IJournals International Journal of Social Relevance & Concern
ISSN No: 2347-9698 Volume 2 Issue 10, October-2014

PLAGIARISM DETECTION USING FUZZY SEMANTIC SIMILARITY MEASURES

DURGA BHAVANI DASARI[1], Dr. VENU GOPALA RAO. K[2]

[1]*Research Scholar, Dept of CSE, Jawaharlal Nehru Technological University, Hyderabad, India*

bhavani.dd@gmail.com

[2]*Professor, Dept of CSE, G. Narayanamma Institute of Technology and Science, Hyderabad, India*

Kvgrao1234@gmail.com

Abstract: *Plagiarism involves reproducing information in modified format or sometimes the original as it is. This is quiet among students, researchers and academicians. In such a scenario of the growing research and development publications, many techniques and methodologies have been developed for the plagiarism detection to evaluate the originality in the research documents both in regards to the web based as well as local repository based contents. Different similarity measures are available for comparison of textual data. These similarity measures are used for plagiarism detection. However this paper focuses fuzzy semantic similarly measures with plagiarism detection.*

Key words: Plagiarism Detection, Fuzzy similarity, Similarity measures, cosine similarity, Vector Space Model.

I. INTRODUCTION

Today's huge volume of digital information is both advantages as well as disadvantageous too. Advantageous in sense that we can get each and every information on the net freely available for reference and hence searching time for required information has reduced a lot. Also it makes more acquaintance of information to people. Disadvantage in the sense that finding duplicate documents has become difficult. Manual detection of them takes more time and hence we seek the help of plagiarism detection system.

In the first step, all the documents need to be processed to perform tokenization, stop words removal, stemming etc. In the next step, a subset of documents that may possibly be the sources of plagiarism need to be selected. Vector Space Model(VSM) can be used for this candidate selection. Similarity between a suspicious document and a source document can be a computed using cosine similarity measure between the document vectors weighted by tf-idf scoring.

Thirdly, a sentence-wise in depth analysis using fuzzy semantic based approach to find the plagiarized parts in the suspicious documents. This can detect similar, yet not necessarily the same, statement based on the similarity degree between words in the statement and the fuzzy set.

This stage entails the computation of fuzzy degree of similarity that ranges between two edges: 0 for completely different sentences and 1 for exactly identical sentences. Two sentences are marked as similar (i.e plagiarised) if they gain a fuzzy similarity score above a certain threshold[1]. The last step is post-processing where consecutive sentences are joined to form single paragraphs/sections[3].

2. RELATED WORKS

Plagiarism detection is computer assisted plagiarism detection approach which can be used in academics. It does not rely on the texts of the given documents but depends on the references given with a particular research papers. It first identifies similar patterns in citation sequences of two academic works [2]. Subsequent nonexclusively containing citations by using citation patterns. Citation patterns are identified based on the factors of similar order and proximity of citations with in the text. In order to quantify the pattern's degree of similarity, some other factors, for example the absolute number or relative fraction of shared citations in absolute number or relative fraction of shared citations in the pattern as well as the probability that citations co-occur in a documents are considered. Stylometry [2] applies some statistical methods in order to determine an author's unique writing style. It is mainly used for identifying the authorship attribution or intrinsic plagiarism detection.

3. SEMANTIC LEXICON

A lexicon is a list of words in a language—a vocabulary—along with some knowledge of how each word is used. A lexicon may be general or domain-specific; we might have, for example, a lexicon of several thousand common words of English or some language. The words that are of interest are usually open-class or content words, such as nouns, verbs, and adjectives, rather than closed-class or grammatical function words, such as articles, pronouns, and prepositions, whose behaviour is more tightly bound to the grammar of the language. A lexicon may also include multi-word expressions such as fixed phrases (by and large), phrasal verbs (tear apart), and other common expressions. Each word or phrase in a lexicon is described in a lexical entry; exactly what is included in each entry depends on the purpose of the particular lexicon. The details that are given may include any of its properties of spelling or sound, grammatical behaviour, meaning, or use, and the nature of its relationships with other words. A lexical entry is therefore a potentially large record specifying many aspects of the linguistic behaviour and meaning of a word.

