Exploitation and use of social annotations in information search

Tran Anh Dung[(1)], Vu Thanh Nguyen[(2)]

Software Engineering Department, University of Information Technology
Ho Chi Minh City, Vietnam

Abstract

Today, with the strong growth of the internet, the search service has been developed rapidly, support web users to easily search for their information. However, with the explosion of information is increasingly enormous, how to return search results satisfy the user remains a difficult problem. Currently, many web-based bookmark systems (such as delicous.com) allows users to easily share and organize their favorite web pages online by using social annotations. The goals of this paper are: 1) exploiting social annotations from delicous.com; 2) using the similarity measure (Social Similarity Ranking) and measure the popularity of the web page (SocialPageRank), build search engine to assist users to search quickly and efficiently. Preliminary experimental results show that the social annotations based search results are very positive.

Keywords: *Social annotation, social page rank, social similarity ranking, information search.*

1. Introduction

Along with the strong development of the internet, the search engine is a very popular and necessary tool of web users. Google is very successful but it was only able to search our questions as keywords. It is always "looking" many unrelated documents, or with existing related documents, Google to find no results or the return is too large. It is difficult to identify the same results,... as the volume of information on the internet is increasing huge. Therefore, in recent years there have been many studies to improve the quality of search engines. Most research focuses on two aspects: 1) Rearrange the order of the web pages according to the query document similarity. There are a number of techniques are applied such as: anchor text generation, metadata extraction, link analysis, search log mining, query extraction profile, web usage profiles,...; 2) Ordering the web pages according to their priority,… It doesn't care the query of user when ranking the priority of web pages. Some techniques are PageRank, HITS, fRank,...

Recently, with the development of web 2.0 technologies, there have many bookmark systems to support web users to easily create annotations for web pages to express concerns and preferences online. So the question is how can we make use of annotations in the search engine. That will support users to search of web resources easily and effectively in the context of the Internet contains almost all the information related to every field, every corner of life.

But it is very large, so large that almost no one can control.

The rest of the paper is organized as follows. Section 2 briefly describes some related works. Section 3 presents how to exploit social annotations from Delicious system[7]. Section 4 describes in detail the social annotation search. Section 5 provides some experimental results. Finally, we conclude with section 6.

2. Related work

In 1998, Google introduced the PageRank[8] algorithm which was considered as a accurately measure the importance of a web page. However, this algorithm only considered rigidly between sites without regard to other characteristics. Based on this the website creators can take advantage of to increase page rank.

In 2006, Pavel A. Dmitriev, Nadav Eiron, Marcus Fontoura, and Eugene Shekita [9], research using community annotations in Enterprise Search. In 2007, Shenghua Bao, Xiaoyuan Wu, Ben Fei, Guirong Xue, Zhong Su, and Yong Yu [11] first mentioned the interest of users by considering public comments. Thereby the authors developed algorithms SocialSimRank and SocialPageRank. This measure reflects a certain relationship between the keywords appear in the web page. In 2008, Ding Zhou et al [3], has been studying and using social annotations in information retrieval and has brought positive results.

Each article contributes a significant part in solving the problem how to improve the effectiveness of information search system on the network. However, how to apply of public social annotations on the search engine to improve search efficiency is also quite sparse.

3. Exploit social annotations

Delicious[7] is a web-based bookmark sharing system developed by Joshua Schachter in late 2003. Now it is part of Yahoo. The main objective of the system is to allow users to store, share, and discover web bookmarks. When users add a bookmark to the system, they can choose to share them. Users annotate a certain URL in three categories: description, note and tags.

The system can recommend tags for URLs that other users have used. These annotations will be used by search engines Delicious. Users can bookmark the page html, audio files, video files, image files, pdf files and doc files - any resource identified by a URL. However, the search results are sorted by time.

Therefore, we have built applications to exploit social annotations from this system to provide search engine. The following figure describes the exploitation of annotations from delicous.com.

