An Unsupervised Approach for the Emergence of Ontologies from Persononomies in Tagging-Based Systems

Carlos Alberto M. Basso, Josiane M. P. Ferreira, Sérgio Roberto P. da Silva
Departamento de Informática
Universidade Estadual de Maringá
Maringá – PR – Brazil
mail.carlosalberto@gmail.com, jmpinhei@din.uem.br, sergio.r.dasilva@gmail.com

Abstract—Tagging-based systems are becoming a widely used tool as they are considered simple and quick to categorize resources. However, due to the free vocabulary used for tagging, and also because of its plane structure, there are some drawbacks inherent to this kind of system, mainly when users do the information retrieval. This paper presents a proposal for the emergence of ontologies from tagging-based systems, seeking to reduce information overload, thus minimizing the cognitive effort of users when processing information retrieval.

Keywords: Ontology emergence, personomies, tagging

I. INTRODUCTION

Recently there has been a huge increase in the volume of information that comes to us through the means of communication, including the World Wide Web. As a consequence, users have been gradually overloaded with information, which demands great cognitive effort to discern relevant things from insignificant ones. It is essential to develop new techniques to improve the organization of information on the web. One way to achieve this is by making use of taxonomies that organize information in a hierarchical form, allowing information to be recovered more efficiently. Unfortunately, taxonomies sometimes make the information classification process expensive, especially when the piece of information the user wants to classify fits nowhere or when it fits in more than one category. Furthermore, when taxonomies are too specialized, they become confusing for the users both in classifying and in retrieving the information [1].

Another way of significantly enhancing the organization of information and its subsequent recovery is the use of ontologies, which represent knowledge structured through concepts, instances, attributes and relations that are modeled as a graph or network [2]. Unlike taxonomies, which allow only the definition of the generalization/specialization relation, ontologies enable the definition of various types of relations such as generalization/specialization, whole/part, cause/effect, and associations, among others [1]. These well structured semantic relations make the information retrieval process less difficult for users, since there are other ways of accessing the same information. The problem of using ontologies for classification and recovery of information is that it brings a high cognitive effort for users. This occurs primarily because both ontologies and taxonomies are difficult to build and maintain [2], and also because the coherence of the ontology must be kept after the classification of new information.

If the task of organizing information in the form of an ontology demands a high cognitive effort, to organize information by making use of tags is quite a simple task. Nowadays, several systems allow users to use tags freely chosen to organize information on the web. Such process is commonly known as tagging [3]. The set of categorizations and tags of a user is called his personomy, and a set of personomies socially available for a community of users characterizes a folksonomy [3]. The value of this tag attribution process is derived from the fact that users make use of their own vocabulary, thus, adding an explicit meaning to the resource, which emerges from their understanding about the information or about the object being categorized [3]. This makes the information categorization a low-cost task.

The use of tagging for organizing information is very interesting due to the easiness in which users categorize a resource. But, on the other hand, the process of information retrieval is hindered by other reasons. Firstly because personomies are affected by organizational problems and ambiguity, despite the fact that well developed controlled vocabularies [4] and hierarchical schemes [3] can effectively improve such drawback. Regarding the problems of ambiguity, it should be emphasized that the current systems do not seem to have been designed to deal with the use of acronyms, homonyms, synonyms and compound words in tags. The second problem comes from the increase in the number of categorizations, because the majority of the tagging-based systems makes use of tag clouds or lists of tags to start the information retrieval process (which may be chaotic when there is a large number of tags) demanding a great user’s cognitive effort. Moreover, there are not many other possibilities for visualizing or accessing information, because the only
relation between tags is the co-occurrence (the relation of two tags that are used together in a categorization) [5], which is semantically weak and generates just a flat structure among tags.

