SoPRa: A New Social Personalized Ranking Function for Improving Web Search

Mohamed Reda Bouadjenek*
PRiSM Laboratory
Versailles University
mrb@prism.uvsq.fr

Hakim Hacid*
Sidetrade
114 Rue Gallieni, 92100
Boulogne-Billancourt, France
hhacid@sidetrade.com

Mokrane Bouzeghoub
PRiSM Laboratory
Versailles University
mokrane.bouzeghoub@
prism.uvsq.fr

ABSTRACT

We present in this paper a contribution to IR modeling by proposing a new ranking function called *SoPRa* that considers the social dimension of the Web. This social dimension is any social information that surrounds documents along with the social context of users. Currently, our approach relies on folksonomies for extracting these social contexts, but it can be extended to use any social meta-data, e.g. comments, ratings, tweets, etc. The evaluation performed on our approach shows its benefits for personalized search.

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

General Terms: Algorithms, Experimentation.

Keywords: Information Retrieval, Social networks.

1. INTRODUCTION

Nowadays, the Web is becoming more and more complex with the socialization and interaction between individuals and objects. This evolution is known as social Web, which includes linking people through the World Wide Web. This is mainly done through platforms such as *Facebook*, *Twitter*, or *YouTube*, where users can comment, spread, share and tag information and resources. The social Web leaded to facilitate the implication of users in the enrichment of the social context of web pages[1]. Especially, it allows users to freely tag web pages with annotations. These annotations can be easily used to get an intuition about the content of web pages to which they are related. Hence, several research works ([4, 7, 9, 19]) reported that adding tags to the content of a document enhances the search quality, as they are good summaries for documents. In particular, tags are useful for documents that contain few terms.

In such a context, classic models of Information Retrieval (IR) should be adapted by considering (i) the social context that surrounds web pages and resources, e.g. their annotations, their associated comments, their ratings, etc. and (ii) the social context of users, e.g. their used tags, their comments, their trustworthiness, etc. Exploiting social information has a number of advantages (for IR in particular). First, feedback information in social networks is provided directly by the user, so user interests accurate information can be harvested as people actively express their opinions on social platforms. Second, a huge amount of social information is published and available with the agreement of the publishers. Exploiting this information should not violate user privacy, in particular social tagging information, which doesn't contain sensitive information about users. Finally, social resources are often publicly accessible, as most of social networks provide APIs to access their data.

In this paper, we are interested in improving the IR model by proposing a new ranking function for documents, while considering the social context of the Web. The approach we are proposing relies on social annotations, which are associated to documents in bookmarking systems but can consider other social metadata, e.g. comments, tweets, etc.

In this context, we propose the following contributions: (1) A Social Personalized Ranking function called *SoPRa*. (2) A method for weighing user profiles and social documents. (3) An extension of SoPRa by considering tagging users individually. (4) An intensive evaluation of *SoPRa*.

The rest of this paper is organized as follows: in Section 2.1, we present the fundamental concepts, and we formally define the problem we tackle. Section 3 presents the related work. Section 4 introduces our approach for ranking documents. Experiments are discussed in Section 5. We conclude and provide some future directions in Section 6.

2. BACKGROUND

2.1 Background and notation

Social bookmarking systems are based on the techniques of *social tagging*. The principle is to provide the user with a mean to freely annotate resources on the Web with tags, e.g. URIs in *delicious*, or images in *Flickr*. These annotations can be shared with others. This unstructured approach to classification is often referred to as a *folksonomy*. A folksonomy is based on the notion of bookmark, which is formally defined as follows:

*This work has been mainly done when the authors was at Bell Labs France, Centre de Villarceaux.

[1]In this paper, we also refer to web pages as documents or resources.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

Definition 1. Let U, T, R be respectively the set of Users, Tags and Resources. A *bookmark* is a triplet *(u,t,r)* such as $u \in U, t \in T, r \in R$, which represents the fact that the user u has annotated the resource r with the tag t.

Then, a folksonomy is formally defined as follows:

Definition 2. Let U, T, R be respectively the set of Users, Tags and Resources. A folksonomy $\mathbb{F}(U, T, R)$ is a subset of the Cartesian product $U \times T \times R$ such that each triple $(u, t, r) \in \mathbb{F}$ is a *bookmark*.

