Intrinsic Plagiarism Detection Approach Utilizing Support Vector Machine in Text Classification

[1]Neelesh Channawar, [2]Dr. S.B.Kishor
Gondwana University, Gadchiroli,Maharastra,India
channawar.nilesh@gmail.com, s.b.kishor.spc@gmail.com

Abstract: Plagiarism is a process of stealing the concepts, ideas, words, methods or results of others without mentioning the correct identification, recognition or citation. In plagiarism, one can confirm work outside of him, or engage in other hard work, rather than giving correct references and recognition. Today, plagiarism day is a serious problem in the field of research. Plagiarism detection use statistical methods or natural language of data. The purpose of each classification is to create a set of predictive models. Document classification is one of these applications and can be used in many categories such as new classifications, speech recognition. In this study, statistical and semantic features were used to determine the function of support vector machine (SVM) in detecting plagiarism. In this research paper we are providing a brief concept of plagiarism detection and text classification method using SVM and n-gram along with their types, functionalities and area of application. We will survey the effectiveness of SVM in text classification for plagiarism detection.

Keywords— *Plagiarism Detection, SVM, K-nearest neighbour, text classification, machine learning approach.*

Introduction

The current plagiarism is a very serious dilemma in the professional environment and even in the education system. Since everyone has access to the Internet, it is easy to use as a source of information. However, copying files from the Internet can be considered as plagiarism: anything on the Internet can come from books, research or articles. Separating content can lead to serious legal issues; many types of software have been discovered today. This software contains many types of information in which you can use databases such as articles, books or search queries, the Internet or comparisons between files.

The advent of the digital age has significantly increased the number of digital resources available on the Internet. Creating, storing and distributing these digital resources today is easy and straightforward. The rapid development of this digital enterprise has increased the possibilities of infringement of copyright and theft. To solve this problem, since 1990 researchers have stolen their experience in various languages, starting with the identification systems of digital publications [1].

However, the identification program began to detect software misuse through theft in the 1970s [2]. Whereas, many methods and tools are available online for theft identity. However, choosing the best crew identification tool or search detection tool in the best manner is very difficult. This can be due to the lack of proper evaluation environment in the field of plug testing. Plagiarism is the theft of another person's job or idea [3]. It can be done by two ways: (1) To obtain text from specific sources, actions or ideas and (2) present it without the recognition of the text source, work or concept (2). Articles can be stolen in various forms. However, there are often two types of stolen articles, such as: (1) text script plagiarism and (2) source code Plagiarism [4][5]. Theft can occur in a single natural language or in two or more different languages. Many researchers or software vendors are still trying to provide effective methods or tools to detect theft. Generally, two types of plagiarism detection techniques are available based on the use of external resources or references [6][5] such as:

1. intrinsic plagiarism detection: where no external references are used
2. Extrinsic plagiarism detection: where external references are used.

Plagiarism is a key issue in academia, and prevention is a very necessary measure. In recent years, a series of studies have been carried out to show that the task of other authors, the imitation of projects and research has increased rapidly. Due to the research work related to the reproduction of other works, the originality of the person has reached its level, which also reaches the level of student integrity and plays an important role in the student's life and academic

goals for student. This low academic integrity reflects the desire to achieve student goals [7].

SVM is a new technology for binary classification tasks that refers to elements and contains elements nonparametric application statistics, neural networks and machine learning. Like classic technology. SVM also classifies the company as solvent or insolvent based on its valuation. The function of the selected financial key figure. However, this feature is neither linear nor parameter. Formal basics a short introduction to SVM follows. Linear SVM case where the score function is still present. First, linearity and parameterization are introduced to clarify the concept of edge and maximization.

I. MACHINE LEARNING APPROACH

Most existing techniques for detecting plagiarism using similarity measure Approach. This technique is actually similar to the technique used in information retrieval (IR) Determine the rank query based on the measure of similarity to a query [8].

