Utilizing Virtual Communities for Information Retrieval and User Modeling

Azza Harbaoui[1], Sahbi Sidhom[2], Malek Ghenima[1] and Henda Ben Ghezala[1]
[1]*Department of Computing, Laboratory RIADI, National School of Computer Sciences, Manouba, Tunisia*
[2]*Laboratory KIWI-LORIA, University of Lorraine, Vandoeuvre, France*

Keywords: Personalized Information Retrieval, User Modeling, User Profile, Virtual Communities.

Abstract: Internet has become the largest library in human history. Having such a large library made the search process more complicated. In fact, traditional search engines respond users by sending back the same results to different users having expressed different information needs and different preferences. A significant part of difficulties,report to vocabulary problems (polysemy, synonymy...). Such problems trigger a strong need to personalize the search results based on user preferences. The goal of personalized information is to generate meaningful results interesting to a number of information users using their profile. This paper presents a personalized information retrieval approach based on user profile. User profile is built from the acquisition of explicit and implicit user data. The proposed approach also presents a semantic-based optimization method for user query. The system uses user profile to construct virtual communities. Moreover, it uses the user's navigation data to predict user's preferences in order to update virtual communities.

1 INTRODUCTION

This paper deals with the personalized information retrieval (IR). In fact, classic IR methods usually intended for simple textual search, faced new heterogeneous documents and rich scalable contents. Consequently, the user is facing these evolutions and is being more and more unsatisfied, searching for IR search results quality. In order to overcome the limits of existing IRS, the main goal of personalized IR is to propose effective solutions, by focusing on the most relevant results to a user query. In this paper, we propose a novel approach able to capture what is relevant to a user. The proposal relies upon four components: (i) a process of disambiguation based on user's votes and domain ontology, (ii) multidimensional profile modeling based on explicit and implicit user's collaboration, (iii) construction of user's profile network, (vi) construction of user's influence network based on user's rate confidence and user's rate share.

2 RELATED WORKS

The main interest of personalized systems when extracting information is the use of a process that considers end-user's interests and preferences. The personalization process needs representing, accessing and storing a users personal information. We present in this section an overview of personalization and user profile.

2.1 Personalization

In fact, existing personalization approaches (Zhang et al., 2014) have contributed to the improvement of information systems use. A large body of research in Information Retrieval (IR) has highlighted that relevance is a complex and a challenging concept. However and despite their widespread, the underlying complexity stems mainly from the fact that relevance is estimated by considering multiple dimensions and that most of them are subjective since they are user-dependent. To answer this challenge, the commonly used approaches are based on specific user's preferences.

Positioning ourselves under a contextual or even personalized IR, access to a resource that is relevant and adapted to user context is a factor of Personalized Information Retrieval Systems (van Rijsbergen, 2013). Involving users in the search process requires modeling their profiles. Indeed, it was found that among the reasons behind lack of performance of the used personalization techniques is the integration of

Harbaoui, A., Sidhom, S., Ghenima, M. and Ghezala, H.
Utilizing Virtual Communities for Information Retrieval and User Modeling.
In *Proceedings of the 12th International Conference on Web Information Systems and Technologies (WEBIST 2016) - Volume 2*, pages 29-34
ISBN: 978-989-758-186-1

user profile regardless of the context of use. However, users are different in nature, with roles and diversified skills ; their preferences can be general, stable or recurrent. Any information forming a profile may not be appropriate for all search circumstances. Personalization systems often focus on partial aspects of users, and in this case the system uses explicit acquisition, where users express their personal data or preferences through the specification of their interests. This has the disadvantage of requiring users to get involved in the process, which imposes on them an additional charge on their use of the system. In addition, users, forced for example to fill out a form or answer a question, can introduce erroneous information, which may lead to insignificant results. Thus, the current IR problem faces real challenges that are closely related among others to the size of the documents on the web. The first challenge is to find an efficient way to provide users with the information they really need. As a result, several studies have considered integrating user profile in the search process. For ten years, research (Maleszka, 2015) on IR has been evolving and has been largely influenced by an interest in this area, wary to add a little more intelligence for improved relevance.

2.2 User profile

Modeling the user is at the center of the implementation of a personalized information retrieval process. The goal of user modeling is to select the most relevant information that reflects user's interests (Min and Jones, 2011). This modeling consists of designating a structure in which we store information that describes essentially :

- User interests ;
- Preferences ;
- Context ;
- Expected goal of the search;
- Individual traits;
- Experience.