In this paper we use the notation summarized in Table 1.

Table 1: Paper's Notation Overview

Variable	Description
u, d, t	Respectively a user u, a document d and a tag t.
U, D, T	Respectively a set of users, documents and tags.
$\mid A \mid$	The number of element in the set A.
T_u, T_d, $T_{u,d}$	Respectively the set of tags used by u, tags used to annotate d, and tags used by u to annotate d.
D_u, D_t, $D_{u,t}$	Respectively the set of documents tagged by u, documents tagged with t, and documents tagged by u with t.
U_t, U_d, $U_{t,d}$	Respectively the set of users that use t, users that annotate d, and users that used t to annotate d.
$Cos(\vec{A}, \vec{B})$	The cosine similarity measure between two vectors.
$\overrightarrow{pu}$	The weighted vector of the profile of the user u.

2.2 Problem definition

We can formalize the ranking problem as follows: Let consider a folksonomy $\mathbb{F}(U, T, R)$ whose a user $u \in U$ submits a query q to a search engine. We would like to re-rank the set of resources $R_q \subseteq R$ (or documents) that match q, such that relevant resources for u are highlighted and pushed to the top for maximizing his satisfaction and personalizing the search results. The ranking follows an ordering $\tau = [r_1 \geq r_2 \geq \cdots \geq r_k]$ in which $r_k \in R$ and the ordering relation is defined by $r_i \geq r_j \Leftrightarrow Rank(r_i, q, u) \geq Rank(r_j, q, u)$, where $Rank(r, q, u)$ is a ranking function that quantify similarity between the query and the resource w.r.t the user [14].

3. RELATED WORK

We distinguished two categories for social results re-ranking that differ in the way social information is used. The first category uses social information by adding a social relevance to documents while the second use it for personalization.

Re-ranking using social relevance: Several approaches have been proposed to improve document re-ranking using social relevance. Social relevance refers to information socially created that characterizes a document from a point of view of its interest, i.e. its general interest, its popularity, etc. Many approaches [1, 10, 13, 18] have been proposed to adapt popular algorithms such as PageRank and HITS.

Personalized re-ranking: In general, users have different interests. Hence, in an IR system, providing the same sorted documents is not really suitable. Thus, a personalized function to sort documents w.r.t each user is expected to improve results. Many approaches have been proposed to personalize the ranking using social information [8, 14, 15, 17]. These approaches consider a matching between the user profile and a document, and a matching between the query and the document annotations as in Figure 1. The approach we are proposing is part of this initiative. However, we consider a new aspect, which is a social matching score between a query and the annotations of documents.

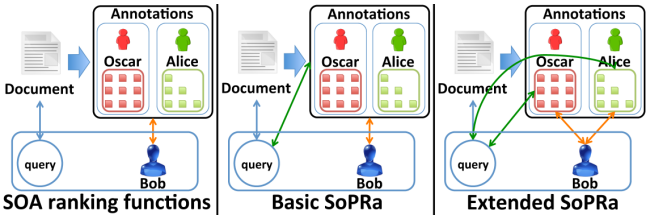


Figure 1: Illustration of the basic differences between the different approaches.

4. SoPRa FUNCTION

In this Section, we first define the *SoPRa* function, and then we present the methods for modeling social documents and users. Finally, we present an extended version of SoPRa.

4.1 Basic SoPRa

We follow the widely used Vector Space Model (VSM), where all the queries and the documents are mapped to be vectors in a universal term space.

On the one hand, we believe that a matching score between a document d and a query q should be based on (i) a textual matching score, and (ii) a social matching score. The textual matching score expresses the similarity between the textual content of d and q. The social matching score expresses how similar the social representation of d is, for q. This social representation is based on the annotations associated to d. More formally, in this work, we consider these two ranking scores as independent evidence, and we propose to merge them using a linear function as follows:

$$Score(q, d) = \beta \times Cos(\overrightarrow{q}, \overrightarrow{T_d}) + (1 - \beta) \times Sim(\overrightarrow{q}, \overrightarrow{d}) \quad (1)$$

where β is a weight that is equal to 0.5, $Sim(\overrightarrow{q}, \overrightarrow{d})$ denotes the textual matching score between d and q (currently computed using the *Apache Lucene* search engine in our implementation), and $\overrightarrow{T_d}$ is the vector that models the social representation of the document d.