Since the problem of information retrieval in tagging-based systems is related to remembering the term associated with a resource, a task in which human beings have cognitive difficulties [6], a possibility to overcome this problem would be to use the advantages of the ontologies as structures for information retrieval, and to use the advantages of the tagging process to organize information. The central idea of our proposal is to define a process to generate an ontology from a user’s personomy, with various relations (hierarchical and non-hierarchical), which may turn the information retrieval process into an easy task from a user’s point of view. This ontology may be used in the future to: (i) improve information retrieval using semantic relations between tags to enable searches by concepts instead of using the exact form the tag was created; (ii) suggest terms for categorization, based on the user’s interest, since the ontology obtained may express the user’s areas of interest; and (iii) improve the user’s “navigation” capability in the tagspace, through structures such as hierarchies or graphs, which may be easily acquired from ontology semantic relations. The main goal of all these advantages is to reduce the user’s cognitive effort when retrieving the desired information.

This paper is organized as follows: In Section 2, we describe some related studies and the way they differ from the present investigation. In Section 3 each step of our proposal is detailed aiming at the emergence of ontologies from personomies including an ontological model for representing the semantic amongst tags, and a proposal for a concept disambiguation. We finish this section evaluating the algorithms proposed. Finally, in Section 4, we show the conclusions, the limitations of our proposal and suggestions for further investigation.

II. Emerging structure from personomies

There are basically two kinds of approaches to extract/emerge structure from tagging-based systems. Some proposals use statistical analysis of the tagspace, based on tag co-occurrence, in order to identify clusters of related tags. Systems with this resource may find related sets of tags which identify different contexts. Their difficulty lies in the fact that due to the lack of hierarchic levels in the tagspace there is no criterion to organize the tags; consequently, just a small set can be shown to the user. Moreover, there is no explicit connection between the tags and their meanings, or semantic relations interconnecting them [7], and this makes the user’s orientation difficult since such resources may direct users in tagspace navigation, helping them in the retrieval of information. Belgeman et al. [8] approach and the Flickr (a tagging-based system for the categorization of photos — http://flickr.com/) are example systems which use clusters to show context-related tags.

Another approach is to follow proposals that use other data sources, in addition to the tagging data, to establish semantic relations among tags. Damme et al. [5] suggested possibilities to map different types of relations between tags in an ontology, however, these possibilities have not been implemented. Another similar study, by Laniado et al. [9], proposed a tool to organize tags of a personomy into a hierarchy of concepts to be shown instead of the Delicious’ (a tagging-based system that allows bookmark categorizations — http://del.icio.us/) lists of tags. Thus, other data source (in addition to a personomy) are necessary to set the hierarchical relations amongst terms, and also a process for tag sense disambiguation to show only relations of user’s interest. Our proposal differs from the one by Laniado et al. [9] since their focus is just on obtaining hierarchical relations important to navigation, and ours is on the emergence of a structure with various kinds of relations between the user’s tags. Finally, there is the proposal by Angeletou et al. [7], which also extract some semantic relations from other data source beyond the folksonomies itself. However, such proposal aims at obtaining semantic web entities in other ontologies available on the web to associate them with the tags. Conversely, our aim is to build a structure with various levels starting from the tags and to take advantage of this ontology in terms of searches, data visualizations and in other aspects.

III. A proposal for the evolution of ontologies from personomies

For the evolution of an ontology from a personomy it is firstly necessary to understand the tagging process. There are some proposals such as Knerr’s [10] and Echarte’s [2] which define ontological models for this task. These models differ a little from each other but both have, at least, the three basic pivots of tagging-based systems, which are: the user who performs the categorization, the resource and the tags used in the categorization. In addition to the three basic attributes for the representation of a tagging, the ontological model can also represent the date/time when the resource was categorized, the system used to categorize the resource (e.g. Delicious, Flickr, etc.), the type of the categorized resource (e.g. a photo, a bookmark, etc.), if the categorization can be publicly viewed, etc. In our proposal it was necessary to extend these models to make possible to give meaning to the tags and relate them to others using semantic relations, as illustrated in Figure 1.
function is to obtain, group and manage information of systems, a system called categorization data, from the various tagging-based process from the personomy. To get the user’s fixed, the next step is to look for the structure building systems. Only relation between tags in traditional tagging-based among tags in relation to co-occurrence, which is the attributes allows people to map stronger semantics occurrence between two tags in a tag with its tokens); broader); hasAlternativeTag (which relates a tag with another if they have the same sense); hasPart/isPartOf (which relates two tags in which one is part of another); isA/kindOf (which relates a narrower tag with a broader); hasToken/isTokenOf (which relates a grouped tag with its tokens); coOccurWith (which maps the co-occurrence between two tags in a categorization). These attributes allows people to map stronger semantics among tags in relation to co-occurrence, which is the only relation between tags in traditional tagging-based systems.

