Finding Real Semantic of Replaced Words Using K-gram and NGD

Sonal N. Deshmukh, Ratnadeep R. Deshmukh and Sachin N. Deshmukh

Abstract— **Semantic Search has become a buzzword in Web Mining. Researchers have developed variety of algorithms for semantic search. Some of the methods use Search Engines hit count of a sentence for similarity measure. Example of this can be Google Distance measures. A problem of word substitution in the text can be solved by using similarity search measures. Generally, word substitution detection has gained utmost importance as terrorist groups are using substitutions for conveying their messages to their counter parts via email. As the substituted words are normal word, it is difficult to automatically recognize it.**

This paper discusses the methods for detection of substituted word based on search counts like Normalized Google Distance (NGD) and k-gram frequency for measurement of similarity.

***Index Terms*—Text Substitution, Semantic Search, NGD, k-Gram.**

I. Introduction

There is a vast development in communication media, especially in India, in last fifteen years. This includes use of telephones, mobile phones, internet, email etc. This facility is proved beneficial for the illicit acts in terrorisms and crimes too. It includes sending text messages via email or SMS to the group members either using fake identification or by hacking/stealing the device or network link. Emails containing sensitive text can be separated by scanning every email for occurrence of sensitive words and then processing it using another level of data mining algorithms.

However, illicit groups started substituting the sensitive word in the email by a normal word in order to hide the meaning of the sentence so that it can be interpreted as a normal mail. Such type of obfuscation also is seen in the bribe cases where both parties communicate in public. Human intervention can detect such substitutions with the help of contextual information and general sense. However, automatic detection of such obfuscated messages is quite difficult. At the same time, it is not possible to manually scan every message.

Apart from email communication, terrorist groups are using websites to publish objectionable material for example, publishing detailed procedure to manufacture bomb. However, in order to hide the actual meaning of the published material, the data uploaded on the website is obfuscated such that it looks normal to the users.

As substituted words are selected without logic in word selection and they are selected such that the substituted message looks like normal.

This paper discusses the approaches to identify such suspects which can then be processes to next level Data Mining algorithms for further analysis. The approaches present here are based on Search Engine hit count. First approach is based on search count of k-gram of the sentences and second is based on Normal Google Distance (NGD), the algorithm presented by Google Research Lab.

Manuscript received March 10, 2010 .

Sonal N Deshmukh with Jawaharlal Nehru Engineering College, Aurangabad the National Institute of Maharashtra INDIA (e-mail: sonal_deshmukh@ ymail.com).

R. R. Deshmukh is with the Department of Computer Science and IT, Dr B A M University, Aurangabad, MS INDIA (e-mail: ratnadeep_deshmukh@yahoo.co.in).

S N Deshmukh is with the Computer Engineering Department, College of Engineering, Pune, MS INDIA on lien from the Department of Computer Science and IT, Dr B A M University, Aurangabad, MS INDIA (e-mail: snd.comp@coep.ac.in).

II. Literature Survey

With a given bag of words, there can be one or more meaningful sentence/s created. The probability of alteration/substitution of word/s in the sentence will be the conditional probability $p(a|b)$ where 'b' is a bag of words and 'a' is a bag of words with alterations. The probabilities $p(a)$ and $p(a|b)$ can be used to discriminate the suspected sentence.

A substitution or replacement of harmful word with any other normal word is difficult to find. Generally criminals use innocuous words in place of sensitive words to hide the meaning of the communication between them. An efficient technique to automatically flag suspicious messages so that they can be investigated either by a more sophisticated data mining techniques or manually is still a research need [1].

An easier variant, the problem of detecting a substituted word with substantially different frequency from the word it replaces was addressed by SzeWang Fong *et al.* [2]. This work is based on a handful of text rather than a sentence. SzeWang Fong *et al.* presented the measures based on Sentence Oddity, k-gram frequencies, Hypernym Oddity and Point wise mutual Information and proved that these families of measures make errors on different sentences so that, when they are combined, the overall detection rates are close to 90 percent or better and the false positive rates fall to around 10 percent.

PMI and HMM are measures used to detect more suspicious or odd messages. PMI is used to measure strength of association between the word and other string. Here this word may be a substituted word.

