User Interest Extraction based on Weighted Tags

Saida Kichou[1], Hakima Mellah[1], Imene Lasbeur[2] and Imen Abdelouahid[2]

[1] Research Center on Scientific and Technical Information (CERIST) Benaknoun, Algiers, Algeria
[2] University of Boumerdes, Boumerdès, Bumardas, Algeria

Abstract. Collaborative tagging systems are based on assigning keywords freely chosen by users, which promotes ressources sharing and organization by the way and improves the information retrieval. The tags allocation by users is illustrated particularly in sites sharing photos or videos (Flickr, YouTube). As navigations and clicks, tags can be good indicators of the user's interests. In this paper, we examine the limitations of previous tag-based profile extraction. We believe that for a better result, tags of a resource must represent well its content. Existing systems consider 'Popularity' as the unique criterion to judge the tag effectiveness. But it does not always reflect its importance and representativeness to the resource. In this paper, we propose a novel approach based on tag strength to represent a user. In which we introduce weighted tags based on user expertise.

1 Introduction

Collaborative tagging has become a very popular way to share, annotate, and discover online resources in Web 2.0. In this way, the user is becoming active; he is involved in the information production where he can enrich the content of these resources.

Collaborative or social Tagging is recently recognized for its potential to leverage collaborative production of information that support a wide range of mechanisms such as social search [24], and recommendation [21], although tagging was originally thought as a technique to improve personal content management.

Tagging system has emerged as a support to organize shared resources. It allows users to participate in content enrichment by adding key words (tags) to describe the resources, for a better categorization. Its simplicity and usefulness to improved information retrieval have attracted a high number of users [19].

As the social media are growing in terms of number of users, resources and interactions, the user may be lost or unable to find useful information. Social elements could avoid this disorientation like tags which become more and more popular and contribute to avoid the disorientation of the user [18]. Users on the Social Web interact with each other, create/share content and express their interests through chosen tags. Tags are new information to create or enrich the user profile [7].

Tags are tools to mark resources, on the one hand for guiding other users to have information [13], on the other hand to receive information about a user due to the history of tagging [12].

Kichou S., Mellah H., Lasbeur I. and Abdelouahid I..
User Interest Extraction based on Weighted Tags.
DOI: 10.5220/0004599300220031
In *Proceedings of the 2nd International Workshop on Web Intelligence* (WEBI-2013), pages 22-31
ISBN: 978-989-8565-63-1

Researches try to represent the user as accurate as possible through different techniques. Several studies are conducted in this area and proposed approaches for profile extraction.

In this paper, we present briefly collaborative tagging systems, in which a considerable number of users annotate shared resources, (text document, image, video) that may be affected several and divergent tags. These tags can enrich the user profile associating them. Then, we present a set of works in tag-based profile extraction, that propose different manner to construct profile, but often, they are based on tags popularity.

Knowing that the tag popularity for a given resource is the number of times it is cited, a popular tag is not necessarily representative of its resource. According to [8], it is very common for a user to repeat the same tags already associated to the resource, this can make the tag repeated popular without been really relevant to the content. To address this issue, we propose to use weighted tags instead of popular ones. According to this, we propose a novel approach to extract user profile by weighting first tags using the weighting method proposed in [15] and then calculate the tag strength based on the weighted tags instead of those based on popularity.

The paper is organized as follows, Section 2 shows briefly what the Collaborative Tagging is, in Section 3 we present the related works in our area. We present our approach in Section 4. Finally, we finish with a conclusion.

2 The Collaborative Tagging

Collaborative Tagging denotes the process of free associating one or more "tags" to a resource (web page, photo, video, blog ...) by a set of users. The term tagging is often associated with folksonomy, it refers to a classification (taxonomy) made by users (Folks) [16], [23], and defined by [3] as a series of metadata created by a collective users to categorize and retrieve online resources.