Once the ontological tagging model to be used is fixed, the next step is to look for the structure building process from the personomy. To get the user’s categorization data, from the various tagging-based systems, a system called TagManager [4] was used. Its function is to obtain, group and manage information of the user’s personomies in the various tagging-based systems that he/she uses.

After the data getting from a user’s personomy it is necessary to enrich the tagspace with semantically stronger relations than the co-occurrence. To achieve these semantically stronger relations it is necessary to mashup with other data source that allows for obtaining such information. As tags are textual elements, a possibility for getting this kind of information is from the Wordnet, which is a large lexical database of the English language, which was developed based on psycholinguistic theories concerning the organization of lexis in the human memory [11]. The Wordnet differs from an ordinary dictionary because it groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms, called synsets, each expressing a distinct concept. The synsets, on their turn, are associated according to their meaning by means of semantic relations, forming a word network [11]. Therefore, the Wordnet is an excellent information source to establish explicit relations between terms that compose tags and terms associated according to their meanings.

A problem with getting the semantic relations from the Wordnet is the different written forms of tags as, for example, synonyms and plurals. Furthermore, a word may have more than one meaning (ambiguity), which can impair the attainment of semantic relations and, subsequently, the retrieval of information.

To identify the Wordnet concepts in which a tag is related, firstly the tags must pass through a cleaning process that consists of a lemmatization and a tokenization. Thus, grouped tags, e.g. “CamelCase”, “programming_language”, etc, are broken into tokens. The grouped tag is broken in places where special characters or upper case characters (indicating camel case) occur. Then these tags and tokens must undergo a lemmatization process, which is necessary to identify the base form of a word. For example, if the lemmatization algorithm receives the plural term “women”, it returns the singular form of the word, i.e. “woman”, which is in the Wordnet database. Afterwards, these processed terms are used to find concepts in the Wordnet. When the lemmatized label is a compound tag, the process is carried out by trying to find concepts keeping all the tag tokens. Sometimes it is necessary to try many different forms, for example, using spaces or hyphens between the tokens. When a compound tag is found in the Wordnet using all its tokens, this written form is associated as the hasLemmatizedLabel attribute of that tag. As an example, we have the tag “programmingLanguage” which is present in the Wordnet as “programming language” (space separated). This would be the lemmatized label. The separated terms “programming” and “language”—also present in the Wordnet—would be associated with the “programmingLanguage” tag as

---

1 Lemmatization is a process which reduces the inflectional forms and sometimes derivationally related forms of a word to a common base form.
hasToken and isTokenOf relations, as observed in Figure 2. When no occurrence of a tag is found in the Wordnet, the only relation between a term with others will be the co-occurrence. However, if more than one concept from a tag is found in the Wordnet, a sense disambiguation is needed to identify the best choice (which is described in Session 3.3). For example, the word “jaguar” can mean the brand of cars or the animal, since no explicit semantic is associated by the user to the tags used in categorizations the automatic definition of linguistic relations is very difficult.

As important as identifying concepts that represent the user’s tags in the Wordnet is to identify and set the semantic relations between them. This may be obtained from the Wordnet relations of synonymy, hypernymy, hyponymy, meronymy and holonymy, which are mapped on the ontological model respectively as hasAlternative, isA, kindOf, hasPart and isPartOf, because these terms would be more easily understood and because they are more suitable to be used with other information sources than the Wordnet. A level of synonyms and more specific terms (holonyms and hyponyms) are obtained, which serve to help the user to remember the used tags, and N levels for broader relations (up to the Wordnet root node), since they are used to help organize data in the form of a hierarchy, allowing this to be shown instead of tag clouds and lists. In Figure 2 we can observe a simplified example of the input and the respective output obtained. We may observe that from the tagspace (“programmingLanguage”, “Java” and “Prolog”), whose only relation is the co-occurrence, an ontology was evolved with other semantic relations among tags. Moreover, terms that are not tags were added to help, for example, in searches for concepts instead of keywords, and to show the tagspace in the form of a hierarchy.