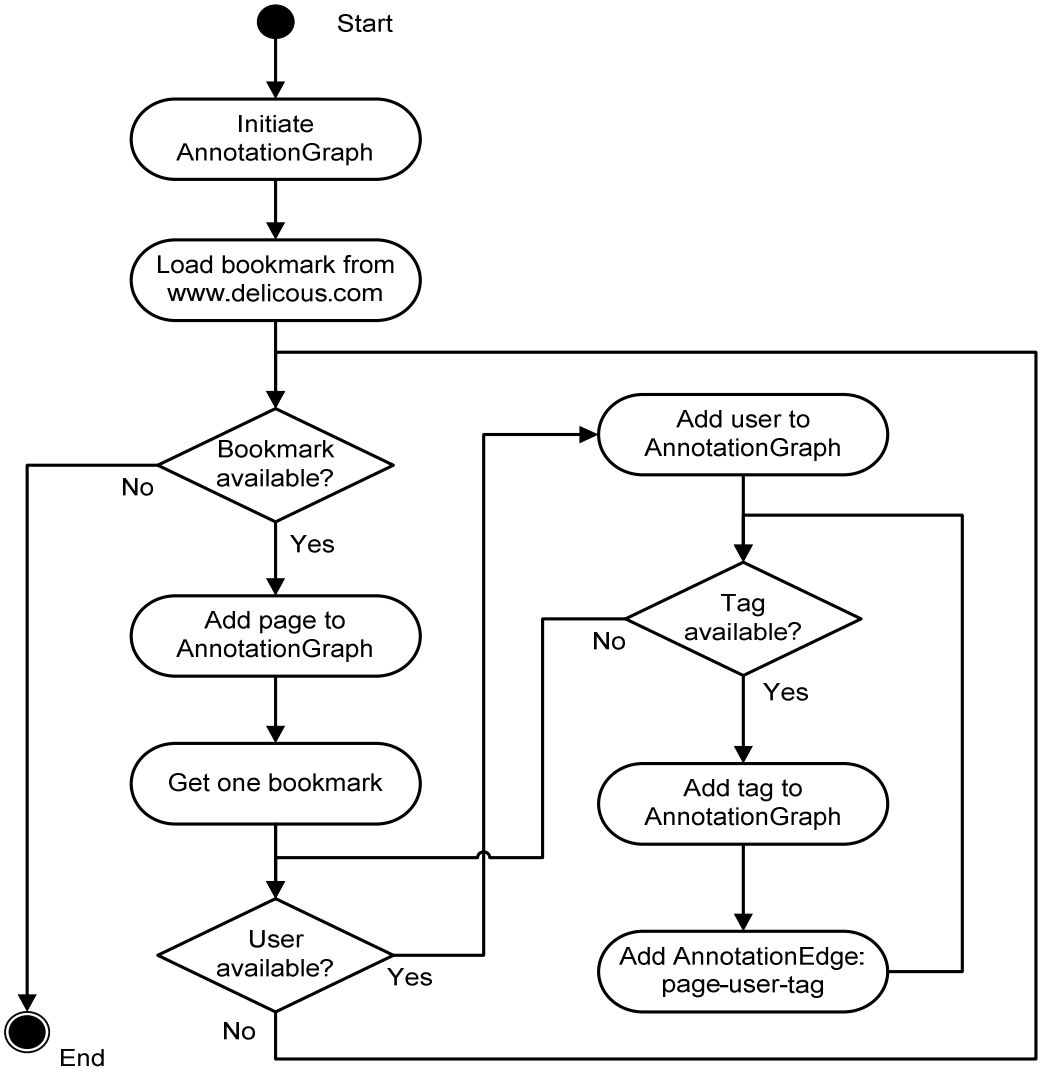


Figure 1: The process of annotation extraction from delicous.com

4. Search engine with social annotation

4.1. Social search system

In the social search systems, there are three kinds of users related:

1) Web page creator: create pages and link the pages with each other to make browsing easy for web users. They provide the basis for search engine.
2) Web page annotator: create annotations for web pages, share annotations for other users.
3) Search engine user: use search engines to get information from the web. They may also become web page annotators if they save and annotate their favorites from the search results.

