, p_j \rangle \in SN_{undirect}$ and se_hits is a number of search engine results n between persons p_i and p_j, where $\langle p_i, p_j, n \rangle \in SE$.

The employment of models $check_{woSocio}$ and $check_{wSocio}$ is not intended to predict plagiarism but rather for ranking of potential pair-wise plagiats that investigator can review and confirm in the latter stages of plagiarism detection.

In the conclusion of our experiment we performed a follow-through interviews with all of the students where they defended their work and evaluators determined if their submitted work is original. The evaluator's decisions were then used for $cheat_{confirmed}$ response variable to evaluate both models.

4.2 Results and Discussion

When building a model $check_{woSocio}$ the results show that the predictor variable $match\_cs$ is significant ($p \leq 9 \cdot 10^{-6}$) in predicting the response variable $cheat\_confirmed$. Next step was to build another model $check_{wSocio}$ with integrated social information, where the results of this second model show that all predictor variables $match\_cs$ ($p \leq .0025$), $match\_fb$ ($p \leq .0289$), $match\_tw$ ($p \leq .0904$) and se_hits ($p \leq .0432$) are significant in predicting the response variable $cheat\_confirmed$. Now we can compare both models,

which consists of variables that all have significant impact on the prediction by performing ANOVA with Likelihood Ratios Test (LRT) on both models.

The measure used for comparison is deviance as a distance between two probabilistic models (in our case with generalized linear model it equals to two time log ratio of likelihoods between two nested models). Deviance can be regarded as a measure of lack of fit between model and data. We can conclude that the residual deviance in $check_{woSocio}$ (model 1) is significantly higher ($p \leq 6 \cdot 10^{-8}$) than in $check_{wSocio}$ (model 2). We can argue that the model 1 has poorer fit to the data and model 2 performs better.

To confirm that results are meaningful we have performed the test for overdispersion for both models that could lead to distort test standard errors and inaccurate test of significance. We have performed fitting of the model twice – once with binomial family and second with quasibinomial family and the results confirmed that overdispersion is not a problem (the non-central Chi-Squared test was not significant with $p_{woSocio} = .977$ and $p_{wSocio} = .990$). We have also assessed the model adequacy by checking for unusually high values in the hat values, studentized residuals and Cook's D statistics. The results of these tests also confirmed that the models are adequate.

5 Conclusion and Future Work

Plagiarism detection approaches mainly focus on the first two stages of FSPDP. However that is not sufficient to successfully reveal authors that are performing unethical acts relating to plagiarism, because we also have to deal with false positive and false negative results from the analysis stage.

To become more effective in the process of plagiarism detection we proposed approach and PDA software tool, that's able to successfully support the human work in confirmation and investigation stage. In confirmation stage we can efficiently narrow the set of potential plagiarists from previous stages and in investigation stage we can visualize the relationships among potential plagiarists with additional semantic information. The evaluation of two different models, on selected case study, demonstrates that the obtained results are significant. This proves that the inclusion of social network information of document's authors facilitates plagiarism detection process against the approach where these information from the social networks and the Web is not employed in manual decision making process of confirmation and investigation of plagiarism performed by human investigator.

Future work will focus on improving the framework by further analysis of social network connections between suspicious authors. The mutual communication will be analysed by using advanced methods for text analysis. We will also try to find the correlation between authors' mutual messages and plagiarism between their submitted documents.

References

1. Ali, A.M.E.T., Abdulla, H.M.D., Snásel, V.: Overview and comparison of plagiarism detection tools. Paper presented at the DATESO 2011: Annual International Workshop on DAtabases, TExts, Specifications and Objects, Pisek, Czech Republic (2011)

2. Alzahrani, S.M., Salim, N., Abraham, A.: Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Trans. Syst., Man, and Cybernetics **42**(2), 133–149 (2012)
3. Alzahrani, S., Palade, V., Salim, N., Abraham, A.: Using structural information and citation evidence to detect significant plagiarism cases in scientific publications. J. Am. Soc. Inform. Sci. Technol. **63**(2), 286–312 (2012)
4. Balaguer, E.V.: Putting ourselves in SME's shoes: automatic detection of plagiarism by the WCopyFind tool. Paper presented at the SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (2009)
5. Buckley, E., Cowap, L.: An evaluation of the use of Turnitin for electronic submission and marking and as a formative feedback tool from an educator's perspective. Br. J. Edu. Tech. **44**(4), 562–570 (2013)
6. Culwin, F., Lancaster, T.: Visualising intra-corpal plagiarism. Paper presented at the 5th International Conference on Information Visualisation. London (2001)
7. Hanny, D.O., Van Wijk, C.: Plagiarism: Punish or prevent? Some experiences with academic copycatting in the Netherlands. Bus. Commun. Q. **74**(2), 196 (2011)
8. Đurić, Z., Gašević, D.: A source code similarity system for plagiarism detection. Comput. J. **56**(1), 70–86 (2013)
9. Hage, J., Rademaker, P., van Vugt, N.: A comparison of Plagiarism Detection Tools: Department of Information and Computing Sciences. Utrecht University, The Netherlands (2010)
10. Hein, J., Zobrist, R., Konrad, C., Schuepfer, G.: Scientific fraud in 20 falsified anesthesia papers: detection using financial auditing methods. Anaesthesist **61**(6), 543–549 (2012)
11. Introna, L.D., Hayes, N.: On sociomaterial imbrications: what plagiarism detection systems reveal and why it matters. Inf. Organ. **21**(2), 107–122 (2011)
12. Joy, M., Luck, M.: Plagiarism in programming assignments. IEEE Trans. Educ. **42**(2), 129–133 (1999)
13. Makuc, Ž.: Methods to Assist Plagiarism Detection. (Bachelor of Science), University of Ljubljana, Faculty of Computer and Information Science (2013)
14. Marsh, B.: Turnitin.com and the scriptural enterprise of plagiarism detection. Comput. Compos. **21**(4), 427–438 (2004)
15. Mozgovoy, M., Kakkonen, T., Cosma, G.: Automatic student plagiarism detection: future perspectives. J. Educ. Comput. Res. **43**(4), 511–531 (2010)
16. Mozgovoy, M., Fredriksson, K., White, D., Joy, M., Sutinen, E.: Fast plagiarism detection system. In: Consens, M.P., Navarro, G. (eds.) SPIRE 2005. LNCS, vol. 3772, pp. 267–270. Springer, Heidelberg (2005)
17. Prechelt, L., Malpohl, G., Philippsen, M.: Finding plagiarisms among a set of programs with JPlag. J. Univ. Comput. Sci. **8**(11), 1016–1038 (2002)
18. Rolfe, V.: Can Turnitin be used to provide instant formative feedback? Br. J. Educ. Tech. **42**(4), 701–710 (2011)
19. Schleimer, S., Wilkerson, D.S., Aiken, A.: Winnowing: local algorithms for document fingerprinting. Paper presented at the ACM SIGMOD International Conference on Management of data (2003)
20. Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic plagiarism analysis. Lang. Resour. Eval. **45**(1), 63–82 (2011)
21. Underwood, J., Szabo, A.: Academic offences and e-Learning: individual propensities in cheating. Br. J. Educ. Tech. **34**(4), 467–477 (2003)
22. Witherspoon, M., Maldonado, N., Lacey, C.H.: Undergraduates and academic dishonesty. Int J. Bus. Soc. Sci. **3**(1), 76–86 (2012)
23. Youmans, R.J.: Does the adoption of plagiarism-detection software in higher education reduce plagiarism? Stud. High. Educ. **36**(7), 749–761 (2011)