Automatic Image Annotation using Semantic Text Analysis

Dongjin Choi[1] and Pankoo Kim[1*]

[1] Dept. Of Computer Engineering Chosun University, Gwangju, South Korea
dongjin.choi84@gmail.com, pkkim@chosun.ac.kr

Abstract. This paper proposed a method to find annotations corresponding to given CNN news documents for detecting terrorism image or context information. Assigning keywords or annotation to image is one of the important tasks to let machine understand web data written by human. Many techniques have been suggested for automatic image annotation in the last few years. Many researches focused on the method to extract possible annotation using low-level image features. This was the basic and traditional approach but it has a limitation that it costs lots of time. To overcome this problem, we analyze images and theirs co-occurring text data to generate possible annotations. The text data in the news documents describe the core point of news stories according to the given images and titles. Because of this fact, this paper applied text data as a resource to assign image annotations using TF (Term Frequency) value and WUP values of WordNet. The proposed method shows that text analysis is another possible technique to annotate image automatically for detecting unintended web documents.

Keywords: Image annotation, Text analysis, WUP measurement, Semantic analysis

1 Introduction

In the last decade, images and videos are the most common contents on the web documents due to the fact those digital cameras and other digital devices became popular over the world. Moreover, lots of Social Network Services (SNS) have been emerged into digital devices especially, Smartphone. The SNS have completely changed human life style into a person who is willing to share his/her current activities through digital photos. However, it has become more difficult to distinguish which data is reliable or not due to huge amount of textural and image data on the web. Moreover, there is a high possibility of leaking personal information of users. Users are able to send any types of data to anyone in anywhere and anytime. This is a serious problem of insider security. The 'insider threat' is an individual with privileges who misuses them or shoes access results in misuse [16]. In order to prevent leakage of personal data, many researchers have been studying recently. [17] proposed a new model of differential privacy for evaluating tables with *k*-anonymity

[*] Corresponding Author

to prevent leakage of personal information. The growing bulk of unstructured data such as text, images and video is needed to be formed into specific predefined manners. In order to satisfy this fact, automatic image annotation method is the first requirement for making structured web data for detecting unintended malicious data such as terrorism. There are two main approaches for image annotation task. First is supervised learning method which was tagged by human hands [1], [2], [3]. These researches applied probabilistic method and ontology scheme to determine which keywords will be precise annotations. The given images were labeled with a common semantic label and classify into corresponding group. These supervised methods guarantee high precision rates though, it requires lots of time and human efforts for labeling manually. For example, if we have an image for animal 'tiger', the system has to discover hypernym and hyponym of 'tiger' concepts. The second approach is an unsupervised automatic image annotation using low-level image features. [4] proposed a method to separate regions of images for detecting objects and describe into small vocabulary of blobs. Automatic image annotation is a popular task in computer vision. Many approaches have been introduced using lots of distinct learning algorithms [7], [8], [9]. Because of the image processing techniques, it is possible to obtain objects in given images. Despite these researches applied different algorithms, all works essentially attempt to learn the correlation between image features and keywords. However, it is still an expensive and challenging task for machine. Hence, automatic image annotation techniques are starting to apply high-level features especially text data [5], [6]. The text data which is surrounding given images and their co-occurring texts have great evidence to discover relevant keywords. This is based on the fact that the surrounding text data of images is likely to describe the given images. For example, let us we have a news document or Wikipedia document. The surrounding texts of images in these web documents explain not only for the given image but also main purpose of documents. It is no doubt that web text data has lots of noisy information. We hereby propose a method to remove irrelevant keywords which were extracted by using Term Frequency (TF) value through WUP similarity in WordNet. WordNet was developed by the Cognitive Science Laboratory of Princeton University and it defines approximately 81,000 noun concepts [10]. WordNet is one of the most well-known Knowledge Base (KB) over the world. So it has been applied to many different fields for finding semantic similarity between terms. Hwang has been studied to grasp semantic similarities and context information from abstract in Wikipedia documents [11], [18]. His research proved that WordNet has valuable information to build semantic network between words for semantic retrieval system. For this reason, we applied modified WUP similarity in WordNet to measure semantic relations between titles and candidate annotations. This paper is organized as follows: Section 2 explains what WUP similarity is. The proposed automatic annotation algorithm is introduced in Section 3. Finally, Section 4 concludes with discussion of future work in this area.