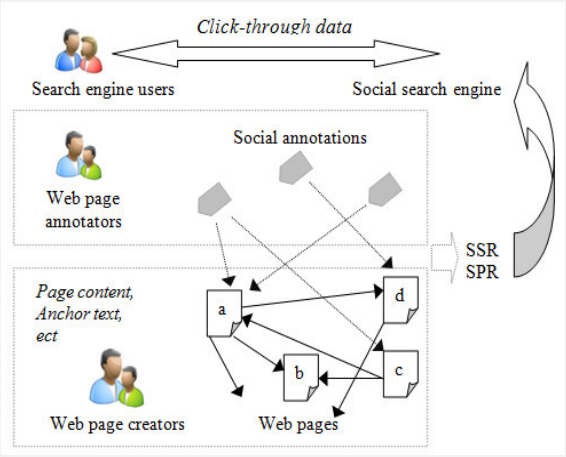


Figure 2: Social search system

4.2. Web page annotator

Web page annotators are web users who use annotation to organize, memories and share their favorite online. They provide a source of clean data usually a good summary of corresponding web pages. Besides, similar or closely related annotations are usually assigned to similar web pages by users with common interests. In the social annotation environment, the similarity among annotations in various forms can further be identified by the common web pages they annotated. Base on this observation, SocialSimRank (SSR) is used to measure the similarity between the query and annotations. In 2007, Shenghua Bao, Xiaoyuan Wu, Ben Fei, Guirong Xue, Zong Su, Yong Yu proposed algorithm SocialSimRank (SSR) and SocialPageRank (SPR).

Assume that there are N_A annotations, N_P web pages and N_U web users. M_{AP} is the $N_A \times N_P$ association matrix between annotations and pages. $M_{AP}(a_x,p_y)$ denotes the number of users who assign annotation a_x to page p_y. Letting S_A be the $N_A \times N_A$ matrix whose element $S_A(a_i, a_j)$ indicates the similarity score between annotations a_i and a_j and S_P be the $N_P \times N_P$ matrix each of whose element stores the similarity between two web pages. Then we have a SocialSimRank algorithm to evaluate the similarity between any two annotations.

Step 1: Init

Let $S_A^0(a_i, a_j) = 1$ for each $a_i = a_j$, otherwise 0

$S_P^0(p_i, p_j) = 1$ for each $p_i = p_j$, otherwise 0

Step 2:

Do{

For each annotation pair (a_i, a_j) do

$$S_A^{k+1}(a_i, a_j) = \frac{C_A}{|P(a_i)||P(a_j)|} \sum_{m=1}^{|P(a_i)|}\sum_{n=1}^{|P(a_j)|} \frac{\min(M_{AP}(a_i,p_m), M_{AP}(a_j,p_n))}{\max(M_{AP}(a_i,p_m), M_{AP}(a_j,p_n))} S_P^k(P_m(a_i), P_n(a_j))$$

For each page pair (p_i, p_j) do

$$S_A^{k+1}(p_i, p_j) = \frac{C_P}{|A(p_i)||A(p_j)|} \sum_{m=1}^{|A(p_i)|}\sum_{n=1}^{|A(p_j)|} \frac{\min(M_{AP}(a_m,p_i), M_{AP}(a_n,p_j))}{\max(M_{AP}(a_m,p_i), M_{AP}(a_n,p_j))} S_A^{k+1}(A_m(p_i), A_n(p_j))$$

}**Until** $S_A(a_i, a_j)$ converges.

Step 3: Output $S_A(a_i,a_j)$

In this algorithm, C_A and C_P denote the damping factors of similarity propagation for annotations and web pages, respectively. $P(a_i)$ is the set of web pages annotated with annotation a_i and $A(p_j)$ is the set of annotations given to page p_j. $P_m(a_i)$ denotes the mth page annotated by a_i and $A_m(p_i)$ denotes the mth annotation assigned to page p_i. Note that the similarity propagation rate is adjusted according to the number of users between the annotation and web page.

Letting $q=\{q_1, q_2, \ldots, q_n\}$ be a query which consists of n query terms and $A(p)=\{a_1, a_2, \ldots, a_m\}$ be the annotation set of web page p. Then we have the formula for calculating the degree of similarity between the query and the annotation as follows:

$$sim_{SSR}(q, p) = \sum_{i=1}^{n}\sum_{j=1}^{m} S_A(q_i, a_j)$$

In our experiments, both C_A and C_P are set to 1.0, and convergence coefficient is chosen ε=0.00001.

4.3. Page Quality Estimation Using Social Annotations

Currently static ranking methods usually measure the quality of web pages from the perspective of the web page creators, or from the perspective of search engine users. The social annotations are the new information can be used to capture the quality of the web page from the perspective of web page annotators. SPR algorithms to measure the quality of web pages indicated by social annotations:

Input:

+ Association matrices M_{PU}, M_{AP}, M_{UA}

+ Vector P_0: The random initial SocialPageRank score P_0

+ ε: parameter convergence

Begin

Do {

1. $U_i = M_{PU}^T * P_i$
2. $A_i = M_{UA}^T * U_i$
3. $P_i' = M_{AP}^T * A_i$
4. $A_i' = M_{AP}^T * P_i'$
5. $U_i' = M_{UA}^T * A_i'$
6. $P_{i+1} = M_{PU} * U_i'$

} Until P_i converges.