It also can be observed in Figure 2 that more than one sense was found for the tag “java”, which are represented by the signs “s1” (“a beverage consisting of an infusion of ground coffee beans”) and “s2” (“a platform-independent object-oriented programming language”), generating an out-of-context branch in the given graph (“java” (s1), “beverage”, “food”, ..., “entity”). For this reason, a tag sense disambiguation process is needed to avoid this kind of undesired information.

A. The Process of Tag Sense Disambiguation

A word has a limited number of senses (often provided by a dictionary or other reference source), and the task of disambiguation is to make a forced choice between these meanings to each use of an ambiguous word, based on its context [12]. On the problem here addressed the various senses of a word are disposed in the Wordnet database, represented by synsets. To identify the context of a tag in a categorization, we can use (i) the co-occurrence tags, (ii) the title of the categorized resource, (iii) an possible description of the categorized resource and, in some cases, (iv) the own categorized resource.

Therefore, we tested the use of a semantic similarity measure called lin [13] and the semantic relatedness measure called adapted leks [14]. Both measures have the same objective: to describe how strong the senses of words are interconnected. For this reason, they were tested separately to identify the most suitable one for the term disambiguation process to be used in categorization. Considering that some Wordnet relations form a hierarchy, the measure proposed by Lin [13] sums the value of Information Content² (IC) of the last subsumer of two concepts with the sum of IC of each one separately. The adapted leks measure [14] works by finding overlaps in the glosses of two compared synsets. The relatedness score is the sum of the squares of the overlap lengths. Its implementation does not use only the synset gloss, but also uses relations between them to compare with glosses of near synsets. The disambiguation, considering just the co-occurrence to identify tags context, is done comparing all the synsets corresponding to tags of a categorization

Figure 2. A simplified example with the input of co-occurrent tags and the emergent ontology.

² Information Content measures the specificity of a given concept being based on pre-determined values obtained from tests in corpus.
among themselves. The strongly connected senses (synsets) are the ones used to get the semantic relations used in the emergence of ontologies from the user’s tags. It would eliminate, for example, the branch related to the tag “Java” that is out of context in Figure 2, which represents concepts of “Java coffee” (s1), instead of the programming language with this name (s2).

As an optional step for the disambiguation process, we may consider the use of the title of the categorization for context identification. For that reason, after the comparison of tags from a categorization among themselves, meaningful words from the title are also compared with the tags to enhance the choice of the synset semantically more appropriated. To use the title in a more efficient way in this process it has to be “cleaned”, i.e. it must have: (i) the punctuation removed; (ii) the words lemmatized; and (iii) it the “stop words” (these are words that do not help to identify the context of a phrase, e.g. terms like ‘a’, ‘but’, ‘for’, etc) removed. After this process, the list of remaining words from the title can be compared with the categorization tags to find the stronger semantic similarity/relatedness between the tag’s senses and to choose the more adequate meaning.

The object description and the object itself were not used for disambiguation, because this information is often not available or, depending on the categorized resource, does not help in automatically recovering the context of a tag.

B. Evaluation

In an initial experiment we performed over the tags of a user, 64% of the tags found in the Wordnet had more than one meaning. After the disambiguation process, using only the co-occurrent tags for the identification of the context, 79% of the ambiguous tags had a positive level of semantic similarity between their meanings using the lin measure and 92% had a positive level of semantic relatedness using the lesk measure. By using the title of the resource to improve the identification of the context, these values reached 90% with the lin metric and 97% with the adapted lesk measure. Unfortunately, despite the identification of some degree of semantic similarity or relatedness, this does not indicate that the choices of the meanings were correct.

To measure the quality of the disambiguation process a random sample of categorizations was extracted from different users, and the meanings chosen were verified manually, according to the synset glosses obtained in the Wordnet and the context of the categorization (using the resource title and the co-occurrent tags). The sample of the categorizations chosen belongs to several subjects such as economy, psychology, politics, cooking, learning, photography, technology and history. Using only co-occurrent tags to determine the context, 64% of the tags that passed through the disambiguation process were correct in the choice of meanings using the lin metric and 68% using the adapted lesk measure. These values changed to 64% and 71% respectively when using both the co-occurrent tags and the resource titles to identify the context, as shown in Figure 3.

