Personalized Access to Scientific Publications: from Recommendation to Explanation

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. Several recommender systems have been proposed in the literature for adaptively suggesting useful references to researchers with different interests. However, in order to access the knowledge contained in the recommended papers, the users need to read the publications for identifying the potentially interesting concepts. In this work we propose to overcome this limitation by utilizing a more semantic approach where concepts are extracted from the papers for generating and explaining the recommendations. By showing the concepts used to find the recommended articles, users can have a preliminary idea about the filtered publications, can understand the reasons why the papers were suggested and they can also provide new feedback about the relevance of the concepts utilized for generating the recommendations.

1 Introduction

Reading scientific literature is a critical step for conceiving and developing scientific projects, but finding appropriate literature for scientific researches is still an expensive task. Recommender systems have been also utilized to support researchers since these tools can filter information according to the personal interests of the users. However, by just filtering a list of scientific papers, current systems provide only a basic support to the user (a list of potentially relevant papers), whereas it would be much more useful to highlight in the recommended paper concepts and knowledge relevant for the user. As a consequence, such approach still leaves a lot of work to the user who both (i) has to read the paper in order to identify the main concepts in the recommended paper and (ii) cannot understand why the paper is actually recommended to him. On the other hand, we claim that more semantic approaches can be integrated for overcoming these drawbacks and, in particular, in this paper we use keyphrases (KP) for: modeling the user interests, computing the utility of a resource for a user, explaining the recommendations, and collecting feedback from users in a quite unobtrusive way.

A keyphrase is a short phrase (typically constituted by one to three/four words) that indicates one of the main ideas/concepts included in a document. A keyphrase list is a short list of keyphrases that reflects the content of a single

S. Carberry et al. (Eds.): UMAP 2013, LNCS 7899, pp. 296–301, 2013.

document, capturing the main topics discussed and providing a brief summary of its content. In this work, a user profile is built by exploiting the keyphrase lists extracted from the papers which are relevant to a specific user. Then, in order to compute the relevance of a new article, the user profile is matched with the keyphrase list extracted from that article. More interestingly, the explicit representation of the scientific papers is used for explaining why the system recommended the documents by showing: (i) the keyphrases which are both in the user model and in the paper and (ii) other keyphrases found in the document which are not yet stored in the user model but can support the user in understanding/evaluating the new paper. The explanation of recommendations by means of keyphrases produces several benefits. First of all, the user satisfaction can be increased since explanations save his time: the user is not forced to read the entire document in order to catch the main contents of the paper. The trust of the users in the system can be increased as well since, by showing the recommended concepts, the user can better understand the criteria utilized by the system for computing the recommendations. Finally, by showing the concepts available in the user model as well as in the recommended papers we provide a simple way to the user for refining his interests: the user can decide to add a new concept to his profile or decrease (or even cancel) the relevance of a concept.

The paper is organized as follows: Section 2 reviews related work, the proposed approach is illustrated in Section 3, the evaluation performed so far is described in Section 4, and Section 5 concludes the paper.

2 Related Work

Several recent works focused on filtering relevant publications from huge collections of papers by exploiting both collaborative filtering approaches [3] and content-based mechanisms [4]. However, as shown in [1], improving the accuracy of the recommendation is not the only goal of researchers who work on scientific paper recommendation. In fact, the access to the knowledge stored in the recommended papers can be simplified by providing new services for accessing recommendations, new navigational interfaces, and new visualizations techniques. By following this research directions, in this paper, we propose a mechanism where the recommendations of scientific papers are explained to the users by showing the main concepts which are in the recommended publications. There is actually a growing interest on explanation of recommendations since, as showed by Zanker, explanations are essential to increase the users satisfaction [7]. The impact of explanations is also shown in [5], where the authors also provide a taxonomy of explanations by identifying three explanation styles: (i) the *human style* which provides explanations by showing similar users, (ii) the *item style* where similar items are reported as explanation (iii), and the *feature style* where the main features of the recommended items are shown. In our work, relevant concepts are extracted from scientific publications for both generating the recommendations and providing feature style explanations. The idea of representing the interests of researchers as concepts extracted from publications was also proposed in [2], where the authors train a vector space classifier in order to associate terms (i.e. unigrams) to the concepts of the ACM Computing Classification System (CCS). The hierarchical organization of the CCS is used to represent user interests and documents as trees of concepts which can be matched for producing recommendations. Our approach, on the other hand, does not need a training phase since we adopt more sophisticated NLP techniques for identifying relevant concepts (i.e. keyphrases constituted by n-grams) included in the papers. Since n-grams provide a more significant and detailed description of the ideas reported in publications, we use them in order to generate the explanation.

