ExpertFOAF recommends experts

Tereza Iofciu
iofciu@l3s.de

Jörg Diederich
diederich@l3s.de

Peter Dolog
dolog@cs.aau.dk

Wolf-Tilo Balke
balke@l3s.de

1 Introduction

FOAF files are often used for publishing simple information about persons and about their community. Our proposal is to extend the user's FOAF files with automatically generated user profiles, which are based on a histogram of the user's interest and which are also semantically enriched using the GrowBag approach [1]. The main assumption is that such user profiles can provide good hints about users' expertise. Such extended FOAF files (called *ExpertFOAF*) can be published on a user's home page, on web pages of institutions or conferences to characterize them. They can be crawled by distributed recommender systems for finding users with similar interests and, hence, expertise in different domains.

We consider a well-defined user profile to express best the user's current interests in a domain, where the domain (such as the research environment with students and professors) is defined by a collection of tagged objects (e.g., publications and deliverables tagged with keywords).

2 Histogram-based User Profiles

In our approach, user profiles are histogram vectors consisting of interest tags and their weights [2]. These tags can, for example, be the tags of a user in a collaborative tagging environment or the author keywords from a publication server, characterizing the publications of a user.

As presented in [2], there are two steps for obtaining the tag-based user profile. We first keep track of an intermediate profile the user creates when selecting objects of interest (either selected explicitly as 'characterizing my interests' or, better, implicitly from the tagging the user has made previously). This profile is then translated into the final user profile format by mapping the objects to their tags. In this manner we also keep track of the cardinality of the tags, i.e. how often they occur in the different objects of interest. We differentiate between two types of profiles based on the nature of the objects selected by the user as relevant, the implicit profile and the explicit one. The most relevant selected objects contribute to the implicit profile and the relevant objects that a user has selected explicitly form the explicit profile (e.g. for publications, it means the articles authored by the user).

The main advantage of this approach is that users have to specify comparatively few objects to generate a reasonably large profile, (assuming that an object has on average more than one tag). Since tags are usually shared (e.g. folksonomies, taxonomies), it is more probable to find commonalities between user profiles with our approach, thus we can provide better recommendations. Furthermore, having a profile on the metalayer rules out the concept of "bad" items in the profile, which may happen at object level (i.e. there may be objects the user showed an implicit interest in, for example, by downloading it to the user's desktop, but which the user finally found bad. The tags of such 'irrelevant' publications may, nevertheless, be somehow relevant to the user).

Such user profiles can, hence, be generated automatically, or at least semi-automatically in domains, where tagged corpora of objects are available. We assume that when starting from a corpora with relevant tags, our approach guarantees the quality of the interest based profiles as the nature of the taxonomy has a great impact on the quality of the user profiles. It is easy to assume that in the publication-keyword domain the quality of the profiles is high due to the accuracy of the keywords assigned to published articles. For the blog-tag domain is more difficult as there is a lot of noise and nor the accuracy of the tags, nor of the blogs themselves can be assured. The noise level can be big even in domain blogs. We will have to use also link analysis to estimate the quality of the tags and blogs, and thus of the profiles.

We proposed to include such user profiles in a user's FOAF file, forming the so-called *ExpertFOAF* files. ExpertFOAF files comprise both, the objects and also the tags defining the users' interests to represent the user on different levels of details. We further want to extend the FOAF format with an attribute for the weights of the tags, to make it possible for the histogram vector to be included in the profile. This way the user can find also other applications for her ExpertFOAF. By having the users publish their profiles on the site of a conference they participate at we can obtain a histogram of topics for the participating community. This would give a good insight on the background knowledge of the conference participants.

In addition to the idea of tag-based user profiles, we propose to enhance a profile with more tags when relations between the domain tags do exist or can be extracted. For this purpose, we want to use the GrowBag approach [1] to further reduce the sparsity of the user profiles. The semantic GrowBag bases on the collection of tagged objects to determine intrinsic relations between the domain tags (i.e. finding super-tags and sub-tags for a given tag) and automatically creating a tag graph with weak and strong relations. The user profiles can be enhanced by adding to the histogram vector, for each tag, its super-tags and/or its tags which are above the respective tag in the graph. We want to add this tags with different subunitary weights, proportional to the importance of the relation and decreasing with the distance in the graph. The rationale behind this is that super-topic relations far down the tag hierarchy are better suited to enrich the user profile than those relations high in the hierarchy. As an example, inferring that a user with interests in 'RDF' has also interests in 'Semantic Web' seems very reasonable while someone having specified interests in 'XML' might not be in the same way interested in any 'markup language'.

This approach is comparable with further adding synsets to profiles, using Wordnet as presented in [3]. The advantage in our approach is that with the GrowBag approach we obtain automatically the 'shallow' semantic tag relations.

3 Research challenges

There are several practical obstacles in refining the data for our approach and for evaluating its results. We need to prove that interest based user profiles can represent expertise.

We apply our approach to the domain of digital libraries, using a subset of the DBLP data set as object corpus, which has been enhanced with 'tags', e.g., the keywords that were manually specified by the authors of the publications. We intend to evaluate our approach to finding experts by assuming that there should be an overlap between the list of coauthors of an user and the list of experts recommended based on her profile. Finding relevant experts on topics is not only good for personal interests, but it is a tool which can be used by conference organizers when selecting appropriate reviewers.

We also want to test our approach for webblogs, where the objects are the blogs and the tags are their topics. Here we can create profiles for users and create ExpertFOAF and also create profiles for blogs and export them to SIOC(creating ExpertSIOC) format. We can apply recommender algorithms to find experts at people level or at blog level.

The problem that arises when dealing with collaborative tagging is that the vocabulary is not controlled, even in domain centered collections, we have to establish different means for cleaning the tag space (e.g., to filter tags like 'good stuff', which are not topic-oriented) and also for automatically identifying tags which represent the same notion(e.g. 'www' is the same as 'world wide web'). Another problem which arises when using folksonomies or taxonomies is that usually the tags' meanings and impact change over time. When applying the GrowBag approach we also have to keep track of the moments of issue of the objects in the user profile. The question is whether we should just use the current time as we aim at representing the current user's interests. For example the topic 'search engine' has evolved from being a sub-topic of 'internet' and 'web' in 1998-1999 to becoming a super-topic in 2003-2004.

References

[1] W.-T. Balke, U. Thaden, and J. Diederich. The Semantic GrowBag Demonstrator for Automatically Organizing Topic Facets. In *Proceedings of SIGIR2006 Workshop on Faceted Search*, Seattle, USA, August 2006.

[2] J. Diederich and T. Iofciu. Finding Communities of Practice from User Profiles Based On Folksonomies. In *Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice (TEL-CoPs'06), co-located with the First European Conference on Technology-Enhanced Learning*, Crete, Greece, 2006.

[3] B. Magnini and C. Strapparava. User Modelling for News Web Sites with Word Sense Based Techniques. In *User Modeling and User-Adapted Interaction*, volume 14, pages 239–257, 2004.