International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.9 (2015), pp.353-362
http://dx.doi.org/10.14257/ijmue.2015.10.9.36

Comparison and Evaluation of Highly Related Tag-Pairs Extraction Methods

DaeHoon Hwang

Dept. of Computer Engineering, Gachon University
hwangdh@gachon.ac.kr

Abstract

Until recently, it was expected that the web users' tagging would be used in classification, recommendation, and information search. However, the users' subjective decision or infrequent tags with high interrelationship have caused inaccurate results.

This research proposes an algorithm to disregard users' subjective tags and utilize infrequent tags with high semantic similarity using WordNet. Furthermore, to enhance the efficiency of the proposed algorithm, a weighted matrix is proposed.

To evaluate the proposed algorithm, the current co-appearing frequency between tag-pairs method, tag-pair semantic similarity extraction algorithm, and tag-pair weight matrix method were analyzed and compared.

Keywords: *tag, tag-pair, co-appearing frequency, semantic similarity*

1. Introduction

Currently, many internet users expect their tags be reused to improve efficiency of categorization, recommendation and searching however, inaccurate results can be obtained in such case if their subjective tags are tagged or highly related and infrequent tags are ignored.

Several methods such as collaborative tagging [1], hierarchical structure of tags [2], tag-based searches and ranking [3-4], and tag-based recommendation [5] have been developed. However, most of the existing tag-based methods use only the tag co-appearing frequency to find the tag-pair semantic similarity; hence, their result of is inaccurate in the translation.

To solve this problem, this research proposes a tag-pair semantic similarity extraction (TSSE) algorithm that uses WordNet [6], and computes tag-pair frequency matrix (TFM) to enhance its efficiency.

This research uses Flickr's Open API [7] to test the proposed method. The collected tag information of the top 500 images that has 'tomato' as the keyword are used to analyze and compare the existing frequency-based method and our proposed research.

2. Highly Related Tag-Pairs

2.1. Tag-pair Weight Matrix Creation System

Figure 1 shows a block diagram of the tag-pair weight matrix creation system using tag similarity suggested by this research.

The system suggested in this research consists of 3 modules: tag frequency extraction module (TFEM), semantic similarity extraction module (SSEM) and tag-pair weight matrix creation module (TWMCM).

Of these, TEEM was carried out in the preceding research [4], and SSEM and TWMCM are suggested in this research.

