Can Your Friends Help You to Find Interesting Multimedia Content on Web 2.0?

Alejandro Corbellini[1,2], Daniela Godoy[1,2], and Silvia Schiaffino[1,2]

[1] ISISTAN Research Institute, UNICEN University,
Tandil, Bs. As., Argentina
[2] CONICET, National Council for Scientific and Technical Research, Argentina
{alejandro.corbellini,silvia.schiaffino,
daniela.godoy}@isistan.unicen.edu.ar

Abstract. Social tagging constitutes one of the defining characteristics of Web 2.0 as it allows users to collectively classify and find diverse resources, such as Web pages, songs or pictures, using open-ended tags. The data structures underlying these systems, also known as folksonomies, suffered an explosive growth on account of the widespread success of social tagging. Thus, it is becoming increasingly difficult for users to find interesting resources as well as filter information streams coming from this massive amount of user-generated content on Web 2.0. In addition, most resources lacks easily extractable content to apply traditional content-based profiling approaches. In this paper we present an approach to build tag-based profiles for multimedia resources (such as songs, pictures or videos) using the social tags associated to resources as a means to describe them and, in turn, user interests. Experimental results show that the tags assigned by members of the community can help to predict the interestigness of a given resource for a user in an effective way.

Keywords: Social Tagging Systems, Folksonomies, Web 2.0.

1 Introduction

Social tagging systems, also known as folksonomies, emerged in the last years as a novel social classification on the Web that contrasts with traditional pre-defined taxonomies or directories usually seen on the Web. This scheme relies on the convergence of tagging efforts of a large community of users to a common categorization system that can be effectively used to organize and navigate large information spaces. In fact, the term *folksonomy* is a blend of the words *taxonomy* and *folk*, and stands for conceptual structures created by the people [6].

Multimedia resources such a songs, videos or pictures are collectively created, annotated and categorized in sites such as *Last.fm*[1], *Flickr*[2] or *YouTube*[3], among others. In these sites, users annotate resources using a freely chosen set of keywords or open-ended tags. In fact, it is argued that the power of tagging lies in the ability for

[1] http://www.last.fm/
[2] http://www.flickr.com/
[3] http://www.youtube.com/

F.V. Cipolla-Ficarra et al. (Eds.): ADNTIIC 2011, LNCS 7547, pp. 253–261, 2012.

people to freely determine the appropriate tags for resources without having to rely on a predefined lexicon or hierarchy.

The downside of tagging is the constantly expanding size of communities using social sites and the completely unsupervised nature of tags that lead to a huge volume of resources to be explored and analysed. In consequence, the discovery of relevant resources becomes a time consuming and difficult task for users. Tag-based user profiling techniques have emerged to help users in selecting appropriate tags to resources, finding relevant information and locating like-minded users [11, 2].

In traditional information retrieval and filtering systems, long-term interests are expressed in user profiles, which are usually learned starting from the textual content of documents exemplifying the user interests. In folksonomies, however, most of the resources have a non-textual content, including music, pictures and video. In the absence of textual content, meta-data can be used to build profiles. Particularly, in tagging systems the most important meta-data associated to resources are the tags or annotations assigned by users.

In this paper, a tag-based profiling approach that exploits social tags as a source for modelling user interests is presented. This approach assumes that users are likely to be interested in additional content annotated with similar social tags to the ones assigned to resources they showed interest in before. Thus, a learning algorithm is used to train a classifier (or user profile) that recognizes potentially interesting resources. This profile can be also applied to filter incoming information from tagging systems (e.g. RSS feeds).

The rest of this paper is organized as follows. Section 2 introduces the proposed approach for user profiling based on social tagging activity. Section 3 describes the empirical study carried out to validate the approach with a dataset from *Lastfm* site. Section 4 reviews related research in the area. Finally, concluding remarks are stated in Section 5.

