Leveraging semantic similarity for folksonomy-based recommendation

Daniela Godoy, Gustavo Rodriguez and Franco Scavuzzo

Abstract

For recommending interesting resources, such as Web pages or pictures available in social tagging systems, assessing their similarity with user profiles is crucial. Here, we analyze the role of semantic similarity to calculate the resemblance between user profiles and published resources in folksonomies. Experiments carried out with data from two social sites showed that associating semantics to tags results in more accurate similarities among elements in tagging systems and, consequently, enhances recommendations.

Keywords: Social tagging systems; folksonomies; similarity measures; recommender systems.

1 Introduction

Social tagging refers to the practice of collaboratively annotate Web resources using textual labels, also known as tags. Del.icio.us, Flickr and CiteULike are examples of social sites in which users share a variety of resources, such as Web pages, pictures, videos and bibliographic references. The result of this collaborative tagging process is a social classification scheme known as folksonomy in which users and resources are related through tag assignments.

The rapidly growing size of communities using social sites as well as the large amount of shared resources available in folksonomies, make the discovery of relevant content a time consuming and difficult task for users. Unsupervised tagging and the lacking of a control vocabulary for annotating resources exacerbate this problem as social tags are naturally noisy and ambiguous, reducing its effectiveness for content indexing and searching.

In this context, recommender systems bringing support to users in tagging, searching and discovering novel resources are becoming not only valuable, but also extremely necessary tools. In fact, multiple recommender systems have emerged for social tagging sites. Traditional approaches for building recommender systems are basically built upon content-based and collaborative filtering techniques, both heavily relying on the notion of similarity. In the content-based approach potentially interesting items are predicted according to their similarity with the items the user liked in the past, while in the collaborative approach, items are recommended if they were interested to people the user shares interests with.

In social tagging user interest profiles can be represented with the tags a user tends to annotate resources with, whereas resources with the tags they tend to be annotated with. Thus, similarity assessments of tag-based representations are affected by the syntactic variations of tags stemming from the use of different verbal forms, plurals, acronyms, synonyms, among other reasons. By relating tags with semantic entities from lexical resources tag-based representations can be
enriched, reducing the effects of the previously mentioned problems.

Quantifying the relatedness among tags, resources and users is essential to the development of recommender systems for folksonomies. Enhancing similarity measures with semantic knowledge extracted from lexical databases such as WordNet allows to get better similarity assessments and, in turn, improve recommendations. In this paper we aim at analyzing the role of semantic similarity in the context of traditional recommendation approaches for suggesting interesting resources in folksonomies. Particularly, our work focused on empirically evaluating the impact of associating concepts to tag-based representations of folksonomy elements (users and resources) on the precision of the recommendations delivered to users.

2 Similarity Measures in Folksonomies

In order to apply classical content-based and collaborative filtering recommendation techniques, the similarity among users and resources or users and other users need to be estimated starting from the folksonomy. Both users and resources are represented according to the third dimension involved in tagging systems, which is the social tags used for annotation purposes. Then, users assigning many of the same tags can be seen as alike, whereas resources are similar to a user profile if they are annotated with tags the user tend to apply.

Figure 1 depicts the recommendation scheme in social tagging systems for suggesting potentially interesting resources. On the one hand, users are assumed to be interested in the resources they annotate, so that other similarly tagged resources would also be interesting for them. On the other hand, users resemble each other if they tend to use the same tags, then one user can receive as recommendation a resource annotated by the other user. In this scheme, similarities among profiles and resources derived from the similarity of their tag-based representations.

Figure 1: Resource recommendation approaches in social tagging systems
2.1 Semantic-Based Tag Similarity

Semantic assessments of similarity between two tags are obtained by associating them to semantic entities or concepts. For this purpose, we use WordNet\(^1\), a large lexical database of English language that groups English words into sets of synonyms called synsets and records various semantic relations between these synonym sets.

For nouns and verbs a subsumption hierarchy can be obtained based on the \(is-a\) relationships that connects hyponyms (more specific synsets) to hypernyms (more general synsets). Since a synset can have multiple hypernyms, the network become a directed acyclic graph to which a top-level node subsuming all the roots of the disconnected hierarchies is added to make the graph fully connected. Hence, approaches to measuring semantic relatedness between concepts base the measure of similarity on the properties of paths in this graph.

