A Combination Indexing for Image Social Bookmarking System to Improve Search Results

Pijitra Jomsri
Department of Information Technology, Faculty of Science and Technology
Suan Sunandha Rajabhat University
Thailand
pijitra.jo@ssru.ac.th

Journal of Digital Information Management

ABSTRACT: *Web 3.0 and social bookmarking have altered the traditional roles of the indexer and user. Recently, web, allows users to create, organize, and search for images and other information sources through social tagging and other method activities. One of the image social bookmarking is such as Flickr. This research examines to increase the efficiency of image search result by creating indexes. The assumption of the experiment is a combination of social tagging and other factor such as image description that can improve web performance. Therefore, three indexers were created and were compared with the search results between a search engine using the indexing method of "Description with Tag" or DT indexer, "Posted Time With Tag" or PT indexer and the "Tag only" or T indexer which is a native method. The retrieval performance of this search engine is evaluated using the mean values of Normalized Discount Cumulative Gain (NDCG). The result illustrates that the search engine with indexer using DT indexer provides better indexer using the T indexer in all the ranks. This primary evaluation in the experiments proves that the chosen heuristic indexer can improve the efficiency of the web image searching on social bookmarking websites.*

Subject Categories and Descriptors
H.3.1 [Content Analysis and Indexing]; I.4 [Image Processing and Computer Vision]; I.4.10 Image Representation

General Terms
Image Processing and Indexing.

Keywords: Image Retrieval, Image Indexing, Joint Indexing

Received: 19 January 2016, Revised 1 March 2016, Accepted 19 March 2016

1. Introduction

Nowadays, the Internet and web technology are increasingly used rapidly; especially the visible feature is the voluminous amount of digital image information produced every year. It is a rough estimate that there are over ten billion images are hosted in the World Wide Web. In addition, most of the images in social bookmarking are created and published by individuals. This estimate excludes those in stock or publicly inaccessible personal photo collections. However, a well-conceived technique can help users for image searching. A small fraction of these images only is indexed for effective and efficient retrieval. Social/collaborative tagging and photo sharing services such as Flickr (http://www.flickr.com) pose new challenges and present tremendous opportunities for designers of indexing and retrieval systems while demanding dynamic sets of solutions [17]. Flickr is almost certainly the best online photo management and sharing application in the world. On Flickr, members upload photos,

share them securely, supplement their photos with metadata like license information, geo-location, people, tags, etc., and interact with their family, friends, contacts or anyone in the community [1]. It was developed by the Ludicrop Company by Caterina Fake and Stewart Butterfield in 2004. In the beginning, Flickr pointing in a chat room shared more photos. After this development, Flickr received more popularity and acquired by yahoo in 2005. Additionally, Flickr became a popular website for users to share and keep personal photographs.

In the community of image searching, users are creators and indexers of multimedia information sources through activities such as tagging. Moreover, users are becoming accustomed to this level of involvement, the problem of how to incorporate user-generated tags and other factors in the process of indexing and retrieval needs urgent attention.

In this paper, we use social tagging to improve the image indexing. The premise is that in social image bookmarking, only social tagging may not be enough to represent the pictures of user interests. An indexing method is proposed by using tagging information together with a description which can call "Description with Tagging" (DT indexer) indexing method. To evaluate the proposed indexing method, it was compared with another indexing approach; tagging information only indexing method which is referred to them as "Tag Only" indexing method (T indexer) and tagging information with posted time which can call "Posted Time With Tag" indexing method (PT indexer) respectively.

The paper is structured as follows. Section 1 mentioned related work. Section 2 illustrated the design Framework for image searching. The Section 3 is Result and Discussion. Finally, the Section 4 addressed the Conclusion and Future work.

2. Related work

Most of the works related to the image searching focus on improving the efficiency of academic web resource searching systems such as Flickr.

Many previous experiments on Flickr such as by Christophe, Dominique, Jean-Samuel, Nicolas, and Pascal study with Flickr used a huge data support that has a large data resource of web structure to survey it [3]. Meeyoung, Alan, Bon, and Krishna investigated the social cascades, or how information disseminates through social links in online social networks by using data set from Flickr [4]. Moreover, Lyndon, Mor, Shane, Rahul, and Tye discovered that the community-contributed media and annotation can enhance and improve our access to multimedia resources – and our understanding of the world [5]. In addition, Mark and Michael study of collections for the MIR community comprising 25000 images from the Flickr observed the real community of users both in the image content and the image tag [6].