On the other hand, in the non-personalized search engines (classic IR models), the relevance between a query and a document is assumed to be only based on the textual content of the document. However, as relevance is actually relative for each user, considering only a matching between a query and documents is not enough to generate satisfactory search results. Thus, we propose to estimate the interest of a user u to a document d by computing a similarity between the profile of u and the social representation of d. Then, we propose to merge this interest value to the previous ranking score computed in Equation 1 for computing an overall score to a document. Formally, the ranking score of a document d that potentially match the query q issued by a user u is computed as follows:

$$Rank(d, q, u) = \gamma \times Cos(\overrightarrow{p_u}, \overrightarrow{T_d}) + (1 - \gamma) \times Score(q, d) \quad (2)$$

where, γ is the weight that satisfies $0 \leq \gamma \leq 1$. The ranking model of the basic SoPRa function is illustrated in Figure 1

for more clarity. In summary, SoPRa ranks documents according to: (i) a textual content matching score of documents and the query, (ii) a social matching score of documents and the query, and (iii) the social interest score of the user to documents.

4.2 Weighting scheme

In this paper, the social representations of documents and the user profiles are estimated by their social annotations and modeled as in the VSM. Hence, if we consider web pages or users as documents and annotations as terms, the above setting is right for the VSM. Even if the VSM has been developed a long time ago, it has shown its effectiveness for IR and remains very competitive and challenging. One of the key points in the VSM is the weighting of terms. We propose to simply weight annotations using the *tf-idf* measure as follows:

$$w_t^d = tf_t \times log(\frac{\mid R \mid}{\mid R_t \mid}),\ w_t^u = utf_t \times log\left(\frac{\mid U \mid}{\mid U_t \mid}\right) \quad (3)$$

where w_t^d is the weight of the term t in the social representation of d, tf_t denotes the tag frequency, w_t^u is the weight of the term t in the profile of u, and utf_u is the user term frequency, i.e. the number of time the user u used the tag t.

4.3 Extended SoPRa

In classic models of IR, the content of a web page is considered as a mixture of homogeneous terms generated by the same creator, i.e. the author of the web page. However, social bookmarking systems allow users to freely assign annotations to documents following their own vocabulary to describe these documents. Hence, unlike the textual content of a web page, annotations can be seen as a mixture of heterogeneous fragments, where each fragment describes the content of the web page with annotations of a particular user. This notion of fragments is illustrated in Figure 1 as clusters of annotations. Consequently, we believe that IR ranking functions may be improved by considering independently each user that annotates a web page. Strengthening annotations provided by similar users to the query issuer can enhance the score of a document. To address this problem, we propose an extension of *SoPRa* by discriminating between users who annotate web pages and by considering their similarities with the query issuer. Hence, we extend the basic *SoPRa* as follows:

$$Rank(d,q,u) = \gamma \times \sum_{u_k \in U_d} Cos(\overrightarrow{p_{u_k}}, \overrightarrow{p_u}) \times Cos(\overrightarrow{p_u}, \overrightarrow{T_{u_k,d}}) + (1-\gamma) \times \left[\beta \times \sum_{u_k \in U_d} Cos(\overrightarrow{p_{u_k}}, \overrightarrow{p_u}) \times Cos(\overrightarrow{q}, \overrightarrow{T_{u_k,d}}) + (1-\beta) \times Sim(\overrightarrow{q}, \overrightarrow{d})\right] \quad (4)$$

where $\overrightarrow{T_{u_k,d}}$ is the vector that models the social representation of the document d based only on the annotations provided by u_k to d. The ranking model of the extended *SoPRa* is illustrated in Figure 1 for more clarity.

In summary, in this section, we presented *SoPRa* in its basic form as well as an extension of *SoPRa*, which individually consider users and their similarities to the query issuer. In the next Section, we present the evaluation we performed.

