A Tag Clustering Method to Deal with Syntactic Variations on Collaborative Social Networks

José Javier Astrain, Francisco Echarte, Alberto Córdoba, and Jesús Villadangos

Dpt. de Ingeniería Matemática e Informática
Universidad Pública de Navarra
Campus de Arrosadía. 31006 Pamplona, Spain
josej.astrain@unavarra.es,
patxi@eslomas.com,{alberto.cordoba,jesusv}@unavarra.es

Abstract. Folksonomies have emerged as a common way of annotating and categorizing content using a set of tags that are created and managed in a collaborative way. Tags carry the semantic information within a folksonomy, and provide thus the link to ontologies. The appeal of folksonomies comes from the fact that they require a low effort for creation and maintenance since they are community-generated. However they present important drawbacks regarding their limited navigation and searching capabilities, in contrast with other methods as taxonomies, thesauruses and ontologies. One of these drawbacks is an effect of its flexibility for tagging, producing frequently multiple syntactic variations of a same tag. Similarity measures allow the correct identification of tag variations when tag lengths are greater than five symbols. In this paper we propose the use of cosine relatedness measures in order to cluster tags with lengths lower or equal than five symbols. We build a discriminator based on the combination of a fuzzy similarity and a cosine measures and we analyze the results obtained.

Keywords: Folksonomies, Fuzzy similarity, Tag Annotation, Tag Clustering.

1 Introduction

Folksonomies offer users an easy way to sort and organize resources by assigning text tags at different resources, such as photos, web pages, documents, etc. User annotations and categorization define collaboratively the semantics of the tags and resources used, causing the emergence of their associated semantics (unstructured knowledge). Folksonomies are based on the assignation of text tags to these resources. Tag annotation allows defining collaboratively the meaning of the annotated resources, and the used tags. Tags provide users new approaches for information search and exploration. Some navigation tools as clouds of tags allow users searching for certain tags and to locate resources tagged by other users. Though folksonomies have a great success in current web, mainly due to their simplicity of use, they also have important drawbacks. The fact of users creating tags and assigning them freely to resources produces the inexistence of any structure among these tags.

M. Gaedke, M. Grossniklaus, and O. Díaz (Eds.): ICWE 2009, LNCS 5648, pp. 434 – 441, 2009.

Users can introduce synonyms, syntactic tag variations or different granularity levels [1] in the tagging process, lowering the quality of folksonomies and making more difficult the exploration and retrieval of information [2,3].

Several works in the literature focus on solving some of the problems associated with folksonomies. Folksonomy browsing is addressed in [4] presenting different ways to display tag clouds; [5] analyze the co-occurrence of labels to improve the quality of tag clouds. Tag clustering is addressed in [6,7]. An in-depth study of semantic tag relatedness is addressed in [8]. The problem of exploring hierarchical semantics from social annotations is studied in [7]; and [6] deals with the conversion of a large corpus of tags into a navigable hierarchical taxonomy using a graph of similarities. Other proposals such as [9] propose to improve the quality of folksonomies supporting users in the task of resource annotation by suggesting tags. Other works as [1,10] relate folksonomies with formal information classification systems as ontologies [11] and personalized recommendation [12].

Most of the above proposals do not take into account that a relevant number of the existing tags corresponds to syntactic variations (erroneous or not) of previously existing tags. The performance of a pre-filtering of the tags before applying an algorithm for tag clustering, as occurs in [13], allows minimizing the effects of syntactic variations increasing the quality of tag clustering. In [13] Specia and Motta create clusters of semantically related tags over a reduced experimental data set, using a previous step in which Levenshtein similarity measure is used to reduce the number of tags identifying syntactic variations. Then the folksonomy is changed replacing each identified variation by a representative tag. Another way to represent these variations is presented in [1, 14]. The use of pattern matching techniques designed to automatically recognize syntactic variations of tags provides mechanisms to improve the quality of folksonomies [15]. Approximate string matching techniques allow dealing with the problem introduced by syntactic variations on folksonomies. The problem consists on the comparison of an observed input string called α, maybe containing errors, and a pattern string ω in order to transform α in ω [16]. Edit operations (insertion, deletion and change of a symbol) allow recovering those errors transforming α in ω. The number of edit operations needed to recover an input string provides a distance measure between the input string and the pattern string. This distance, known as edit distance, can be expressed in terms of similarity and distance (dissimilarity) measures between strings. Imperfect pattern matching techniques perform better when dealing with syntactic variations of tags as indicated in [17]. The use of a fuzzy automaton with ε-moves (FA_ε), as described in [15], allows obtaining correct tag clustering rates greater than 95% when considering large data sets.