ISSN: 1975-0080 IJMUE

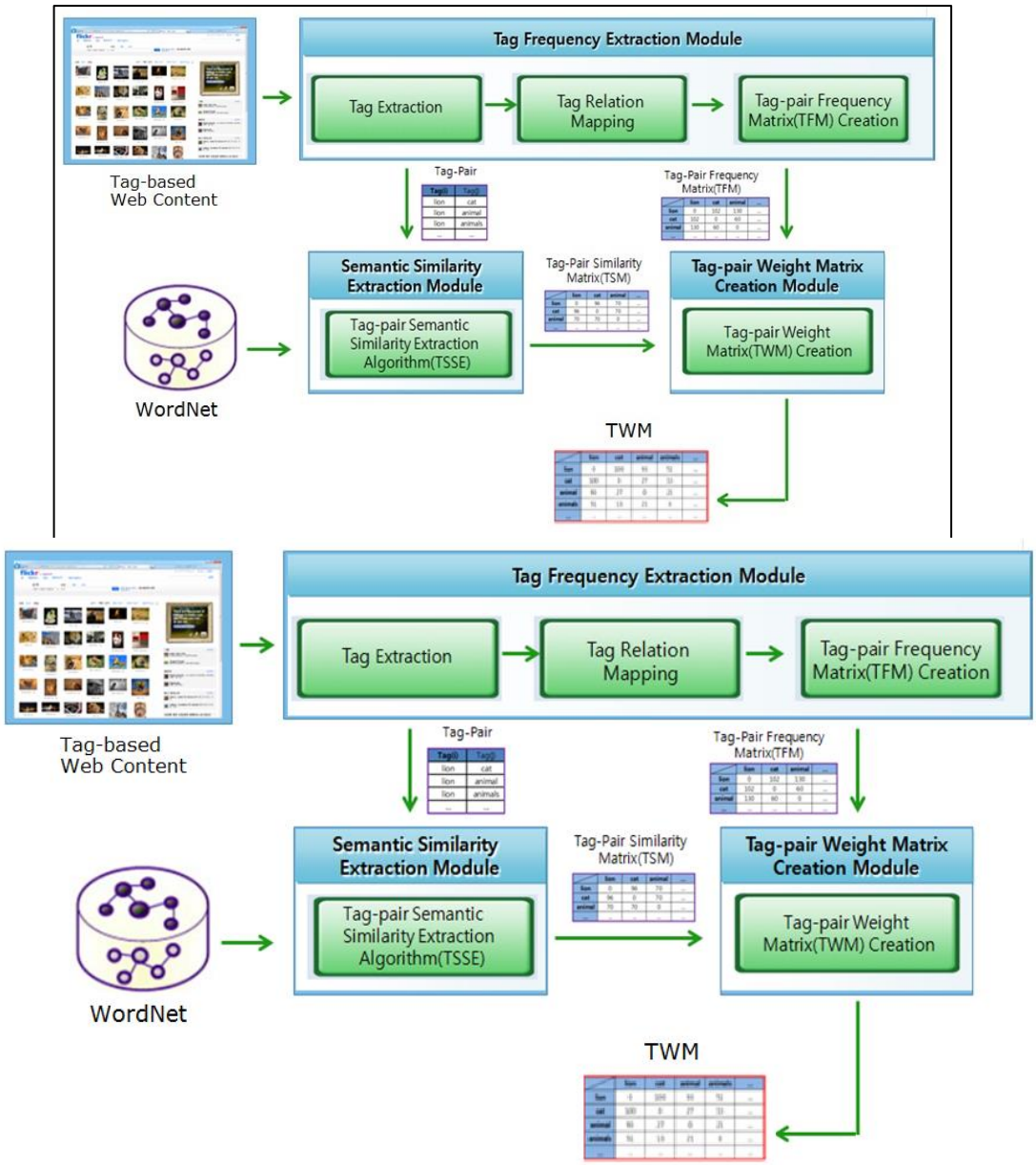

Figure 1. Tag-Pair Weight Matrix Creation system

2.2. Tag Frequency Extraction

The tag frequency extraction module is the first step to extract similarity between tags which are tagged on contents. This module extracts tag co-appearing frequency and generate a tag-pair frequency matrix. To perform the above purpose, this module is comprised of 3 steps.

Step 1: Tag Extraction

Tag extraction module collects contents and tag information from tag based site such as flickr, delicious, and buzzillions using Open API.

Step 2 : Tag Relation Mapping

Tag relation mapping module performs tag mapping according to the relation of different tags which are tagged on the contents collected in the Step 1. The tags ‘lion’, ‘cat’, ‘nature’, ‘africa’, and ‘animals’ of Figure 2’s image A are the tags that the users

tagged after seeing the image A. These tags have a certain relation each other, and Figure 2 represents the relation graph between these tags.

Co-appearing tags means that different tags are tagged simultaneously on an image of a same subject, and these tags have a high relation each other.

For example, 'animal' tag and 'lion' tag are tagged simultaneously several times, and have high relation each other. Figure 2 shows the extraction process of co-appearing tag 'lion' and 'nature' by mapping the relation tags of A image and B image.

