J Braz Comput Soc (2012) 18:283–298
DOI 10.1007/s13173-011-0055-1

WTI

An approach to enrich users' personomy using the recommendation of semantic tags

Sérgio R.P. da Silva · Marcelo R. Borth · Josiane M.P. Ferreira · Valéria D. Feltrim

Received: 15 June 2011 / Accepted: 25 December 2011 / Published online: 28 January 2012

Abstract Tagging-based systems are a popular and convenient way to organize information on the Web. Despite the alleged advantage of the free choice of words used to categorize Web resources in this kind of systems, it also brings some disadvantages due to the difficulty to remember freely chosen tags when users need to retrieve tagged resources. This paper presents a new approach to improve the quality of the categorizations performed by users in tagging-based systems by means of the recommendation of semantic tags. Our approach combines three sources of information for selecting the recommended tags: the Web resource been categorized, the tagging-based system folksonomy and the user personomy. By using these sources, we combine some features of the context of the categorization, the social opinion about the resource been categorized and the users' vocabulary preferences. The use of the Web resource helps to solve the *cold start problem*, and the recommendation of more contextualized and personalized tags helps to develop a better personomy for the user, which could relieve the users' cognitive effort when retrieving tagged resources.

Keywords Semantic tag recommendation · Tagging-based systems · Ontology

A previous version of this paper appeared at the III International Workshop on Web and Text Intelligence (WTI 2010).

S.R.P. da Silva (✉) · M.R. Borth · V.D. Feltrim
DIN – Departamento de Informática, UEM, Av. Colombo, 5790, Zona 7, 87020-900, Maringá, PR, Brazil
e-mail: sergio.r.dasilva@gmail.com

M.R. Borth
e-mail: marceloborth@gmail.com

V.D. Feltrim
e-mail: valeria.feltrim@gmail.com

J.M.P. Ferreira
CPGEI – Pós-graduação em Engenharia Elétrica e Informática Industrial, UTFPr, Av. Sete de Setembro, 3165, Rebouças, 80230-901, Curitiba, PR, Brazil
e-mail: josianempf@gmail.com

1 Introduction

Due to technical difficulties and high costs of implementation, it becomes impractical to have qualified experts evaluating and controlling all published content on the Web. This lack of schemes, or measures, to ensure the quality and organization of information results in a problem called *information overload* [22].

The tagging technique [36] represents an initiative to help in the organization and assignment of meaning to resources available on the Web. This technique adopts the principle that, if someone is able to publish a resource, she/he is also able to organize and assign meaning to other resources. Thus, it passes over to users the responsibility of organizing and labeling Web resources using tags, as they believe to be most convenient, and without any kind of control [33], eliminating the need for experts. As a result, this technique became an interesting alternative for an open and highly changeable environment as the Web, where one cannot maintain a hard scheme of control and organization.

In tagging-based systems (TBSs), the set of tags and tagged resources of a user comprises his/her *personomy*, which generally reflects the user's vocabulary, preferences, interests, and knowledge. Some systems allow users to share their personomies with each other, producing what is known as a *folksonomy* [36], which generally reflects the social view of a Web resource.

Although the freedom and dynamicity are positive characteristics of TBSs, they are also the main cause of problems

when users try to retrieve tagged resources [17]. This freedom allows for the introduction in the user's personomy of alternative forms of writing, synonymy, polysemy, different lexical forms, and different levels of accuracy in the vocabulary, what makes it difficult for users to remember the tags used in the categorization and complicates the retrieval of the tagged resources [14]. Most of these problems are directly related to not taking into account semantic [15, 43] and contextual information [29] during the categorization process.

As pointed by [1, 8, 13], the fact that cooccurrence is the only relation among tags is also a limiting factor when retrieving information because it is semantically weak. The absence of stronger semantics makes it difficult to solve ambiguities caused by synonymy and polysemy which are common to natural languages. In order to improve the tagging technique the creation of mechanisms has been proposed, in order to bring strong semantics to tags by recommending tags based on ontologies derived from the folksonomy of the TBS [2]. Although these recommendation bring valid suggestions, they rely only on the folksonomy as the source of information for creating semantics, leaving out contextual information provided by the Web resource itself.

Like some other authors [23, 40], we believe that it is necessary to consider as much information as possible about the resource and the users' preferences if we want to help the user to create and use better tags, which has high influence in the personomy tagspace convergence [30] and in making the resource retrieval process easier. Thus, aiming at improving the tag recommendation process, we propose the combination of three sources of information—the Web resource, the TBS's folksonomy and the user's personomy—to improve the quality of the recommended tags [5]. The combination of these sources is an attempt to take into account the following aspects of the tagging process: *contextual* (i.e., the Web pages of the resource being tagged); *social* (i.e., the TBS folksonomy view of the resource); and *personal* (i.e., the users' personomy which reflects is vocabulary preferences).

In order to combine these three sources of information and bring stronger semantics to the tags, we propose an algorithm that analyzes lightweight ontologies [10] generated from each source and extracts the most relevant concepts that are common to them to use as semantic tags to be recommended. To get the three ontologies, we show how the emergence of a lightweight ontology from tags belonging to the user personomy proposed by Basso et al. [3] can be adapted to emerge lightweight ontologies from tags of a TBS folksonomy and from terms of a Web resource. We also show how these three ontologies can be combined to generate tags for a TBS.

This paper is organized as follows: In Sect. 2, we discuss the relevance of the three sources of information available for a semantic tag recommender system and briefly review their use in current literature. In Sect. 3, we briefly discuss some alternatives for the emergence of ontologies from terms. In Sect. 4, each step of our semantic tag recommendation proposal is described. In Sect. 5, we analyze results derived from some experiments with real users. Finally, in Sect. 6, we show the conclusions and limitations of our proposal, along with suggestions for further investigation.

2 Sources of information for a tag recommender system

A tagging-based system usually has three sources of information available when it needs to recommend tags: (i) the Web resource, (ii) the folksonomy of the TBS, and (iii) the user's personomy. Each one of these has a particular importance in the recommendation process.

The Web resource is the main element of a categorization, and its content can be available in many forms, such as text, pictures, videos, flash animations, etc. In this work, we considered only textual content (i.e., any Web page with some text). One of the most important aspects of a Web resource content that is generally forgotten is the fact that it can express some features of the *context* of the resource categorization. For any Web resource, there are several factors influencing the context in which it could be used, but for a Web page the context can usually be determined by the way in which the vocabulary is employed by their author in order to expose the content to the reader. The task of getting the context of a Web page is not trivial, since the text of the page may present misspellings, synonymy, polysemy, bending terms, parts that do not refer to the content (e.g., header, footer, menu columns, advertisements), among others. Taking all these aspects into account, we decided to represent the context of a Web page by the set of keywords that is most representative of the characteristics and properties of the Web page, i.e., the most relevant terms contained in its term-vector[1] [26]. It is also possible to use Web resource metadata when it is available to represent a summary of the content [30], but as any summary it does not convey the richness of the content itself for the generation of the candidate tags.

A TBS folksonomy normally reflects the vocabulary that is common to the system's users [28], providing a *social view* of the categorized resources, which a single user could never have by themselves. Thus, using the TBS folksonomy data to recommend tags can always be a valid alternative, since the user is categorizing a resource that others in the community have already categorized and, therefore, there

[1] A term-vector is a vector of pairs of keywords and their frequency of occurrence in the text.

may be a common interest, which guarantees the utility of the folksonomy tags. Also, the idiosyncrasies present in a folksonomy may benefit the information retrieval process, as they represent alternative and interesting terms for the users (which makes the *serendipity* effect possible [38]).

The user's personomy can bring together a wide diversity of knowledge about the individual, since in a categorization users express, by the used tags, their knowledge, intentions and terminology preferences related to the content of each resource [32]. From the analysis of a user's personomy, it is possible to guide the tag recommendation to target the user's vocabulary, offering terms according to their preferences [34, 41].

Reviewing the recent literature about tag recommendation, we found that current approaches generally focuses on the system folksonomy as its main source of information [27]. A small number of approaches also uses the Web resource content to assist in the recommendation, among them we can quote Lu et al. [24], which analyzes the Web resource content and combines it with tags of similar resource for generating the candidate recommendations; Song et al. [37], which extract the document vector of the Web resource and applies statistical techniques over a bipartite graph of the words, tags, and resources to generate the candidate tags; and Heymann et al. [16], which uses the Web resource text, anchor text, and surround hosts for generating the recommending tags. Another small number of approaches employ information about the user together with the information of the folksonomy and Web resource metadata to select the tags that will be recommended, among them we can quote Lipczak [23], which uses the Web resource title and cooccurrence analysis to expand the set of candidate tags filtering it by using the user personomy to obtain the final recommendation; Musto et al. [30], which use the Web resource metadata to generate the candidates and the user information to personalize the final recommendation; and Tatu et al. [40], which use the Web resource to extract a combination of semantic and statistical characteristics to construct models of users and documents that are used to generate and select the recommended tags (they also employing the *WordNet*[2] to *standardize concepts*).