The main contribution of this paper is the introduction of a discriminator that combines a syntactic similarity measure based in a fuzzy automaton with ε-moves (FA_ε), and a cosine relatedness measure. This combination improves significantly the performance of the syntactic variations detection, even when considering short length tags (lower or equal than three symbols).

The rest of the paper is organized as follows: section 2 describes the tag clustering process; section 3 describes the experimental scenario and the results obtained; and finally, conclusions and references end the paper.

2 Syntactic Variations Clustering

Folksonomies can be described following different approaches [6,8,17]. A folksonomy F can be defined as a tuple: $F=(U,R,T,f_a:U\text{x}R\text{x}T\text{x}\ldots\text{x}T)$, where U, R and T are respectively the finite sets of users, resources and tags defined in the folksonomy; and where the annotation relation f_a relates a user, with a resource and with the set of tags employed by the user to annotate the resource. In order to represent the syntactic variations of tags, this definition must be extended. Then, a folksonomy is defined as a tuple: $F'=(U,R,T,T',f_a:U\text{x}R\text{x}T\text{x}\ldots\text{x}T,f_g:T'\rightarrow T\text{x}\ldots\text{x}T)$. In this model U, R and T keep their meaning and a new set with name T' is used to represent the clustering of T elements, being $T'\subseteq T$. This model allows clustering tag variations included in T in a new set of tags T'. Relation f_a keeps the same meaning, relating a user with a resource and a set of tag variations used to annotate the resource by the user. Function f_g represents the relation between T' groups of tags and T tags variations.

A fuzzy finite state automaton with ε-moves FA_ε, is a sixtuple $(Q,\Sigma,\mu,\mu_\varepsilon,\sigma,\eta)$ where Q is a non-empty finite set of states; Σ is a non-empty finite set of input symbols (*input alphabet*) where Σ^+ is the set of all non-empty strings over Σ, and $\Sigma^* = \Sigma+U\{\varepsilon\}$; $\mu: Q\text{x}Q\text{x}\Sigma \rightarrow [0,1]$ is the state transition function; σ and η are fuzzy sets on Q; and μ_ε is a reflexive binary fuzzy relation on Q representing the state transition function by empty string. For q, $p\in Q$ and $x\in\Sigma$, the value $\mu(q,p,x)\in [0,1]$ represents the degree to which the automaton in state q and with the input symbol x may enter to state p. For $q\in Q$, σ (q) indicates the degree to which q is an initial state, and $\eta(q)$ indicates the degree to which q is a final state.

(1) $\hat{\mu}: \Im(Q)x\Sigma \rightarrow \Im(Q)$

(2) $\hat{\mu}_\varepsilon : \Im(Q) \rightarrow \Im(Q)$ is the fuzzy state transition function by empty string. Given a fuzzy state $V \in \Im(Q), \hat{\mu}_\varepsilon(V) = V \circ \mu_\varepsilon^T$, where μ_ε^T is the T-transitive closure of μ_ε.

