Inferring User Intent in Web Search by Exploiting Social Annotations

Jose M. Conde, David Vallet, Pablo Castells
Universidad Autónoma de Madrid
Cantoblanco, 28049 Madrid, Spain
jose.conde@estudiante.uam.es,{david.vallet, pablo.castells}@uam.es

ABSTRACT

In this paper, we present a folksonomy-based approach for implicit user intent extraction during a Web search process. We present a number of result re-ranking techniques based on this representation that can be applied to any Web search engine. We perform a user experiment the results of which indicate that this type of representation is better at context extraction than using the actual textual content of the document.

Categories and Subject Descriptors H.3.3 Information Search and Retrieval – *retrieval models, information filtering*.

General Terms Algorithms, Experimentation.

Keywords Web search, folksonomy, context.

1. INTRODUCTION

With the advent of the Web 2.0, social tagging systems have exponentially grown both in terms of users and contents. These systems encourage users to tag different types of content items in a way that enhances their organization and sharing. The domains of these systems are varied, including music (e.g., Last.fm), photo streams (e.g. Flickr), or Web pages (e.g. Delicious). However, even with the aid of currently available tools, users still have problems accessing the ever increasing information available on the Web, which is known as the information overload problem.

Personalized and contextualized content access has aimed in recent years to alleviate the information overload problem on the Web by taking both long and short term interests of users into account. Additionally, the underlying semantic information generated by users in social tagging systems, known as *folksonomies*, has enabled the research and development of new effective retrieval and personalized methods. Folksonomies have been recently exploited in order to, for instance, improve [1] or personalize [4] Web search, by exploiting the implicit user and document profiles that can be extracted by mining the social tagging actions of users.

In this paper, we investigate whether social tagging systems can be a new information source for the runtime construction and representation of users' search intent, i.e. if this information can be interpreted as a new source of (implicit) search context. We focus our research on Web search and the use of Delicious[1], a social bookmarking service. We hypothesize that social tagging systems such as Delicious can be a valuable source of information in order to elucidate the semantics involved in the search process of the user. Context-aware models based on implicit feedback rely on processing the content of the executed queries and accessed documents by users during search activities [5]. This information is used to build a representation of the current search context of the user that can be leveraged in subsequent queries of a search session. Here we investigate the application of such techniques to a folksonomy-based representation of Web documents.

This use of social tagging information for user intent extraction has not been investigated in depth before. An initial approach to the exploitation of Delicious as a context source has been presented by Schmidt et al. [3], who extract the tags related to the documents opened by the user, and apply them in query expansion, using an ad-hoc decay factor that gives more importance to recently opened documents. However, no evaluation is provided in this work, so the feasibility of folksonomy-based context representation lacks a formal empirical study, to the best of our knowledge. Rather than on query expansion [3],[5], our approach is based on result re-ranking, as we focus on raising the precision of the top results presented to the user rather than improving recall, as users are prone to only inspect the top results returned by Web search engines. We also investigate more formal context construction strategies [5], such as the ostensive model [2].

Similar to [1],[4], we exploit Delicious as the social annotation corpus to improve Web search, but we focus on providing a context-aware search to the user, rather than implementing a popularity measure [1] or a personalized search [4]. Differently from [4], our approach does not require the user to have a valid user profile in the system. Instead, we build a short term profile at runtime, based on the implicit information provided by the user during the search process and the available tagging information provided by users of the social tagging service.

2. A Folksonomy-based Context Model

We define a folksonomy $\mathcal{F}$ as a tuple $\mathcal{F} = \{\mathcal{T}, \mathcal{U}, \mathcal{D}, \mathcal{A}\}$, where $\mathcal{T} = \{t_1, \dots, t_L\}$ is the set of tags that comprise the vocabulary expressed by the folksonomy, $\mathcal{U} = \{u_1, \dots, u_M\}$ and $\mathcal{D} = \{d_1, \dots, d_N\}$ are respectively the set of users and the set of documents that annotate and are annotated with the tags of $\mathcal{T}$, and $\mathcal{A} =\subset \mathcal{U} \times \mathcal{T} \times \mathcal{D}$ is the set of annotations of the form (u_m, t_l, d_n) where t_l is a tag assigned by a user u_m to a document d_n by a user . The profile of document d_n is defined as a vector $\overrightarrow{d_n} = (d_{n,1}, \dots, d_{n,L})$, where $d_{n,l} = |\{(u, t_l, d_n) \in \mathcal{A} | u \in \mathcal{U}\}|$ is the number of times the document has been annotated with tag t_l. In our Web search scenario, the set of documents $\mathcal{D}$ represents the resources present on the Web, and are identified by a URL.

In order to represent the search process of the user, we define a query trail, $qt = (q, \{d_1, \dots, d_Q\}), d_i \in \mathcal{D}$, as a query q and an ordered set of documents in the result set that the user has accessed after issuing the query. We can then define a session trail, $st = \{qt_1, \dots, qt_S\}$ as an ordered set of query trails that are related

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

[1] Delicious - Social bookmarking, http://delicious.com/

to the current search session of the user. When the user executes a query, the current session trail can be analyzed in order to infer a folksonomy-based representation of the user's search context.

