Producing relevant interests from social networks by mining users' tagging behaviour: A first step towards adapting social information

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

Producing relevant interests from social networks by mining users' tagging behaviour: A first step towards adapting social information

Manel Mezghani[b,a,*], André Péninou[b], Corinne Amel Zayani[a], Ikram Amous[a], Florence Sèdes[b]

[a] *Sfax University, MIRACL Laboratory, Sfax, Tunisia*
[b] *IRIT, Université de Toulouse, CNRS, INPT, UPS, UT1, UT2J, France*

Keywords:
User interests
Tagging behaviour
Resource
Indexation
Social network
Adaptation
Indexing methods
Semi-structured data and XML

A B S T R A C T

Social media provides an environment of information exchange. They principally rely on their users to create content, to annotate others' content and to make on-line relationships. The user activities reflect his opinions, interests, etc. in this environment. We focus on analysing this social environment to detect user interests which are the key elements for improving adaptation. This choice is motivated by the lack of information in the user profile and the inefficiency of the information issued from methods that analyse the classic user behaviour (e.g. navigation, time spent on web page, etc.). So, having to cope with an incomplete user profile, the user social network can be an important data source to detect user interests. The originality of our approach is based on the proposal of a new technique of interests' detection by analysing the accuracy of the tagging behaviour of a user in order to figure out the tags which really reflect the content of the resources. So, these tags are somehow comprehensible and can avoid tags "ambiguity" usually associated to these social annotations. The approach combines the tag, user and resource in a way that guarantees a relevant interests detection. The proposed approach has been tested and evaluated in the *Delicious* social database. For the evaluation, we compare the result issued from our approach using the tagging behaviour of the neighbours (the egocentric network and the communities) with the information yet known for the user (his profile). A comparative evaluation with the classical tag-based method of interests detection shows that the proposed approach is better.

1. Introduction

Social media has been successful in recent years, with millions of users visiting sites like *Facebook*, *Twitter*, *Flickr*, *Delicious*, etc. These social media sites principally rely on their users to create content, to annotate others' content with tags,[1] ratings, and comments and to make on-line relationships. As social media sites continue to grow and social content continue to evolve, the users are having more difficulty finding the information they need. To avoid this problem, researchers have chosen adaptation as a classic solution like [1]. We also adopt this solution and apply it in a social context.

Adaptation is a process strongly related to user modelling. In fact, each user has specific needs and then requires specific adaptation. So, we adapt the resources of social media according to each user profile. Adaptation could be reached through a

* Corresponding author at: IRIT Laboratory, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9.
E-mail addresses: mezghanni@irit.fr (M. Mezghani), peninou@irit.fr (A. Péninou), corinne.zayani@isecs.rnu.tn (C.A. Zayani), ikram.amous@isecs.rnu.tn (I. Amous), sedes@irit.fr (F. Sèdes).

[1] Social annotations.

http://dx.doi.org/10.1016/j.datak.2016.12.003

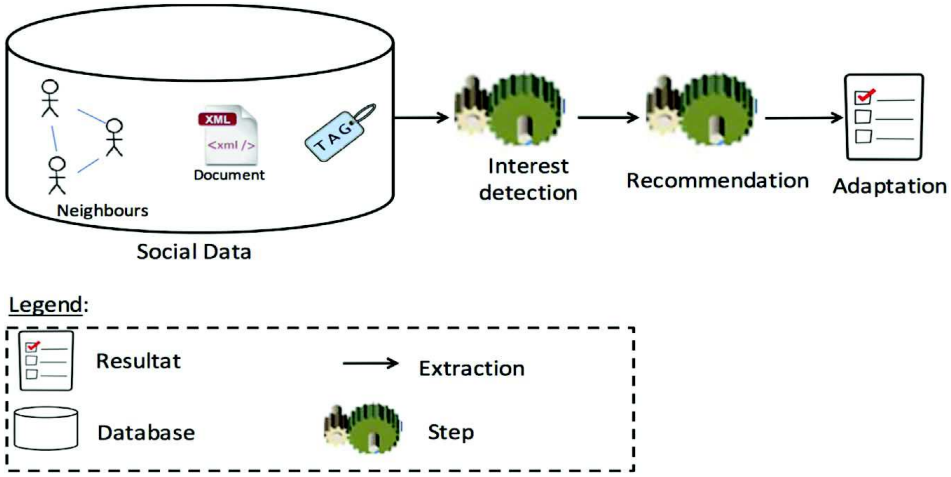

Fig. 1. Interest detection in adaptation process.

technique of recommendation or personalization. We focus on recommender systems research area which has recently put a lot of attention in social networks by developing a new class of systems called social recommender systems [2–6]. However, to achieve a reliable recommendation, we should first detect the accurate user interests (see Fig. 1). We must highlight that developing a recommender system is out of the scope of this paper. Detecting accurate interests will be useful for proposing a robust recommendation system in further works. So, we develop, in this paper, a method of social interests detection that can be used for a recommendation purpose.

There are many sources from which we can extract user interests. In a classical way, user interests are extracted from his own profile (e.g. interest attribute). The user profile contains informations provided explicitly by the user himself and stored in the database of the system (e.g. name, surname, age, profession, etc.). Interests are also extracted from his social behaviour (e.g. tagging behaviour) or his social network (e.g. friends). However, in a social context, many parameters make the detection of interests a crucial problem. We focus on some problems that affect the interest detection process:

1. *Lack of information in the explicit user profile*: The user generally does not give all the information related to his interests and then the explicit user profile can never be considered fully known by a system. So, we cannot focus on analysing the profile to detect relevant interests [7].
2. *User activity*: The user is more and more active (participate in discussions, comment and tag resources, etc.). Consequently, detecting his needs becomes harder [8]. Indeed, the user can describe his interests through different manner and then the choice of the behaviour to analyse may be challenging.
3. *A lot of information*: Detecting relevant social interests has to face with the evolutionary aspect of social networks. In fact, the quantity of information (users and content) is in exponential growth. For the users, many relationships may be established (friend relationships, users belonging to the same group, etc.). For the content, many types of information are available in social networks such as images, web pages, videos, etc. This variety makes the interest detection harder, since the user may interact with several contents.
4. *The influence of other users*: In social networks, the user may be influenced in a negative way by his neighbours (e.g. other users in the network such as his friends, users tagging the same resource, etc.). In fact, spammers orient the user to follow specific resources by tagging them with the same tag. However, neighbours may be a benefit information for detecting relevant interests [4] and then for a better adaptation [2]. The complexity of detecting "good" neighbours leads to the complexity of finding accurate interests from these users.

Our approach is proposed from the hypothesis that social environment and especially neighbours provide an information from which user interests may be extracted [7]. This hypothesis was proved in the context of finding social user profiles from social networks. So, we analyse neighbours in order to detect the most relevant interests to each user. This analysis aims to infer information that reflects better user interests, through analysing social information of neighbours. Neighbours could be the explicit friend relationship, users sharing some commons behaviours (e.g. visiting or tagging the same resource), the egocentric network, users belonging to the same community, etc. Through this paper we call neighbour as a generic term and we will specify it in the proposed approach part. We focus on some social information of neighbours such as:

- The tagging behaviour (of the neighbours), which reflects user opinion about a resource [9]. This information has proven its utility to detect user interests [2,4].
- The tagged resources content (of the neighbours), which is considered as an information that could be used to detect interests. In fact, resource-centred approaches are more robust because of the richer information contained in the resources compared to user-centred approaches [6].

The proposed approach treats mainly the textual resources (semi-structured resources, plain text, etc.) that are present in almost all main social networks such as *Delicious* by analysing the tagged URL, *Twitter* by analysing the *tweets*, etc. Our approach does not deal with pictures (for example the case of *Flickr*).

