International Journal of Latest Trends in Engineering and Technology
Vol.(10)Issue(1), pp.287-291
DOI: http://dx.doi.org/10.21172/1.101.50
e-ISSN:2278-621X

PLAGIARISM..? THEORY AND MACHINE LEARNING

Dheeraj kumar Sahni[1], Varun kumar[2]

Abstract-In simple words plagiarism means to CHEAT someone's data and publish the data as the cheater himself owner. Plagiarism is basically cheating or you can say copying of other's thought idea, presentation, and document and represent them as their own. Many years ago a term came into lime light and that was Intellectual rights. intellectual rights are those rights which are given to an intellect of a new invention so that no one other than the inventor can claim the invention be her/him and no one can copy the method or process used in the invention , if someone is caught red handed then fine and presentiments are the punishments. Similarly to intellectual's rights there is a term called Copyright. Copyright is used in authentication of text documents which have their authors not inventors, authors lay down their ideas into text, they are provided with Copyrights by which no one other than the author himself/herself can claim the data to be theirs. If someone caught red handed then he/she will be suspect to punish. The breach of copyright law by someone else in any way whether it is copy pasting whether it is copying by modifying is termed as plagiarism. In these way intellectuals rights and copyrights are given to inventors and authors respectively to ensure the safety of their patented document or invention.

Keywords: - Plagiarism, Machine – Learning, Patent etc.

1. INTRODUCTION

1.1 Plagiarism Detection

Plagiarism detection is a method to detect the plagiarised text documents or part of a document which is doubtful to be plagiarised. There are two types of plagiarism:-

First, the plagiarism which is carried out by copying the same original text document as it is, this type of plagiarism is pretty much easy to detect.

Second, the plagiarism which is carried out by modifying the original document and then copies it, this type of plagiarism if very much difficult to detect as the plagiarist uses synonyms and antonyms to modify the document which become difficult to detect. Plagiarism detection is carried out by some algorithmic programs which are framed to do the plagiarism detection of text documents. We can do the plagiarism detection by two means-

1) By manual approach, but it requires large memorisation, high efforts and time complexity will also be higher .This method is not that much effective in case of lengthy suspicious documents.
2) Using machine to detect the plagiarism, machines come to be very effective and efficient way in detection. Machines are given algorithm software programs to detect plagiarism in the suspicious document.

Why Plagiarism Detection

1) In today's era, an era of internet, Where almost everything is available on the internet made it very easy to plagiarise the text i.e. copy paste the text which is easy to copy.
2) It is not limited to academic sector but it is extended in sector of research, arts, science, ethics etc.
3) The documents of copyrights can also be downloaded and copy paste and can be claim of plagiarist.
4) It is very much easy to plagiarise because everything is made available publically in huge amount of information, relatively it becomes very much difficult to detect the plagiarised text documents or part of document which are suspicious to be plagiarised.

[1] UIET, MDU, Rohtak, India
[2] UIET, MDU, Rohtak, India

5) These type of doings breach the copyright rule and there is necessary need to detect the plagiarise text to prevent it from fake author.
6) Plagiarist can use the original documents by modifying it which becomes very difficult to detect.

1.2 Indexing of Plagiarism

Intrinsic Plagiarism

1) A plagiarism in which there is no source documents available digitally i.e. they are in books but not present digitally to compare with suspicious document is termed as intrinsic plagiarism.
2) Intrinsic plagiarism is somewhat difficult to detect because there is no reference documents to compare with.
3) Also, when there are no documents present digitally the other way is to detect plagiarism manually but manual detection is not an effective way to deal with plagiarism.
4) Therefore we need to do it with machines , there are some algorithms which deals with intrinsic plagiarism
    - Plag -inn algorithm
    - Genetic algorithms

1.3 Extrinsic Plagiarism

1) In extrinsic plagiarism there are source documents available for reference purpose of detection of plagiarism.
2) Extrinsic plagiarism is relatively simple as compared to intrinsic plagiarism.
3) Extrinsic plagiarism contains a reference corpus from which plagiarism is detected.
4) Firstly source documents are collected for comparisons of document with suspicious documents.
5) There are certain algorithms for detecting extrinsic plagiarism.