Despite the recent critical assessments of the optimistic views about and impact of the social tagging and Flickr [19], it is viewed that the social tagging is more useful and crucial for some collections than others. For instance, images in the cultural heritage domain could be tagged and annotated by users who are knowledgeable about local history and language [13].

Some researchers explained that social tagging and professional indexing will have their own separate but complementary roles to play in information organization and retrieval in the future [14, 32, 33]. For example, tagging and tags could be beneficial to users when they need to browse for information resources, a process more helpful than either a directed search or a known item search. An information organization system that relies on controlled vocabularies for indexing purposes might yield different results [15]. Borkur and Roel used tag or so called keywords for searching, which user was using to describe the content of the photo or provide additional contextual and semantic information [8]. The Tag has become a popular means to annotate various web resources, such as web page bookmarkers academic publication, and multimedia object. The tag provides a meaningful descriptor of the object, and the entire user to organize and index them content. Patrick Schmitz represented some promising initial results in inducing from the Flickr tag and described the initial of ontology as a supplement to a tagging system [9]. Livia and Ross applied tags from Flickr to describe city cores [10]. Dong Liu proposed a relevance-based ranking scheme for social image search, aiming to automatically rank images according to their relevance to the query tag. Experimental results in a real Flickr image collection demonstrated the effectiveness [20]. Shaowei Liu et al., suggested an image ranking approach to web image search to improve the relevance between returned images and user intentions [21] whereas Dan Lu et al., improved social re-ranking system for tag-based image retrieval with the consideration of an image's relevance and diversity. The Inverted index structure was created for the social image dataset to accelerate the searching process. Moreover, the result is effective and efficient than Flickr [22]. Zheng Xu et al. [23] measured the semantic relatedness of Flickr images. The data sets including 1000 images from Flickr are used to evaluate the method. Two data mining tasks including clustering and searching are performed, which showed the effectiveness and robustness. Cattuto et al. [24] investigated the network features of the social tags system, which is seen as a tripartite graph using metrics adapted from classical network measures. Lambiotte and Ausloos [25] described the social tags systems as a tripartite network with users, tags, and annotated items. The proposed tripartite network was projected into the bipartite and the unipartite network to discover its structures. In [26], the social tags system was modeled as a tripartite graph which extends the traditional bipartite model of ontologies with a social dimension.

Recently, many researchers investigated the applications of social tags in information retrieval and ranking. In [27], the authors empirically studied the potential value of social annotations for web search. Zhou et al. [28] proposed a model using latent dirichlet allocation, which incorporated the topical background of documents and social tags. Xu et al. [29] developed a language model for information retrieval based on the metadata property of social tags and their relationships to annotated documents. Bao et al. [30] introduced two ranking methods: *SocialSimRank*, which ranked pages based on the semantic similarity between tags and pages, and *SocialPageRank*, which ranked returned pages based on their popularity. Schenkel et al. [31] developed a top- algorithm which ranked search results based on the tags shared by the user who issued the query and the users who annotated the returned documents with the query tags.

Some researchers developed index by combining the social tagging with other factors whereas Pijitra, Siripun., and Worasit combined social tagging with title and abstract or "TTA indexer" to create an index of research paper searching and found that it is more effective [2],[7] and Abebe showed a fundamental difference between the type of tags and type of index term used [11].

These papers used different views to adjust search results of image searching by creating two heuristic indexers to find the optimal indexing.

3. Research Design Frameworks

In this section, the experimental design and evaluation method are discussed. Figure 1 illustrates the framework of this study. The experiment was divided into five steps as follows.