(3) $\mu^* : \Im(Q)x\Sigma^* \rightarrow \Im(Q)$ is the extended transition function for the fuzzy finite state automaton with ε-moves. It is defined as follows:

a) $\mu^*(V,\varepsilon) = \hat{\mu}_\varepsilon(V) = V \circ \mu_\varepsilon^T, \forall\, V \in \Im(Q)$

b) $\mu^*(V,\alpha x) = \hat{\mu}_\varepsilon(\hat{\mu}(\mu^*(V,\alpha),x)) = (\mu^*(V,\alpha) \circ \mu[x]) \circ \mu_\varepsilon^T, \forall\, V \in \Im(Q), \alpha \in \Sigma^*, \text{and } x \in \Sigma$

The language accepted by FA_ε, denoted $L(FA_\varepsilon)$, is the fuzzy set on Σ^* such that $L(FA_\varepsilon)(\alpha) = \max_{q\in Q}\left(\mu^*(\sigma,\alpha)(q) \circ \eta(q)\right), \forall\, \alpha \in \Sigma$.

Two tags co-occur if both of them are used by a user to annotate a certain resource. Given a folksonomy $F=(U,R,T,f_a:U\text{x}R\text{x}T\text{x}\ldots\text{x}T)$, we can define a co-occurrence graph as a weighted undirected graph whose set of vertices is the set T of tags. Two tags t_1 and t_2 are connected by an edge, if there exists at least one annotation with a f_a relation corresponding to user u, resource r, containing both tags. The weight of an edge $w(t_1,t_2)$ is determined by the number of co-ocurrences of the two tags connected (t_1 and t_2). According to co-occurrences, each tag t can be encoded as a weights vector $v_t \in R^T$ where each position is associated to a tag t' and whose value is determined by the weight of the co-occurrences between both tags (t and t'): $v_{t,t'} = w(t, t')\ \forall\, t \neq t' \in T$, where $v_{tt} = 0$. This encoding allows measuring the semantic similarity between to

tags using the cosine measure. Given two tags t_1 y t_2 represented by $v_1, v_2 \in R^T$, their cosine similarity is defined as:

$$Similarity\ (t_1, t_2) := \cos(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1| \cdot |v_2|} \quad (1)$$

When comparing two tags encoded following their co-occurrences vectors, the measure provided supplies values in the closed interval [0,1] representing the existing angle between both vectors (v_1 and v_2). This measure is independent of the tag length, and its value is 0 whenever both vectors are orthogonal and 1 when both vectors have the same direction.

In [15] we proposed a method that allowed the classification of tags (containing syntactic variations) based on a discriminator which computed similarity measures among a candidate tag and a set of pattern tags contained in a dictionary. The main drawback of similarity measures based on dictionary comparisons is their poor performance when considering short length chains. The proposed fuzzy similarity based on fuzzy automata with ε-moves FA_ε provides low recognition rates whenever the tag lengths are lower or equal than three symbols. In a folksonomy, a syntactic variation in a short length tag (e.g.: *cut* is transformed in *cat*) can imply a great impact in the meaning represented by this tag. In order to deal with syntactic variation of tags grant an adequate clustering, we propose the use of the cosine measure to increase the reliability of the fuzzy similarity when dealing with short length tags. The main problem is how to identify if a tag is a syntactic variation of a pre-existing tag or not. The cosine measure allows identifying if a candidate tag is semantically similar to a pattern tag. Cosine measure allows discriminating a great number of false positives that fuzzy similarity measures can introduce when dealing with short length tags.

In such way, we propose to assist the discriminator with the cosine relatedness similarity measure between tags. Figure 1 describes the process followed by a new candidate tag that is introduced in the system for the first time. The discriminator computes the fuzzy syntactic similarity and the cosine measure among the observed tag and the set of existing tags stored in a dictionary. The occurrence of a new tag not included in the dictionary implies a clustering process. If the discriminator identifies

