Identifying User Interests in Folksonomies

Elias Zavitsanos[1], George A. Vouros[1], and Georgios Paliouras[2]

[1]Dpt. of Information and Communication Systems Engineering
University of the Aegean, Greece
izavits@iit.demokritos.gr, georgev@aegean.gr
[2]Inst. of Informatics and Telecommunications, NCSR Demokritos, Greece
paliourg@iit.demokritos.gr

Abstract. This paper proposes a probabilistic method for classifying folksonomy users to specific domains and for identifying their specific interests to these domains. The proposed method uses a hierarchical probabilistic topic modeling approach that exploits tags to induce hierarchies of latent topics. These hierarchies represent domain conceptualizations of specific domains that are either collective or user-specific. We propose two alternative methods that exploit the induced hierarchies for classifying and identifying users' interests to specific domains and provide preliminary evaluation results.

1 Introduction

Folksonomy tags depend totally on the interests, preferences, conceptualization, nomenclature, whim and personal style of users. Therefore, there is a great potential for acquiring knowledge about folksonomy users by exploiting the tags they introduce and the relations between folksonomy components ([4], [8], [7], [5], [11]). Works regarding the identification of folksonomy users' interests build user profiles by either exploiting external resources and existing ontologies [2],[9], by exploiting the tags and the tagged objects [3], by clustering users exploiting their tagging activity [6], or by exploiting users' tags and other users expertise profiles to infer user's expertise [1]. Our approach is to induce collective and user-specific domain topic hierarchies (domain conceptualizations), and exploit these hierarchies to classify users and to identify their specific interests to these domains. Specifically, this work contributes towards (a) The automated induction of topic hierarchies from tags in a statistical and parametric-less way, without requiring external resources or prior knowledge, and (b) The classification and identification of users' particular interests to specific domains, by exploiting the induced domain topic hierarchies. To a great extent than existing approaches, the proposed overall method is generic, unsupervised, parametric-less, language-agnostic, and does not require training data.

Specifically, we propose the use of a hierarchical probabilistic topic modeling approach for inducing domain conceptualizations from tags. Probabilistic topic models identify latent features (topics) that maintain a probability distribution over the tag space. Topics represent essential components regarding the content of tag chains (sets of tags related to a specific resource), thus, they reflect users'

conceptualization of the tagged resources. Hierarchical probabilistic topic models order the induced topics in hierarchies that may constitute the backbones of domain ontologies. The proposed method induces collective (and user-specific) topic hierarchies (conceptualizations), as these are reflected by tags introduced by sets of users (resp. individual users) related to a domain. We then propose two alternatives for user classification which compare the collective domain conceptualizations either with the users' tag chains directly, or with the induced user-specific topic hierarchies.

We provide experimental results using datasets gathered from the Del.icio.us collaborative tagging system: Crawling Del.icio.us for a two-month period we have gathered the tag chains of resources related to specific "domains" delineated by the tags **design**, **software**, **programming** and **web**. The crawler takes as input a single tag characterizing a domain (e.g. **programming**), and a number specifying the depth of the crawling process. For instance, for depth equal to 0, only the tag chains of the first page for the input tag are gathered. For depth equal to 1, the tag chains of the first page are gathered, and next, for each tag of each tag chain, the tag chains of the first page of that tag are also gathered. Each tag chain is treated as a separate ("virtual") document. The above crawling process is performed without considering the individual users tagging the resources. We have been running the crawler for crawling depths from 0 to 3. Following a similar process, we have also gathered the tag chains for each of 300 users per domain. These corpora of documents (tag chains) provide all the necessary information in order to induce collective and users'-specific conceptualizations and further classify the users to the domains.

2 The Proposed Method

The induction of (collective or user-specific) topic hierarchies is essential to any of the proposed classification methods.

The **hierarchy learning algorithm** is originally proposed as a generic method for inducing topic hierarchies given a corpus of documents [12]. A document (in our case, a tag chain) is assumed to have been generated by some latent topics. These topics have been drawn by a Dirichlet Process base measure, which in turn has been drawn from a Global Dirichlet Process that applies to the whole corpus of documents, assuring the sharing of topics among documents. The topics maintain a multinomial probability distribution over the tags of the corpus. We are interested in the process that computes the topics and their hierarchical relations. According to the proposed method, each level of the topic hierarchy is associated with a HDP [10] . The dataset provides the observations for the inference of the latent hierarchy. The process starts by inferring the lowest level of the hierarchy: Tags are assigned to leaf topics. Having inferred the leaf topics, their mixture proportions for the documents is known. In other words we can infer which topics have contributed, and to what degree, to the "generation" of each tag chain. Furthermore, the assignment of a tag to a specific topic constitutes the observation for the inference of the next level up. At the next levels up, following the same procedure, each inferred topic maintains a distribution over

the tags of the tag chains and over the topics at the level below. The procedure is repeated until it converges to a single topic, which serves as the root of the hierarchy. More details may be found in [12].

