Research Journal of Information Technology

ISSN 1815-7432

Academic Journals Inc.

www.academicjournals.com

Research Journal of Information Technology 6 (4): 406-412, 2014
ISSN 1815-7432 / DOI: 10.3923/rjit.2014.406.412

A Combined Schema for Surveillance

[1]Priya Govindarajan and [2]K.S. Ravichandran
[1]SASTRA University, Kumbakonam, India
[2]SASTRA University, Thanjavur, India

Corresponding Author: Priya Govindarajan, SASTRA University, Kumbakonam, India

ABSTRACT

As recent practice while accessing the popular Web sites indicates, the electronic communication assisted by many malicious sites may provide direction to perceive and thwart vicious intrigue. Posting content which are online and which assist the act of terrorism, such as redeployment on how to formulate grenades etc., are considered as illicit and methods to sense those acts have been survived. The words of similar frequencies are being replaced which is projected as the outline of this study. The current working model doesn't predict the original sentences where as it detects the sentences that differs from the normal perspective. Using the left context of the word (current) and by calculating its co-occurrence statistics the work of analyzing the original sentence is being carried off. The word to be substituted based on the frequency is detected using left-to-right processing technique.

Key words: Analyzing, content, co-occurrence, frequencies, left-to-right technique, method

INTRODUCTION

Surveillance is the scrutinizing the behavior, tricks, or other varying information, usually of people for the purpose of persuading, managing, routing, or protecting them. We do have different types of surveillances carried off via computers, telephone, cameras, social networking, biometric, aerial surveillance and Global Positioning System (GPS) (Wang *et al*., 2013). This study portrays a methodology for tracking information transmitted as "Text content".

The anticipated word is tracked via speech recognition algorithm or else if it is sufficiently unlike then the backup technique is used as a diverse method. For detecting the word for substitution strong left context is mostly is used. There is a limitation in the usage of resources for detection. These differences make it to differ from other detection methods. Letter reversal is used to detect misspelled or the word which is out of the box or out of the context.

This model is not for only tracking lexically formed errors which is used for predicting spellings that are misspelled (Traxler *et al*., 2014) rather than it focuses on the visual quality transforming property of the content. Unusual conversations are very rare and using rare words in an unusual conversation looks unusual. One can predict or track the conversation where words can be substituted only by using the frequency/No. of times a word appears in the context. If a word is among the words, where it is supposed should not be-are tracked as the words to be substituted.

The preceding work were done by replacing (Fong *et al*., 2008) noun with noun, we have elongated the work by replacing verb and other auxiliary part of the sentences within the same perception. The word to be replaced is also considered to be rare in the perspective of the context.

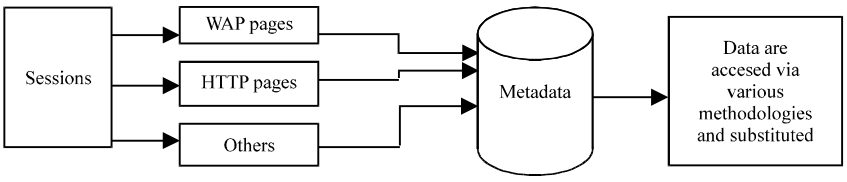

Fig. 1: Converting the metadata for easy access

SCHEMA

Word surveillance can be perceived with different perspective and with diverse terminologies. Surveillance is much needed in the arena of government as well as private sectors, for scrutinizing social activities, threats and for illegal activities. Many surveillance systems (Mehta *et al*., 2010) consider that some of the existing tools may protect society from terrorists and criminals. There are various perspective for surveillance some perceive it as "You have zero privacy anyway, get over it." While others perceive it as "Don't fear, if you aren't wrong in any perspective". It is considered very important in today's environment but it should be carried out without affecting the privacy of an individual.

We symbolize a system or an application (Fig. 1) which tracks words that can be substituted in the due process to fetch its original content or meaning.

Extort data-as resource: We use the e-mail, web source, HTTP pages (dataset) are considered as a source (Pickering *et al*., 2006) for detecting methodology. As an exemplar, consider the e-mail of the employees in the span of 3 years before the crumple of the firm. These documents are considered as unceremonious content and never ever can be brought to the spotlight. These interceptions are needed for the artificial data to track the valid concept within the original context.