There are several definitions of user modeling in the literature (Cheung et al., 1998), (Tanudjaja and Mui, 2002). We retain some below:
"A user model is a knowledge source in a natural-language-dialogue system which contains explicit assumptions on all aspects of the user that may be relevant for the dialogue behavior of the system" (Esparza et al., 2012)
"User model is an explicit representation of the system of a particular user's characteristics that may be relevant for personalized interaction." (Wen et al., 2004)
"The process of gathering information about the users of computer systems (Treur and Umair, 2011)and of making this information available to systems which exploit it to adapt their behavior or the information they provide to the specific requirements of individual users has been termed as user modeling."

Several techniques were developed in the literature to model the user. They differ according to the approach of profile representation and construction (Micarelli et al., 2007). As our main objective is to provide the end user with personalized search results, the proposed approach first builds user profile. It connects then user profiles to each other to construct the users virtual communities. The assumption is that when searching for a relevant document, the search system should use in addition to specific user's needs and previous searches, knowledge on other users. Search personalization is achieved by returning to the user ranked results using their profile and query.

As for our objectives, we propose an approach (*cf.* Figure 1) with four main components: (1) a user profile construction component, (2) an acquisition of user's data navigation component, (3) a virtual communities construction component and (4) an influence network component.

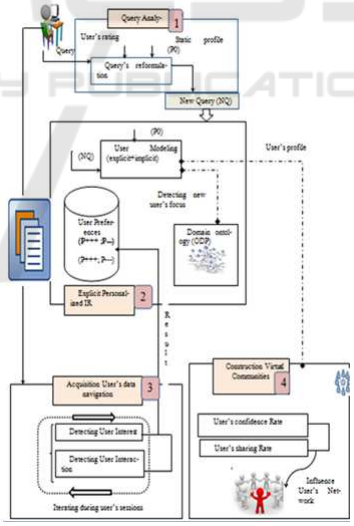

Figure 1: Proposed approach.

User profile is based on an implicit interaction with the user: implicit because the user is not directly asked to give opinion, and interactive through using navigation to measure interest in a given query.

2.3 Virtual Communities based User Profile

The social aspect of user's personal life is becoming increasingly important to examine. We define a virtual community as a group, often informal, of individuals with a common interest in sharing knowledge with members of the same community. They can view the search history, documents relating to previous visits, tags as well as part of the user profile visible to the public. A virtual community may be represented in many ways, we show in what follows its most common graphic representation (Yakoubi and Kanawati, 2013): a community represented by a graph is a set of points, some of which are directly connected by one or more links. It can be represented in many ways: networks waves ; hypertext Graphs; and social networks. In our paper, we used graphs to represent the user's network. Indeed, each node is simply a user profile. Each arc represents a relationship between two users. Each edge is weighted by a weight representing the degree of similarity between the edges previously connected by this arc. We also used community detection to identify the community to which a user of our system belongs. Automatic identification of communities has attracted much attention in recent years and many algorithms have been proposed to model them. These algorithms consider only networks structural data regardless of any other attributes specific to edges for example. Most communities detection methods are based on the intuition that the community structure of field graphs is naturally a hierarchical structure (a community consists itself of sub-communities that consist themselves of sub-sub communities and so on). In addition, if all the methods do not explicitly target maximization of modularity it is mainly used to compare the performance of methods. We opted for the approach of "Louvain" (Newman, 2004)to detect communities.

2.3.1 Louvain's Method

Louvain's method (Newman, 2004) has significant advantages compared to other community detection methods. It has been validated and tested successfully on several projects, including its speed (although this is not proven, the algorithm has a linear function or almost), allowing it to process graphs up with billions of links, its multi-scale aspect, which enables it to discover communities at different scales, and its excellent accuracy compared to other methods. For these reasons, we used this method to detect virtual communities.

2.3.2 Construction of User's Influence Network

An influence network is a virtual community consisting of user profiles, some of which may be influential profiles. These members, through their profile, allow for influencing members who belong to the same community, and thus influence their research. Accordingly, the virtual community is reduced to a subcommunity or even an influence network where the search space is reduced, thereby improving the search results. To build this network, we defined two variables that we called trust rate and sharing rate. These will be described in the following section, applying an iterative two-stage succession.

2.3.3 Sharing Rate

Sharing rate, denoted ω_p, is a numeric value that ranges between 0 and 1. specifies sharing rate of a document by any user within his/her community. ω_p determines thus the rate of occurrence of a document shared by a given user in the community to which he/she belongs. Its formula is defined as follows:

$$\omega_p = \frac{number\_friends\_user(i)}{total\_number\_user's\_community} \quad (1)$$

2.3.4 Trust Rate

Under certain conditions, a community is the focus of a user. Trust level of a user j in a user i is denoted by ω_{cij} and given by the following formula:

$$\omega_{ci,j} = \frac{\sum_{j=0,\ i\neq j}^{n} \Omega_{c_{i,j}}}{n-1} \quad (2)$$

2.3.5 Influence Network Constructing Steps

The construction of an influence network allows for extracting in a clear way the various friendship links between a user as represented by his/her profile and other members of the community. To do this, we used the following approach: For a given user U, we extract the network of all his/her friends as a first stage and then, for each profile of the community, we define an influence rate, deduced from equations. The system derives the nature of the relationship that links user U to his/her friends. This deduction is based on thresholds according to well-defined criteria, namely;

- If influence rate $>$ **0,9** : this friend is dominant.
- If 0.6 $<$influence rate **0,9** : this is a helping Friend.
- If 0.3 $<$influence rate $<$**0,6** : this is a cooperative friend.

