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

Today, FOAF files are only utilized for publishing simple information about persons and their respective community. Our proposal is to extend a user's FOAF file by an automatically generated user profile. If a general classification of interests (for instance in the form of a taxonomy) is given, these profiles can be represented in an IR-style fashion by histograms expressing the degree of interest in each topic. Our main assumption is that such user profiles can then provide good hints about the users' expertise.

We consider a well-defined user profile to express best the user's current interests in each domain, where a domain is basically defined by a collection of tagged objects (e.g., publications and deliverables tagged with keywords). The resulting tag clouds can also be semantically enriched, using for instance the GrowBag approach [2] to define a hierarchical structure on tags. Eventually, 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.

2 Histogram-based User Profiles

In our approach, user profiles are represented as histogram vectors consisting of predefined interest tags and their weights [1]. These tags can, for example, be the set of annotations provided by a user in a collaborative tagging environment or the author keywords from a publication server, characterizing the publications of a user.

As presented in [1], there are basically two steps for obtaining a tag-based user profile. We first record the intermediate profile a user creates when selecting objects of interest (either selected explicitly as 'characterizing my interests' or, better, implicitly from the taggings a user has made previously). This profile is then translated into a final user profile format by mapping the objects to their tags and keeping track of the cardinality of the tags.

Moreover, we always distinguish two types of resulting profiles based on the nature of the objects selected by the user as being of interest: the implicit profile and the explicit one. Only those objects that a user has selected explicitly as being relevant form the explicit profile (for instance with respect to publications, only those articles actually authored by the user).

The main advantage of this approach is that users have to specify comparatively few objects to generate a reasonably complete profile, (assuming that an object has on average more than just a single tag). Since tags are usually shared (e.g. folksonomies, taxonomies), it is very probable to find commonalities between user profiles with our approach. Thus, we can provide high quality recommendations. Furthermore, building the final profile on the metalayer (i.e. on tags) decreases the impact of ”bad” items in the profile, which can easily happen at object level (consider for instance objects in which a user showed interest by downloading them to the desktop, but which the user subsequently recognized as being irrelevant. The tags of such irrelevant publications may, nevertheless, be somehow relevant to the user).

Such user profiles can be generated automatically, or at least semi-automatically in domains, where tagged corpora of objects are readily available. We assume that when starting from a corpus with high quality tags, our approach also guarantees the quality of the interest-based profiles. It is sensible to assume that in the publication-keyword domain the quality of the tagging is high due to the accuracy of the keywords assigned to published articles by the authors. In the blog-tag domain it is more difficult as there is a lot of noise and neither the accuracy of the tags, nor the correctness of the blogs' contents can be assured, not even in domain blogs. We plan to use link analysis in this case to estimate the quality of the tags and blogs, and thus of the profiles.

Our final ExpertFOAF files still comprise both, the objects and also the tags defining the users' interests to represent the user on different levels of detail. We plan to extend the FOAF format with an attribute for the relative weights of the tags, to enable the inclusion of the histogram vector into the profile. In this way users can find also other applications for their ExpertFOAFs: For example, by having users publish their profiles on the site of a conference they participate in, one could obtain a cumulated histogram of topics important for the participating community. This would cast a light on the average background knowledge of the conference participants.

In addition to the idea of tag-based user profiles, we propose to enhance profiles when relations between individual domain tags do exist or can be extracted. For this purpose, we plan to use the GrowBag approach [2] to further reduce the sparsity of the user profiles. This approach is comparable to adding synsets to profiles, e.g. using Wordnet as presented in [3]. The advantage in our approach is that with the GrowBag approach we automatically obtain the 'shallow' semantic tag relations. The semantic GrowBag algorithm 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 creates a tag graph with weak and strong relations. The user profiles can then be enhanced by adding hierarchical information to the histogram vector, for each tag, its super-tags and/or tags which are above the respective tag in the graph can be added. 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. For 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 other ’markup languages’.

3 Research challenges

There are several practical obstacles in refining the data for our approach and for evaluating its results. We eventually need to prove that interest-based user profiles can really represent expertise. One problem that arises when dealing with collaborative tagging is that the vocabulary is not controlled. Hence, even in domain-specific 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 for 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 top level topic in 2003-2004.

In a first use case we already applied our approach to the domain of digital libraries, using a subset of the DBLP data set as object corpus. This corpus is enhanced with tags, i.e. the keywords that were manually specified by the authors of the publications. We intend to evaluate our approach of finding experts by assuming that there should be an overlap between the list of co-authors of an user and the list of experts recommended based on his/her profile. Finding relevant experts on topics is not only good for personal interests, but it is for example a tool that can be used by conference organizers when selecting appropriate reviewers.

We also want to test our approach for Web-blogs, where the objects are the blogs and the tags are their respective topics. Here we can create profiles for users, create ExpertFOAFs 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.

References

[1] 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.

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