User Classification based on Maximum Likelihood : This method uses the hierarchy learning algorithm to compute the collective topic hierarchy for each of a set of domains. Then, users are classified by computing the log-likelihood of each domain model, given the user-specific tag chains. The user is classified to the domain whose model has the maximum likelihood, since it is assessed that this model is most likely to have "generated" his/her tag-chains. It must be pointed out that, as a consequence of this computation, the log likelihood of the specific topics that may have generated users' tag chains are also computed: Doing so, the interest of users to specific domain topics in the collective hierarchy is revealed.

User Classification based on Hierarchy Comparison : The second alternative for user classification, in conjunction to the collective conceptualizations, induces a topic hierarchy for each user, using the hierarchy learning algorithm presented above with input the user-specific dataset. Then, the classification process continues as follows: having the collective model of each domain and the user-specific domain model, the topic hierarchies are aligned and the corresponding user is classified to the domain whose model is "closest" to the user's model. In order to align the two hierarchical topic models, we use the DMA distributional ontology alignment method proposed in [13]. The main idea is to align the two ontologies, and based on the correspondences to derive scores that measure their "closeness". In our problem case, given that all topics in both (collective and user-specific) hierarchical models are represented as multinomial probability distributions over the tags of the dataset, for the computation of the similarity between different topics we use the Total Variational Distance (TVD), ranging in $[0, 1]$: $TVD = \frac{1}{2} \sum_i \mid P(i) - Q(i) \mid$ ($P(\cdot)$ and $Q(\cdot)$ are the multinomial distributions over tags in the compared topics). Finally, Matching Precision MP, Matching Recall MR and the Matching F-measure MF provide an assessment of user's topic hierarchy "closeness" to the collective topic hierarchy. The formulae for these measures are given and explained in [13]. The extensive experimental tests in [13] show that this method succeeds to reflect the deviation between the two hierarchies, taking also into account the differences between the hierarchies' structures and the deviations of the induced topics.

3 Empirical Evaluation

The empirical evaluation of the proposed classification methods concerns the classification of the different users into the four main domains: **design**, **programming**, **software** and **web**. Given the tags of users and the computed collective conceptualizations for each of these domains, we have asked three evaluators to classify the users into the four domains, in order to use this classification as ground truth. The evaluators have agreed for the classification of 285 users - out of 300 - per domain. For evaluation purposes, the datasets for all users were put in a single directory. Each user was classified to only one domain

(multi-label classification is left for future work). Table 1 provides experimental results for both classification alternatives using the datasets for crawling depth equal to 1.

Table 1. Evaluation results for the two classification approaches.

	LogLikelihood Approach			DMA		
Domain	Precision	Recall	F-measure	Precision	Recall	F-measure
Design	0.98	0.82	0.89	0.91	0.80	0.85
Programming	0.99	0.67	0.80	0.96	0.68	0.80
Software	0.75	0.97	0.85	0.75	0.96	0.84
Web	0.83	1.0	0.90	0.82	0.92	0.87

It must be pointed out that, regarding the second classification alternative, we may change the calculation of the MF measure, so as to increase the importance of having a large number of topic correspondences between hierarchies rather than a few precise correspondences. This increases the F-measure for **design** to 1.00 for **programming** to 0.82, for **design** to 0.86 and for **web** to 0.99.

References

1. A. Budura, D. Bourges-Waldegg, and J. Riordan. Deriving expertise profiles from tags. In *CSE (4)'09*, 2009.
2. F. Carmagnola, F. Cena, L. Console, O. Cortassa, C. Gena, A. Goy, I. Torre, A. Toso, and F. Vernero. Tag-based user modeling for social multi-device adaptive guides. *User Modeling And User-Adapted Interaction*, 18(5):497–538, 2008.
3. J. Diederich and T. Iofciu. Finding communities of practice from user profiles based on folksonomies. In *EC-TEL 2006 Workshop Proceedings*, 2006.
4. S. Golder and B. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.
5. H. Halpin, V. Robu, and H. Shepard. The dynamics and semantics of collaborative tagging. In *SAAW'06*, 2006.
6. X. Li, L. Guo, and Y. Zhao. Tag-based social interest discovery. In *WWW*, 2008.
7. A. Mathes. Folksonomies - cooperative classification and communication through shared metadata, December 2004.
8. P. Mika. Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics*, 5(1):5–15, 2007.
9. M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt. Semantic modelling of user interests based on cross-folksonomy analysis. In *ISWC*, 2008.
10. Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006.
11. H. Wu, M. Zubair, and K. Maly. Harvesting social knowledge from folksonomies. In *Proc. of the 17th Conference on Hypertext and Hypermedia*, 2006.
12. E. Zavitsanos. *Learning Ontologies from Text Collections and Evaluating them Against Gold Standards*. PhD Thesis, University of the Aegean, 2009.
13. E. Zavitsanos, G. Paliouras, and G. A. Vouros. Gold standard evaluation of ontology learning methods through ontology transformation and alignment. *TKDE*, `http://doi.ieeecomputersociety.org/10.1109/TKDE.2010.195` (to appear).