- If influence rate <**0,3** : this is a submissive friend.

We report in the following, an illustrative example of a user. The output of the system after closure of friends and the influence network deduced is shown in (*cf.* Figure 2).

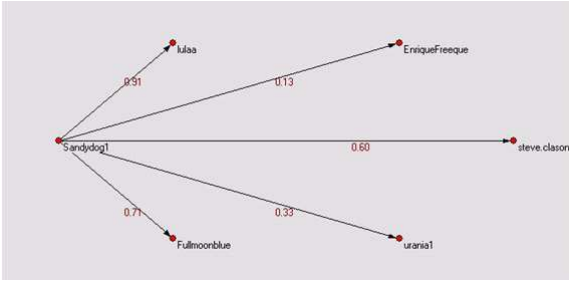

Figure 2: Example of Sandydog's influence network.

2.4 Case Study

After constructing the influence network, the user focuses on his/her dominant friend and launches a research through a Case-based reasoning process (CBR). To do this, the user will enter the identifier of his/her friend, the identifier of his/her query and the assigned ratings. Thus, the system will return 20 documents deemed relevant to the dominant friend of the user. This process is supported by the Jcolibri framework. We opted this object-oriented framework, because it, inter alia, facilitates the construction of CBR systems. This framework ensures the management of the case base (add, edit, and delete cases). Just give the case a structure (in this case, information about the dominant friend), and then call these operations to manage the case base. It also offers a rich set of predefined, reusable similarity functions to calculate similarity between ontological concepts.

3 EXPERIMENTATION AND RESULTS INTERPRETATION

For experimentation and evaluation purposes, we empirically tested our approach on the Book Social Search Dataset (Koolen et al., 2012). The Dataset consists of 2,8 million book records from Amazon, extended with social metadata. First, the experimental users were asked to notify their corresponding relevant results and then precision, recall, Mean Average Precision (MAP) were calculated. The evaluation protocol was designed to tune the experimentation parameters and then to evaluate the effectiveness of our personalized approach. It is based on three stages :

The first is to measure the impact of a user's implicit interest in a set of documents. Thus, in addition to the users explicit knowledge, we were able to highlight a set of implicit knowledge. The second, in addition to the query processing process, will measure the impact of the integration of user profile, on improving search performance. The last is to test the contribution of integrating virtual communities on improving the research process. The objective of these experiments is twofold: first, to prove the applicability of the approach and second to compare and validate each contribution against the baseline.

3.1 The Scenario Process

We proceed to the evaluation of the personalized information retrieval approach through the following stages :

1. Analysis of the query (reformulation and disambiguation) by using the rates and tags added by users after a first search. Disambiguation is done by querying the data dictionary WordNet ;
2. Construction of the user profile by adopting a multidimensional representation and hybrid acquisition of navigation traces (implicit and explicit aquisition). Detection of new interests of users is done by querying the ODP-domaine ontology ;
3. Construction of virtual communities through the acquired user profiles and restarting the search process to assess the impact of integration in the research process ;
4. Detection of influence networks (reduced virtual communities).

This network is integrated back into a process of indexing and searching in order to evaluate the performance of our method and its impact on the performance of indexing in information search. To validate the proposed approach, we conducted tests, using the experimental toolkit LEMUR[1] and the weighting scheme $tf \times idf$.This configuration is used as the baseline for our comparative evaluation. Next, we calculated improvement across the different scenarios and the baseline system.

3.2 Experimental Results

Through our experiments, we try to evaluate the effectiveness of our proposal over various real user's queries. To estimate the quality of the results obtained under different scenarios, we used standard measures of precision @X, specifically**P@10**, **P@20** and**P@30** documents which are respectively average

[1] http://www.lemurproject.org/

precision of the 10, 20 and 30 first documents returned, the NDCG (Normalized Cumulative Gain) and MAP (Mean Average Precision) of all 40 selected queries. For each query, the first 100 records are returned by the experimental SRI Lemur and average precisions are calculated to measure the relevance of the system. The presented results compare precision, MAP and NDCG values between the personalized approach and the baseline. We notice that a precision at 5 of a personalized search is better than the baseline. The sheet at (*cf.* Figure 3). presents the search results quality measurement with virtual communities. We calculate **P@5**, **P@10**, and MAP(*cf.* Figure 4) for a set of 40 queries. We can see in this sheet that **P@10** presents more accurate results. Moreover, we see that performance improvement is better for **P@10** or **P@15** than **P@30** and **P@100**. This improvement can be explained by the fact that there is less irrelevance when we consider the first 10 or 15 results. We can also conclude that the personalized search including virtual communities improves system precision.

