Reena Kharat, Preeti M. Chavan, Vaibhav Jadhav, Kuldeep Rakibe / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 3, May-Jun 2013, pp.077-080

Semantically Detecting Plagiarism for Research Papers

Reena Kharat, Preeti M. Chavan, Vaibhav Jadhav, Kuldeep Rakibe
Department of Computer Engg, Pimpri Chinchwad College Of Engg., Pune.

ABSTRACT

Plagiarism means copying of published work without proper acknowledgement of source. Plagiarism is a major concern, in an academic environment, which affects both the credibility of institutions as well as its ability to ensure quality of its student. Plagiarism detection of research papers deals with checking similarities with other research papers. Manual methods cannot be used for checking research papers, as the assigned reviewer may have inadequate knowledge in the research disciplines. They may have different subjective views, causing possible misinterpretations. Therefore, there was an urgent need for an effective and feasible approach to check the submitted research papers with support of automated software. A method like- text mining method came into picture to solve the problem of automatically checking the research papers semantically. Our proposed system uses Term Frequency- Inverse Document Frequency (TF-IDF) and Latent Semantic Indexing (LSI) to semantically find plagiarism.

Keywords - Decision Support Systems, Latent Semantic Indexing (LSI), Term Frequency- Inverse Document Frequency (TF-IDF), Text Mining

I. INTRODUCTION

Plagiarism by students, professors, industrialist or researcher is considered academic fraud. Plagiarism is defined in multiple ways like copying others original work without acknowledging the author or source. Original work is code, formulas, ideas, research, strategies, writing or other form. Punishment for plagiarism consists of suspension to termination along with loss of credibility. Therefore, detecting plagiarism is essential. Research paper selection is recurring activity for any conference or journal in academia. It is a multi-process task that begins with a call for papers. Fig. 1 shows research paper selection process. Call for paper is distributed to communities such as universities or research institutions. They are then assigned to experts for peer review. The review results are collected, and the papers arethen ranked based on the aggregation of the experts review results.

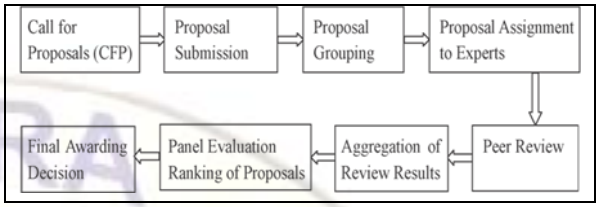

Figure 1*:* Research paper selection process

Expert reviewer may have inadequate knowledge in research discipline. Plagiarism detection software will help him to detect plagiarism quickly. Proposed system requires the database with existing research papers. When the call for papers (CFP) [5] is made from the end-user, the system accepts the research paper submitted by end-user. The system then finds the similarity between the paper submitted and existing research papers.

The proposed method aims to make manual process of checking plagiarism of research Papers computerised. The system allows an agency to ensure the ambiguity of the research Paper submitted by end-user. It helps agency to find semantically similar research papers. Proposed method makes use of TF-IDF and LSI.

II. LITERATURE REVIEW

There are many existing formal methods available for plagiarism detection. W. M. Wangn, C. F. Cheung [6] had proposed- Semantic based intellectual property management system, an automated system for assisting the inventors in patent analysis. It incorporated semantic analysis and text mining techniques for processing and analysing the patent documents. But, this method proposed a hybrid knowledge-based approach to assign reviewers the clustered research papers. MOSS stands for “Measure Of Software Similarity” was developed by Alex Aiken at UC Berkeley. MOSS employs a document fingerprinting technique to detect textual similarity. MOSS is a command line tool and is not easy to use. YAP [8] stands for Yet Another Plague, tries to find a maximal set of common contiguous substrings to detect plagiarism, proposed by Wise. It has three different versions -YAP1, YAP2 and YAP3. Chen et al discussed SID [9] stands for Shared Information Distance or Software Integrity Detection, detects similarity between programs by computing the shared information between them. Prechelt, Malpohl, and Phlippsen has discussed JPlag [10],

that finds plagiarism in source code written in Java, C, C++ and Scheme. The use of minimal match length in JPlag misses some matches.