From the above review, we can observe that very few systems has tried to use the combination of the three sources of information together. Also, in spite of the fact that some systems had used a semantic approach, the level of semantic they explored is shallow. Most of them make use of complex statistical techniques to identify some level of semantic relations among the concepts and use these relations to inform the selection of candidates to the recommending tags. We propose to make use of these three sources together with an approach based on the recommendation of semantic tags that explores relations among the concepts. The three sources of information taken into account in this proposal can be combined in various ways, which we will discussed in the next section.

2.1 Possible scenarios for a tag recommendation process in a TBS

Taking into account the three sources of information discussed above, there are eight different scenarios that could happen in a tag recommendation process, as shown in Fig. 1. We can divide them in two groups: those that do not analyze the Web resource (1 through 4), and those that do it (5 through 8). Therefore, the only variables are whether or not there are sufficient amount of data in the user's personomy and the TBS' folksonomy to be used by the recommender system.

Let us first consider the scenarios where the Web resources are not used as a source of information, since this is the common case for the recommender systems in current TBSs.

Scenario 1: A user, without any information on his/her personomy, is trying to categorize a resource using a TBS without information on its folksonomy about that resource. This is the worst case scenario for a tag recommender system, and represents the situation confronted by a new system user categorizing a new resource. Since we do not have access to any of the three sources of information there is no way to generate recommendations. This is an instance of the *cold start problem* (i.e., the problem to generate recommendations for a resource without any source of information from where to take the terms to recommend) and happens in most of the recommender systems in current TBSs.

Scenario 2: A user, with information on his/her personomy, is trying to categorize a resource using a TBS without information on its folksonomy about that resource. This is the typical case of a user trying to categorize a new resource in a TBS. As a personomy has only information about resources already categorized by the user, there is no way to

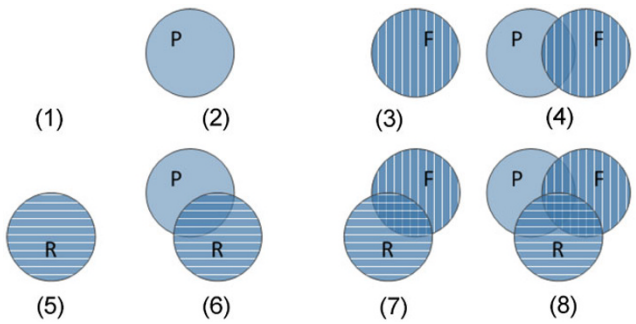

Fig. 1 Scenarios for a tag recommendation in a TBS (**R** stands for *resource*, **P** for *personomy* and **F** for *folksonomy*)

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

obtain information about the current resource being categorized and, therefore, we can only generate recommendations based on the global users' interests, given by the most used vocabulary in his/her personomy. This kind of data is most of the time of little use, generating recommendations of low quality. This is another instance of the *cold start problem* and also happens in most recommender systems in the current TBSs.

Scenario 3: A user, without information on his/her personomy, is trying to categorize a resource using a TBS with information on its folksonomy about that resource. This is another possible scenario for a new system user, but once the resource has already been evaluated and categorized by other system users, it is possible to generate recommendations from a social point of view. However, it will not be possible to personalize these recommendations to match the user's vocabulary, since there is no information in their personomy. The use of the folksonomy as the unique source of information is the common case for the majority of the recommender systems in current TBSs.

Scenario 4: A user, with information on his/her personomy, is trying to categorize a resource using a TBS with information on its folksonomy about that resource. This could be considered a desirable scenario for a recommender system, since it would be possible to use the folksonomy's social point of view to generate recommendations, giving priority to the terms most used by the community; and also to use the user's personomy to further personalize the recommendation data to match their vocabulary. The use of the folksonomy together with the personomy as source of information is the configuration used by the *Delicious* system.[3]

For the other four scenarios, we will assume that the Web resource was analyzed and a representation of some features of the context of the categorization is available.

Scenario 5: A user, without information on his/her personomy, is trying to categorize a resource using a TBS without information on its folksonomy about that resource. Unlike what happens in Scenario 1, once we have information from the Web resource it will be possible to generate recommendations from the extracted contextual data, avoiding the *cold start problem*.

Scenario 6: A user, with information on his/her personomy, is trying to categorize a resource using a TBS without information on its folksonomy about that resource. Again, contrary to what happens in Scenario 2, once we have information from the Web resource it will be possible to generate recommendations from the extracted contextual data, avoiding the *cold start problem*. In addition, in this scenario it would be possible to personalize the recommendations to the user's preferences, based on the information contained in their personomy.

Scenario 7: A user, without information on his/her personomy, is trying to categorize a resource using a TBS with information on its folksonomy about that resource. What makes this scenario different from Scenario 3 is that it will be possible to use the folksonomy data to filter the contextual data extracted from the Web resource, giving priority to the terms most used by the community.

Scenario 8: A user, with information on his/her personomy, is trying to categorize a resource using a TBS with information on its folksonomy about that resource. In this scenario, besides the social filter employed in the last scenario, we could also personalize the recommendation data to the user's preferences based on the vocabulary of his/her personomy. In this way, this could be considered the best case scenario for a recommender system.

Although it is possible to make recommendations without analyzing the Web resource, as discussed in Scenarios 2, 3, and 4, using it as a source of information could lead to better recommendations. This will take place because the recommended tags will come from terms present in the Web resource, which normally makes it easier for the use to remember. Even better, if the folksonomy information is available, it would be possible to further improve the tag's quality by applying a social filter to them, given priority to the tags most used by the community. In addition, if the personomy information is available, it would be possible to increase the memory of the recommended tags by taking into account the user's vocabulary preferences, which will certainly contribute to the retrieval of the resource.

One more aspect is worth mentioning. As discussed in Scenarios 5 and 6, the variations of the Scenarios 1 and 2, where the *cold start problem* normally happens in the recommender systems of the current TBSs, the use of the Web resource as a source of information allows the system to deliver recommendations to users, avoiding the *cold start problem*.

Taking all these aspects into consideration, we claim that using these three sources of information to recommend *semantic tags* to the users of a TBS can enrich the quality of the user's personomy. The adoption of semantic tags would avoid the use of mistaken terms and ambiguity and would improve the quality of the user's tag from the beginning. To obtain semantic tags we develop an algorithm that analyze and combine ontologies emerged from the three sources discussed, which will be presented in the following sections.

3 On the emergence of ontologies from tags and Web pages

There are basically two approaches to extract/emerge structure from a set of terms, such as Web pages, folksonomies,

[3]Delicious is the most famous TBS nowadays http://www.delicious.com/.

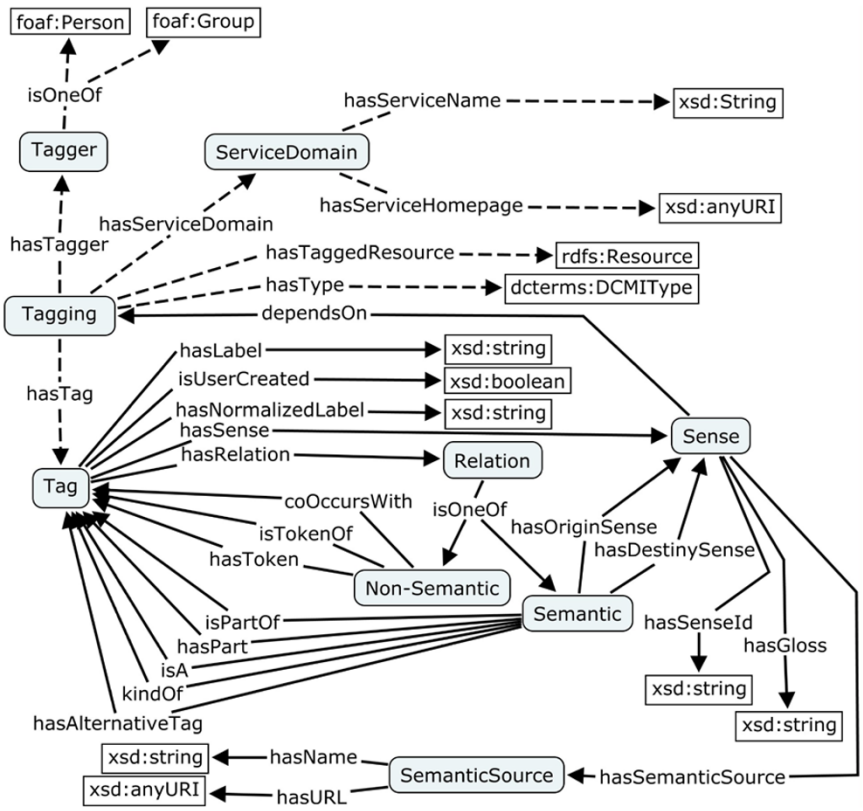

Fig. 2 Ontology model to represent the process of tagging, Basso et al. [3]

and personomies. Some proposals make a statistical analysis of the terms, based on cooccurrence, to identify clusters of related terms [4, 42]. Other proposals use external sources of data to establish the semantic relations among terms. In the context of TBSs, van Damme et al. [7] suggest possibilities to map different types of relations among tags in an ontology; Laniado et al. [21] propose a tool to organize the tags of a personomy into a hierarchy of concepts to be displayed in place of the *Delicious* tag list; Angeletou et al. [2] extract semantic relations from another data source in addition to the folksonomy data; and Basso et al. [3] emerge an ontology from the tags of a user's personomy using the *WordNet* as the external data source.