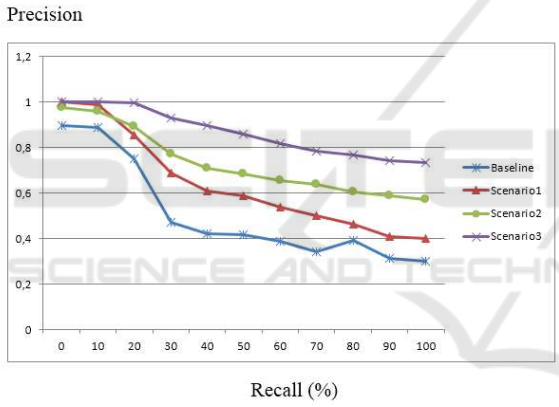

Figure 3: Recall/Precision Curve.

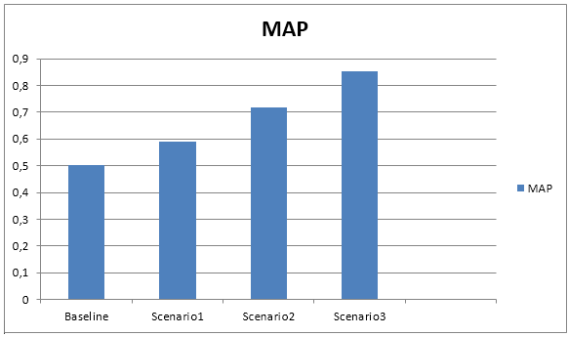

Figure 4: MAP histogram.

4 CONCLUSIONS

We began this paper with an overview of personalization. Then we presented our proposal of using virtual communities to model users for information retrieval. The proposed approach includes the creation of user profiles, the construction of virtual communities and influence networks. We conducted an experimentation and evaluation phase on INEX Book Social Search Dataset (Koolen et al., 2012). Evaluation shows that the system improves search results when integrating virtual communities and influence networks. However, these results show some limitations. In fact, the process of user profile construction and that of virtual communities construction are not synchronized. The two processes cannot take place simultaneously. Profile construction is always done in batch mode in the case of a cold start.

REFERENCES

Cheung, D. W., Kao, B., and Lee, J. (1998). Discovering user access patterns on the world wide web. *Knowledge-Based Systems*, 10(7):463–470.

Esparza, S. G., OMahony, M. P., and Smyth, B. (2012). Mining the real-time web: a novel approach to product recommendation. *Knowledge-Based Systems*, 29:3–11.

Koolen, M., Kazai, G., Kamps, J., Doucet, A., and Landoni, M. (2012). Overview of the INEX 2011 Books and Social Search Track. In Geva, S., Kamps, J., and Schenkel, R., editors, *Focused Retrieval of Content and Structure : 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2011*, volume 7424 of *Lecture Notes in Computer Science*, pages 1–29. Springer.

Maleszka, B. (2015). An adaptation method for hierarchical user profile in personalized document retrieval systems. In *Intelligent Information and Database Systems*, pages 107–116. Springer.

Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S. (2007). Personalized search on the world wide web. In *The adaptive web*, pages 195–230. Springer.

Min, J. and Jones, G. J. (2011). Building user interest profiles from wikipedia clusters.

Newman, M. (2004). Detecting community structure in networks. *European Physical Journal*, 38:321–330.

Tanudjaja, F. and Mui, L. (2002). Persona: A contextualized and personalized web search. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, pages 1232–1240. IEEE.

Treur, J. and Umair, M. (2011). An agent model integrating an adaptive model for environmental dynamics. *International Journal of Intelligent Information and Database Systems*, 5(3):201–228.

van Rijsbergen, K. (2013). The roots of the theoretical basis for information retrieval. In *International Conference on the Theory of Information Retrieval, ICTIR*

'13, Copenhagen, Denmark, September 29 - October 02, 2013, page 19.

Wen, J.-R., Lao, N., and Ma, W.-Y. (2004). Probabilistic model for contextual retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 57–63. ACM.

Yakoubi, Z. and Kanawati, R. (2013). Leader-driven approach for community detection in complex network. In *proceedings of the international conference on intercations in complex systems*.

Zhang, D., Song, T., Li, J., and Liu, Q. (2014). A linked data-based framework for personalized services information retrieval in smart city. In *Web Information Systems Engineering–WISE 2013 Workshops*, pages 461–473. Springer.