5. EVALUATION

To evaluate our approach, we have selected a *delicious* dataset, which is public, described and analyzed in [16]. Before the experiments, we performed four data preprocessing tasks: (1) We remove annotations that are too personal or meaningless, e.g. "toread", "Imported IE Fa-vorites", etc. (2) The list of terms undergoes a stemming by means of the Porter's algorithm in such a way to eliminate the differences between terms having the same root. (3) We downloaded all the available web pages while removing those which are no longer available using the *cURL* command line tool. (4) Finally, we removed all the non-english web pages. Table 2 gives a description of the resulted dataset.

Table 2: Details of the delicious dataset

Bookmarks	Users	Tags	Web pages	Unique terms
9 675 294	318 769	425 183	1 321 039	12 015 123

5.1 Evaluation methodology

Making evaluations for personalized search is a challenge since relevance judgments can only be assessed by end-users [8]. This is difficult to achieve at a large scale. However, different efforts [3, 4, 11] state that the tagging behavior of a user of a folksonomy closely reflects his behavior of search on the Web. In other words, if a user tags a document d with a tag t, he will choose to access the document d if it appears in the result obtained by submitting t as query to the search engine. Thus, we can easily state that any bookmark (u,t,r) that represents a user u who tagged a document d with tag t, can be used as a test query for evaluations. The main idea of these experiments is based on the following assumption:

For a query $q = \{t\}$ issued by u with query term t, relevant documents are those tagged by u with t.

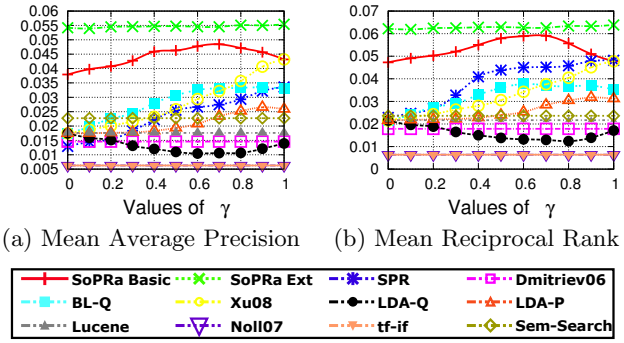
Hence, for each evaluation, we randomly select 2000 pairs (u,t), which are considered to form a personalized query set. For each corresponding pair (u,t), we remove all the bookmarks $(u,t,r) \in \mathbb{F}, \forall r \in R$ in order to not promote the resource r (or document) in the results obtained by submitting t as a query in our algorithm and the considered baselines. By removing these bookmarks, the results should not be biased in favor of documents that simply are tagged with query terms and making comparisons to the baseline uniformly. Hence, for each pair, the user u sends the query $q = \{t\}$ to the system. Then, we retrieve and rank all the documents that match this query using our approach or a specific baseline, where documents are indexed based on their textual content using the *Apache Lucene*. Finally, according to the previous assumption, we compute the Mean Average Precision (MAP) and the Mean Reciprocal Rank (MRR) over the 2000 queries. The random selection was carried out 10 times independently, and we report the average results.

5.2 Comparison with baselines

We compare our approach to several personalized and non-personalized baselines, in which the social based score is merged with the textual based matching score using a linear function with a γ parameter. The results are illustrated in Figure 2, while varying γ.

5.2.1 SoPRa VS non-personalized ranking approaches

We compare our approach to: SocialPageRank (SPR) [1], Dmitriev06 [9], and the Lucene naive score. We also compare our approach to an approach where the matching score is computed as in Equation 1, and we refer to this approach



(a) Mean Average Precision (b) Mean Reciprocal Rank

Figure 2: Comparison with the baselines.

as BL-Q. The last approach use LDA [5], where we construct a model using documents. Then, for each document that match a query, we compute a similarity between its topic and the topic of the query (inferred by considering queries as documents) using the cosine measure. The obtained value is merged with the textual ranking score and we refer to this approach as LDA-Q.

The results show that *SoPRa* and its extension is much more efficient than all the non-personalized approaches for all values of γ. Hence, we conclude that the personalization efforts introduced by *SoPRa* improve the search quality. We also notice that most of the non-personalized approaches decease their performance for high values of γ. This is certainly due to the fact that they are not designed for personalized search, since these approaches fail in discriminating between users in spite of their preferences.