Plagiarism detection based on similarity can be divided into three groups: Similarity (cosine, fingerprint, etc.) [9, 10], graphic similarity (ontology, etc.) [11, 12] and line Correspondence (bioinformatics, etc.) [13]. All techniques are based on similarity measures Returns a similarity degree value of 0 to 1. A value equal to 1 is the largest value of the average degree of similarity is 100%. The lowest value is a sign that you are moving away from the same thing. The problem is the number of places where a reasonable degree of similarity can be considered. as plagiarism? A value of 90% can be considered a minimum threshold for plagiarism, but if you want a higher level of similarity than category of plagiarism, you can also get 95%. It means that the measurement of plagiarism on the basis of similarity requires an adjustment of the similarity threshold [14]. However, the threshold is not required for machine learning. The degree of similarity in machine learning was never visible due to a plagiarism decision, depends on the results of the learning experts presented in digital form value. The experts were asked to rate a series of rate comparisons and assessments. Then he had to determine if it was plagiarism or not. The decision data will be to be an intelligence of machine learning. K Nearest Neighbour (KNN), Vector Learning (SVM) and Artificial Neural Network (ANN). KNN is a simple theory, but the accuracy is very good. In this technique, members of X are classified as their nearest neighbours. The number of nearest neighbours that determine the classification result is usually greater than 1, but not too much [8].

In order to optimize the accuracy of the n value change, you need to try. If X has a neighbour that is primarily a plagiarized category, then X is a plagiarized member and vice versa. Although KNN has high precision, this technique is attributed to the computer calculates the distance to all adjacent members. Therefore, this technique is also known as an inert classifier. SVM is a classifier that proves itself especially in text-related cases. This technique is based on statistical methods. Basically, the formula for this technique is to find the boundary between the two classes. The learning method is to find the optimal threshold for each class whose target distance is the farthest from the two boundaries. The set of coordinates that determine this limit will be referred to below as Support vector. In this case, two types of plagiarism are plagiarism rather than plagiarism [8].

II. TEXT CLASSIFICATION

Text classification is a commonly used technique as a basis for applications in document processing and Visualization, web mining, surveillance technology, patent Analysis, etc. Evaluation of different methods of Experiments, the basis for the choice of a classifier, is a Solution for a specific problem. Not a single classifier is always better [15], so we for practical reasons need to develop a methodology for efficient operation. Text categorization, also known as text Categorization, refers to the problem automatically assign predefined text passages (paragraphs or documents) in predefined categories. It is the task of the text categorization classifies documents based on predefined class's content automatically.

The widespread and increasing availability of text documents in electronic form increase the importance of the use of automatic methods to analyze the content of text documents. The method of using experts in the field also identifies and defines new text documents. Categories are time consuming, expensive and have their own restrictions. As a result, identification and classification of textual documents based on their content Imperative. A series of statistical learning techniques and machines designed for the classification of text, including the regression model, the nearest k-neighbour, the decision tree, Naive Baye, Support Vector Machines, with N-grams and many others. Such techniques are used in many parts of English language as the language of identification, proof of Plagiarism, type of text categorization, news categorization, recommendation systems, spam filtering and many more [16].

III. SUPPORT VECTOR MACHINE

The SVM model is a supervised machine learning technique which is based on the statistical theory. It was first proposed by Cortes and Vapnik(1995) [17] from their original work on Structural risk

minimization and later modified by Vapnik(1998) [18].

Essentially, the SVM serves as a linear separator between two data points to identify two different classes in a multi-dimensional environment. SVM uses a very large set of nonlinear functions regardless of the task. You have a clever way to prevent a match. You have a very smart way to use many functions without the computational overhead you need. The main goal of this approach is to maximize the distance between the classes and to minimize the distance between the hyper plane points. SVM essentially defines the interaction processing in terms of features and repetitive features. The SVM splits the data set into two vector sets under the vector "n" of the dimension space. The SVM algorithm basically constructs a hyper plane environment so that each element is compared to a separate linear line. A hyper plane concept is proposed to separate data based on maximum distance analysis to identify classes. To reduce the error rate, the largest marginal classifier is defined [19].

