A Study on the Image Search System using Tag Ranking

Si-Hwa Lee*, Sae-Hong Cho**, Dae-Hoon Hwang***

*Medipia Tech. Inc., Ltd.
leesihwa@gmail.com

**Dept. of Multimedia Engineering, at Hansung University
389 Samsun-Dong 3-Ga, Sungbuk-Gu, Seoul, Korea
choshi@hansung.ac.kr

***Dept. of Computer Science, at Gachon University
1342 Sungnamdaero Sujung-gu, Sungnam-Si, Kyunggi-Do, Korea
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, 3, 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 ranking and search system using tag similarity to solve those problems. Major modules which consists a system are as follows:

First, this paper suggests tag-pair weight matrix (TWM)-based tag cluster (TBTC) algorithm to solve the cause 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 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[2,3,4]. In the tags tagged by users, considerable number of tags with users' subjective tendency exist by their intuitive judgement, 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 pilot study and this study conducted research on tag-pair weight matrix creation technique through TSSE and on search utilization through TRM.
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 pilot study[12]. The TFM creation formula is presented below:

$$TFM(i, j) = \sum_{k=1}^{m} TFM_k(i, j)$$  \hspace{1cm} (1)

3.2 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.

$$TWM(i, j) = TSM(i, j) \times TFM(i, j)$$  \hspace{1cm} (2)
3.3 TWM-based Tag Clustering & TBTC-based Ranking Algorithm

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 pilot study.

CBTR (cluster-based tag ranking) algorithm is suggested 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.

4 Conclusion

In this paper, we developed an algorithm for a cluster-based tag ranking and search system in order to solve the problems which the currently existing tag-based system and algorithms

References

3. del.icio.us, http://del.icio.us/