Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 65 (2015) 911 – 919

Procedia Computer Science

International Conference on Communication, Management and Information Technology (ICC 2015)

Text mining and similarity search using extended tri-gram algorithm in the reference based local repository dataset

Ranjeet Kumar*, R.C.Tripathi

Indian Institute of Informtation Technology, Deoghat Jhalwa, Allahabad, India

Abstract

In the emerging technological scenario world is becoming a digital hub where data is easily accessible on the internet. Whe take into consideration the academic and research & development fields, the digital resources become more important. Acad research and their publications in current environment can be easily accessed through internet. Easy availability of such res work attracts academic literature dishonesty or plagiarism. Many of the research papers published in the several Confer proceedings and Journals may have some percentage of plagiarized contents. At the same time, the author(s) may cite irrel references in the research paper. In the present research paper, a reference based extended trigram approach has been report check the textual plagiarism of the text written in the research papers. References form the pivotal part of any research pap dissertation which defines the area of research and the state of the art of the research based on which originality o contribution is adjudged. The present article also discusses the behavior of the referencing in three major research categorie Research papers, Master Dissertations, and the Doctoral Dissertations.

Peer-review under responsibility of Universal Society for Applied Research

Keywords: Reference Search; Trigram Technique; Local Repository Search; Textual Similarity Search; Textual Plagiarism

1. Introduction

In the research & development (R&D) field, the publications are the most valuable aspects to recognize the aut

* Corresponding author. Tel.: +91-5322922161; fax:91-532-2922125.
E-mail address: ranjeet@iiita.ac.in

Peer-review under responsibility of Universal Society for Applied Research
doi:10.1016/j.procs.2015.09.062

through their works. The work presented in the research publications represents their academic or res achievements. Accordingly they are acknowledged for their quality work. These researchers when author quality research publication, they keep eye on every contemporary development and maintain the originality quality of their work against this background. There are so many factors to define quality in a research public one of them is reference of that area presently. It shows that the present work is consistent and connected t research work done presently. Another matter of concern for the research publications is how the published w thoughtful and useful for real life applications and extensive. The focus of these fundamental issues has addressed earlier to derive methodological concepts related to them. One of the most thoughtful works report Martin[1] distinguishes between research quality, importance, and impact. By thinking of citations as meas "impact", criticisms revolve around the notion of "unrecognized" or "innate" quality of a work. The quality o research work is recognized to be high if the citation of the research work is followed several times. In the cu publication scenario, majority of the research works are first published in electronic form through institut repositories, digital libraries, publisher's websites, author's WebPages, etc and then in the print media form. the work becomes available online all the time and therefore the citations can be tracked much easily compar old system of getting the work only through print media.

The importance of the citation of the research work can be thought of as relevancy of the research con The research contents are relevant with the cited work or not, are a critical issue of the current research p Number of cases can be found wherein a five pages research article had more than 30 reference citations. S cases can also be found wherein a 14-16 pages research article may have more than 90 reference citations. In of these cases, only 30%-35% references can be justified based on the contents of the research article. Striki most relevant references may be found missing in such research articles.

Regardless of subject examined, it is found that authors cite only a fraction of the real references w particularly influence them reported by [Mac Roberts et. al] [2, 3,4,5,6]. While it is undoubtedly impossible to dete influence on an article, a large percentage can be detected. Thus, instead of one part of influences there are those that are cited and those that are not cited. By [Kostoff et al] [7, 8], in some typical cases of the current work is found that there are only a few authors who have studied scientific articles to determine whether influence cited or not. Szava-Kovats[9] studied physics articles and found that there is an "over-abundance of rel literature". Szava-Kovats[10] also found that the vast majority of influences on articles are not cited.

This research paper is organized first to outline the concerned literature review in the section 2 there the proposed methodology in section 3. In the subsection 3.1, extraction of the references and in subsection references categorization has been discussed. This is followed by the test cases, results, discussions and concl in the section 4 and section 5 respectively.