Word detection schema: Frequency tends to present number of occurrences of a repeating word or character per unit time. One may find N-number of ways for the manipulation of frequencies or No. of occurrences (Deshmukh *et al*., 2013) of particular word in the sentences. Primarily, for calculating the No. of occurrences projected via API as well as via web interface is quite different from one another. If consistency is the tag word, then one may move with API. Secondarily, one may also consider the indexing system but the outcome are not persistent because of the updating which is made with respective to index. Thirdly, we wish to track the stop words which are never translucent, therefore one never knows the spot to start the detecting mechanism.

Recurrence schema: Number of occurrences of the word that has been substituted should be less (excepted result). The frequency (Chen, 2014) for detecting n-gram is more difficult. Therefore, we target the occurrences of a single word in more generalized way. The substituted words recurrence is calculated in Table 1, by manipulating the first non-stop word to its right and left which are known as Right occurrence and Left occurrence, respectively.

As an exemplar, taking into account the sentence "Twenty kilogram is very heavy to lift", the Left occurrence for 'kilogram' is "Twenty kilogram is very heavy" and the right occurrence for 'heavy' is "Heavy to lift". The number of occurrences of the projected output is manipulated for the entire document which we call as Recurrence schema.

Table 1: Detection of targeted sentences using recurrence

Target sentence	Recurrence
Blast may happen	r = 489
Blast may happen tonight	r = 1
Blast may happen	r = 204
Except that the blast may happen	r = 0
Meeting may happen	r = 26
Meeting may happen tonight	r = 0
Meeting may happen	r = 0

The recurrence (r) of the sentence should be less or zero if it has to be substituted or if there is route for substitution. Table 1 depicts the target sentence using recurrence for which substitutions cannot be brought about, since their recurrence value (r) is either less than zero or equal to zero.

The list of stop word from Wordnet 3.1 were implemented in the example, the left occurrence of "Blast" is "Except that the blast" (f = 60), the excepted right occurrence is "Blast may happen" (f = 8,160). The left occurrence of the substituted sentence of "Meeting" is "Except that the meeting" (f = 87) and the right occurrence for "Meeting may happen" (f = 122).

The recurrence (r) should be smaller for the substituted sentences but in the above example the prediction conceded only with the right occurrence. The conclusion drawn is that the recurrence (r) depends upon the sentence structure and also the way in which the language (English) is understood.

Rarity schema: The frequencies of a sentence (Donnelly *et al.*, 2012) without the substituted word are anticipated to be high, while the frequency of a sentence with substitution should be low (expected outcome). This spots the difference between the original content and the content with masqueraded words. The measures that are applied to the word for detecting are projected by the Rarity of the (ER) Sentence (RS) and enhanced rarity of the sentence.

$$RS = \frac{\text{Bag of words and its recurrence, removing target word}}{\text{Entire bag of words and its recurrence}}$$

$$ER = \frac{\text{Bag of words and its recurrence, excluding target word}}{\text{Entire bag of words and its recurrence}}$$

From Table 1, the frequency of the entire sentence (The meeting may happen) is 2.02 M, when the word "Meeting" is removed, the frequency increases as 5.07 M. By which the rarity of the sentence is calculated as 2.5 (5.07/2.02). If the sentence is substituted then the frequency of the sentence is 1.06 M, the calculated sentence rarity is 4.7 (5.07/1.06). As predicted, Table 2 rarity of the sentence (RS) with substitution is higher than the sentence without substitution.

The rarity of the sentence (RS) for the above calculated sentence in Table 2 where the value of the rarity increases once the sentence is being substituted. It also projects the difference of rarity before and after the substitution in the sentence.

While manipulating the value of RS, some of the words are present both in numerator and denominator, to eliminate the above situation ER (Enhanced Rarity) for the sentence is being defined.