2 WUP Similarity Measurement

WUP similarity [19] is one of the popular methods to measure similarity of nodes. It is a function of the path length from the least common subsumer (LCS) of the two given concepts C_1 and C_2, which is the most specific concept that they share as an ancestor. This similarity value is scaled by the sum of the path lengths from the individual concepts to the root. For example, if C_1 was 'China.n.01[1]' and C_2 was 'Xinjiang.n.01[2]' then the LCS would be 'administrative district.n.01[3]'. The WUP similarity between node C_1 and C_2 is calculated by following formula 1.

$$sim_{wup} = \frac{2 \times depth(LCS(C_1, C_2))}{depth(C_1) + depth(C_2)} \ . \tag{1}$$

where, *depth*(*C*) is the depth of concept *C* in the WordNet hierarchy. The value of this method goes to thigh when two concepts share an ancestor with long depth. The semantic relations between sense 'China.n.01' and sense 'Xinjiang.n.01' defined in WordNet are shown in following Figure 1.

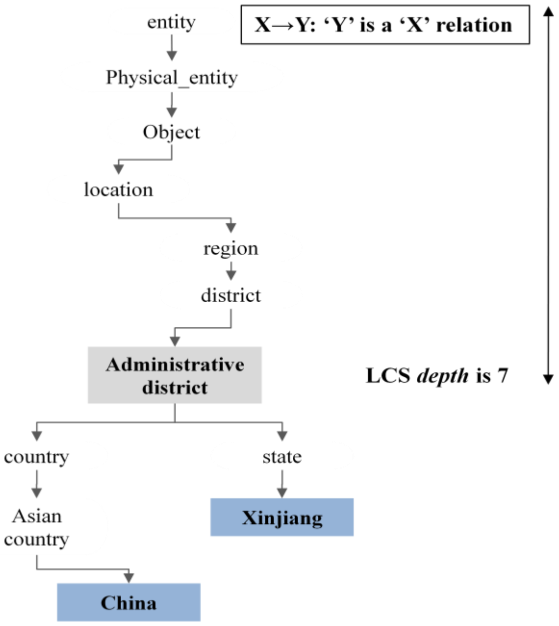


Fig. 1. The concept hierarchy in WordNet

The similarity between 'China.n.01' and 'Xinjiang.n.01' is 0.7368 which means that these two words share many ancestors so, it can be considered as a relevant concept

[1] A communist nation that covers a vast territory in eastern Asia, the most populous country in the world

[2] An autonomous province in far northwestern China on the border with Mongolia and Kazakhstan

[3] A district defined for administrative purposes

each other. In order to grasp lexical similarities or semantic similarities between concepts, lots of studies have been applied WordNet hierarch for building semantic relations. [12] proposed new similarity measurement method for analyzing web documents based on WordNet sense network to make computer can understand human language. Also, [13] tested a semantic similarity using diverse measurement method and compared their accuracy, precision and recall rate, respectively. This research shows that there is no best technique to discover semantic similarities for machine like human does. For example, human can easily distinguish differences between 'bat.n.01[4]' and 'bat.n.02[5]' but machine cannot. Also, human can understand 'jaguar' as a 'vehicle' but computer may misunderstand 'jaguar' as a 'big cat'. This is a major problem when machine tried to understand human language because natural human language is still complicated for machine. In order to overcome this limitation, we applied modified WUP measurement to find most relevant annotation from extracted candidates. The following formula 2 indicates modified WUP similarity.

$$sim_{wup} = \frac{2 \times depth((LCS(C_1,C_2)))^2}{depth(C_1)+depth(C_2)} \ . \tag{2}$$

when the sim_{wup} value is higher than 0.5, *depth*(LCS) will be multiplied again to the numerator. When WUP value goes higher than 0.5, it means that given two concepts are sharing half of all concept hierarchies. So we can emphasize relevant concepts using modified WUP measurement. Eventually, the standard deviation of similarities between two given concepts using *m*_WUP value will be higher than simple WUP value [15].