Apiratikul focussed on Document Fingerprinting Using Graph Grammar Induction (DFGGI) [11], which uses a graph-based data mining technique to find fingerprints in the source code. Authors analysed the advantages and limitations that are currently available with systems for detecting plagiarism and concluded that text-mining, [3] technique can be used to check research papers based on their similarities.

III. TECHNICAL PRELIMINARIES

A. TF-IDF

TF-IDF encoding describes a weighted method based on inverse document frequency (IDF) [7] combined with the term frequency (TF) to produce the feature v, such that

vi = tfi *log(N/dfi)

The weights are assigned using above formula.

Here, N is the total number of papers in the discipline, tfi is the term frequency of the feature word wi and dfi is the number of papers containing the word wi.

TF increases the weight of term and IDF decreases weight of term. The Term-Document matrix is created in this step.as shown in fig

A=	d1	d2	d3
t1	0.58	0	0
t2	0.1	-0.3	0
t3	0	0	0.98

Weighted term doc matrix

B. LSI

LSI is mathematical technique which uses Singular value decomposition SVD. It accepts Term-Document Matrix from TF-IDF and applies SVD on that matrix has many applications like clustering, vector dimension reduction and it is used in making search engines.

	D1	D2
D1	2	5
D2	3	3

	t1	t2
t1	3	5
t2	2	6

From Matrix A. another two matrices are derived namely B,C.where B is document by document matrix,which stores weight of terms which are common to both documents and is calculated using

B = A*A^T C = A^T*A

Where C is term-term matrix which stores weight of both terms which are occurring together in same document. One more matrix is created which is derived from B as square root of Eigen values of principle diagonal matrix, which is denoted as ∑.

And final matrix is build by using,

A = S*∑*U^T.

Where S is matrix obtained from B as Eigen values of B. and U is matrix obtained from C as Eigen values of C.

Now, A can be used for weighting Document vectors And Term vectors.

C. Detecting Plagiarism

The formula required for detecting plagiarism:

Plagiarism= Total number of matched sentences in matched proposal ÷Total number of sentences in the input proposal

IV. PROPOSED METHOD TO DETECT PLAGIARISM

The Plagiarism system checks the similarity of the input paper submitted by end-user, with the existing research papers of the respective discipline. The system finally outputs the Best Matching Unit (BMU). The system provides Best 5 matched papers with respect to the input research paper, in the descending order, with the ordered best matched paper, as shown in Figure 2.

After the research papers are submitted by the end-users, the papers in provided discipline are checked using the text-mining technique, as shown in Figure 3. The main plagiarism process consists of four steps, as:

Step 1) *Text document collection*:

The existing research papers are stored in the text format, within the database.

Step 2)*Text document preprocessing:*

The contents of papers are usually non-structured. The pre-processing analyzes, extracts, and identifies the keywords in the full text of the papers and tokenizes them. Here, a further reduction in the vocabulary size is achieved, through the removal of frequently occurring words referred as stop-words, via- stop file. This is called as filtering phase of removal of stop-words.

Step 3) *Text document encoding*:

On filtering text documents they are converted into a feature vector. This step uses TF-IDF algorithm. Each token is assigned a weight, in terms of frequency (TF), taking into consideration a single research paper. IDF considers all the papers, scattered in the database and calculates the inverse frequency of the token appeared in all research papers. So, TF is a local weighting function, while IDF is a global weighting function.