In order to validate our finding, we compare the founded relevant tags (our approach applied to user's neighbours) with the user's tags (real tagging behaviour). Our approach is experimented on the *Delicious* social database. Different forms of neighbours are considered in our validation (an egocentric network[2] and users belonging to the same community[3]). Our findings are compared with the approach using directly tags provided by the users (tag-based approach). This approach is the continuation of our work already done in [10].

The rest of this paper is structured as follows. In the second section we present related works. In the third section we present a synthesis of the state of the art. In the fourth section we present and describe the proposed approach. In the fifth section we present and comment the results of our experiment on *Delicious*. In the last section we conclude and present the perspectives of our work.

2. Related works

Interests could be implicitly deduced from user's behaviour. The advent of social networks has created new behaviours associated with the user reflecting his interests. In fact, the social user is no longer belonging to the audience but is becoming an active contributor to the creation of the social content. The social user is more and more active (exchanges information, takes part in groups, etc.) and curious (compares to get the best information, seeks opinions, etc.). So, the methods for detecting user interests are focusing on social information rather than focusing on the user himself.

According to [9], interests could be deduced from the social environment based on the **user**, the **resource** or even the **tag**. We present some researches focusing on each element.

2.1. Interest detection from users

The user-based interest detection could be deduced from other users in the networks (neighbours) [4,7]. The neighbours are considered as an important source of data since they have proved their utility to overcome the "cold-start" problem for new users in the system, to reflect the user interests [7] and also to enrich the users' profiles for recommendation purpose [2,4].

Neighbours reflect the social relation of the user with other users. This relation could be explicit (friend relationship) or implicit (e.g. users who interact in the same resource, users sharing commons interests, etc.). This social relationship is recently and well detailed in Musial and Kazienko [11]. Neighbours of the user in social context are described through ties. Where, "*a tie between two users aggregates all types of the relations that exist between these two persons*" [11]. Some studies analyse neighbours in order to detect users considered close to the user, in term of interests. Neighbours are detected by several metrics such as cosine similarity [4], the Jaccard similarity [12], "X-compass" [8], etc. Other studies detect neighbours through observations like the work of [4], which enriches the user profile with tags of his friends not included in the profile based on the observation that two people share common tags are considered close people and may well have interests in common.

Other researchers try to combine different parameters in order to detect the similarity between users. Cabanac [13] calculates the similarity between authors by analysing their proximity, their connectivity and the number of paper in common. Guy et al. [14] calculate the score of proximity through different criteria: (i) more people and/or tags within the user profile related to the item, (ii) the stronger relationship of these people and/or tags to the user, (iii) the stronger relationships of these people and/or tags to the item, and (iv) the freshness to the item. Roth et al. [15] detect the implicit relationship between users through their mail exchange. They calculate the proximity through the frequency of interaction between user, the freshness of the interaction and the direction of the interaction.

Neighbours could also be detected in graph-based context, where [7] analyse the egocentric networks to detect interests and [16] detect communities to analyse their dynamics.

To summarize, the detection of neighbours depends on the context of the work. Also, a neighbour is an information which may contain "good" persons who influence the user in a good way (to enrich user profile, to recommend relevant resources, etc.) or "bad" persons who influence the user in a negative way like spammers (who disorient the user).

2.2. Interest detection from resources

Interests are deduced based on the objects/resources that the user accesses [8,17]. The resource can be any type (URL, video, image, etc.). In [8] user interests are discovered by keywords extraction and analysis from each single source (sources are *Facebook*, *linkedIn*, etc.). In [17] user interests are discovered from the analysis of the user historical behaviour of visiting resources, time spending on a web page, etc.

Although these works are resource-based, they do not analyse the content of the resource. To analyse the content of the resource, different techniques exist such as indexation that is used in order to extract 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 [18].

[2] The explicit relationships between the user with the other users.

[3] The community is generated through a classical predefined algorithm, out of the scope of this paper.

The accuracy of a query (regarding a resource) may be scored through different scoring function applied in information retrieval such as TF*IDF, BM25, etc. These scores are the result of an indexing process, which invoke a query and a collection of resources. The use of these methods has shown their utility and robustness in the information retrieval [19].

In social context, the query can be a tag. Content of tagged resources have been analysed in recommendation purpose from a machine learning perspective of view in [6]. Also, [20] proposes a recommendation approach for social tagging systems (modelled as a Latent Dirichlet Allocation approach) that combines content and relation analysis in a single model.

However, most of researches do not consider the accuracy of the tags with the resource content. This accuracy reflects if the user is really interested with the content or not. For example if a user tag a politic resource with a politic tag, this reflects that the user is interested with the politic thematic. In the opposite scenario, we can assume that a user may be a spammer or it is not interested in the content of the resource. Spam may be treated (filtered) implicitly while determining the most significant tags according to resources.

2.3. Interest detection from tags

Several researches have focused on detecting social user interest from different social information and especially from tags. The utility of tags has been proved to detect user interests [4]. Also, tags are considered as a powerful tool to reflect user's opinion about a resource [2]. In fact, every user could describe his opinion on the content of a resource with his own keywords. Tag-based user profile modelling, in an adaptation context, has been detailed in [21].

Tags are generally used to overcome the lack of information in the explicit user profile. But, tags do not follow any rules and then, they could contain information not relevant to the resource content like personal tag, spam, etc. Tags which are considered as personal, reflect the “feeling” of the user and not the content of the resource like “good”, “awesome”, etc. Also, this could be harmful facing spammers (persons who influence the users in a negative way to follow specific information) or when users assign words not understood by the community.

The set of tags assigned to resources is called *folksonomy* [2,5,18]. *Folksonomy* is a powerful tool for capturing the collective knowledge. Unlike ontology, *folksonomy* is not structured. This characteristic leads to an ambiguous vocabulary which may influence the understanding of the user interests by the system or even by other people in the network. To avoid these problems, many metrics are envisaged like *SpamFactor*, *SpamRank*, *spamClean*, etc. [5]. Also, several techniques are used such as clustering, converting *folksonomy* to an ontology, or in a classical way by using a natural language processing tool likes WordNet.[4] These techniques aim to structure the folksonomy in a comprehensible way to use it a recommendation, personalized, etc., purpose. More details about treating tag's ambiguity are explained in [5,21].

3. Synthesis

The motivation of our approach is that a relevant aspect of a social adaptation system needs to capture the user interests using relevant social information. So, when adaptation is produced, the estimated interests of a user may be considered as irrelevant, due to the inefficiency of the used information. To overcome this problem, our approach makes a selective use of the available information about interests to produce an accurate interests list for each user.

In order to develop our approach, we analyse the tagging behaviour of the neighbours. This choice is motivated by (i) the studies which promote the collective knowledge to reflect the user interests [2,4,7] and (ii) by the lack of the information in the explicit user profile. To be more precise, our approach detects relevant interests for a user based on the analysis of the set of resources (e.g. bookmarks) tagged by his neighbours. So, relevant interests for a user are part of the set of neighbours' tags that reflect the content of resources.

To summarize, our approach tries to combine the tag, user and resource in a way that guarantees a relevant interests detection approach. Our approach uses the neighbours' tags and treats them according to the content of their respective resources. The relevant tags are those reflecting the resources content.

In order to validate the potential detected interests, we compare them with the user interests (his tags considered as his interests). So, the relevant interests (issued from the analysis of the neighbours) are stated accurate for a user since they exist in the user profile [7].