![Figure 3. Results from the disambiguation process.](image)

As shown in Figure 3, the correctness of the results is not significantly improved using the resource title to help in the context identification, but the quantity of tags for which some degree of similarity was obtained is improved. Due to these results, we decided to use the adapted lesk metric, once better results were obtained in quantity and quality of sense choices.

Analyzing the problems in tag sense identification, it was noticed that the categorizations related to technology were those with more errors in disambiguation choices. In addition, more abstract terms such as "design", which has many meanings in the Wordnet, did not get satisfactory results. We have noticed that many synsets of these words are very similar between them and in the future we will try to compare the similarity of these senses and maybe (depending on the similarity level) set more than one meaning for the correspondent tag.

In order to test the proposed methodology, we applied our algorithm for the emergence of ontologies over a random sample of 2,100 personomies, totaling approximately 1,730,056 tags. After the tag enriching step, 53% of them were recognized in the Wordnet. The tags not found in the Wordnet will continue being related to others just by co-occurrence, not achieving the semantic benefits. According to Laniado et al. [9], “the most popular tags have a much higher probability of belonging to the Wordnet”.

We believed that more tags could be recognized in the Wordnet if the users employed a tool like TagTydier [4], whose main goal is the detection of inconsistencies in the tagspace, providing a cleaner and more organized vocabulary for the emergence of an ontology and representing an important initiative to assist in the control and organization of personomies. Also, the use of other information sources than the Wordnet (described in Session 4), and/or the
recommendation of tags belonging to a semantic source (to prevent errors in typing and ambiguity) in the moment of the categorization could help to improve the recognition rate.

C. Applications

To verify the usefulness of the emergent ontology some tests were done on two possibilities: (i) to improve the process of information retrieval in persononomies, and (ii) to make it possible to show tags in the form of a hierarchy alternatively to tag clouds and tag lists.

In respect to information retrieval, we had an improvement in the result quality by using the semantic relations of the ontology. This could be attributed to the fact that “after processing a linguistic message, people normally remember just its meaning and not its exact wording” [6], which can hinder information retrieval in a personomy. Thus, the ontology could help the users by allowing the search engine to use concept instead of just words. The semantic relations in the emergent ontology enables searches with any alternative term of a concept, instead of the exact form a tag was written. For example, in a search for “airplane”, returned resources categorized with the tags “plane” and “aero plane” could be returned, since they are related by the hasAlternative relations in the ontology.

Another possibility is to search for narrower concepts in the ontology than the specified term, which could return all the resources categorized with more specific tags or tags that represents parts of a whole. For instance, in a search for the term “vehicle”, even if this term is not used in the categorizations, results categorized with the tags “car”, “truck”, bike, etc could be shown. In the same way it is possible to search for broader concepts, which could return all the resources categorized with general tags or tags which constitute a whole. For example, in a search for the term “pedal”, which is not a user created tag, the search engine could return results categorized with “bike”, which is related with “pedal” by the relation hasPart/isPartOf and is a tag used to categorize resources.

When more than one term is entered in the search field, the search could be made by using the intersection between the users’ created tags related to the input terms. Figure 4 shows an example of a search for the terms “tires” and “vehicle”. In this hypothetical search, the input term “vehicle” is not a tag from the user’s personomy, but specializing the search using semantic relations some user created tags were found (“car” and “bus”). The other input term for the hypothetical search was “tire” that is a user created tag. The overlap between the three user created tags (“car”, “bus” and “tire”) can be used to return the results relative to “tires of a vehicle”.

