An Algorithm for the Improvement of Tag-based Social Interest Discovery

José Javier Astrain, Alberto Córdoba, Francisco Echarte, Jesús Villadangos
Dept. Ingeniería Matemática e Informática
Universidad Pública de Navarra
Campus de Arrosadía. 31006 Pamplona, Spain
Email: {josej.astrain,alberto.cordoba}@unavarra.es, patxi@eslomas.com, jesusv@unavarra.es

***Abstract*—The success of Web 2.0 has generated many interesting and challenging problems as the discovering of social interests shared by groups of users. The main problem consists on discovering and representing the interest of the users. In this paper, we propose a fuzzy based algorithm that improves the Internet Social Interest Discovery algorithm. This algorithm discovers the common user interests and clusters users and their saved resources by different interest topics. The collaborative nature of social network systems and their flexibility for tagging, produce frequently multiple variations of a same tag. We group syntactic variations of tags using a similarity measure improving the quality of the results provided by the Internet Social Interest Discovery algorithm.**

***Keywords*-Social interest discovering, syntactic variations, collaborative tagging systems**

I. INTRODUCTION

Nowadays, one of the problems of social networks in Web 2.0 is the discovering of common interests shared by user communities. Users of a community use to have same interests. In this sense, social communities growth is limited by the definition of scalable and well adapted communities to user interests.

The discovering of social interests shared by groups of users can be focused following three different approaches. The user-centric approach focuses on detecting social interests based on the social connectivity among users [1], [2]. Those works analyse user's social or on-line connections to discover users with particular interests or expertise for a given user. Recent works, as [3], [4] represent the three types of entities that exist in a social tagging system (users, items and tags) by a 3-order tensor, on which latent semantic analysis and dimensionality reduction is performed using both the Higher Order Singular Value Decomposition (HOSVD) method and the Kernel-SVD smoothing technique. In the object-centric approach, [5], [6] explored the common interest among user based on common objects they fetched in peer-to-peer networks. However, without other information of the objects, it cannot differentiate the various social interests on the same object. Furthermore, in Internet social networks such as del.icio.us, most of objects are unpopular. Thus, it is difficult to discover common interest of users on them [7]. The tag-centric [7], [8], [9] approach focuses on directly detecting social interests or topics analysing user annotations. This approach avoids the limitation of object-centric approach [7]. In [10], a tag-centric approach is used to provide semantic resource classification.

Tagging techniques have been widely used in many different social networks. As introduced in [8], the proportion of frequencies of tags within a given site tend to stabilize with time (due to the collaborative tagging by all users). Furthermore, the distribution of frequency of tags for popular sites follows the power law as proved in [11]. This reinforces the need to discover the interests of users since although they use automatic tagging systems, they do so using an uncontrolled vocabulary.

Resource classification can be performed by using clustering techniques using both keywords [12] and tags [7], [10], [13]. An internet social discovery system (ISID) is developed in [7], which cluster users and their saved URLs based on their annotations. Although users may have different interests for an item (and items may have multiple facets) the fact is that tags implicitly describe the users' interests. We discover common interest shared by groups of users in social networks by utilizing user tags. Our approach is based on the insightful study and observation on the user generated tags in social networks systems such as del.icio.us [7]. As users annotate resources, the occurrence of common tags reinforce their common interests.

A same resource can receive different tag annotations from different users; then we consider that a resource has converged when its distribution of tags converges rapidly to a remarkably stable heavy-tailed distribution. Although an increase on the number of annotations also includes an increase on the number of different tags involved in the annotation process, we can observe that most of users agree on the more relevant tags. Those tags are used in a great number of annotations, and by a great number of users. So, a quantification of this agreement degree aids to define a certain threshold in charge of identifying resource convergence. The set of aggregated user tags on a resource is quite compact and stable enough to characterize the same main resource.