2 Our Approach

A folksonomy can be defined as a tuple $\mathsf{F} := (U, T, R, Y, \pi)$ which describes the users *U*, resources *R*, and tags *T*, and the user-based assignment of tags to resources by a ternary relation between them, i.e., $Y \subseteq U \times T \times R$ [6]. In this folksonomy, π is a user-specific sub-tag/super-tag-relation possible existing between tags, i.e., $\pi \subseteq U \times T \times T$.

The collection of all tag assignments of a single user constitute a *personomy*, i.e., the personomy P_u of a given user $u \in U$ is the restriction of F to *u*, i.e., $\mathsf{P}_u := (T_u, R_u, I_u, \pi_u)$ with $I_u := \{(t, r) \in T \times R \mid (u, t, r) \in Y\}$, $T_u := \pi_1(I_u)$, $R_u := \pi_2(I_u)$, and $\pi_u := \{(t_1, t_2) \in T \times T \mid (u, t_1, t_2) \in \pi\}$, where π_i is the projection on the *i*th dimension. In social tagging systems, tags are used to organize shared information within a personal information space.

The interests of a user are extracted from the user own personomy. This is, the resources the user annotated in the past are assumed to reflex the user preferences. To describe these resources, content in the form of text is sometime not available, but meta-data is. In social tagging systems the richer meta-data information about resources is the tags assigned to resources by the user community. Using social tags for user profiling allows users to capitalize on the associations made by persons who have assigned similar tags to other resources.

Section 2.1 describes the pre-processing techniques applied to tags in order to reduce syntactic variations. Section 2.2 introduces the learning technique employed to learn a user profile starting from the resources in the user personomy and the tags associated to them.

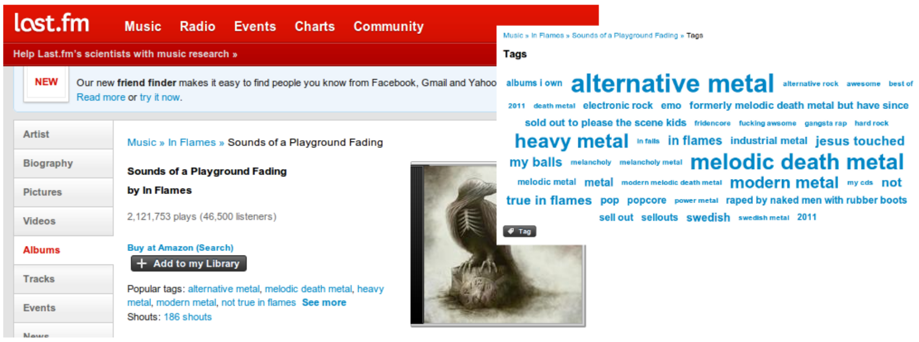
2.1 Tags Pre-processing

Each resource in the user personomy is considered in the proposed approach as an example of the user interests and it is described with the tags other users assigned to the resource. This is known as the full tagging activity (FTA) associated to resources, i.e., all tags assigned by members of the community to the.

Even though the success of tagging systems was greatly due to the possibility of freely determine a set of tags for a resource without the constraint of a controlled vocabulary, lexicon, or pre-defined hierarchy; the free-form nature of tagging also leads to a number of vocabulary problems. In this paper we deal with two common variations [17]:

- inconsistently grouping of compound words consisting of more than two words. Often users insert punctuation to separate the words, for example *ancient-egypt*, *ancient_egypt*, and *ancientgypt*;
- use of symbols in tags, symbols such as #, -, +, /, :, _, &, ! are frequently used at the beginning of tags to cause some incidental effect such as forcing the interface to list some tag at the top of an alphabetical listing.

To prevent syntactic mismatches due to these reasons original raw tags were filtered to remove symbols such as #, -, +, /, :, _, &, ! , which at the same time allows joining compound words. After these pre-processing step tags are weighted according to the number of users that assign the tag to the resource so that the more frequently a tag is used to annotate a resource the more important it is to describe the resource content. Figure 1 shows an example of an album in *Last.fm* and the tag cloud associated to this resource, that summarized its full tagging activity.



