An Approach to Enrich Users’ Personomy Using Semantic Recommendation of Tags

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

Universidade Estadual de Maringá
Av. Colombo, 5790, Zona 7, Maringá – PR – Brazil
marcelo@borth.com.br, {sergio.r.dasilva, josianempf, valeria.feltrim}@gmail.com

Abstract. Tagging-based systems are a popular and convenient way to organize information on the Web. However, the free vocabulary used to categorize Web objects brings some disadvantages to this type of system, especially when users need to retrieve the tagged object. This paper presents an approach to semantics recommendation of tags for tagging-based systems that aims at improving the quality of the categorizations performed by the users. Our approach combines three different resources for recommending tags: the Web page been categorized, the tagging-based system folksonomy and the user’s personomy. By combining these resources, we seek to reduce the problems in the user’s vocabulary, recommending representative tags that may relieve the users’ cognitive effort when retrieving tagged objects.

Keywords: Tag Recommendation, Ontology, Tagging-based Systems.

1 Introduction

Due to technical difficulties and the high costs of implementation, it becomes impractical to have qualified experts controlling and evaluating all content that is published on the Web. This lack of schemes, or measures, to ensure the quality and organization of information results in a problem called information overload [1].

The tagging technique [2] represents an initiative to help the organization process and assigning meaning to content available on the Web. This technique adopts the principle that, if someone is publishing a content, s/he is able to organize and assign meaning to content. Thus, it delivers to users the responsibility for organizing and labeling the Web content with tags, as they believe to be most convenient, and without any kind of control [3], 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. Furthermore, this technique has a low cognitive cost to users.

Although the freedom and dynamicity of tagging to assign meaning to objects are good characteristics, they are also the main cause of problems when users try to retrieve tagged information. Errors or alternative forms of writing, polysemy,
synonymy, different lexical forms, different levels of accuracy and the types of tag-resource association hamper the retrieval of the tagged objects. These problems are related to the lack of semantic information in the categorization process [4, 5, 6].

As pointed by [7, 8, 9], the fact of the unique relation among tags to be co-occurrence is a limiting factor when retrieving information because its semantic is weak. We also believe that the availability of tools that bring semantics to tags can advance the tagging technique, improving the tag recommendation process, so that users can use tags of better quality in their categorizations, which would improve the retrieval of tagged objects.

In tagging-based systems (TBSs), the set of tags and tagged objects of a user comprise his/her personomy. The personomy reflects the user’s vocabulary, preferences, interests and knowledge. Some systems allow their users to share their personomies with other users. The effect of this sharing is known as a folksonomy [2].

In this paper, we show how the emergence of an ontology from tags belonging to the user personomy, proposed by Basso et al. [10], can be adapted to emerge ontologies from terms of a Web page and from tags of a folksonomy in a TBS, aiming at improving the tag recommendation process. We propose the combination of three ontologies to improve the quality of the tags that are recommended for the user when s/he is categorizing an object.

This paper is organized as follows: In Section 2, there is a brief description of some alternatives for the emergence of ontologies from terms. In Section 3, each step of our semantic recommendation of tag proposal is described. In Section 4, we present and discuss some initial results. Finally, in Section 5, we show the conclusions and limitations of our proposal, along with suggestions for further investigations.

2 Emerging Structure from Tags

There are two different approaches to extract/emerging structure from terms of a TBS. Some proposals make a statistical analysis of the user’s tag-space, based on co-occurrence, to identify clusters of related tags. Other proposals use external sources of data to establish the semantic relations among tags. In this context, Damme et al. [11] suggest possibilities to map different types of relations among tags in an ontology; Laniado et al. [12] propose a tool to organize the tags of a personomy into a hierarchy of concepts to be displayed in place of the Delicious tags list; Angeletou et al. [13] extract semantic relations from another data source in addition to the folksonomy data; and Basso et al. [10] emerge an ontology from the tags of a user’s personomy.