In the present paper, reference section is the focus part of our research for finding the textual similar determine if the research paper copied the ample amount of text and cited this as a reference, or if copied the a amount of text but did not provide its reference. The test cases are presented for three main categories of te contents of the academia i.e. (a) Research Articles/papers, (b) Master Dissertations, and (C) Doctoral Disserta The test cases reveal distinct patterns of referencing for each of these copyrightable works. The extended tri technique has been used and found useful for the similarity search and plagiarism detection approach in these w

2. The literature review

Right since the early days, research papers confined to one typical investigation in a particular domain maintained the continuity of the previous work in various contexts for which the references played an impo role. There are so many research works reported which focus on citation issues of the research. In this regar two core principles reported are i) Bibliographic Coupling[11], where two documents are said to be coupled if share one or more references, and ii) Co-citation Analysis[12], where similarity between documents A and

measured by the number of documents that cite later on both A and B. Many research papers were reported in re to citation indexing, citation clustering and co-citations coupling amongst the works. In the year of 1974 H Small[13], published two issues of Science Studies which related co-citation and Bibliographic Coupling. In the ar "Science citation index"[14], a new dimension in indexing was advocated based on analysis of citation links an scholarly articles. [R. Mercer et al] [15] and [H. Nanba et al] [16] reported, the relationship of the two research docun have been analyzed by the citation text and citation sentences around them. They extracted the text in ord determine the relationship between the two research articles connected by that citation, called the citation func In the context based citation information processing [A. Elkiss et al] [17] and [R Kumar et al] [18], have proposed s useful analysis. They provided a quantitative analysis of the benefits of citation contexts with regards to c applications such as summarization and information retrieval. In their proposed methodology, they exam relationship between abstract and citation contexts of given research articles and analyzed that citation context have given some extra focus which may not be present in an abstract. In such a case, they suggested that cit contexts can be utilized as a different kind of supplementary summary to the traditional abstract.

The researchers in this regard therefore have developed some important applications in recent years to overcom existing problems of automatic extraction of the context from the citations given in the research article. The re developments by D Bergmark[19] and [B Powlev et al] [20] propose that a collection of different sources of int evidences about entities from documents may be parsed to arrive at the reference list in the order of author n year and integrate it for further processing. Another important development in this regard is to identify bibliogr items and retrieve citation contexts from a plain text file as introduced by [I G Councill et al] [21] and [R Kumar [22]. They introduced "ParCit"; a system that depends on a machine learning method coupled with a heu processing framework. The system models to identify the bibliography items and match them with the text fea to find out the relevant citation contexts. They use tokenization process for parsing of the reference list that is b on several metadata fields such as author, title etc. For every reference item, one or more regular expression produced in order to match the citation contexts in the body of the text. These expressions can handle ex citation styles, such as square bracket or parenthetical markers, and implicit citation styles which use the au names and year of publication.

3. The proposed methodology

In the prior art of the present proposed methodology, the plagiarism related to citation problems were defin numerous manners and each one has been used gainfully in different research contexts. The citation inde bibliographic coupling, citation based key phrase extractions and many more applications were developed in regard. The present paper is focusing on major issue of plagiarism which is based on references given in research article. In a query research article, there are number of references listed in its "references" sec Sometimes it is about 100 in a 10 pages research article. All references given in any research article may be ma related to that particular article or may have some references which are not related to the current research artic the major concern of the present paper. The present research paper proposes the methodology for checkin plagiarism from amongst the used references in any research article. For this, one finds out the *under drawn* as as the *over drawn* references as the first major problem addressed in this context. In the present methodology, we have categorized the citations by the author in three different categories. These are termed as a) Under Citation b) Normal Used Citation and c) Over Drawn Citation. The figure 1 given below is the flow diagra working of the proposed methodology.