The SVM classifiers are based on the hyper plane class (w.x) + b = 0w ∈ RN, b ∈ R, according to the decision functions f (x) = sign ((w.x) + b). We can show that the optimal hyper plane is defined as the one with the maximum separation margin between the two classes. In practice, the user specifies the kernel function. The transformation φ (.) is not explicitly specified. Given a function of the kernel K (xi, xj), the transformation φ (.) is given by its own functions (a concept in function analysis). Eigen functions can be difficult to build explicitly. Therefore, we specify the kernel function without worrying about the exact transformation. There is another view that the kernel function as an inner product is actually a measure of the similarity between objects [19].

Some of the kernel functions are [19],

a) Polynomial kernel with degree d, K (X, Y) = (XT Y + 1) d

b) Radial basis function kernel with width σ,

K (X, Y) =exp (- ‖X -Y‖2 / (2 σ 2))

- Closely related to radial basis function neural networks
- The feature space is infinite-dimensional.

c) Sigmoid with parameter θ and K (X, Y) = tanh (kxTy + θ), The 'd', 'σ' and 'θ' are parameters chosen by the user.

Kernel Selection for Support Vector Machine

Training vectors xi are mapped into a higher (may be infinite) dimensional space by the function Φ. Then SVM finds a linear separating hyper plane with the maximal margin in this higher dimension space .C > 0 is the penalty parameter of the error term. Furthermore, K (xi, xj) ≡ Φ(xi)T Φ(xj) is called the kernel function [2]. There are many kernel functions in SVM, so how to select a good kernel function is also a research issue. However, for general purposes, there are some popular kernel functions [17] & [19] [20]-

- Linear kernel: K (xi, xj) = xi T xj.
- Polynomial kernel: K (xi, xj) = (γ xiT xj + r) d, γ > 0
- RBF kernel: K (xi , xj) = exp(-γ ‖xi - xj‖2) , γ > 0
- Sigmoid kernel: K (xi, xj) = tanh (γ xiT xj + r)

Here, γ, r and d are kernel parameters. In these popular kernel functions, RBF is the main kernel function because of following reasons [17] [20]:

1) The RBF kernel nonlinearly maps samples into a higher dimensional space unlike to linear kernel.

2) The RBF kernel has less hyper parameters than the polynomial kernel.

3) The RBF kernel has less numerical difficulties.

Mode of selection in svm

Model selection is also an important topic in SVM. Recently, SVM has performed well in data classification. Their success depends on the setting of several parameters that affect the generalization error. We often refer to this parameter setting process as model selection. If a linear SVM is used, then only the cost parameter C needs to be adjusted. Unfortunately, linear SVMs are typically applied to linearly separable problems.

Many problems can be separated in a nonlinear way. For example, satellite data and shuttle data can not be separated linearly. Therefore, we often use non-linear kernels to solve classification problems. Therefore, we have to choose the cost parameter (C) and the core parameters (γ, d) [21] and [22]. Normally, we use the Cross-Validation raster search method to select the best parameter set. This parameter set is then applied to the training record and the classifier is retrieved. The classifier is then used to classify the test data set for generalization accuracy.

IV. CONCLUSION

The report presented an overview of the theory of SVM. This paper covers various topics in the finding text classification and the role of SVM for plagiarism detection approach. These classification techniques that are useful for detection of Plagiarism. The efficiency of the calculation consists in reducing the mechanical stresses and various suggestions were made. SVM are motivated through statistical learning theory. The theory characterizes the performance of

learning machines using bounds on their ability to predict future data. SVM has been successfully used for text classification technique in computer science. SVMs are set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize the geometric margin.