HMM is popular in speech recognition. It estimates the probability of occurrences of a word based on the preceding adjacent region. PMI may be better measure compared with HMM [1]. We considered Hidden Markov Model for further research work.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

III. K-Gram Frequencies

An ***n*-gram** is a contiguous sequence of *n* items from a given sequence of text. It is expected that if n-gram is searched in a search engine, the frequencies for the n-gram with original must be greater than the frequencies of the n-gram having substituted word, as a combination of bad of words with substituted word is less likely to occur. We can consider 1 gram, 2 gram, 3 gram string and so on. But it has been observed that more than 3 gram or 4 gram string does not occur on search engine with some frequency [3].
However, as calculation of n-gram may increase the time complexity, a more general form of n-gram, k-gram is proposed to be used [4]. The k-gram of a substituted word is the string containing that word and its context up to and including the first non-stopword to its left, and the first non-stopword to its right. The threshold for classifying the substituted word needs to be decided. Generally, if hit count is 4 then the word is considered to be original otherwise a possible substitution.

SzeWang Fong *et al.* got 90% success rate in the detection with 10% of false positive rate for k-gram frequency methods. We carried out the experimentations as given in the paper.

IV. Normalized Google Distance

NGD (Normalized Google Distance) is an approximation of NID (Normalized Information Distance) [5]. It is a semantic relativity measure derived from the number of hits returned by Google search engine for a given set of words. NGD value is between 0 and 1, value 0 indicates closely related words and value 1 indicates loosely related words. Normalized Google distance between two search terms x and y is

$$NGD(x,y) = \frac{\max\{\log f(X), logf(y)\} - logf(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where M is total number of web pages searches by Google, f(x) and f(y) are the number of hits for searched terms x and y respectively and f(x, y) is the number of web pages on which both x and y occurs.
NGD can be calculated for a pair of words. Being it gives relative distance between a pair of words, NGD can be used to detect the substitution of words problem. It is obvious that NGD calculated for the original word and its every adjacent word in a sentence except a stop word should be less than the substituted word and its adjacent words in the sentence.

V. Experiments

We made a survey of two measures NGD and K gram frequency for detection of substitution of word in test dataset. For experimentation of above mentioned, we used Google and Yahoo! Search engines search count. Google is considered to be the most used and effective search engine [6], [7]. Google uses 'Trust Rank' algorithm to create a personalization vector in Google matrix that decreases the harmful effect of link spamming [8]. So here we decided to consider Google as search engine for testing the data. Behavior of Google search engine is peculiar to the use of punctuations used along with the search term. Use of double quotes to the keyword results in different hit count than that we get with no quotes. This in turn is different than that we get when conjunction 'AND' is used. In our experimentations, we considered all possible ways of giving keywords. For many test cases, use of quotes revealed in hit count of zero only, hence such observations were not considered.

Using NGD to Detect Substitution:
NGD is tested for set of two words and we tried to quantify the strength of relationship between these words. In order to find combined frequency of terms x and y, *i.e.* f(x, y), to calculate NGD is taken on various basis. NGD values ranges from 0 to 1. If NGD of words is 0, we can conclude that there is strong relationship between these two words and if it is 1 then these words are not related.
Consider test data given in Table 1 and Table 2. Search count of these terms is calculated by using '*space*' between words for combined frequency of x and y in Google search engine. Figure 1 shows the result for NGD calculated for related strings such as "Mahatma Gandhi", "United States" etc. and Figure 2 shows the result for unrelated terms like "cocoon trump" etc.

