Web Image Retrieval Re-ranking with Wikipedia Semantics

Seongjae Lee and Soosun Cho

Dept. of Computer Science & Information Engineering,
Korea National University of Transportation,
380-702, Chungju, Chungbuk, Korea,
lnew1004@gmail.com, sscho@ut.ac.kr

Abstract

Nowadays, to take advantage of tags is a general tendency when users need to store or retrieve images on the Web. In this article, we introduce some approaches to calculate semantic importance of tags attached to Web images, and to make re-ranking the retrieved images according to them. We have compared the results from image re-ranking with two semantic providers, WordNet and Wikipedia. With the semantic importance of image tags calculated by using Wikipedia, we found the superiority of the method in precision and recall rate as experimental results.

Keywords: *Web Image Retrieval, Semantic Relatedness, Flickr, Image Tags, Wikipedia, WordNet.*

1. Introduction

Generally, most photo images stored on the Web have lots of tags added with user's subjective judgments not by the importance of them. So, in tagged Web image retrieval, they have become the cause of precision rate decrease on simple matching of tags to a given query. Therefore, if we can select semantically important tags and employ them for the image search, the enhanced retrieval result could be expected. In this article, we propose a method to make image retrieval re-ranking with the prior tags which share more semantic relatedness to a given query. To calculate the semantic relatedness, we employ the WordNet [1] or Wikipedia [2] as the semantics provider. Test was carried out with Flickr [3] images to confirm how much our method is effective.

Metadata creation of Web images using tags can provide flexible and variety of classification methods, but also has basic limits. One example is that the search work using these tags often result in output with very different image from sought ones. To resolve these problems, we have worked for utilizing the semantic relatedness between tags and a given query acquired from WordNet synonym or hypernym set[4]. And we found that the image retrieval with our semantic relatedness showed some enhancement over simple Flickr retrieval. So we have thought about the utilization of Wikipedia as a semantic provider instead of WordNet. Wikipedia is huge web encyclopedia service composed of around 10 million cases of articles. The data of Wikipedia can be accessed and modified freely by anybody in all around World, but the service is being executed as a mode of reviewing modified contents and then being implemented. That is why many studies were carried out in various fields using Wikipedia.

For example, a research was made to calculate similarity between two Wikipedia articles by comparing corresponding vectors after constructing variable weighted vector with all the terms included in each article [5]. However, more effective method has been published by

Milne & Witten [6]. They applied TF-IDF method on the vectors which are composed of weights of links included each Wikipedia article. In our approach, by applying the method of link weights, relatedness between a tag and a given query is calculated as similarity between corresponding titles of two Wikipedia articles.

Wikipedia is composed of links between articles. The mode of expressing articles as link vectors and calculating the similarity between articles with these links is very simple and effective. As has pointed out in research [6], the measure is defined by the angle between the vectors of the links found within the two articles of interest. The weight of a link is defined by the total number of links to the target article over the total number of articles. Thus if *s* and *t* are the source and target articles, then the weight *w* of the link $s \rightarrow t$ is:

$$w(s \rightarrow t) = \log(\frac{|W|}{|T|}) \qquad if \quad s \in T, \quad 0 \quad otherwise$$

where *T* is the set of all articles that link to *t*, and *W* is the set of all articles in Wikipedia. In other words, the weight of a link is the inverse probability of any link being made to the target, or 0 if the link does not exist. Thus links are considered less significant for judging the similarity between articles if many other articles also link to the same target.

After constructing all vectors of Wikipedia articles using link weight above, to calculate relatedness between a given query *Q* and a tag *T*, cosine similarity of two vectors is used as below.

$$similarity = \cos(\theta) == \frac{Q \cdot T}{\|Q\| \|T\|} = \frac{\sum_{i=1}^{n} Q_i \times T_i}{\sqrt{\sum_{i=1}^{n} (Q_i)^2} \times \sqrt{\sum_{i=1}^{n} (T_i)^2}}$$

Because vectors for around 10 million total Wikipedia articles should be constructed, actual calculation is very hard and not efficient. Therefore, serial numbers of alphabetic order are given on entire articles and corresponding link numbers along with link weight were stored. Cosine similarity between vectors of a query and a tag can be calculated fast and efficiently by multiplying and summing up the link weights whose serial numbers are same.