In respect to an alternative exhibition of the tagspace, we work in the derivation of a hierarchy starting from the root node of the generalization/specialization relations of the ontology. The hierarchy developed shows the relations between the concepts obtained from the Wordnet until the tags of the user’s personomy are reached. However, the Wordnet relations are very fine grained, providing a hierarchy with many levels, which may hinder the user navigation. Thus, we needed to employ a “cleaning” process to remove some levels of unnecessary nodes. Similarly to the proposal by Laniado et al. [9], our approach consists of making a depth first search removing all nodes with less than two sub-nodes which do not represent a user’s personomy tag. After a node removal its sub-nodes are added into its super-node. The higher level nodes as, for example, “entity”, “physical entity”, “abstraction”, etc. can also be concealed from being exhibited, as they are too abstract. A problem in displaying a hierarchy of terms is that in a first glance the user may not be familiar with the used terminology, which in parts is taken from the Wordnet. However, when users get used to the terminology, the hierarchy helps them abstract the large quantity of tags that appear in traditional tag clouds and lists, which requires great cognitive effort to identify the desired terms for the recovery of categorized resources. In addition, hierarchies, due to their abstraction levels, may be better displayed on a web browser and this helps the user to recognize the terms used in categorizations, a cognitively easier task than to remember tags used in categorizations [6]. An example of a hierarchical structure for navigation in tagspace resulting from a users’ vocabulary through an emergent ontology is shown in Figure 5.

A third possibility is to use the emerged ontology to recommend terms for the categorization of new resources. We cannot say much about this yet, since our experiment is in its early steps, but we are studying various possibilities. There is a possibility for when the user enters with the first tag/term that he/she wishes to use in a categorization to suggest other terms from the ontology relations, reached with the use of a spreading activation algorithm. This could be associated with the recommendation of the first entered terms (input) from other sources as, for example, from co-occurrent tags in folksonomies, or from indexing the categorized

Figure 4. A hypothetical search for “tire and “vehicle”. 
resource, etc. The tag recommendation could be shown in the form of graphs derived from the emergent ontology, with many terms between the input term and the tags already used in other categorizations. This suggestion becomes even more interesting because as tags are chosen in the graph they are already associated with a concept (a Wordnet synset) and, for this reason, the disambiguation process is not necessary when the user selects one of them.

IV. CONCLUSIONS AND FINAL CONSIDERATIONS

Tagging is a good approach to reduce the cognitive effort from users in information categorization process. However, when using tagging the information retrieval process is hindered by several aspects. In this paper we proposed an unsupervised approach for the emergence of ontologies from personomies aiming at the improvement of the information retrieval process. This improvement comes from more advanced search possibilities achieved using the emergent ontologies and the possibility to show tags in more organized forms than in tag clouds and lists.

The proposed process is an unsupervised approach, which is very important and valuable because it does not need the users’ intervention, and, thus, it does not add to the user’s cognitive effort in categorizing information. Unfortunately, as a side effect, it becomes difficult to automatically identify the meanings of the tags. For this task, we proposed a disambiguation algorithm based on the adapted lesk semantic relatedness measure.

The quality of the users’ vocabulary has great influence in the disambiguation process and the resulting ontology. For this reason, we suggest that for the emergence of a quality ontology from a personomy it could be better to use recommended tags in the categorization process from semantic concepts described by existent ontologies instead of allowing just free text labels. In the case of using free tags we recommended that they are “cleaned” before generating the ontology.

A restriction of our proposal is that we only studied tags in the English language. We choose this because this is the base idiom of the Wordnet, but the result could be easily generalized for other idioms.

As a future work we intent to use other information sources to recognize more tags (in addition to those recognized in Wordnet), and to obtain other semantic relations besides those available in the Wordnet. Among the information sources that are being investigated are the ConceptNet (http://conceptnet.media.mit.edu/) that makes use of commonsense knowledge to relate concepts; the DBpedia (http://dbpedia.org), which is a community effort to extract structured information from Wikipedia (http://wikipedia.org) and to make this information available on the web in the format of an ontology, and an acronym dictionary.

The proposal for the emergence of ontologies discussed in this paper will be used in the future by the TagManager [4] system for the exhibition of the tagspace in an alternative to tag clouds and lists, and to improve the information retrieval process, allowing searches by concept instead of using only the written form of tags. We believe that the generated ontologies can also be used by other information retrieval systems and by content recommendation systems, since the ontology may have information about the user’s interest.

ACKNOWLEDGMENT

We would like to thanks CAPES for the scholarship granted to Carlos A. M. Basso during his Master degree from which this paper is a partial result.

REFERENCES