Personalized Social Query Expansion Using Social Bookmarking Systems

Mohamed Reda Bouadjenek[1,2], Hakim Hacid[1],
Mokrane Bouzeghoub[2], Johann Daigremont[1]
[1]Bell Labs France, Centre de Villarceaux, 91620, Nozay
[2]Laboratoire PRiSM, Université de Versailles , 45 Avenue des États Unis, 78035, Versailles France
[1]FirstName.LastName@alcatel-lucent.com, [2]FirstName.LastName@prism.uvsq.fr

ABSTRACT

We propose a new approach for social and personalized query expansion using social structures in the Web 2.0. While focusing on social tagging systems, the proposed approach considers (i) the semantic similarity between tags composing a query, (ii) a social proximity between the query and the user profile, and (iii) on the fly, a strategy for expanding user queries. The proposed approach has been evaluated using a large dataset crawled from *del.icio.us*.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Information Storage and Retrieval, Information Search and Retrieval

General Terms: Algorithms, Experimentation.

Keywords: Personalization, Social Information Retrieval, Social networks.

1. INTRODUCTION

Nowadays, social networking sites, collaborative tagging, and micro-blogging sites are certainly the most representative and the most adopted applications in this new era of Web 2.0 for interacting with peers, exchanging messages, sharing resources like photos and videos, commenting on news, creating and updating profiles, etc. Valuable information is generally exchanged on these platforms but is not necessarily leveraged due to the dynamics of the ecosystem, the huge amount of information, information volatility, etc. In this context, it is natural that finding relevant pieces of information is a recurrent need for users. However, finding such relevant information becomes harder for end-users since: (i) by definition, the user doesn't necessarily know what she is looking for until she finds it, and (ii) even if the user knows what she is looking for, she doesn't necessarily know how to formulate the right query.

Query expansion comes then as a means to reduce the impacts of such problems. It enriches the user's initial query with additional information that could be relevant to the initial query so that the system may propose suitable results that better satisfy the user's needs. We consider in this work social bookmarking systems, also called *folksonomies*, which are generally modeled as a tripartite social graph with three types of entities: users, documents, and tags. We propose a new approach to query expansion through the combination of semantic and social information to reduce some of the limitations of the existing approaches, e.g., personalization.

The problem we are addressing can be formalized as follows: *For a given user u who issued a query $Q = \{t_1, t_2, ..., t_m\}$, how to provide for each $t_i \in Q$ a ranked list of its related terms $\{t_{i1}, t_{i2}, ..., t_{ik}\}$, such that the gap between user's expectations and system's offerings is minimized.* We translate this by the fact that the ranking function needs to take into account the semantic link between t_i and t_{ij} $(1 \leq j \leq k)$, and the social link between t_{ij} and u, which translates the closeness of the user firing the query with other users in the system. Thus, the objective is to transform Q into a new query Q' such that: (i) Q is necessarily included in Q', (ii) the results of Q are included in those of Q', and (iii) the obtained results with Q' should increase the accuracy of the results and doesn't decrease the user's satisfaction.

2. METHOD

To provide social and personalized expansions of a query term t with related term t_j, we propose to take into account two main features: (i) the similarity between t and t_j, i.e., the semantic strength between the two terms, and (ii) the similarity between t_j and the user profile expressing the extent to which a tag t_j is likely to be interesting to the considered user. We define a user profile as a weighted vector $\vec{p_u} = \{w_{t_1}, w_{t_2}, ..., w_{t_n}\}$, where w_{t_i} is the *user term frequency, inverse user frequency (utf-iuf)* that evaluates how important a term is to a user inside a set of users, i.e., similar to the tf-idf measure.

Once these two similarities are computed, a merge operation is necessary to obtain a final ranking value that indicates the similarity of t_j with t w.r.t. the user u. For this, we choose the *Weighted Borda Fuse (WBF)* as summarized in Equation 1, where $0 \leq \gamma \leq 1$ is a parameter that controls the strength of the semantic and the social parts.

$$Rank_t^u(t_i) = \overbrace{\gamma \times Sim(t, t_i)}^{\texttt{Semantic Part (i)}} + \underbrace{(1-\gamma) \times \frac{\sum\limits_{t_j \in p_u}^{m} Sim(t_i, t_j) \times w_{t_j}}{m}}_{\texttt{Social Part (ii)}} \tag{1}$$

where $Sim(t, t_i)$ is a similarity computed between the query term t and t_i, m the profile's length, and w_{t_j} the weight of t_j in the user profile. In order to consider specific constraints in our approach, we propose a similarity measure that takes into account the credibility of entities in a folksonomy based on their popularity obtained using SPR [3], where the credibility of an entity e (i.e., user, document, or tag) is given by $-log(SPR(e)$. Then, we reduce into three bipartite graphs,

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

i.e., User-Tag, User-Document, and Tag-Document, where we consider only the User-Tag graph (see [4]), to compute the similarity between tags inspired from the *Jaccard similarity* using Formula 2.

$$Sim(t_i,t_j)=\frac{\sum_{u\in N(t_i)\cap N(t_j)} Min(\omega(t_i,u),\omega(t_j,u))\times(-log(SPR(u)))}{\sum_{u\in N(t_i)\cup N(t_j)} Min(\omega(t_i,u),\omega(t_j,u))\times(-log(SPR(u)))} \quad (2)$$

Where $N(t_i)$ is the set of users that are the direct neighbors of t_i (resp. t_j) in the bipartite graph User-Tag, and $\omega(t_i, u)$ the number of times t_i interacts with u obtained by their aggregation over documents. Finally, we reduce the bipartite graph to obtain a graph of tags G_{tag}, which represents similarity between tags computed with Formula 2.