Resnik’s information-based approach \([11]\) is based on the intuition that the shorter the path from one concept to another, the more similar they are, which in an \(is-a\) taxonomy can be determined by inspecting the relative position of the most specific concept that subsumes both concepts.

Let \(C\) be the set of concepts in an is-a taxonomy permitting multiple inheritance. The taxonomy is augmented with a function \(p : C \rightarrow [0,1]\), such that for any \(c \in C\), \(p(c)\) is the probability of encountering an instance of concept \(c\). Following the standard definition from information theory, the information content of \(c\), denoted \(IC(c)\), is then:

\[
IC(c) = -\log p(c)
\]  

where \(p(c)\) is the probability of finding \(c\) in a given corpora:

\[
p(c) = \frac{\sum_{w \in W(c)} frequency(w)}{N}
\]

where \(w\) is a word, \(W(c)\) is the set of words describing the concept \(c\) and \(N\) is the total number of words in the corpora. In this way, polysemous words contribute with the frequency of all their meanings. In other words, \(IC\) is obtained through statistical analysis of a corpora from where probabilities of concepts occurrences are inferred. In the experimentation reported in this work, \(IC\) is derived from SemCor, a manually sense-tagged subset of the Brown Corpus under the assumption that each word is used in its most often occurring sense.

The semantic similarity of a pair of concepts \(c_1\) and \(c_2\) is then defined as:

\[
sim_{\text{Resnick}}(c_1, c_2) = \max_{c \in S(c_1, c_2)} IC(c)
\]

where \(S(c_1, c_2)\) is the set of concepts that subsume both \(c_1\) and \(c_2\).

Following this idea Jiang and Conrath \([6]\) proposed a similarity measure postulating that the

\(^1\) http://wordnet.princeton.edu/
semantic distance of the link connecting a child-concept \( c \) to its parent-concept \( c_p \) is
proportional to the conditional probability \( p(c|c_p) \) of encountering an instance of \( c \) given an
instance of \( c_p \). From this postulate derives the following formula for the semantic similarity
between concepts \( c_1 \) and \( c_2 \):

\[
sim_{JCC}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) + 2 \cdot \text{sim}_{Rsnick}(c_1, c_2)}
\]

(4)

This measure demonstrates to be the most appropriate to measure concept relatedness in
WordNet [3].

2.2 Similarity of Tag-based Representations

User profiles and resources from social tagging systems are translated to a bag-of-words
representation in which each element is identified by a feature vector with a numerical value or
weight indicating its importance. Each element (user or resource) from a folksonomy \( F \) is then
identified by a vector \( v \) in the \( t \)-dimensional space, where each vector component \( w_{ij} \) represents
the weight of the tag \( t_i \) in the element \( j \):

\[
\overrightarrow{v_j} = (w_{1j}, w_{2j}, \ldots, w_{|T|j})
\]

(5)

where \( t \in T \), \(|T|\) represents the total number of tags in the folksonomy and weights are
assumed to be zero if the tag is not present in the element description.

For representing a user \( u \) in the folksonomy a vector \( \overrightarrow{v_u} \), constituting the user profile, is
created based on all tags employed by this user to annotate resources. The amount of non-zero
entries of this vector corresponds to the set of all tags from \( T_u \) in the user personomy \( uP \), the
restriction of the folksonomy to a single user. Tag weights are assigned to each of these tags
according to their frequency of appearance in the resources annotated by the user.

Likewise, a resource \( r \) in the folksonomy is represented by a vector \( \overrightarrow{v_r} \) in which non-zero
entries correspond to all tags assigned to the resource by members of the community. Tag
weights are assigned according to the number of users in the folksonomy \( F \) that annotated the
resource \( r \) with each tag, i.e. how many times the tag was assigned to the resource.

In this work we evaluated three similarity measures to determine the degree of resemblance of
two users or a user with a resource for generating recommendations.

2.2.1 Overlap Similarity

Overlap similarity simply measures the amount of tags two vector representations have in
common, where a vector represents either a user or a resource. This measure is defined as:
where $V_i$ and $V_j$ are the sets of all tags of user $i$ and $j$, respectively.