4. The interest detection approach

In this section, we detail our approach of detecting users interests. The problems that affect the interest detection process are listed in Section 1. Our approach tries to overcome these issues as follows:

- For **User activity** and **Lack of information in the explicit user profile** issues, the approach focuses on the user behaviour. This behaviour concern mainly his tagging behaviour in order to benefit from this explicit information provided by the user that may reflect his interests. The approach focuses on user's social behaviour to infer his interests.
- With respect to the **A lot of information** issue, the approach analyzes the neighbours and mainly the egocentric network and

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

the communities in order to reduce the spectrum of analysis and then avoid the scalability issue of social data.
- For **The influence of other users** issue related to spammers, the approach analyses of tags and their relevance to the associated resource. So, tags not describing the content are discarded and then the possible tags of spammers are reduced.

In our approach, we analyse the tags assigned to the resources to detect user interests. Let us note:

- $U=\{u_1,...,u_n\}$, the set of users in the social network, where n is the number of users.
- $R=\{r_1,...,r_m\}$, the set of resources in the social network, where m is the number of resources.
- $T=\{t_1,...,t_h\}$, the set of tags, where h is the number of tags and $u \in U$.
- $N_u=\{n_{u1},...,n_{uj}\}$, the set of neighbours of the user u, where j is the number of neighbours and user $u \in U$.
- $I_u=\{i_{u1},...,i_{uk}\}$, the set of relevant interests of the user u, where k is the number of relevant interests and user $u \in U$. This is the result of our algorithm.

The approach of interest detection is performed through two main steps. First we prepare the data that we will use. Then, we proceed to apply our approach. This latter aims to generate the relevant resources according to each tag, then to score these resources and to select the top-k resources. If the tag assigned by the user to a resource that is in the top-k, then the tag is considered an accurate interest. We explain each step below.

4.1. Data preparation

Before explaining our approach, we proceed to prepare the data used as an input for detecting user interests.

We extract in the first step the data from the social network. This data concern mainly: (i) The tagging behaviour relations $\langle U, T, R\rangle$, which are composed of the tags applied to the resources by users. (ii) The neighbours N_u (the egocentric network, the communities, etc.). In this next section, we propose the algorithm of interests detection according to a non-predefined form of set of neighbours. We will test in Section 5 different form of set of neighbours in order to show their influence on the results. (iii) The content of resources in the network (i.e. URLs).

After extracting the data, we index the extracted resources. Indexation aims to describe the content of a document by keywords. The resources are indexed (as semi-structured resources or plain text) using the *Lucene* API.[5] *Lucene* is a tool for indexing and searching technology. *Lucene* is a field-based indexation technique. This characteristic allows indexing the resources according to one or more fields. For example, fields could be the *title*, the *content*, the *URL*, etc. We have taken into account only the *content* field.

The indexing process is as follows: Lucene indexes resources by dividing them into a number of terms. Terms are generated using an analyser that converts each word in its root form. Then, it stores the terms in an index file (*IndexFile*), where each term is associated with the resource content.

4.2. Approach

We present the general algorithm of our approach in Table 1 and then the detail of each function. This algorithm is applied for all users U. The function *Add(param*1, *param*2), allows us to add the *param*2 into the *param*1. So, there no overwriting of the *param*1.

We begin with generating the relevant resources R' to a given tag, where $R' = \{r'_1, \ldots, r'_v\}$ the set of relevant resources and v the number of relevant resources and $R' \subseteq R$. We use the function Add() in order to add each relevant resource into R'. This step interrogates the *IndexFile* (the output of the indexation step). When a request/query is made it is treated by the same analyser used to build the index and then used to find the corresponding term(s) in the index. This provides a list of resources matching the query. In our context, a query is considered as a tag throughout the rest of this paper. We present the algorithm of generation resources relevant to a given tag $t_h \in T$ (see Table 2).

After generating relevant resources (R') according to a specific tag (t_h), a score is assigned to each resource according to the assigned tag. The purpose of using such score is to separate the most relevant resources related to a specific tag. This score is the result of a function of similarity which takes into consideration the resource (textual) and the tag. Many similarity functions exist in the literature such as the similarity function supported by *Lucene*. We choose a predefined function[6] of similarity which is a variant of the TF-IDF scoring model. The choice of such a model is due to the fact that TF-IDF is an efficient and simple algorithm for matching words in a tag to resources that are relevant to that tag. However, the main limitation of such a model is that it does not take into consideration the relations between words (e.g. synonyms). The similarity function is described through the formula (1) as follows:

$$score(q, r) = coord \cdot queryNorm(q) \cdot \sum_{t \in q} \{tf(t \in r) \cdot \mathrm{idf}(t)^2 \cdot t.\ getBoost()(t, r)\} \tag{1}$$

The term t is the result of the resource indexation process. Each term t is associated with a resource r. The elements of this

[5] http://lucene.apache.org/core/
[6] http://lucene.apache.org/core/3_5_0/scoring.html

Table 1
The general algorithm of the interest detection approach for a specific user u.

Input: N_u, T_{nuj}, *IndexFile*
// T_{nuj} is the set of tags of the neighbours
Output: I_u
$I_u = \emptyset$, $R' = \emptyset$, $R'' = \emptyset$
1. For each $n_{uj} \in N_u$
2. $R' = \text{GenerationResourcesRelevantToTag}(T_{nuj}, \textit{IndexFile})$
3. $R'' = \text{Scoring}(R', T_{nuj})$
4. Add (I_u, SelectionRelevantTag(T_{nuj}, R''))
5. End For
Return I_u

Table 2
The algorithm of the generation of the resources relevant to each tag.

GenerationResourcesRelevantToTag (T_{nuj}, *IndexFile*)

Input: N_u, T_{nuj}, *IndexFile*
Output: R'// Set of the resources relevant to a each $t_h \in T_{nuj}$
$R' = \emptyset$
1. For each $t_h \in T$ do
2. Add (R', LuceneGeneration (t_h, *IndexFile*))
/* Generate List of resources R' relevant to the tag.*/
3. End For
4. **Return** R'

predefined scoring function are described as follows:

- $score(q, r)$ is the score affected to a specific resource r according to a specific query q.
- $coord(q, r)$ is a score factor based on how many of the query q terms are found in the specified resource r.
- $queryNorm(q)$ is a normalizing factor used to make scores between queries comparable.
- $tf(t \in r)$: Term Frequency of the term t in the resource r. It is defined as the number of times term t appears in the currently scored resource r.
- $idf(t)$: Inverse Document Frequency measure the importance of a term t in all the collection of resources.
- $t.\ getBoost()$ is a search time boost of term t in the query q. The boost is 1.0 by default.
- $norm(t, r)$ is a value of different boost and length factors: (i) Document boost sets a boost factor for hits on any field of the current resource. This value will be multiplied into the score of all hits on this resource. (ii) Field boost sets the boost factor hits on the current field. This value will be multiplied into the score of all hits on this field of a resource. (iii) lengthNorm(field): computed when the resource is added to the index in accordance with the number of tokens of this field in the resource, so that shorter fields contribute more to the score. The returning value is a normalization factor for hits on this field of this resource.

We run this scoring function according to the field *content*. This function provides a result of the top-k resources R'' relevant to the query q considered as a tag, where $R'' = \{r''_1, \ldots, r''_w\}$, the set of top-$k$ relevant resources, where w is the number of relevant resources and $R'' \subseteq R'$. We use the function Add() in order to add each relevant resource according to a tag into R''. The scoring algorithm of the resources is described in Table 3.

Table 3
The algorithm of scoring the resources for a given tag.

Scoring (R', T_{nuj})

Input: R', T_{nuj}
Output: R''// Set of the $\text{top}-k$ $r_{v'} \in R'$ relevant to $t_h \in T_{nuj}$
$R'' = \emptyset$
1. For each $r_{v'} \in R'$ do
2. For each $t_h \in T_{nuj}$
3. score[]=score ($r_{v'}$, t_h) // Lucene scoring function
4. End For
5. Add (R'', Top-k Generation($r_{v'}$, score[]))
6. End For
7. **Return** R''

Table 4
The algorithm of selection of relevant tags.