End.

Output: P*: the converged SocialPageRank score.

Let M_{PU} be the $N_P \times N_U$ association matrix between pages and users, M_{AP} be the $N_A \times N_P$ association matrix between annotations and pages and M_{UA},the $N_U \times N_A$ association matrix between users and annotations. Element $M_{PU}(p_i,u_j)$ is assigned with the count of annotations used by user u_j to annotate page p_i. Elements of M_{AP} and M_{UA} are initialized similarly. Let P_0 be the vector containing randomly initialized SocialPageRank scores.

In Step 2, P_i, U_i, A_i denote the popularity vectors of pages, users, and annotations in the ith iteration. P'_i, U'_i, A'_i are intermediate values. As illustrated in Figure 3, the intuition behind above Equation is that the users' popularity can be derived from the pages they annotated (1); the annotations' popularity can be derived from the popularity of users (2); similarly, the popularity is transferred from annotations to web pages (3), web pages to annotations (4), annotations to users (5), and then users to web pages again (6). Finally, we get P* as the output of SocialPageRank (SPR) when the algorithm converges. Sample SPR values are given in the Figure 3.

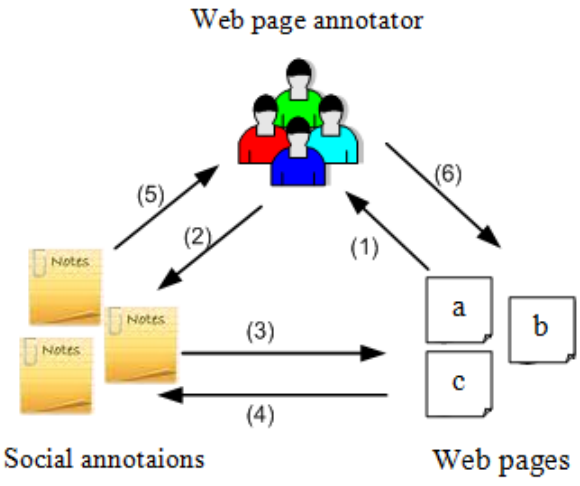


Figure 3: Illustrates the process of calculating SPR

5. Experimental

5.1. Social Annotation Data

Currently, there are many web-social bookmark system. Delicious[7] system is one of the most popular bookmarks. For experiments, we have built an application to perform automatic extraction data from this system. To August, 2012 we have collected 133,044 web pages, 906,142 users and 380,836 different annotations.

5.2. Evaluation Results

To evaluate the results, we have selected a number of any keywords and evaluated through metrics MAP (Mean Average Precision). Let Q = $\{q_1,q_2,\ldots, q_N\}$ is a set of queries. R =$\{r_1, r_2, \ldots., r_k\}$ is the top K result set returned, and r_i is encoded as follows:

$$r_i = \begin{cases} 1, if\ relevance \\ 0, otherwise \end{cases}$$

Evaluation of a link returns are correlated/uncorrelated depending on the user's opinion, in the first selection of 10 empirical results to evaluate.

Average Precision (AveP): measure the accuracy of the result returned:

$$AvP(q_i) = \frac{\sum_{j=1}^{K} \frac{|\{r_k, r_k = 1, k \le j\}|}{j}}{|\{r_k, r_k = 1\}|}$$

$\{r_k, r_k=1, k\le j\}$: is the number of matches to j

$|\{r_k, r_k=1\}|$: is the total number of matches

Mean Average Precision (MAP): average accuracy of the query

$$Mean\ Average Precision(MAP) = \frac{\sum_{i=1}^{N} AvP(q_i)}{|Q|}$$

|Q|: is the number of queries

5.3. Experimental results

To evaluate the results of search engines, we have experimented on a set of collected data and do a search on any number of keywords. Each keyword search conducted and selected 10th first results to calculate Average Precision.