3 Recommendation and Explanation by Using Keyphrases

In order to support our claims we developed a recommender system, named *Recommender and Explanation System* (*RES*), which is aimed at supporting researchers by adaptively filtering the scientific publications stored in a database called *SPC* (Scientific Paper Collection). Each paper uploaded in the SPC is processed by using the Dikpe KP extraction algorithm (described in [6]) in order to represent each paper as a set of KPs. Given a paper, Dikpe extracts from it a list of keyphrases where each KP has a weight (called *Keyphraseness*) that summarizes the several lexical and statistical indicators exploited in the extraction process. Higher is the Keyphraseness, more relevant is the KP.

Keyphrases are used to represent documents as well as to model user interests. More specifically, user models and documents are represented by a network structure called *Context Graph (CG)*. For each document stored in the SPC, a CG is built by processing its KP list. User profiles, on the other hand, are obtained by collapsing the CGs built for the documents marked as interesting by the user (as shown in Figure 1) and, possibly, enriched with other KPs gathered via relevance feedback.

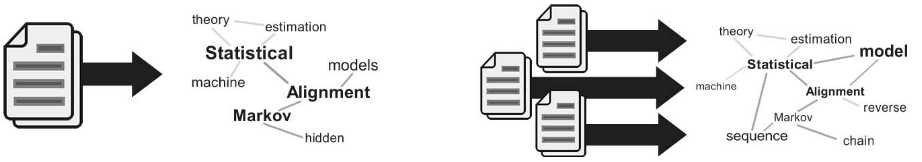


Fig. 1. Document content model and user profile construction

CGs are built by taking into account each single term belonging to each KP: each term is represented by a node of the graph and if two terms belong to the same KP their corresponding nodes are connected by an arc. Both nodes and arcs are assigned a weight which is the normalized sum of the Keyphraseness values associated to each KP containing the corresponding terms (such values are computed by the KP Extraction module). Heavy KPs will generate heavy nodes and arcs; term occurring in several KPs will generate heavy nodes and heavy arcs denoting frequent associations. Terms that never appear together in a KP won't have any direct link, but, if used together in the same context, may be connected indirectly, through other nodes, allowing the system to infer implicit KPs: for example the KPs "Markov alignment" and "hidden alignment" produce arcs that make possible the matching of the "hidden Markov alignment" KP. Finally, terms and phrases that are not used in the same context, won't be connected, creating isolated groups in the CG. Breaking KPs and then organizing terms in such a structure allows us to build, for each term, a meaningful context of interest, making it possible to disambiguate polysemic words in a better way than by matching phrases. Recommendations are provided to the users in three steps, as shown in Figure 2: Matching/Scoring, Ranking, and Presentation.

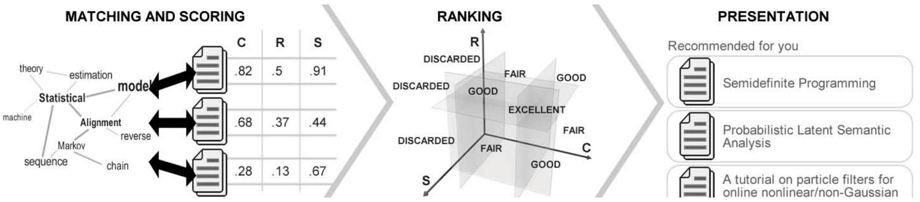
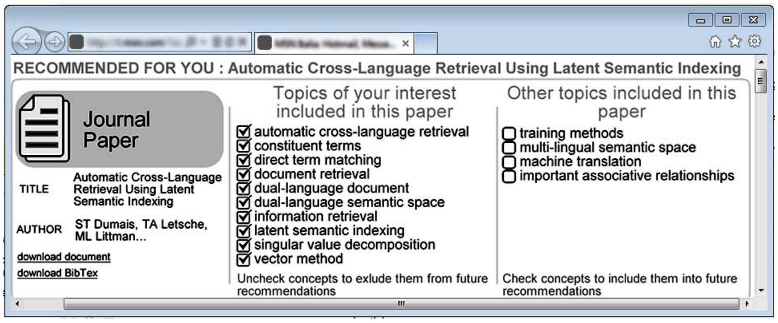


Fig. 2. The three steps of the recommendation process

In the first step every document in the SPC is matched against the user model by calculating the following parameters: *Coverage (C)*, *Relevance (R)* and *Similarity (S)*. C is the count of shared nodes between user and document CG, divided by the number of nodes in the document CG; by default, documents under a 10% coverage threshold are discarded, since the shared nodes are not enough for a meaningful ranking. R is the average TF-IDF measure of shared terms. S is the sum of the weights of shared arcs divided by the sum of the weights of all arcs occurring between shared nodes in the user CG. This last parameter is intended to assess the overlap between the two CGs and to measure how relevant are the shared arcs. In this way, each document is considered a point in a 3-dimensional space where each dimension corresponds to one of the three above parameters. In the Ranking phase, the 3-dimensional space is subdivided into several subspaces according to the value ranges of the three parameters, identifying in such a way different regions in terms of potential interest for the user. High values for all three parameters identify an excellent potential interest, while values lower than specific thresholds decrease the potential interest. Five subspaces are identified from *excellent* to *discarded* and each document is ranked according to where its three-dimensional representation is located. Being the system an experimental testbed, such threshold values have been manually tuned. Finally, in the Presentation step, documents are sorted by descending ranking order and the top ones are suggested to the user. 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 keyphrase (KP) lists are presented to the user: (i) KPs