REFERENCES

[1]. S. Brin, J. Davis, H. Garcia-Molina, Copy detection mechanisms for digital documents, in: ACM SIGMOD Record, Vol. 24, ACM, 1995, pp. 398-409. 24

[2]. A. Parker, et al., Computer algorithms for plagiarism detection.

[3]. M. S. Anderson, N. H. Steneck, The problem of plagiarism, in: Urologic Oncology: Seminars and Original Investigations, Vol. 29, Elsevier, 2011, pp. 90-94.

[4]. N. Charya, K. Doshi, S. Bawkar, R. Shankarmani, Intrinsic plagiarism detection in digital data.

[5]. Hussain A Chowdhury, Dhruba K Bhattacharyya, Plagiarism: Taxonomy, Tools and Detection Techniques.

[6]. S. M. Alzahrani, N. Salim, A. Abraham, Understanding plagiarism linguistic patterns, textual features, and detection methods, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (2) (2012) 133-149.

[7]. Nikhil Ghode, Shubham Jadhav, Sampada Moon, Ashmina Khan, Shrutika Bhalkar , Detecting Plagiarism In Academics Using Levenshtein Distance Algorithm And Semantic Similarity, International Journal on Future Revolution in Computer Science & Communication Engineering ISSN: 2454-4248 Volume: 4 Issue: 3,pp. 471-473.

[8]. Imam Much Ibnu Subroto, Ali Selamat2: Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine. Vol.12, No.1, March 2014, pp. 209~218ISSN: 1693-6930.

[9]. MS Pera, Y.-K. Ng. SimPaD: A word-similarity sentence-based plagiarism detection tool on Web documents. Web Intelli. and Agent Sys. 2011; 9: 27-41.

[10]. N Gustafson, et al. Nowhere to Hide: Finding Plagiarized Documents Based on Sentence Similarity. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WIIAT'08). 2008; 690-696.

[11]. P Foudeh, N. Salim. A Holistic Approach to Duplicate Publication and Plagiarism Detection Using Probabilistic Ontologies. Advanced Machine Learning Technologies and Applications (A. Hassanien, et al., Eds.). Springer Berlin Heidelberg. 2012; 322: 566-574.

[12]. C Liu et al. GPLAG: detection of software plagiarism by program dependence graph analysis. 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA. 2006.

[13]. C Xin et al. Shared information and program plagiarism detection. IEEE Transactions on Information Theory. 2004; 50: 1545-1551.

[14]. C.-Y. Chen et al. Plagiarism Detection using ROUGE and WordNet. Journal of Computing. 2010;2(3): 34-44.

[15]. Manning C. D. and Schutze H., 1999. Foundations of statistical Natural Language Processing [M]. Cambridge: MIT press[23].

[16]. NileshChannawar, Dr.S.B.Kishor, Implicit ascertain for plagiarism detection and text classification, Indian J.Sci.Res. 17(2): 163-167, 2018, ISSN: 0976-2876 (Print) ISSN: 2250-0138(Online),

[17]. V. N. Vapnik, The Natural of Statistical Learning Theory, Springer, New York, NY, USA, 1995.
A. V. N. Vapnik, Statistical Learning Theory, Wiley, New York, NY, USA, 1998.

[18]. Padmavathi Janardhanan, Heena L., and Fathima Sabika, Effectiveness of Support Vector Machines in Medical Data mining, journal of communications software and systems, vol. 11, no. 1, march 2015.

[19]. Chih-Wei Hsu, Chih-Chung Chang, and Chih- Jen Lin. “A Practical Guide to Support Vector Classification” . Deptt of Computer Sci. National Taiwan Uni, Taipei, 106, Taiwan http://www.csie.ntu.edu.tw/~cjlin 2007.

[20]. Durgesh k. Srivastava, lekha bhambhu, data classification using support vector Machine, Journal of Theoretical and Applied Information Technology.

[21]. C.-W. Hsu and C. J. Lin. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 13(2):415-425, 2002.

[22]. Chang, C.-C. and C. J. Lin (2001). LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm .