Open Learning, the Issue of Plagiarism - Efficient Algorithm

YAHIA JAZYAH
Department of Information Technologies and Computing
Arab Open University, Kuwait Branch
Ardiya, KUWAIT
yahia@aou.edu.kw

Abstract: - E-learning is a big pace in education and has a significant role in instruction of students in higher education. Plagiarism is one of the most serious challenges for lecturers to detect; several techniques and algorithms are exist that try to solve this problem or mitigate the bad effect of it. In this research, a comprehensive overview about e-learning focusing on plagiarism is introduced, and presenting an efficient algorithm for detecting plagiarism.

Text Matching is the basic of the proposed algorithm; it analyzes texts and searches for matches, not just looking for what in common. When applying the proposed algorithm on two files containing questions and answers. Results show that the algorithm can detect fair and real matching between text files even when the files are including the questions themselves.

Key-Words: - E-Learning, Open Learning, Plagiarism, Algorithm.

1 Introduction

The traditional educational system in universities for a long time has been a classroom with a professor giving speeches to students while students listening and taking notes.

The huge development of ITC Information and Communication Technologies has provided great opportunities for education suppliers to introduce new methods of delivering educational materials. E-learning is one of the new alternative and innovative learning environments compared with traditional learning.

E-learning [1] identifies various types of computer-aided learning; usually using modern technological means such as CD-ROM. it is expanding especially in the sphere of distance education and corporate training.

Blended learning [2] is another development of ITC that encompasses both classroom-based and extracurricular educational activities with the use of complementary technologies of traditional and e-learning. In blended learning, the time allotted to work on e-learning courses can range from 20% to 80%.

Both blended and E-learning are facing problem of plagiarism [3]. Several methods have been designed in order to detect plagiarism; this research develops an efficient algorithm to detect plagiarism that achieves high accuracy.

One of the most serious problems facing open learning is plagiarism. It aims to detect the unauthorized copy of text published on the internet. Improving the detection of plagiarism will enhance and ease the education which reflects on the quality of students as output of the educational process.

Several algorithms have been designed to detect plagiarism which based on matching the text of documents whatever the contents are. Some students rewrite the question inside the answer document which is considered as plagiarism from the view of detecting algorithms; this is a weak point that can raise the similarity value. The proposed algorithm can detect the question and ignore it, in other words, it will not be considered as matching and so not plagiarism.

After testing the algorithm, it shows the ability to detect real plagiarism and ignoring the questions themselves.

The remaining of this paper organized as follows: part 2 presents various definitions of e-learning, part 3 shows the advantages and benefits of e-learning, part 4 presents factors that make e-learning more effective, part 5 presents e-learning issues, part 6 focusing on plagiarism; algorithms that detect it and applications software as well, part 7 presents the proposed algorithm and the results of the proposed algorithm, and finally part 8 is the conclusion.

2 Definition of E-learning

There are several definitions of e-learning summarized below [4]:

- "E-Learning identifies various types of computer-aided learning, usually using modern technological means; such as CD-ROM".

- "E-learning can be understood as an educational process, using information and communication technologies to create training, to distribute learning content, communication between students and teachers and for management of studies".

- E-learning challenges the traditional ways of training and learning, and provides new solutions for problems. For instance, the role of teachers is probably changing from importers of knowledge to expeditors of knowledge. And it can be a very good learning practice that can exceed the education you may experience in a crowded classroom. E-learning contains different types of educational tools in learning and educating. It has the same meaning with technology-enhanced learning (TEL), computer-based instruction (CBI), computer-based training (CBT), computer-assisted instruction or computer-aided instruction (CAI), internet-based training (IBT), web-based training (WBT), online education, virtual education, and virtual learning environments.

- Blended learning is the technique were traditional lessons are mixed with virtual remotely/e-learning lessons. This kind of scenarios opens up several advantages that are not commonly available in traditional lessons. The employment of Web as medium in part of the lessons broadens the knowledge and makes use of state of art technological advancements possible [5].

3 Advantages and Benefits of E-learning [6]

Every student has the luxury of choosing the place and time that suits him/her. And so, it is flexible when issues of time and place are taken into consideration.

Enhancing the efficacy of knowledge and qualifications via ease of access to huge amount of information.