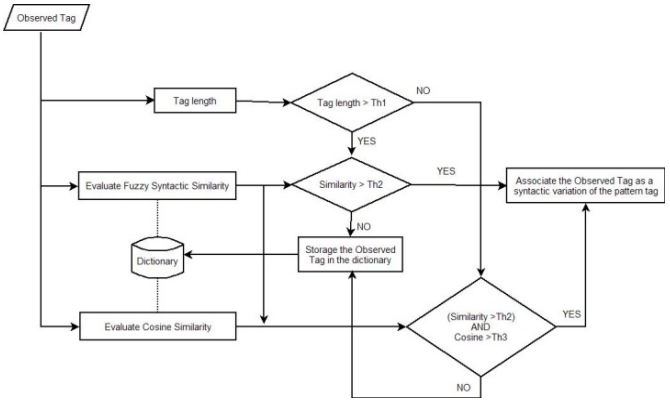

Fig. 1. Syntactic tag variation discrimination, flow diagram

the tag as a syntactic variation of an existing tag, it assigns this new tag to the cluster whose cluster-head is the pattern tag with the higher similarity value (pattern). According to the tag length, the discriminator uses the fuzzy similarity or the fuzzy & cosine similarities. Thresholds Th_1, Th_2 and Th_3 represent the tag length threshold, the fuzzy similarity threshold and the cosine threshold, respectively. Whenever the tag length is greater than Th_1, the discriminator uses the fuzzy similarity measure for the tag clustering process. In other case, the cosine measure is also considered by the discriminator in conjunction with the fuzzy similarity measure. If both, fuzzy and cosine measures provided values greater than Th_1 and Th_2 respectively, then the discriminator identifies the tag as a variation of a certain pattern tag, and performs the tag clustering according to this result. When fuzzy and cosine measures do not agree (values lower than thresholds) the discriminator includes the tag in the dictionary.

3 Experimental Results

In order to evaluate our proposal, we have collected data from the social web *del.icio.us* during the first weak of the year 2009[1], collecting 2,296,300 annotations. Each annotation consists on a tag assigned by a user to a resource, on a given date. We have obtained the 1,000 tags more widely used among the set of annotations. Although these tags only are the 0.64% of the total set of tags (a very small sample size), they represent: (i) the 66.92% of the total set of annotations, (ii) the 78.24% of the set of resources and (iii) the 87.85% of the total sets of users. We have analyzed one by one the 1,000 tags (with fuzzy similarity and cosine measures) adding them to the dictionary (initially empty) when the discriminator identifies them as new tags, and clustering them when the discriminator identifies them as syntactic variations of existing ones. A first experiment focuses on the selection of the adequate threshold values for the hybrid method (fuzzy and cosine). A second experiment focuses on the hybrid method (fuzzy similarity and cosine measures) validation. Figure 2 represents the tag length distribution for the initial set and for the subset of 1,000 annotations more frequently used, respectively. The rate of occurrences of small length tags (lower or equal than five symbols) is near a 15% for the initial set of tags, and increases to 35% when considering more frequently used subset.

The fuzzy similarity measure provides good clustering rates of tag including syntactic variations [15]. Considering the related set of 1,000 annotations, the fuzzy similarity measure provides a correct classified rate (*OK*) of 91.4% for a threshold value of 0.0003. To improve this rate, mainly for short length tags, we analyze the threshold values concerning the cosine measure. Figure 3 (left) shows the correct clustering rate obtained for different threshold values. Better results are obtained for a threshold of 0.7, obtaining a correct clustering rate of 95.5%. As Figure 2 (right) shows, the hybrid method improves the results provided by the fuzzy similarity even if the cosine threshold is not selected properly. The threshold considered by the hybrid method (see Figure 4) determines the weight assigned by the hybrid method to the fuzzy similarity according to the length of the tag considered. The goal of the hybrid method is to improve the correct clustering rate provided by the fuzzy

[1] http://www.eslomas.com/index.php/publicaciones/tagsvariationscombinedmethod

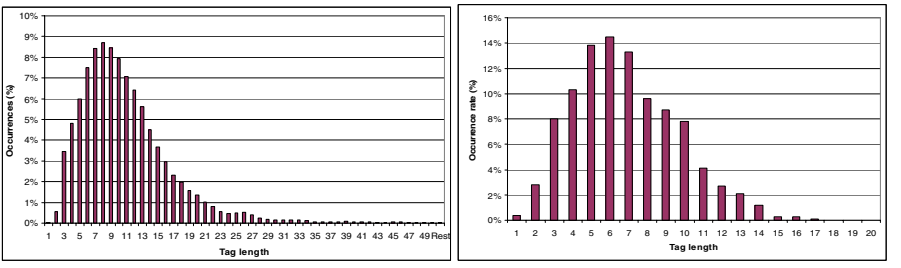

Fig. 2. Tag length distribution for the initial (left) and experimental (right) sets respectively