Table 1: Accuracy 1st experimental result set

		Similarity of the ith results												AveP
No	**Keyword**	1	2	3	4	5	6	7	8	9	10	Sum	Rel	
1	Teaching	1.0	1.0	1.0	0.0	0.8	0.8	0.9	0.9	0.0	0.0	**6.37**	**7**	**0.91**
2	Digital marketing strategy	0.0	0.5	0.7	0.8	0.0	0.0	0.6	0.0	0.0	0.0	**2.49**	**4**	**0.62**
3	E-learning	0.0	0.0	0.3	0.5	0.6	0.7	0.7	0.8	0.8	0.0	**4.34**	**7**	**0.62**
4	HD movie free	1.0	1.0	1.0	1.0	0.0	0.8	0.9	0.0	0.0	0.0	**5.71**	**6**	**0.95**
5	JQuery images stretch	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.6	0.0	0.0	**4.63**	**5**	**0.93**

MAP = 0.81

Figure 4: Performance accuracy 1st experimental result set

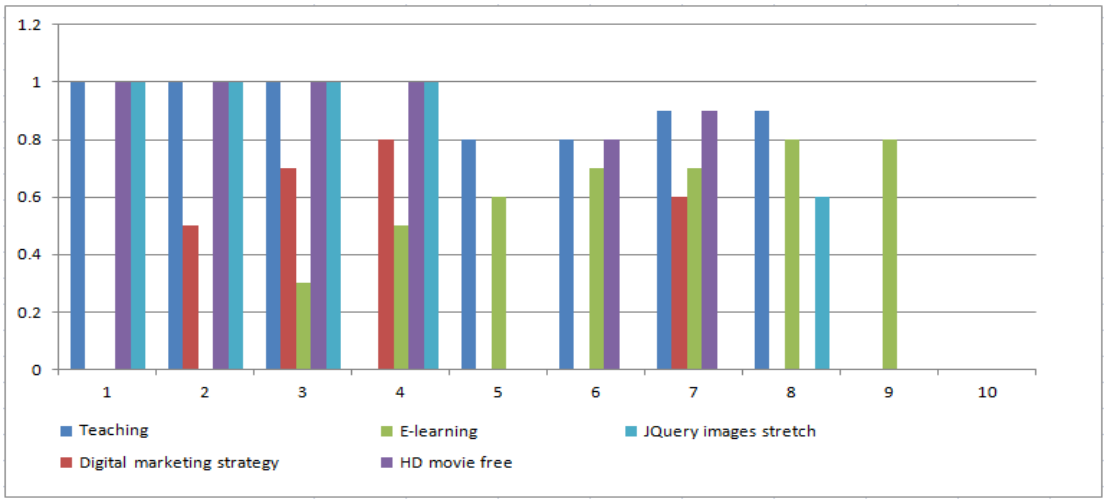
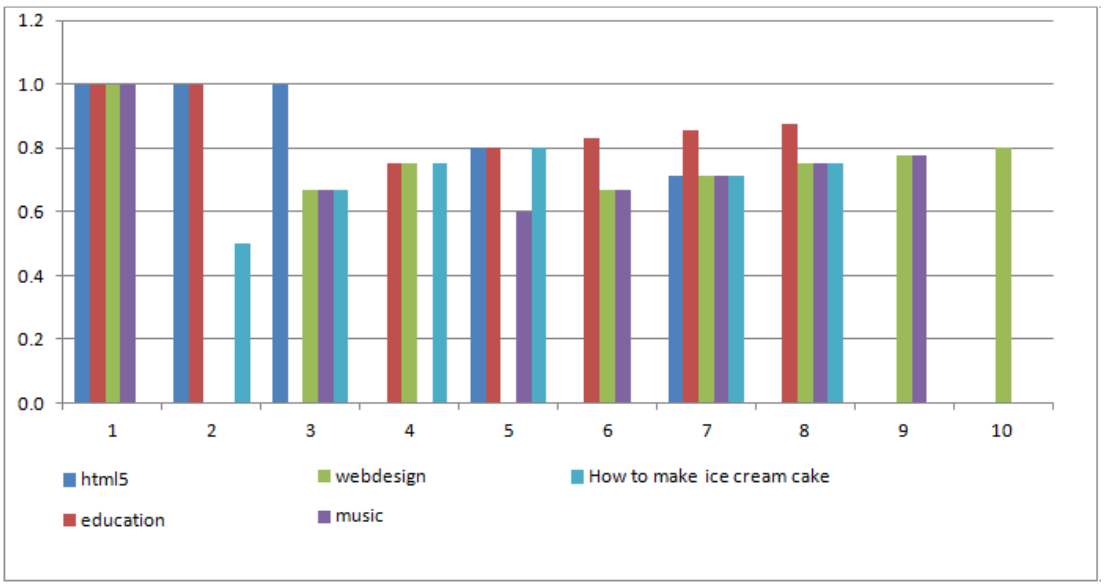