TABLE 1: List of Related Terms

Sr. No	X	y	Sr. No	X	Y
1	Radha	Krishna	9	Ram	Laxman
2	Rahul	Gandhi	10	Tube	Light
3	Tom	Jerry	11	Yellow	Pages
4	Charlie	Chaplin	12	Sun	Rise
5	Hair	Oil	13	Sun	Set
6	Sonia	Gandhi	14	Mobile	Phone
7	News	Paper	15	Krishna	Balram
8	Spider	Man	16	Jawaharlal	Nehru

TABLE 2: List of Unrelated Terms

Sr. No	X	y	Sr. No	X	Y
1	Table	Sky	4	Shirt	Mango
2	Dhruv	Parth	5	Bus	Tea
3	Pen	Aeroplane			

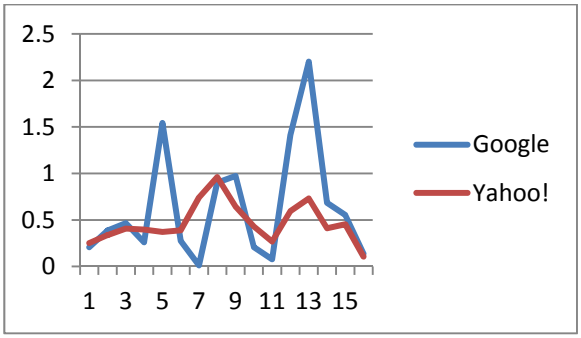

Figure 1: Related Data by Using Space

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

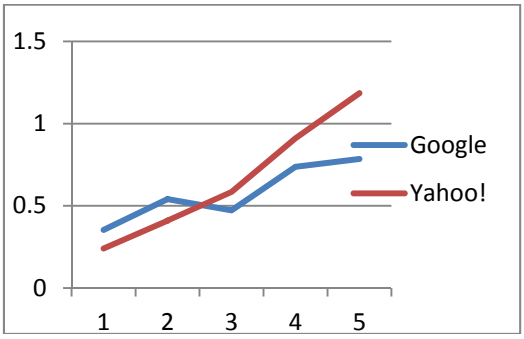

Figure 2: Unrelated Data by Using Space

By using *'and'* between two words, NGD of all sets are near to 0 for Google search engine. But again a set having unrelated words also results near to 0. "Table" and "sky" are unrelated words resulting NGD 0.5409. For yahoo search count, NGD of all sets is near to 0 even for unrelated data. Figure 3 below shows the result for NGD calculated for related strings Figure 4 shows the result for unrelated terms.

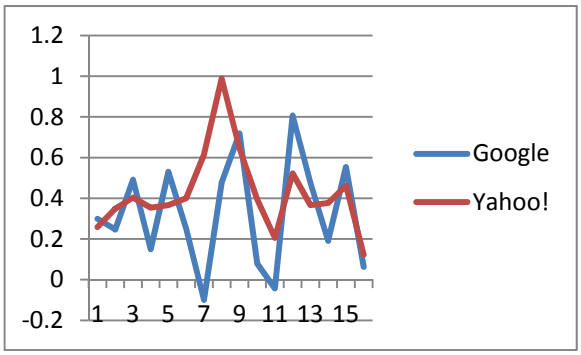

Figure 3: Related Data by Using 'and'

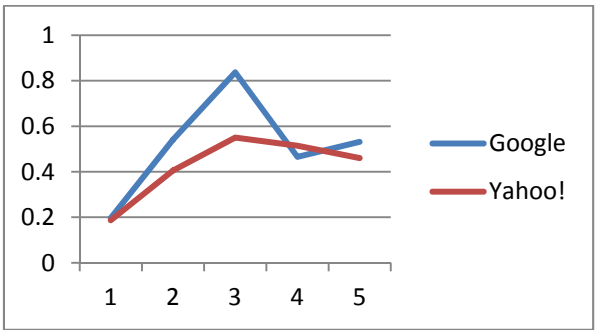

Figure 4: Unrelated data using 'and'

Quoted string gives slight different results than above. If we take two words 'Jawaharlal' and 'Nehru' then we get combined frequency of "Jawaharlal Nehru" as 1880000 in Google and in Yahoo! it is 1850000. Here Google returns more hits than Yahoo! In this case 50% of NGD values are more than 1 for both related and unrelated data in both of the search engines. Figure 5 below shows the result for NGD calculated for related strings Figure 6 shows the result for unrelated terms.