Our new **friend finder** makes it easy to find people you know from Facebook, Gmail and Yahoo
Read more or try it now.

Music » In Flames » Sounds of a Playground Fading

Sounds of a Playground Fading
by In Flames

2,121,753 plays (46,500 listeners)

Popular tags: alternative metal, melodic death metal, heavy metal, modern metal, not true in flames **See more**
Shouts: 186 shouts

Music » In Flames » Sounds of a Playground Fading » Tags

Tags

albums i own alternative metal alternative rock awesome best of 2011 death metal electronic rock emo formerly melodic death metal but have since sold out to please the scene kids fridencore fucking awsome gangsta rap hard rock heavy metal in falls in flames industrial metal jesus touched my balls melancholy melancholy metal melodic death metal melodic metal metal modern melodic death metal modern metal my cds not true in flames pop popcore power metal raped by naked men with rubber boots sell out sellouts swedish swedish metal 2011

Fig. 1. Example of representing an album in Last.fm using the tag cloud

2.2 One-Class Classification

In order to profile a user annotating multimedia resources, such as songs or pictures, the tags or annotations in the user personomy are taken as examples. The annotated resources in a personomy constitute positive examples of the user interests that can be easily collected from folksonomies. On the contrary, to identify representative negative examples or non-interesting resources is more complex since users might not tag a potentially interesting resource because of multiple reasons.

Since only positive examples of the user interests are available, the task of determining whether a resource is interesting for a user basing learning exclusively on these examples can be seen as a one-class classification problem. One-class classification differs in one essential aspect from conventional classification as it assumes that only information of one of the classes, the target class, is available. The boundary between the two classes has to be estimated from data of only the normal, genuine class. Then, the task is to define a boundary around the target class, such that it accepts as many of the target objects as possible, while it minimizes the chance of accepting outlier objects.

One-class classification based on SVMs (Support Vector Machines) are used in this paper because they showed superior performance than other classifiers in a comparative study [5]. SVMs are a useful technique for data classification, which has been shown to be perhaps the most accurate algorithm for text classification, it is also widely used in Web page classification. Schölkopf et al. [14] extended the SVM methodology to handle training using only positive information and Manevitz and Yousef [8] applied this method to document classification and compare it with other one-class methods.

For training one-class SVM classifiers, the origin is considered the only member of the negative class as well as a certain number of data points of the positive class. SVM approach proceeds by determining the hyperplane that separates most of the negative data from the origin of the hypersphere containing the examples of the target class, separating a certain percentage of outliers from the rest of the data points. Then the standard two-class SVM techniques are employed. Figure 2 depicts this procedure. In this work we used LibSVM[4] [1] implementation of one-class SVM.

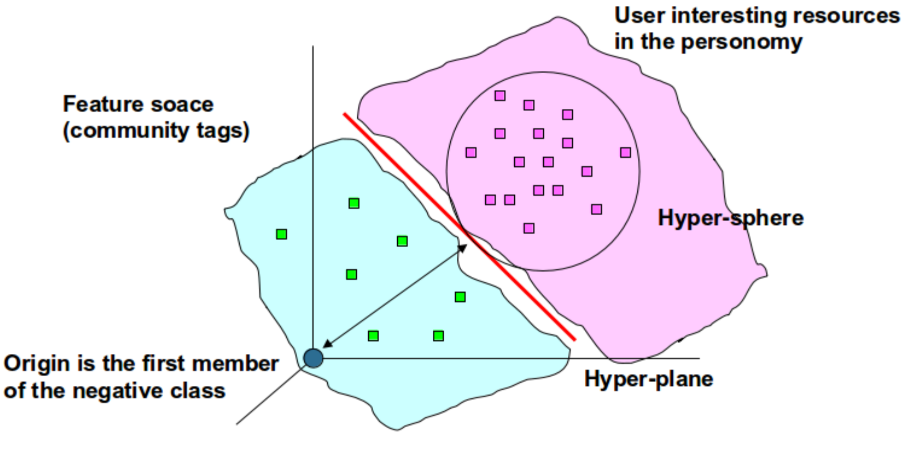

User interesting resources in the personomy
Feature soace (community tags)
Hyper-sphere
Origin is the first member of the negative class
Hyper-plane

