ISSN: 2321-7782 (Online)
Volume 1, Issue 6, November 2013
International Journal of Advance Research in Computer Science and Management Studies
Research Paper
Available online at: www.ijarcsms.com

A Plagiarism Detection Mechanism using Reinforcement Learning

Sudhir D. Salunkhe[1]
M. Tech research Scholar
Dept. of Information Technology
Bharati Vidyapeeth University, College of Engg.
Pune - India

S. Z. Gawali[2]
Assistant Professor
Dept. of Information Technology
Bharati Vidyapeeth University, College of Engg.
Pune - India

Abstract: ***This paper introduces a new approach for plagiarism detection in text documents by using reinforcement learning technique. Detecting plagiarism in a document is finding the dishonesty of the author. In reinforcement learning technique, temporal difference method is used to retrieve the text from document and then to check it with other documents content. This anti-plagiarism mechanism finds results with more accuracy and requires less time because of the use of temporal difference method. The system extracts sentences in the suspected document and then it compares every sentence with database to get the result. Every sentence is checked for grammatical structure too, to get more accurate results.***

Keywords: Plagiarism Detection, Anti-plagiarism mechanism, Reinforcement learning, Temporal difference.

I. INTRODUCTION

Plagiarism is known as copying others work without giving credit to original author. In academics it becomes a major problem and it is difficult to find it manually so a system is envisioned that can find it more accurately and fast.

Plagiarism also deals with copying others thoughts, ideas, concepts etc. without giving credit to the original author or failing to give citation while writing any document. Plagiarism is of two types one is copying text and other is copying ideas [1].

Plagiarism should be detected to find out the dishonesty in document writing. Plagiarism detection for text has two types one is external and other is intrinsic. [2], in external plagiarism detection system the suspected document is compared with all documents such as documents on web or any other database based on the similarity criteria and various detection methods are used whereas in intrinsic plagiarism detection method document is only inspected without comparing it with external documents.

In our system we are using Temporal Difference (TD) learning for text plagiarism detection as an external detection system. TD learning is a combination of Monte Carlo ideas and dynamic programming (DP) ideas. [3] TD resembles a Monte Carlo method because it learns by sampling the environment according to some policy. TD is related to dynamic programming techniques because it approximates its current estimate based on previously learned estimates (a process known as bootstrapping).[4].

In this paper we are discussing plagiarism detection, reinforcement learning and some existing plagiarism detection techniques and also elaborate on our proposed plagiarism detection system using Temporal Difference.

© 2013, IJARCSMS All Rights Reserved

II. PLAGIARISM DETECTION

Plagiarism should be detected to avoid the dishonesty in document writing. In [5] it is stated that plagiarism detection usually is based on comparison of two or more documents. In order to compare two or more documents and to reason about degree of similarity between them, there is a need to assign numeric value, so called, similarity score to each document.

Plagiarism detection, also known as text copy detection, is designed to determine whether a document is copied from other documents in whole or in part without any reference indicated. Besides copying text without any change, changing the order of the original text and replacing synonym are also regarded as plagiarism. [6]. There are many techniques available plagiarism detection which are as in next section.

III. PLAGIARISM DETECTION TECHNIQUES MENTIONED IN LITERATURE

In [7] cluster-based plagiarism detection method is proposed, which has been used in South China University of Technology to check plagiarism in network engineering related courses, also the same is used to detect external plagiarism in PAN-10 competition and the proposed method uses Winnowing's fingerprint extraction algorithm.

In [8] the technique proposed by using document clustering. This technique used to reduce the searching time only and this is not enough to judge the plagiarism in document. The main approach is to reduce time and simplifying the result for source code manual plagiarism detection method. In this method they used Singular Value Decomposition (SVD) for effective clustering of the documents by creating a new matrix with fewer dimensions used for clustering the original i.e. source document and a suspicious document. Neural Network is used for local matching and comparison between a suspicious document and a source document. They used of Kohonen maps for visualization and comparison of the final result.

In [9] a web based semantic plagiarism detection technique is proposed. Which used semantic based similarity detection? They have used google translate API for translating the document for plagiarism checking. The semantic comparison used here determines the similarity level based on only two roles of a sentence i.e. nouns and verbs not on adjectives, prepositions, time and location. It is required to include all roles of sentence to improve the result also there is issue of time required for sematic comparison in this system.