E-learning has discussion forums. They provide opportunities for relations between learners, and help eliminate barriers that have the potential of hindering participation including the fear of talking to other learners.

E-learning is cost effective. There is no need for the students or learners to travel. And it offers opportunities for maximum number of learners with no need for many buildings.

E-Learning allows self-pacing. For instance the asynchronous way permits each student to study at his or her own pace and speed whether slow or quick. It therefore increases satisfaction and decreases stress

4 How to make e-learning more effective:

One of challenges of e-learning and blended learning is how to make it more effective.

[7] summarizes conditions in order to improve E-learning effectively. The Availability of hardware (particularly computers), improving the software, faster internet connectivity/improved bandwidth with lower prices, provision of technical support for e-learning at a range of scales, appropriate content in appropriate languages, improved training for teachers in e-learning at all levels, and awareness raising about the value of e-learning.

5 e-learning issues:

Any institution adopts e-learning faces some important issues [8]:

- Institutions must provide an adequate and reliable technical infrastructure to support e-learning activities;

- Teachers and students must possess the technical skills to use e-learning tools;

- Professors must redesign their courses to incorporate e-learning effectively into their pedagogy.

- Plagiarism is another issue not only in e-learning but also in other types of education.

6 Plagiarism

6.1 Plagiarism Algorithms

Plagiarism is one of the most serious ethical problems in education. It occurs when a writer deliberately uses someone else's ideas or other original material without acknowledging its source or crediting it.

Plagiarism problem can be minimized through the integration of plagiarism checking tools and other checking methods into e-learning systems [9].

Many researches implement algorithms in order to detect plagiarism. [10] implements a web based plagiarism detection system for academic activities.

The algorithms, normally, used in plagiarism detection software are string tiling, Karp-Rabin algorithm, Haeckel's algorithm, k-grams, string matching algorithm [11]. In [12], the authors describe two algorithms that are used to test for efficiency in plagiarism detection.

In [13], the authors propose a system that is based on properties of assignments that course instructors use to judge the similarity of two submissions instead of the popular text-based analyses. This system uses neural network techniques to create a feature-based plagiarism detector and measures the relevance of each feature in the assessment [14].

Two popular methods by Levenshtein and Damerau [15] define edit distances that can be used to compare the similarity of two strings of characters with each other. These distances are used in a variety of applications ranging from DNA analysis to plagiarism detection.

[16] uses Levenshtein distance to compare word n-gram and combine adjacent similar grams into sections. In another approach [15], the Levenshtein distance and simplified Smith-Waterman algorithm are merged as a single algorithm for the identification and quantification of local similarities in plagiarism detection. In [17] the researchers used the LCS distance combined with other POS syntactical features to identify similar strings locally and rank documents globally.

[18] proposes an approach that is based on eliminating correct references from scientific papers to make them as plagiarized passages, it can correctly simulate real cases of text re-use.

[19] presents an efficient plagiarism detection tool, CPLAG, for C programming language codes. The tool assesses the structure of the C programs based on a set of attributes and performs a binary encoding of the C code statements. Subsequently, it utilizes computationally inexpensive bitwise operations to detect similarity between the given C programs. The design of CPLAG considers the commonly used techniques to avoid detection of plagiarism for delivering an efficient performance. Moreover, it avoids the extensive computations as used by existing tools for plagiarism detection

[20] proposes to use word2vec model to detect the semantic similarity between words in Arabic language which can help in detecting plagiarism. Word2vec is a deep learning technique that is used to represent words as features of vectors with high precision. It uses OSAC corpus for training word2vec model.

[21] proposes a plagiarism detection algorithm based on approximate string matching to be specified in "copy and paste"-type plagiarisms, and a speed improvement to an implementation of the algorithm. Most of the computations required in the algorithm are omitted by two kinds of approximations of the output used for plagiarism detection, while the decrease of accuracy caused by the approximations is acceptable.

[22] proposes approach based on four well-known models namely Bag of Words (BOW), Latent Semantic Analysis (LSA), Stylometry and Support Vector Machines (SVM). The proposed approach works by capturing usage patterns of the most common words (MCW) from books of 25 authors. Stylistic features for each author were harnessed in the method by adjusting the LSA weighting technique. The adjusted LSA method was trained in a novel manner using the leave-one-out-cross-validation technique and compared with the traditional LSA method

[23] proposes a multi-features fusion method based on Logical Regression model for the high-obfuscation plagiarism seeds identification. This method uses Logical Regression model to combine lexicon features, syntax features, semantics features and structure features extracted from suspicious text fragments pairs.