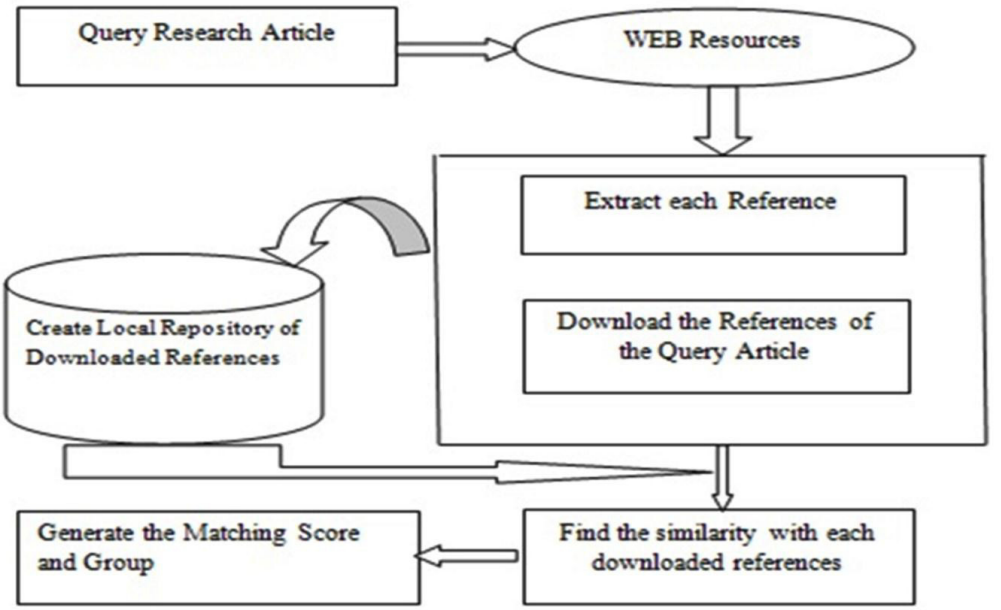


Figure 1. Shows the system architecture for working of proposed methodology

3.1. Extraction of the references from the research article

The extraction of the references from a query research article is the first and foremost step in the current worl the given references in the research article are extracted from the article and processed for their searching o internet so as to create the repository for the query research document. To extract the citations from query res article, "Parscit" citation extractor has been used. The extraction phase comprises the process of identifying extracting reference strings in the bibliography or reference section of a given document and parses them into logical components such as title, author(s), publisher, year of publication, page to page. The output of the extr references is stored in an xml file which also contains some relevant sections of information contents for the document. These are used for further processing. Complete working is described as below:-

(a) Parsed: The parsed section contains header information of the query article like – title(whether of research or a book if any), author(s), Institute(s), year of publication

b) Parscit: This section contains citations present in the query paper. Each citation is present in the <citation and contains title, author, year of publication, etc.

Thus extracted title is matched to the citation title found above and a score is generated. If T1 and T2 are title L1 and L2 are their respective lengths, here LCS stands for Longest Common Subsequence, and then their mat score is:

$$Score_A = \frac{((\text{Length of LCS } (T1,T2))^2}{(L1 * L2)}$$

Scores, in similar fashion are computed for $Score_B$(based on, Authors), $Score_C$(based on Year) and $Score_D$(b on place of publication). Then the cumulative value of score is computed for each of referred document

Score = $Score_A + Score_b + Score_C + Score_D$

Whichever pdf document has the highest score is our starting record in the pdf 's repository and then all those files who have significant score (i.e. > 0.05) are stored to form the local shortlisted repository for fu processing.

3.2. References categorization

In the present methodology, we target to identify those research papers which are i) under drawn, ii) normal iii) over drawn. For this purpose, after downloading of all these research papers referenced in the query rese paper, as described in previous section, we have found out their ranked list for content similarity with that o query research article. We then categorize them in three categories i) Under used citation, ii) Normal used cita and iii) Over drawn citations. For this, the steps of processing are described as below

(a) To extract keywords from a pdf, we first extract text from it using Apache PDFBox library for extraction and then fetch out the keywords using Alchemy API keyword extraction service. This API l limit on the text content which it can process at a time. For large text files, we recursively split it into p until keywords are extracted from the split parts and then all unique keywords are taken as a set fo current citation. Call this set K1.
(b) Extract keywords from query research article, following the same steps as in step (a). Call this set K.
(c) Matching score is generated for set K and K1 and results are classified into three categories based or similarity score.

The above process is repeated for finding similarity score of query research article in succession with each o references.

Then the final Score is normalized by the formula

$$\text{Normalized final score S} = \frac{final\ score}{size\ of\ K + size\ of\ K1}$$

After generating score S for each of the reference, it is matched upon the pre computed values to decide in w category it lies. The three categories are defined as below:

If 'S' is the matching score of citation 'C'

$0.00 < S < 0.05$ then 'C' is under-used citation
$0.05 < S < 0.25$ then 'C' is normal-used citation
$0.25 < S < 1.00$ then 'C' is over-drawn citation

In the above process we have found out the most relevant references given by the author and after matchin the keywords of the references and the query article the similarity of the content was also found using exte trigram technique for the textual plagiarism.

Now, we generate query of trigram like fragments of words 1-3, 4-6, and so on and then search these e generated query for the similar content on the selected references which have normal and over drawn score. For downloaded documents are saved in the local database in the .txt form for further processing of the te similarity.