One of the main problems of the tag-centric approach is the existence of a high number of syntactic variations (erroneous or not) of other existing tags. A pre-filtering of the tags, as occurs in [14] where the Levenshtein similarity

 ISBN: 978-1-61208-104-5

measure is used to reduce the number of tags (identifying syntactic variations), allows increasing the quality of tag clustering minimizing the effects of syntactic variations. In previous works, we proved that the utilization of a fuzzy algorithm (FA_ε) [13] provides best classification results than the obtained when using classical distances as the Levenshtein and Hamming distances; and in [15], we improved those results adding a semantic measure (cosine) to the fuzzy automaton. Cosine similarity, traditionally used in information retrieval [16], measures the similarity between a couple of vectors of n dimensions by finding the cosine of the angle between them. The semantic similarity is obtained comparing the vectorial representation of a couple of terms. In this paper, we present the Fuzzy based Internet Social Discovery algorithm (F_b-ISID), which increases the discovery results provided by ISID [7] by using the fuzzy automaton with ε-moves in conjunction with the cosine similarity to remove the syntactic variations of tags on a folksonomy (*tag cleaning*). The good results obtained show the convenience of re-clustering the tags in order to remove the syntactic variation of tags. The tags containing syntactic variations are clustered in their representative tags preserving their semantic information and reinforcing the tag relevance in the whole set of tags.

The tag cleaning process improves the interest discovering results obtained. Finally, we consider that the appliance of a tag cleaning process must be performed for all the algorithms, which use the related tag-centric approach.

The rest of the paper is organized as follows: Section II describes the F_b-ISID algorithm; Section III describes the experimental results obtained; and finally, conclusions, acknowledgements and bibliographical references end the paper.

II. F_b-ISID DESCRIPTION

In order to deal with the large amount of syntactic variations of tags usually existent in folksonomies, we present the *Fuzzy based-ISID* (F_b-ISID) method. F_b-ISID is based on the pre-filtering of the posts with the aim of increasing the interest discovering results obtained by ISID. For such purpose, F_b-ISID clusters the syntactic variations of tags reducing the entropy of the posts by means of the fuzzy and cosine similarity measures above described. F_b-ISID improves the search of topics of interest against ISID introducing a new component Syntactic Variations which is in charge of the elimination of syntactic variations on tag-centric systems. This section is devoted to describe the components of the F_b-ISID algorithm. The main characteristic of F_b-ISID consists on the introduction of a component, called *SyntacticVariation*, which avoids the syntactic variations of the posts.

The F_b-ISID architecture provides functions as finding topics of interests, resource clustering, and topics of interest indexing.

1) *Finding topics of interest*. For a given set of bookmark post (Bookmark Post is a social bookmarking site alowwing users to submit their blog post and other stories to share with others and make them popular), find all topics of interest. Each topic of interests is a set of tags with the number of their co-occurrences exceeding a given threshold. Those sets of tags, which do not reach this threshold do not give rise to a new topic of interests;
2) *Clustering*. For each topics of interests, find the URLs and the users such that those users have labelled each of the URLs with all the tags in the topic. For each topic, a user cluster and a URL cluster are generated;
3) *Indexing*. Import the topics of interests and their user and URL clusters into an indexing system for application queries.

The components of this architecture are: DATASOURCE, SYNTACTICVARIATION, TOPICDISCOVERY, CLUSTERING and INDEXING.

1) *DataSource*: F_b-ISID inputs are users' posts obtained from social networks as a stream of posts p = *(user,URL, tags)*, where the combination of *user* and *URL* uniquely identifies a post p, and *tags* is the set of tags that the user uses to label the referred URL.
2) *SyntacticVariation*: the discriminator included in this component is in charge of grouping syntactic variations of tags. It computes the fuzzy similarity and the cosine measures among the observed tag and the set of already existing tags (stored in a dictionary) in order to discover syntactic variations of tags. The dictionary includes all the tags that have been used by users in their annotations provided that they are not syntactic variations of other pre-existing tags. The occurrence of a new tag not included in the dictionary implies a clustering process. The identification of a tag as a syntactic variation of an existing tag by the discriminator, implies the assignation of a new tag to the cluster whose cluster-head is the pattern tag with the higher similarity value (pattern). According to the tag lengths, the discriminator calculates the fuzzy similarity or both fuzzy and cosine similarities. Three thresholds Th_1, Th_2 and Th_3, which represent the tag length threshold, the fuzzy similarity threshold and the cosine threshold, respectively, are considered. 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.

 ISBN: 978-1-61208-104-5

When fuzzy and cosine measures do not agree (values lower than thresholds) the discriminator includes the tag in the dictionary.