We define a context representation $C(st, t_l)$ as a weighting function that scores the importance of a given tag t_l based on its degree of relation to the user's current session trail st. We based our context representation techniques on the ostensive model [2]:

$$C_o(st, t_l) = \sum_{i=S}^{1} \alpha^{S-i} \cdot \sum_{d_n \in qt_i} w(d_n, t_l),$$

where α is a weight factor that takes into consideration the order of the query trails, i.e. how recently within the current session the query trails occurred. $w(d_{n,l})$ is a weighting function that indicates how well tag t_l represents document d_n. By varying the α parameter we can obtain different implicit context representation models [5]. For instance, for $\alpha < 1$ the context function represents the standard ostensive model [2], where documents that the user opened recently are taken more into consideration for the context representation. In this case α is regarded as a reduction factor. With $\alpha = 1$ we represent an accumulative model where all opened documents have the same importance when representing the user's context. With an $\alpha > 1$ documents opened at the beginning of the search session have more importance when representing the user's context. We analyze two different weighting functions: 1) a weighting function $tf\text{-}idf(d_n, t_l)$, that uses the $tf\text{-}idf$ value of a tag using $d_{n,l}$ as the frequency value; and 2) a weighting function that takes also into consideration the viewing time of the document, by multiplying the previous function by a factor $t(d_n)$, given by the time in seconds that the user spent viewing the document.

We then use this context representation to re-rank the results of the Web search system. Result documents are re-ordered by the similarity value between the document d_n and the current context of the user, taking into account the original ranking. We define this similarity value as a scalar product: $sim(d_n, st) = \sum_l d_{n,l} \cdot C_O(st, t_l)$. Finally, in order to take into account the actual query of the user, the new ordered result list is aggregated with the original result list using CombSUM with rank based normalization.

3. EXPERIMENTS

In order to test different context-aware adaptation techniques, we defined three search tasks and monitored 14 users while performing each task for 15 minutes. The tasks were performed on a common Web browser. All user interactions were collected using a query and action log toolbar[2]. In order to act as a baseline, we instructed users to use Microsoft's Bing[3] Web search engine when performing a Web search. As a way to obtain relevance judgments, users were instructed to bookmark those results they found interesting for their tasks, but they were not asked to find as many relevant documents as they could. Our intention was that users deviate as little as possible from their normal search methodology. The three search tasks were designed to emulate common search situations. These were: 1) planning a holiday trip; 2) searching for a new electronic product you are interested in buying; and 3) an open task were users could search on any topic of their interest. When each task was finished users had 5 additional minutes to judge an aggregated list of results that appeared during their search session, in order for us to obtain more relevance judgements. The above information constituted our evaluation topics. It is worth noting that the coverage of Delicious was 38% for accessed documents and 51% for accessed documents that users marked as relevant.

Figure 1 presents the results of the evaluation. In order to test our hypothesis, we compare the ranking results of the approaches presented in the previous section (tag source) with the same approaches when using the actual textual content of the documents as its representation (text source). The combination of both sources is not shown as it did not produce better results. As a baseline, we use the original ranking provided by Bing (baseline). Some of the evaluated approaches were our representation techniques with varying values of α: $\alpha = 0.5$ (ostensive), $\alpha = 1.0$ (accumulated), and $\alpha = 1.5$ (begin). These three techniques used the $tf\text{-}idf$ weighing function. We also show the performance of the accumulated technique with the time weighting factor (acum + time). In addition, we adapted the technique presented by Schmidt et al. [3], although their work was originally designed for query expansion. As a metric we show MAP at a cut-off point of the top 50 results. Other metrics such as P@10 and NDCG give similar results.

Table 1: MAP@50 performance of context-aware techniques

Sour	baseline	osten	[3]	acum	begin	acum+ti
tag	0.0969	0.1005	0.0998	0.0998	0.1005	**0.1013**
text		0.0881	0.0880	0.0880	0.0887	0.0870

The results show that the context-aware approaches had significantly better results when using the folksonomy-based representation of documents than their actual content. These differences were statistically significant for all approaches (Wilcoxon, $p < 0.05$). Additionally, only the folksonomy-based approaches resulted in an improvement over the baseline, also with statistical significance. From the analysis of the results we can conclude that in our experiment the variations of the α parameter produced no significant effect. The best performing approach was the combination of the accumulated approach and the time sensitive weighting function, which achieved a 15% improvement over the best technique based on textual content and a 4.5% increase over the baseline.

To conclude, we have presented a number of approaches in order to evaluate folksonomy-based context representation techniques. Our results validate our hypothesis that the folksonomy-based representation and exploitation of the user context is more effective than traditional approaches based on the content of the document.

4. ACKNOWLEDGEMENTS

This research was partially supported by the Spanish Ministry of Science and Education (TIN2008-06566-C04-02) and the Regional Government of Madrid (S2009TIC-1542).

5. REFERENCES

[1] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. Optimizing web search using social annotations. In *WWW '07*, pages 501-510, 2007.

[2] Campbell, I. and van Rijsbergen, C. J.. The ostensive model of developing information needs. In *COLIS'96,* pages 251-268, 1996.

[3] Schmidt, K. U., Sarnow, T., and Stojanovic, L. Socially filtered web search: an approach using social bookmarking tags to personalize web search. In *SAC'09*, pages 670-674, 2009.

[4] Xu, S., Bao, S., Fei, B., Su, Z., Yu, Y. Exploring folksonomy for personalized search. In SIGIR'08, pages 155-162, 2008.

[5] White, R. W., Ruthven, I., Jose, J. M., and Van Rijsbergen, C. J. Evaluating implicit feedback models using searcher simulations. *ACM TOIS.*, 23(3), pages 325-361, 2005.

[2] Lemur toolbar: http://www.lemurproject.org/querylogtoolbar/

[3] Microsoft Bing: http://www.bing.com