2.1 An Approach for the Generation of an Ontology from Tags

In this work we adopt Basso’s et al. [10] approach, which establishes an extended ontological model based on the works of Knerr [14] and Echarte [15]. This approach emerges ontologies from the user’s personomy, making it possible to give meaning to tags and to relate them to each other using semantic relations.
An Approach to Enrich Users’ Personomy Using Semantic Recommendation of Tags

Because on TBSs the only relation among tags is co-occurrence, to establish semantic relations among tags it is necessary to undertake a mashup with another source of data. Once tags are textual elements, Basso et al. suggest consulting the WordNet [16], which is a large lexical database of the English language whose structure emerged from neurolinguistics theories of human lexical memory. WordNet differentiates itself 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 WordNet corresponding to the user’s tags, Basso et al. identifies the set of relations among them, which are all mapped to the ontological model of Figure 1.

In Figure 1, the gray relations represent the ontological model proposed by Knerr [14] and the black relations represent the extension proposed by Basso et al. [10], 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 to be not identified as any concept (synset) in the WordNet. In this case, the only relation considered will be the co-occurrence, not generating semantic benefits. According to Laniado et al. [12], the probability of most popular tags to belong to the WordNet is high. However, an experiment [10] has shown that, on average, only 53% of 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 still been registered in the lexical database. To solve (i) and (ii) it is necessary to detect inconsistencies in the personomy, by performing a cleaning of erroneous terms. To solve (iii), it is necessary to use a more
dynamic source of information than WordNet. We believe that the use of semantic recommendation of tags at the time of categorization, aiming at avoiding misspelling errors and ambiguity, could improve the recognition rate of tags in the WordNet.

2.2 Extending the Ontology Emergence Approach

Basso’s et al. approach was developed for recognizing only nouns present in WordNet. However, for recommending tags, it may be of use to recognize other parts of speech such as 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, like medicine, engineering, computer science, etc., were used and at different levels of knowledge. Using WordNet, we have identified 76% of the terms found in the Web pages.

For the identified terms, the class that had the higher recognition rate was noun (82.5%) followed by adjective (8.3%), verb (7.9%) and adverb (1.3%). Considering that, there is a reasonable amount of verbs, adjectives and adverbs in the content of Web pages, and that these may be representative to the page context, we adapted the algorithm proposed by [10] to emerge ontologies by using them. As a result, we have gotten an increase in the recognition rate of terms of approximately 17%. In order to increase even more the amount of recognized terms, it will be necessary to make a mashup with other information sources.

These results have shown that it is interesting to extract terms from Web pages for all four grammatical classes present in WordNet in order to represent the page context. In the next section, it is shown how the use of semantic recommendation of tags can help in the emergence of ontologies from a personomy/folksonomy and in the information retrieval.

3 Improving the Recommendation of Tags

If we want to increase the precision of object retrieval in TBSs, it is necessary to improve the quality of the user’s personomy, since it is the source of the keywords used in the retrieval phase.

Hence, we propose the use of semantic recommendation of tag based on the emergence of ontologies from a blend of three different information sources: (i) the user’s personomy, (ii) the terms of the folksonomy, and (iii) the terms of the Web page. The reason why we chose to use these three types of resources is that each one has a specific goal. The user’s personomy can identify his/her vocabulary preferences. The terms of the folksonomy provide an overview of what the community of users is thinking about the object, i.e., it gives a collaborative outlook. Finally, the content of the Web page contains the current context of the object being categorized, i.e., it is the real object and reason of the categorization.

One benefit of our semantic approach to recommendation is that it allows describing and organizing objects using concepts, making the categorization free from ambiguity. In addition, unlike most current TBSs that force the users to remember the
exact lexical form of the tag in order to retrieve an object, the semantic relations
present in an ontology enable searches using alternative terms than the selected tag.
Besides, it also allows making searches using more generic or more specific concepts
for the desired term, what contributes to reduce the cognitive effort on the user when
retrieving a categorized object. Another characteristic of this approach is that it
provides recommendations based on the context of a Web object and not based only
on its URL. The steps of our proposal are illustrated in Figure 2 and detailed below.