3) *TopicDiscovery*: this component is in charge of finding the frequent tag patterns for a given set of post. F_b-ISID uses association rules algorithms to identify the frequent tag patterns for the post.
4) *Clustering*: this component collects the posts that contain the tag set (topic), inserting into two collections of clusters (identified by topics) the resources (URLs) and the users of the posts. Its main problem is its complexity, since the algorithm used matches each topic against each post. Then, for a set of n tags, there are 2^n possible topics to check. In order to reduce this complexity, we build a prefix tree over the merged topics. The clustering algorithm for a given set T of topics and a given set P of post is described in Fig. 1.
5) *Indexing*: this component provides some simple query services for applications:
   - For a given topic, listing all URLs that contain this topic.
   - For a given topic, listing all users that are interested in this topic.
   - For given tags, listing all topics containing the tags.
   - For a given URL, listing all topics that are concerned this URL.
   - For a given URL, and a topic, listing all users that are interested in this topic and have saved the URL.

1: **for all** topic $t \in T$ **do**
2: $t.user \leftarrow \emptyset$
3: $t.url \leftarrow \emptyset$
4: **end for**
5: **for all** post $p \in P$ **do**
6: **for all** topic t of p **do**
7: $t.user \leftarrow t.user \cup p.user$
8: $t.url \leftarrow t.url \cup p.url$
9: **end for**
10: **end for**

Figure 1. F_b-ISID clustering algorithm.

Figure 2 illustrates the F_b-ISID architecture, where the *Syntactic variations* function is added to the ISID architecture. The posts obtained by *DataSource* are processed by *SyntacticVariation*, which clusters the posts avoiding the syntactic variations of tags. The resultant posts (Posts') are then processed by *TopicDiscovery*, which provides the topics of interest. The *Clustering* component clusters these topics and provide the results to the *Indexing* component.

In [13] we proposed a method to group syntactic variations of tags using pattern matching techniques. The aim

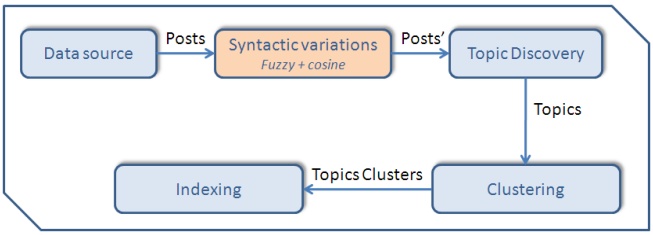

Figure 2. F_b-ISID architecture.

is to cluster in a single centroid all the tags that can be considered as a syntactic variation of a given tag. This centroid represents all the tags included in this cluster, which are syntactic variations of it. In particular, the proposed fuzzy similarity measure (a fuzzy automaton with ε-moves, FA_ε) offers better classification results than other classic techniques (Hamming and Levenshtein measures) after comparing them over a large real dataset. The identification of syntactic variations depends on the length of the tags. Similarity measures perform well for tag lengths equal or greater than five symbols, providing poor results in other cases. In [15] we proposed an hybrid method which adds to the related fuzzy similarity measures a cosine measure in order to improve the clustering process when dealing with short length tags. The use of both cosine and fuzzy similarity measures ensures recognition rates greater than 95% over datasets including large and small length tags. Results have been validated by experts outside the project. By adding the cosine similarity, the tag clustering performed ensures a higher semantic clustering.

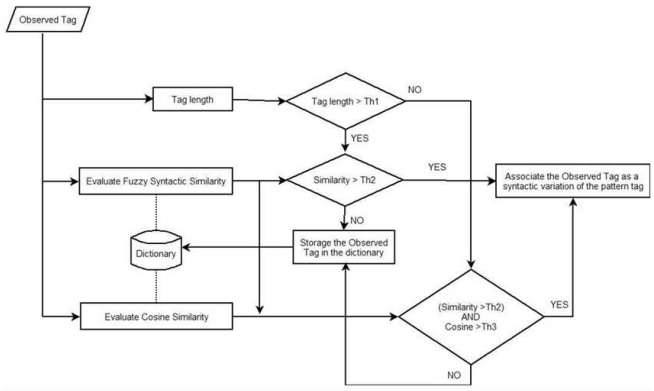

Figure 3. Syntactic variation discovering.

The discriminator used to cluster syntactic variations of tags (see Figure 3) computes the fuzzy similarity and the cosine measures among the observed tag and the set of already existing tags (stored in a dictionary) in order to discover syntactic variations of tags. The occurrence of a new tag not included in the dictionary implies a clustering process. The identification of a tag as a syntactic variation of an existing tag by the discriminator, implies the assignation of

 ISBN: 978-1-61208-104-5

a new tag to the cluster whose cluster-head is the pattern tag with the higher similarity value (pattern). The discriminator uses the fuzzy similarity or the fuzzy and cosine similarities according to the tag length. Three thresholds Th_1, Th_2 and Th_3, which represent the tag length threshold, the fuzzy similarity threshold and the cosine threshold, respectively, are considered. 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.