SelectionRelevantTag(T_{nuj}, R'')

Input: T_{nuj}, R''
Output: I_u
I_u=∅
1. For each $t_h \in T_{nuj}$ do
2. If($\exists\ r_{v''} \in R''$, $t_h\ \exists\ \langle U, T, R''\rangle$)
3. Add (I_u, t_h). // Add the tag t into the set of the relevant interests of the user u.
4. End If
5. End For
6. **Return** I_u

This algorithm generates the set of relevant resources (R'') from the previous set (R') according to the specific tag (t_h) and the top-k resources having the higher score. For example, using a tag=“math”, one resource belonging to R'' is associated to the one with the title=“IXLMath” and its URL=“http://www.ixl.com/”.

After scoring the resources, we test if the resource tagged by q exists in the top-k result provided by the scoring function. If it is the case, the tag q is stated as relevant to the resource. This step is iterated for all tags of each neighbour. This process is described in Table 4.

This algorithm generates a list of relevant interests (I_u), for each user, as a list of tags that better describe the content of the tagged resource.

5. Evaluation approach

In order to validate our proposal, we consider users for which we have got an existing profile. The profile is defined as the list of tags assigned by the user and it is considered as a ground truth. So, we compare tags of the user (issued from his profile) with tags issued from our approach (calculated from neighbours). In our social analysis, we have built a list of interests (tags) without considering the user himself. So, the data used as ground truth (user profile) is different from the data for computation (neighbours). The calculated profile is validated through comparing it with the user interests (the ground truth). The evaluation process is described through Fig. 2.

Given a target user $u \in U$, our approach builds a candidate set I_u={i_{u1},...,i_{uk}} (see Section 4.2) of potential interests. I_u is the calculated profile. For each $i_{uk} \in I_u$, we analyse the existence of the interest i_{uk} in the profile of the target user u. All the interests i_{uk} satisfying this test are the potential interests that are considered as correct (“real” interests). They are grouped in the set C_u, where C_u={c_{u1}, ..., c_{uy}} and y the number of correct interests, and each $C_{uk} \in I_u$ (so, $C_u \subset I_u$).

The validation of our proposal is done through an existence test of the user interests in the potential interests calculated by our approach. This test is done through two methods:

- By a simple comparison of the tags contents (i.e. if user-tag=“picture” and neighbour-tag=“picture”, then the tag “picture” is considered relevant). We will call this technique in the rest of this paper as “simple comparison”.
- By taking into account the synonyms or the related word or even the form of the tag (singular/plural) (i.e. if user-tag=“picture”

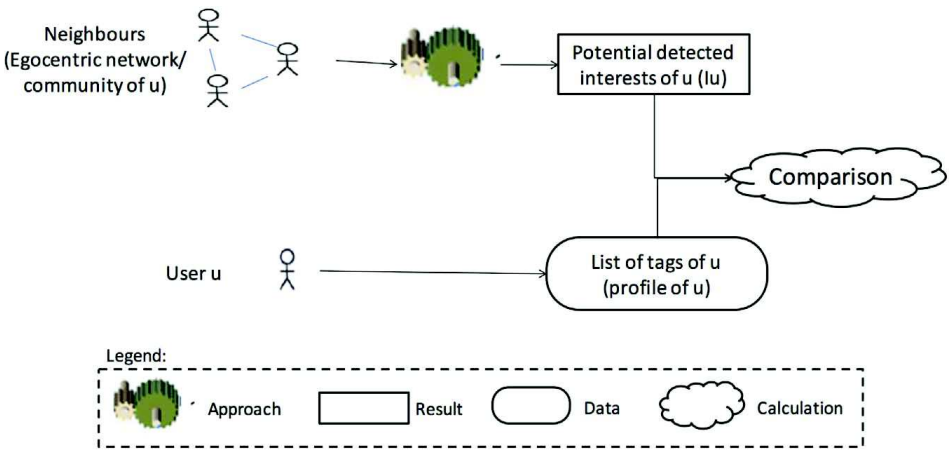

Fig. 2. Evaluation process.

and neighbour-tag="pictures" or "photo", then the tag "picture" is considered relevant). The synonyms and related words are detected by interrogating Wordnet.

6. Experiment on *Delicious*

In this section, we present the used *Delicious* dataset and the evaluation of the results of our experiment.

6.1. Dataset

The *Delicious*[7] database contains social networking, bookmarking, and tagging information. This dataset is extracted from [22]. It provides information about the friend relationships and the tagging information. The users are described through their ID. For example: Users={1, 2, 3, 4, 5, 6}

The resources are described through their ID, title and URL. For example: Resources: {

1 IFLA – The official website of the International Federation of Library Associations and Institutions http://www.ifla.org/,

7 EdSelect http://www.edselect.com/,

8 Cool Canada (Collections Canada) http://www.collectionscanada.gc.ca/cool/index-e.html,

9 Kidsreads.com http://www.kidsreads.com/, }

The tags are described through their ID and value. For example: Tags: { 1 collection_development, 2 library, 3 collection, 4 development, 5 lesson_plan, 6 war, 7 veterans, 8 discover, 9 canada, 10 read_ books }

We present some statistics of the data of this dataset in Table 5.

We consider in our work a user profile as tags assigned by the user. For example: User profile={bullying, bullyingvideo, design, designs, florida, journals, minimal, package, pakage, women }

Neighbours are users connected (as his egocentric network or community) to a current user. For example: Neighbours: { 1, 45, 6, 3456, 3001 }. The neighbours (users) are described through their ID.

6.2. Evaluation

Our approach is evaluated according to two criteria:

First, we study the influence of social environment of the user, and mainly the influence of the chosen set of neighbours, in the precision of the results. We have evaluated our approach in two ways using the user egocentric network and also according to the user communities (issued from a specific community detection algorithm). According to Section 5, we test two methods of validation, and retain the one which provides the best results to do the rest of the evaluations. We tested the influence of the value of k that select the top-k resources relevant to a tag. We retain the value that provides better results.

Second, we compare our approach with the approach that uses the tag information of the neighbours without any pre-treatment (classical tag-based approach). To be more precise, we compare our approach that analyzes tags according to their relevance to the content of the resources with the approach that takes directly into consideration tags provided by the user.

6.2.1. Evaluation according to the set of neighbours

We run our approach on all the users of the database. These users have different number of neighbours (which may vary from 1 to 90 neighbours). The number of tags, resources and tagging relations is different for each user. This number may roughly vary from 3 to 800 for the tags, from 10 to 450 for the resources, and from 20 to 500 for the tagging relations. We calculate the precision of the detected interests according to the tags produced by our approach and using neighbours (formula (2)). The precision $P(u)$ for each user $u \in U$ is calculated according to the number of really accurate tags ($Cu \subset Iu$), that exist in both user profile and calculated results (from neighbours' profiles), and the total number of tags provided as accurate (Iu) (through our approach):

$$P(u) = |Cu|/|Iu| \tag{2}$$

We calculate also the recall of the detected interests according to the tags calculated (through our approach) from the neighbours' profiles (formula (3)). The recall $R(u)$ for each user $u \in U$ is calculated according to the number of accurate tags ($Cu \subset Iu$) and the total number of tags in the user profile (Tu).

$$R(u) = |Cu|/|Tu| \tag{3}$$

We have calculated the average precision of the users according to the egocentric network and also according to the communities. The egocentric network is defined as the set of users connected explicitly with a given user with one degree of separation [7]. The definition of the community is proposed by [16] and used in [7]. The community is detected through an algorithm called "iLCD" which have proven his utility. We use this algorithm in order to generate communities associated with our dataset. The community may contain a set of users who could be present in the egocentric network of the given user. We calculate the average precision for all users (formula (4)) provided from the precision formula $P(u)$ (formula (2)) for the user u, where n is the number of the users (in our case n=1867):

[7] www.delicious.com

Table 5
Dataset presentation.