[24] presents a different approach for measuring semantic similarity between words and their meanings; it suggests new strategies for detecting the plagiarism in the user document using the semantic web.

6.2 Plagiarism Software

Software and websites (paid-for and free) are available for detecting plagiarism. The most popular software are Turnitin and iThenticate. Turnitin was designed by John Barrie and a group of his UC Berkeley colleagues. This software uses digital technology to conduct a meta-search of the internet to locate sources from where the document might have been plagiarized. Teachers receive an Originality report that cites the degree of originality and links to Internet webpages that help determine what Web resources students have tapped [25, 26].

All techniques target to detect similarities among several files and ignore the questions themselves (in some cases answer document includes the questions), which is considered as matching, and so a plagiarism. None of the previous techniques consider this issue. The proposed algorithm detects the questions and ignores them off the calculation of similarity value.

7 Proposed algorithm and analysis

TurnItIn doesn't differentiate between the introduction, Questions and student answers. These three points (Introduction, Questions and Answer Keys) are expected to be repeated in all students answer files. As a result the similarity value driven by Turnitin is high, which is false value.

The proposed algorithm solves the previous mentioned problem in order to provide accurate and fair similarity value based on student's answer (ignoring the introduction and question); see fig. 1.

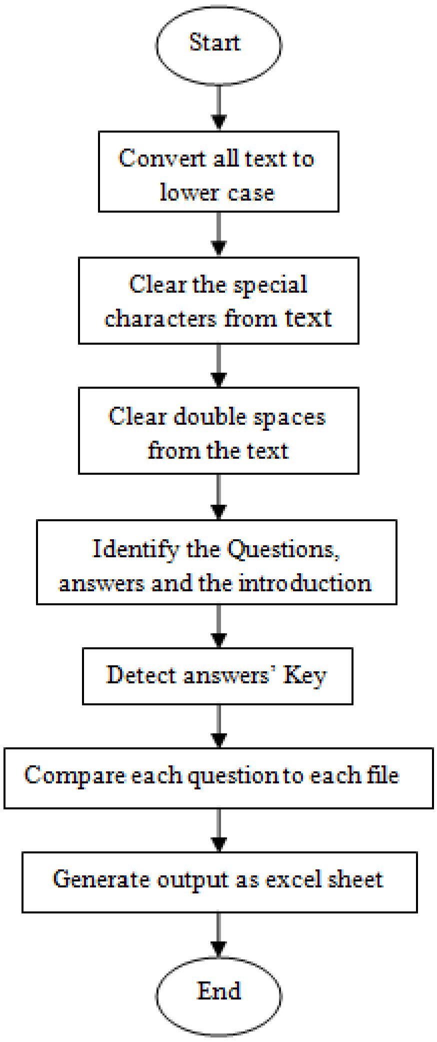
Start
Convert all text to lower case
Clear the special characters from text
Clear double spaces from the text
Identify the Questions, answers and the introduction
Detect answers' Key
Compare each question to each file
Generate output as excel sheet
End

Fig. 1 Flowchart of the proposed algorithm.

The followings are the basic steps of algorithm:

- prepare the text file ready for comparison:-
    - Convert all text to lower case.
    - Clear the special characters from text (i.e. ^, $, \n, etc).
    - Clear double spaces from the text.
- Identify the Questions and their answers along with the introduction.
    - After Identifying the Answers, the system starts detecting answers' Keys along with the Common words used in English such as (“the”, ”us”,” from”…etc.).
- Start comparison stage (where each question from each file has to be compared with its peers in all other files).

The algorithm is implemented using Microsoft Visual C#. The output is a Microsoft Excel file which includes:

- File Names: file names which the system does the comparisons to them.
- Distinct Words: the number of distinct words of the tested file.
- Common: the number of the shared words between the tested file and the other inspected files.
- Match: the number of matches between the tested file and the other inspected file.
- Match percentage: the percentage of words that are shared between the inspected files and tested file.