In this work, we adopt Basso's et al. [3] approach, which establishes an extended ontological model based on the works of Knerr [18] and Echarte [9]. This approach emerges a lightweight ontology [10] from the terms of the user's personomy, making it possible to give meaning to tags and to relate them to each other using a set of semantic relations, like "is a," "a kind of," "is part of," "has part," etc. To overcome the problem that on TBSs the only relation among tags is cooccurrence, they have done a *mashup* with another source of ontological data to establish the semantic relations among tags. Once tags are textual elements, they make use of the *WordNet* [11], which is a large lexical database of the English language, whose structure emerged from neurolinguistics theories of human lexical memory. *WordNet* is different from a common dictionary because it groups nouns, verbs, adjectives, and adverbs in cognitive synonyms sets called *synsets*, each one expressing a different concept. Therefore, after identifying concepts (*synsets*) in the *WordNet* corresponding to the user's tags, Basso et al. identify the set of relations among them that are all mapped to the ontological model of Fig. 2, in which the dashed relations represent the ontological model proposed by Knerr [18] and the continuous relations represent the extension Basso et al. proposed, expressing the knowledge of how the tagging process should be modeled with attributes and semantic relations among tags.

There is a possibility for a term not to be identified as any concept (*synset*) in the *WordNet*. In this case, the only relation considered will be the cooccurrence, which will not generate any semantic benefits. According to Laniado et al. [21], the probability of most popular tags belonging to the *WordNet* is high. However, an experiment made by Basso et al. [3] has shown that, on average, only 53% of a TBS users' tags are identified in the *WordNet*. In general, tags that are not identified can be: (i) misspelling errors; (ii) acronyms; and/or (iii) a recent concept that has not been registered in the lexical database yet. One way to solve (i) is to detect the inconsistencies in the user's personomy, by performing a cleaning of "mistaken" terms [6]. To solve (ii), we can use an acronyms dictionary and expand the acronyms to full words. To solve (iii), it is necessary to use a more dy-

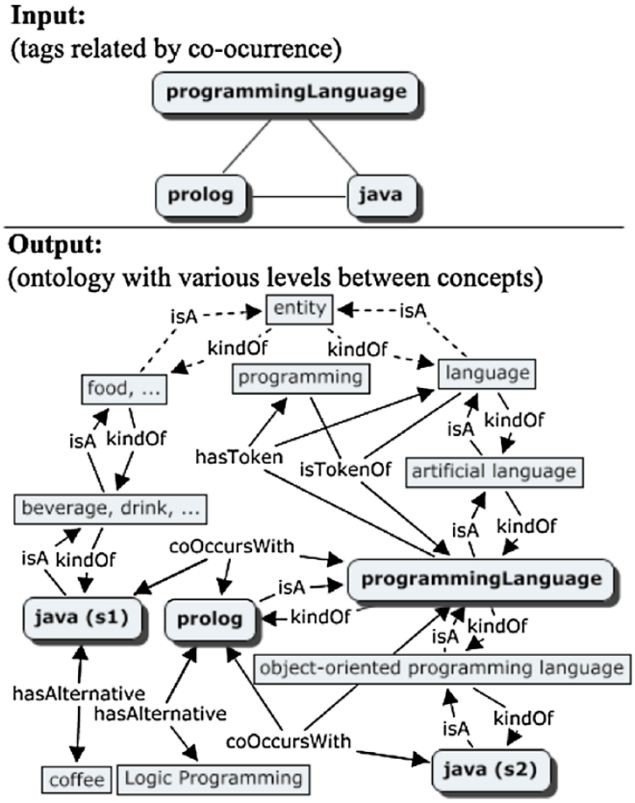

Fig. 3 An ontology generated by Basso's et al. algorithm

namic source of information than the *WordNet*, such as the *DBpedia*,[4] or some extensions of the *WordNet* such as *Stanford WordNet Project*.[5]

Basso's et al. approach to emerge lightweight ontologies from the users' personomy was developed for recognizing only nouns present in *WordNet*, resulting in ontologies like the one shown in Fig. 3. However, for recommending tags using the three sources of information, it may be of use to recognize other parts of speech such as verbs, adjectives, and adverbs as well.

To investigate the usage patterns of terms in Web pages, as well as its recognition rate in the *WordNet*, we conducted an experiment where Web pages of various topics such as medicine, engineering, computer science, etc., were used. We consulted 238 Web pages (randomly selected from a *Delicious* system folksonomy[6] data set of about 160 thousand bookmarks we have downloaded in February 2010) obtaining a total of 48,066 terms, from which we recognized in the *WordNet* 75.9%. For the recognized terms, the class that had higher recognition rate was nouns (82.5%), followed by adjectives (8.3%), verbs (7.9%), and adverbs (1.3%). Since for characterizing the Web page context verbs and adjectives are also important, and considering that there is a reasonable amount of verbs and adjectives in the content of Web pages, we adapted the algorithm proposed by Basso et al. [3] to emerge ontologies using nouns, adjectives, and verbs. As a result, we got an increase of approximately 17% in the recognition rate of terms. In order to increase even more the amount of recognized terms, it will be necessary to make a *mashup* with ontological sources other than the *WordNet*.

4 The process for selecting the recommended semantic tags

In this section, we describe the process we employed to generate semantic tags from the three sources of information we suggested to use. The process is composed of five steps, as shown in Fig. 4, and is presented as follows.

Step 1. Extracting the term-vector from a Web page: For the processing of the Web pages we adopted a statistical approach [26]. Therefore, we first extract its title[7] and body term-vectors applying a cleaning process to remove punctuation, extra symbols, and *stop words*.[8]

As in a document words can have variants such as plural, singular, words with suffixes, etc., which should not change its semantic representation, we use a conflation technique [20] to merge the words that have lexical variations to a single word. As a consequence, each pair of lexical variants of a term, which is identified with the same semantics, is represented by just one of them. In this way, the document will be represented by a single set of terms.

The most common conflation processes are stemming and lemmatization [26]. In this work, we choose to apply lemmatization as it reduces a word to its corresponding canonical form, keeping its morphological category. Therefore, after obtaining a term-vector, which is a tool that applies a linguistic process of lemmatization that employs the *WordNet* is used [12], reducing the lexical variances among terms that have the same meaning and joining their frequencies in the term-vector.

After the cleaning and lemmatization task, a similarity verification is performed among the terms of the Web page's body and title. This task aims to identify the words in the body that have similar meanings with the word in the title, so that we could increase its relevance (i.e., frequency). For this, we used the semantic similarity metric called *Lesk* adapted to *WordNet* [19], whose goal is to measure how

[4] http://dbpedia.org/About.

[5] http://ai.stanford.edu/~rion/swn/.

[6] *Delicious* will be taken as the baseline in this work due to its importance in the area of TBSs.

[7] The title's terms will be used to improve the relevance of the Web page's terms, by evaluating its similarities and increasing the Web page's term frequencies.

[8] A *stop word* in this context is a word that does not add semantics to a document, such as "a," "but," "for," etc.

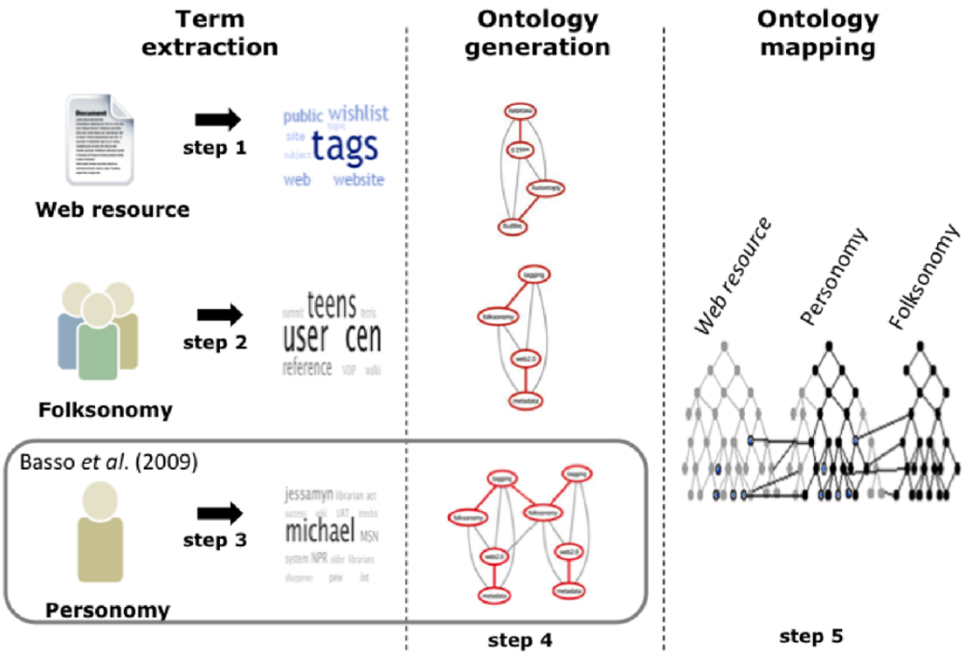

Fig. 4 The semantic tag recommendation process

strongly the meanings of two words are interconnected. The result of comparing pairs of terms using *Lesk* is a value representing the degree of similarity between them. The terms classified with "high similarity" (greater than 0.5) will have their frequencies increased and for the ones with "low similarity" (smaller than or equal to 0.5) the frequency value is not updated.