IV. PROPOSED WORK

Plagiarism detection is important for every document to check its originality. In the product development we have used .NET framework and C# as developing language. We have developed a plagiarism detection system for text in which Reinforcement Learning is used. Reinforcement Learning can be used for many applications such as robot control programming, game programming information retrieval etc. Temporal Difference (TD) as one of the method of Reinforcement Learning is used here for Information Retrieval. For separating sentences in the document every sentence is separated after '.', '?', '!' characters and then separated sentences are used for comparison and for making the tree of the sentences for comparison Stanford parser [13] is used.

TEMPORAL DIFFERENCE (TD) LEARNING:

TD algorithm was proposed by Sutton in 1988 [10]. This algorithm combines the Monte Carlo algorithm and the dynamic programming technology [11]. TD is a prediction method in which rewards given to the state and new states are estimated. Simplest TD(0) method is as follows as given in [12],

$$V(S_t) \leftarrow V(S_t) + \alpha[r_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

Where V(S_t) is the estimate of nonterminal state S_t at time t. r_(t+1) is the observed reward, V(S_(t+1)) is the estimante at time t+1. And α is a constant step-size parameter. TD methods wait only until the next time step for rewarding.

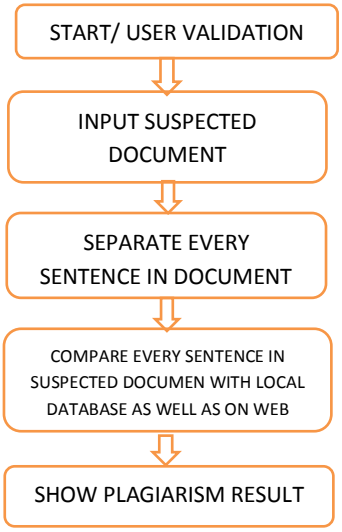
© 2013, IJARCSMS All Rights Reserved

FLOW OF THE SYSTEM:

The system had gone through the following steps:

1. Validation of user, in which user can register or registered user can login to the system.
2. The document such as .doc, .txt or .pdf type is accepted for plagiarism checking. Immediately after a document is given as input, it is separated into separate sentences by using boundary detection algorithm.
3. All stop words in the sentences are removed. Stanford Parser is used for making tree of sentences. Every sentence is now ready to compare with database, either local database or global database over web it also checks for grammatical changes made by the author.
4. For checking document for plagiarism with global database over web Google search API has been used as it can extract sentences in all types of documents and compares the sentences with database over web.
5. Display of result after combining results of offline as well as online plagiarism checking.

Following figure shows diagrammatic flow of system.



ALGORITHM:

Terms Used:

Suspected document – Q;

Sentences in Q – {q1,q2,…qn};

Sentences in original document in database (either local or global database) – {d1, d2,…dn};

Plagiarism set - P=Null;

INPUT: Q.

For Q

Separate sentences Q= {q1,q2,….,qn};

For every q in Q,

Compare with database document D,

If (q==d)

Add sentence to plagiarized set P,

Update result. P=P+q;

End if

© 2013, IJARCSMS All Rights Reserved

End for

If (P==NULL)

Display “document is plagiarism free”

Else

Display,

set P as a plagiarized sentences.

Highlight all P in Q as plagiarism text.

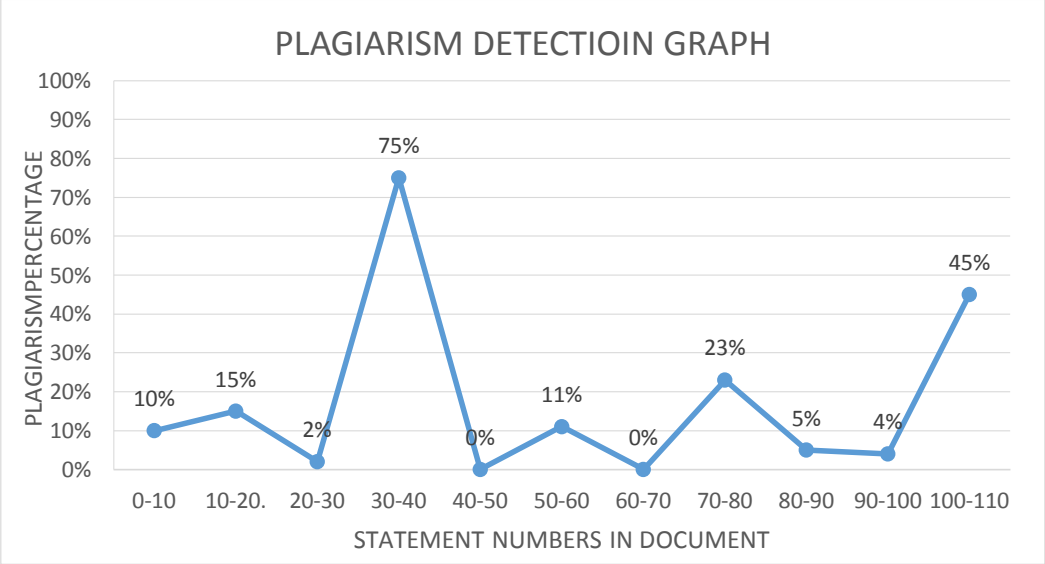
Show graph.