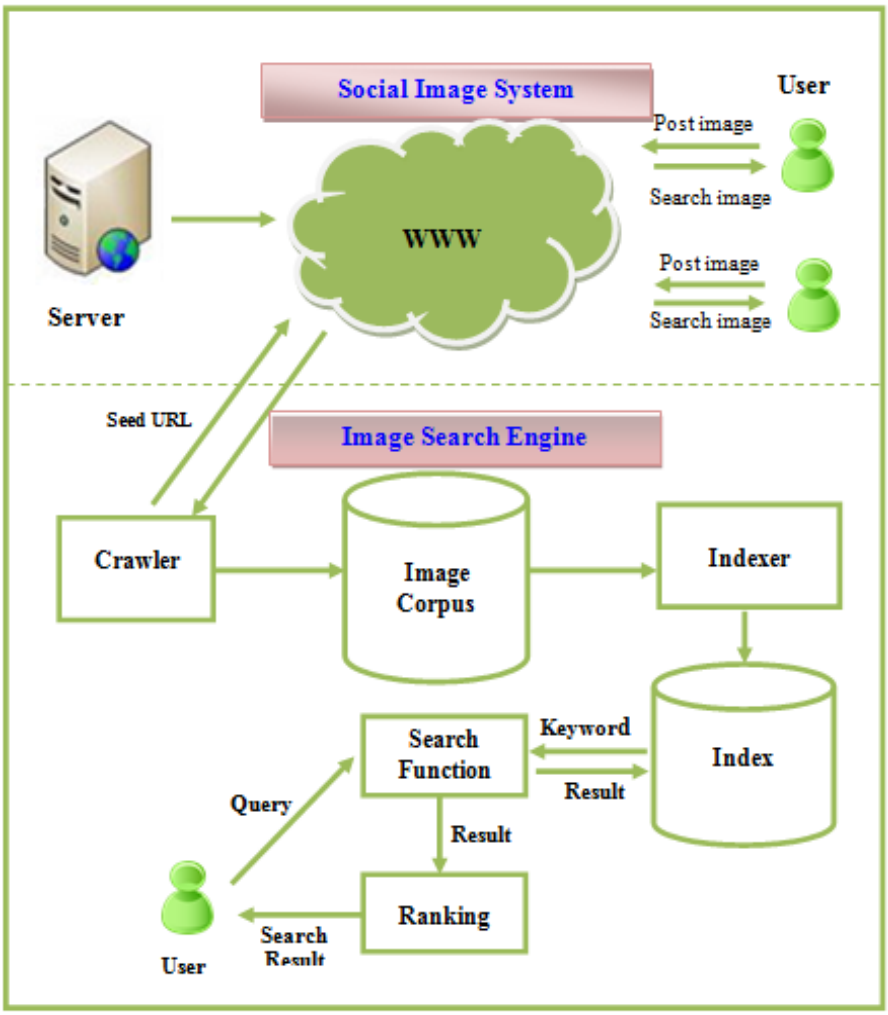

Figure 1. A modified framework for image searching

3.1 Image Crawling

Java programming is used to implement a crawler on the research documents. The crawler collected data from Flickr during July 2011 to July 2014. The collected documents consist of 63,345 images. Each record in the image corpus contains: image ID, image name, tag of each image, link for viewing full image, descriptions, number of viewer, and posted time.

3.2 Image Indexing

In research experiment, three different indexers were developed. The equation (1),(2) and (3) show a modified

term Frequency/Inverse Document Frequency (*tf/idf*) formula for the different indexers, where *T* is "Tag only" , *DT* is "Description with Tag" , and *PT* is "Posted Time with Tag":

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|\{T\}|}{|\{t_i \in d\}|} \quad (1)$$

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|\{DT\}|}{|\{t_i \in d\}|} \quad (2)$$

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|\{PT\}|}{|\{t_i \in d\}|} \quad (3)$$

Let $n_{i,j}$ be the number of occurrences of the considered term in document d_j, $|T|$ is the total number of "Tag" documents in the corpus, and $|DT|$ is the total number of "Description with Tag" documents in the corpus.

$|\{t_i \in d\}|$ is number of the tag documents where the term t_i appears (that is $n_{i,j} \neq 0$) , $|\{t_i \in DT\}|$ is number of description with tag document where t_i appears and $|\{t_i \in PT\}|$ is number of posted time with tag document where t_i appears. If the term is not in the corpus, this will lead to a division –by-zero. It is therefore common to use $1 + |\{t_i \in d\}|$.

