I J C T A, 10(8), 2017, pp. 585-588 ISSN: 0974-5572

1-hop Greedy Cite Order Plagiarism Detection

R. Siva*, G.S. Mahalakshmi and S. Sendhilkumar*****

ABSTRACT

Quite often, researchers do not find time to narrate the step-by-step experiences of their research, and at the last hour, ends up writing the research article / thesis by leaning over the shoulders of already published articles. The rephrasing approach they use makes the text plagiarism detection tougher. Citation pattern analysis would be helpful to identify whether the portion of article is a 'modified copy' of the already published research article. The problem grows magnificent as the researcher attempts to copy the survey segments of various survey papers. This paper proposes 1-hop cite order plagiarism detection assuming that the problem taken is a research offence.

Keywords: Citation Tiling, Greedy, Citation Plagiarism, Research Article, Citation Pattern

1. INTRODUCTION

The way research articles are being written is an interesting phenomenon. A researcher attempts to mature his/her own idea, brings it to a form of an acceptable research issue and proceeds for implementation. However, due to the very nature of research, success is not reaped at the first attempt and before the researcher proceeds for completion of implementation, the time to get the idea published is fast approaching. There may be various valid reasons for the above issue; however, the hard truth has to be accepted that the researcher does not finds time to narrate the step-by-step experiences of their research, and therefore, ends up writing the research article / thesis in no time. This impacts the quality of the research article/thesis.

Towards the climax, researchers tend to write their idea on their own and leans over others' shoulders, when it comes to writing literature surveys. The survey part (or Related Work, technically to name so) is very much essential to substantiate the motivation behind the idea conveyed in the research article. In spite of the researcher getting inspired by a valid publication, when it is about doing 'Black-and-White', many researchers do not find enough time and patience, and therefore, forget or least bothers to recollect the paper that inspired them, and ends up searching for a relevant survey paper in the similar area or addressing similar research issue.

There are other valid reasons for this act of researchers. A paper is not valued for the idea it conveys. More than the semantic clarity, on these days, the outer cover of research article, like publisher, journal, author and sometimes, the newly coined words, attract the researcher's attention. Therefore, such so-called-popular & recent articles are again downloaded by the researcher for further use.

At this stage, the newly downloaded article is only used for filling the 'Related work' part, and the researcher ends up re-phrasing the contents of the copied area, thus retaining the citation pattern. Therefore, citation pattern analysis would be useful here to identify whether the survey is a 'modified copy' of the existing research article. Sometimes, the researcher does not include the recently downloaded article in the

* Department of Computer Science & Engineering, K.C.G. College of Engineering, Chennai, Tamilnadu, INDIA, *Email: sivavb6@yahoo.co.in*

** Department of Computer Science & Engineering, Anna University, Chennai, Tamilnadu, INDIA, *Email: gsmaha@annauniv.edu*

*** Department of Information Science & Technology, Anna University, Chennai, Tamilnadu, INDIA, *Email: ssk_pdy@yahoo.co.in*

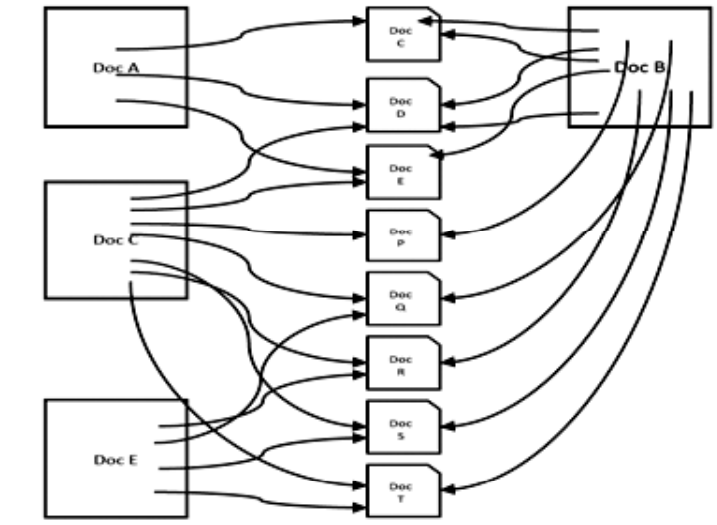
References section, and only retains the references pertaining to the copied segments. The problem grows magnificent as the researcher attempts to copy the survey segments of various survey papers.

This paper addresses the issue by utilising1-hop cite order plagiarism detection since the act of copying the citation pattern inline results in increased chances of the communicated paper's acceptance. This bypassing approach shall be superficially analyzed using Bibliographic Coupling (BC) [1], however, coincidences of BC match could also occur, given the fact that, there are only few articles published remarkably in the respective research area. Therefore, we emphasis, analyzing cite order similarities extending to 1-hop down the reference levels.