1.4 Intrinsic Plagiarism Detection Algorithm

1) Plag -Inn Algorithm

   a > Plag - inn algorithm is a unique approach in which the only the suspicious document can be analysed because of intrinsic plagiarism where there is no digitally available source files for plagiarised or suspicious to be plagiarised text.

   b > In this algorithm the grammatical differences in sentences of a document is carried out to find the plagiarised text.

   c > This algorithm is coded on the idea that different authors have different grammatical rules for their sentence writing ,So this algorithm is totally based on finding that sentence grammatical differences.

   d > the very first step is to dissolve the whole document into single sentences.

   e > After dissolution of sentences word by word distillation is held, each word is separated in sentences for further process, words are separated in categories verbs , nouns , adjectives etc.

   f > finally a grammar syntax tree will be generated.

   g > A distance matrix of document will be formed which contains the distance between comparing pair of trees.

   f > the resultant vector is calculated for every doubtful for each row in matrix.

   g > by the resultant vector we will calculate the Gaussian Normal Distribution which find outs the predicted mean and standard deviation value.

f >final step is to group those doubtful sentences and find out the plagiarised by Gaussian normal distribution values.

->to optimize the algorithm a reference corpus has been used which holds many English documents for plagiarized detection.

1.5 Genetic Algorithm

1.5.1 Less Computable Algorithm

1) As stated above the plag inn algorithm needs high computations to find out the plagiarised text in documents.
2) To reduce the plag inn algorithm's complexity genetic algorithm was introduced.
3) Genetic algorithm uses fewer computations to find out the plagiarised text.
4) Genetic algorithm is based on natural process of change and development.
5) In this genetic algorithm chromosomes are used as biological evolution parameter and chromosomes are set of genes where a gene is aeon as a parameter.

1.5.2 Genetic Algorithm

a > first of all, in this algorithm a random value is assigned to every gene of a chromosome.

b > chromosomes are expected of size p.

c > all the genes develop a resultant parameter which is used to evaluate the fitness of each chromosome.

d > after then the algorithm changes the parametric value assigned to each genes of 50% of the chromosomes so that the size remains p.

e > the algorithm will check out the healthy gene after the alteration and take it in concentration.

f > repeat the algorithm for specified evolution number otherwise stop.

1.5.3 Documents subsets

1) It is almost inadequate to detect the plagiarised text manually, although to an extent we can compare the documents manually but for lengthy documents this method will prove inadequate.
2) To improve the complexity algorithm we make the subsets of a pre-evaluated length of subsets of sentences.
3) For examples, if we have 300 sentences in a document, we need to choose a splitting number for the subsets, let we choose 155 as splitting number.
4) After choosing 155 as splitting number two subsets will be formed which will contain the sentences below 155th n
5) And the next subset will contain the remaining other sentences above 155th sentence.

1.5.4 Plagiarised Text Documents Are of Two Types

A > exactly same copied documents, i.e. plagiarism took place by copying the exactly same text from original document

- ➔ However, this type of plagiarism is very easy to detect, less efforts and time is consume to find out this type of plagiarism.

B > Plagiarism by modification of original source document is restively difficult task to find out the plagiarist text sentences because synonyms and antonyms are used to modify the document.

- ➔ This type of plagiarism is very much difficult because the words are replaced by their synonyms in suspicious file; hence it comes out to be very difficult to carry out the detection.

2. REFERENCE CORPUS METHOD

1) The documents which are suspicious to be plagiarised are compared with the sources of original documents.
2) The sources of plagiarised documents are found out and comparison between suspicious and original source documents will be held.
3) There will be large number of comparisons to locate the plagiarised text.
4) The source documents which are used as reference for the comparison are named as reference corpus.
5) This method is relatively very much complex and requires many computations.
6) An easy to implement method on the basis of calculations is N-GRAMS method.

3. N-GRAMS METHOD

1) METER (measuring text reuse) corpus find out that n-gram method is more reliable for the plagiarism detection.
2) In n-gram method n number of word comparisons is carried out to find out the plagiarised parts of a document.
3) The suspicious document is dissolved into n number of words.
4) For example , look at this sentence – it is raining outside

   Let us consider n grams to be 2, and then the segmentation will be-

   It is, raining outside.

