Knowledge-Based Systems 000 (2016) 1–15

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Use of textual and conceptual profiles for personalized retrieval of political documents

Eduardo Vicente-López, Luis M. de Campos*, Juan M. Fernández-Luna, Juan F. Huete

Departamento de Ciencias de la Computación e Inteligencia Artificial, E.T.S.I.I.T., CITIC-UGR, Universidad de Granada, 18071-Granada, Spain

ARTICLE INFO

Article history:
Received 24 May 2016
Revised 4 August 2016
Accepted 7 September 2016
Available online xxx

Keywords:
User profile
Personalization
Information retrieval
Privacy-enhanced
E-Government

ABSTRACT

The amount of information we are exposed to on a daily basis is increasing exponentially. Besides, Information Retrieval Systems (IRSs) return the same results for a given query regardless of who submitted it. In order to address the problems of finding useful, relevant information and adapting the results to the user, the use of personalization techniques is now more necessary than ever. They are not, however, particularly popular in live environments as users remain unconvinced about their reliability and, more importantly, are concerned about privacy issues. We have developed and compared six generic user profile representations in order to improve the personalization process and address the problem of privacy. We propose a new weighting scheme to build the profiles and a new personalization technique to join the advantages of some of the previous profiles. A comprehensive evaluation study of the proposed generic user profiles was performed and this revealed very good personalization performance results and some interesting conclusions about their use in a political context, more specifically with official documents from the Andalusian Parliament.

1. Introduction

Nowadays, most information is created and exchanged in digital format with an exponential increase in recent years [9]. The e-Government framework is a specific yet important area of application. In this article, we focus on the parliamentary context in which vast amounts of information have been published. An abundance of information is pointless, however, if citizens and politicians are unable to find the relevant documents that match their information needs, generally related to problems affecting their daily lives.

Most parliaments have two main official publications: the records of the parliamentary proceedings (plenary and committee sessions) and the official bulletins. In a plenary session, political groups will present their proposals which are then debated and put to the vote. Committee sessions, on the other hand, cover particular fields such as agriculture, education or the economy. Each parliamentary proceedings document contains the full transcriptions of the speeches given by the members of the parliament in each session. The main component of these documents is the *initiative*, which presents a detailed discussion of a specific issue. Each initiative is then manually tagged by expert parliamentary documentalists with one or more subjects from the EUROVOC[1] thesaurus in order to classify the content.

Our research group has collaborated with the Andalusian Parliament since 2005 and has had access to their official publications. These are in XML format and some comprise the document collection used in the evaluation process in this article. As a result of this collaboration, we have built the *Seda*[2] IRS [11] in order to improve public access to these official parliamentary documents.

In most cases, traditional IRSs are used to access to parliamentary documents facing the following main problems: a large amount of information is available, users tend to formulate short and ambiguous queries [30], and little is known either about the users or their information needs except for their query keywords. As a result, IRSs tend to retrieve the same results for a given query regardless of the user. This issue is known as the *one size fits all* problem and personalization [4,5,18,31,36] offers a possible solution. In personalization, both the user and the query are important in the retrieval process. The main objective of personalization is to retrieve results which best suit the user to better satisfy the user's specific information needs, thereby improving the user satisfaction with the IRS.

* Corresponding author. Fax: +34 958243317.
E-mail addresses: evicente@decsai.ugr.es (E. Vicente-López), lci@decsai.ugr.es (L.M. de Campos), jmfluna@decsai.ugr.es (J.M. Fernández-Luna), jhg@decsai.ugr.es (J.F. Huete).

[1] http://eurovoc.europa.eu/
[2] http://irutai2.ugr.es/SEDA/

http://dx.doi.org/10.1016/j.knosys.2016.09.005
0950-7051/

Although we did intend to introduce various personalization features into *Seda*, we encountered certain privacy issues since Parliament does not allow any personal data to be collected about members of parliament or the public. This is the norm rather than the exception and is increasingly becoming a major barrier to personalized IRSs [28]. According to [19], approximately 85% of users are concerned about the privacy or security of their online personal information, 90% have at some time refused to provide online personal information and 35% supply false online personal information.

This article therefore has two main objectives: 1) to provide an alternative option for implementing personalization in privacy-constrained environments, with a good performance and without the need to collect any personal information. This would be achieved by means of *generic profiles* which are learned from the document collection content; in our particular case from the Andalusian Parliament *committee sessions* which cover specific areas of interests, and 2) to be able to select the best representation of the previous generic profiles and configuration parameters to be used for any given personalization technique and retrieval scenario.