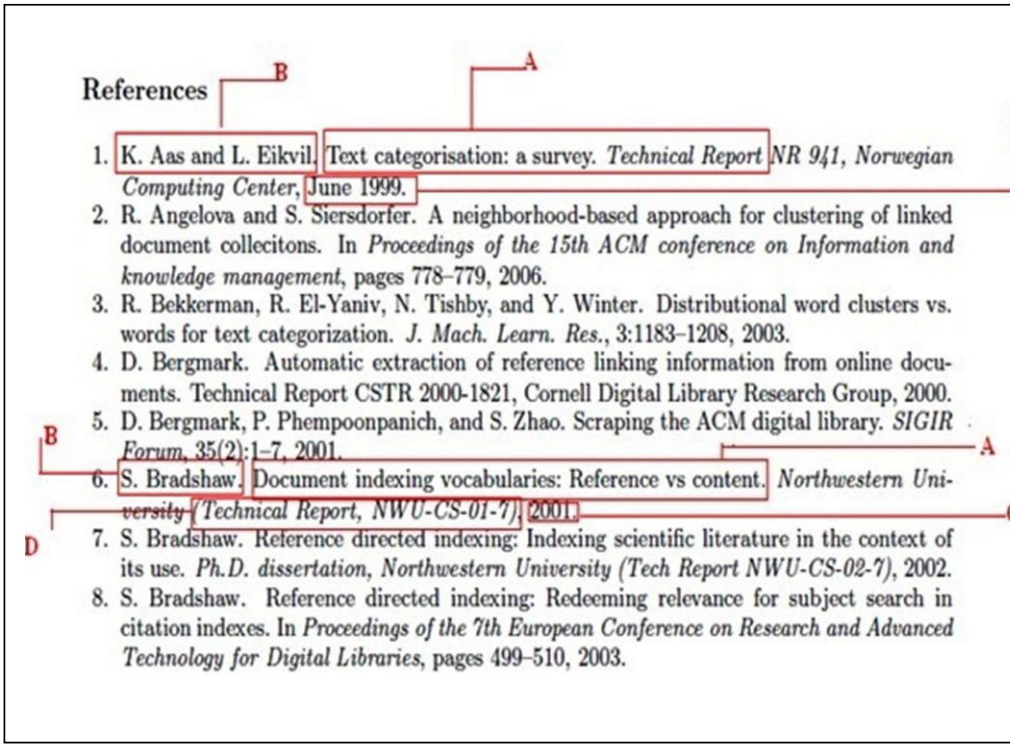
To find the suspected plagiarized portions from the downloaded documents, the sentences are fragmented trigram words sequences. Let the suspicious document *s be* split into sentences (*si*). Now *si* is split into wo grams. The set of *n*-grams represent the sentence. Thus a query document *d* is not split into sentences, but si into word *n*-grams; and each sentence $si \in s$ is searched singleton over the set of downloaded documents called l order to determine if *si* is a plagiarized portion of $d \in D$, we compare the corresponding sets of *n*-grams.

$$C\ (si \mid d) = \frac{|N(si) \cap N(d)|}{|N(si)|}$$

where $N(\cdot)$ is the set of n-grams in $(\cdot)$. If the maximum $C(si \mid d)$, after considering every $d \in D$, is g
than a given threshold, si becomes a portion plagiarized in the query document d.

4. The test cases- data sets and the results

In the proposed methodology, the reference based plagiarism checking is the second target. The idea is to
the query research paper whether it has manifested any copyright violations. For this, we download each and
reference given in the query research paper and also download its referred research articles from the interne
then with the help of tool for information access and web mining, we performed keyword based search wh
defined in the section 3.2 on all the downloaded the documents. To download referenced research paper
consider A) the title of the paper, B) Author of the paper, C) the year of the publication and D) the public
conference or journal name. Based on all keyword matches on these queries, a repository of web searched pap
downloaded to form the local short listed repository. The process is automated and for each reference, it is rep
to have downloads of all reference given in the query document. Figure 2 shows the options taken for the pr
keyword matching for the given set of references.



Once the downloaded research articles are deposited in the local repository, the keyword extraction proces
each and every document is performed. These are then matched to find out the similarity score of query refe
article with each of its referenced research articles. Based on the similarity score thus calculated, we then calc
the three specified sections of the categorization process. We find out the under used citations, normal citation
overused citations.

