Detecting user's interests based on the accuracy of collaborative tagging information

Manel Mezghani, André Péninou, Corinne Amel Zayani, Ikram Amous, Florence Sèdes

Detecting user's interests based on the accuracy of collaborative tagging information

Manel Mezghani[1,2*], André Péninou[2*], Corinne Amel Zayani[1*], Ikram Amous[1+] and Florence Sèdes[2*]

[1]Sfax University, MIRACL Laboratory, Sfax, Tunisia. [2]Paul Sabatier University, IRIT Laboratory.

*surname@irit.fr [+]*ikram.amous@isecs.rnu.tn*

Abstract. A user profile has to reflect the user's needs according to his characteristics (personal data, interests and preferences), his context, and his situation. In this article, we focus on the problem of adaptation quality in social networks, which is affected by the accuracy and relevance of the user's interests. The originality of our approach is the proposal of a new technique of interests' detection by analyzing the accuracy of the tagging behaviour of the user in order to figure out the tags which actually reflect the resources' content. Our approach has been tested and evaluated on the Delicious social database.

Introduction

Social information is permanently growing. Consequently, the adaptation process becomes more complex. The adaptation is a process strongly related to user's profile modelling. A profile that reflects the appropriate characteristics (interests, preferences, etc.) could avoid cognitive overload and disorientation of the user when accessing the information space. In our work, we are interested in detecting the user's interests that will be used in further works for an adaptation purpose.

Detecting social user's interests' is a non trivial problem (Milicevic *et* al., 2010). In fact, the user's profile building process suffers from the lack of information provided by himself. Indeed, the user generally doesn't give all the information related to his interests. So his profile can never be considered fully

known by a system.In order to overcome such a problem, the researchers have analyzed the social environment of the user such as his neighbours (the persons connected to the user explicitly or implicitly), his tagging behaviour (the collaborative action of tagging resources), or even the objects (the resources) he interacts with (see for example (Astrain *et* al., 2010)).

In this paper, we firstly present some existing works integrating the social environment of the user to detect interests. Then, we show the differences of our approach compared to the other approaches in the same context. We then describe our proposal for detecting interests and the experiments done to validate it. Finally, we conclude and discuss some future works.

Related works

According to (Astrain *et* al., 2010), interests could be deduced from the social environment based on the **user**, the **object** or even the **tag**. The collaborative tagging behaviour is described as the connection of these three elements: it represents the action of tagging a resource (object) by each user.

For the **user**, interests could be explicitly provided in the user's profile (Zayani *et* al. 2007), or implicitly deduced from his behaviour of navigation (Rebai *et* al., 2012) or behaviour of tagging (Kim *et* al., 2011). The user-based interest could be deduced from other users in the networks (neighbours) (Kim *et* al., 2011) (Tchuente, 2013).

For the **object**, interests are deduced based on the objects that the user accesses (White *et* al., 2009) (Ma *et* al., 2011). Objects could be any type of resource (URL, web page, image, etc.). Although these works are object-based, they do not analyze object's content. To analyze resource content, different techniques exist such as the indexation technique. Indexation is used in order to extract the significant terms from resources. After indexing resources different scoring function could be applied in order to detect the most relevant resource according to a specific query (Vallet *et* al., 2010).

For the **tag**, its utility has been proven to detect user's interest (Kim *et* al., 2011). **Tag**-based interest detection could be deduced by analyzing used tags (De Meo *et* al. 2010) or by analyzing the semantic of tags (Kim *et* al., 2011).

Synthesis

After presenting some researches done to analyze the tagging behaviour elements, we now discuss the main differences between our approach and the other researches: i) Unlike most of researches which focus on the tag content considered as an interest (by analyzing the semantic of the tags for example), we will focus on analyzing the accuracy of the tags with the resources' content. ii) We focus on analyzing the object-based rather than the user-based interest

detection. In fact object-based interest detection provides richer information than the user-based method (Song *et* al., 2011). iii) for object-based interest detection, most of researches do not consider the accuracy of the tags with the object (resource) content. This problem has been addressed in (Milicevic *et* al., 2010). However, the proposed approaches use techniques such as clustering, semantic processing, etc. and none of them use the resources' content analyze in their works. iv) dealing with the accuracy of the tag could overcome problems related to the nature of these social annotations. The main problem is the ambiguity associated to these tags since they are user generated keywords and do not follow any rules. This problem has been explicitly addressed in some researches (see (Mezghani *et* al., 2012) for more details). In our approach, this drawback will be treated automatically while detecting the accurate tags.

To summarize, our approach uses the users' tags and treats them according to the content of their respective resources. The accurate tags are those reflecting the resources' content. In order to validate our research, we will use the social environment that reflects the user's interests. The interests are stated accurate for a user since they exist in his neighbours' profile (Tchuente, 2013).