2. Test and Evaluation

2.1. Test with Flickr Images

For our image re-ranking by using semantic relatedness, tags are arranged in the order of high relatedness with a given query, and sequential search is executed based on this order. That is, after selecting some prior tags with higher relatedness per each image, these are utilized for searched images re-ranking. To calculate semantic relatedness based on Wikipedia, English Wikipedia data on 15th Jan. 2011 were used in this test. This data was composed of pages-articles type single XML file around 30GB. The number of articles collected was 10,861,570 cases (including repetition) and the number of meaningful links was 75,261,480 cases (including repetition).

Also, to use Flickr images for test, Flickr API [7] was employed. The number of images extracted from Flickr using each query such as ‘bird’, ‘car’, ‘house’, or ‘sea’ was 1,500 and the total number of images was 6,000. To compare with results from our former research [4], same test was carried out with new 6,000 selected images on which relatedness calculation based on synonym and hypernym set of WordNet was applied.

2.2. Test Results

Figure 1 shows each search results targeted 1,500 ‘house’ images for the performance evaluation. For judging collected 1,500 ‘house’ images from Flickr, it was determined as ‘correct’ if it is corresponding to appearance of building, otherwise it was determined as ‘incorrect’. As has seen from Figure 1, while 14 images were judged as ‘correct’ in WordNet-based search, all 24 images were judged as ‘correct’ during Wikipedia-based search. Since users want to find required result from front page promptly, by applying our re-ranking method users’ satisfaction can be improved remarkably.

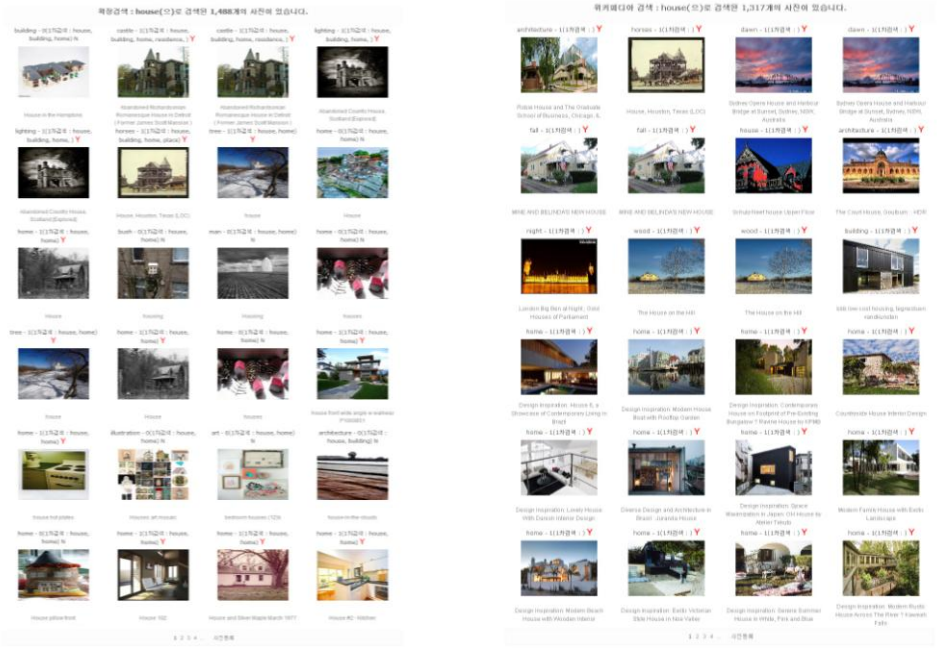


Figure 1. Left:14 images were judged as ‘correct’ in WordNet-based search, Right: All 24 images were judged as ‘correct’ in Wikipedia-based search