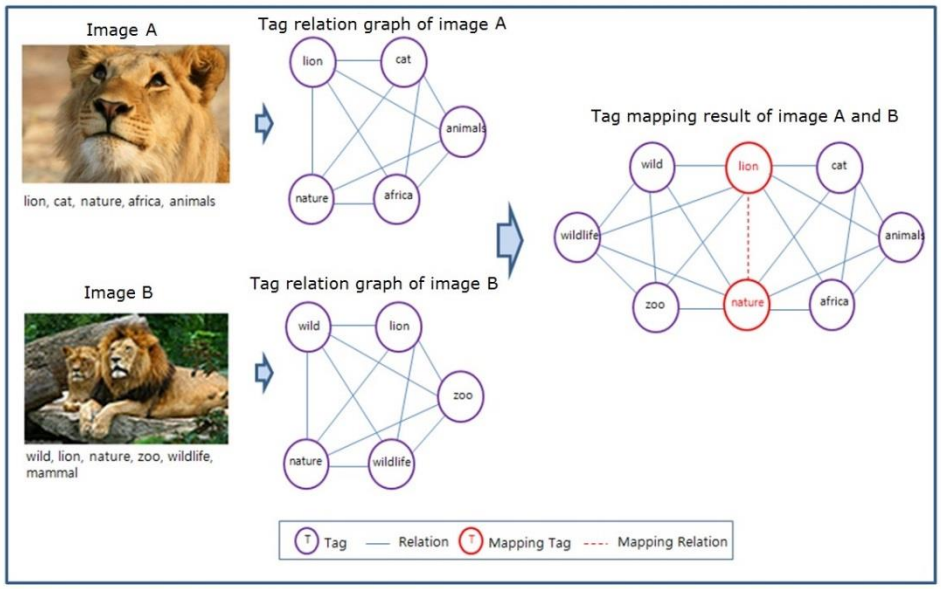

Figure 2. Example of Tag Relation Mapping

(1) Step 3 : TFM Creation

TFM creation module create the TFM matrix using tag co-appearing frequency which is extracted at the tag relation mapping process of Step 2. The method to create TFM is defined as equation (1) and (2). The ***TM(i,j)*** of equation (1) and (2) is the ***ij*** element of tag graph's weight matrix, and means o-appearing or not.

TM(i,j)=0: tag *i* and tag *j* do not appear simultaneously on a certain contents (1)

TM(i,j)=1: tag *i* and tag *j* appear simultaneously on a certain contents (2)

Equation (1) is used when tag *i* and tag *j* do not appear simultaneously on a certain content or image, and means that the two tags do not have a relation each other. Equation (2) is used when two tags appear simultaneously, and the two tags have a relation each other.

$$TFM(i,j) = \sum_{k=1}^{m} TM_k(i,j) \quad (3)$$

The TFM matrix of equation (3) means the co-appearing frequency which tag *i* and tag *j* appear simultaneously on *m* images, and is computed by summation of tag-pair matrix $TM_k(i,j)$ of each image.

Figure 3 shows the creation process of TFM matrix.

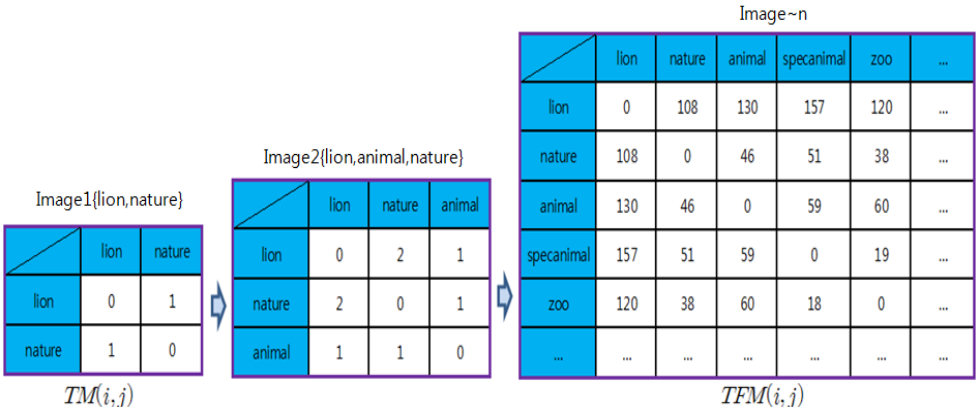

Figure 3. Creation Process of TFM Matrix

2.3. Tag-Pair Semantic Similarity using WordNet

The original tag-based search system uses only co-appearing frequency to extract the similarity between tags. The following are a few problems that exist in these methods:

First, users' subjective tags create inaccurate tags.

Second, unorganized tags are the main factors for inefficient information search, because users do not consider the relationship and priority order between tags.

To correct the subjective and infrequent tags that have highly related meanings, WordNet is applied to extract tag-pair semantic similarity. There are two categories for the methods that use WordNet. One method is the information content based method [8-9], which uses the minimum distance between words, and the other is a node based method [10-11].

Lin [8] and Resink [9] used the tag-pair's semantic similarity which is based on information content. These methods combine the semantic information extracted from the original corpus and concept information obtained from the WordNet.

2.4. TSSE Algorithm

Lin [8] and Resink [9] extracted tag-pair's semantic similarity by including the similarity and difference at the same time; however because of ambiguity of words, directly-opposed problems occur between some words.

The TSSE algorithm with WordNet is suggested in this research to overcome the existing tag-based research problems.

