International Journal of Software Engineering and Its Applications
Vol.8, No.2 (2014), pp.229-234
http://dx.doi.org/10.14257/ijseia.2014.8.2.22

Development of Tag Ranking-Based Image Search System

Si-Hwa Lee[1], Sae-Hong Cho[2] and Dae-Hoon Hwang[3]

[1]*Medipia Tech. Inc., Ltd.*
[2]*Dept. of Multimedia Engineering, at Hansung University*
389 Samsun-Dong 3-Ga, Sungbuk-Gu, Seoul, Korea
[3]*Dept. of Computer Science, at Gachon University*
1342 Sungnamdaero Sujung-gu, Sungnam-Si, Kyunggi-Do, Korea
leesihwa@gmail.com, chosh@hansung.ac.kr, hwangdh@gachon.ac.kr

Abstract

The existing tag-based systems offer search results with low accuracy, as they provide search results through a single tag matching by using content-tagged tags. Because users do not consider correlations and priorities between tags upon content tagging, the content and relevant information that the tags have are not efficiently offered. In this context, this study suggests a cluster-based tag ranking and search system using tag similarity to solve the problems mentioned above. Through such a system, the existing single tag matching problem can be solved.

Keywords: *Tag, Tag Similarity, WordNet, Tag Ranking*

1. Introduction

Tagging currently makes a great appeal to many Internet users and widely applies from Web documents like blog to multimedia data such as image and video [1]. However, unsatisfactory results are demonstrated in reality, due to limitations of tag, unlike an expectation that the tags used for tagging may maximize efficiency through their being reused for information classification, recommendation and search [2-4]. Such results are ascribed to low search results, derived from inaccurate tags, and inefficient information navigation, owing to unstructured tags. This paper designs a cluster-based tag raking and search system using tag similarity to solve those problems.

Major modules consisting the system are as follows:

First, this paper suggests tag-pair weight matrix (TWM)-based tag cluster (TBTC) algorithm to solve the of low search results, arising from inaccurate tags, which can be the first problem of the existing tag-based systems.

Second, this paper suggests cluster-based tag ranking (CBTR) algorithm to solve the problem of inefficient navigation, due to unstructured tags, which can be the second problem, based on clustering results.

Third, this paper suggests tag-pair semantic similarity extraction (TSSE) algorithm to solve the problem of simultaneous appearance frequency of tags, owing to the tags that a user subjectively assigned and the tags with low frequency but high semantic relations between tags.

2. Existing Systems and Related Studies

Although, a tag is very useful for positioning certain information to wide range of

ISSN: 1738-9984 IJSEIA

categories, it shows an inefficient result in searching for accurate information that a user wants.

The existing tag-based search systems[2, 3, 4] have the following problems:

Firstly, a problem of search results having low accuracy caused by inaccurate tags: Search results can be inaccurate as well, when inaccurate tags are tagged to content.

Secondly, a problem of inefficient information navigation, due to unstructured tags: The information related to content that the tags have is not offered to users efficiently.

The causes can be low search results, derived from inaccurate tags and inefficient information navigation, due to unstructured tags [2-4]. In the tags tagged by users, considerable number of tags with users' subjective tendency exist by their intuitive judgment, and some of the tags have high frequency. In the existing studies that extract and use tags having high frequency, the following problem is demonstrated: the tags are defined having high correlations by user's subjectivity. And, the tags having low simultaneous appearance frequency but high semantic relations between tags are defined without correlations between tags, due to low frequency.

Many studies including collaboration tagging [5], tag clustering [6, 7], tag hierarchy creation [8] and tag-based search have been conducted to solve the problems of the existing tag-based search systems.

3. Cluster-based Tag Ranking and Search System

Figure 1 shows a block diagram of the cluster-based tag ranking and search system using tag similarity suggested by this study.

The system suggested in this study consists of five modules: Tag Frequency Extraction Module (TFEM), Semantic Similarity Extraction Module (SSEM), Tag-pair Weight Matrix Creation Module (TWMCM), Tag Clustering Module (TCM) and Tag Ranking Module (TRM).

Of these, TFEM and TCM were carried out in the preceding study [10] and this study conducted research on tag-pair weight matrix creation technique through TSSE and on search utilization through TRM.