Also in Table 1 and Table 2, precisions and recalls are compared from first 4, 8, 12, 16, 20, 24, and 28 pages to accurately evaluate the performance of 2 semantics providers. From the results, we can find the facts as follows:

Firstly, Wikipedia is definitely better semantics provider than WordNet with test images from category ‘bird’ or ‘house’. In these categories, the precisions and recalls are higher with Wikipedia semantics than them with WordNet semantics, for all retrieved pages. On the other hand, in testing with images from category ‘car’ or ‘sea’, WordNet semantics shows more powerful results in precision and recall rates from first 4 pages. But we can also find that Wikipedia overcomes the inferiority on the first 12 pages of ‘car’ images and first 24 pages of ‘sea’ images, respectively.

Table 1. Precision and recall rates with WordNet semantics (%)

pages	bird		car		House		sea	
	precision	recall	precision	recall	precision	recall	precision	recall
Top4	59.4	6.8	80.2	8.3	76.0	7.8	85.4	8.9
Top8	56.3	12.9	73.4	15.1	64.6	13.3	84.9	17.6
Top12	54.5	18.8	68.8	21.2	61.8	19.1	83.0	25.8
Top16	52.3	24.0	67.4	27.8	58.1	23.9	82.3	34.1
Top20	54.8	31.5	66.5	34.2	58.1	29.9	75.8	39.3
Top24	55.4	38.2	66.7	41.2	59.0	36.4	71.9	44.7
Top28	56.0	45.0	64.9	46.8	60.3	43.4	71.7	52.1

Table 2. Precision and recall rates with Wikipedia semantics (%)

pages	bird		car		House		sea	
	precision	recall	precision	recall	precision	recall	precision	recall
Top4	76.0	8.7	71.9	7.4	91.7	9.4	68.8	7.1
Top8	72.9	16.7	68.2	14.1	80.2	16.5	71.9	14.9
Top12	74.0	25.5	69.1	21.4	80.6	24.9	72.9	22.7
Top16	72.1	33.1	70.1	28.9	83.6	34.4	74.0	30.7
Top20	72.3	41.5	68.5	35.3	84.4	43.4	75.0	38.9
Top24	73.1	50.4	67.7	41.8	85.1	52.5	74.1	46.1
Top28	73.2	58.9	67.9	48.9	86.2	62.1	73.1	53.0