There are several tagging systems on the web such as Delicious for web pages, Flickr for images, YouTube for videos, Technorati for blogs and CiteULike for scientific papers.

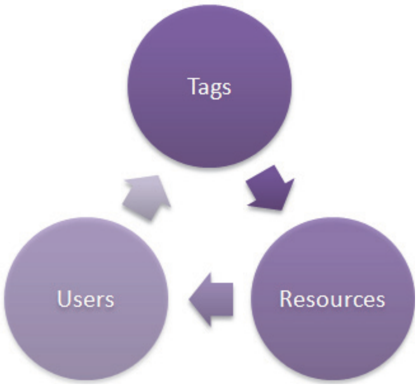

Fig. 1. Tripartite structure of tagging system.

In a tagging model, there is mainly three entities, users, tags and resources (Fig1). Links can be found between resources (such as links between web pages), and between users (social network) [17].

3 Related Work

The social user is characterized with his social activity like sharing information, communicating with other users and tagging resources. Several studies are conducted in profile extraction. [6] Considers that tags are a new type of user feedback, and can be a very important indicator of his preferences. Different approaches of profile construction based on tagging are presented.

[8] Presents the naïve approach and the co-occurrence one, that are used to construct a user profile. The first one results a generic profile, the second results a better one but tags are not weighted. [8] Presented also a new approach based on a user tags graph creation, which takes into account the age of the tag.

A hybrid approach was implemented in [15], it is a combination of naïve and co-occurrence approaches. It seems more efficient in that it results more specific profile with weighted tags.

To create a specific and dynamic user profile basis on tags, [14] introduced the concept of tag capacity to represent a resource based on two factors, the order of tagging (ie. the position of the tag in the list cited by the user) and popularity. According to [11] the first tag given by a user for a resource is more representative than the following. Huang in [14] used this theory to calculate the tag strength to describe a person (user).

Abel in [1] represents a tag-based profile as a set of weighted tags for cross-system user profile, as well as [10] creates a tag-based profile used for a better music recommendation on Last.fm, based on a logarithm function.

All these approaches are based on tags popularity, indeed tags representing a user are those cited by him weighted by their popularity. When we consider that popular tag can be not relevant to the resource, we consider that the use of popular tags is not enough to represent the user profile. Actually, such a definition of user profile induces a notion of similarity between a user profile and a resource profile [9].

Schöfegger in [20] tests in addition to the popularity binary values indication whether or not the user has used the tag.

Researches try to represent the user as accurate as possible, [9] enriches user profiles by “authoritative” tags, which are tags considered as important (for example tags having a high PageRank). This approach is graph-based, where we find two graphs: the *tag resource graph* called TRG and the *tag user graph* called TUG. These graphs are used in order to filter qualitative tags (i.e.: fun. good, etc.) then generate list of candidate tags by means of the IDDS (Iterative Deepening Depth First Search). Finally, the user profile is enriched by tags from these candidate lists. Although this approach has shown its usefulness, the tags filtering used is based only on popularity.

A user manipulates different tags for different resources; this implies that there can be a multitude of interests that cannot be restricted to a set of popular tags. These interests will have different importance and weight, the difference is due to the user preferences dedicated indirectly for resources.

For this, we propose an approach to extract profile by calculating tags strengths using tags weights inspired from our previous work [15].

4 User Interest Extraction Approach based on Weighted Tags

In this section, we present a new approach which aims to build an accurate user profile based on the user's tags.

The idea is to use our weighting methods, presented in [15], and combined with the strength formula presented in [14]. Our motivation is to improve the constructed profile precision, and knowing that a popular tag is not necessarily a good one to describe a resource, it is wiser to improve first the quality of the capacity notion (tag capability to represent a resource) and then calculate the tag strength to represent a person. For this, we first present a model of the user profile to contain their personal information, activity and expertise, and a construction method of this profile based on tags.

Then, we propose our formula to calculate the tag capacity to describe a resource and tag strength to describe a user. Obtained tags are classified by descending order of strength and the top n tags form the profile.