Algorithm 1 *Effective Social Query Expansion SoQuES*

Require: A Social folksonomy Graph G; u : a User; Q : a Query;

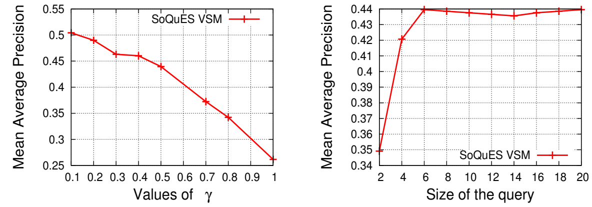
1: $P_u[m] \leftarrow$ extract profile of u from G
2: **for all** $t_i \in Q$ **do**
3: $\quad l \leftarrow$ list of neighbor of t_i in tag graph G_{tag}
4: $\quad$ **for all** $t_j \in l$ **do**
5: $\quad\quad t_j.Value \leftarrow Rank^u_{t_i}(t_j)$
6: $\quad$ Sort l w.r.t. to $t_j.Value$ and let only the top k terms in l
7: $\quad$ Make a logical OR ($\vee$) between t_i and all terms of l,
8: $\quad$ Update Q'
9: **return** Q'

After getting the user's profile as explained above (Line 1), the purpose is to enrich each term t_i of Q with related terms (Line 2). Then, the objective is to get all the neighboring tags t_j of t_i in the tag graph G_{tag} (Line 3). After that, in Line 4, we compute for each t_j, the ranking value that indicates its similarity with t_i w.r.t. the user u (Line 5). Next, the neighborhood list has to be sorted according to the value of $Rank[t_j]$ and keep only the k top tags (Line 6). Finally, t_i and its remaining neighbors must be linked with the OR ($\vee$) logical connector (Line 7) and updated in Q'. *As an example, if a user u issues a query $Q = t_1 \wedge t_2 \wedge ... \wedge t_m$, it will be expanded to become Q'=($t_1\vee$ $t_{11}\vee$ ...$\vee$ t_{1l}) $\wedge$ ($t_2\vee$ $t_{21}\vee$ ...$\vee$ t_{2k}) $\wedge$ ...$\wedge$ ($t_m\vee$ $t_{m1}\vee$ ...$\vee$ t_{mr}).*

3. EVALUATIONS

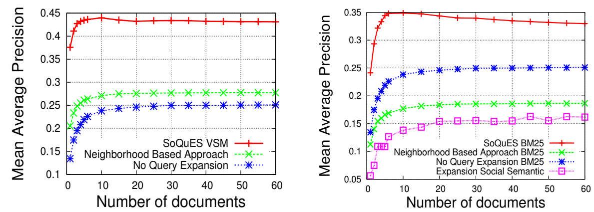
The evaluation has been performed on data crawled from *del.icio.us* on September 2010, which consist of about 150×10^3 bookmarks. We have used the same evaluation protocol like in [1], which consists of: *the relevant documents for a personalized query $Q = (u, t)$ fired by user u with query term t are those tagged by u with t.* Q is enriched and transformed into Q', and documents that are indexed with tags of Q' are retrieved, ranked and sorted according to two IR model: (i) using the *cosine* similarity measure between Q' and each document d_j (Vectorial model), with values of Equation 1 as weights of Q'; or (ii) using the *Okapi BM25* similarity measure (probabilistic model). This is made on *1000* couples (u,t), over which we compute the *Mean Average Precision (MAP)*. We first begin by computing the optimal values of γ and the number of terms added to the queries (Figure 1).

We choose to fix $\gamma = 0.5$ for giving the same importance to the semantic similarity between terms, and to the social proximity with the user profile, and the query size = 6 per query term to compare with other approaches.



(a) MAP over different values of γ (size query=6) (b) MAP over different values of query size ($\gamma = 0.5$)

Figure 1: MAP for different values of γ, and query size, averaged over 1000 queries, using the VSM.



(a) Comparison with approaches based on the VSM. (b) Comparison with approach based on the Okapi BM25.

Figure 2: Comparison of the proposed approach, while fixing $\gamma = 0.5$ and query size to 6.

The comparison with other approaches is made while using the *Vectorial Model*, and the probabilistic model based on *Okapi BM25*. The results for our method are illustrated as *SoQuES VSM* and *SoQuES BM25* in Figure 2(a) and 2(b) respectively. These approaches are: (i) **Non expanded queries** based on VSM and BM25, (ii) **Neighborhood based approach**, which consists of enriching the query Q with the most related terms to t_i without considering the user profile, based on VSM and BM25, and (iii) **Expansion Social Semantic**, an approach proposed in [2], where documents are ranked using BM25 and tag similarity scores. We implemented this strategy and evaluated it over our *del.icio.us* dataset.

4. CONCLUSION

We described in this paper a new query expansion approach based on social personalization to transform an initial query Q to another query Q' enriched with close terms that are mostly used by a given user and her social relatives. A formal evaluation of the quality of the results, using datasets crawled from *del.icio.us*, has shown the benefits of this approach in comparison to other approaches. Reinforce the evaluation by including real users, improve the execution time, and combine our approach with existing techniques are among our plans for future work.

5. REFERENCES

[1] David Carmel et al. Personalized social search based on the user's social network. In *CIKM*, 2009.
[2] Matthias Bender et al. Exploiting social relations for query expansion and result ranking. In *ICDE Workshops*, 2008.
[3] Shenghua Bao et al. Optimizing web search using social annotations. In *WWW*, pages 501–510, 2007.
[4] Peter Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15, 2007.