Number	Description	i.e.
1867	Users	
7668	Bi-directional user relations	15,328 (user_ i, user_ j) pairs and average of 8.236 relations per user.
69226	URLs	
38581	Principal URLs	www.delicious.com
69226	Tags	
437593	Tag assignments (tas)	Tuples [user, tag, URL], an average of 234.383 tas per URL and an average of 6.321 tas per tags.
104799	Bookmarks	Distinct pairs [user, URL] obtained from tas, an average of 56.132 bookmarked URLs per user and an average of 1.514 users bookmarking a URL.

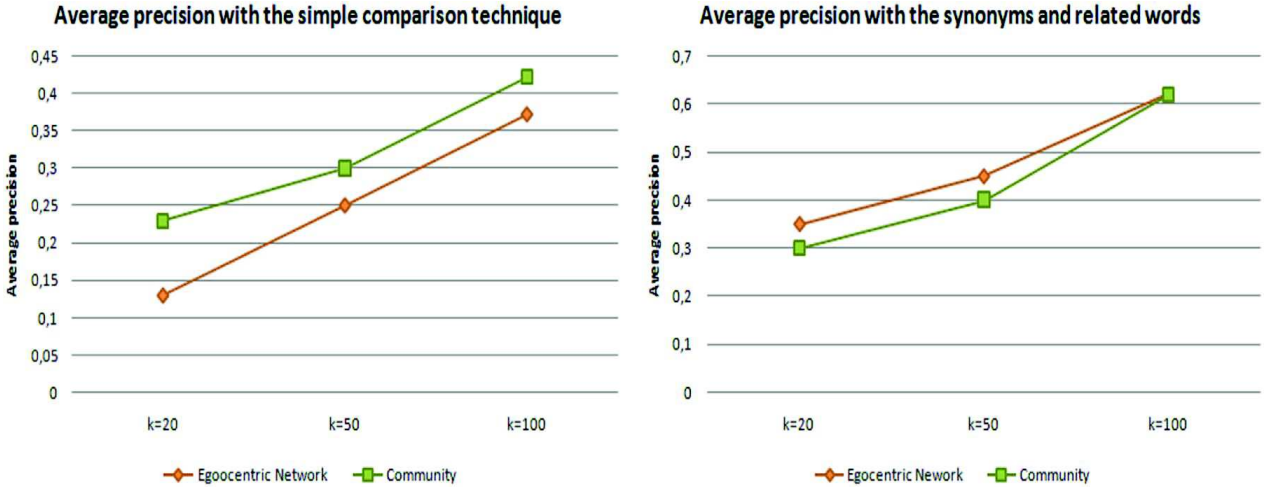

Fig. 3. (Left) The average precision according to k=20, k=50 and k=100 according to the simple comparison technique. (Right) The average precision according to k=20, k=50 and k=100 according to the synonyms and related words.

$$Average\ Precision = \sum_{i=1}^{n} P(u)/n \tag{4}$$

We calculate also the average recall for all users (formula (5)) provided from the precision formula $R(u)$ (formula (3)) for the user u, where n is the number of the users (in our case n=1867):

$$Average\ Recall = \sum_{i=1}^{n} R(u)/n \tag{5}$$

Our approach has been tested with different values of the top-k such as k=20, k=50 and k=100. We calculate the average precision according to the egocentric network and the average precision according to the communities for both methods of evaluation: the

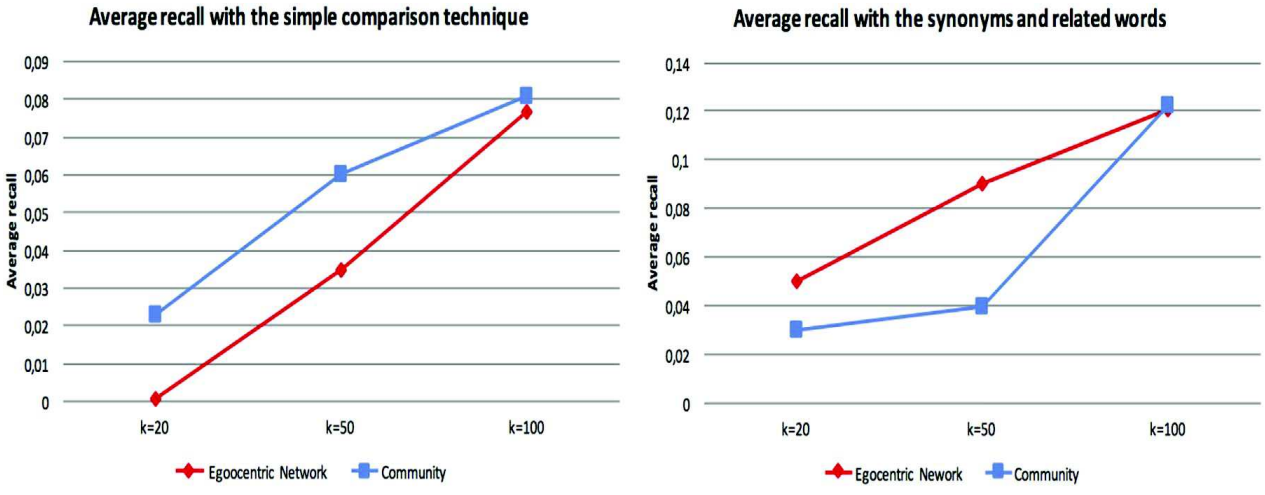

Fig. 4. (Left) The average recall according to k=20, k=50 and k=100 according to the simple comparison technique. (Right) The average recall according to k=20, k=50 and k=100 according to the synonyms and related words.

Table 6
The average precision of all users according to the egocentric network and the communities (k=100).

Average precision	Simple comparison technique	S and RW
Egocentric network	0.3830	0.6038
Communities	0.4042	0.6125

Table 7
The average recall of all users according to the egocentric network and the communities (k=100).

Average recall	Simple comparison technique	S and RW
Egocentric network	0.0766	0.1207
Communities	0.0808	0.1225

simple comparison technique and the synonyms and related words (Fig. 3).

We calculate also the average recall according to the egocentric network and the average recall according to the communities for both methods of evaluation: the simple comparison technique and the synonyms and related words (Fig. 4).

From these two tests, we clearly see that the precision and the recall that take into consideration the synonyms and related words are better than the simple comparison technique. This is an expected result since users may have the same interests (tags) but they may describe them differently, using different tags.

We choose k=100 (of the top-k resources relevant to a tag) for the rest of the evaluation since it provides better results (Figs. 3 and 4). We calculate the average precision of all users in the database according to the simple comparison technique and according to the consideration of the synonyms and related words (S and RW). Table 6 shows the different values in terms of precision according to the egocentric network and according to the communities (extracted from Fig. 3).

We calculate the average recall of all users in the database according to the simple comparison technique and according to the consideration of the synonyms and related words (S and RW). Table 7 shows the different values in term of recall according to the egocentric network and according to the communities (extracted from Fig. 4).

We notice that the consideration of the synonyms and related words provides higher precision and recall values for both the communities and the egocentric network. Also, for the simple comparison technique, the average precision and recall are higher for the communities. The precision and recall according to the synonyms and related words is higher for the egocentric network than for the communities.