In the rest of this paper, we consider the following notations:

- $U=\{u_1, u_2,\ldots,U_m\}$ the set of users.
- $R=\{r_1, r_2,\ldots,r_n\}$ the set of shared resources in the system ;
- $T=\{t_1,t_2,\ldots,t_l\}$ the set of tags ;
- $Y=\{(u, t, r)_1, (u, t, r)_2,\ldots,(u, t, r)_p\}$ the set of annotations (the tagging actions, p is the number of actions) with $(u, t, r) \in U \times T \times R$;

We present below the different steps starting by the profile modeling, then the tags weighting and finally the strength calculation.

4.1 User Profile Modeling

To make our idea, we first define a user model that will contain his different information, and then we present the different steps of building the user profile. Here we use our model adopted in our last work [15], in the following a brief description.

Profile Representation. The user profile is a structure of heterogeneous information, which covers broad aspects such as cognitive environment, social and professional users [22]. This heterogeneity is often represented by a multidimensional structure. Eight dimensions in the literature are defined for the user profile [2], [4]: the personal data, interests, expected quality, customization, domain ontology, the return of preferences (feedback), the security and privacy and other information. A user profile is constructed either in a static way, by gathering information that rarely changes like name, age, etc., or in a dynamic way, by gathering information that frequently changes. Information about a user is obtained explicitly by the user himself or implicitly by observing his behaviour during his session (history, clicks, pages visited, etc.). The user profile contains information such as [25]: 1) Basic information which refers to the name, age, address, etc. 2) Knowledge of the user which is extracted generally from his web page navigation, 3) Interests which are defined through a set of keywords or logical expression, 4) History or feedback which design collected information form user's activity and could be deduced from number of clicks, time allowed in consulting resource, etc. and 5) Preferences.

Defining the profile of a particular user for a given application is equivalent to select the dimensions considered useful [4]. In our work, a user is defined by three dimensions. The first containing personal information, the second represents its interests and the latest information is the degree of expertise in the domain.

- *The Personal Dimension*: is used to identify the user (username, name, login, password ...). These information are introduced by the user.
- *The Expertise Dimension*: expert users in a given area, use specific terms to tag since they have a perfect mastery of the concepts in this area. This dimension is the degree of mastery of the user in tagged resources domain. It depends on the tag levels in the domain ontology used for this purpose. More the expertise is great more the user is close to the resource context. For more details see [15].
- *The Interest Dimension*: Interest dimension tells us about the user interests and preferences. This is what we will calculate using our formula. The interest dimension is represented as

Int (u_i) = {(t_1, strength$_1$), (t_2, strength$_2$) ... ... (t_j, strength$_j$)}. Where t is the tag, strength is the tag force to represent a user; 'j' is a chosen threshold.

Profile Construction. The user profile construction is building dimensions Int (u) and exp (u) based on tagging operations the user performs.

Construction of the Expertise Dimension. An expert user in one domain has a perfect mastery of specific terms in this domain. Therefore, he associates these terms with specific resources that he tags (eg in pharmacy, an expert associates the name of a drug molecule, whereas a novice just associates the term 'medicament').

Expertise is defined in [15] as the average depth of the user tags and it is calculated as follows:

$$\boldsymbol{Exp(ui)} = \frac{\sum \boldsymbol{prof(tj)}}{|\boldsymbol{Tu}|} \quad (1)$$

Where Prof (t) is depth of tag tj, that is the number of nodes separating it from the root in a given ontology (we use in our case Wordnet); T_u is a set of the user's tags that it has associated to resources, defined as follows:

$$T_u = \{t_j \mid (u_i, t_j, r) \in Y\}.$$

Construction of the Interest Dimension. Our calculation concerns this dimension. Initially we calculate tags weight based on the number of users used them, in a second step we calculate strength of each tag to describe a user. In the following we detailed these different steps.