This step has a great significance in our approach because by extracting the most representative terms of the Web page, we can ensure that the ontological model will be well-formed and will make a valid context representation.

Step 2. Retrieving terms from the folksonomy: There are two approaches to retrieving the folksonomy term-vector referent to a Web resource. We could use the API of the TBS requesting the terms used to categorize a given URL, or we could develop a *screen scrapper* for the TBS Web page that presents the terms used to categorize a given URL. However, using the last alternative one can usually get only a limited number of tags for a resource (e.g., accessing *Delicious* to extract the terms of a Web resource returns only the 30 most used tags by the community). For both cases, it is possible that the system's users have not yet categorized the Web resource the user is categorizing, and consequently, there will be no data available for emerging the ontology from the folksonomy. In this work, we adopt the first approach using the API of the *Delicious* system.

Another important observation about this step is that we are considering a TBS that uses the approach proposed in this work, making it unnecessary to further process the folksonomy term-vector. If this is not the case, it will be necessary to submit the folksonomy term-vector to the same preprocessing step which was applied to the Web page terms.

Differently from the processing of the Web page term-vector, the frequency of the terms will be given by the frequency of the tags employed by the user community. Also, it will be necessary to equalize the frequencies of the Web page and folksonomy term-vectors, since their range are almost always very different due to the fact that a Web page has a nearly stable content and a folksonomy changes with its use.

Step 3. Retrieving the term-vector for the personomy: The recovery of the user term-vector from the TBS is carried out in the same way as the retrieval of the folksonomy term-vector, but now the frequency of the terms is accounted only for the current user. Since we will be adopting the approach of prioritizing the terms preferred by the user, we do not equalize this term-vector. Instead, we make a mapping from the Web page and folksonomy ontologies generated in the next step over the personomy ontologies, as explained in Step 5.

Step 4. Generating the three ontologies: One serious problem when generating an ontology from a Web page, the folksonomy data or the personomy data is its very large number of terms. Therefore, to maintain the performance of an ontology-based system manageable it is essential to find a way to reduce the number of terms used in its generation without losing to much quality.

Looking for a solution for this problem, we analyzed the relationship among the distribution of terms in the three sources of information we suggested the use of. First, we observed that Zipf's law [31] shows that the frequency distribution of terms in a document (e.g., a Web page) follows a power law. Second, Lux et al. [25] shows that the distribution of terms in a folksonomy also follows a power law.

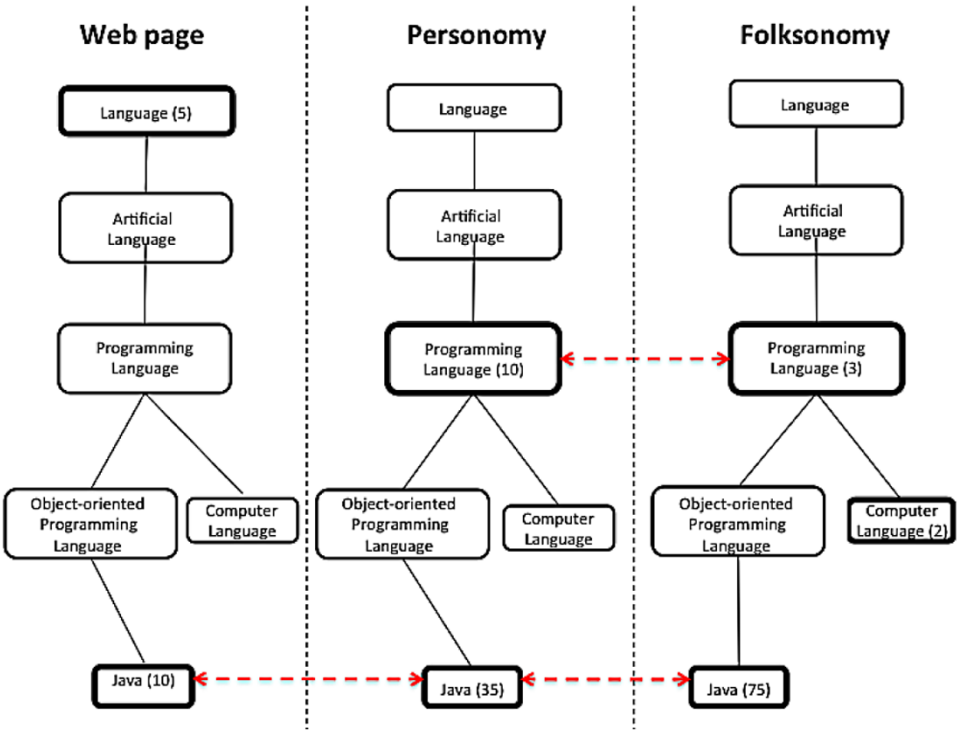

Fig. 5 Mapping between Web page and folksonomy ontologies over the personomy ontology (*dashed lines* indicate a concept match and the numbers indicates the equalized score of the concepts)

Third, Zhang et al. [44] shows that the distribution of terms in a personomy also follows a power law. In addition, we noticed that this law is also associated to the Pareto's principle [31], which states that 80% of the causes are responsible for 20% of the effects, while 20% of the causes are responsible for 80% of the effects.

Thus, putting all these observations together, we decided to apply Pareto's principle by reducing the amount of terms of each resource (the Web page, the folksonomy and the personomy) by 80%, expecting to have a reduction in the quality of our ontologies in approximately 20%. This seems to be a reasonable decision, since we will be using the 20% of the terms that are the most significant ones. Therefore, the gain in computational time for generating the ontologies will be certainly greater than the loss in its quality.

Once the set of terms are defined, an ontology is generated for each source of information (the Web page, the folksonomy and the personomy) based on our extension of the Basso et al. [3] proposal.

Step 5: Mapping the terms among the three ontologies: To get the final ranking of the concepts, we use a procedure that determines a mapping between the ontologies to reinforce the concepts that have a matching. Our understanding of ontology mapping is defined by [39] as: "Given two ontologies A and B, mapping one ontology with another means that for each concept (node) in the ontology A we try to find the corresponding concept (node), one which has the same or similar semantics, in the ontology B and vice versa." For the mapping between the ontologies of our approach, we create a procedure that corresponds to the previous definition.

The matching process in this case is much simpler than a common ontology matching process because the ontologies are generated from the same source, i.e., they are generated based on the *WordNet* structure. Thus, they all have the same root entity, and most of the internal paths are similar, which avoids the costly calculation necessary to determine if two concept from the different roots are the same. Thus, to identify concepts that are similar in two ontologies, our algorithm compares the equality between the *synsets*. In this task, all *synsets* in the Web page and the folksonomy ontologies are compared, one by one, with each *synset* of the personomy ontology. When two *synsets* are equal, the concepts are mapped between the ontologies, i.e., for each concept of an ontology that is found in the personomy ontology, a mapping relation between them is created, as shown in Fig. 5.

For each matching, the concepts in both ontologies are reinforced. Consequently, since we are prioritizing concepts in the personomy ontology, a concept in this ontology can be reinforced twice, once for matching the Web ontology and once for matching the folksonomy ontology. Thus, at the time of recommendation, a concept that is present in the three ontologies will have two mapping relations, and consequently, its relevance will be greater than the others that have just one, or none. The result of this process will be the identification of the most related concepts that tend to be the most relevant for the user. This will also eliminate,

Listing 1 Tag selection algorithm

```
def tagsSelector(TBS, user, URL, numRecommendations):

# Step 1 – Retrieve the Web page term–vector
titleTV = extractTermVector(URL, "title")
for term in titleTV:
        lematization(lexicalCleaning(term))

bodyTV = extractTermVector(URL, "body")
for term in bodyTV:
        lematization(lexicalCleaning(term))
        term = SimilarityDetection(term, titleTV)

# Step 2 – Retrieve the folksonomy term–vector
folkTV = retrieveTerms(TBS, "folksonomy", "")
equalize(bodyTV, folkTV)

# Step 3 – Retrieve the user personomy term–vector
userTV = retrieveTerms(TBS, "personomy", user, URL)

# Step 4 – Generate the ontologies
pageOnto = generateOntology(paretoReduction(pageTV))
folkOnto = generateOntology(paretoReduction(folkTV))
userOnto = generateOntology(paretoReduction(userTV))

# Step 5 – Map the ontologies
for userConcept in userOnto:
        userConceptSS = getSynset(userConcept)
        for pageConcept in pageOnto:
                pageConceptSS = getSynset(pageConcept)
                if userConceptSynset == pageConceptSynset:
                        mapSynsets(userConceptSS, pageConceptSS)
        for folkConcept in folkOnto:
                folkConceptSS = getSynset(folkConcept)
                if userConceptSynset == folkConceptSynset:
                        mapSynsets(userConceptSS, folkConceptSS)

# Step 6 – Recommend tags with highest priority
return getStrongerTags(pageConceptSS, folkConceptSS,
      userConceptSS, numRecommendations)
```

for instance, concepts present in the ontologies that are not related to the user's interest.