End else

End for

V. Experimental Results

In this system we have used Temporal Difference learning [10] to improve performance of the system. The experiment carried out on a text document. The system gives result in terms of total percentage of document plagiarized. It also highlights the sentences which are plagiarized by changing the background color with yellow and also gives the detected references of the original text according to that sentence. Graph related to the document plagiarism result is shown as below in which result shown is for text document in which total sentences are 107. The graph shows percentage of plagiarism in the group of 10 sentences. The final result for total plagiarism in document is given in figures like 18%



VI. Conclusion

In this paper we have proposed a new plagiarism detection method in which Temporal Difference learning technique is used. Temporal Difference learning is used to improve the speed of system for retrieving the data from database. Also the system improves accuracy of plagiarism detection. We are firstly separating every statement in the document and then Stanford Parser is used for tree formation of sentences after this all sentences are compared with the local and global database to detect plagiarism.

References

1. M. Bouville, "Plagiarism: Words and ideas," Science & Engineering Ethics, vol. 14, pp. 311-322, 2008.

2. Martin Potthast, Benno Stein, Alberto Barr´on-Cede˜no, and Paolo Rosso. An Evaluation Framework for Plagiarism Detection. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, August 2010.

© 2013, IJARCSMS All Rights Reserved

3. Richard Sutton and Andrew Barto . Reinforcement Learning. MIT Press. ISBN 0-585-02445-6, 1998.

4. http://en.wikipedia.org/wiki/Temporal-Difference-learning ,retrieved on 4-11-2013, 2:15 pm.

5. Romans Lukashenko, Vita Graudina, Janis Grundspenkis, "Computer-Based Plagiarism Detection Methods and Tools: An Overview" in International Conference on Computer Systems and Technologies – CompSysTech, 2007.

6. Bao Jun-Peng,Shen, Jun-Yi,Liu Xiao-Dong,Song, Qin-Bao, "A Survey on Natural Language Text Copy Detection[J]",Journal of Software, vol.14, No.10, pp.1753-1760(Ch), 2003.

7. Du Zou, Wei-Jiang Long, Zhang Ling, "A Two-Phase Plagiarism Detection Method" in NSFC (National Natural Science Foundation of China, ID: 60603022), CNGI (China's Next Generation Internet, ID: 2008-122).

8. Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, Vaclav Snasel, Ivo Vondrak "Using Kohonen Maps and Singular Value Decomposition for Plagiarism Detection", Third International Conference on Computational Intelligence, Communication Systems and Networks, IEEE, 2011.

9. Chow Kok Kent, Naomie Salim "Web based Cross Language Semantic Plagiarism Detection" Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2011.

10. Sutton R S. Learning to Predict by the methods of temporal differences. Machine Learning, 1988.Wang Qiang, Zhan Zhongli "Reinforcement Learning Model, Algorithms and Its Application" International Conference on Mechatronic Science, Electric Engineering and Computer, 2011.

11. Richard S. Sutton and Andrew G. Barto "Reinforcement Learning: An Introduction" MIT Press, Cambridge, MA, 1998.

12. http://nlp.stanford.edu/downloads/lex-parser.shtml, retrieved on 20-11-2013 at 4:05 pm.

AUTHOR(S) PROFILE

Sudhir D. Salunkhe, has completed degree in B.E. Information Technology from Shivaji University, Kolhapur in 2011 and currently pursuing the M.Tech. in Information Technology from Bharati Vidyapeeth University College of Engineering, Pune in Maharashtra (INDIA).

S. Z. Gawali, has completed M.Tech. in IT from Bharati Vidyappeth University, College of Engineering, Pune and currently pursuing Ph.D. from Bharati Vidyapeeth University, Pune. Currently working as Assistant professor in Information Technology Department at Bharati vidyapeeth University College of engineering, Pune.

© 2013, IJARCSMS All Rights Reserved