Fig. 2. One-class SVM

[4] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

3 Experimental Results

In this paper, we use the *Last.fm* dataset described in [13] that collects neighbour and friend relationships in addition to tagging activity. In this site, friends are contacts in the social networks, while neighbours are users recommended by the system as potential contacts based on their music playing histories. The dataset also contains annotations in the form of triples (user,item,tags) and group membership information that was crawled from *Last.fm* site. The entire dataset contains 99.405 users, 52.452 of these users are considered active, i.e. have at least one annotation. The 10.936.545 triples annotate 1.393.559 items with 281.818 tags, belonging to 66.429 groups.

To represent a resource into the profile of a user the full tagging activity of the resource was used, including the tags assigned by all the community. In addition, smaller groups conformed by users having some relationship with the target user were considered in this study. Other relationships available in the dataset that were used for experimentation are:

– *friends*: extracted from a fraction of *Last.fm* social network, friendship relationships are symmetric;
– *group members*: users that belonging to the same groups as the user in consideration;
– *neighbours*: neighbourhood of users with similar tastes, neighbourhood relationship is not symmetric.

For experimentation 100 users were randomly selected from those having a minimum of 10 friends, group memberships and neighbours in the dataset and also at least 100 annotated resources. Evaluation was carried out using a holdout strategy that split data into a 66% for training and a 34% for testing. This is, each personomy was divided into a training set used to learn the one-class classifier and a testing set used to assess its validity. To make the results less dependent of the data splitting, the average and standard deviation of 5 runs are reported for each user, with error-bars indicating standard deviations.

Figure 3 shows the accuracy of classifiers for detecting interesting examples in the testing sets. This is, how many times the classifier deemed a resource as relevant and it effectively was. It can be observed in the figure that classifiers built using the full tagging activity of the community reach a good performance, only improved using tagging activity of users belonging to the same groups than the target users. Friends instead are not good predictors of the user interests. Likewise, classification of resources using neighbour tags was the worst performing scheme.

The performance reached for the different classifiers can be explained by the reduction in the dimensionality space during learning when information provided for less users than the entire community is considered. Table 1 summarizes the number of unique tags involved in learning the classifiers in each case. Also, the minimum, maximum and average number of tags used for the classifiers are reported in the table. Table 2 shows the number of users in the community, friends per personomy, number of groups the user belong to and neighbours suggested by *Last.fm*.