2. PROPOSED WORK: 1-HOP GREEDY CITATION TILING

Greedy Citation Tiling (GCT) [1] is an adaptation of a text string similarity function Greedy String Tiling (GST) [2]. GST aims to identify all matching sub-strings with individually longest possible size in two sequences. Corresponding individually longest matches in both sequences are permanently linked with each other and stored as a so called tile. Bibliographic Coupling blindly finds the intersection of 'References' of both the papers under investigation. The absolute or relative number [1] indicates the level of plagiarism. However, the semantic content might have been rephrased indicating the originality of author expression. Therefore, above BC, another approach measuring citation pattern similarity has to be examined.

Greedy citation tiling attempts to match the maximum number of repeated patterns. As an extension to greedy citation tiling, we propose one-hop or one-level down the matched references of Greedy citation



Doc A	C	D	Ins.	E	Ins.	Ins.	Ins.	Ins.	Ins.	Ins.
Doc B	C	P	Q	D	C	E	R	S	T	D
Doc C	Ins.	Ins.	P	Q	Ins.	P	Q	Ins.	Ins.	Ins.
Doc E	R	Ins.	R	S	T	Ins.	Ins.	Ins.	Ins.	Ins.

Figure 1: 1-hop Greedy Citation Pattern Analysis

tiling results. We tend to consider the references of references which were nominated in citation tiling. In figure 1, P and Q are the references of Doc C, and, R, S and T are the references of Doc E. Greedy citation Tiling attempts to discover the matching citation patterns with Doc C, Doc D, and Doc E.1-hop approach extends the idea over the references of Doc C, and Doc E. Figure 1 depicts the problem proposed in the paper.

The reason is that researchers tend to go one level down the references (we mean, references of references) to fetch more research articles. This is common in 'happening research areas' like wireless sensor networks where research articles in three or four levels down the references tend to fall on the same time period, thus preventing timeline based reference paper filtration. Such voluminous and multi-level greedy citation tiling shall be addressed by using optimization approaches [3], which form our future work.

3. RESULTS

The dataset used for this approach is Journal of Informetrics [4]. There are 663 research articles, across 8 volumes with 4 issues per volume. We analysed all research articles with respect to their references confining only to 'Related Work' sections, for 1-hop Greedy citation tiling analysis. We applied regular expression to extract the titles of articles in references, used string matching approaches to find BC couples and then went on with pattern based approaches for 1-hop analysis inside the text. The results obtained are quite interesting.

374 research articles were found to possess matching citation tiles in 1-hop approach when compared to 211 research articles in Greedy Citation Tiling approach. As an evaluation, we have also analysed the semantic similarity of citation segments labelled by 1-hop approach. Interestingly, the cosine similarity used for semantic analysis produced less detection when compared to the seed publication. This is because the 1-hop references were hardly matching semantically with that of the seed paper prone for citation plagiarism detection.

Table 1
Comparison of 1-hop Greedy Citation Tiling & Evaluation

# of articles in the dataset	# of articles Matched (Greedy Citation Tiling)	# of articles Matched (1-hop Greedy Citation Tiling)	# of articles Matched (Evaluation of citation contexts via cosine similarity)
663	211	374	156

4. CONCLUSION

This paper discussed the interesting aspect of applying greedy algorithms for finding 1-hop citation tiling.We have used cosine similarity approaches to evaluate the performance of the 1-hop approach. However, the problem is getting complex, when multi-hop references are considered. Therefore, including semantic similarity approaches before proceeding to subsequent levels would fetch convincing results.

REFERENCES

[1] Gipp, Bela, and Norman Meuschke. "Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence." Proceedings of the 11th ACM symposium on Document engineering. ACM, 2011.

[2] Ahtiainen, A., Surakka, S., and Rahikainen, M.,Plaggie: GNU-licensed source code plagiarism detection engine for Java exercises. In Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006 (New York, NY, USA, 2006), Baltic Sea '06, ACM, pp. 141–142.

[3] Norman Meuschke and BelaGipp. Reducing computational effort for plagiarism detection by using citation characteristics to limit retrieval space. Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on.IEEE, 2014.

[4] G.S. Mahalakshmi, G. Muthuselvi and S. Sendhilkumar, A Bibliometric analysis of Journal of Informetrics – A decade Study, in proceedings of Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM'17), organized by Department of Computer Science and Engineering, University College of Engineering Tindivanam, 2017 (accepted)