### 2.2.2 Cosine Similarity

Commonly used in information retrieval area, the cosine measure estimates the similarity between two vectors as the cosine of the angle the form in a vector space. Given two vectors $v_i$ and $v_j$, each representing either a user or a resource, the cosine similarity is computed as follows:

$$
\text{sim}_\text{cos}(v_i, v_j) = \frac{\sum_{t \in T} w_{ti} \cdot w_{tj}}{\sqrt{\sum_{t \in T} w_{ti}^2} \cdot \sqrt{\sum_{t \in T} w_{tj}^2}}
$$

where $w_{ti}$ is the weight of the tag $t$ in the vector representation $v_i$ of a user or resource $i$.

### 2.2.3 Semantic Similarity

For defining a semantic similarity, we extended the cosine similarity to consider concepts associated to tag in WordNet instead of simply tag matching. Normally, the cosine similarity will only consider the products of those dimensions having an exact match in both vectors, since for non-matching tags the corresponding dimensions in the second vector will be zero.

For considering semantically enriched tags, if a tag in the first vector match other in the second vector their weights are multiplied. Otherwise, the tag weight is multiplied by the weight of the most similar tag in the second vector found with the Jiang and Conrath similarity, provided it exceeds a certain threshold. If there is not similar enough tag, the tag in the first vector is multiplied by zero.

Figure 2 depicts an example of how semantic similarity is calculated. Let suppose a user, represented by all the tags he has assigned to its resources, and a resource, represented by all the tags users in the community annotated it with. In the example, the tag technology is present in both vector representations. For the tags music and business some similar tags can be found, which are songs and investments, respectively. In contrast, the travel and news tags have neither a direct matching nor a semantically similar tag in the other vector.
More formally, the semantic similarity measure can be defined as:

$$\text{sim}_{\text{sem}}(v_i, v_j) = \sum_{t \in T} w_{ij}$$  \hspace{1cm} (8)

where $w_{ij}$ is calculated according to Equation 9.

$$w_{ij} = \begin{cases} w_{pi} \cdot w_{pj} & \text{if } t_i = t_j, \\ w_{ti} \cdot w_{tk} & \text{if } t_i \sim t_k, \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (9)

and

$$t_k = \arg \max_{t \in T} \text{sim}_{J&K}(t_i, t)$$  \hspace{1cm} (10)

For setting the threshold of relatedness of two tags we considered the study of Busanitsky and Hirst [3] that analyzed how well several similarity measures reflect human judgments of semantic relatedness. According to Jiang and Conrath findings, distance values up to $\sim 10$ means high semantic similarity between two terms in WordNet, then we apply this threshold in our experiments. However, it remains to determine how this threshold impacts on the precision of recommendation approaches.

3 Empirical Evaluation

3.1 Description of Datasets

For the empirical evaluation and comparison of similarity measures we used two datasets gathered from different social tagging systems: CiteULike, a social bookmarking system for tagging academic papers, and CABS120k08 [10], a collection of Web pages with annotations.
extracted from one of the main social bookmarking sites on the Web as Del.icio.us. The first is a domain-specific social bookmarking service and the second a general-purpose one, so that together they provide a good perspective of the type of folksonomies that can be found on the Web. CiteULike dump used for experiments counts with 45,028 users, 299,112 tags and 1,464,648 articles, related by a total of 5,323,631 tag assignments, collected in the period 11-2004 to 04-2009. CABS120k08 dataset contains 117,434 URLs with additional meta-data, it is composed of 388,963 users, 175,910 unique tags and 117,434 Web pages, related by a total of 3,528,875 tag assignments, collected in the period 01-2004 to 10-2007.

Social tagging systems on the Web own their success to the opportunity of freely determine the tags for resources without the constraint of a controlled vocabulary, lexicon or pre-defined hierarchy. Uncontrolled vocabulary also leads to a number of problems in the resulting tags such as misspellings, synonyms and morphological variety. In turn, syntactic mismatches in tags interfere with the task of recommendation algorithms.

We applied a filtering approach to alleviate these problems. First, a filter of compound words replaces symbols such as -. _ or & by whitespace characters to divide individual terms. Second, a dictionary filter verifies whether each individual term exists in an English dictionary. If the word appears in the dictionary it passed directly to the stemming filter, if not its spelling is checked using the Spelling Suggestion Web service offered by Yahoo!, which provides a suggested spelling correction. The last filter corresponds to the stemming of the remaining words using Porter algorithm to solve morphological variations.