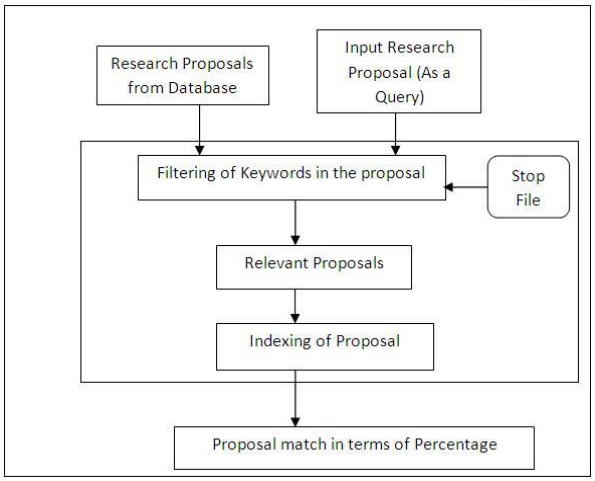

Figure 2: System Architecture

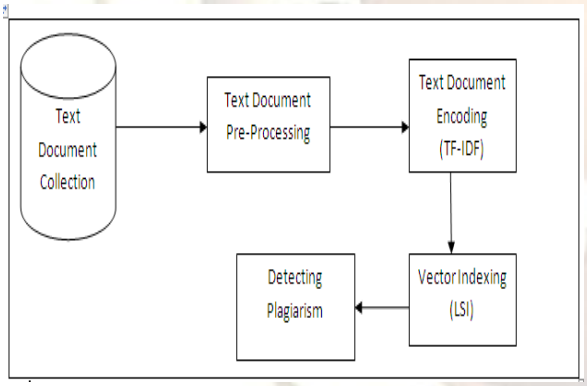

Figure 3: Main process of text mining

Step 4)*Semantically Indexing research Papers:*

LSI creates the semantic relations among the keywords, after gaining feature-vector. LSI is a technique for substituting the original data vectors with shorter vectors in which the semantic information is preserved. A term-by-document matrix is formed, which is decomposed into a set of eigenvectors using singular-value decomposition. The eigenvectors that have the least impacts on the matrix are then discarded. Thus, the document vector formed from the term of the remaining eigenvectors has a very small dimension and retains almost all of the relevant original features. Hence, the system outputs semantically best matched research papers in descending order.

V. CONCLUSION

Today, competition requires timely and sophisticated analysis on an integrated view of data. A new technology leap is needed to structure and prioritize information for specific end-user problems.

Our method facilitates text-mining and optimization techniques to cluster research papers based on their similarities. The proposed method can be used to expedite and improve the paper grouping process.

Plagiarism Detection Method for checking Papers can make this leap. It facilitates text-mining technique to check research papers based on their similarities. It can be used in College Universities to find ambiguity in the SRS, submitted by the students. It can be used for Patent Analysis, for supporting the Intellectual Property Rights. Thus, the future of the proposed system lies in constructing the automated decision-making system for detecting plagiarism.

REFERENCES

[1] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map*", IEEE Trans. Neural Netw, vol. 11, no. 3*, May 2000, 586–600.

[2] Juan Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries", Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855.

[3] R. Feldman and J. Sanger, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data". *New York: Cambridge Univ. Press*, 2007.

[4] Zukas, Anthony, Price, Robert J., "Document Categorization Using Latent Semantic Indexing White Paper" Content Analyst Company, LLC.

[5] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and OuLiu "An Ontology-Based Text-Mining Method to Cluster Papers for Research Project Selection," IEEE transactions on systems, man, and cybernetics—part a: systems and *humans, vol. 42, no. 3*, may 2012.

[6] W.M. Wangn, C.F.Cheung ,"A Semantic-based Intellectual Property Management System (SIPMS) for supporting patent analysis" ,*Knowledge Management Research Centre*, Department of Industrial and Systems Engineering,The HongKong Polytechnic University, HungHom,

Kowloon, Hong Kong.

[7] Milic-Frayling, N., 2005. "Text processing and information retrieval",In Zanasi, A. (Ed.), Text Mining and its Applications to Intelligence, CRM and Knowledge Management. *WIT Press*, Southampton Boston, pp. 1–45.

[8] Wise, M., "YAP3: improved detection of similarities in computer program and other texts", *Proceedings of twenty seventh SIGCSE technical symposium on computer science education, Philadelphia, USA. 130-134*, 1996.

[9] Chen, X., B. Francia, M. Li, B. Mckinnon and A. Seker, "Shared Information and Program Plagiarism Detection", *IEEE Transactions on Information Theory, vol. 50, pp.1545-1551*, 2004

[10] Prechelt, Lutz, Guido Malpohl, Michael Phlippsen, "JPlag: Finding plagiarisms among set of programs", *Technical Report 2000-1*, March 28, 2000

[11] Apiratikul, P., "Document Fingerprinting Using Graph Grammar Induction*", Masters Thesis submitted to the Department of Computer Sciences, Oklahoma State University*,2004