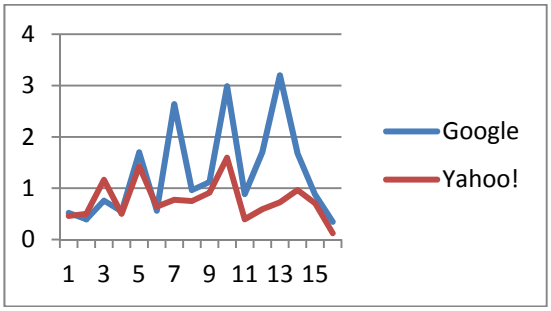

Figure 5: Related Data by Using Quotes

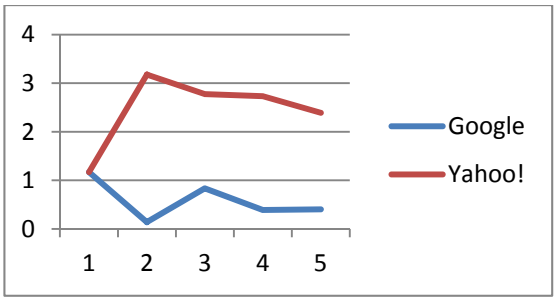

Figure 6: Unrelated Data by Using Quotes

If we consider k-gram of the sentences then the results are different. In our experimentations, we divided the sentence into left k-gram and right k-gram. Instead of using single word frequency for x and y, we used left k-gram and right k-gram frequencies of a sentence. This method works fine for getting NGD for the sentences having maximum four words. Sentence having less number of words can have lesser frequency value resulting in NGD value nearing to 0.

When we used 'and' we got same result indicating that the sentence having more words cannot be verified with this method.

Using K gram to Detect Substitution:

Another measures that we used are sentence oddity (SO) and hyponym oddity (HO) [2]. In our experimentation, we got result zero for both original sentence and substituted sentence. This implied that use of SO and HO leads to no conclusion regarding the relationship between two sentences. Hence we tried an approach of k gram to detect word substitution. Here we divided the sentence into two parts, since it is not very usual to find k gram for whole sentence directly. We calculated left k gram and right k gram starting from a noun in the sentence. Left k gram is starting from considered noun towards left till the start of the sentence is reached and right k gram is starting from considered noun to rightwards till the end of sentence is reached. Considering original and substituted sentence for testing data, we got following results for left k gram and right k gram for original and substituted sentences.

List of sentences and associated results used to test k-gram method is given below in the Table 3. Also the behavior of the right and left k-grams is given graphically in the figure 7, 8 and 9.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