III. Experimental results

The comparison between F_b-ISID and ISID is performed implementing all the components of both algorithms following the same process and the same presentation of the results than those of [7].

For such purpose, we have retrieved web pages annotated by users from Del.icio.us using its *Recent Bookmarks* page. We consider those resources bookmarked by at least 250 users, storing the URL of those resources. Information has been retrieved during the period from 1-15 March, 2010. A total amount of 419,891 resources, with 2,296,300 annotations, 197,148 users and 156,897 tags have been obtained. We have randomly generated some subsets of posts from a total amount of 779,674 posts. Table I shows the number of tags, users and resources (URLs) for each subset of posts. For example, the subset containing 100,000 posts refers to 36,603 different tags, 44,051 different users and 72,567 different resources. Repetitions are not considered.

TABLE I
Number of different tags, users and URLs for each subset of posts.

Posts	Tags	Users	URLs
5,000	4,546	3,060	4,545
50,000	22,111	23,659	38,984
100,000	36,603	44,051	72,567
200,000	59,424	74,502	134,424
300,000	59,398	99,144	192,462
400,000	97,430	121,498	245,929
500,000	114,358	146,972	293,685
600,000	130,451	166,546	339,875
700,000	145,372	183,542	384,663
779,674	156,897	197,148	419,891

The execution of the *SyntacticVariation* component over the dataset retrieves a total amount of 991 syntactic variations (4.31% of the 779,674 posts) with a recognition rate of the 96.97%, which has been verified manually with the aid of Wordnet and Wikipedia. Table II shows the distribution of the syntactic variations of tags.

TABLE II
Syntactic variations distribution.

Syntactic variation	Occurrence distribution	
	Number	Percentage
Number (singular/plural)	772	77,90
Delimiters	109	10,99
Synonyms	59	5,95
Misclassification	30	3,03
Other	21	2,12
Total	991	100

Table III shows the number of subsets generated by both algorithms. It can be seen that the number of tag subsets generated by F_b-ISID is lower than the number of tag subsets generated by ISID. We observe how F_b-ISID (347,985,324) obtains a 24.10% less of subset tags than ISID (458,452,178). The suppression of syntactic tag variations causes the clustering of those concepts scattered in many different terms. That allows the apparition of new topics of interest since the new subset of tags reach the threshold required to become a topic of interest.

TABLE III
Number of tag subsets.

Posts	Number of subsets		Variation
	ISID	F_b-ISID	(%)
5,000	6,590,055	5,602,342	14,99
50,000	60,905,524	50,694,270	16,77
100,000	113,015,128	95,877,191	15,16
200,000	203,577,716	173,393,490	14,83
300,000	283,734,103	245,137,336	13,60
400,000	338,641,100	256,157,047	24,36
500,000	340,920,037	309,822,299	9,12
600,000	408,342,102	323,623,856	20,75
700,000	437,792,108	336,625,280	23,11
779,674	458,452,178	347,985,324	24,10

A comparative study between F_b-ISID and ISID is presented in Table IV, which shows the results obtained for the grouping of contents: Topics & Users, Topics & URLs, Users & URLs, Topics of interests, Users and URLs.

1) Topics & Users: F_b-ISID improves the classification a 11.16%.
2) Topics & URLs: F_b-ISID improves the classification a 11.40%
3) Topics of interests: F_b-ISID improves the classification a 20.15%
4) Users & URLs: similar results for both F_b-ISID and ISID (0.04%).
5) Users: similar results for both F_b-ISID and ISID (0.07%).
6) URLs: similar results for both F_b-ISID and ISID (0.06%).

 ISBN: 978-1-61208-104-5

TABLE IV
GROUPING: TOPICS, USERS AND URLS FOR EACH SUBSET OF POSTS.

