An Approach to Anti- Plagiarism Using Machine Learning Methods

DURGA BHAVANI DASARI, Dr. VENU GOPALA RAO. K

Abstract: **Being a growing problem, plagiarism is generally defined as literary theft and academic dishonesty in the literature, and it is really has to be prevented and stick to the ethical principal. To detect such dishonesty in document writing and Anti-Plagiarism system is proposed. In which Machine Learning Methods can be used to get fast response of plagiarism in suspected documents.**

Key words: **Plagiarism, Anti-Plagiarism, Machine Learning Methods, suspected documents.**

I. INTODUCTION

1.1 *Plagiarism:*

The plagiarism is known as the copying ideas or words from original document without giving any credit to them. In [9] all of the following are considered plagiarism:

- Turning in someone else's work as own work.
- Copying words or ideas from someone else without giving credit.
- Failing to put a quotation in quotation marks.
- Giving incorrect information about the source of a quotation.
- Changing words but copying the sentence structure of a source without giving credit.
- Copying so many words or ideas from a source that it makes up the majority of your work.

A. Plagiarism Detection

Now many students are publishing documents for their academics and authors are publishing many documents and books but many of them are copying text, ideas from other one's documents without giving any reference or credit to them. This will lose the originality of documents. Manual checking for such dishonesty of document writing is not possible.

To avoid plagiarism an anti-plagiarism system is propose in which the suspected document is compared with local database which contains collection of the original documents and also the suspected document is compared with global data base through web.

There are many methods for plagiarism detection those can be applied to source code, to free text, or to both for plagiarism detection. Our work related to plagiarism detection in text document. For that we are using machine learning methods .

II . Machine Learning Methods

The process machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension -- as is the case in data mining applications – machine learning uses that data to improve the program's own understanding. Machine learning programs detect patterns in data and adjust program actions accordingly.

A. Types of Methods

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system.
They are:

- Supervised learning. The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- Unsupervised learning, no labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.
- In reinforcement learning, a computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal or not. Another example is learning to play a game by playing against an opponent.

Between supervised and unsupervised learning is semi-supervised learning, where the teacher gives an incomplete training signal: a training set with some (often many) of the target outputs missing. Transduction is a

Manuscript received January 09, 2015

Durga Bhavani Dasari, Research Scholar, Dept of CSE, Jawaharlal Nehru Technological University, Hyderabad, India,, +91-9490988482
Dr. Venu Gopala Rao. K, Professor, Dept of CSE, G. Narayanamma Institute of Technology and Science, Hyderabad,India,+91-9849025342

special case of this principle were the entire set of problem instances is known at learning time, except that part of the targets are missing.

III . Reinforcement Learning

In our system we are going to reinforcement learning and much of the computational theory of reinforcement learning(RL) appears in Sutton and Bartow's classic textbook[3]. Reinforcement learning is the concept of artificial intelligence so that it will be able to give best performance to our system.

A. Temporal Difference(TD) Learning

Temporal difference is the one of method related with reinforcement learning. In [10] it is stated that, TD learning is a prediction method. It has been mostly used for solving the reinforcement learning problem. "TD learning is a combination of Monte Carlo ideas and dynamic programming (DP) ideas." 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). As a prediction method, TD learning takes into account the fact that subsequent predictions are often correlated in some sense. In standard supervised predictive learning, one learns only from actually observed values: A prediction is made, and when the observation is available, the prediction is adjusted to better match the observation so TD is best option for plagiarism detection technique [13].

TD learning solves the problem of temporal credit assignment, i.e. the problem of assigning blame for error over the sequence of predictions made by the learning agent. The simplest implementation of TD learning employs a lookup table where the value of each state or state-action pair is simply stored in a table, and those values are adjusted with training. This method is effective for tasks which have an enumerable state space. [14].

Three methods of Temporal Difference are stated in [3] which are as Q-learning, Eligibility Tracing and Actor-Critic Methods. We will use one of these method for developing the plagiarism detection technique. The system will able to detect the plagiarism in suspected document with the help of this technique with more speed and correctness. It will shows the final result in terms of percentage of document plagiarized and also with the help of graph so that it is easy to understand to the user of this system.