3.2.1 Algorithm of indexing Method

Index Algorithm

- two basic algorithms:
  1. make an index for a single document
  2. merge a set of indices
- incremental algorithm:
  - maintain a stack of segment indices
  - create index for each incoming document
  - push new indexes onto the stack
  - let b=Z be the merge factor; M=∞

for (*size* = 1; *size* < M; *size* *= b) **{**
 if (there are b indexes with *size* docs on top of the stack) **{**
  pop them off the stack;
  merge them into a single index;
  push the merged index onto the stack;
 } else {
  break;
 }
}

- optimization: single-doc indexes kept in RAM, saves system calls
- notes:
  average b*$\log_b$(N)/2 indexes
  N=1M, b=2 gives just 20 indexes
  fast to update and not too slow to search
- batch indexing w/ M=∞, merge all at end equivalent to external merge sort, optimal
- segment indexing w/ M<∞

Table 1. Algorithm of indexing Method

Indexing process is one of the core functionalities provided by Lucene [16]. The Figure1 illustrated the indexing process and use of classes. *IndexWriter* is the most important and core component of the indexing process. In addition Table 1 shows a step by step process of the Algorithm of indexing Method**.**

3.2.2 The hypothesis

1) The null hypothesis:
H_0: There is no statistical difference among the means of NDCG at K=1-20 of the three indexing, *T*, *DT*, and *PT*.

$$(\mu_{Tindex} = \mu_{DTindex} = \mu_{PTindex})$$

2) The alternate hypothesis:
H_1: Not all approaches are equal

$$(\mu_{Tindex} \neq \mu_{DTindex} \neq \mu_{PTindex})$$

3.3 Image Searching

Search engines are based on two developed indexers. Equation (1) is applied to the first search engine for creating the web page developed in the experiment. In addition, Equation (2) and (3) is applied to another search engine respectively. The interface is shown in the Figure 2. Here, the subject can specify the search engine. The number of the results per page can also be defined. In addition, the subject can view the results by image ID, small image, and link to full image.

The comparison of a query with the image indexer is computed, a cosine similarity measurement is used to retrieve and rank search results. The similarity score of query *q* for a document is defined as in equation (4).

Figure 2. The interface of the Image Search Engine web page

$$score(q,d) = \sum_{t \in q} \left(tf(t \in d) \times idf(t)^2 \times B_q \times B_d \times L \right) \times C \quad (4)$$

Where

$$B_q = getBoost(t \ \ field \ in \ q)$$

$$B_d = getBoost(t \ \ field \ in \ d)$$

$$L = lenghtNorm(t \ \ field \ in \ d)$$

$$C = coord(q,d) \times queryNorm(s)$$

Where B_q and B_d is the field boost and which is set during

indexing. *L* is the normalization value of a field, given the number of terms with the field; *C* is a value from coordination factor, based on the number of query terms the document contains multiplied with the normalization value for a query, given the sum of the squared weights of each of the query term. Note that getBoost is a function in Lucene [16], which is used to generate indexes for the experiments.

3.4 Experimental Setting

As mentioned in problem, the experiment is designed to solve the main problem of this research. The goal of the experiment is to validate the proposed methodology through the search results in image social searching by using the combination of indexing techniques. A governing metric is used to gauge the experimental outcomes. Search engines based on three indexers were developed. Equation (1) was applied to the first search engine to create the index. Equation (2) was applied to the second search engine based on DT indexer, and Equation (3) was applied to the last search engine based on PT indexer. An accompanying interface web page was also developed for the experiment. Twenty Subjects who are considered as experts in the field participated in the experiment. Therefore, their relevancy ratings are assumed to be perfect for each query.

In the study setting, each subject is assigned to investigate the image obtained from the search engines. Each subject asks three different queries. Each query is applied to all search engines. The first 20 documents for each search engine for relevancy are displayed. Finally, the subjects were asked to rate the relevancy of the search results on a five-point scale:

Score 1 is not relevant at all.

Score 2 is probably not relevant.

Score 3 is less relevant.

Score 4 is probably relevant.

Score 5 is extremely relevant.

The relevancy ratings of each resource in the result set were recorded and used to rank the result set which in turn were used as the normalization constants for NDCG computation using K=1-20. One-way ANOVA was applied to measure the mean difference of NDCG scores at K= 1-20 from the three indexing. If the results from the F-value indicated a significant difference at the 0.05 level, the null hypothesis would be rejected. In addition, Levene's test was used to assess the equality of variances in the samples. If the significance value is greater than 0.05, this meant that the variance is equal. Results from ANOVA showed that there were significant differences among the groups as a whole. The subtle differences among the groups were further amplified by multiple comparisons. Moreover, the Tukey post-hoc and LSD tests were employed to test the equality of variances. The Dunnett T3 was used to test whether there would be any differences in the variances.

