RES: A Personalized Filtering Tool for CiteSeerX Queries Based on Keyphrase Extraction

Dario De Nart, Felice Ferrara, and Carlo Tasso

Artificial Intelligence Lab
Department of Mathematics and Computer Science
University of Udine, Italy
{dario.denart,felice.ferrara,carlo.tasso}@uniud.it

Abstract. Finding satisfactory scientific literature is still a very time-consuming task. In the last decade several tools have been proposed to approach this task, however only few of them actually analyse the whole document in order to select and present it to the user and even less tools offer any kind of explanation of why a given item was retrieved/recommended. The main goal of this demonstration is to present the RES system, a tool intended to overcome the limitations of traditional recommender and personalized information retrieval systems by exploiting a more semantic approach where concepts are extracted from the papers in order to generate and then explain the recommendation. RES acts like a personalized interface for the well-known CiteSeerX system, filtering and presenting query results accordingly to individual user's interests.

1 Introduction

Reading scientific literature is a critical step for conceiving and developing scientific projects, however it still remains an expensive task. Actually, systems such as CiteSeerX, Google Scholar, and Mendeley allow researchers to discover new knowledge by querying and browsing millions of publications. Many systems have been developed to provide personalized access to such a huge amount of information, but all of them present their results as a bare list of items, occasionally displaying some lines extracted from the abstract, leaving to the user the burden of checking whether the recommendation is relevant or not by reading the paper. Moreover, most of those systems use collaborative mechanisms, meaning that the recommendation is driven by other users' behaviour, a principle that may not work well for small research communities due to sparsisty and cold start issues [4]. We claim that a more semantic and content-based approach, assuring that the recommendation is driven by the actual content included in the paper, and a brief, yet detailed and informative, explanation of the recommendation could save much time and effort and could lead to a greater user satisfaction. In order to support our claim, we present RES, a Recommendation

S. Carberry et al. (Eds.): UMAP 2013, LNCS 7899, pp. 341–343, 2013.

and Explanation System using a completely content-based approach, based on the use of Keyphrases (KP), i.e. short phrases of up to three words. Using KPs instead of keywords allows to preserve information about the context in which terms are used and, moreover, KPs have a high cognitive plausibility. KPs are automatically extracted from texts by means of Dikpe, a Keyphrase Extraction tool previously developed in our laboratory [2].

2 Related Work

In [5] several collaborative techniques for recommending papers of CiteULike are presented and discussed. In [1], the relations involving users, publications, tags, and other metadata are used to produce a graph for computing personalized suggestions, without analysing the document content. Many other examples, here omitted due to shortage of space, feature other kinds of collaborative and/or metadata-based approaches. Content-based and hybrid approaches are used as well: in [3], the authors propose a filtering system based on keyphrase extraction for identifying potentially relevant documents, yet there is no explanation of the resulting recommendation. [4] points out how recommendation strategies based on explicit decision models are able to offer adequate explanations. Finally, [6] discusses the benefits of explanations on user satisfaction.

3 System Overview

The RES system includes a database called *SPC* (Scientific Paper Collection) and the following three main modules:

1) A *Web User Interface Module* devoted to: (i) let the user create and edit his profile, (ii) query CiteSeer, and (iii) access the recommended items. Recommendations are presented as a ranked list of documents where the top items are those that better match the user profile. For each document two lists are presented: KPs appearing in both the user profile and in the document and relevant KPs present in the document but not in the user profile.

2) A *KP Extraction Module*, devoted to: (i) gather CiteSeer results,(ii) extract KPs from each article, and (iii) store its representation in the SPC.

3) A *Recommendation Engine Module* devoted to: (i) build and maintain personalized user profiles representing specific interests of the user and (ii) matching user profiles against document representations stored in the SPC.

Our recommendation strategy is document-centric: in order to create a user profile the system requests the user to enter one or more sample articles or paragraphs that summarize his/her interests and then filters search results according to the similarity between them and the profile. Document contents are represented as lists of KPs, which are split into single terms in order to create a graph representation where terms are nodes and the arcs represent co-occurrence in the same KP. This modelling technique allows RES to build, for each term, a meaningful context of interest by simply checking its adjacency list. The same

term used in the same context in different articles should reasonably refer to the same adjacent concepts, showing in such a way a certain degree of similarity: more shared concepts indicate higher similarity. The full recommending algorithm takes into account also the TF-IDF weights of shared terms in order to penalize trivial associations.

4 Evaluation and Conclusions

In the first development stage of the system, we have performed a limited number of formative tests, mainly aimed at experimenting different system parameterizations. A set of over 300 scientific papers dealing with Recommender Systems and Adaptive Personalization was stored in the SPC and manually classified according to 16 topics. Later, 200 uncategorized documents dealing with several random ICT topics were added in order to create noise in the data. 250 user profiles were automatically generated for each one of the 16 topics by means of groups of 2, 4, 6, and 10 seed documents; then, for each user profile, RES and a baseline reference system (TF-IDF based ad-hoc developed) were compared. For each recommendation, every recommended item dealing with the same topic as the seed document was considered as a good recommendation. We have defined the *accuracy* as the average percentage of good recommendations over the total recommended items. Results gathered so far are very promising since RES outperformed the baseline mechanism: accuracy of 57% vs baseline accuracy of 42% with 2 seed documents, up to accuracy of 72% vs baseline accuracy of 60% with 10 seed documents. Despite encouraging results, the improvement of RES is ongoing, focusing on: adapting it to other scientific collections, to reduce the time needed for content analysis, and to envisage more effective procedures for creating profiles. Evaluation is ongoing and in the future it will address the quality and the impact of the produced explanations.

References

1. Doerfel, S., Jäschke, R., Hotho, A., Stumme, G.: Leveraging publication metadata and social data into folkrank for scientific publication recommendation. In: Proceedings of the 4th ACM RecSys Workshop on Recommender Systems and the Social Web, pp. 9–16. ACM, New York (2012)
2. Ferrara, F., Pudota, N., Tasso, C.: A keyphrase-based paper recommender system. Digital Libraries and Archives, pp. 14–25 (2011)
3. Govindaraju, V., Ramanathan, K.: Similar document search and recommendation. Journal of Emerging Technologies in Web Intelligence 4(1), 84–93 (2012)
4. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender systems: an introduction. Cambridge University Press (2010)
5. Parra, D., Brusilovsky, P.: Evaluation of collaborative filtering algorithms for recommending articles on citeulike. In: Proceedings of the Workshop on Web, vol. 3. Citeseer (2010)
6. Zanker, M.: The influence of knowledgeable explanations on users' perception of a recommender system. In: Proceedings of the Sixth ACM Conference on Recommender Systems, pp. 269–272. ACM, New York (2012)