Although our algorithm gives priority to adjusting the recommendation to the user's interests, it does not favor terms of the personomy only. Once a concept of the Web page or the folksonomy ontology has a higher frequency than those contained in the personomy, even if it has little relational mappings, it may be recommended. Consequently, the concepts that are not in the personomy but are frequently used in the text or by the community (i.e., concepts that are interesting to add to the user's personal vocabulary) will also be recommended.

At the end of the process, the recommendation will show the concepts that have the greatest relevance based on their priority (i.e., the number of mappings between ontologies) and then by the frequency of each concept. The algorithm in Listing 1 summarizes the process discussed above.

5 Evaluation

This section presents three experiments employed to analyze the quality of the semantic tags recommended by the proposed approach. The first experiment automatically assesses the relevance of the semantic tags recommended by this approach comparing them with the tags used in the folksonomy of a current TBS. The second experiment assesses the acceptance by real users of the semantic tags recommended by this approach considering only the most relevant concepts selected from the sources used. The third experiment assesses the acceptance by real users of the semantic tags recommended by this approach considering the most relevant concepts selected from sources used and their hypernyn concepts.

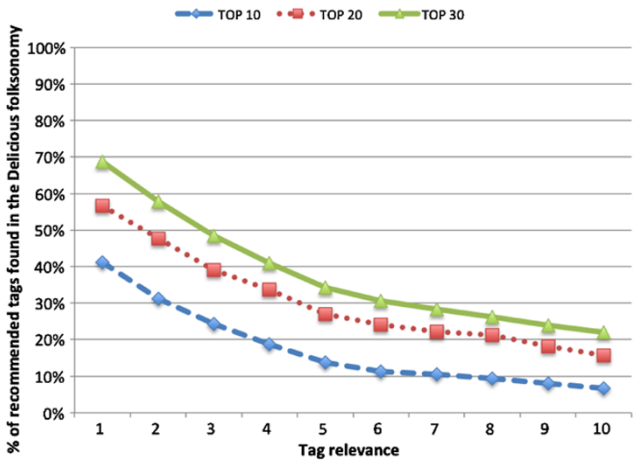

Fig. 6 Recommended semantic tags present in the *Delicious* folksonomy

5.1 Experiment 1: Comparison of the recommendation of semantic tags with the *Delicious* system folksonomy

In this experiment, we automatically analyzed 1,304 bookmarks, randomly selected from a *Delicious* data set containing 3,401 users with 2.613.446 categorizations and 8.208.402 tags (1.804.282 unique URLs and 305.948 unique tags), aiming at accessing the *relevance of the recommendations*. Between the selected URLs there were web pages that contain a great variety of content (e.g., text, images, video, and a combinations of them), making it a very good approximation of a real situation. We should notice that, due to this test set characteristics, the percentage of terms found in each web page has a great variation, which has an impact on the final quality of the suggested terms. It was our decision to use this test set to evaluate the behavior of our approach in a real situation. Thus, for each URL, we analyzed the percentage of the first 10 recommended tags produced by our approach that were found between the TOP 10, 20, and 30 tags of the *Delicious* folksonomy.

As we observe in Fig. 6, for the first three recommended tags (the average number of tags used in the *Delicious* system) the percentage of tags found in the TOP 10 tags of the

Delicious system lies between 40–24%, but this range increases to 68–48% in the TOP 30, as expected. Although these numbers are not high, they may be considered acceptable, as there must be a discount due to the characteristic of the test set, which is not ideal for our approach. However, to ensure that our approach recommends tags that are well accepted by users, we had to perform experiments with real users.

5.2 Experiment 2: Evaluation of the recommendation of semantic tags with real users

The purpose of this experiment was to assess the *acceptance by real users* of the semantic tags recommended by the proposed approach. It is important to justify why we choose to do tests with real users instead of using a standard collection, such as the ones used in ECML PKDD discovery challenge. We take this decision due to the fact that the choice of a tag by a user in a tagging process is strongly biased by the tags s/he receives as recommendation from the system. Thus, if we compare the tags generated by our approach for a given resource with the ones selected by users' of another system, we will be comparing the other system bias with ours. But this does not imply that this would be the user's choice if s/he had originally be submitted directly to our bias, i.e., we will not be allowing our recommender to influence the users' choice, which is exactly what we want to measure.

In this experiment, the three sources of information (the Web resource—**R**; the *Delicious* system folksonomy—**F**; and the users personomy—**P**)[9] were considered in the analysis. Thus, we were able to analyze all but the first scenario discussed in Sect. 2.1, since there is not enough information in this scenario to generate recommendations. For Scenarios 2, 3, and 5, where we have only one source of information we recommended syntactic tags based on the frequency of the terms. For Scenarios 4, 6, 7, and 8, where we have more than one source of information, we recommended semantic tags based on the combination of the ontologies of the sources through our algorithm, as discussed in Sect. 4.

In the experiment, ten users evaluated each scenario. Each participant received 70 predefined URLs randomly selected for categorization, being 10 for each scenario. This URLs are equally divided in two knowledge areas: computer science and general knowledge. The scenarios are also randomly presented so that there was no possibility for the users to identify which scenario they were evaluating. For each URL, 10 tags were recommended. The participants had the freedom to select the recommended tags and/or to inform additional tags. The developed system can distinguish

Fig. 7 Presentation of a tag recommendation considering its degree of relevance

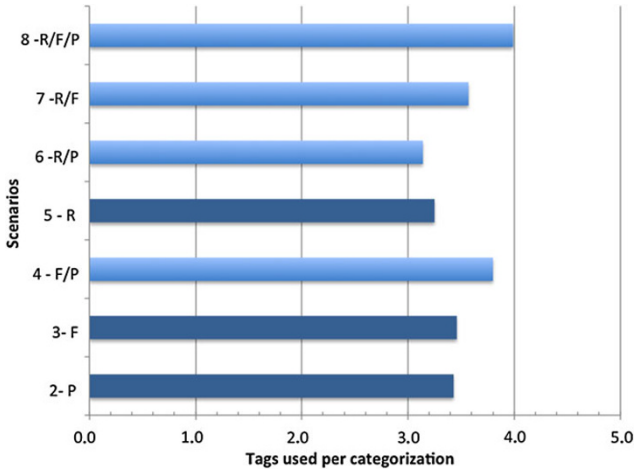

Fig. 8 Average number of tags used in each scenario

which tags were recommended, which were selected by the user, and which ones were added. We understand that each tag has a degree of relevance for the Web resource, so they were presented in the recommendation interface in descending order of their relevance from left to right, as shown in Fig. 7.

Analyzing the average number of tags used in each categorization, shown in Fig. 8, we observed the following facts. First, among the syntactic approaches (Scenarios 2, 3, and 5) there is no significant difference in the number of tags used. Second, among the semantic approaches that use the Web resource (Scenarios 6, 7, and 8) there is an increase in the number of tags used as we consider the personal and social aspects, and both, which indicates an improve in the recommendation quality. Third, Scenario 4 presents a good use of the recommended tags. One explanation for this result is that the use of semantic tags in this scenario helps to avoid the possible ambiguities that normally happens given a more informed choice for the user. Last, Scenario 8 improves the average number of tags used in a categorization by 15%, when compared to the baseline Scenario 3, which is the most common scenario for TBSs.

Once a user has the freedom to use any tag on a categorization, they can select recommended tags or enter additional tags directly. When users inform additional tags in a categorization, it means that the recommendation was not interesting enough to meet their needs. For this reason, we also evaluate the average number of categorizations in which users reported additional tags and observe the follow-

[9] We will use the notation **R/F/P** to represent the sources available in each scenario.