After testing on various research articles, the average score has been calculated in the most of the document
is given below in Table 1. It is the overall system performance for the first target of the proposed methodology.

Table 1 The overall citation results of the system for References in 3 different cases.

Main Category	Accuracy
Under used	(25/29) = 86.5%
Normal Used	(42/47) = 89.3%
Over Drawn	(20/23) = 86.9%

Thus proposed method is quite accurate to arrive at i) under used, ii) normal used, and iii) over drawn refere for a given query research paper.

In the present work, thus we have developed the references based textual similarity finding of the research pa for checking the plagiarism of the contents amongst two generations of the given research papers. Therefore approach is new and the results are very satisfying. The possible uses of the present application seem very va near future in regard to the detection of copyright plagiarism cases. The tables 2, 3, 4 given below are the re arrived at using our methodology for the 150 test cases including 50 research papers, 50 master dissertations, an doctoral dissertations respectively for a newly established institution to quantify the behavior of the references in the research works of its students having no facility for any prior check up of plagiarism.

.

Table 2 Research Articles

Percentage of Plagiarism Found	Total Number of cases	From Single Source and References is not Given	From Single Source and References is Given	Multiple Sources and References is not given	Multiple Sources and References is given
10%-20%	**25**	**15**	**05**	**03**	**02**
20%-30%	**15**	**10**	**02**	**02**	**01**
30%-40%	**05**	**03**	**00**	**02**	**00**
40%-50%	**03**	**02**	**00**	**01**	**00**
50%-Above	**02**	**01**	**00**	**01**	**00**

Table 3 Master Dissertations

Percentage of Plagiarism Found	Total Number of cases	From Single Source and References is not Given	From Single Source and References is Given	Multiple Sources and References is not given	Multiple Sources and References is given
10%-20%	**15**	**05**	**02**	**06**	**02**
20%-30%	**20**	**10**	**05**	**04**	**01**
30%-40%	**08**	**04**	**01**	**02**	**01**
40%-50%	**05**	**03**	**00**	**01**	**01**
50%-Above	**02**	**01**	**00**	**01**	**00**

Table 4 Doctoral Dissertation

Percentage of Plagiarism Found	Total Number of cases	From Single Source and References is not Given	From Single Source and References is Given	Multiple Sources and References is not given	Multiple Sources and References is given
10%-20%	**28**	**05**	**18**	**01**	**04**
20%-30%	**15**	**05**	**05**	**03**	**02**
30%-40%	**03**	**01**	**01**	**01**	**00**
40%-50%	**02**	**01**	**00**	**01**	**00**
50%-Above	**02**	**01**	**00**	**01**	**00**

5. Conclusion

In the presently emerging and growing technology scenario of the world, many applications developed fo prevention of copyright misconduct have been reported in regards to the academic and research & develop publications. Copyright is a major issue and therefore a focus area for all academics and the publication ho Textual similarity based on word to word, sentence to sentence, semantic analysis have been studied earli finding the plagiarism in the given two documents or query document with other larger set of documents. Hov an important issue needing deeper explorations is the reference based applications. Herein also many applica have been developed and proposed for the different purposes. Citation indexing and bibliographic coupling a most used and referred applications in this regard.

The present research work, proposes a new methodology based on references used in any research paper/ res document first to find out i) under used, ii) normal used and iii) overdrawn references of it and then to plagiarism in it. The reference based textual/content matching of the concerned documents has been then eval for 150 test cases of a newly established institution. In the present paper, based on the testing of these 150 test c it is found that in its research papers and master dissertations, major portion were plagiarized in lower perce from single source and they mostly did not give the references from where they have plagiarized. However i doctoral dissertations, the portions from where they were found plagiarized in lower percentage, the authors n cited the references. Again in overall, there was only insignificant no. of cases which were found not to have the references. Finally a score table is arrived at for finding behaviour of students in regard to referencin categories of a) Research papers in Conference proceedings b) Master Degree Dissertations and c) Doc Dissertations. The analysis of the work has been tested and used to arrest plagiarism cases and enable the ins get free from evils of plagiarism.

6. Future work

In the recent years, research and developments publications have grown rapidly. Many researchers are working different aspects of the publications. In the present research paper references based textual similarity search has investigated and found some interesting facts regarding different level of research works to have different pa The referencing style and their pattern analysis seem to have future scope of the present work for the actual analysis of research works being received for publications.