Table 2: Rise in rarity after substitution in the sentence

Frequency of the entire sentence	Target word excluded	Rarity
2.02 (before substitution)	5.07	2.5
1.06 (after substitution)	5.07	4.7

Table 3: Calculation of Hypernym f_H after substitution

Bag of words	Frequency
We except that the blast may happen today	f = 2.02 M
We expect that the meeting may happen today	f_H = 1.03 M
We except that the match may happen today	f_H = 1.41 M
We except that the gathering may happen today	f_H = 1.87 M

As an exemplar, consider the frequency (of 2 sentence) from Table 1 for which the numerator is 3.12 M, then an ER may be predicted as 1.5 (3.12/2.02). If frequency (numerator) in perspective to the substituted sentence is 4.04 M and then an ER may be 3.8 (4.04/1.06). Again the sentence with substitution has the higher Enhanced Rarity (ER) as predicted.

Hypernym measures and ranking: Hypernym of a word drags (Liu *et al*., 2013) use to a point of describing in a generalized manner about "Classes of objects". Consider, place is a hypernym for 'College', 'Airport', 'Firm' etc. The hypernym for the word that has undergone substitution are expected to be more appropriate:

Hypernym (H) = Sentence frequency-Sentence with hypernym

Hypernym may seem inappropriate if the sentence is concept oriented or else the hypernym may seem usual in the context.

The present calculation for a sentence should be nearby zero or negative value then the sentence may be considered (De Maio *et al*., 2014) as concept-centric or else the sentence is considered as contextually inappropriate and dragged in for substitution.

From Table 3, one possible hypernym for "Blast" is "Meeting" and another possible hypernym are "Match", "Gathering". The hpernym score is towards negative values for ordinary sentences and positive for the sentences with excepted substitution. Based on the corresponding value, the sentences are being identified for substitution.

SIMPLE IMPLEMENTATION WITH AN EXEMPLAR

Two measures which are applied to the word for detecting the oddity of the sentence (RS) and enhanced oddity of the sentence (ER) are conducted.

Perception: There is a rise in frequency, if a non-conceptual word is taken off from the text and it becomes low if conceptual word is taken off from the context.

Rise in frequency→Possibility of substitution

Example:

Meeting→Attack

The substituted word should be appropriate in perspective to (Zhou and Zhu, 2010) semantic and syntactic measures which is based on the frequency of the word too.

Practical issues: For implementing the methodologies a popular web service search interface with 1.77 billion indexed pages was used.

Many practical issues were encountered while accessing the pages. First, the handling of the stop words by the search interface was very inconclusive\opaque. For example, at Google, the bag of words for the string "Chase the dreams" occurs 1,77,00,000 times where as the bag of words for the string "Chase dreams" occurs 6,77,00,000 which is very counterintuitive.

Next, the frequencies by Yahoo and Google are taken into accounts which are different in all perspectives. For example, Google generates the quote "Chase the dreams" 5 times in greater volume than Yahoo's frequency. These issues indirectly sense that the data with low frequency must be dealt cautiously.

RESULT AND DISCUSSION

The above mentioned measures were applied to two sets of data (Liu and Curran, 2006), one derived from corpus Brown and other from the Delicti corpus. Both are huge corpus with more than one million words in different perspectives. Some sentences are depicted as an example but the substitutions are not only limited to noun and the Table 4 portrays the schema from the dataset for substitution (Sentence a).

- The person has to be taken off after a ride
- The person has to be murdered after a ride
- During the meeting there will be a presentation
- During the meeting there will be a blast

Table 4 measures the comparisons of different schema and the outcome are being structured (After substitution), for Rarity of the sentence (RS) and Enhanced Rarity (ER), for example for RS the rarity increased after substitution as 3.6 (before substitution>1.5). The left/right occurrence the recurrence (r) slopped down to 122 and 412 after substitution (before which was 90 and 300, respectively). Hypernym projected with a negative value (concept centric sentences-no route for further substitution) after substitution. The entire schema portrayed the predicted values. Next the focus shifts to execution speed of the system, scrutinizing of URL and about multiple substitution.

During the course of the implementation, primarily for extracting the content the aptness of the URL was validated via a testing methodology (Unit test). As the next step the web page was

Table 4: Comparisons of various schemas

Methodology	End-result
Rarity of the sentence (RS)	3.60
Enhanced Rarity (ER)	0.82
Left occurrence	122
Right occurrence	412
Hypernym	-6

processed, by removing the predefined code/algorithm then the text contents within the web page was extracted (expected result). The unwanted grammatical terms were eradicated from the web page like articles, preposition, suffix, prefix. Each and every form of page (with various sizes) with diverse contents was scrutinized.