In order to test the system, 50 scenarios are designed; each one consists of three word document files (A, B, and C) all of them have an intro and two questions along with their answers. Two of them have identical answers (A and B) and the third one (B) has one different Answerr. The third file (B) has an answer with one word (in order to test the answer when adding this word as answer key). There are three tests to validate the system:

- Regular test: testing if the system will detect the identical files (A and B).

- Testing files when removing the answer key: because of the third file (B) has one answer of one word and the other answer is identical, the matching analysis result is 50%.

- Question pattern test: testing questions that are different (answer key is not existing) and the result is mismatch.

Several different tests are conducted based on different input pattern of text; the system detects matching all the time.

7.1 System's Features:

- The system can search the input folder that contains word document only.

- The system can make reports for every file in the input folder.

- The system has two types of reports.

    - Single file: investigating who is the most matching and who is the least matching. It is detecting group of files and the user has the option to exclude some files from the comparison.

    - Final report: which includes the basic result (a number expressing the un-originality of the student's answer)

8 Conclusion

Plagiarism is one of the most serious problems facing the educational process specially e-learning and blended learning. In this research, an efficient method is presented to detect the originality of student's work based on several steps that can solve the problem of high similarity value between two files when some texts, which are not related to the answer, are existing. The system can detect such problem efficiently when many scenarios are applied.

Acknowledgment

This research was supported and funded by the research sector - Arab Open University - Kuwait Branch under decision number 18100. We thank everybody who assisted us to improve the study.

Ethics

We testify that this research paper submitted to the International Journal of Computers - International Association of Research and Science, title: Open Learning, the Issue of Plagiarism - Efficient Algorithm has not been published in whole or in part elsewhere.

This research project was conducted with full compliance of research ethics norms of Arab Open University - Kuwait.

References:

[1] Safiyeh Rajaee Harandi, "Effects of e-learning on students' motivation", *3rd International Conference on Leadership, Technology and Innovation Management*, volume 181, 2015, pages: 423 – 430

[2] Olga V. Yanuschika, Elena G. Pakhomovaa, Khongorzul Batbold, "E-learning as a Way to Improve the Quality of Educational for International Students", *International Conference for International Education and Cross-cultural Communication. Problems and Solutions (IECC-2015),* 09-11 June 2015, Tomsk, Russia. Pages: 147 – 155

[3] Emil Marais, Ursula Minnaar, David Argles, "Plagiarism in e-learning systems: Identifying and solving the problem for practical assignments", *Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)*, 5-7 July 2006, Kerkrade, Netherlands, Pages: 822 – 824, DOI: 10.1109/ICALT.2006.1652567

[4] Safiyeh Rajaee Harandi, "Effects of e-learning on students' motivation", *3rd International Conference on Leadership, Technology and Innovation Management*, volume 181, 2015, pages: 423 – 430

[5] Frederico M. Schaf, Carlos E. Pereira, Renato V. B. Henriques, "Blended Learning using GCAR-EAD Environment: Experiences and Application Results", *Proceedings of the 17th World Congress The International Federation of Automatic Control*, Seoul, Korea, July 6-11, 2008

[6] Valentina Arkorful and Nelly Abaidoo, "The role of e-learning, the advantages and disadvantages of its adoption in Higher Education", *International Journal of Education and Research*, Vol. 2 No. 12 December 2014

[7] Olojo Oludare Jethro, Adewumi Moradeke Grace, Ajisola Kolawole Thomas, "E-Learning and Its Effects on Teaching and Learning in a

Global Age", *International Journal of Academic Research in Business and Social Sciences,* Vol. 2, No. 1, January 2012, ISSN: 2222-6990

[8] Andreea-Maria Tîrziua, Cătălin Vrabie, "Education 2.0: E-Learning Methods", *5th World Conference on Learning, Teaching and Educational Leadership, WCLTA 2014*, pages: 376 – 380.

