Information Sciences 281 (2014) 399–417

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Exploiting social bookmarking services to build clustered user interest profile for personalized search

Harshit Kumar *, Sungin Lee, Hong-Gee Kim [1]

*BiKE Lab, Seoul National University, Yeongun Campus, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 5 January 2013
Received in revised form 19 March 2014
Accepted 8 May 2014
Available online 24 May 2014

*Keywords:*
User profiling
Clustering
Personalized search
Social bookmarking service

ABSTRACT

Search engine users tend to write short queries, generally comprising of two or three query words. As these queries are often ambiguous or incomplete, search engines tend to return results whose rankings reflect a community of intent. Moreover, search engines are designed to satisfy the needs of the general populace, not those of a specific searcher. To address these issues, we propose two methods that use Singular Value Decomposition (SVD) to build a Clustered User Interest Profile (*CUIP*), for each user, from the tags annotated by a community of users to web resources of interest. A *CUIP* consists of clusters of semantically or syntactically related tags, each cluster identifying a topic of the user's interest. The matching cluster, to the given user's query, aids in disambiguation of user search needs and assists the search engine to generate a set of personalized search results. A series of experiments was executed against two data sets to judge the clustering tendency of the cluster structure *CUIP*, and to evaluate the quality of personalized search. The experiment results indicate that the *CUIP* based personalized search outperforms the baseline search and is better than the other approaches that use social bookmarking services for building a user profile and use it for personalized search.

© 2014 Elsevier Inc. All rights reserved.

## 1. Introduction

The abundance of information available on the Web has made search engines (SEs) an indispensable tool. Higher availability of information means that there is a greater chance of finding sought-after information on the Web, but with increased complexity of discovering relevant information. While SEs do a good job of ranking results to maximize global happiness, they fail to do a very good job for specific individuals [39]; it appears that the rankings reflect a community of intent rather than the goals of individuals. There are many reasons for the ineffectiveness of SEs. First, user queries are of poor quality: the average length of user queries ranges between two to three words [36]; such short queries cannot effectively describe the user search intent or user information needs. Second, some queries are polysemous [33]: they have different meanings in different contexts; hence it is impossible for the SE to judge the user intent from the short polysemous queries.

The major shortcoming of SEs is the inability to incorporate user modeling with search and unadaptiveness to individual users. Personalization has emerged as an appealing approach when dealing with the issues caused by the variation of on-line behaviors and individual differences observed in user interests, information needs, search goals, query contexts, and others [5]. Many methods [1,4,8,15,16,26,28,41,44,45] are proposed to study user search behavior and use it to build a profile of

* Corresponding author.
*E-mail address:* kumar@bike.snu.ac.kr (H. Kumar).
[1] Principal corresponding author.

http://dx.doi.org/10.1016/j.ins.2014.05.008
0020-0255/© 2014 Elsevier Inc. All rights reserved.

user interests based on the user interactions on the Web. These methods focus on analyzing the content of queries and web pages, but in some cases there are no suitable descriptors available, such as topics and genres, that can be used to accurately describe user interests. Moreover, these methods rely on mining data sources, such as user's email [38], click-through history [4], desktop files [10], and bookmark history [17], all of which tend to be noise-infested. Information sources, such as social bookmarking services, which are low on noise, and provide precise information, are very much desirable. Social bookmarking services, such as flickr, Delicious, and Pinterest, allow users to annotate resources; this facilitates management, organization, and sharing of resources [41], and also provides an indication of sources of user interests. Noll and Meinel [28], Xu et al. [45] proposed solutions that use a social bookmarking service to extract tags from resources of user interests, and use the tags for building a User Interest Profile (*UIP*). User interests can be viewed as contextual variants that may help to disambiguate the user query intent when the original query is vague or there are too many search results that the user has to wade through to find the most relevant ones.