5.2.2 SoPRa VS personalized ranking approaches

Here we compare our approach to: Xu08 [17], Noll07 [12], tf-if [14], and Semantic Search [2]. We also use an approach based on LDA, where we construct a model using documents. Then, for each document that match a query, we compute a similarity between its topic and the topic of the user profile (inferred by considering users as documents) using the cosine measure. The obtained value is merged with the textual ranking score and we refer to this approach as LDA-P. Here, the obtained results also show that our approach is much more efficient than all the baselines for all values of γ. Especially, our approach outperform the LDA-P approach and the Xu08 approach, which we consider as the closest works to our. We also notice that the Noll07 and the tf-if approaches give poor results. This is certainly due to the fact that they fail in ranking documents that doesn't share tags with users, since in our experiment we remove the triplets that associate the user, the query terms and documents.

Finally, we note that the better performance are obtained for $\gamma \in [0.6, 0.8]$ for the basic *SoPRa*, a compromise between the user interest matching score and the query affinity matching score. As for the extended *SoPRa,* it seems that γ has no impact on the results. This show that the extension proposed takes full advantage of the user interest matching score and the query affinity matching score. We also note that the extension of *SoPRa* provides better performance than the basic one. This shows that considering users individually and their similarities to the query issuer provide a better estimation of the relevance of documents.

6. CONCLUSION AND FUTURE WORK

This paper discusses a contribution to the area of IR modeling while leveraging the social dimension of the web. We proposed a new documents ranking function called *SoPRa,* which uses social information to enhance and improve web search. The experiments performed show the benefit of *SoPRa* while comparing it to the closest works. *SoPRa* can be improved in different way. First, the temporal dimension of social users' behavior has not been deeply investigated yet in the literature, e.g. considering the evolution of the taste of users in the ranking function. Second, considering a social relevance score factor, which characterizes documents from a point of view of interest, is a possible improvement of *SoPRa*, e.g. their popularities. Finally, performing an online user evaluation in order to validate our results is an ongoing task. SoPRa has been developed and integrated to the LAICOS [6] platform.

7. REFERENCES

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

[2] M. Bender, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, R. Schenkel, and G. Weikum. Exploiting social relations for query expansion and result ranking. In *ICDE Workshops*, 2008.

[3] D. Benz, A. Hotho, R. Jäschke, B. Krause, and G. Stumme. Query logs as folksonomies. *Datenbank-Spektrum*, 10:15–24, 2010.

[4] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *CIKM*, 2008.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[6] M. R. Bouadjenek, H. Hacid, and M. Bouzeghoub. LAICOS: An Open Source Platform for Personalized Social Web Search. In *KDD*, 2013.

[7] M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, and A. Vakali. Using Social Annotations to Enhance Document Representation for Personalized Search. In *SIGIR*, 2013.

[8] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, and S. Chernov. Personalized social search based on the user's social network. In *CIKM*, 2009.

[9] P. A. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita. Using annotations in enterprise search. In *WWW*, 2006.

[10] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, 2006.

[11] B. Krause, A. Hotho, and G. Stumme. A comparison of social bookmarking with traditional search. In *ECIR*, 2008.

[12] M. G. Noll and C. Meinel. Web search personalization via social bookmarking and tagging. In *ISWC'07/ASWC'07*, 2007.

[13] T. Takahashi and H. Kitagawa. A ranking method for web search using social bookmarks. In *DASFAA*, 2009.

[14] D. Vallet, I. Cantador, and J. M. Jose. Personalizing web search with folksonomy-based user and document profiles. In *ECIR*, 2010.

[15] Q. Wang and H. Jin. Exploring online social activities for adaptive search personalization. In *CIKM*, 2010.

[16] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *ECAI*, 2008.

[17] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring folksonomy for personalized search. In *SIGIR*, 2008.

[18] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Towards improving web search by utilizing social bookmarks. In *ICWE*, 2007.

[19] X. Zhang, L. Yang, X. Wu, H. Guo, Z. Guo, S. Bao, Y. Yu, and Z. Su. sdoc: exploring social wisdom for document enhancement in web mining. In *CIKM*, 2009.