The precision reflects how many selected interests are relevant. The recall reflects how many relevant interests are selected. Since, our aim is to detect relevant interests for each user, the precision metric is more adequate in this context. So, for the rest of the paper we focus only on precision metric.

We present the box plot of our precision result in Fig. 5. These box plots reflect the distribution of the precision values in the results (according to four quantiles). They are more representative than simple average of the precisions. For each method of comparison, the superior extremity of the continuous line represents the maximal value of the obtained values, whereas the extremity inferior represents the minimal value. Concerning the rectangle, it recovers all the values situated between the first and the third quartile, that is the values of 25% of the data are situated below the first quartile, and 25% of the data are situated above the third quartile. The gap interquartile thus corresponds to 50% of the values situated in the central part of the distribution, and is thus used as indicator of dispersal.

According to Fig. 5, we notice that: (i) For the precision according to the synonyms and related words, the distribution is almost

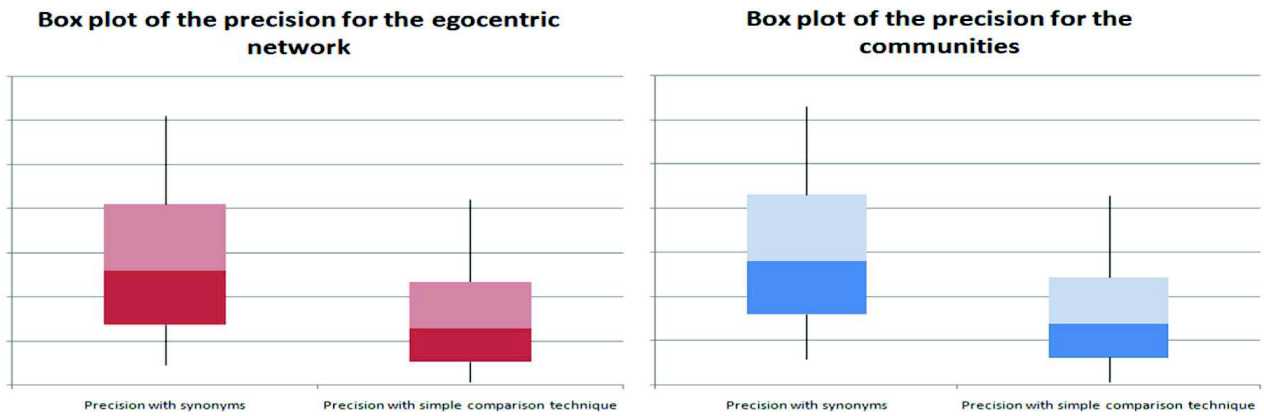

Fig. 5. Box plot of our approach according to the egocentric network (left) and the communities (right).

in the middle for both the egocentric network and communities. This reflects that most of users have the same average precision. (ii) For the precision according to the simple matching technique, the distribution is below the distribution of the synonyms and related words, for both the egocentric network and communities. This reflects that most of the users have the same lower precision.

From this evaluation, we clearly see that the precision that takes into consideration the synonyms and related words is generally better than the simple comparison technique. This is an expected result because users may have the same interests but they may describe them differently.

To summarize the obtained results, we have done some statistics for all 1867 users that show the difference between the two techniques of validation: (i) for the egocentric network, the precision is 100% higher for the synonyms and related words than the simple comparison technique, (ii) for the communities, the precision is 98.01% higher for the synonyms and related words than the simple comparison technique. These values show the relevance of the synonyms and related words to reflect the user interests.

Moreover, we have noticed that the precision (for the two methods of computation) varies according to different cases: (i) the precision is higher for active users (having a lot of neighbours and a lot of tagging behaviour); (ii) the precision is less higher for less active users; (iii) the precision equal to zero (in both cases) due to the gap of the number of tags provided by the user versus his neighbours. For example, in the one hand, the number of tagging relations of a user is equal to 20. In the other hand, the number of tagging relations of all his neighbours is equal to 500. This difference reduces the precision rates.

We calculate the difference between results obtained from synonyms and related words and from simple comparison technique. We call this difference a distance. The distance is calculated for the egocentric network and for the communities. Let us note:

- $PEgo.\ SRW_u$ the precision value of the user $u \in U$ according to his egocentric network with the consideration of synonyms and related words.
- $PEgo.\ SC_u$ the precision value of the user $u \in U$ according to his egocentric network with the simple comparison technique.
- $PCom.\ SRW_u$ the precision value of the user $u \in U$ according to his communities with the consideration of synonyms and related words.
- $PCom.\ SC_u$ the precision value of the user $u \in U$ according to his communities with the simple comparison technique.

The distance between two precision values is defined as the value of the difference of these values. The average distance (AD) is the average value of all the distances. The average distance is detailed for the egocentric network $AD\_\ Ego$ (formula (6)) and for the communities AD_Comm (formula (7)) as follows:

$$AD\ Ego = \sum_{i=1}^{n} (PEgo.\ SRW_u - PEgo.\ SC_u)/n \tag{6}$$

$$AD\ Comm = \sum_{i=1}^{n} (PCom.\ SRW_u - PCom.\ MT_u)/n \tag{7}$$

The values of the average distance (for all users of the dataset) between the results for the egocentric network $AD\_\ Ego$=0.2208 and for the communities $AD\_\ Comm$ = 0.2082.

The average distance is a positive value which reflects the relevance of the method considering the synonyms and related words for both the egocentric network and the communities.

Moreover, we tested if our approach dealt with the ambiguity of the resulting tags. We note that the accurate interests provided by our approach are comprehensible keywords which reflect really the content of the resource like “technology”, “foursquare”, “history”, etc. This is an advantage since the tags are user-generated keywords. Our approach has filtered the ambiguous tags (e.g. “gis”) that are not comprehensible by other users. The tags ambiguity has decreased (for this set of users) from 35% to 10% according to WordNet. So, the gap of tags ambiguity between the original data (before treatment) and the results (after treatment) equals to 71.25%.

6.2.2. Evaluation according to tag-based approach

Using the same set of users that we have used the previous section, we compared our approach with the classical tag-based approach. This latter considers the tags as the users interests [9,23]. We compare the result provided by our approach with the result of the approach that uses all the tags of the neighbours (without considering their relevance to the associated resources). The comparison is done according to the egocentric network of the user and also according to his communities. We compare according to k=100 of our approach (since it provides better results). Moreover, we compare by taking into consideration only the synonyms and

Table 8
The average precision of our approach and the tag-based approach.

Average precision	Our approach	Tag-based approach
Egocentric network	**0.6038**	0.3459
Communities	**0.6125**	0.3259

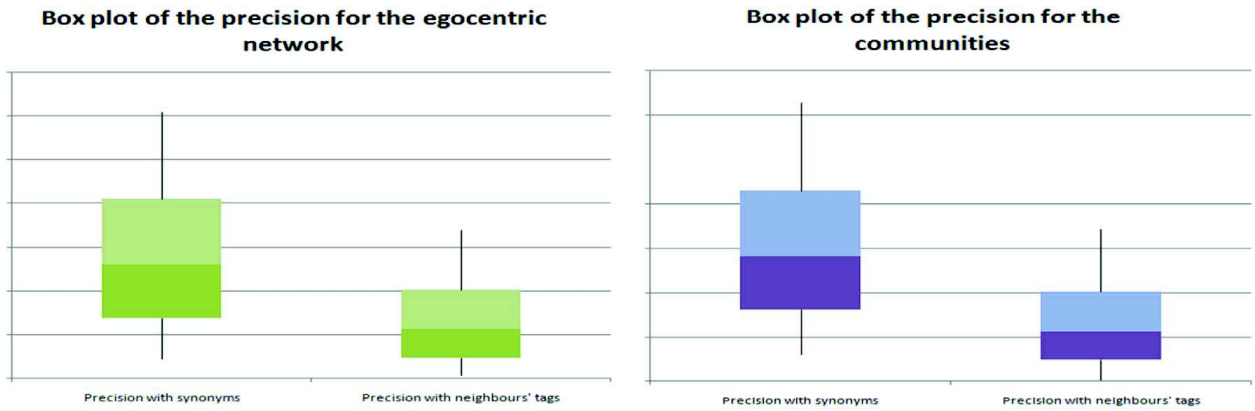

Fig. 6. Box plot of our approach according to the egocentric network (left) and the communities (right).

the related words (since it is better than the simple comparison technique). We calculate the average precision of all users in the database and compare it with the average precision provided by our approach.