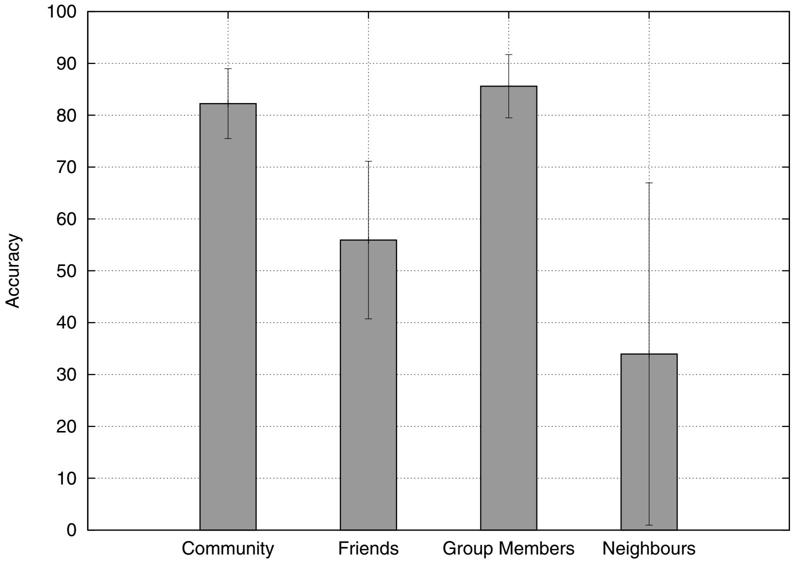


Fig. 3. Classifiers accuracy in identifying interesting resources

It is worth noticing that when classifiers are learned using the tags assigned by other users in the same groups as the target user, accuracy improves in spite of the slight reduction in the number of tags. This implies that the complexity of the learning problem is diminished, consequently reducing the times of training and classification.

Table 1. Summary of tag dimensions according to the users considered

Node	Community	Friends	Group members	Neighbours
# unique tags	45778	4027	40328	172
# minimum	1090.0	67	808.0	2
# maximum	21808.0	2178.0	7908.0	84
Mean ± SD	9155.6±7058.08	447.44±635.00	4480.8±2455.93	19.11±24.00

Table 2. Summary of number of user relationships

	# minimum	# maximum	Mean ± SD
# community	-	-	99405
# friends	113	873	381.33±244.89
# groups	52	300	106.77±72.58
# neighbours	59	60	59.55±0.49
# resources	108	3546	692.22±1031.82

4 Related Works

Tag-based profiling approaches enable personalize resource recommendation in social tagging systems [2, 11]. User profiles consisting of weighted tags vectors are obtained using the tags frequency of occurrence in the user resources as well as the inverse user frequency as a measure of their relative importance [12, 3, 16].

Firan et al. [4] compared tag-based user profiles with more conventional profiles based on song and track usage in the music search portal *Last.fm*. The results showed that tag-based profiles significantly improve the quality of recommendations. Au Yeoung at al. [18] investigated an algorithm which performs graph-based clustering over the network of user tagged documents to identify interest topics and extract tag vectors. In [15], tag clustering is used to group tags with similar meanings as the basis for a personalization algorithm for recommendation in folksonomies. Both users and resources are modelled as weighted tag vectors and tag clusters are intermediary between them. Vectors describing resources are used to detect relevant resources for a given cluster of tags, whereas similarity among a user vector and a tag clusters allows to recommend resources.

Graph representations were also proposed to model the relationships among tags in a user profile. Michlmayr et al. [9] compared tag-based profiles consisting of a single vector of weighted tags with graph representations in which nodes correspond to tags and edges denote co-occurrence or other relationships among them. *Add-A-Tag* [10] algorithm, extends this model to include temporal information by updating the weights of edges in the graph using an evaporation technique known from ant algorithms for discrete optimization. The idea of using semantic relationships among tags in tag-based profiles has also been explored in [7], in which the semantic distance between two tags is calculated based on co-occurrence statistics and common sense reasoning.

User profiles in these works model the user preferences in terms of the tags a user employed to annotate its resources in the past. Instead, in this research a user profile models the type of resources a user is interesting in based on the social tags attached to them, i.e. using a collective description of the resource. Thus, the proposed approach does not rely on the degree of coincidence between user tags and tags assigned by other members of the community to the resources.

5 Conclusions and Future Work

In this paper a tag-based profiling approach that exploits social tags for identifying relevant resources from folksonomies according to the interests of individual users was presented. One-class classifiers were used to learn the user interests from resources in the user personomy and the tags collectively assigned to them. Thus, collective knowledge extracted from folksonomies contributes to automatic, personal Web document classification.

Experimental results obtained with a dataset from *Last.fm* site showed that tag-based classifiers accurately recognize interesting resources. In these experiments, the use of other relationships among users were also explored in order to reduce the tag space, thereby diminishing the complexity and times involved in learning classifiers. The results demonstrate that friends and neighbours do not provide enough information to efficiently classify novel resources. In contrast, information about user membership to one or more groups allows to limit the number of users classifiers are learned from and, at the same time, to improve the accuracy of classifiers.