**Step 1. Extracting Terms from Web Pages**: To work with the context of Web
pages we must extract the keywords that may be the most relevant to the object. In
our work, we adopted a statistical approach [17] in which the processing of the Web
pages is done by extracting the terms of its content and the terms of its title, which are
very helpful for disambiguating concepts. Then, a cleaning process is done to remove
punctuation and symbols, and then the *stop words* (i.e., words that do not help in
identifying the context of a document, for example, ‘a’, ‘but’, ‘for’, etc.) are removed.

Words in a document can have variants like plural, singular, words added of
postfix, etc. The various forms of inflecting a word should not change its semantic
representation, since this would cause a problem when indexing the terms for the
document. To alleviate this problem, we use a conflation technique, whose goal is to
unite the words that have lexical variations to a single word. By consequence, the
document is represented by a single set of terms, and each pair of lexical variants of a
term, which is identified with the same semantics, is represented by just one of them,
allowing joining their frequencies. The most common conflation processes are
stemming and lemmatization [17]. In this work, we choose to apply lemmatization, as
it reduces a word to its corresponding canonical form, keeping its morphological
category. Therefore, after the cleaning process, a linguistic process of lemmatization
that employs the *WordNet* is used in attempt to prevent lexical variances among terms
having the same meaning.

Then it is performed a similarity verification among the terms in the page’s body
and in its title, aiming at identifying the words of the text that have similar meanings
with the title’s words of the Web page and increasing its relevance (frequency). For
this, we used the semantic similarity metric called *Lesk* [18], adapted to *WordNet*, which goal is to measure how strong the meanings of two words are interconnected.

The result of comparing pairs of terms (one term to another in the Web page and in the title) using *Lesk* is a value representing the degree of similarity between them. The terms classified with “high similar” will have their frequencies increased and for the ones having “low similarity” the frequency value are not updated.

This step has a great importance for our approach because extracting terms that might be the most representative of the Web page context can ensure that the ontological model will be well formed.

**Step 2. Retrieving Terms from the Folksonomy:** In order to retrieve tags from the folksonomy that are relevant to a Web page we use the *TagManager* [19], a system whose goal is to group and manage personomy information for various TBSs owned by a user, and apply the following steps to generate the ontology. Another approach to recover terms from a folksonomy is to develop an application that makes the *screen scrapper* of the terms used in a categorization to a particular Web object in a TBS like *Delicious*. However, using this alternative (i.e., accessing a *Delicious* URL to extract the terms) one can only get the 30 most used tags by the user community for an object. For both cases, it is possible that the system’s users have not yet categorized the object and, consequently, there will be no data available for emerging the ontology from the folksonomy.

**Step 3. Retrieving Terms from the Personomy:** The recovery of the terms from the personomy is performed in the same way that the recovery of terms from the folksonomy was carried out, but the difference now is that the categorizations are retrieved only for one user.

**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 that its number of terms is very large. Thus, to maintain the performance of an ontology-based system manageable it is essential to find a way to reduce its number of terms. For this task, Zipf’s law [21] ensures that the frequency distribution of terms in a document follows a power law curve, which is associated to the Pareto principle [21], 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. The variation used in our study corresponds to the 80-20 rule, where 80% of tags are used in 20% of categorizations, while 20% of tags are used in the other 80% of categorizations. Thus, we assume that the 20% of the most frequent terms are more representative than others.

By applying Pareto’s principle, we decide to reduce 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 ontology 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. 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 resource type (the personomy, the folksonomy and the Web page) based on our extension of the *Basso et al.* [10] proposal.
Step 5: Mapping among the Three Ontologies: To obtain the final ranking of the concepts, we use a process that determines a mapping between the ontologies to reinforce the concepts that have a matching. Our understanding of ontology mapping is defined by [22] 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 process that corresponds to the previous settings.