3.5 Evaluation metric

NDCG (Normalized Discounted Cumulative Gain) as originally proposed by Jarvelin and Kekalainen [18], was used to evaluate the performance of each search engine. This metric is a retrieval measurement devised specifically for web search evaluation. The NDCG is computed as in the equation (5).

$$NDCG_q = M_q \sum_{j=1}^{k} \frac{\left(2^{r(j)} - 1\right)}{\log(1+j)} \qquad (5)$$

Where *k* is a truncation or threshold level, *r (j)* is an integer representing the relevancy given by the subject and M_q is a normalization constant calculated so that the perfect ordering would obtain a NDCG of 1. NDCG rewards relevant documents appearing in the top ranked search reducing their contributions to NDCG

4. Result and Discussion

4.1 Experimental Results

The top twenty ranks of average NDCG scores were present. Where *k* is the level of the rank, *T* indexer is "Tag only" indexing method, and *DT* indexer is the "Description with Tag" and *PT* indexer is the "Posted Time with Tag" indexing method as shown in Table 2. The NDCG average scores for *DT* indexer are in the range of 0.622-0.689, average scores for *PT* indexer are in the range of 0.620-0.673., whereas, average of the NDCG scores for *T* indexer are in the range of 0.619- 0.663.

Rank No.	*DT*	*PT*	*T*
1	0.622018	0.620068	0.619068
2	0.644743	0.630595	0.616595
3	0.649545	0.630127	0.613127
4	0.676848	0.6358	0.6158
5	0.671225	0.645928	0.616593
6	0.662621	0.647751	0.612678
7	0.662447	0.654118	0.624118
8	0.671747	0.652731	0.632731
9	0.670909	0.659918	0.639918
10	0.669698	0.64627	0.63627
11	0.677076	0.659389	0.639389
12	0.675105	0.660355	0.640355
13	0.677903	0.665195	0.645195
14	0.679384	0.666749	0.646749
15	0.676583	0.666907	0.656907
16	0.672542	0.669709	0.659709
17	0.674101	0.669851	0.659851
18	0.679677	0.673017	0.663017
19	0.681004	0.673004	0.663004
20	0.689989	0.673647	0.663647

Table 2. Average of the NDCG scores for the first 20 ranks of three different rankings

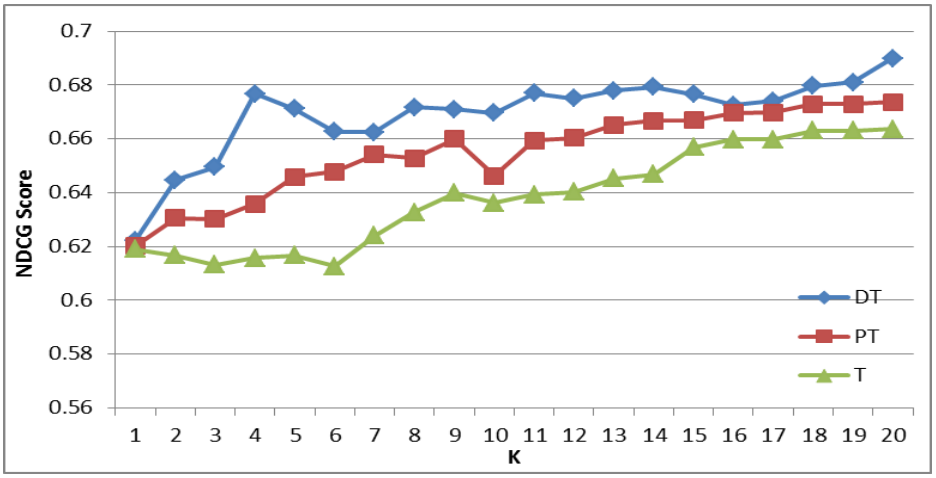

Figure 3. Compare of the average NDCG for three indexing methods

Figure 4. Example of The top results of query “bird” with the proposed relevance based indexing strategy

Figure 5. Example of The top results of query “Thailand” with the proposed relevance based indexing strategy

Figure 4-6 illustrates the top 20 results of query “bird” ,”Thailand” and “Nature” with this research method. The results suggest that the top results are more relevant with other ranking methods.

Figure 6. Example of The top results of query "Nature" with the proposed relevance based indexing strategy

The averages of NDCG score on three different search engines were shown in Figure 3. The x-axis denotes the first 20 ranks of the search results, whereas the y-axis represents the average NDCG score. It suggests that the "Description with Tag" indexing method provide a better set of search results compared with other indexing method. To test the mean difference of the NDCG score of three indexers, *One-Way ANOVA* is employed for top 20 ranks.