References

1. MARTIN, B. R. The use of multiple indicators in the assessment of basic research, Scientometrics, (1996) Vol. 36, page 343-362
2. Mac Roberts, M.H., & Mac Roberts, B.R. Another test of the normative theory of citing. Journal of the American Society for Infor Science, (1987b). 38, pages 305–306.
3. Mac Roberts, M.H., & Mac Roberts, B.R. Author motivation for not citing influences: A methodological note. Journal of the An Society for Information Science, (1988) 39, pages 432–433.
4. Mac Roberts, M.H., & Mac Roberts, B.R. Problems of citation analysis: A critical review. Journal of the American Society for Infor Science, (1989) 40, pages 342–349.
5. Mac Roberts, M.H., & Mac Roberts, B.R. Problems of citation analysis. Scientometrics, (1996) 36, pages 435–444.
6. Mac Roberts, M.H., & Mac Roberts, B.R. Citation content analysis of a botany journal. Journal of the American Society for Infor Science, (1997a) 48, pages 274–275.
7. Kostoff R.N, Science and technology transition metrics. Unpublished report online, Office of Naval Research, Arlington, (200 Retrieved September 29, 2009, from http://www.onr.Navy.mil/sci_tech/33/332/ docs/metrics_book_review1.doc
8. Kostoff, R.N., Morse, S.A., & Oneu, S. Seminal literature of anthrax research. Critical Reviews in Microbiology, (2007) 33, page 181.

9. Szava-Kovats, E. Indirect-collective referencing (ICR) in the elite journal literature of physics: II. A literature science study on the le communications, Journal of the American Society for Information Science and Technology, (2002) 53, pages 47–56.
10. Szava-Kovats, E. Phenomenon and manifestation of the "Author's effect of showcasing" (AES): A literature science study, I. Emerg causes and traces of the phenomenon in the literature, perception and notion of the effect. Journal of Information Science, (2008) 34, 30–44.
11. M. M. Kessler Bibliographic coupling between scientific papers. American Documentation, (1963) 14: pages 10-25.
12. H. Small Co-citation in the scientific literature: A new measurement of the relationship between two documents. Journal of the Am Society of Information Science, (1973) 24(4): pages 265-269.
13. H. Small (1973). Co-citation in the scientific literature: A new measurement of the relationship between two documents. Journal American Society of Information Science, 24(4): pages 265-269
14. E. Garfield Science citation index, a new dimension in indexing. Science, (1964) 144(3619): pages 649– 654.
15. R. Mercer and C. D. Marco A design methodology for a biomedical literature indexing tool using the rhetoric of science. In Proceedi the Bio-Link workshop in conjunction with Human Language Technology conference / North American chapter of the Associatic Computational Linguistics annual meeting (HLT/NAACL), (2004)., pages 77–84.
16. H. Nanba, N. Kando, and M. Okumura Towards multi paper summarization using reference information. In Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI-99), (1999) pages 926–931.
17. A. Elkiss, S. Shen, A. Fader, G. Erkan, D. J. States, and D. R. Radev Blind men and elephants: What do citation summaries tell us al research article? JASIST, (2008) 59(1): pages 51–62.
18. Ranjeet Kumar, R.C.Tripathi, An Analysis of Automated Detection Techniques of Textual Similarity in Research Documents, Interna Journal of Advanced Science and Technology, Vol. 56, July, 2013, pages 99-110
19. D. Bergmark Automatic extraction of reference linking information from online documents. Technical Report CSTR2000-1821, (Cornell Digital Library Research Group.
20. B. Powley and R. Dale Evidence-based information extraction for high-accuracy citation extraction and author name recognitic Proceedings of the 8th RIAO International Conference on Large-Scale Semantic Access to Content (2007).
21. I. G. Councill, C. L. Giles, and M. Y. Kan. Parscit An open-source CRF reference string parsing package. In Proceedings of Lan Resources and Evaluation Conference (2008): (LREC 08).
22. Ranjeet Kumar, R.C.Tripathi, "A Trigram Word Selection Methodology to Detect Textual Similarity with Comparative Analysis of S Techniques, 4th IEEE International Conference on Communication Systems and Network Technologies (CSNT), 7-9 April, 2014 Page 387