Table 8 shows that our approach overcomes the classical tag-based approach in terms of precision. This is due to the consideration of the content of the resources analysed for the selection of relevant tags. The selection process implicitly filters ambiguous tags that may not be comprehensible for other users. Consequently, we obtain a higher precision than the tag-based approach.

Similar to the previous section, we present the box plot of our result according to the precision values in Fig. 6. According to Fig. 6, we notice that: (i) For the precision according to the synonyms and related words, the distribution is almost in the middle for both the egocentric network and communities. This reflects that most of users have the same average precision. (ii) For the precision according to the neighbours' tags, the distribution is below the distribution of the synonyms and related words, for both the egocentric network and communities. This reflects that most of the users have the same lower precision.

Through these comparisons, we notice that our approach performs generally better than the tag-based approach. Also, we notice that the better results are related to the active users. In fact, for the egocentric network, the precision is 88.10% higher for our approach with the synonyms and related words than the tag-based approach (for all users). Also, for the communities, the precision is 91.10% higher for our approach with the synonyms and related words than the tag-based approach (for all users).

We calculate the difference between results obtained from our approach with synonyms and related words and from the tag-based approach. We call this difference a distance. Similar to Section 6.2.1, the distance is calculated for the egocentric network and for the communities. Let us note:

- $PEgo.\ SRW_u$ the precision value of the user $u \in U$ according to his egocentric network with the consideration of synonyms and related words.
- $PEgo.\ TB_u$ the precision value of the user $u \in U$ according to his egocentric network with the tag-based approach.
- $PCom.\ SRW_u$ the precision value of the user $u \in U$ according to his communities with the consideration of synonyms and related words.
- $PCom.\ TB_u$ the precision value of the user $u \in U$ according to his communities with the tag-based approach.

The average distance is detailed for the egocentric network $ADEgo'$ (formula (8)) and for the communities $AD.Comm'$ (formula (9)) as follows:

$$AD\ Ego' = \sum_{i=1}^{n} (PEgo.\ SRW_u - PEgo.\ TB_u)/n \tag{8}$$

$$AD\ Comm' = \sum_{i=1}^{n} (PCom.\ SRW_u - PCom.\ TB_u)/n \tag{9}$$

The values of the average distance (for all users of the dataset) between the results for the egocentric network $ADEgo' = 0.2742$ and for the communities $AD.Comm' = 0.2930$. The average distance is a positive value which reflects the relevance of our method for both the egocentric network and the communities.

7. Conclusion

In this paper, we have proposed an approach for detecting accurate user interests based on the social environment. The goal was to infer users interests from content of the tagged resources in order to figure out the tags really reflecting the thematic of the resources. The originality of our approach is based on the proposal of a new technique of interests detection by analysing the

accuracy of the tagging behaviour of a user in order to figure out the tags which really reflect the content of the resources. So, these tags are somehow comprehensible and can avoid tags "ambiguity" usually associated to these social annotations. This is done through an indexation technique followed by an algorithm that score tags assigned to resources. This score reflects the relevance of the tag according to a resource. From this score, we have selected the most relevant resources (top-k). If the tag assigned by the user to a resource that is in the top-k, then the tag is considered an accurate interest.

The experiments have been done on the *Delicious* database. The validation aims to compare two different forms of set of neighbours: the egocentric network and the communities, with the user profile. The precision and the recall results have shown that the egocentric network and the communities reflect more the user interests if we consider the tags synonyms and the related words.

According to the result the values of k (that selects the top-k resources relevant to a tag) influences the precision values. The best value is k=100 and it is used to do the evaluation.

The experiment shows that our method provides a comprehensible set of interests. Consequently, our approach could be used for a purpose of adaptation (e.g. enrichment of the user profile, recommendation, etc.), since it provides a solution for detecting relevant user interests.

The results have proved that the consideration of the tagged resources to detect the relevant user interests (our approach) is better than considering directly the tags assigned by the users (classical tag-based approach). In fact, our approach has treated the tag ambiguity and then, has provided better results.

Our approach has treated problems that affect the interest detection process (listed in Section 1) as follows: For the **User activity** and the **Lack of information in the explicit user profile** issues, the approach has focused on the user behaviour and mainly his tagging behaviour. This social behaviour is almost present in every social network. So, our approach could be applied in other networks. In case of non-existence of the tags information, we may consider other textual information that reflects the user interests such as comments, "like" button, etc. With respect to the **A lot of information** issue, our approach has analysed the neighbours (the egocentric network and the communities) and then has avoided the scalability issue of social data. In fact, the choice of a set of users has reduced the spectrum of analysis and then has reduced the quantity of information to analyse. For **The influence of other users** issue, our approach has analysed tags and their relevance to the associated resources. As a result, our approach has filtered tags not describing the content. In consequence, the approach has reduced the possible tags of spammers.

In future works, we will do some other experiments in order to determine the set of neighbours that reflects more the user interests (from the egocentric network and the communities). We will also test our approach on other databases to improve our results.

We will deal with the drawbacks of our method which are the following: (i) our approach is not able to treat the non-textual information (e.g. image, video, etc.). We will propose then a more generic solution for the interests detection. This could be done for example by analysing the metadata present in these types of information; (ii) also, we have noticed that the analysis of the active users provides better results than the less active ones. So, our approach performs worse for the new users in the system. This is due to lack of information initially provided by these categories of users. So, the case of the not very active users should be treated separately in order to provide them better results.

Acknowledgements

This work was financially supported by the "PHC Utique" program of the French Ministry of Foreign Affairs and Ministry of Higher Education and Research and the Tunisian Ministry of higher education and scientific research in the CMCU project number 30540XK.

References

[1] R.Z. Rebai, C.A. Zayani, I. Amous, An adaptive navigation method for semi-structured data, in: T. Morzy, T. Hrder, R. Wrembel (Eds.), , Advances in Databases and Information Systems, Advances in Intelligent Systems and Computing vol. 186, Springer, Berlin, Heidelberg, 2013, pp. 207–215 URL: ⟨http://link.springer.com/chapter/10.1007/978-3-642-32741-4_19⟩.

[2] P.D. Meo, G. Quattrone, D. Ursino, A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy, User Model. User-Adapt. Interact. 20 (2010) 41–86.

[3] P.d. Meo, E. Ferrara, F. Abel, L. Aroyo, G.-J. Houben, Analyzing user behavior across social sharing environments, ACM Trans. Intell. Syst. Technol. 5 (2014) 14:1–14:31.

[4] H.-N. Kim, A. Alkhaldi, A. El Saddik, G.-S. Jo, Collaborative user modeling with user-generated tags for social recommender systems, Expert Syst. Appl. 38 (2011) 8488–8496.

[5] A.K. Milicevic, A. Nanopoulos, M. Ivanovic, Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions, Artif. Intell. Rev. 33 (2010) 187–209.