3 Automatic Semantic Text Analysis

Consider we have news documents consisted of titles, images and surrounding texts. Each of news documents describes current issued information corresponding to given titles and images. There is a traditional problem that the given images are not labelled. If so, annotations were labelled by human hands. This is a disturbing task for human so it has to be automated. For this reason, we propose an algorithm to analyze images and their co-occurring text data. The following Figure 2 shows proposed process for automatic image annotation system. In order to annotate given images automatically, titles of news documents have to be extracted, at first. After extracting a title of given document, stopwords will be deleted. The stopwords are terms that appear so frequently in text that they lose their usefulness as search terms. The stopping is a simple task of removing common words from the stream of token. The most common words are typically function words that help form sentence structure but contribute little on their own to the description of the topics covered by the text [14]. The most popular "the," "a," "an," "that," and "those" are *determiners*. These words are part of how we describe nouns in text, and express concepts like location or quantity. After

[4] Nocturnal mouselike mammal with forelimbs modified to form membranous wings and anatomical adaptations for echolocation by which they navigate

[5] A turn trying to get a hit

stopping process, we extract only noun type of words due to the fact that nouns or proper nouns have significant meaning and they are subjects or objects of sentences. Finally, we can obtain title word list $title_t = \{t_1, \ldots, t_n\}$. Following Table 1 shows the extracted title word lists compare to the original title sentences.

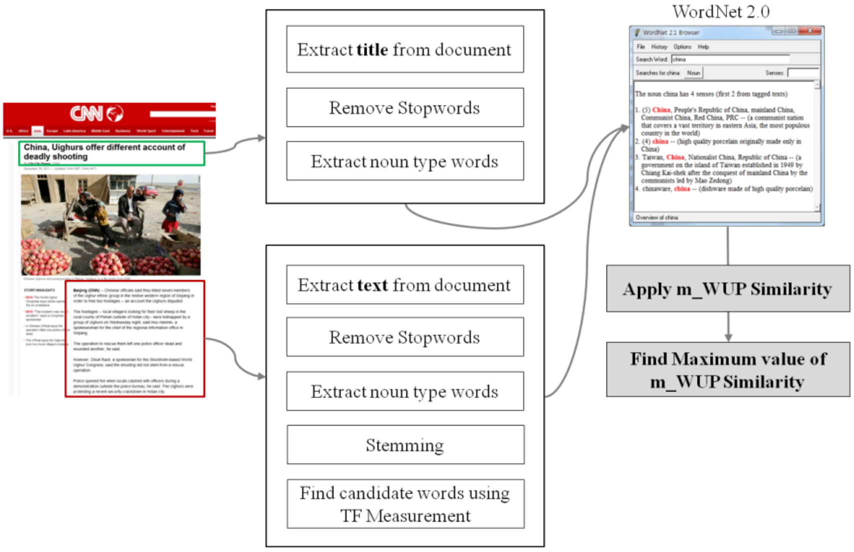

Extract title from document
Remove Stopwords
Extract noun type words
WordNet 2.0
Extract text from document
Remove Stopwords
Extract noun type words
Stemming
Find candidate words using TF Measurement
Apply m_WUP Similarity
Find Maximum value of m_WUP Similarity
China, Uighurs offer different account of deadly shooting

Fig. 2. The proposed system architecture

Table 1. Extracted word list from title

No.	Title	$title_t$
1	China Uighurs offer different account of deadly shooting	{China, Uighurs, offer, account}
2	Donations pour into Philippines in wake of deadly storm	{Donations, Philippines, wake, storm}
3	Top U.S. diplomat to visit North Korea's neighbors	{Top, US, diplomat, North, Korea, neighbors}
4	China Uighurs offer different account of deadly shooting	{China, Uighurs, offer, account}
…	…	…

The next step is extracting surrounding text from documents. In this step we applied *TF* value to determine candidate words. This weight is a value often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a corpus. It is followed by given formula 3.

$$Term\ Frequency\ w(x) = \frac{tf(x)}{\max tf(n)}\ . \tag{3}$$

where, n is a total number of word and max *tf*(n) denotes the maximum frequency of the given document. Thus, the expression computes a term ratio for each term in a surrounding text.