Number of posts	ISID					
	Topics & Users	Topics & URLs	Users & URLs	Topics	Users	URLs
5,000	4,517	5,562	3,034	92	1,959	2,726
50,000	98,334	107,774	39,375	1,250	19,558	30,167
100,000	261,435	268,875	82,166	3,140	37,915	58,671
200,000	637,717	632,257	170,052	7,331	66,314	112,560
300,000	1,054,851	1,042,349	258,396	11,277	89,478	163,302
400,000	1,498,064	1,447,532	347,008	14,574	110,577	210,256
500,000	1,914,231	1,779,109	433,676	14,836	133,972	250,608
600,000	2,378,987	1,856,432	554,765	16,786	145,786	287,654
700,000	2,944,428	2,624,802	612,216	20,567	169,052	330,955
779,674	3,014,563	2,765,498	686,056	22,012	172,987	348,765
Number of posts	F_b-ISID					
	Topics & Users	Topics & URLs	Users & URLs	Topics	Users	URLs
5,000	4,822	5,890	3,108	100	2,001	2,782
50,000	110,045	119,629	39,579	1,393	19,642	30,323
100,000	294,191	298,999	82,202	3,483	37,923	58,697
200,000	731,529	714,237	170,089	8,402	66,313	112,589
300,000	1,219,574	1,188,299	258,740	13,243	89,538	163,566
400,000	1,574,956	1,652,038	345,351	14,701	110,269	209,100
500,000	2,262,048	2,059,656	434,885	18,735	134,220	251,424
600,000	2,528,314	2,197,279	559,234	24,765	145,876	288.765
700,000	2,972,815	2,630,037	604,828	15,837	169,806	331,045
779,674	3,393,105	3,121,478	686,056	27,566	173,102	348,964

TABLE V
INDEXING: "BLOG" AND "HTTP://ANIMOTO.COM".

Number of posts	ISID					F_b-ISID				
	blog			animoto		blog			animoto	
	URLs	Users	Topics	Topics	Users	URLs	Users	Topics	Topics	Users
5,000	124	106	1	5	0	164	137	1	7	0
50,000	1,213	920	26	53	1	1,513	1,200	46	57	2
100,000	2,244	1,818	60	88	5	2,845	2,361	107	96	7
200,000	4,163	3,344	159	232	8	5,301	4,329	274	251	14
300,000	6,189	4,866	257	297	14	7,761	6,235	463	312	25
400,000	8,042	6,357	361	377	27	10,078	8,134	513	351	40
500,000	9,848	8,015	364	472	32	12,329	10,248	693	498	49
600,000	11,461	9,467	424	502	35	13,564	12,087	484	521	55
700,000	13,257	10,937	504	593	44	16,539	14,363	581	613	63
779,674	14,573	12,146	564	626	50	18,323	15,315	629	643	71

One can note that F_b-ISID provides best classification results when grouping results by topic, while classification results remain unchanged when only considering User and URL clustering.

The topic *blog* and the URL *http://animoto.com* are used to build the following basic queries:

1) For the topic *blog*, list all the URLs associated with the tag *blog*.
2) For the topic *blog*, list all users that are interested in this topic.
3) For tag *blog*, list all topics containing the tag *blog*.
4) For the URL *http://animoto.com*, list all the topics containing the resource *animoto*.
5) For the URL *http://animoto.com* and the topic *blog*, list all the users interested in the topic *blog* that have saved the URL *animoto*.

Table V shows the results obtained for the subsets of posts, namely URLs, Users and Topics for *blog*, and Topics and Users for *http://animoto.com*. The results obtained show that F_b-ISID:

a) obtains a 25,73% of URLs containing the topic *blog*;
b) increases the number of users interested in *blog* (29,05%);
c) increases the number of topics containing *blog* (11,52%);
d) increases the number of topics that are related to *http://animoto.com* (4,10%);
e) increases the number of users interested in *blog*, which use *http://animoto.com* (23,44%).

Fig. 4 shows that F_b-ISID provides best results for the number of URLs and Users related with the topic *blog* for each of the post sets.

 ISBN: 978-1-61208-104-5

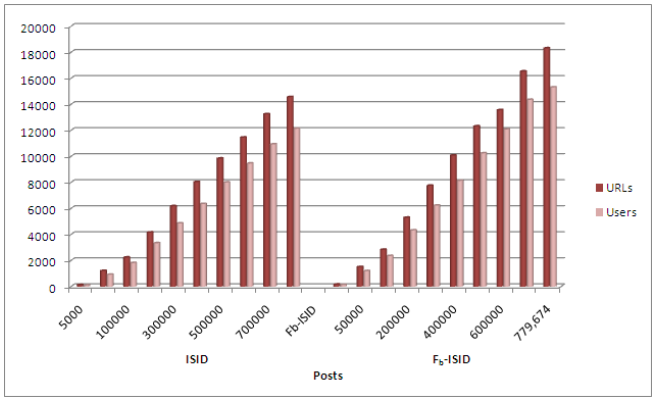

Figure 4. Comparative of URLs and users related with the topic blog.