A lexicon can -be viewed as an index that maps from the written form of a word to information about that word. This is not a one-to-one correspondence, however. Words that occur in more than one syntactic category will usually have a separate entry for each category; for example, flap would have one entry as a noun and another as a verb. Separate entries are usually also appropriate for each of the senses of a homonym-refers to lexemes with the same form but unrelated meanings and the second term polysemy –refers to the notion of a single lexeme with multiple related meanings. A lexicon may be just a simple list of entries, or a more-complex structure may be imposed upon it. For example, a lexicon may be organized hierarchically, with default inheritance of linguistic properties. Lexical entries include the linguistic behaviour or use of a word its phonetics, morphology, written forms, behaviour; it‗s relative frequency and all other aspects of its meaning. The word semantic properties include relationship between the meaning of the word and those of other words. The lexicon possesses the inheritance properties we can inherit a lexicon by other one. The ―classical‖ lexical relationships pertain to identity of meaning, inclusion of meaning, part–whole relationships, and opposite meanings. Identity of meaning is synonymy: Two or more words are synonyms (with respect to one sense of each) if one may substitute for another in a text without changing the meaning of the text. The lexicon has a highly semantic structure that governs what words can mean, and how they can be used. This structure consists of relations among words and their meanings, as well as the internal structure of individual words. The linguistic study of this systematic, meaning related, structure is called lexical semantics. We have used the lexeme, an individual entry in the lexicon. A lexeme should be thought of as a pairing of a particular orthographic and phonological form with some form of symbolic meaning representation. The lexicon is a finite set of lexemes. This allows us to include compound nouns and other compositional phrases as entries in the lexicon. We have proposed and created a semantic lexicon based on the feedback data. The semantic lexicon includes the organization entities their properties as well as the words or phrases that define the entity with their properties. The words or phrases are extracted from the pos tagged data by forming the chunk of the data with the help of predefined grammar. The format for the database file is words, entity, properties.

4. SIMILARITY MEASURES

Different similarity measures are available which can be used for comparison of textual contents. Following table

illustrates some of the popular similarity measures along with the proposed measure:

Table1: Different similarity measures

Measure	Equation	Range
Jaccard	$J(x,y) = \frac{\lvert x \cap y \rvert}{\lvert x \cup y \rvert}$	0 to 1
Dice	$D(x,y) = \frac{2(x \cap y \rvert}{\lvert x \cup y \rvert}$	0 to 2
Cosine	$Cos(x,y) = \frac{\sum_i (x_i, y_i)}{\sqrt{\sum_i (x_i)^2} \sqrt{\sum_i (y_i)^2}}$	0 to 1
Matching Coefficient	$M(x,y) = \lvert x\rvert - \lvert x-y\rvert$	0 to |x| where |x|=|y|
Proposed	$P(x,y) = 1 - \frac{\lvert x - y\rvert}{\lvert x\rvert}$	0 to 1

The proposed similarity measure listed in the last row of the above able yields better results in comparison to the other measures. To illustrate this, Let us consider the following two contents:

Content1: Cache memory has maximum hit value of one.[8 words]

Content2: Cache memory has minimum miss value with zero. [8 words]

Total unique words in these two contents is 12. Content1 has 4 words similar to content2. As 4 words out of 8 words of content1 are same as content2's, it can be said that content1 is 50% similar with content2.

4.1 Jaccard Similarity:

S= | X ∩ Y | / | X U Y |

Let us check similarity of content1 and content2 using jaccard similarity.

X={ cache, memory, has, maximum, hit, value, of one}

Y= { cache, memory, has, minimum, miss, value, with, zero}

X ∩ Y = { Cache, memory, has, value}

X U Y = { cache, memory, has, maximum, hit, value, of one, minimum, miss, value, with, zero}

| X ∩ Y | = 4 and | X U Y | = 12

So, Jaccard similarity

S= | X ∩ Y | / | X U Y | = 4/12=0.33

It is seen that, although half of the total words were same, similarity percentage calculated using jaccard formula is 33 % only.

4.2 Dice's similarity:

S= 2 x | X ∩ Y | / | X U Y |
For the contents, dice similarity is
2 x .33 =.66
Here it is seen that, although half of the total words were same, similarity percentage calculated using Dice's formula is 66%.