Table 2 shows how surrounding text was changed into candidate annotation. This is the preprocessing step to find candidate terms through removing special characters, stopwords and extracting noun types of words list $text_t = \{k_1, \dots, k_n\}$. Now, we are ready to calculate *TF* weigh of filtered surrounding text. Following Table 3 shows the result of *TF* weight. The total number of terms in new document number 1 is 122.

Table 2. Results processed by each step

Step	Result
Surrounding Text	Beijing (CNN) -- Chinese officials said they killed seven members of the Uighur ethnic group in the restive western region of Xinjiang in order to free two hostages -- an account the Uighurs disputed.
Remove Special Characters	Beijing CNN Chinese officials killed seven Uighur ethnic restive western region Xinjiang free hostages account Uighurs disputed …
Remove Stopwords	Beijing CNN Chinese officials killed seven Uighur ethnic restive western region Xinjiang free hostages account Uighurs disputed …
Extract noun type words	Beijing CNN Chinese officials Uighur western region Xinjiang hostages Uighurs …

Terms which appeared more than twice were shown in Table 3. It is clear that most relevant terms appeared in surrounding text more often. However, this is not always true. The term ‘rescue’ is close to the title “China Uighurs offer different account of deadly shooting’ even though its occurring frequency is two. Moreover, a word ‘terror’ was discarded due to the fact that it only appeared once although ‘terror’ was relevant to given title. In order to overcome this drawback, we multiplied WUP value between t_i and k_j to the *TF* value. Following formula 4 express the semantic weight.

Table 3. *TF* results of news documents #1

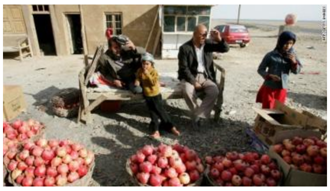
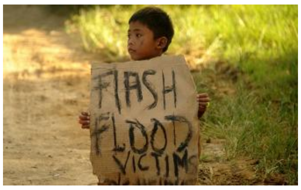
Terms	TF	Terms	TF
Uighur	13/122	Pakistan	2/122
security	4/122	city	2/122
Xinjiang	3/122	government	2/122
crackdown	3/122	hostages	2/122
region	3/122	militants	2/122
Beijing	2/122	operation	2/122
Chinese	2/122	police	2/122
Han	2/122	population	2/122
Hotan	2/122	rescue	2/122

$$SW_{t_i k_j} = w(x) \times Sim_{m\_wup} \tag{4}$$

So, we are able to compare t_i and k_j and determine how much they are closed to. Eventually, we can obtain final annotation for given image through proposed process. The following Table 4 the news images and its annotation grasped automatically.

The annotations are different from other traditional research that described object in given images. Recognizing an object in images is not cover major meaning of images. The proposed approach in this paper focused on the annotations which describe core meaning of given image. Hence, annotated words are semantically related to titles and images. We believe that this annotation can represent not only documents but also images. However, traditional approaches only can detect object in image so, results will be 'human,' 'apple,' 'boy,' 'girl,' and so on for first image in Table 4.