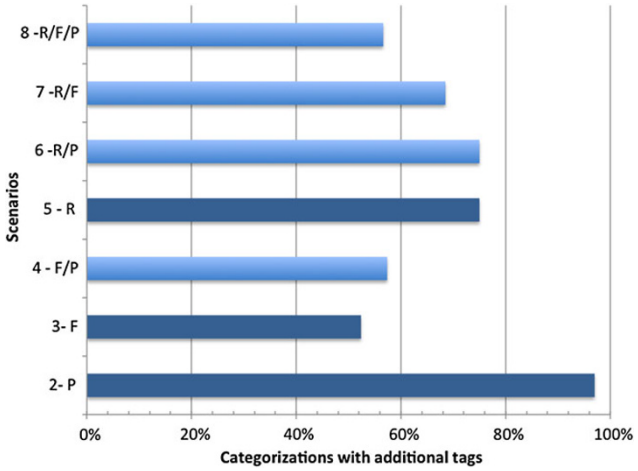

Fig. 9 Categorizations that used additional tags recommended by users

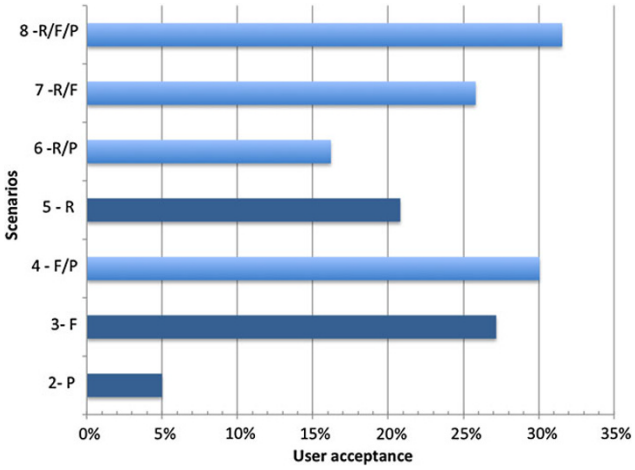

Fig. 10 User's acceptance of the recommended tags in each scenario

ing facts. First, analyzing Fig. 9, we can see that, as was predicted in Sect. 2.1, Scenario 2 shows the worst result (97% of the tags was added by the users), as the personomy does not have enough information about the resource to generate good recommendations. Second, the use of syntactic tags based only in terms extracted from the Web resource text ordered by its frequency does not help reducing the number of additional tags used (Scenario 5), even when we change to the use of semantic tags with the consideration of the personal aspect the results are not satisfactory (Scenario 6). A reasonable explanation for this is that most of the time the highest frequency terms in the text of the Web resource are very general and users prefer to add their own tags instead of a general term. The addition of the personal aspect does not help because the personomy has no information about the categorized resource. Third, the effect of the social and personal aspects were observed again between the scenarios that use a semantic approach with a decrease in the number of additional tags used. Last, the Scenarios 4 and 8, which use the combinations of more than one source an semantic tags have good results but no better than the syntactic tags of the folksonomy alone.

Two issues should be considered in this experiment to explain part of the results' behavior. First, we observe that great part of the tags added by the users were compound words, and most of them were composed of tags recommended to the user. This contributes to increase the number of the additional tags used and to decrease the user acceptance of the recommended tags. We did not recommend compound words in this experiment due to the complexity to choose among the possible term combinations. This shall be done as a future work. Second, we suppose in our approach that the personomy and folksonomy used were generated by the proposed process. This means that they are "clean" in the sense of having been through a normalization process, which improves their recognition rate in the *WordNet*. Unfortunately, it was not possible to accomplish this at the time of this experiment, which also influence the acceptance of the recommended tags. This also shall be done as a future work.

Analyzing the overall user acceptance for each scenario (i.e., the acceptance of the recommended tags over all categorizations), as shown in Fig. 10, we observed the following facts. First, for the scenarios that employs a syntactic approach (Scenarios 2, 3, and 5) the resulting behavior was as expected. When taken separately, the personomy should generate worst quality tags comparing to the Web resource, which should generate worst quality tags when compared to the folksonomy. Second, the social aspect has greater power for filtering the Web resource in the semantic approach than the personal aspect when taken separately, with a significant improvement in the user acceptance ($\alpha = 0.05$ and $t = 7.3689$ between Scenarios 6 and 7).[10]

Third, when the three sources of information are considered within the semantic approach, they produce a significant improvement in the user acceptance ($\alpha = 0.05$ and $t = 2.6954$ between Scenarios 7 and 8). But there is no significant difference if we leave the Web resource out when using the semantic approach ($\alpha = 0.05$ and $t = 0.6757$ between Scenarios 4 and 8). Last, the use of the three sources of information results in a significant improvement in the user acceptance ($\alpha = 0.05$ and $t = 1.9865$) over the (Scenario 3) the folksonomy syntactic approach (our baseline), with an increase of 16% in the overall user acceptance.

[10]We have employed a one tail t-test for this experiment because the populations are unknown but we can consider their standard deviation equal, given that the tested subjects are the same for the two approaches.

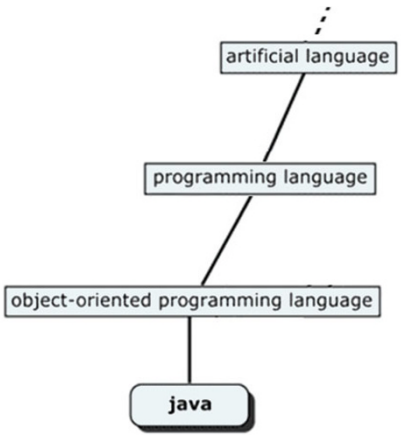

Fig. 11 Partial hierarchical structure of the tag "Java" based on relations of the *WordNet*

5.3 Experiment 3: Evaluation of the recommendation of hypernyn tags with real users

In this third experiment, we followed the same methodology as in the second experiment, namely we used the same URLs and the same evaluation metrics, but a smaller number of users was employed per scenario in the analysis due to resource restrictions. So, the results of this experiment should be considered as preliminary. The main difference between the two experiments is that in this one the focus was on assessing the recommendation of semantic tags based on the hierarchical structure of the sources ontology. Thus, we considered only Scenarios 8 (R/F/P), 7 (R/F), and 4 (F/P), which use semantic tags and have shown better results in the second experiment.

In the previous experiment, we observed that, for any of the evaluated scenarios, there is expressive user acceptance for the first four more relevant tags of a recommendation (which will be called *main tags*). Thus, this experiment employs only the first four main tags proposed by our algorithm plus its hypernyn concepts in a total of 8 recommended tags. The hypernyn tag is the first more generic concept of the main tag in the ontology hierarchy. For example, in Fig. 11, if the tag "java" is recommended, the "object-oriented programming language" is also recommended because it is the direct hypernyn concept of term "java".

After collecting the data, we observed that, contrary to experiment 2, Scenario 7 obtained a result 25% better than Scenario 8 for the number of tags used per categorization, as shown in Fig. 12. This results should be taken with caution, but one possible explanation for this behavior is that the use of the personal aspect could have a negative effect if the user's personomy has not been created from semantic concepts from the beginning. This happened due to the user's freedom for choosing the tags to use in the categorization. Such freedom allows a user to choose words that are idiosyncratic and that do not generalize well to fit in a more formal ontology. We need further investigation to confirm this possibility.

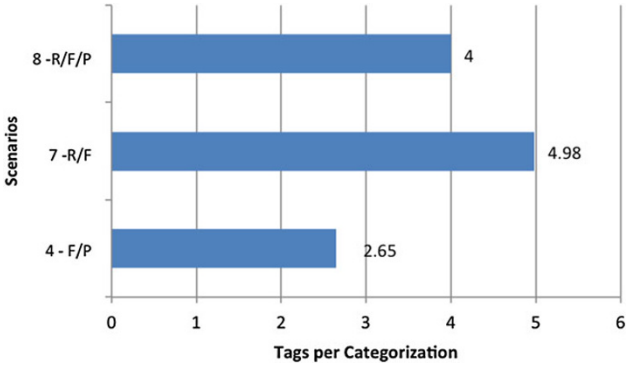

Fig. 12 Average number of tags used in each scenario considering hypernyn concepts

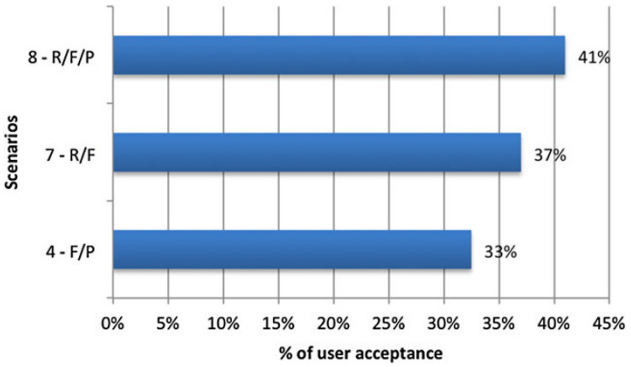

Fig. 13 User's acceptance of recommended tags for all classes of tags in each scenario

Aiming at understanding the contribution of the hypernyn tags, we analyzed the overall user acceptance of the recommended tags. Comparing the results of Fig. 13 with Fig. 10, we noticed that there is an order change between Scenarios 4 and 7 with regarding which one is best, but the difference has no statistical significance. However, in general, the user acceptance has similar values, considering the limited number of tests. This is a good evidence that the quality of the tags has not changed. But, it is necessary to make more experiments to guarantee that it is worth spending time to get the hypernyn tags, unless we have another use for them such as to use them in a visualization interface for the user's personomy.