[6] Y. Song, L. Zhang, C.L. Giles, Automatic tag recommendation algorithms for social recommender systems, ACM Trans. Web 5 (2011) 4:1–4:31.

[7] D. Tchuente, M.-F. Canut, N. Jessel, A. Peninou, F. Sedes, A community-based algorithm for deriving users' profiles from egocentrics networks: experiment on facebook and DBLP, Soc. Netw. Anal. Min. 3 (2013) 667–683.

[8] Y. Ma, Y. Zeng, X. Ren, N. Zhong, User interests modeling based on multi-source personal information fusion and semantic reasoning, in: Proceedings of the 7th International Conference on Active Media Technology, AMT'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 195–205. URL: ⟨http://dl.acm.org/citation.cfm?id=2033896.2033923⟩.

[9] J.J. Astrain, A. Cordoba, F. Echarte, J. Villadangos, An Algorithm for the Improvement of Tag-based Social Interest Discovery, 2010, pp. 49–54. URL: ⟨http://www.thinkmind.org/index.php?view=article & articleid=semapro_2010_3_10_50021⟩.

[10] M. Mezghani, A. Péninou, C.A. Zayani, I. Amous, F. Sedes, Analyzing tagged resources for social interests detection, in: ICEIS 2014 – Proceedings of the 16th International Conference on Enterprise Information Systems, vol. 1, Lisbon, Portugal, 27–30 April, 2014, pp. 340–345. http://dx.doi.org/10.5220/0004971303400345.
[11] K. Musial, P. Kazienko, Social networks on the internet, World Wide Web 16 (2013) 31–72.
[12] H.-N. Kim, A. Roczniak, P. Lévy, A. Saddik, Social media filtering based on collaborative tagging in semantic space, Multimed. Tools Appl. 56 (2012) 63–89.
[13] G. Cabanac, Accuracy of inter-researcher similarity measures based on topical and social clues, Scientometrics 87 (2011) 597–620.
[14] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, E. Uziel, Social media recommendation based on people and tags, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10, ACM, New York, NY, USA, 2010, pp. 194–201. http://dx.doi.org/10.1145/1835449.1835484. URL: ⟨http://doi.acm.org/10.1145/1835449.1835484⟩.
[15] M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, R. Merom, Suggesting friends using the implicit social graph, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, ACM, New York, NY, USA, 2010, pp. 233–242. http://dx.doi.org/10.1145/1835804.1835836. URL: ⟨http://doi.acm.org/10. 1145/1835804.1835836⟩.
[16] R. Cazabet, F. Amblard, C. Hanachi, Detection of overlapping communities in dynamical social networks, in: 2010 IEEE Second International Conference on Social Computing (SocialCom), 2010, pp. 309–314. http://dx.doi.org/10.1109/SocialCom.2010.51.
[17] R.W. White, P. Bailey, L. Chen, Predicting user interests from contextual information, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, ACM, New York, NY, USA, 2009, pp. 363–370. http://dx.doi.org/10.1145/1571941.1572005. URL: ⟨http://doi.acm.org/10.1145/1571941.1572005⟩.
[18] D. Vallet, I. Cantador, J.M. Jose, Personalizing web search with folksonomy-based user and document profiles, in: C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rger, K.v. Rijsbergen (Eds.), , Advances in Information Retrieval, Lecture Notes in Computer Science vol. 5993, Springer, Berlin, Heidelberg, 2010, pp. 420–431 URL: ⟨http://link.springer.com/chapter/10.1007/978-3-642-12275-0_37⟩.
[19] Y. Cai, Q. Li, Personalized search by tag-based user profile and resource profile in collaborative tagging systems, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, ACM, New York, NY, USA, 2010, pp. 969–978. http://dx.doi.org/10.1145/1871437.1871561. URL: ⟨http://doi.acm.org/10.1145/1871437.1871561⟩.
[20] B. Zhang, Y. Guan, H. Sun, Q. Liu, J. Kong, Survey of user behaviors as implicit feedback, in: 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE), vol. 6, 2010, pp. 345–348. http://dx.doi.org/10.1109/CMCE.2010.5609830.
[21] M. Mezghani, C.A. Zayani, I. Amous, F. Gargouri, A user profile modelling using social annotations: a survey, in: Proceedings of the 21st International Conference Companion on World Wide Web, WWW '12 Companion, ACM, New York, NY, USA, 2012, pp. 969–976. http://dx.doi.org/10.1145/2187980.2188230. URL: ⟨http://doi.acm.org/10.1145/2187980.2188230⟩.
[22] I. Cantador, M. Szomszor, H. Alani, M. Fernandez, P. Castells, Enriching ontological user profiles with tagging history for multi-domain recommendations, in: 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008), 2008. URL: ⟨http://eprints.ecs.soton.ac.uk/15451/⟩
[23] X. Li, L. Guo, Y.E. Zhao, Tag-based social interest discovery, in: Proceedings of the 17th International Conference on World Wide Web, WWW '08, ACM, New York, NY, USA, 2008, pp. 675–684. http://dx.doi.org/10.1145/1367497.1367589. URL: ⟨http://doi.acm.org/10.1145/1367497.1367589⟩.

Manel Mezghani received her Ph.D. degree in Computer Science, in 2015, from the Paul Sabatier University, Toulouse, France and from Faculty of Economics and Management of Sfax, Tunisia. She is a Researcher at Toulouse Research Lab in Computer Science (Institut de Recherche en Informatique de Toulouse, IRIT). Its current research areas include user profile, social network analysis, semi-structured document, adaptive systems, multimedia information retrieval and mobile social networks.

André Péninou is currently an Associate Professor in Computer Science at the University of Toulouse 2 Jean-Jaurès and researcher at Toulouse Research Lab in Computer Science (Institut de Recherche en Informatique de Toulouse, IRIT). He got his Ph.D. degree in Computer Science from the Paul Sabatier University, in 1993. His research area is multimedia and semi-structured document personalization. His current research topics include metadata extraction from multimedia documents (video), search-oriented document annotation, user interests detection in egocentric social networks, and thus relevant documents filtering for users.

Corinne Amel Zayani is an Associate Professor in Computer Science at the Faculty of Sciences in the University of Sfax in Tunisia. She has received the master's degree in Computer Science from University of Valenciennes and Hainaut-Cambresis, in 2001. She got his Ph.D from University Paul Sabatier at Toulouse 3, in 2008. She is a Member of MIRACL Laboratory of Sfax University. Its current research areas include distributed and social adaptive systems, semi-structured documents and Web services.

Ikram Amous Ben Amor received her Master's degree in Data Processing from Paul Sabatier University (Toulouse III, France), France, in 1999. She obtained her Ph.D. in informatics from the Paul Sabatier University, in December 2002. She is currently an Associate Professor at the Higher School of Electronic and Telecommunication of Sfax in Tunisia (Enet'Com). She is also a Member of MIRACL laboratory of Sfax University, Tunisia. Her research interests include social network and user profile analysis, adaptive semi-structured document, etc. She has participated in several program committees of national and international conferences.

Florence Sèdes is a Full Professor in Computer Science at the University Toulouse 3. She got her Ph.D. degree in 1987.

From 2011 to 2013, she was appointed as the Deputy Director of the Toulouse Research Lab in Computer Science (Institut de Recherche en Informatique de Toulouse, IRIT), a joint unit with the National Scientific Research Center (CNRS). She leaded the National Research Network i3 (Information, Intelligence, Interaction). She is a Member of the National University Council of the French Ministry.

Her research areas concern data management and information systems. She leads international projects about Web services, documents, multimedia metadata and data security. Applications in IoT, CCTV, forensic and social interactions illustrate her contribution on social media ecosystems. http://scholar.google.com/citations?user=Dmv9sz8AAAAJ & hl=fr.