B. Mathematical formulation for TD

Let $r_t$be the reinforcement on time step *t*. Let $V_t$be the correct prediction that is equal to the discounted sum of all future reinforcement. The discounting is done by powers of factor of γsuch that reinforcement at distant time step is less important.

$$\bar{V}_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$$

where $0 \leq \gamma < 1$. This formula can be expanded

$$\bar{V}_t = r_t + \sum_{i=1}^{\infty} \gamma^i r_{t+i}$$

by changing the index of i to start from 0.

$$\bar{V}_t = r_t + \sum_{i=0}^{\infty} \gamma^{i+1} r_{t+i+1}$$

$$\bar{V}_t = r_t + \gamma \sum_{i=0}^{\infty} \gamma^i r_{t+i+1}$$

$$V_t = r_t + \gamma V_{t+1}$$

Thus, the reinforcement is the difference between the ideal prediction and the current prediction.

$$r_t = \bar{V}_t - \gamma \bar{V}_{t+1}$$

TD-Lambda is a learning algorithm invented by Richard S. Sutton based on earlier work on temporal difference learning by Arthur Samuel. This algorithm was famously applied by Gerald Tesauro to create TD-Gammon, a program that learned to play the game of backgammon at the level of expert human players.The lambda (λ) parameter refers to the trace decay parameter, with $0 \leq \lambda \leq 1$. Higher settings lead to longer lasting traces; that is, a larger proportion of credit from a reward can be given to more distant states and actions when $\lambda$is higher, with $\lambda = 1$producing parallel learning to Monte Carlo RL algorithms.

IV. CONCLUSION

In this paper we introduced some techniques of plagiarism detection. And we represented one new concept of using reinforcement learning method for plagiarism detection. Temporal difference is the method to be used for developing the system to detect the plagiarism in suspected documents and this method will improve the performance of the system.

REFERENCES

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

[2] Michael Tschuggnall, G¨unther Specht "Detecting Plagiarism in Text Documents through Grammar-Analysis of Authors"

[3] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998.

[4] Daniele Anzelmi, Domenico Carlone, Fabio Rizzello, Robert Thomsen, D. M. Akbar Hussain "Plagiarism Detection Based on SCAM Algorithm" IMECS 2011.

[5] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla , Vaclav Snasel, Ivo Vondrak "Using Kohonen Maps and Singular Value Decomposition for Plagiarism Detection" , 2011.

[6] CLEF (Notebook apers/LABs/Workshops) 2010

[7] S. Brin, et al., "Copy detection mechanisms for digital documents," presented at the Proceedings of the 1995 ACM SIGMOD international

conference on Management of data, San Jose, California, United States, 1995.

[8] Ahmed Hamza Osman1, 2, Naomie Salim1, Mohammed Salem Binwahlan1, Ssennoga Twaha1, Yogan Jaya Kumar1 andAlbaraa Abuobieda "Plagiarism Detection Scheme Based on Semantic Role Labeling", 2012.

[9]. Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla , Vaclav Snasel, Ivo ondrak "Using Kohonen Maps and Singular Value Decomposition for Plagiarism Detection" , 2011.

[10]. S. Brin, et al., "Copy detection mechanisms for digital documents," presented at the Proceedings of the 1995 ACM SIGMOD international conference on Management of data, San Jose, California, United States, 1995.

[11]. Ahmed Hamza Osman1, 2, Naomie Salim1, Mohammed Salem Binwahlan1, Ssennoga Twaha1, Yogan Jaya Kumar1 and Albaraa Abuobieda "Plagiarism Detection Scheme Based on Semantic Role Labeling", 2012.

[12]http://plagiarism.org/plagiarism-101/what-is-plagiarism retrieved on 15-07-2013 at 10:50am

[13]http://en.wikipedia.org/wiki/Plagiarism_detection retrieved on 15-07-2013 at 2:05pm

[14]http://www.stanford.edu/group/pdplab/pdphandbook/hand bookch10.html #x 26-1380009.4 retrieved on 17-07-2013 at 12:15pm

[15]. http://plagiarism.org/plagiarism-101/what-is-plagiarism

[16]. http://en.wikipedia.org/wiki/Plagiarism_detection

[17] http://www.stanford.edu/group/pdplab/pdphandbook/hand bookch10.html #x 26-1380009.4