// *c* : concept, which conceptually includes certain tag-pair
// *word(c)* : all synsets, which exist in information content file of concept c
// *count(c)* : frequency of synset elements, which belong to *word(c)* words
// *freq(c)* : frequency summation of all synset elements
// *Pr(c)* : concept probability
// *N* : number of nouns among the *word(c)* words
// *IC(c)* : information content
// $S(c_1, c_2)$: set of concept, which conceptually includes both c_1 and c_2
// $Weight : 0 < Weight < 1, Weight_1 + Weight_2 = 1$
// *TS* : tag-pair's similarity, *i.e.*, semantic similarity between tag pair

Find *word(c)* by concept *c*
for (i=1: $i \leq |word(c)|$; i ++) {
$freq(c) = freq(c) + count(i)$
}
$Pr(c) = freq(c) / N$
$IC(c) = 1 / \log Pr(c)$

$$Sim_1(c_1, c_2) = \max_{c \in S(c_1, c_2)} IC(c)$$

Compute $IC(c_1)$ and $IC(c_2)$
$Sim_2(c_1, c_2) = 2 * IC(S(c_1, c_2)) / (IC(c_1) + IC(c_2))$
$TS = \{Sim_1 \times Weight_1 + Sim_2 \times Weight_2 \} / 2$

Figure 4. TSSE Algorithm

2.5. Tag-Pair Weight Matrix

To extract semantic similarity of a tag-pair using WordNet, we require the tag-pair words prior to extraction. Tag co-appearing frequency, which is an existing tag similarity extraction method, has an advantage over other method that users can easily extract the tags correlated to a specific keyword, on the basis of tagged tags. However, it also has a disadvantage that the user-subjective tags account for high frequency. Further, another problem with this method is that highly related tags with low frequency between tag-pair are not utilized.

This research proposes a tag-pair weight matrix (TWM) using equation (4). The matrix is proposed in order to retain all advantages of WordNet's semantic similarity and tag co-appearing frequency and overcome the disadvantages.

$$TWM(i,j) = TSM(i,j) \times TFM(i,j) \quad (4)$$

TSM (i,j) is the *ij*th entry of tag-pair similarity matrix, and *TFM(i,j)* is the *ij*th entry of the tag co-appearing frequency matrix.

3. Experiment and Analysis

To test the proposed method, the Flickr's Open API [7] was used. Top 500 images including the tag 'tomato' and tag information are collected from Flickr, and 11,088 tags are collected from these images.

3.1. Extraction of Tag-Pair Co-Appearing Frequency

The tag-pair co-appearing frequency extraction method [6] was used to extract the related tag-pairs. Further, 15,867 related tag-pairs were extracted from the 'tomato' keyword. Figure 4 shows the top 20 tag-pairs and their co-appearing frequency.

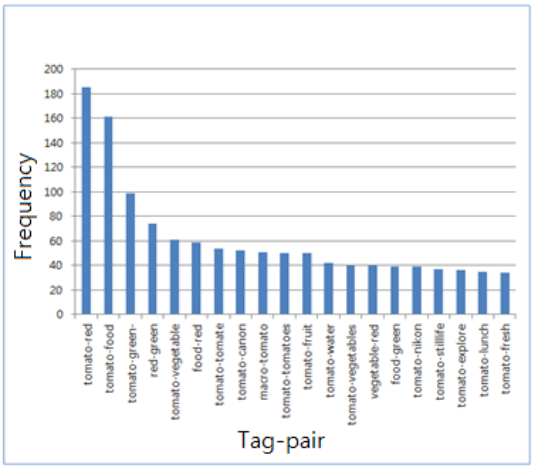

Figure 5. Top 20 Tag-Pairs using the 'Tomato' Keyword

In Figure 5, the most frequent tag-pair is 'tomato-red' and occurs 183 times. Furthermore, tags such as 'food', 'green' and 'vegetable' had high co-appearing frequency. The tags that frequently co-appeared in the 500 'tomato' images were highly related tags with the 'tomato' keyword, and the low-frequency tags were slightly related tags.

The 'tomato-green' or 'tomato-canon' tag-pair had high co-appearing frequency; however, the tag-pair had low semantic similarity and users' subjective tags.

3.2. Adaptation of TSSE

In tag-pair co-appearing frequency method, there was a drawback that users' subjective tags might appear in the top tag-pairs and highly related infrequent tags might not be selected. To disregard users' subjective tags and utilize infrequent tags that had highly related meanings, the TSSE algorithm was applied using WordNet.