Assume that the sample comes from populations that are approximately normal with equal variances. The level of significance is set to 0.05 (a=0.05). The results can be summarized for K = 1 to 20 (All 20 rank), the Table 3 shows that the statistical testing result indicates the significant difference in the NDCG mean scores of all indexers at a=0.05. In other words, the mean scores of NDCG of *DT* indexer, *PT* and *T* indexer are not the same.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.634	2	1.718	15.413	.00
Within Groups	77.633	672	.127		
Total	81.359	673			

Table 3. ANOVA for Experimental Results

4.2 Discussion

The result of the heuristic indexing method by using DT Rank can improve image search engine because the method is based on user behavior. In the current study, the researcher focused on "Description with Tag" or DT indexer, "Posted Time With Tag" or PT indexer and the "Tag only" or T indexer that contained the posted image. We observed that most of subject testers would like to read the description about the more just-posted image. However, the content of the Tag for this particular study is still important. These factors will help system to adjust the ranking and improve search results of image social bookmarking.

5. Conclusion and Future work

This preliminary study focused on the comparison of a heuristic search engine with two indexers. Here, the heuristic indexer implemented used "tag only", "description with tag" and "posted time with tag".

Twenty subjects are assigned to investigate the image obtained from the search engines. Each subject specified three different queries. Each query is applied to these two search engines. The first 20 images of each search engine for relevancy are displayed. Finally, the subjects were asked to rate the relevancy of the search results on a five-point scale.

The results show that *DT* indexer returns a higher NDCG score than other indexer. This implies that *DT* indexer has a better performance than *T* indexer.

However, the number of subjects is considered to be small in the experiment. In order to confirm the finding, more subjects may be needed in the experiments. In addition, the experiment should be extended to different search domains. Improving indexing not only enhances the performance of academic image searches, but also all image searches in general.

Future research in the area consists of extending the scale of experiments, developing ranking, as well as optimizing the parameters.

Acknowledgment

The authors would like to thank the Suan Sunandha Rajabhat University for scholarship support. The study is not possible without the data from Filckr.

References

[1] Flickr, *http://www.flickr.com*

[2] Jomsri, P. (2011). A Combination of Similarity Ranking and Time for Social Research Paper Searching. *In*: Proc.World Academy of Science, Engineering and Technology,78.

[3] Prieur, C., Cardon, D., Samuel, J., Beuscart, N.Pissard, and Pons, P. The Stength of Weak cooperation: A Case Study on Flickr. *In: LIEFA*,University Paris-Diderot.

[4] Cha, M., Mislove, A., Adams, B., Gummadi, K.P. Characterizing Social Cascades in Flickr. (2008) *In:* Proc. the first workshop on Online social networks, Saarbrucken, Germany.

[5] Kennedy, L., Naaman, M., Ahern, S., Nair, R., Rattenbury, T. How Flickr Helps us Make Sence of the World: Context and Content in Community-Contributed Media Collection. *In:* Proc. Yahoo! Research Berkeley, CA, USA.

[6] Huiskes, M.J., Le. M.S. (2008). The MIR Flickr Retrieval Evaluation. *In:* Proc. the 1st ACM international conference on Multimedia information retrieval, p. 39-43.

[7] Jomsri, P., Sanguansintukul, S., Choochaiwattana, W. (2009) .Improve Research paper Searching with Social Tagging – A Preliminary Investigation, *In:* Proc. Natural Language Processing, 2009 (SNLP '09).

[8] Sigurbjornsson, B., van Zwol, R. (2008). Flickr Tag Recommendation base on Collective Knowledge. *Proc. of the 17th international conference on World Wide Web.*

[9] Schmitz, P. (2006) .Inducing Ontology from Flickr Tags. Proc. of the Collaborative Web Tagging Workshop (WWW '06).

[10] Hollenstein, L., Purves, R.S (2010) .Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, Volume 1.

[11] Rorissa, A. (2010). A Comparative Study of Flickr Tags and Index Terms in a General Image Collection. *Journal of the American Society for Information Science and Technology*.

[12] Farooq, U., Kannampallil, T.G., Song, Y., Ganoe, C.H., John, M.C, Giles, L. (2007). Evalating Tagging Behavior in Social Bookmarking Systems: Metrics and design heuristics, *In*: Proc. of the 2007 international ACM conference on Supporting group work (GROUP'07), Sanibel Island, Florida, USA, p.351-360.