While these generic profiles might be considered *'unrealistic'*, since they do not represent real users, they are a valid approach [10,29] for possible users interested in certain areas. This is particularly true in our political context, where one politician may serve on several committees. Without the use of generic profiles, the user might also wish to include additional query terms to try to describe the committee session content. This might be difficult for the user and may also trigger the *query-drift* problem [43]: the inclusion of possibly unrelated terms in the original query might result in the retrieval of unexpected results which might not contain the original query terms. The user might also choose to filter out all documents that do not belong to the committee sessions but in doing so, certain relevant results might be omitted (about 25% according to our studies). Furthermore, a filtering approach is not a valid solution since relevant documents might be found not only in committee sessions but also in plenary sessions and official bulletins, and since these are not implicitly classified, they will not therefore be retrieved. As a result, the final percentage of possible relevant missed documents will be much higher than the previous value. The best approach is therefore to use some kind of personalization.

In order to achieve this article main objectives, we propose six different ways to represent the generic user profiles learned from the document collection content. These are based on general topic areas which are subsequently selected by the users according to their interests and preferences. It should be noted that these areas are quite well defined and characterize parliamentary activities. These profiles are ideal for introducing personalization into privacy-constrained environments where users are reluctant to reveal personal information, as occurs in the case of the Andalusian Parliament.

Although these profiles have been used for a political context, they can also be applied in other privacy-constrained retrieval environments. The only requirement for building our generic user profiles is to have a collection where at least a subset of its documents can be classified into different areas of interest or categories, that future users might find interesting. If this were not the case, a clustering process could be used to find clusters of similar documents according to their content, and subsequently a classification process can assign new documents to the corresponding clusters. In our case, since each document in the document collection belongs to one committee session, we have an implicitly classified document collection.

We next expose our contributions to achieve the objectives of this article: 1) the development of user profiles based only on terms from documents belonging to a given area of interest (committee sessions) irrespective of where they appear in the document; 2) the proposal of a new weighting scheme for profile items called *diffFreq* and we have shown how this is superior to the common *tf*idf* approach [26], at least in these category-based generic profiles; 3) the construction of user profiles based on the EUROVOC thesaurus subjects which are manually assigned to each initiative. Although this approach might seem promising, it has not been confirmed by our results. Since we still believed that subjects should add value, we designed different ways to obtain the maximum benefit from subjects and terms simultaneously; 4) the development of a new personalization technique which uses subjects and terms from the previous profiles with reasonably good results; 5) a *hybrid* user profile (with four variations) comprising both subjects and terms. We shall explain how each approach should be used and the results obtained; and 6) a comprehensive evaluation and comparative study of all of the proposed user profile representations.

Although generic profiles are frequently used in personalized contextual evaluation environments, e.g [29,33]., most personalized IRSs are not validated with real world experiments [42], since they are extremely difficult due to their complexity and the potential costs involved. However, these experiments are necessary to demonstrate the true effectiveness and improvements of any personalized IRS over other systems. Various efforts have been made to solve this problem, such as for example [40] where the authors present an easy automatic methodology to evaluate these personalized IRSs. With the previous comprehensive evaluation (sixth contribution) we provide the best generic profile representation and configuration parameters to be used for a given personalization technique and retrieval scenario. We have also obtained very good personalized results with a retrieval performance improvement of up to 80.17% on the non-personalized search, together with some interesting conclusions about the merits of using these generic profiles.

The remainder of the article is organized as follows: Section 2 reviews the different user profile approaches in the literature; Section 3 describes profile construction, use and the results obtained for the newly developed term-based and subject-based profiles; Section 4 explains how subjects and terms can be combined to work together, firstly with the newly developed personalization technique, and secondly with the *hybrid* profiles; Section 5 compares all of the developed profiles and presents some interesting conclusions; and finally Section 6 outlines the general conclusions of the article and proposals for future research.

2. Related work

There are three main stages to any IR personalization process: the first is to acquire and represent the user context in the user profile; the second is to exploit the user profile information in the retrieval process as well as possible; and the third is to evaluate the entire personalization process. Some additional issues may also be considered such as privacy when collecting or managing personal data [19], or different ways of presenting the personalized results [2] as simply and as intuitively as possible.

The quality of the personalized results is highly dependent on the quality of the user profile and how well its information is exploited in the retrieval process. The user profile building process is therefore an extremely important step in order to obtain good personalized results. We can see the importance of building accurate user profiles even applied to other domains such as social media [21] or IR related fields such as recommender systems [3].

The authors in [14] outline the following three main stages within the IR user profile building process related to the user information: 1) to collect it; 2) to represent it; and 3) to keep it up-

Please cite this article as: E. Vicente-López et al., Use of textual and conceptual profiles for personalized retrieval of political documents, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.09.005