[9] Emil Marais, Ursula Minnaar, David Argles, "Plagiarism in e-learning systems: Identifying and solving the problem for practical assignments", *Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06),* 5-7 July 2006, Kerkrade, Netherlands, Pages: 822 – 824, DOI: 10.1109/ICALT.2006.1652567

[10] Dinesh Kumar Saini, Lakshmi Sunil Prakash, "Plagiarism Detection in Web based Learning Management Systems and Intellectual Property Rights in the Academic Environment", *International Journal of Computer Applications* (0975 – 8887) Volume 57– No.14, November 2012. Pages: 6 – 11

[11] P. Clough., "Plagiarism in natural and programming languages: an overview of current tools and technologies", June2000.

[12] B.-R. A. Z. Su, K.-Y. Eom, M.-K. Kang, J.-P. Kim,and M.-K. Kim,, "Plagiarism detection using the levenshtein distance and smith-waterman algorithm," in ICICIC '08 *Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control*. , Washington, DC, USA, 2008, p. 569.

[13] V. L. S. Engels, and M. Craig. , "Plagiarism detection using feature-based neural networks.," in *Proceedings of the Thirty-Eighth SIGCSE Technical Symposium on Computer Science Education*, Covington, Kentucky, March 2007, pp. 34-38.

[14] R. M. Federica Mandreoli, Paolo Tiberio, "A document comparison scheme for secure duplicate detection", *International Journal of Digital Libraries*- Springer-Verlag 2004, Volume 4, Issue 3, November 2004, pp 223–244

[15] Z. Su, B. R. Ahn, K.Y. Eom, M. K. Kang, J. P. Kim, and M. K. Kim, "Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm", *3rd IEEE International Conference on Innovative Computing Information And Control*, 2008. ICICIC '08. 18-20 June 2008, Dalian, Liaoning, China, DOI: 10.1109/ICICIC.2008.422

[16] V. Scherbinin and S. Butakov, "Using Microsoft SQL server platform for plagiarism detection", Stein, Rosso, Stamatatos, Koppel, Agirre (Eds.): PAN'09, pp. 36-37, 2009.

[17] C. Barnbaum, "Plagiarism: A Student's Guide to Recognizing It and Avoiding It", 2002. Available at http://www.cpalms.org/Public/PreviewResourceUrl/Preview/25915

[18] Mohtaj S, Asghari H, Zarrabi V. "Compiling a text re-use detection corpus from scientific papers with semi-real cases of plagiarism". *International Conference In Asian Language Processing (IALP),* 2017 Dec 5 (pp. 227-230). IEEE.

[19] Jain S, Kaur P, Goyal M, Dhanalekshmi G. "CPLAG: Efficient plagiarism detection using bitwise operations". *Tenth International Conference on Contemporary Computing (IC3),* 2017 Aug 10 (pp. 1-5). IEEE.

[20] Suleiman D, Awajan A, Al-Madi N. "Deep Learning Based Technique for Plagiarism Detection in Arabic Texts". *International Conference on New Trends in Computing Sciences (ICTCS)* 2017 Oct 1 (pp. 216-222). IEEE.

[21] Baba K. "Fast plagiarism detection based on simple document similarity". *Twelfth International Conference on Digital Information Management (ICDIM)*, 2017 Sep 12 (pp. 54-58). IEEE.

[22] AlSallal M, Iqbal R, Amin S, James A, Palade V. "An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection". *9th International Conference on Developments in eSystems Engineering (DeSE),* 2016 Aug 31 (pp. 203-208). IEEE.

[23] Kong L, Lu Z, Qi H, Han Z. "High obfuscation plagiarism detection using multi-feature fusion based on Logical Regression model". *4th International Conference on Computer Science and Network Technology (ICCSNT),* 2015 Dec 19 (Vol. 1, pp. 355-359). IEEE.

[24] Agarwal J, Goudar RH, Kumar P, Sharma N, Parshav V, Sharma R, Srivastava A, Rao S. "Intelligent plagiarism detection mechanism using semantic technology: A different approach". *International Conference on Advances in Computing, Communications and Informatics (ICACCI),* 2013 Aug 22 (pp. 779-783). IEEE.

[25] Gallant, T. B., & Drinan, P., "Organizational theory and student cheating:Explanation, responses, and strategies." *Journal of Higher*

Education, volume 77, issue ,5 pages: 839. 2006

[26] Hart, M., & Friesner, T., "Plagiarism and poor academic practice - A threat to the extension of e-learning in higher education?" *Electronic Journal of e-Learning*, volume 2, issue1, 2004.