For enabling the calculation of semantic similarity of tag pairs both must be present in WordNet dictionary. In CiteULike collection 89% of the top-100 tags (order by frequency) are covered by WordNet and 97% in CABS120k08, considering the top-500 the percentages are 90% and 90.8% and in the top-1000 percentages drop to 88.3% and 89%, respectively. The high level of WordNet coverage ensures that semantic similarity will be calculated on the basis of available semantic relationships.

3.2 Resource-to-User Recommendation

In folksonomies, the introduction of tags caused the emergence of tag-based profiling approaches, which assume that users expose their preferences for certain contents through tag assignments. Resource-to-user recommendation approach focus on building a representation of resources published in folksonomies and learning profiles of user interests so that resources matching a profile can be recommended.

User interest representations in this approach are obtained by creating a vector of all tags the users employed to annotate their resources weighted according to their frequency. Likewise, resource representations are given by the overall set of tags assigned to resources in the system with the same weighting strategy. Both representations are compared using one of the above similarity measures, resources exceeding some similarity threshold are recommended to users.

For the experiments we used personomies in both datasets, which provide ground-truth about relevant resources, assuming that a user is interested in the resources he tagged. Oppositely,
irrelevant resources are not available as this would require to acquire explicit user judgments. For this reason, we focused experimental evaluation in determining the impact of semantic similarity in recognizing interesting resources that can be recommended to the user with some certainty. We conducted experiments using a holdout strategy that randomly splits a user personomy into 80% for training, used to learn the user interest profiles or vector representation of the user, and 20% for testing, used for validation. In order to make the results less dependent of data splitting, in all experiments the average and standard deviation of 10 runs for each user are reported.

Quality of recommendations was evaluated considering the number of hits, i.e. the number of resources in the test set that were also present in the list of recommendations. Hit-rate grants high values to an algorithm that is able to predict the user interests and low values otherwise. If $N$ is the size of the test set, the hit-rate of a recommendation algorithm is computed as $\frac{\text{number of hits}}{N}$.

Figure 3 depicts the results obtained for the resource-to-user recommendation scheme in CABS120k08 and CiteULike datasets. It can be observed that a simple overlap measure performs poorly recommending resources. Although considerably better, the cosine similarity measure based on tag matching can only recommend less of a half of the resources the user is interesting in. Once enriched with semantic knowledge, tags lead to a higher number of hits.

![Figure 3: Hit-rates of the content-based approach for recommendations in CABS120k08 and CiteULike datasets](image)

3.3 User-to-User Recommendation

User-based recommendation considers that users assigning similar tags share information interests. For recommendation, this approach learns the user profiles by building the weighted tag vectors with the tags the users assigned to their resources and searches the $K$ more similar users...
by comparing the profile of the active user with the remaining users in the folksonomy. Finally, the resources annotated by the $K$ more similar users are employed to generate recommendations ordering them according to two factors: the distance of the user providing the candidate resource (i.e. a resource is more important if it is suggested by a nearer user) and the number of votes the resource received (i.e. the more neighbors have annotated the resource, the more important).

Like in the previous experiments a holdout strategy was used. To reduce the number of users involved in the search for neighbors, only those having at least a resource tagged in common with the active user were considered.

Since a ranked list of suggestions is produced, the order in which the recommendations are presented becomes important. We used MAP (Mean Average Precision), a metric that emphasizes ranking relevant recommendations higher, to evaluate the results. This measure averages the precisions computed at the point of each of the relevant recommendations in the ranked sequence.

Figures 4 (a) and (b) show the results obtained for several values of $K$ in CABS120k08 and CiteULike datasets. User-to-user results show the same trend than resource-to-user ones. Semantic similarity of tag vectors outperforms cosine and overlap similarities based on simple matching.

![Figure 4: MAP scores of the user-to-user approach](image)

### 3.4 Hybrid Recommendation

A well-known disadvantage of content-based recommendation approaches (resource-to-user) is that they produce recommendations very similar to the items already seen by the user in the past. In a hybrid approach, the combination with collaborative recommendations (user-to-user) adds diversity to the list of suggestions.