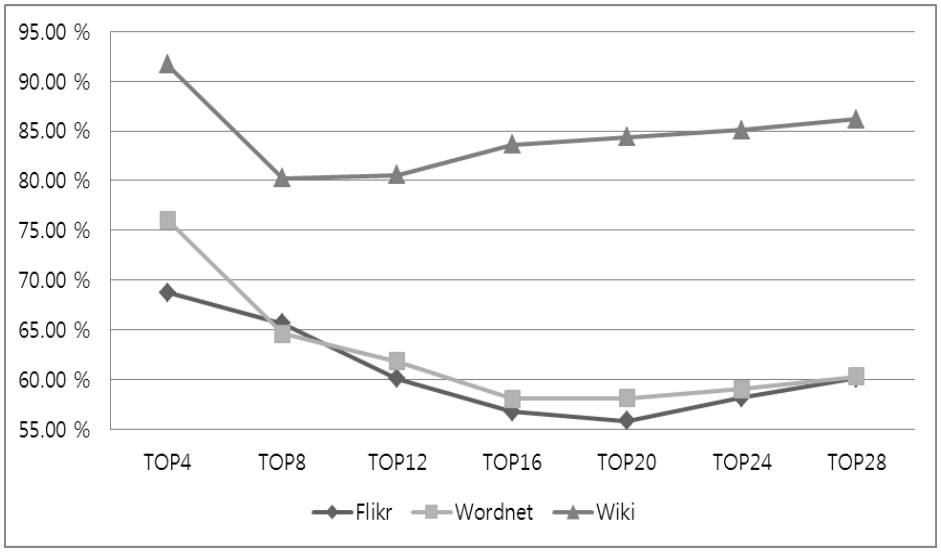


Figure 2. Precisions of 'house' images

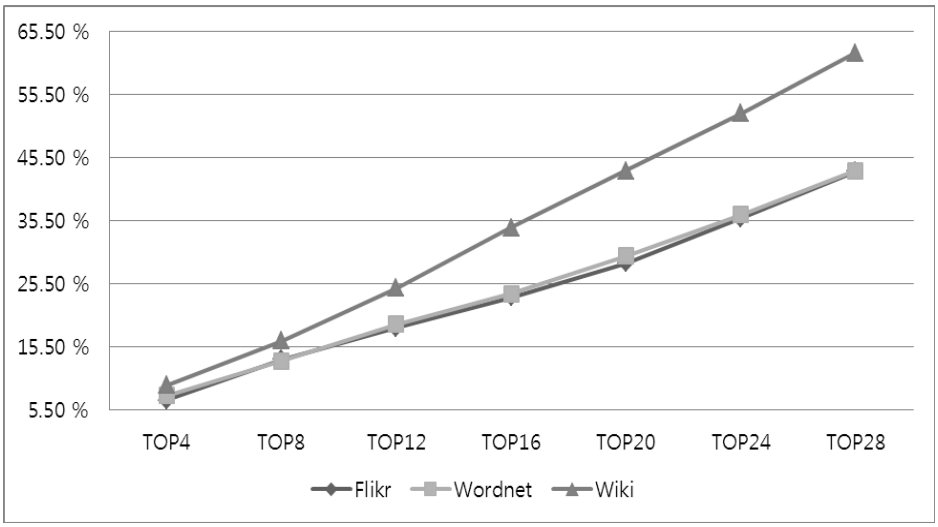


Figure 3. Recalls of 'house' images

3. Conclusion

In Figure 1 and Figure 2, we can also see the graphs indicating Wikipedia-based method as the winner with 'house' images. In tagged Web image retrieval, users prefer finding desired images as soon as possible. If more desired images are being displayed from front pages like upper ranked 12 pages etc., user's satisfaction can be improved. Therefore, re-ranking of retrieved images using Wikipedia-based semantic relatedness suggested in this article can be one of effective methods.

Acknowledgements

This research was financially supported by the Ministry of Education, Science and Technology (MEST) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation. This work was supported by National Research Foundation grant funded by the Korea government (MEST) (No. 2011-0005288).

References

[1] G. A. Miller, "WordNet: An On-line Lexical Database", International Journal of Lexicography. 3, 4 **(1990)**.

[2] Wikipedia, http://www.wikipedia.org.

[3] Flickr, http://www.flickr.com.

[4] D. Kwon, D. Hong and S. Cho, "Web Image Retrieval using prior Tags based on WordNet Semantic Information", Journal of Korea Multimedia Society. 12, 7 **(2009)**.

[5] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis", Joint Conference on Artificial Intelligence, **(2007)** January 6-12; Hyderabad, India.

[6] D. Milne and I. H. Witten, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links", Association for the Advancement of Artificial Intelligence: Workshop on Wikipedia and Artificial Intelligence, **(2008)** July 13-14; Chicago, USA.

[7] Flickr App Garden, http://www.flickr.com/services/api/.

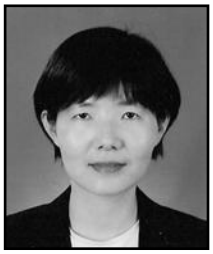
Authors

Seongjae Lee

2011 Chungju National University, Dept. of Computer Science and Information Engineering (B.E.)

2011~ Now KNUT, Dept. of Computer Science and Information Engineering (M.S. Candidate)

※ Research Area : Data Mining, Mobile Web, HTML5



Soosun Cho

1987 Seoul National University, Dept. of Computer Science and Statistics (B.S.)

1989 Seoul National University, Dept. of Computer Science and Statistics (M.S.)

2004 Chungnam National University, Dept. of Computer Science (Ph.D)

2004~ Now KNUT, Dept. of Computer Science and Information Engineering (Professor)