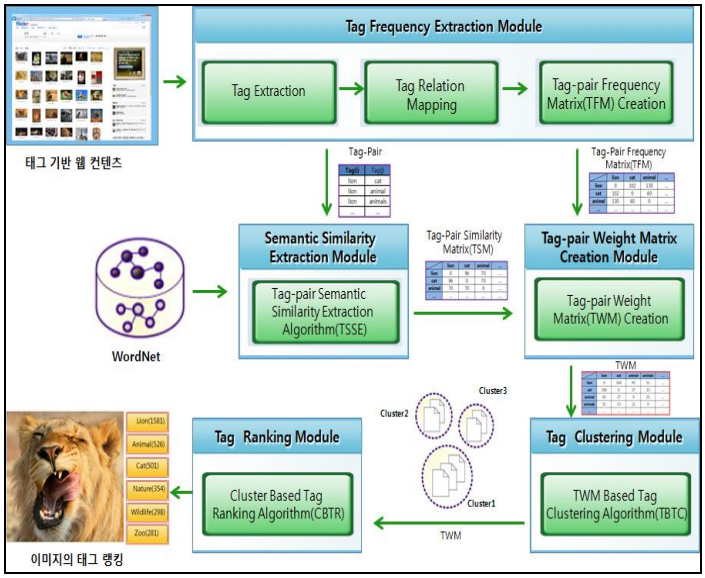

Figure 1. Cluster-based Tag Ranking and Search System

3.1. Tag Frequency Extraction

The tag frequency extraction module (TFEM) is the first method to extract similarity between content-tagged tags. TFEM plays a role in extracting tag simultaneous appearance frequency and creating tag-pair frequency matrix (TFM) and was conducted in the preceding study [12]. The TFM creation formula is presented below:

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

3.2. Semantic Similarity Extraction

The TSSE algorithm using the WordNet suggested by this paper is to overcome existing tag-based studiers' problems and Figure 2 shows the proposed TSSE algorithm.

// c: Conceptually included concept of a specific tag pair.
// $word(c)$: All synsets in the information content file of concept c
// $count(n)$: Frequency of each synset factor belonging to $word(c)$
// $freq(c)$: Sum of all synset factors' frequencies
// $\Pr(c)$: Concept probability
// N: Total number of nouns of $word(c)$ words
// IC: Information Content
// $S(c_1, c_2)$: Collection of concepts containing concept c_1 and c_2 conceptually
//Weight : $Weight$ ($0 < Weight < 1$, $Weight_1 + Weight_2 = 1$)
//TS: Tag Similarity, that is, Semantic Similarity between Tags
Find $word(c)$ by Concept c
for ($i = 1$ to $|word(c)|$) **{**

$freq(c) = freq(c) + count(i)$

}
$\Pr(c) = freq(c)/N$
$IC(c) = 1/\log Pr(c)$
$Sim_1(c_1, c_2) = \max_{c \in S(c_1, c_2)} IC(c)$
Compute $IC(c_1)$ and $IC(c_2)$
$Sim_2(c1, c2) = 2 * IC(S(c_1, c_2)) / IC(c_1) + IC(c_2)$
$TS = \{ Sim_{1X} Weight_1 + Sim_{2X} Weight_2 \} / 2$

Figure 2. TSSE Algorithm

3.3. Tag-Pair Weight Matrix Creation

Tag simultaneous appearance frequency, which is an existing tag similarity extraction method, has a merit that users can easily extract the tags correlated to a specific keyword, based on tagged tags. But, it has a demerit that user-subjective tags account for high

frequency and that there is a problem that high semantic tags having low frequency between tags are not utilized.

This study created tag-pair weight matrix (TWM), which is a similarity measure between final tags by applying formula (2) in order to solve the similarity problem between two tags.

$$TWM(i,j) = TSM(i,j) \times TFM(i,j) \quad (2)$$

3.4. TWM-based Tag Clustering

Tag clustering module (TCM) plays a role in clustering the tags having high correlation by applying TBTC (TWM-based tag clustering) algorithm. This paper proved the accuracy of the tag clustering algorithm suggested in the preceding study [10]. Figure 3 shows a short summary of TBTC algorithm.

```
// k: Cluster Number
// C_k: kth Cluster
// TWM(i,j):
// Max(i,j):
// CWM_k: (Cluster Weight Matrix)
k=1
Initialize CWM_k
Repeat {
Select Max(i,j)
Add tag i and tag j to C_k
Add element Max(i,j) to CWM_k
While( TWM(i,j) ≥ θ) {
Add tag i and tag j of TWM to C_k
Add element of TWM(i,j) to CWM_k
}
k= k+ 1
} until (all tags of TWM which larger than θ are included in C_k)
```

Figure 3. TBTC Algorithm

3.5. TBTC-based Ranking Algorithm

This section suggests CBTR (cluster-based tag raking) algorithm to solve the existing tag-based systems' problem of inefficient information navigation, due to unstructured tags.