Proposed approach

In this section, we will propose our approach for detecting accuracy of the user's interest. This approach is based on the hypothesis that a user, who tags a resource with keywords reflecting its content, is really interested with the thematic of this resource. This observation will be experimented and validated on the Delicious social dataset.

Description

In our approach, we analyze the tags assigned to the resources to detect user's interest. The resources are generally a set of URLs describing them. We extract in the first step the tagging behaviour relations, composed by the tags applied to the resources by each user. Generally this activity is represented in a tripartite model which describes the users U={*u1,...ul*}, the resources being tagged R={*r1,...,rm*} and the tags T={*t1,...tn*} :

$$\text{Tagging relation} : <U, T, R> \quad (1)$$

where *l* the number of users, *n* the number of tags and *m* the number of resources. In the second step, we extract the content of these URLs and index them as semi-structured (XML) files, using the Lucene indexing tool API[1]. We will use it in order to figure out the most accurate tags with regard to the content of the tagged resources. Lucene relies on a field-based indexation technique. This characteristic

[1] http://lucene.apache.org/

enables indexing the documents according to one or more fields. Our indexation process is done according to the fields: title, content and URL. After indexing the content of the resources, we assign a rank to each resource according to the assigned tag. This rank is computed from a similarity between the resource (as a XML file) and the query (as a tag). Many similarity functions exist in the literature such as the similarity function supported by Lucene[2].
We run this scoring function according to the field content. After ranking the resources, we test if the resource tagged by the query exists in the top-k result provided by the ranking function. If it's the case, we state the tag as relevant to the resource. This step is iterated for all tags of each user's neighbour. In order to validate the relevant tags list, we compare the founded relevant tags (of the user's neighbours) with the user's tag (real tagging behaviour). The validation step will be detailed in the next section. Figure 1, describes the interest's detection process.

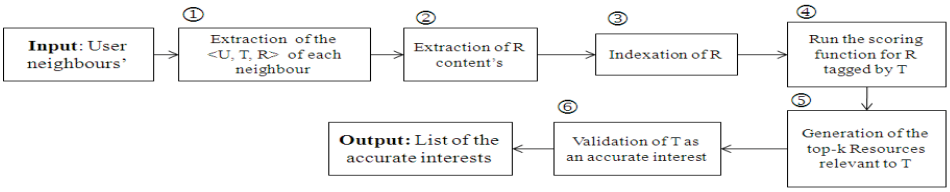

Figure. 1. The interest detection process.

Validation

We validated our approach upon the Delicious database that contains social networking, bookmarking, and tagging information. It provides information about the user's friend relationships and the tagging relation information <U, T, R>. The users U are described through their ID. The resources R are described through their ID, URL and title. The tags T are described through their ID and value. We have tested our approach on a set of 100 users. These users have different number of neighbours (varying from 1 to 20). The number of tags, documents and tagging relations is different for each user. This number may roughly vary from 10 to 500 for the tags, from 10 to 500 for the documents, and from 20 to 600 for the tagging relations. For the result of the top-k documents relevant to a query, we have chosen k=10000. The value of k is chosen according to the largest possible value, as-we wanted to test (in this first stage) with the maximum of results achievable (even those with lower scores). Also, the choice of the k value is proportional to the number of resources (69226 URLs) and tags (53388 tags) in the database.

Let's take as an example the tag "math" assigned by a user to different resources. This tag has a higher score according to the resource's title "IXL Math", which contains math related thematic, then the resource title "Online Dice Roller", which does not contain any information related to the thematic. So, according to this example, the tag "math" is relevant to the resource "IXL Math". After

[2] http://ipl.cs.aueb.gr/stougiannis/default.html

detecting this relevant tag, we will validate this result by using the user's neighbours. The validation objective is to show if this relevant tag is accurate to the user or not.

In this experiment, the neighbours are the explicit friendship relation (the user's egocentric network). The method of validation uses the social environment of the user (the neighbours) to detect interests. In fact the neighbours provide information that reflect the user's interests (Tchuente, 2013). We calculate the precision of the detected interests according to the tags in the neighbours' profiles. The precision is calculated according to the number of accurate tags (which exist in the user's neighbours profile) and the total number of tags provided as accurate.

$$\text{Precision} = \frac{\text{Number of accurate results}}{\text{Number of accurate results} + \text{Number of inccurate results}}$$

This precision is calculated for each single user's neighbour. The overall precision is the average of all the neighbours' precision. Figure 2, shows the overall precision, for this set of users, between the calculated relevant tags and the user's tag (real tagging behaviour).

