Information Sciences 336 (2016) 115–128

Contents lists available at ScienceDirect

Information Sciences

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

Integrating social profile to improve the source selection and the result merging process in distributed information retrieval

Zakaria Saoud*, Samir Kechid

LRIA, Computer Science Department, USTHB, Algiers, Algeria

ARTICLE INFO

Article history:
Received 26 January 2015
Revised 5 November 2015
Accepted 15 December 2015
Available online 22 December 2015

Keywords:
Information retrieval
Social profile
Folksonomy
Source selection process
Merging process

ABSTRACT

In this paper we present a new personalized approach, which integrates a social profile into a distributed search system. Most previous studies on distributed information retrieval are based on textual information, and rarely consider any social information. Based on this observation, we propose an approach which exploits the social profile and the different relations between social entities. We believe that this method can: (i) enhance a query expansion, (ii) personalize and improve both the source selection and the result merging process in distributed information retrieval systems.

1. Introduction

With the increasing volume of the web content, the amount of information covered by a centralized search engine naturally decreases [18]. Therefore a centralized approach of information retrieval no longer satisfies users' needs. A possible way of addressing this challenge is that considering the meta-search engines [9] which aim at increasing coverage by combining resources coming from different centralized search engines. Compared to a centralized approach, the big potential of a Distributed Information Retrieval (DIR) approach is the possibility of getting the information from different sources [37]. The two major problems in DIR that we study in this paper are: (1) source selection and (2) result merging. In fact, the large amount of results returned by meta-search engines is a potential source of noise. One possible way of reducing the noise is adopting personalization. Personalized meta-search engines use profiles of users to refine the results lists in order to return the relevant documents which better satisfy the users' information needs. Many kinds of personal data can be used to build the user profile such as: a user's manual selected interests [26], user's search history [23], etc. Internet growth and the advent of web 2.0 gave birth to different types of large scale social networks, for instance Facebook, Tumblr, Twitter, Flickr, Instagram, etc. These are now recognized as an important means for information dissemination [29], which allow internet users, to communicate and exchange ideas between them [7]. Many social networks are considered as social tagging systems. These systems allow users to provide annotations (tags) to resources, in order to label resources or documents by topic. Several social bookmarking services, such as Flickr[1] and Delicious,[2] are considered online folksonomy services. The term to describe these social tagging data, is known as folksonomies [11]. The set of tags can be used as a source of personal

* Corresponding author. Tel.: +213 662943801.
E-mail addresses: zakaria.saoud@live.fr (Z. Saoud), skechid@usthb.dz (S. Kechid).

[1] Flickr – photo sharing, http://www.flickr.com/.

[2] Delicious – social bookmarking, http://delicious.com/.

http://dx.doi.org/10.1016/j.ins.2015.12.012
0020-0255/

data to build the user profile. In this work, we propose novel personalization approach, which exploits the social relations between items and tags by utilizing a user profile. This profile is defined by a social tagging system to expand the queries and further to improve the result merging and source selection process. This paper is a revised and expanded version of the conference paper [30], which lacks the source selection process and the profile adaptation technique. The rest of the paper is organized as follows: Section 2 describes the related work. Section 3 introduces the details of our proposed personalized approach. Section 4 describes the experimental setup and results obtained. Finally, conclusion of current work and possible future work are listed in Section 5.

2. Related work

2.1. Social information retrieval

Social tagging systems are web-based systems that allow Internet users to add, edit, and share bookmarks of web documents. In social bookmarking services, such as Delicious and Flickr, the users can annotate their bookmarks with arbitrary keywords known as tags. The collection of a user's tags constitutes his/her personomy, and the collection of all users' personomy constitutes the folksonomy. A folksonomy is a tuple $F := (U, T, D, Y)$ where $U = \{u_1, \ldots, u_M\}$ is a set of users, $T = \{t_1, \ldots, t_L\}$ is a set of tags, and $D = \{d_1, \ldots, d_N\}$ is a set of resources or web documents, and Y is a ternary relation between U, T and D, i.e.: $Y \subseteq U \times T \times D$, whose elements are called tag assignments [11,39]. In our case, the elements of D represent the different web resources and are identified by their URL. Users are identified by their user ID. In social information retrieval, many studies have been proposed in the context of personalization search, and most of these studies are based on the folksonomy structure [4,17,32,39,41,42].

Schenkel et al. [31] developed a framework for exploiting such social relations to improve the effectiveness of search and recommendation. They created a scoring model that exploits social relations and semantic/statistical relations among items and tags. This scoring model gives a great importance to users who have a high score of friendship strength with the query initiator. The score of friendship strength is a linear combination of the spiritual friendship strength, the social strength friendship, and the global friendship strength. Rather than item recommendation, our personalized retrieval models, is composed of several web search engines to re-rank the lists of search results according to the user profile. This is applicable to meta-search engine.

Bender et al. [3] exploited the different entities of social networks (users, documents, tags) and social relations between these entities, to expand the queries by adding the similar tags to the query keywords, and make a social expansion to give an advantage to documents that tagged by the user's close friends. Jeon et al. [14] developed an approach that used the links and similarities between the user profiles in the filtering algorithm results. This approach is called collaborative filtering. The principal advantage of this approach is the enlargement of the coverage of research using similar profiles. For example, in the case where the user does not obtain satisfactory results for a query, we can then use the most similar profiles to expand the current user profile, to enlarge the search and retrieve more relevant results. Vallet et al. [39] presented a personalization model that exploits folksonomy structure. Two measures were developed to calculate the relevance between a user profile and a document to re-rank the list of results returned by a search engine. Lin et al. [20] proposed a novel personalized news recommendation framework called " PRemise". This framework integrated content information, user feedback, and social information, in order to resolve data sparsity and cold-start problems, to improve the recommendation efficiency. Kumar et al. [17] proposed two methods: SVD and modSVD, to build for each user, a Clustered User Interest Profile (CUIP), using the set of tags. Each (CUIP) contains many clusters, and each cluster identifying a topic of the user's interest. The matching cluster to a given user's query will be used to personalize the search results. Yu et al. [43] proposed a new friend recommendation method ACR-FoF (algebraic connectivity regularized friends-of-friends), which extends existing friend recommendation algorithms to achieve both recommendation relevance and content spread in a social network. Liu et al. [21] proposed a novel event-based recommendation approach, which combines the event trend analysis and personal preference to recommend blog articles of popular events that suit user interests. Shafiq et al. [32] proposed a new Web search personalized approach, which extracts the user preferences from the social network, to re-rank the results from a search engine.

2.2. Source selection approach

Source selection is a decisive step in the metasearching process [12]. It aims to reduce the number of selected sources for a given query, keep the search efficiency when there are many sources of information available, and only select the relevant sources to the user's query. Several source selection methods have been developed, and they can be classified into two main categories: manual selection methods, and automatic selection methods. Since we only focus on automatic selection, the manual selection methods are not discussed in this paper. In automatic selection, several approaches have been defined. Callan et al. [5] proposed an approach called CORI (Collection Retrieval Inference network), which considers a collection as a meta-documents, and the selection is made according the similarity between the user query and the collection. They used an inference network for ranking the collections, in which leaves denote the document collections, and nodes represent the terms that occur in the collection. Similar to documents ranking method, the collections are ranked on a score $p(Q|C_i)$ which