4.2 Weighting Tags based on User Expertise

The tag weight is calculated according to the user who issued it. The same tag will be assigned two different weights, if it is assigned by two different users. On the other hand, for the same user, the tags associated with a resource should have different weights. We use an adaptation of our previous tag weight definition in [15] depending on the user expertise. With the aim to introduce the subjective aspect of the tag, the user feedback is introduced via a rating.

The tag weight is calculated as follows:

$$w_t^r = \sum_{i=0}^{k} [w_{order} * \mathbf{exp}\ (u_i)]^{conf(u_i,r)} \tag{2}$$

Where exp is the expertise already computed, k is the number of users associating tag t for the resource r.

W_{order} is the tag weight based on order used in [14], calculated as follow:

$$w_{order}(T) = \begin{cases} e^{-i/_{10}}, & i \leq 10 \\ e^{-1}, & i > 10 \end{cases} \tag{3}$$

This formula promotes first tags given by a user, after the tenth, the following have the same weight (e^{-1}).

Conf (u_i, r) represents the degree of trust (or confidence) of the user in his tag. This is achieved via a rating from one (01) to five (05) every time he tagged a resource. It is calculated as follows:

$$Conf(u_i, r) = \frac{d}{5}, (d \in \{0,1,2,3,4,5\}) \tag{4}$$

The degree of user confidence in the tag associated to the resource is used as a kind of weight regulator. If the user is at all not sure of his tag, he assigns a rating of 0 and the calculated weight becomes a simple popularity calculation, while if the user assigns the maximum score, his expertise is fully used in the tag weight. So it is the introduction degree of user expertise in the tag weight calculation.

4.3 Strength Calculation based on Weighted Tags

Inspired by the force formula proposed in [14], and once the tag weight is calculated based on the tag order, the user expertise and the confidence, the tag t strength to represent user u ($strength_u$ (t)) is calculated in our case as:

$$Strength_u(t) = \sum_{r \in R_u} w_t^r \tag{5}$$

With R_u a set of tagged resources by the user u. w_t^r is the tag weight calculated previously. So the interest dimension int(u) is:

Int(u)= {(t_1, Strength(t_1)), (t_2, Strength(t_2)) ... ... (t_j, Strength(t_j))}.

5 Experimentations

To test our approach we have implemented it, and implemented another approach based only on popular tags, to see the improvement.

We conducted tests on a collection of 100 URLs extracted from Delicious, tagged by 30 users with different expertise, using 223 tags. WordNet was used to calculate the tag depth.

5.1 Evaluation Process

We have, first, removed all tags that do not appear in WordNet. We calculated the corresponding depth using Wordnet for existing tags. Apply our formula to calculate the weight by calculating the needed parameters (order weight, expertise and confidence). For each tag we calculated its strength in the set of the user's resources.

5.2 Results and Discussion

Several tests were performed on the collection. We present in Table 1 an example of user's u interests presented in popular tags classified by descending order of popularity:

Table 1. User interests based on popularity.

Tags	Popularity	Tag	Popularity	Tag	Popularity
design	15	php	6	Books	3
information	12	ajax	5	Links	3
Website	10	travel	5	Programming	2
url	9	game	5	Data	1
java	8	football	5		
html	8	css	3		

If we consider j=5 (the threshold), the user interest is: intP(u)= {(design,15), (information, 12), (web site, 10), (url, 9), (java, 8)}.

The table2 illustrious the user interests with the proposed approach, classified by the strength.

Table 2. User interests based on strength.

Tag	Strength	Tag	Strength	Tag	Strength
java	31.85	url	11.2	Books	3.6
html	31.81	information	10.5	Links	2.6
php	30.65	css	9.72	Data	2.1
ajax	27.59	travel	9.2	Programming	1.9
design	24.15	game	9.01		
website	20.87	football	4.3		

If we consider j=5, the user interest is: int(u)= {(java,31.85), (html, 31.81), (php, 30.65), (ajax, 27.59), (design, 24.15)}.