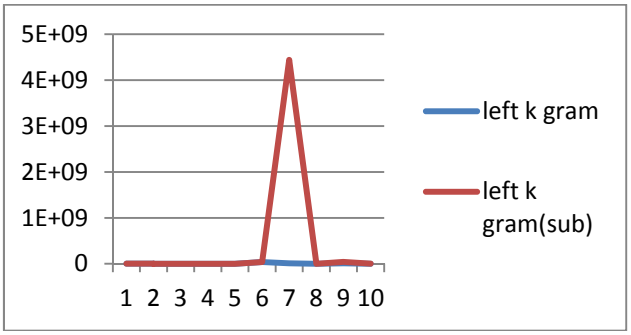

Figure 7: Left k gram for original and substituted sentences

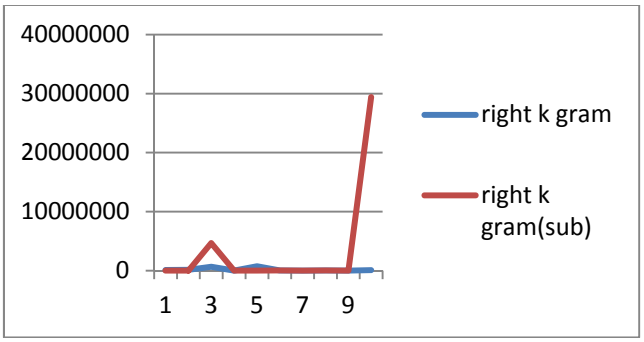

Figure 8: Right k gram for original and substituted sentence

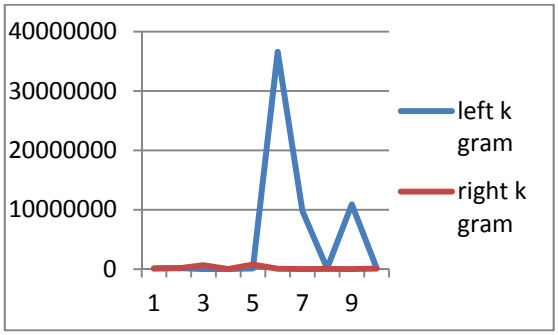

Figure 9: Left and right k gram for original sentences

Figure 7 shows frequencies of left k-gram for original sentences and word substituted sentences. We observed that result varies for both k-grams. Same result is obtained for right k-grams for original and word substituted sentences given in figure 8. We cannot get any idea by comparing frequencies of k grams.

In another test, we used a set of expected sensitive words in the sentence. For a single sentence there may be number of sentences for a set of sensitive words. While comparing left k gram of original sentence and left k gram of word substituted sentence of each sentence, we observed that frequencies for word substituted sentence is greater than original sentence for a set but it is exactly opposite for right k gram.

Table 3: K-gram calculations for sample sentences

Sr.No.	Original sentences	Left k gram	Left k gram frequency	Left k gram freq for substitution	Right k gram	Right k gram freq	Right k gram freq for substitution
1	you will get bomb at Delhi (chocolate)	"will get bomb"	67400	1400000	"bomb at Delhi"	111000	8
2	we have to do murder in Mumbai (felicitation)	“have to do murder"	159000	0	"murder in Mumbai"	142000	2930
3	ramesh will come to collect explosive material(cotton)	"come to collect explosive"	4020	17500	"explosive material"	650000	4670000
4	we expect that attack will happen tonight(marriage)	"we expect that attack"	15000	19900	"attack will happen tonight"	5250	5
5	give training to attack on city(rain)	"give training to attack"	98700	385	"attack on city”	746000	19400
6	the bomb is in position(flower)	"the bomb"	36600000	39800000	"bomb is in position"	46400	39900
7	burn the train tomorrow(colour)	"burn"	9840000	4.44E+09	"burn the train tomorrow"	1	0
8	spread violence as soon possible(happiness)	"spread violence"	144000	1070000	"violence as soon as possible"	32100	21400
9	our next target will be business center(picture)	"next target"	10900000	40900000	"target will be business center"	0	0
10	keep the bomb in the car(bag)	"keep the bomb"	160000	2670000	"bomb in the bag"	93600	29400000

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

VI. Conclusion

This paper used k-gram and NGD for probable detection of substitution of text. The measures uses search count returned by search engine for the given phrase/s. While entering keywords we used 'and'ed strings, used quotations and also searched without quotation. We observed that the result thus obtained proves that k-gram and NGD can be used to detect substitution. However, use of PMI, Cosine similarity and HMM may improve the results, which is a future scoep of the research.

References

[1] Dmitri Roussinov, Szewang Fong, David Skillcorn, "Detecting word substitution: PMI vs HMM", SIGIR 2007, Amsterdam proceedings.

[2] SzeWang Fong, Dmitri Roussinov and David B Skillicon, "Detecting Word Substitution in Text", IEEE transaction on Knowledge and Data Engineering, Vol 20, No. 8, August 2008, pp. 1067-1076

[3] Xiaojin Zhu and Ronald Rosenfeld,"Improving Trigram Language Modeling with the World Wide Web", School of ComputerScience,Carnegie Mellon University,5000 Forbes Avenue,USA

[4] SW. Fong, D.B. Skillicorn and D. Roussinov, "Detecting Word Substitution in Adversarial Communication" 6th SIAM internationalconference on data mining(2006)

[5] Rudi L.Cilibraci and Paul M B Vitanyi, "The Google SimilarityDistance ", IEEE transaction on Knowledge and Data Engineering, Vol 19,March 2007.

[6] Celeste Biever, " Rival engines finally catch up with Google", New Scientist, 184(2004), no. 2474, 23.

[7] Our Search:Google Technology at http://www.google.com/technology/index.html

[8] Rebecca S.Wills, "Google's Page Rank The Math Behind Search Engine", The mathematical Intelligencer @2006 Springer Science

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)