In future works we are planning to experiment with other types of social networks existing on the Web in which relationships among users are of different nature. For example, followers/followee relation in micro-blogging networks as well as friends and groups in Facebook, among others. We will also experiment with more semantic representation of tagging activities. Instead of simply sintactic modification to tags as the one used in this paper, semantic ones will be aoolied with the help of dictionaries and other lexical resources.

Acknowledgments. This work has been partially funded by CONICET through project PIP No. 114-200901-00381.

References

1. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines(2001), Software, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
2. Dattolo, A., Ferrara, F., Tasso, C.: The role of tags for recommendation: a survey. In: 3rd International Conference on Human System Interaction (HSI 2010), pp. 548–555. IEEE Press, Rzeszow (2010)
3. Diederich, J., Iofciu, T.: Finding communities of practice from user profiles based on folksonomies. In: 1st International Workshop on Building Technology Enhanced Learning Solutions for Communities of Practice (TEL-CoPs 2006) (2006)
4. Firan, C., Nejdl, W., Paiu, R.: The benefit of using tag-based profiles. In: 2007 Latin American Web Conference (LA-WEB 2007), pp. 32–41. IEEE Press, New York (2007)
5. Godoy, D.: Comparing One-Class Classification Algorithms for Finding Interesting Resources in Social Bookmarking Systems. In: Lacroix, Z., Vidal, M.E. (eds.) RED 2010. LNCS, vol. 6799, pp. 88–103. Springer, Heidelberg (2012)
6. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information Retrieval in Folksonomies: Search and Ranking. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 411–426. Springer, Heidelberg (2006)
7. Huang, Y.-C., Hung, C.-C., Yung-Jen Hsu, J.: You are what you tag. In: AAAI Spring Symposium on Social Information Processing (AAAI-SIP), pp. 3–41 (2008)
8. Manevitz, L.M., Yousef, M.: One-class SVMs for document classification. Journal of Machine Learning Research 2, 139–154 (2002)
9. Michlmayr, E., Cayzer, S.: Learning user profiles from tagging data and leveraging them for personal(ized) information access. In: Workshop on Tagging and Metadata for Social Information Organization, Banff, Alberta, Canada (2007)
10. Michlmayr, E., Cayzer, S., Shabajee, P.: Add-A-Tag: Learning adaptive user profiles from bookmark collections. In: 1st International Conference on Weblogs and Social Media (ICWSM), Boulder, Colorado (2007)

11. Milicevic, A.K., Nanopoulos, A., Ivanovic, M.: Social tagging in recommender systems: A survey of the state-of-the-art and possible extensions. Artificial Intelligence Review 33(3), 187–209 (2010)
12. Noll, M.G., Meinel, C.: Web Search Personalization Via Social Bookmarking and Tagging. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 367–380. Springer, Heidelberg (2007)
13. Schifanella, R., Barrat, A., Cattuto, C., Markines, B., Menczer, F.: Folks in folksonomies: social link prediction from shared metadata. In: 3rd ACM International Conference on Web Search and Data Mining (WSDM 2010), pp. 271–280. ACM Press, New York (2010)
14. Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation 13(7), 1443–1471 (2001)
15. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In: ACM Conference on Recommender Systems (RecSys 2008), Lausanne, pp. 259–266 (2008)
16. Stoyanovich, J., Yahia, S.A., Marlow, C., Yu, C.: Leveraging tagging to model user interests in del.icio.us. In: AAAI Spring Symposium on Social Information Processing (AAAI-SIP), pp. 104–109 (2008)
17. Tonkin, E., Guy, M.: Folksonomies: Tidying up tags? D-Lib. 12(1) (2006)
18. Au Yeung, C.M., Gibbins, N., Shadbolt, N.: A study of user profile generation from folksonomies. In: Social Web and Knowledge Management, Social Web 2008 Workshop at WWW 2008 (2008)