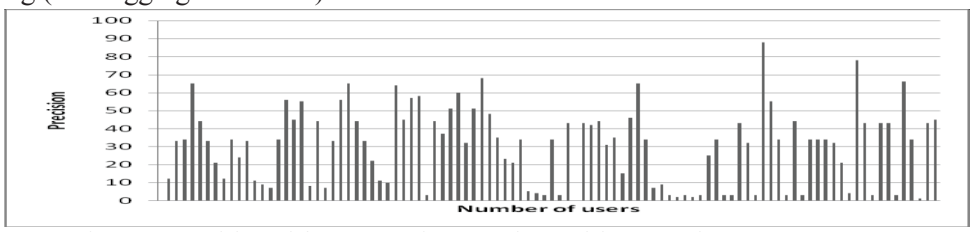

Figure. 2. Precision of the accurate interests detected for a set of 100 users.

Discussion

From this set of users, we have found that the precision vary according to different cases: i) for users who have a lot of friends, the precision is higher than those who have less friends, ii) the test has provided a precision for a few users equal to zero. This is due to the fact that a user may be friend to another user without sharing with him common interests. We have found that this special case is related to the users who have a little number of neighbours.

Also, the accurate interests provided by our approach are comprehensible keywords which reflect really the resource's content like "technology", "foursquare", "history", etc. This is an advantage since the tags are user-generated keywords. Our approach has filtered the ambiguous tags (i.e:"gis") that are not comprehensible by other users. The tags' ambiguity has decreased from 52% to 23% according to WordNet[3].

Conclusion

In this paper, we have proposed an approach for detecting accurate user's interests

[3] http://wordnet.princeton.edu/

based on the social environment. We have exploited the content of the tagged resources in order to figure out the tags reflecting really the thematic of the resources. We have validated our approach through the tagging behaviour of the neighbours (his egocentric network).

In future works, we will test our approach on a larger population of users in order to have more scalable results. Also, we will test other forms of neighbours such as, users tagging the same resources, or even users belonging to the same "community". In fact, a user may share common interests with other people than his explicitly friend relationship. Our approach could be used for an adaptation purpose (i.e.: enrichment of the user's profile, recommendation, etc.), since it provides a solution for detecting user's interests.

References

Astrain, J. J., Cordoba, A., Echarte, F. and Villadangos J. (2010): "An algorithm for the improvement of tag- based social interest discovery". SEMAPRO: The Fourth International Conference on Advances in Semantic Processing. 2010, pp. 49-54.

De Meo, P., Quattrone, G. and Ursino D. (2010):"A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy". User Modeling and User-Adapted Interaction. 2010, 20(1), pp. 41–86.

Kim, H.-N., Alkhaldi, A., Saddik, A. E. and Joi G.-S. (2011): "Collaborative user modeling with user- generated tags for social recommender systems". Expert Systems with Applications, 2011, pp. 8488–8496.

Ma, Y., Zeng, Y., Ren, X., Zhong, N. (2011): "User Interests Modeling Based on Multi-source Per-sonal Information Fusion and Semantic Reasoning". Active Media Technology (AMT) 2011: 195-205.

Mezghani, M., AmelZayani, C. A., Amous, I. and Gargouri, F. (2012): "A user profile modelling using social annotations: a survey".WWW (Companion Volume), 2012, pp. 969-976.

Milicevic, A. K., Nanopoulos, A. and Ivanovic, M. (2010):"Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions". Artif. Intell. Rev. Vol. 33 no. 3, 2010, pp. 187-209.

Rebai, R.Z, Zayani, C. A. and Amous, I. (2012): "An Adaptive Navigation Method for Semi-structured Data". Advances in Databases and Information Systems ADBIS (2), 2012, pp. 207-215.

Song, Y., Zhang, L., and Giles, C. L. (2011): "Automatic tag recommendation algorithms for social recommender systems". ACM Trans., Web.Vol.5, no.1, Article 4, 2011, pp. 1-31.

Tchuente, D., (2013): "Modélisation et dérivation de profilsutilisateursà partir de réseauxsociaux :approche a partir de communautés de réseaux k-egocentriques". Doctoral thesis, University of Toulouse, 2013.

Vallet, D., Cantador, I. and Jose, J. (2010): "Personalizing Web Search with Folksonomy-Based User and Document Profiles Advances in Information Retrieval". Advances in Information Retrieval, 2010, Vol. 5993, pp. 420-431.

White, R., Bailey P. and Chen, L. (2009): "Predicting user interests from contextual information". International Conference on Research and Development in Information Retrieval (SIGIR), 2009, ACM, New York, NY, USA, pp. 363–370.

Zayani, C. A., Péninou, A., Marie-Françoise, C. and Sedes, F. (2007): "Towards an Adaptation of Semi- structured Document Querying". Proceedings of the CIR'07 Workshop on Context-Based Information Retrieval CIR 2007, CEUR-WS.org, Vol. 326.