Table 4. Extracted annotation using proposed method

Image		
Annotation	Uighurs, security, Xinjiang, China, terrorism, Asian, crackdown, police, Beijing, Pakistan	Philippines, storm, donation, Asia, Children, China, rain, flood, Australia, Europe

4 Conclusion

The amounts of data which are a mixture of different media have been dramatically increasing. Also the there is a high possibility of personal information leakage. This is a big issue and has to be protected in advance. Automated way to index text, images, audio, and video data is necessary for not only homeland security but also future semantic services. Future semantic web has to annotate image automatically and build semantic relationships between documents and surrounding images to prevent insider threat. Semantic annotation allows us concept search instead of keyword search. In order to make further step for getting close to semantic web and homeland security issues, this paper proposed semantic image annotation approach to analyze images and co-occurring text. We applied modified WUP similarity measurements when values satisfy the predefined condition. The proposed method is simple though still gives possible approach for building semantic image annotation. The costing time of our suggested method is cheaper than traditional annotation system using image recognition techniques. Also it can be applied to another system directly and easily. The most common methods to extract annotation were image object recognition. So, they gave only names of the objects in images. Our approach not only gives context information of documents but also support semantic relationship between title, images, and surrounding text. The weakness of this research is that it is hard to prove whether our approach is adequate or not. For this reason, we have to apply different semantic measurements to enhance reliability of this research. Moreover, when we combine image object recognition technique over our proposed method, the results will be more faithful than current work.

Acknowledgments.
This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation.

References

1. G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 394-410, 2007.
2. A. Th. Scheiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga, "Ontology-Based Photo Annotation," IEEE Intelligent Systems, vol. 16, pp. 66-74, 2001.
3. L. Hollink, G. Schreiber, J Wielemaker, and B. Wielinga, "Semantic Annotation of Image Collections," In Workshop on Knowledge Markup and Semantic Annotation, KCAP'03, 2003.
4. J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic Image Annotation and Retrieval using Cross-Media Relevance Models," Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 2003.
5. Y. Feng and M. Lapata, "Topic Models for Image Annotation and Text Illustration," The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 831-839, 2010.
6. P. Tirilly, V. Claveau, and P. Gros, "News image annotation on a large parallel text-image corpus," 7th Language Resources and Evaluation Conference, pp. 2564-2569, 2010.
7. David M. B. and M. I. Jordan, "Modeling annotated data," Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 127-134, 2003.
8. V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the Semantics of Pictures," Advances in Neural Information Processing Systems 16 NIPS, 2004.
9. K. Barnard and M. Johnson, "Word sense disambiguation with pictures," Journal Artificial Intelligence, vol. 167, pp. 13-30, 2005.
10. T. Deselaers and V. Ferrari, "Visual and Semantic Similarity in ImageNet," CVPR 2011.
11. M. Hwang, C. Choi, and P. Kim, "Automatic Enrichment of Semantic Relation Network and Its Application to Word Sense Disambiguation," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, pp. 845-858, 2011.
12. M. Hwang, D. Choi, J. Choi, H. Kim, and P. Koo, "Similarity Measure for Semantic Document Interconnections," An International Interdisciplinary Journal, vol. 13, no. 2, pp. 253-267, 2010.
13. S. Fern and M. Stevenson, "A Semantic Similarity Approach to Paraphrase Detection," Computer and Information Science, 2008.
14. W. B. Croft, D. Metzler, T. Strohman, "Search Engines: Information Retrieval in Practice"

15. D. Choi, J. Kim, H. Kim, M. Hwang, P. Kim, “A Method for Enhancing Image Retrieval based on Annotation using Modified WUP Similarity in WordNet,” 11th WSEAS International conference on Artificial Intelligence, Knowledge Engineering and Data Bases, 2012.
16. J. Hunker and C. W. Probst, “Insiders and Insider Threats-An Overview of Definitions and Mitigation Techniques,” Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications, vol. 2, no.1 pp. 4-24, 2011.
17. S. Kiyomoto and K. M. Martin, “Model for a Common Notion of Privacy Leakage on Public Database,” Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications, vol. 2, no. 1, pp. 50-62, 2011.
18. M. Hwang, D. Choi, and P. Kim, “A Method for Knowledge Base Enrichment using Wikipedia Document Information,” An International Interdisciplinary Journal, vol. 13, no. 5, pp. 1599-1612, 2010.
19. Z. Wu and M. Palmer, “Verb Semantics and Lexical Selection,” ACL '94 Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133-138, 1994.