RECOMMENDED FOR YOU : Automatic Cross-Language Retrieval Using Latent Semantic Indexing

Journal Paper

TITLE Automatic Cross-Language Retrieval Using Latent Semantic Indexing

AUTHOR ST Dumais, TA Letsche, ML Littman...

download document

download BibTex

Topics of your interest included in this paper

- ☑ automatic cross-language retrieval
- ☑ constituent terms
- ☑ direct term matching
- ☑ document retrieval
- ☑ dual-language document
- ☑ dual-language semantic space
- ☑ information retrieval
- ☑ latent semantic indexing
- ☑ singular value decomposition
- ☑ vector method

Uncheck concepts to exlude them from future recommendations

Other topics included in this paper

- □ training methods
- □ multi-lingual semantic space
- □ machine translation
- □ important associative relationships

Check concepts to include them into future recommendations

Fig. 3. Recommendation screenshot

appearing in both the user profile and in the document and (ii) relevant KPs present in the document but not in the user profile. This information, shown in Figure 3, serves two goals: it briefly explains why a document was recommended by highlighting its main concepts and, secondly, it offers the user a way to provide relevance feedback on the concepts extracted from each article.

4 Evaluation

In the first development stage of the system, we have performed a limited number of formative tests, mainly aimed at exploring different system tunings. A set of over 300 scientific papers dealing with Recommender Systems and Adaptive Personalization was collected and 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 and the whole set was processed, generating a test SPC. 250 user profiles were automatically generated for each one of the 16 topics using groups of 2, 4, 6, and 10 seed documents respectively; then, for each user profile, RES and a baseline reference system (ad-hoc developed) based upon the well-known and established TF-IDF metric, produced the 10 top-recommended items. For each recommendation, every recommended item dealing with the same topic as the seed document was considered a good recommendation. We have defined the *accuracy* as the number of good recommendations over the total recommended items, and averaging the accuracy values. Results gathered so far are very promising since RES outperformed the baseline mechanism when the user profile was built by using 2 seed documents (RES accuracy=0.57, baseline accuracy=0.42), 4 seed documents (RES accuracy=0.66, baseline accuracy=0.53), 6 seed documents (RES accuracy=0.70, baseline accuracy=0.55), and 10 seed documents (RES accuracy=0.72, baseline accuracy=0.60). Future evaluations will address the quality and the impact of the produced explanations.

5 Conclusion

By just filtering collection of papers, state-of-the-art recommender systems still leave a heavy work to researchers who have to spend efforts and time for accessing the knowledge contained in scientific publications. This issue is faced in this paper, where we propose a mechanism where concepts are automatically extracted from papers in order to generate and explain recommendations.

According to our first experiments the extraction of concepts can produce accurate recommendations and, at the moment, we are evaluating the effectiveness of the explanations in an on-line evaluation scenario, exploiting the system to filter CiteSeer query results. Future works will also use domain ontologies for identifying concepts/explanations by following the approach described in [6]. Finally, we will also address the possible advantages of utilizing our ideas in other scenarios such as news recommendations.

References

1. Sciencerec track at recsys challenge (2012), `http://2012.recsyschallenge.com/tracks/sciencerec/`
2. Chandrasekaran, K., Gauch, S., Lakkaraju, P., Luong, H.P.: Concept-based document recommendations for citeSeer authors. In: Nejdl, W., Kay, J., Pu, P., Herder, E. (eds.) AH 2008. LNCS, vol. 5149, pp. 83–92. Springer, Heidelberg (2008)
3. Ferrara, F., Tasso, C.: Extracting and exploiting topics of interests from social tagging systems. In: Bouchachia, A. (ed.) ICAIS 2011. LNCS, vol. 6943, pp. 285–296. Springer, Heidelberg (2011)
4. Govindaraju, V., Ramanathan, K.: Similar document search and recommendation. Journal of Emerging Technologies in Web Intelligence 4(1), 84–93 (2012)
5. Papadimitriou, A., Symeonidis, P., Manolopoulos, Y.: A generalized taxonomy of explanations styles for traditional and social recommender systems. Data Mining and Knowledge Discovery 24(3), 555–583 (2012)
6. Pudota, N., Dattolo, A., Baruzzo, A., Ferrara, F., Tasso, C.: Automatic keyphrase extraction and ontology mining for content-based tag recommendation. International Journal of Intelligent Systems, Special Issue: New Trends for Ontology-Based Knowledge Discovery 25, 1158–1186 (2010)
7. 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)