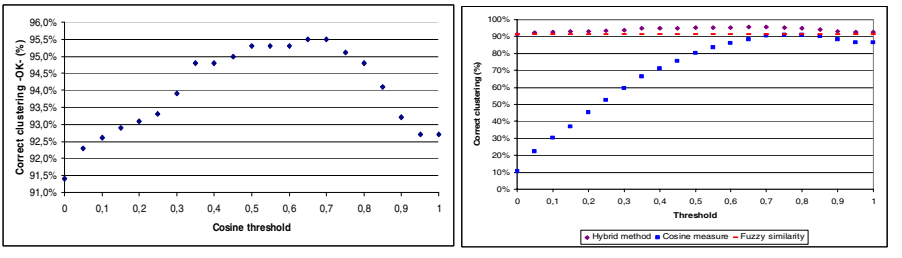

Fig. 3. Threshold selection for the cosine measure (left) and correct clustering rates (right)

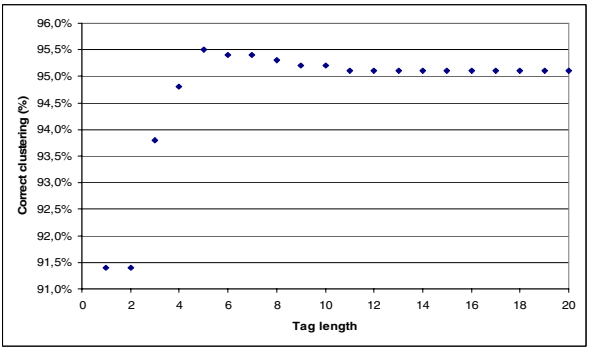

Fig. 4. Hybrid method threshold referred to the tag lengths

similarity when dealing with short tag length. The 91.4% of recognition rate provided by fuzzy similarity increases notably when dealing with tag lengths in the interval [3,10]. Tags with lengths lower than three symbols still provide worse results. A syntactic variation of a symbol often implies a semantic change.

In order to validate the hybrid method, we consider thresholds of 0.0003 and 0.7 for the fuzzy similarity and cosine measures, respectively. These values have been obtained experimentally as described above. In the same way, the threshold value fixed for the hybrid method is five, in order to improve the clustering of tags with lengths of three, four and five symbols. Figure 5 shows the tag clustering results obtained for the experimental subset of 1,000 annotations. Label OK represents the

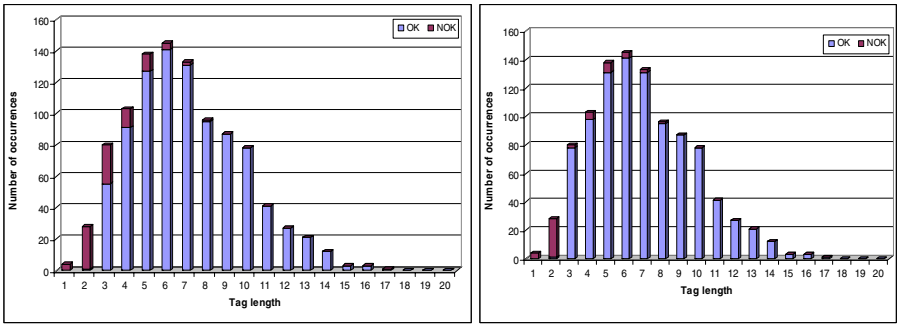

Fig. 5. Tag clustering results provided by the hybrid method for the experimental subset of 1,000 annotations

number of tags correctly grouped using the hybrid method (see Section 2). Label NOK represents the number of misclassified tags. The correct clustering rate (*OK*) has been obtained by comparing carefully one to one all tags. The reduced subset of tags (a thousand) makes possible this comparison. Figure 5 (left) shows the results obtained when only considering the fuzzy similarity measure, and (right) shows the results obtained when considering the hybrid method. The hybrid method improves notably the clustering rates when considering tag lengths between three and five.