The tag and weight values ranked by this algorithm are all different according to image and are defined as the search value of the concerned image. The defined search value becomes the measure to be searched upon search request. Figure 4 shows CBTR algorithm.


```
// m: Number of Cluster
// C_m: mth Cluster
// p: Number of Tag for mth Cluster
// CWM_m: p × p Weighted Matrix for mth Cluster
// SCWM_m(i):
// t: Number pf image
// I_t: tth Image
// q: Number of Tag for tth Image
// TM_t: q × 2 Tag Matrix for tth image
// CWM_m
for(i =1 to p) {
SCWM_m(i) = 0
for(j =1 to p) {
SCWM_m(i) = SCWM_m(i) + CWM_m(i,j)
}

}
for(i =1 to t) {
for(j =1 to q) {
TM_t(j,2) = 0// TM_t 을 초기화
for(k =1 to p) {
if(TM_t(j,1) ==CWM_m) {
TM_t(j,2) = SCWM_m(k)
break;
}
}
}
//
Descending sort TM_tby 2nd column
}
```

Figure 4. CBTR Algorithm

4. Conclusion

In this paper, we developed a cluster-based tag ranking and search system using tag similarity in order to solve the problems which the currently existing tag-based system and algorithms, and suggested three algorithms: TBTC algorithm, CBTR algorithm, and TSSE algorithm. TWM-based TBTC algorithm was designed to solve a insufficient search result problem due to incorrect tag. CBTR algorithm was designed to solve inefficient information navigation problem due to unstructured tag. And, the tag and weight values ranked by TSSE algorithm are all different according to image and are defined as the search value of the concerned image.

Acknowledgements

This work was supported by the Gachon University research fund of 2014.”(GCU-2014-R029).

References

[1] S. H. Lee and Y. M. Roh, "Technology Trend for Automatic Image Tagging", National IT Information Promotion Agency, Weekly Technology Report, no. 1427, **(2010)**, pp. 1-8.
[2] Flickr, http://www.flickr.com.
[3] del.icio.us, http://del.icio.us/.
[4] Technorati. http://technorati.com/.
[5] J. Ma, "The sustainability and stabilization of tag vocabulary in CiteULike: An empirical study of collaborative tagging", online information review, vol. 36, no. 5, **(2012)**, pp. 655-674.
[6] M. Leginus and F. Durao, "Methodologies for Improved Tag Cloud Genera-tion with Clustering", Lecture Notes in Computer Science, no. 7387, **(2012)**, pp. 61-75.
[7] C. Lu and J. Park, "Exploiting the Social Tagging Network for Web Clus-tering", IEEE Transactions on Systems, Man and Cybernetics, vol. 41, no. 5, **(2011)**, pp. 840-852.
[8] X. Li and D. Tang, "nducing Taxonomy from Tags: An Agglomerative Hier-archical Clustering Framework", Lecture Notes in Computer Science, no. 7713, **(2012)**, pp. 64-77.
[9] M. H. Lee, "Multi Tag-based Search Technique for Efficient Web Contents Search in Web 2.0", Ph.D. Thesis, Kyungwon University, **(2012)**.

Authors

Si-Hwa Lee
Media-Tech. Inc., Ltd.
Interest Area: Web2.0 e-Learning, Tag-Clustering
e-mail: leesihwa@gmail.com

Sae-Hong Cho
Professor, Hansung University
Dept. of Multimedia Engineering
Interest Area: Multimedia, Virtual Reality, Big Data, Digital Contents
e-mail: chosh@hansung.ac.kr

Dae-Hoon Hwang (Corresponding Author)
Professor, Gachon University
Dept. of Computer Media
Interest Area: Multimedia, Big Data, Virtual Reality, Digital Contents
e-mail: hwangdh@gachon.ac.kr