We found that the obtained interest with the proposed approach contains more specific tags with different strengths unlike the popularity, that results generic tags and sometimes with identical popularities, (eg. 5 for ajax, travel, game and football).

To see more clearly our results, we introduce the tag depth which tells us on the specificity of the tag.

The following curves are obtained by illustrating tags that compose intP(u) (Popularity) and intS(u) (Strength) with their depths. The intP(u) curve shows first interests with shallower tags, the deeper tag has a depth less than 8. The intS(u) curve, shows more deeper tags in the top of the interests exceeding a depth of 8, then depth decreases for the following tags that is logically true.

Therefore we can say that the Interest obtained with the strength approach is more specific than the Interest obtained with popularity.

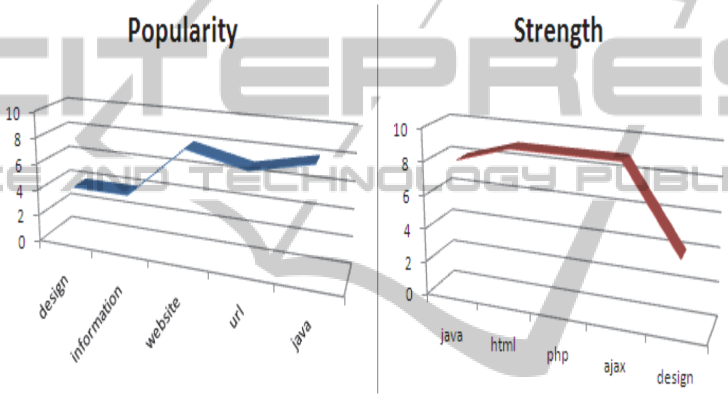

Fig. 2. Comparison between the strength approach and popularity.

6 Conclusions and Future Work

There are several approaches to extract the profile-based tags, which are generally based on popularity. This classification may result less representative tags, but other parameters can contribute to a better ranking. In this paper we described a technique for building a user's interests; our aim is to extract an accurate and dynamic profile so as to take into account changes in preferences over time with a more accurate manner. For this, we adapted our previous work [15], and we were inspired by [14] formula for the tag strength calculus, taking into account the tagging order and the user expertise and confidence. The proposed approach was evaluated on a collection extracted from Delicious, which is considered as reference system because of the huge number of registered users and the richness of tagging.

As a future work, we plan to exploit not only the WordNet hierarchy, but also domain ontologies in order to realize a more powerful technique and evaluate it on real data.