4 Conclusions

In this work, we propose a hybrid method to cluster tags using a fuzzy similarity and a cosine measures. The fuzzy similarity discovers syntactic variations of tags allowing the clustering of tags. The cosine measure allows improving the clustering process when dealing with short length tags. A syntactic variation in a short length tag often implies a change in the meaning of the tag, and the cosine measure allows discovering if that occurs. A high cosine similarity value in a short length tag indicates that this tag is a syntactic variation of an existing one, while a low cosine value indicates that this tag must be considered as a new tag. We tune the threshold values and analyze the clustering rates obtaining that the hybrid method improves the tag clustering process when considering tag lengths lower or equal than five symbols.

Acknowledgements

Research partially supported by the Spanish Research Council under research grants TIN2006-14738-C02-02 and TIN2008-03687.

References

1. Echarte, F., Astrain, J.J., Córdoba, A., Villadangos, J.: Ontology of Folksonomy: A New Modeling Method. In: Semantic Authoring, Annotation and Knowledge Markup, Whistler, British Columbia, Canada (2007)

2. Mathes, A.: Folksonomies - Cooperative Classification and Communication Throught Shared Metadata. Computer Mediated Communication (2004)
3. Guy, M., Tonkin, E.: Folksonomies - Tidying up Tags? DLib Magazine 12(1) (2006)
4. Kaser, O., Lemire, D.: TagCloud Drawing: Algorithms for Cloud Visualization. In: Work. Taggings and Metadata for Social Information Organization, Banff, Alberta, Canada (2007)
5. Hassan-Montero, Y., Herrero-Solana, V.: Improving tag-clouds as visual information retrieval interfaces. In: International Conference on Multidisciplinary Information Sciences and Technologies, Mérida, Spain (2006)
6. Heymann, P., García-Molina, H., Collaborative Creation of Communal Hierarchical, Taxonomies in Social Tagging Systems. Stanford Info. Lab. Tech. Report 2006-10 (2006)
7. Zhou, M., Bao, S., Wu, X., Yu, Y.: An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 680–693. Springer, Heidelberg (2007)
8. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
9. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the Semantic Web: Collaborative Tag Suggestions. In: Workshop on Collaborative Web tagging, Edinburgh, Scotland (2006)
10. Passant, A.: Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs: Theoretical background and corporate use-case. In: International Conference on Weblogs and Social Media, Boulder, USA (2007)
11. Gruber, T.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2), 199–220 (1993)
12. Shepitsen, A., Gemmel, J., Mobasher, B., Burke, R.: Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering. In: 2nd ACM Conference on Recommender Systems, Lausanne, Switzerland, pp. 259–266 (2008)
13. Specia, L., Motta, E.: Integrating Folksonomies with the Semantic Web. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 624–639. Springer, Heidelberg (2007)
14. Kim, H.L., Scerri, S., Breslin, J.G., Decker, S., Kim, H.G.: The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies. In: 8th Int. Conference on Dublin Core and Metadata Applications, Berlin, Germany, pp. 128–137 (2008)
15. Echarte, F., Astrain, J.J., Córdoba, A., Villadangos, J.: Improving Folksonomies Quality by Syntactic Tag Variations Grouping. In: 24th Annual ACM Symposium on Applied Computing, Honolulu, USA, vol. 2, pp. 2016–2020 (2009)
16. Navarro, G.: A Guided Tour to Approximate String Matching. ACM Computing Surveys 33(1), 31–88 (2001)
17. Echarte, F., Astrain, J.J., Córdoba, A., Villadangos, J.: Pattern Matching Techniques to Identify Syntactic Variations of Tags in Folksonomies. In: Lytras, M.D., Carroll, J.M., Damiani, E., Tennyson, R.D. (eds.) WSKS 2008. LNCS (LNAI), vol. 5288, pp. 557–564. Springer, Heidelberg (2008)