Table 1: Accuracy 2nd experimental results

		Similarity of the ith results												AveP
No	**Keyword**	1	2	3	4	5	6	7	8	9	10	Sum	Rel	
1	Html5	1.0	1.0	1.0	0.0	0.8	0.0	0.7	0.0	0.0	0.0	**4.5**	**5**	**0.90**
2	Education	1.0	1.0	0.0	0.8	0.8	0.8	0.9	0.9	0.0	0.0	**6.1**	**7**	**0.87**
3	Webdesign	1.0	0.0	0.7	0.8	0.0	0.7	0.7	0.8	0.8	0.8	**6.1**	**8**	**0.77**
4	Music	1.0	0.0	0.7	0.0	0.6	0.7	0.7	0.8	0.8	0.0	**5.2**	**6**	**0.86**
5	How to make ice cream cake	0.0	0.5	0.7	0.8	0.8	0.0	0.7	0.8	0.0	0.0	**4.2**	**7**	**0.60**

MAP = 0.80

Figure 5: Performance accuracy 1nd experimental result set



6. Conclusion

In this paper, we study how to exploit and use of social annotations in information search. Annotations provide not only content but also summary, that indicate the popularity of the web page. Especially beneficial social annotations for finding information in both similarity ranking and static ranking. Paper take advantage of the attention and interest of web users to assist users to quickly search the information that they need. Search results shows the application model into social annotations search engine is a very viable research direction and has high application potential for search engines.

7. References

[1] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme (2006), Information Retrieval in Folksonomies: Search and Ranking, In: Proc. of ESWC 2006.

[2] Brush, B. Annotating digital documents: anchoring, educational use, and notification, In CHI '02 Extended Abstracts on Human Factors in Computing Systems. April 20 - 25, 2002, Minneapolis, Minnesota, United States.

[3] Ding Zhou, Jiang Bian, Shuyi Zheng, Hongyuan Zha, and C. Lee Giles (2008), Exploring social annotations for information retrieval, In WWW '08: Proceeding of the 17th international conference on World Wide Web, pages 715–724, New York, NY, USA, 2008. ACM.

[4] E. Agichtein, E. Brill, and S. Dumais (2006), Improving Web Search Ranking by Incorporating User Behavior Information, In Proc. of SIGIR 2006.

[5] Frank McSherry (2005), A uniform approach to accelerated PageRank computation, In: Proc. of WWW 2005.

[6] Freyne J., Farzan R., Brusilovsky P., Smyth B., and Coyle M, Collecting Community Wisdom: Integrating Social Search & Social Navigation. In Proceedings of International Conference on Intelligent User Interfaces, January 28-31, 2007, Honolulu, Hawaii, United States.

[7] http://delicious.com/

[8] Lawrence Page , Sergey Brin , Rajeev Motwani , Terry Winograd (1998), The PageRank citation ranking: bringing order to the web, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1768

[9] Pavel A. Dmitriev, Nadav Eiron, Marcus Fontoura, and Eugene Shekita (2006), Using annotations in enterprise search, In WWW '06: Proceedings of the 15th international conference on World Wide Web, pages 811–817, New York, NY, USA, 2006. ACM

[10] Qing Cui, Alex Dekhtyar (2005), On Improving Local Website Search Using Web Server Traffic Logs: A Preliminary Report.

[11] Shenghua Bao, Xiaoyuan Wu, Ben Fei, Guirong Xue, Zhong Su, and Yong Yu (2007), Optimizing Web Search Using Social Annotations, World Wide Web Conference Committee, Canada

[12] Siegfried Handschuh, Steffen Staab, Authoring and Annotation of Web Pages in CREAM, Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany

[13] Xian Wu, Lei Zhang, and Yong Yu (2006), Exploring social annotations for the semantic web, In Proc. of the 15th International Conference on World Wide Web (WWW'06)