This paper makes the following contributions:

1. We propose two methods to build a *CUIP* for personalized search: one that uses Singular Value Decomposition (SVD) to generate *svdCUIP*, and the other a variation of SVD, modSVD, to geneate a *modSvdCUIP*. A set of pairs of the form $(t, tw)$, where $t$ is a tag and $tw$ is the accumulated weight of the tag $t$, constitutes a User Interest Profile (*UIP*). A *CUIP* is defined as a set of term clusters, where each term cluster consists of semantically related tags of user interests and tag weights.
2. An automatic evaluation method is proposed to test the proposed methods with the baseline search and folksonomy based personalized search approaches.
3. We performed experiments to evaluate the proposed methods on two different data sets. The first data set, called custom data set, was created from the search histories of 12 volunteers. This data set was organized to establish the ground truth for the evaluation of clustering tendency and clustering accuracy of *CUIPs* generated by the proposed methods. The second data set is a much bigger data set harvested from the AOL search query log. This data set was used to test the improvement in personalized search for the two proposed methods, and their comparisons with other methods.
4. Our results show that personalized search using the *modSvdCUIP* is better than using the *tfUIP* (*term frequency UIP*) [28] and *tfIdfUIP* (term frequency Inverse Document Frequency UIP) [45], and exhibits modestly better performance than the *tfIdfCUIP* [34] and *svdCUIP*. Each cluster, in the cluster structure *CUIP*, identifies a topic, and the application of *CUIP* helps disambiguate the context of user query, which is particularly needed for vague queries.

The rest of the paper is organized as follows. Section 2 discusses the related work starting with the traditional approaches to user profiling for personalized search, followed by the current approaches to user profiling that involve social bookmarking services. Section 3 presents our proposed methods. The experiments are detailed in Section 4, and the paper is concluded in Section 5.

## 2. Background and related work

In this section, we first present the state-of-the-art in building a *UIP*, and discuss the most recent approaches to building a *UIP* from social bookmarking services for personalized search. Differences in approaches are tabulated in Table 1. Finally, we discuss two well-known approaches to obtaining personalized search results.

**Table 1**
A comparison summary of the proposed approaches with the other similar approaches that uses folksonomy for personalized search. (a) Source of terms for constructing a UIP, (b) Web document representation, (c) similarity measure, (d) first-order co-occurrence, (e) second-order co-occurrence, (f) clustering of terms in a UIP, (g) UIP and resource length normalization factor, (h) adaptation/no adaptation/proposed.

| | *tfUIP* [28] | *tfIdfUIP* [45] | *tfIdfCUIP* [34] | *svdCUIP* | *modSvdCUIP* |
|---|---|---|---|---|---|
| (a) | User annotations to Web documents | User annotations to Web documents | User annotations to Web documents on delicious | Annotations by the community of users to Web documents clicked by the user | Annotations by the community of users to Web documents clicked by the user |
| (b) | Resource profile (folksonomy based) | Resource profile (folksonomy based) | Resource profile (folksonomy based) | document contents | Document contents |
| (c) | Dimensionless cosine similarity measure, Eq. (1) | Calculates the tfIdf cosine similarity between a UIP and a resource profile of the Web document, Eq. (2) | Calculate the cosine similarity between a global cluster matching a user query and the resource profile of the Web document | Calculates the cosine similarity between the matching cluster in the CUIP to the user query and the document contents, Eq. (8) | Calculates the cosine similarity between the matching cluster in the CUIP to the user query and the document contents, Eq. (8) |
| (d) | Yes | Yes | Yes | Yes | Yes |
| (e) | No | No | No | No | Yes |
| (f) | No | No | Yes | Yes | Yes |
| (g) | No | Yes | Yes | Yes | Yes |
| (h) | No adaptation | No adaptation | Adaptation | Proposed | Proposed |

Download English Version:

https://daneshyari.com/en/article/393513

Download Persian Version:

https://daneshyari.com/article/393513

Daneshyari.com