The appropriateness of protocols, web address and DNS name were dissected with the predefined methodologies, through which invalid URL can be isolated. Processing or extracting content from more than one website at a time is being averted. A single processing cycle can progress one webpage and its corresponding child node. During the sequence we detected that the predefined methods can trigger only the text content (txt, doc, html etc) or else exception was encountered. The entire technique can be justified by the fact that, in today's world most of the information is being routed via Protocols, URL, DNS and Text.

CONCLUSION

The main intention was to explore the viability of the approach. Scrutinizing the search engines with multiple queries was a time consuming process which can be reduced/improved via direct access to the index of the search engine which indirectly also increases the run time of the system. A number of practical limitations are encountered and there is a plan of overcoming those issues in our future research. Even a word substitution can be prolonged to multiple word substitutions in a single sentence with increased accuracy. But these substitutions are very useful and handy for deidentification of data, deception detection and analysis of various domain reports and anti terrorism.

This drags to the fact that, there are some impacts for surveillance such as trust, privacy and autonomy which should be considered as well as justified in all aspects. One may encounter the differences in the thought of deontologists and consequentialists. Deontologists considers, "Surveillance-violates the rights of an individual" where as consequentialists will project its benefits (to the society) rather than the individual's right. The requirement of the surveillance can be limited with the key factors like when, where, why (it is needed) and the importance of its need, by which one can replace "Necessity" with "Requirement". At the end, surveillance is one most vital way to secure a society, as it relocates power from the surveilled to most surveillant.

REFERENCES

Chen, H., 2014. Mining top-*k* frequent patterns over data streams sliding window. J. Intell. Inform. Syst., 42: 111-131.

De Maio, C., G. Fenza, M. Gallo, V. Loia and S. Senatore, 2014. Formal and relational concept analysis for fuzzy-based automatic semantic annotation. Applied Intell., 40: 154-177.

Deshmukh, S.N., R.R. Deshmukh and S.N. Deshmukh, 2013. Finding real semantic of replaced words using K-gram and NGD. Proceedings of the World Congress on Engineering, Volume 3, July 3-5, 2013, London, UK., pp: 1555-1559.

Donnelly, N., K. Cornes and T. Menneer, 2012. An examination of the processing capacity of features in the thatcher illusion. Attention Percept. Psychophys., 74: 1475-1487.

Fong, S.W., D. Roussinov and D.B. Skillicorn, 2008. Detecting word substitutions in text. IEEE Trans. Knowl. Data Eng., 20: 1067-1076.

Liu, C.Y., W. Sun, W. Chao and W. Che, 2013. Convolution neural network for relation extraction. Proceedings of the 9th International Conference on Advanced Data Mining and Applications, December 14-16, 2013, Hangzhou, China, pp: 231-242.

Liu, V. and J.R. Curran, 2006. Web text corpus for natural language processing. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, April 3-7, 2006, Trento, Italy, pp: 233-240.

Mehta, S., U. Eranna and K. Soundararajan, 2010. Surveillance issues for security over computer communications and legal implications. Proceedings of the World Congress on Engineering, June 30-July 2, 2010, London, UK., pp: 480-484.

Pickering, M.J. and R.P.G.V. van Gompel, 2006. Syntactic Parsing. In: Handbook of Psycholinguistics, Traxler, M.J. and M.A. Gernsbacher (Eds.). 2nd Edn., Elsevier, Amsterdam, pp: 455-503.

Traxler, M.J., D.P. Corina, J.P. Morford, S. Hafer and L.J. Hoversten, 2014. Deaf readers response to syntactic complexity: Evidence from self-paced reading. Memory Cognition, 42: 97-111.

Wang, F., L. Huang, Z. Chen, W. Yang and H. Miao, 2013. A novel text steganography by context-based equivalent substitution. Proceedings of the IEEE International Conference on Signal Processing, Communication and Computing, August 5-8, 2013, KunMing, pp: 1-6.

Zhou, G.D. and Q.M. Zhu, 2010. Kernel-based semantic relation detection and classification via enriched parse tree structure. J. Comput. Sci. Technol., 26: 45-46.