Analyzing the user's acceptance of the individual class of tags, as shown in Fig. 14, we observed that the pattern of Fig. 13 repeats itself in the three scenarios. However, we observed that the acceptance of the hypernyn tags is very low, which shows that the use of this approach is not interesting due to its high cost of generation for the tags. We also observe that the effect of personalization in this kind of tag is low. We believe this happens due to the fact that in most cases the concepts in the higher hierarchical level of an on-

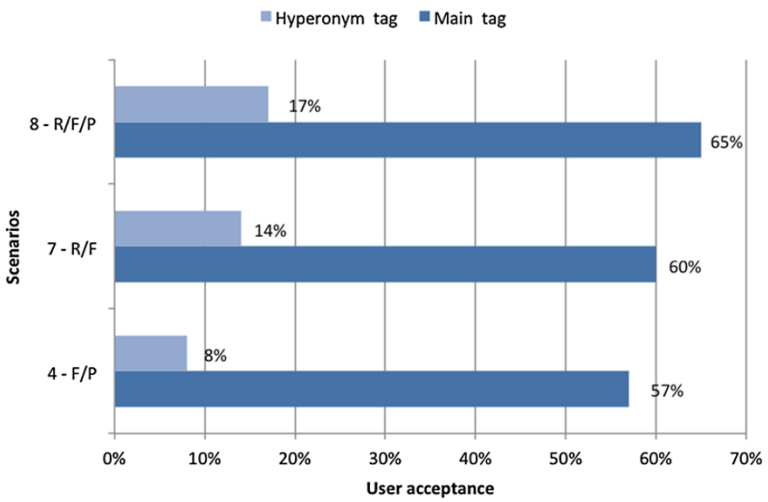

Fig. 14 User's acceptance of the recommended tags for each class of tag in each scenario

tology are very formal, not reflecting the users vocabulary. This formality also reflects in the hypernyn concepts recommended for each main tag, thereby reducing the percentage of the acceptance of hypernyn tags.

6 Conclusions and final thoughts

In this paper, we aimed at addressing the problem of improving the quality of tags used in TBSs. To reach this objective, we proposed an approach to the recommendation of semantic tag for TBS that takes into account the contextual, social, and personal aspects of a Web resource. For this, we suggested using the content of the Web resource, the TBS folksonomy data and the users' personomy data as sources of information for the recommender system.

To obtain the semantic tags for the recommendation, we proposed a process in which, after preprocessing the Web resources, the folksonomy and the personomy terms, we extract a lightweight ontology from them and create a ranking based on the terms frequency and a mapping between the Web resource and folksonomy ontologies over the personomy ontology. In this way, we prioritize the personomy terms but also consider new frequent terms that could appear in the Web resource or the folksonomy. The emergence of the lightweight ontologies was done by extension of the Basso's et al. [3] approach to emerge ontologies from the Web pages and the TBS folksonomy. In this extension, we included adjectives and verbs, which give us a better identification rate for terms in the *WordNet*. We also show how to get a reasonable trade-off between the candidate tag quality and the performance of the ontology generation process using the Pareto's principle over the set of concepts. To evaluate our proposal, we carried out three experiments, one automatic and two with real users.

The first experiment was designed to access the relevance of the semantic tags generated by our approach. It automatically compared the first 10 tags generated by our approach with the top 10, 20, and 30 most relevant tags from the *Delicious* system for a resource. As a result, it shows that although the raw results were not high, they are acceptable, given the test set used where the web pages had content of any kind (i.e., text, images, video, etc.), which certainly affects the quality of the tags generated by our approach.

However, as this experiment was automatic, it did not take into account the bias induced in the user when s/he saw the set of recommended tag during the tagging process. Thus, we decided to do another experiment with real users to access the user acceptance of the recommended tags.

In this second experiment, we used 10 users with varying experience and 7 scenarios (from 2 to 8) discussed in Sect. 2. For the scenarios that use only one source of information, we adopted a syntactic approach based on the frequency of the terms and for the others we used a semantic approach based on our proposal. As its main results, we found that:

- Concerning the number of tags used per categorization, there is no difference in the number of tags used among the scenarios using the syntactic approach. On the other hand, there is a significant difference in the number of tags used among the scenarios using the semantic approach, the best one being the one that uses the three sources of information, with an increase of 15% in the number of tags used in relation to baseline scenario. Among the scenarios that use two sources of information the best was the one using the folksonomy and the personomy as its sources, which was also significantly better than any scenario that used the syntactic approach.
- Concerning the number of additional tags (i.e., tags freely chosen by the users) used in the categorizations, among the scenarios using the syntactic approach, the best one was the one using the folksonomy as its source, followed by the Web resource and them the personomy, a result that was already expected from the discussion in Sect. 2.

Among the scenarios using the semantic approach, the behavior was similar to the number of tags used, but this time there was no significant difference between the scenario using the folksonomy and personomy as its sources and the one using the three sources, also these scenarios were not better than the syntactic approach using only the folksonomy as its source. This shows that the social aspect are essential to the recommendation process in both approaches.

– Concerning the user acceptance of the recommended tags, among the scenarios using the syntactic approach, the behavior was similar to the number of additional tags used, which was also expected from the discussion in Sect. 2. The same has occurred with the scenarios that used the semantic approach, but in this case there was not statistical difference between the scenarios using the three sources and the one using the folksonomy and the personomy. The best scenario in this experiment was the one using the three sources of information with a significant difference from the baseline scenario (Scenario 3) with an increase of 15% in the user acceptance.

In general, this experiment gave us a better understanding of the influence of each aspect of the resource over the recommendation process. We could see that the social aspect shows greater power to filter the Web resource than the personal aspect. Also, that the combination of both produces the best results, as we had already predicted in the discussion in Sect. 2.

The last experiment was designed to address the effect of exploring the ontology structure in the recommendation, for this we used the same procedures of the second experiment, with a more limited number of users in the test, and selected only 8 tags (4 normal tags and 4 hypernyns). Even with such restrictions, we observed that the use of hypernyn concepts brings no direct advantage to the recommendation process when compared with the cost of producing the candidate tags.

All in all, we can say that the proposed approach definitely produces significant results concerning the improvement of the quality of the recommended tags as measured by its user acceptance. Moreover, this approach does not suffer from the *cold start problem*, which is common to most current TBSs, since it was able to use the terms extracted from the Web resources to guide its first recommendations.

We can also visualize other benefits for this approach to tag recommendation as, for example, the exploration of the structure provided by its ontology for navigation on and visualization of the users' personomy. In addition, unlike most current TBSs which make users remember the exact lexical form of the tag in order to retrieve a resource, the semantic relationships present in their ontologies enable searches using alternative terms than the selected tag. Besides that, it also allows for making searches using more generic or more specific concepts for the desired term by walking in the ontology structure, which possibly contributes to reduce the cognitive effort on the user when retrieving a categorized resource.

Despite the benefits we acknowledge for our approach it needs further work before being used in a production system. At the moment, the time necessary for processing the sources is high for on-line Web systems. Thus, further work is necessary in its implementation architecture to improve its performance. Moreover, as observed in the analysis of experiment two, it is necessary to attack the problem of generating compound word tags, since this represents a great number of the additional tags used. Also, it is necessary to aggregate a tag management system, like the *TagManager* system [35], to our system to allow for the maintenance of the tag quality, which will improve its recognition by the *WordNet*, or any other anchor ontology from the Web, resulting in a better ranking for the candidate tags. Last but not least, we intend to address the quality of our recommendation from the point of view of the retrieval task. This will require new experiments with users retrieving the resources they have previously categorized using the proposed approach. Furthermore, it would be interesting to explore the effect that the user interface has over a semantic recommendation.

We believe that the availability of a semantic tag recommender system can enhance the categorization process, so that users can create and use more meaningful tags in their categorizations, which as a consequence, could stabilize the users personomy faster and improve the whole tagging process by making the retrieval of tagged resources easier.

Acknowledgements Our thanks to CAPES for the scholarship granted to Marcelo Rafael Borth during his Master degree from which this paper is a result.