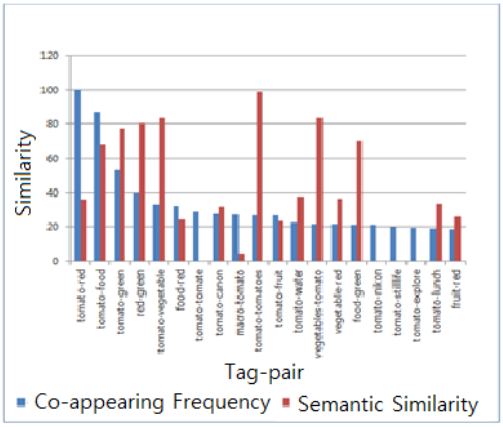

Figure 6. Comparison of the co-appearing frequency and semantic similarity

The red bar in Figure 6 shows the top 20 tag-pair's semantic similarity and the blue bar

shows the co-appearing frequency. A tag-pair like ‘tomato-vegetable’ or ‘tomato-tomatoes’ had high semantic similarity with low co-appearing frequency. The TSSE algorithm is a suitable method to find the semantic similarity; however, it cannot be used independently. If two methods were integrated, the co-appearing frequency method could be complementary to the semantic similarity method.

3.3. TWM Matrix with Weight

To solve the problem in Section 3.2, the TWM generation method which is proposed in Section 2.3 was adapted for the ‘tomato’ keyword. Therefore, the TWM matrix with 3,786 tags was generated. Figure 7 shows a part of the final TWM matrix.

	tomato	food	green	red	vegetable	tomatoes	vegetables	cheese	yellow	pepper	basil	white	orange	canon	water	beans
tomato	0	100	69	61	46	45	30	19	2	16	15	10	15	15	14	14
food	100	0	25	13	19	12	12	13	1	0	7	2	6	6	6	10
green	39	25	0	54	19	15	10	3	14	9	3	8	6	5	6	0
red	61	13	54	0	13	11	7	2	18	5	2	15	11	6	14	1
vegetable	46	19	19	13	0	10	7	0	1	1	1	3	5	1	3	0
tomatoes	45	12	15	11	10	0	7	2	0	3	3	2	4	2	2	0
vegetables	30	12	10	7	7	7	0	1	0	2	2	1	3	2	1	0
cheese	19	13	3	2	0	2	1	0	0	2	3	0	0	1	0	0
yellow	2	1	14	18	1	0	0	0	0	1	0	3	8	1	3	0
pepper	16	0	9	5	1	3	2	2	1	0	2	1	1	1	0	0
basil	15	7	3	2	1	3	2	3	0	2	0	1	0	1	0	0
white	10	2	8	15	3	2	1	0	3	1	1	0	1	2	3	0
orange	15	6	6	11	5	4	3	0	8	1	0	1	0	2	0	0
canon	15	6	5	6	1	2	2	1	1	1	1	2	2	0	2	2
water	14	6	6	14	3	2	1	0	3	0	0	3	0	2	0	0
beans	14	10	0	1	0	0	0	0	0	0	0	0	0	2	0	0
bacon	13	11	0	0	0	0	0	0	0	0	0	0	0	2	0	8
onion	13	6	3	1	2	0	3	2	0	3	1	0	1	1	0	0
bread	13	12	3	1	0	1	0	4	1	1	0	0	0	1	0	1
color	9	5	10	13	3	1	1	0	4	1	1	3	3	1	2	0

Figure 7. TWM Matrix for the ‘Tomato’ Keyword

Figure 8 shows the top 20 tag-pairs computed using equation (4). In this figure, users’ subjective tag-pairs or error-tags such as ‘food-red’ and ‘tomato-tomate’ were disregarded, and tag-pairs such as ‘green-vegetable’ and ‘food-vegetable' were utilized.