References

1. Abel, F., Araújo, S., Gao, Q., &Houben, G. J.: Analyzing Cross-System User Modeling on the Social Web. International Conference on Web Engineering (ICWE'11), Vol 6757, pp28-43. Springer. DOI=http://dx.doi.org/10.1007/978-3-642-22233-7_3. (2011)
2. Amato, G., Straccia, U.: User Profile Modeling and Applications to Digital Libraries.In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Paris, France (1999)
3. Broudoux, E.: Folksonomie et indexation collaborative, rôle des réseaux sociaux dans la fabrique de l'information. In: Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, Scotland (May 2006)
4. Bouzeghoub, M., Kostadinov, D.: Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils.In: Proceedings of Actes de la Conférence francophone en Recherche d'Information et Applications CORIA 2005, pp. 201–218 (2005)
5. Cattuto, C., Schmitz, C., Baldassarri, A., Servedio, V.D.P., Loreto, V., Hotho, A., Grahl, M., Stumme, G.: Network properties of folksonomies. AI Communications Journal, Special Issue on Network Analysis in Natural Sciences and Engineering (2007)
6. Carmagnola, F., Cena, F., Console, L., Cortassa, O., Gena, C., Goy, A., Torre, I.: Tag based User Modeling for Social Multi-Device Adaptive Guides. Special issue on Personalizing Cultural Heritage Exploration (2008)
7. Carmagnola, F., Cena, F., Cortassa, O., Gena, C., Torre, I: Towards a tag-based user model: how can user model benefit From tags? In: Proceedings of the International Conference on User Modeling.Corfù, Greece. Lecture notes in Computer Science, pp. 445–449. Springer. (2007)
8. Cayzer, S., Michlmayr, E.: Adaptive user profiles: Chapitre de livre Collaborative and social Information Retrieval and Access; ISBN-13: 9781605663067,(2009)
9. De Meo,P., Quattrone,G., Ursino, D.: A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. In User Modeling and User-Adapted Interaction 20:41–86 DOI 10.1007/s11257-010-9072-6,(2010)
10. Firan.S, Nejdl.W, Paiu.R.: The Benefit of Using Tag-Based Profiles. Proceedings of the 2007 Latin American Web Conference LA-WEB, page 32-41. Washington, DC, USA, IEEE Computer Society, (2007)
11. Golder Scott, A., Huberman, B.A.: The Structure of Collaborative Tagging System. Journal of Information Science 32(2), 198–208 (2005)
12. Gupta, M., Li, R., Yin, Z., Han, J. 2010. : Survey on social tagging techniques. In SIGKDD Explorations 12(1): 58-72. DOI=http://doi.acm.org/10.1145/1882471.1882480,(2010)
13. Helic, D., Trattnery, C., Strohmaier, M., Andrews, K.. On the Navigability of Social Tagging Systems. In socialCom/PASSAT, 161-168. IEEE Computer Society, (2010).DOI= http://dx.doi.org/10.1109/SocialCom.2010.31,(2010)
14. Huang, Y., Hung, C., Hsu, J.: You are what you tag. In Association for the Advancement of Artificial Intelligence, http://www.aaai.org.(2008)
15. Kichou, S., Mellah,H., Amghar,Y,. Dahak,F.: Weighting Tags Approach Based on User Profile . International Conference on Active Media Technology (AMT 2011), Lanzhou, China September 7-9, (2011).
16. Mathes, A.: Folksonomies - Cooperative Classification and Communication Through shared Metadata. Rapport interne, GSLIS, Univ. Illinois Urbana- Champaign (2004)
17. Marlow, C., Mor, N., Danah, B., Marc, D.: Tagging, taxonomy, flickr, article, toread. In: Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, UK (2006)
18. Mezghani,M,. A User Profile Modelling Using Social Annotations: A Survey, WWW 2012 – MultiAPro'12 Workshop, Lyon. France,(2012)

19. Michlmayr,E., Cayzer,S. : Learning User Profiles from Tagging Data and Leveraging them for Personalized Information Access. WWW2007, May 8–12, 2007, Banff, Canada.
20. Schöfegger.K, Körner.C : Learning User Characteristics from Social Tagging Behavior. HT'12, June 25–28, 2012, Milwaukee, Wisconsin, USA. (2012)
21. Sigurbjörnsson,B., and Zwol,R,V.: Flickr tag recommendation based on collective knowledge, in WWW'08. In Proceedings of the 17th international conference on World Wide Web (WWW '08). ACM, New York,USA, (2008)
22. Tamine-Lechani, L., Zemirli, N., Bahsoun, W.: Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'informations. In: Actes de la Conférence francophone en Recherche d'Information et Applications (CORIA 2006), Lyon, France (2006).
23. Vanderwal, T.: Explaining and Showing Broad and Narrow Folksonomies, http://www.vanderwal.net/random/entrysel.php?blog=1635, (2005)
24. Yahia,S. et al. "Efficient network aware search in collaborative tagging sites", In VLDB'08, pp. 710-721, (2008)
25. Zayani, C. A. Contribution à la définition et à la mise en oeuvre de mécanismes d'adaptation de documents semi-structurés. Doctoral Thesis. University of Toulouse. (2008)