   Let us consider n grams to be 1, and then the segmentation will be –

   It, is, raining, outside.
5) There are three terms to distinguish n gram comparisons-
   - Exactly matching word by word – in this method suspicious document is compared word by word.
   - Comparisons of two attributes by finding out the similar words between them.
   - By comparing all the words without segmentation present in the suspicious document.
6) This type of segmented methods comes very effective and efficient.
7) This method specified the plagiarist text by breaking the sentences into n words.
8) This method is more reliable than others.
9) Unigrams, bigrams, trigrams etc comparisons are used to carry out the detection.
10) Unigram comparison is not that much effective because it is somewhat similar to compare all the words of the document.

4. MACHINE LEARNING

This technique is similar to technique used in information retrieval (IR) which is to determine the rank retrieval based on measuring the similarity to a query. The similarity-based plagiarism detection can be divided into 3 groups, namely text based, graph similarity and line matching. In machine learning approach no threshold value is required for similarity check. The level of similarity depends on the outcome of learning from the experts that has been presented in a numeric value. Several techniques have proven its performance that is k-NN, SVM, ANN. k-NN is a simple theory with a very good accuracy. To optimize the accuracy of variation the ‘n’ needs to be tried. If X has neighbours mostly lying in plagiarism category then X is a member of plagiarism and vice-versa. SVM is a classifier that also proved to be best especially in cases related to text. The learning technique is for finding an optimal threshold for each class with the goal furthest from boundaries. ANN is a classifier based on modelling the brain works with mathematical models, in this case the machine refer to relationship between neurons with other neurons in other layer. A function f(x) is defined as the function of composition of other functions g(x). ‘K’ is the activation function.

Plagiarism Data Representation

A sentence is generally in natural language. This is enough to construct an algorithm based on comparison of level of sentence. Validation of data starts with the collection of data in standard formats PDF, DOC and TXT. Sentence filtering excludes bibliography, table, no meaning phrase, etc. Pre-processing includes Doc to Text, sentence parsing; Sentence to VSM Data is represented in Vector Space Model.

Hybrid Machine Learning

This work will experiment 4 combinations of K-NN – SVM, K-NN – ANN, SVM – ANN and ANN – SVM. This can sometimes conclude that this may not only increase performance but sometimes can decrease the performance too. ANN-SVM method is the best method in overall performance.

5. REFERENCES:

[1]. D. Gusfield. Algorithms on Strings, Trees, and Sequences. Cambridge University Press, 1997.

[2]. G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Distributed Representations, pages 77-109. MIT Press, Cambridge, MA, USA, 1986.

[3]. R. W. Irving. Plagiarism and collusion detection using the smith-waterman algorithm. Technical report, 2004.

[4]. R. Lukashenko, V. Graudina, and J. Grundspenkis. Computer-based plagiarism detection methods and tools: An overview. In Proceedings of the 2007 International Conference on Computer Systems and Technologies, pages 1-6. ACM, 2007.

[5]. C. D. Manning, P. Raghavan, and H. Schütze. Introduction to information retrieval. Cambridge University Press, 2008. 6. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pages 3111-3119, 2013

[6]. B. G. Vasudevan, et al. Flowchart knowledge extraction on image processing. In proceedings of IEEE International Joint Conference on Neural Networks, (IEEE World Congress on Computational Intelligence) 2008; 4075-4082.

[7]. A.-M. Awal, et al. First experiments on a new online handwritten flowchart database. Document Recognition and Retrieval 2011; 7874.

[8]. D. Zhang, G. Lu. Review of shape representation and description techniques. Pattern Recognition 2004; 37: 1-19.

[9]. K. S. Candan, M. L. Sapino. Data management for multimedia retrieval. London Cambridge University Press 2010.

[10]. R. Deriche. Using canny criteria to derive a recursively implemented optimal edge detector. Int. J. Comput. Vis. 1987; 1(2): 167–187.

[11]. Merin Paul, Sangeetha Jamal. An improved SRL based plagiarism detection technique using sentence ranking. Procedia Computer Science 2015; 46: 223-230.