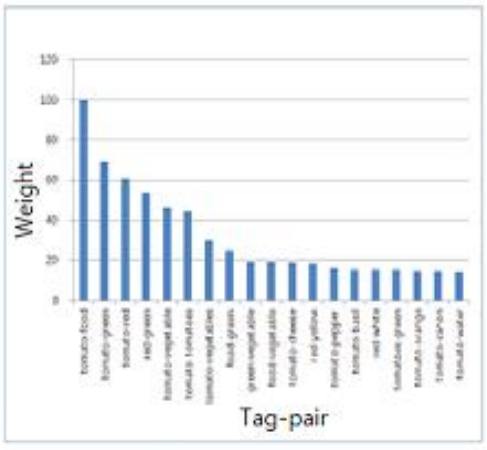

Figure 8. Top 20 Tag-Pairs with Weight Value

The utilized tag-pairs had low co-appearing frequency; however, they were highly related tag-pairs that had high weight value and ranked in the top 20 tag-pairs.

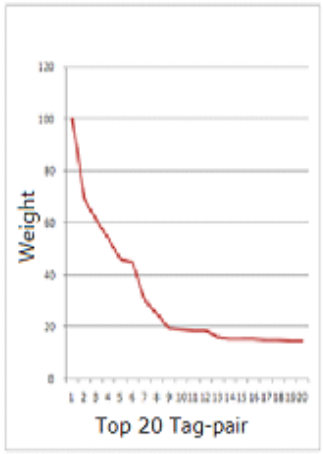

Figure 9. Weight Distribution of the Top 20 Tag-Pairs

Figure 10 shows the tag co-appearing frequency based on the sorted weight value. The x-axis shows the weight value that is sorted in descending order and the numbers on the x-axis are the typical tag-pair's ID number derived from the TWM. The y-axis shows the co-appearing frequency of each tag-pair.

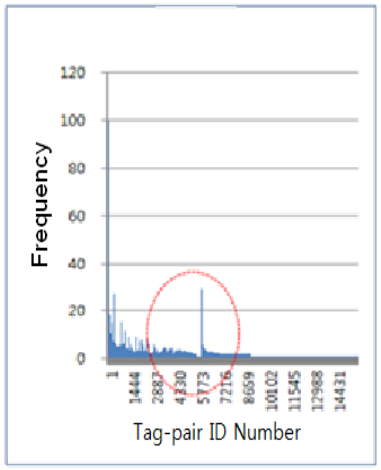

Figure 10. Tag-Pair Co-Appearing Frequency Based on the Weight Value

In Figure 10, the red circle shows the tag-pairs that had high co-appearing frequency with low tag-pair weight value. These tag-pairs were disregarded in our research.

4. Conclusion

In this study, the TSSE algorithm using WordNet and TWM matrix were proposed to overcome the problems that exist in the current tag-based methods and enhance the efficiency of these algorithms. In the experiment, the user-subjective tags were disregarded, and highly related tag-pairs with low co-appearing frequency were utilized.

Acknowledgements

This work is supported by the Gachon University research fund for 2015.

References

[1] H. Halpin and H. Shepherd, "The complex dynamics of collaborative tagging", 16th international conference on World Wide Web, **(2007)**, pp. 211-220.
[2] Brooks C. H. and N. Montanez, "Improved annotation of the blogosphere via autotagging and hierarchical clustering", 15th international conference on World Wide Web, **(2006)**, pp. 625-632.
[3] S. J. Park, S. Lee and D. H. Hwang, "Web Contents Ranking System using Associative Tag and Similar User Weight", Journal of Multimedia Society, vol. 14, no.4, **(2011)**, pp. 567-576.
[4] M. H. Lee, "Multi Tag-based Search Technique for Efficient Web Contents Search in Web 2.0", PhD. thesis, Kyungwon University, **(2012)**.
[5] I. Cantador and D. Vallet, "Content-based Recommendation in Social Tagging Systems", Proceedings of the fourth ACM conference, **(2010)**, pp. 237-240.
[6] WordNet 3.1, "WordNet, a lexical database for the English language", http://wordnet.princeton.edu/, **(2012)**.
[7] Flickr, http://www.flickr.com
[8] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", International Joint Conference on Artificial Intelligence, **(1995)**, pp. 448-453.
[9] D. Lin, "An information-theoretic definition of similarity", International Conference on Machine Learning, **(1998)**, pp.296-304.
[10] Wu Z. and Palmer M. "Verb semantics and lexical selection In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics", **(1994)**, pp.133-138.
[11] Leacock C. and Chodorow M. "Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database", **(1998)**, pp.265-283.

Author

DaeHoon Hwang,Professor, Gachon University Dept. of Computer Science Interest Area: Multimedia, Big Data, Digital Contents e-mail: hwangdh@gachon.ac.kr.