To identify the similar concepts between two ontologies, our algorithm compares the equality between synsets. In this task, all synsets present in the ontologies of the Web page and the folksonomy are compared, one by one, with each synset of the ontology of the personomy. When two synsets are equal, they are mapped between the ontologies, i.e., for each concept of an ontology that is found in the ontology of the personomy it is created a mapping relation between them. For each matching, both concepts are reinforced. In this manner, a concept of the personomy ontology can be reinforced twice, once for each matching. Thus, at the time of recommendation, the concepts that are present in the three ontologies will have two mapping relations and consequently, its relevance will be greater than the others concepts that have just one, and so on. Therefore, the result of this process will identify the most related concepts that tend to be the most relevant for the user characteristics. This will also eliminate, for instance, concepts present in the ontologies that are not related to the user interest.

Although our proposal adjusts the recommendation to the user’s interests it does not reserve only to terms of personomy. 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 relation mapping, it may be recommended. Consequently, the concepts that are not present in the personomy but are frequently used (i.e. are interesting to be added to the user’s personal space) will be recommended.

At the end of the process, the recommendation will show the concepts that have the greatest relevance based on their priority (number of mappings between ontologies) and then by the frequency of each concept.

4 Evaluation

Aiming at evaluating our approach for the semantic recommendation of tags, we performed an experiment based on an automatic analysis of about 400 bookmarks of various topics and different levels of knowledge. The experiment addressed two types of scenarios, one in which the user does not have tags in his/her personomy (i.e. a novice user) and other in which the user has a large number of tags (i.e. an expert user).

To verify the recommended tags to the bookmarks, we use as a comparison the Delicious tags database, which has a section that displays a history of the 30 previously most used tags by others users for each bookmark. Therefore, all recommended tags by our approach were compared with the tags presented within the Delicious folksonomy section. We did not take into account bookmarks that had no
tags in the folksonomy, because we would not have parameters to perform the comparison automatically.

In Figure 5 we can see that, despite the difference between the two types of users examined, the recommendation showed an effective result. In the graph, the horizontal axis displays the amount of recommended tag, while the vertical axis shows the percentage of effectiveness or accuracy of the tag recommendation (in comparison with the Delicious folksonomy). Based on this evaluation, we could recognize that the recommendation between 5 and 7 results presents the best tradeoff concerning the users’ cognitive effort to choose among different recommendations and the efficiency of the recommendation.

![Graph showing the evaluation of semantic recommendation of tags](image)

**Fig. 3.** Evaluation of the semantic recommendation of tags in comparison with the tags in the Delicious folksonomy.

This result shows that our approach can generate good recommendations at the time of a Web page categorization, since the data in the Delicious folksonomy are representative information generated by a user community. However, we still cannot say anything about the quality of the recommended tags for the retrieval task. In order to do so, we need to perform another experiment with actual users to measure the precision of the recovery.

## 5 Conclusions and Final Thoughts

In this paper we propose an approach to semantic recommendation of tag for TBS based on the analysis of three types of distinct information sources: the content of a Web page, the folksonomy terms and the users’ personomy data. We expand Basso’s *et al.* [10] approach to include the adjective, verbs and adverbs grammatical classes, what give us a better identification rate for terms in the WordNet. We also show how to get a reasonable tradeoff between the users’ cognitive effort to choose among recommended tags and the performance of the ontology generation process using the Pareto’s principle over the list of terms.
An Approach to Enrich Users’ Personomy Using Semantic Recommendation of Tags

As an evaluation result, we show that this approach produces good results if compared with the well-established system Delicious. In addition, it has a great advantage of not suffering from the “cold start problem”, which is common to most of current TBSs, since it could use the terms extracted from the Web pages to guide the first recommendations.

Currently, we are directing our efforts in doing experiments with real users to measure the actual effectiveness of our approach in terms of cognitive effort and quality of the recommendation. Subsequently, we want to assess the accuracy of our recommendation for the retrieval task. This will require new experiments with the same users retrieving the resources they have previously categorized. Furthermore, it would be interesting to explore the effect that the user interface has over a semantic recommendation.

The proposed approach to semantic recommendation of tags will be used in the future by the TagManager system [19], aiming at strengthening the user personomy and improve the information retrieval. We also believe that this approach can be used by other systems with the goal of improving the process of tagging and the information retrieval.

6 Acknowledgements

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

7 References