IV. Conclusions

In this paper, we have introduced a new algorithm, named F_b-ISID, for the discovering of the interest of users in collaborative tag-based systems. The experiments performed show that F_b-ISID obtains better results than the ISID algorithm. This good behaviour is due to the fact that the architecture of F_b-ISID contains one component not included in ISID. This component filters or groups together syntactic variations of the tags contained in the initial posts. In this way, F_b-ISID obtains more topics of interests and performs basic queries more efficiently than ISID. Finally, we consider that the clustering of syntactic variation in the data sources of social systems, improves the performance of algorithms for interests discovery, based on tag-centric approaches.

Acknowledgments

Research partially supported by the Spanish Research Council under research grants TIN2008-03687 and TIMI-2006-CENIT-4834.

References

[1] M. F. Schwartz and D. C. M. Wood, "Discovering shared interests using graph analysis," *Commun. ACM*, vol. 36, no. 8, pp. 78–89, 1993.

[2] N. Ali-Hasan and L. A. Adamic, "Expressing social relationships on the blog through links and comments," in *International Conference on Weblogs and Social Media (ICWSM)*, 2007. [Online]. Available: http://www.icwsm.org/papers/2–Ali-Hasan–Adamic.pdf

[3] P. Symeonidis, "User recommendations based on tensor dimensionality reduction," in *Artificial Intelligence Applications and Innovations III*, ser. IFIP Advances in Information and Communication Technology. Springer Boston, 2009, vol. 296, no. 296, pp. 331–340. [Online]. Available: http://dx.doi.org/10.1007/978-1-4419-0221-4-39

[4] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, "A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis," *IEEE Transactions on Knowledge and Data Engineering*, no. To appear, 2010.

[5] K. Sripanidkulchai, B. M. Maggs, and H. Zhang, "Efficient content location using interest-based locality in peer-to-peer systems," in *INFOCOM'2003: 22nd Annual Joint Conference of the IEEE Computer and Communications Societies*. IEEE, march 2003, pp. 2166–2176.

[6] L. Guo, S. Jiang, L. Xiao, and X. Zhang, "Fast and low-cost search schemes by exploiting localities in p2p networks," *J. Parallel Distrib. Comput.*, vol. 65, no. 6, pp. 729–742, 2005.

[7] X. Li, L. Guo, and Y. E. Zhao, "Tag-based social interest discovery," in *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 675–684.

[8] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *J. Inf. Sci.*, vol. 32, no. 2, pp. 198–208, April 2006. [Online]. Available: http://dx.doi.org/10.1177/0165551506062337

[9] Z. Yin, R. Li, Q. Mei, and J. Han, "Exploring social tagging graph for web object classification," in *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2009, pp. 957–966.

[10] F. Echarte, J. J. Astrain, A. Córdoba, J. Villadangos, and A. Labat, "Acoar: a method for the automatic classification of annotated resources," in *K-CAP '09: Proceedings of the fifth international conference on Knowledge capture*. New York, NY, USA: ACM, 2009, pp. 181–182.

[11] H. Halpin, V. Robu, and H. Shepherd, "The complex dynamics of collaborative tagging," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 211–220.

[12] C. H. Brooks and N. Montanez, "Improved annotation of the blogosphere via autotagging and hierarchical clustering," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM, 2006, pp. 625–632.

[13] F. Echarte, J. J. Astrain, A. Córdoba, and J. Villadangos, "Improving folksonomies quality by syntactic tag variations grouping," in *SAC '09: Proceedings of the 2009 ACM symposium on Applied Computing*. New York, NY, USA: ACM, 2009, pp. 1226–1230.

[14] L. Specia and E. Motta, "Integrating folksonomies with the semantic web," in *ESWC '07: Proceedings of the 4th European conference on The Semantic Web*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 624–639.

[15] J. J. Astrain, F. Echarte, A. Córdoba, and J. Villadangos, "A tag clustering method to deal with syntactic variations on collaborative social networks," in *ICWE'09: Proceedings of the 9th International Conference on Web Engineering*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 434–441.

[16] A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani, "Algorithmic detection of semantic similarity," in *WWW'2005: Proceedings of the 14th international conference on World Wide Web*. New York, NY, USA: ACM, 2005, pp. 107–116.

 ISBN: 978-1-61208-104-5