For obtaining hybrid recommendations, the candidate resources of the user-to-user approach were ranked based on their resource-to-user similarity with the user profiles. In this setting, there are two similarities to be assessed: (1) the user-user similarity of the collaborative approach to
obtain the $K$ nearest neighbors, and (2) the similarity of the candidate resource (only those not already tagged by the target user) with the user profile to determine the resource position in the ranked list of recommendations.

Experiments using different combinations of cosine and semantic similarities to estimate (1) and (2) were carried out, overlap was discarded as it had the poorest performance. Figures 5 (a) and (b) depict the MAP scores achieved with CABS120k08 and CiteULike, respectively. MAPs scores using semantic similarity in one or both approach steps outperform the use of cosine similarity exclusively. The best performing combination is the one using semantic similarity in both steps.

![Figure 5: MAP scores for the hybrid approach](image)

4 Related Works

The problem of inferring relatedness between social tags has been addressed in recent works due to its potential for a wide range of applications, such as tag clustering, ontology learning, query expansion, and search assistance besides recommendation. In [13], tag weighting models taking advantage of the three dimensions of folksonomies to base similarity on were analyzed using classical (non-semantic) similarity measures such as Dice, cosine and mutual information. Co-occurrence and its distributional version are used in [9] to define relatedness directly on the network structure of folksonomies instead of using a lexical resource. WordNet was used as gold-standard in the previous works to evaluate the inferred relationships. In contrast to these works, we exploit semantic relationships in WordNet structure for assessing tag similarity. More importantly, we empirically evaluate the impact of semantics in the calculation of user-user and user-resource similarities in the context of classical recommendation approaches.

Closely related to our work is Bogers et al. [2] study on content-based and collaborative methods using different similarity metrics. For item-based filtering, using tags to calculate similarity alleviates sparsity and then improves the results of recommendation. User-based filtering did not lead to the same results. Our work differs from this study in that also a semantic
5 Conclusions

This paper focuses on resource recommendation in social tagging systems, i.e. predicting which unseen resources a user might like (Web pages, pictures, videos, bibliographic references, etc.) based on their comparison with the user profile or the identification of like-minded users. For this task traditional content-based and collaborative filtering techniques are commonly applied with the incorporation of tags as a means of binding together resources and users.

Critical to the problem of searching nearest neighbors and related resources in a folksonomy is the capacity of attaining accurate similarity assessments. In this regard, overlap and cosine similarity measures based on tag matching were compared with a semantic similarity measure in a resource-to-user and user-to-user approaches for recommendation. A hybrid approach combining these approaches outperforms both of them as it involves semantic similarity for locating users and ranked recommendation according to the user profiles. A similarity measure builds upon the relatedness of concepts associated to tag pairs extracted from WordNet dictionary was employed to overcome the noise and ambiguity of social tags. The empirical study carried out showed that the incorporation of semantics produced better quality recommendation lists so that it can be expected to enhance further recommendation approaches involving similarity calculations. Other sources of semantic information such as Wikipedia will be considered and compared with WordNet results in future research.

Acknowledgments

This research was partially supported by the National Scientific and Technical Research Council (CONICET) under grant PIP Nro. 114-200901-00381.

References


Daniela Godoy is a full-time professor in the Computer Science Department at the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN), Argentina. She is a member of ISISTAN Research Institute and researcher at the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). She has a PhD in computer science from UNICEN. Her research interests include user profiling, text mining and recommender systems. Contact her at daniela.godoy@isistan.unicen.edu.ar; http://www.exa.unicen.edu.ar/~dgodoy.

Gustavo Rodriguez received a BSc. degree in Systems Engineering from the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN), Argentina, in 2011, where he was also a graduate teaching assistant. He is currently working on the private sector. His interests are about social tagging systems and recommendation technologies. Contact him at rodriguezgustavoandres@gmail.com.

Franco Scavuzzo received a BSc. degree in Systems Engineering from the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNICEN), Argentina, in 2011, where he was also a graduate teaching assistant. He is currently working on the private sector. His interests are about social tagging systems and recommendation technologies. Contact him at francoscavuzzo@gmail.com.