[13] Matusiak, K.K. Towards user-centered indexing indigital image collections. OCLC Systems & Services, 22 (4) 283–298.

[14] MacGregor, G. , McCulloch, E. (2006) .Collaborative tagging as a knowledge organisation and resource discovery tool Library View, 55 (5)291–300.

[15] Mathes, A. (2007). Folksonomies—cooperative classification and communication through shared metadata. Retrieved January 30from http://blog.namics.com/2005/Folksonomies_Cooperative_Classification.pdf

[16] Hatcher, E. Gospodnetic, O. (2006). Lucene in Action. Manning Publications Co., United States of America.

[17] Enser, P.G.B. (2008). The evolution of visual information retrieval. *Journal of Information Science*, 34 (4) 531–546.

[18] Jaschke, R. Marinho, L. B., Hotho, A., Schmidt-Thieme, L., Stumme, G. (2007). Tag Recommendations in Folksonomies, *In:* Proceedings of PKDD 2007, volume 4702 of Lecture Notes in Computer Science, Springer Verlagpp. 506–514.

[19] A.M C., (2008) Flickr: Acase study ofWeb 2.0. *Aslib Proceedings*, 60(5),493–516.

[20] Liu,D. Hua,X., Wang ,M., Zhang,H. (2009). Boost search relevance for tag-based social image retrieva. *IEEE International Conference on Multimedia and Expo*, 28 June-3 July 2009.

[21] Liu,S., Cui,P. Luan,H., Zhu, W., Yang,S. , Tian ,Q. (2013).Social Visual Image Ranking for Web Image Search, *Advances in Multimedia Modeling* ,Volume 7732 of the series Lecture Notes in Computer Science. p 239-249 .

[22] Lu,D., Liu, X. ,Qian, X. (2016).Tag-Based Image Search by Social Re-ranking.*IEEE Transactions on Multimedia* 18 (8)1-1 · August 2016.

[23] Xu,Z., Luo, X., Liu,Y. Mei,L.,Hu,C. (2014). Measuring Semantic Relatedness between Flickr Images: From a Social Tag Based View. *The Scientific World Journal*.Volume 2014.

[24] Cattuto, C., Schmitz, C., Baldassarri A. (2007), Network properties of folksonomies. *AI Communications*, 20 (4) 245–262.

[25] Lambiotte R., Ausloos, M. (2006). Collaborative tagging as a tripartite network. *Computational Science*, vol. 3393 of Lecture Notes in Computer Science, p. 1114–1117, 2006.

[26] Maulik, U., Bandyopadhyay, S., Saha, I. (2010). Integrating clustering and supervised learning for categorical data analysis. *IEEE Transactions on Systems*, Man, and Cybernetics A, 40 (4) p. 664–675, 2010.

[27] Ramage, D., Heymann, P., Manning, C. D, Garcia-Molina, H. (2009). Clustering the tagged web," *in Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM '09)*, p. 54–63, February 2009.

[28] Zhou, D. , Bian, J. , Zheng, S. , Zha, H., Lee Giles, C.L. G. C. (2008). Exploring social annotations for information retrieval, *In:* Proceedings of the 17th International Conference on World Wide Web (WWW '08), p. 715–724, April 2008.

[29] Xu, S., Bad, S. , Cao, Y., Yu, Y. (2007).Using social annotations to improve language model for information retrieval, *In:* Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM '07), p. 1003–1006, November 2007.

[30] Bao, S. , Xue, G. , Wu, X., Yu, Y. , Fei, B., Su, Z. (2008). Optimizing web search using social annotations. Proceedings of the 16th International World Wide Web Conference (WWW '07). p. 501–510, May 2007.

[31] Schenkel, R. , Crecelius, T. , Kacimi M. (2008). Efficient top-k querying over social-tagging networks, *In*: Proceedings of the 31st Annual International ACM SIGIRConference on Research and Development in Information Retrieval (ACM SIGIR '08), p. 523–530, July 2008.

[32] Heidarian, Arash., Dinneen,Michael J. (2015). A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering, *Journal of E - Technology,* 6 (3) 78-92.

[33] Tian, Xianyun., Yu, Guang., Yu, Yongtian., Li, Pengyu., Pei,Jiayin (2016). A New Method for Sentiment Classification on Weibo, International Journal of Computational Linguistics Research, 7 (1) 12-24.