4.3 Cosine Similarity:

For cosine similarity determination, following word set is constructed

{ cache, memory, has, maximum, hit, value, of, one, minimum, miss, value, with, zero}

Numeric representation of content1 is {1,1,1,1,1,1,1,1,0,0,0,0}

Numeric representation of content2 is {1,1,1,0,0,1,0,0,1,1,1,1}

Cos(content1,content2)= $4/\sqrt{8}x\sqrt{8}$ =4/.95 = .503

Here cosine similarity yields 50.3% similarity.

4.4 Matching coefficient:

For the example contents, matching coefficient is 8-4=4. But as maximum similarity value is not within any fixed range, this measure is not normalized.

4.5 proposed approach:

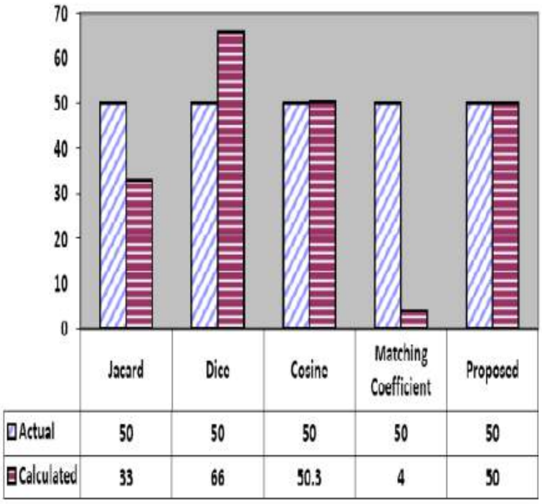
In this paper, a new formula for similarity detection is proposed, which is

$$P(x,y) = 1 - \frac{|x - y|}{|x|}$$

Considering the test content-
|x –y| =4. This is signifying number of words which are in X but not in Y. content1 has 4 words
{maximum, hit, of, one} which are absent in content2.
|X|=8. So, similarity factor,
S=1-4/8=.5

This exhibits that content1 is 50% similar to content2. So, it may be concluded that the proposed similarity measure yields better result in comparison to the other measures. Different commercial plagiarism detection tools are developed based on the measures discussed above. Comparative representation of the existing and proposed measures with graph and data are shown in table2.

Table2: Comparative results of different measures



5 CONCLUSION

In this paper, we have proposed an plagiarism detection using fuzzy semantic similarity measures are described.
Moreover, demonstration of proposed similarity measure and its accuracy over the other measures is explained. This frame work is capable of detecting literal plagiarism where in plagiarists do not spend much time in hiding the academic crime they committed.

6. REFERENCES

[1]. S.M.Alzaharani, N. Salim and A. Abraham, "Understanding plagiarism linguistic patters, Textual Features and detection methods" in IEEE traction on System, man and Cybermetics, PartC: Application and review. Vol.PP,2011, pp.1-1
[2]. S. Meye Zu Eissen and B.Stein, Intrinsic plagiarism detection, in Advances in Information Retrieval 28th European Conference on IR Research, ECIR 2006, pp 565-569.

[3]. S.Alzaharani and N. Salim, " Fuzzy semantic based string similarity for extrinsic plagiarism detection:Lab Report for PAN at CLFE'10", in Proc 4th Int.Workshop PAN-10, padua Italy,2010.

[4]. Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu, ―Using of jaccard Coefficient for Keyword Similarity ||, published in Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong

[5]. SPLAT:A system for self plagiarism detection|| by Christian collbarg, steven Koubhorov,Josua Louie and Thomas slattery, dept. of Computer Science, University of Arizona, Tuscan, AZ 85721.

[6]. B. Karthikeyan, V. Vaithiyanathan, C. V. Lavanya of Sastra University, India – —Similarity Detection in Source Code Using Data Mining Techniques|| published in European Journal of Scientific Research ISSN 1450-216X Vol.62 No.4 (2011), pp. 500-505 © EuroJournals Publishing, Inc. 2011

[7]. Salha M. Alzahrani, Naomie Salim, and Ajith Abraham, Senior Member, IEEE, Understanding Plagiarism Linguistic Patterns,Textual Features, and Detection Methods, published in Ieee Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 42, No. 2, March 2012