References

1. Adrian B, Sauermann L, Roth-Berghofer T (2007) Contag: a semantic tag recommendation system. In: Proceedings of the I—Semantics' 07. JUCS, Graz, Austria, pp 297–304
2. Angeletou S, Sabou M, Motta E (2008) Semantically enriching folksonomies with flor. In: Proceedings of the collective intelligence and the semantic web workshop (CISWeb 2008) at ESWC'08. CEUR-workshop proceedings, vol 351, Tenerife, Spain, pp 65–79
3. Basso CAM, Ferreira JMP, da Silva SRP (2009) An unsupervised approach for the emergence of ontologies from personomies in tagging-based systems. In: Proceedings of the 2009 Latin American web congress, LA-WEB'09. Universidad Autónoma de Yucatán (UADY), IEEE Computer Society, Mérida, YUC, México, pp 193–200. doi:10.1109/LA-WEB.2009.35
4. Begelman G, Keller P, Smadja F (2006) Automated tag clustering: improving search and exploration in the tag space. In: Proceedings of the WWW collaborative web tagging workshop, Edinburgh, Scotland, pp 22–26

5. Borth MR, da Silva SRP, Ferreira JMP, Feltrim VD (2010) An approach to enrich users' personomy using the recommendation of semantic tags. In: Proceedings of the III international workshop on web and text intelligence (III WTI)
6. Côgo FR, da Silva SRP (2008) Uma proposta de organização do vocabulário de tags dos usuários de sistemas baseados em folksonomia. In: Proceedings of the VIII Brazilian symposium on human factors in computing systems, IHC'08, Sociedade Brasileira de Computação, Porto Alegre, RS, Brazil, pp 288–291. URL http://portal.acm.org/citation.cfm?id=1497470.1497509
7. van Damme C, Hepp M, Siorpaes K (2007) Folksontology: an integrated approach for turning folksonomies into ontologies. In: Proceedings of the bridging the gap between semantic web and web 2.0 workshop, ESWC'07, Deri Innsbruck, Innsbruck, Austria, pp 57–70
8. Dill S, Eiron N, Gibson D, Gruhl D, Guha R, Jhingran A, Kanungo T, Rajagopalan S, Tomkins A, Tomlin A, Zien JY (2003) Semtag and seeker: bootstrapping the semantic web via automated semantic annotation. In: Proceedings of the 12th international conference on world wide web, WWW '03. ACM, New York, pp 178–186. doi:10.1145/775152.775178
9. Echarte F, Astrain JJ, Córdoba A, Villadangos JE (2007) Ontology of folksonomy: a new modelling method. In: Proceedings of the semantic authoring, annotation and knowledge markup (SAAKM'07). CEUR-workshop proceedings, vol 289. British Columbia, Canada
10. Giunchiglia FMM, Zaihrayeu I (2007) ncoding classifications into lightweight ontologies. In: Journal on data semantics VIII. LNCS, vol 4380. Springer, Berlin, pp 57–81
11. Fellbaum C (1998) WordNet: an electronic lexical database, 1st edn. MIT Press, Cambridge
12. Finlayson MA (2009) MIT Java WordNet interface. URL http://projects.csail.mit.edu/jwi/
13. Fountopoulos GI (2007) Richtags: a social semantic tagging system. PhD thesis, School of Electronics and Computer Science at University of Southampton, Southampton, UK. URL http://eprints.ecs.soton.ac.uk/15109/
14. Golder S, Huberman BA (2006) Usage patterns of collaborative tagging systems. J Inf Sci 32(2):198–208. URL http://www.hpl.hp.com/research/scl/papers/tags/
15. Guy M, Tonkin E (2006) Folksonomies: tidying up tags? D-Lib Magazine 12(1). URL http://www.dlib.org/dlib/january06/guy/01guy.html
16. Heymann P, Ramage D, Garcia-Molina H (2008) Social tag prediction. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'08. ACM, New York, pp 531–538. http://doi.acm.org/10.1145/1390334.1390425
17. Hotho A, Jäschke R, Schmitz C, Stumme G (2006) Information retrieval in folksonomies: search and ranking. In: Sure Y, Domingue J (eds) The semantic web: research and applications. LNCS, vol 4011. Springer, Budva, pp 411–426. doi:10.1007/11762256
18. Knerr T (2006) Tagging ontology—towards a common ontology for folksonomies. URL http://code.google.com/p/tagont/
19. Korenius T, Laurikkala J, Kalervo J, Juhola M (2004) Stemming and lemmatization in the clustering of Finnish text documents. In: Proceedings of the 13th ACM international conference on information and knowledge management, CIKM '04. ACM, New York, pp 625–633. doi:10.1145/1031171.1031285
20. Kosinov S (2001) Evaluation of n-grams conflation approach in text-based information retrieval. In: Proceedings of 8th international symposium on string processing and information retrieval—SPIRE'01, IEEE Computer Society, Laguna de San Rafael, Chile, pp 136–142
21. Laniado D, Eynard D, Colombetti M (2007) A semantic tool to support navigation in a folksonomy. In: Proceedings of the 18th conference on hypertext and hypermedia, HT'07. ACM, New York, pp 153–154. doi:10.1145/1286240.1286282
22. Levy DM (2005) To grow in wisdom: Vannevar bush, information overload, and the life of leisure. In: Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries, JCDL'05. ACM, New York, pp 281–286. doi:10.1145/1065385.1065450
23. Lipczak M (2008) Tag recommendation for folksonomies oriented towards individual users. In: Proceedings of ECML PKDD discovery challenge, Bled, Slovenia, pp 84–95. URL http://www.kde.cs.uni-kassel.de/ws/rsdc08/pdf/10.pdf
24. Lu YT, Yu SI, Chang TC, Hsu JYj (2009) A content-based method to enhance tag recommendation. In: Proceedings of the 21st international joint conference on artificial intelligence, IJCAI'09. Morgan Kaufmann, San Francisco, pp 2064–2069. http://dl.acm.org/citation.cfm?id=1661445.1661775
25. Lux M, Granitzer M, Kern R (2007) Aspects of broad folksonomies. In: Proceedings of the 18th international conference on database and expert systems applications. IEEE Computer Society, Washington, pp 283–287. doi:10.1109/DEXA.2007.40
26. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval, 1st edn. Cambridge University Press, Cambridge
27. Marinho LB, Nanopoulos A, Schmidt-Thieme L, Jäschke R, Hotho A, Stumme G, Symeonidis P (2011) Social tagging recommender systems. In: Ricci F, Rokach L, Shapira B, Kantor PB (eds) Recommender systems handbook. Springer, Berlin, pp 615–644. doi:10.1007/978-0-387-85820-3_19
28. Mathes A (2004) Folksonomies—cooperative classification and communication through shared metadata. Tech. rep., Graduate School of Library and Information Science at University of Illinois Urbana-Champaign. URL http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html
29. Melucci M (2008) A basis for information retrieval in context. ACM Trans Inf Sys 26(14):1–41. doi:10.1145/1361684.1361687
30. Musto C, Narducci F, De Gemmis M, Lops P, Semeraro G (2009) A tag recommender system exploiting user and community behavior. In: Workshop on recommender systems & the social web—ACM RecSys'09. ACM, New York
31. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Contemp Phys 46(5):323–351. doi:10.1080/00107510500052444
32. Rashid AM, Albert I, Cosley D, Lam SK, McNee SM, Konstan JA, Riedl J (2002) Getting to know you: learning new user preferences in recommender systems. In: Proceedings of the 7th international conference on intelligent user interfaces, IUI'02. ACM, New York, pp 127–134. http://doi.acm.org/10.1145/502716.502737
33. Riddle P (2005) Tags: What are they good for? Tech. rep., School of Information at University of Texas at Austin. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.125.8221
34. Shepitsen A, Gemmell J, Mobasher B, Burke R (2008) Personalized recommendation in social tagging systems using hierarchical clustering. In: Proceedings of the ACM conference on recommender systems, RecSys'08. ACM, New York, pp 259–266. doi:10.1145/1454008.1454048
35. da Silva JV (2009) Gerenciamento do vocabulário do usuário em sistemas baseados em tagging. Master's thesis, Departamento de Informática at Universidade Estadual de Maringá, Maringá, PR, Brazil
36. Smith G (2008) Tagging: people-powered metadata for the social web, 1st edn. New Riders Press, Berkeley
37. Song Y, Zhuang Z, Li H, Zhao Q, Li J, Lee WC, Giles CL (2008) Real-time automatic tag recommendation. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08. ACM, New York, pp 515–522. http://doi.acm.org/10.1145/1390334.1390423

38. Sturtz DN (2004) Communal categorization: the folksonomy. URL http://www.davidsturtz.com/drexel/622/sturtz-folksonomy.pdf
39. Su X (2002) A text categorization perspective for ontology mapping. Position paper. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.107.4082
40. Tatu M, Srikanth M, D'Silva T (2008) Tag recommendations using bookmark content. In: Proceedings of ECML PKDD discovery challenge, Bled, Slovenia, pp 96–107
41. Wetzker R, Said A, Zimmermann C (2009) Understanding the user: personomy translation for tag recommendation. In: Eisterlehner F, Hotho A, Jäschke R (eds) ECML PKDD discovery challenge (DC09). CEUR workshop proceedings, vol 497. Bled, Slovenia, pp 275–284
42. Wu H, Zubair M, Maly K (2006) Harvesting social knowledge from folksonomies. In: Proceedings of the seventeenth conference on hypertext and hypermedia, HYPERTEXT '06. ACM, New York, pp 111–114. doi:10.1145/1149941.1149962
43. Xu Z, Fu Y, Mao J, Su D (2006) Towards the semantic web: collaborative tag suggestions. In: Proceedings of the collaborative web tagging workshop, WWW'06, Edinburgh, Scotland
44. Zhang D, Mao R, Li W (2009) The recurrence dynamics of social tagging. In: Proceedings of the 18th international conference on world wide web, WWW'09. ACM, Madrid, pp 1205–1206. doi:10.1145/1526709.1526930