CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2012; **24**:2193–2206
Published online 28 June 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.2892

SPECIAL ISSUE PAPER

One-class support vector machines for personalized tag-based resource classification in social bookmarking systems

Daniela Godoy*,†

ISISTAN Research Institute and CONICET, Argentina CP 7000, Tandil, Buenos Aires, Argentina

SUMMARY

Social tagging systems allow users to easily create, organize, and share collections of Web resources in a collaborative fashion. Videos, pictures, research papers, and Web pages are shared and annotated in sites such as *Del.icio.us*, *CiteULike*, or *Flickr*, among others. The rising popularity of these systems leads to a constant increase in the number of users actively publishing and annotating resources and, consequently, an exponential growth in the amount of data contained in their folksonomies, the underlying data structure of tagging systems. In turn, the user task of discovering interesting resources becomes more and more difficult and time-consuming. In this paper, the problem of filtering resources from social tagging systems according to individual user interests using purely tagging data is studied. One-class support vector machine classification is evaluated as a means to identify relevant information for users based exclusively on positive examples of their information preferences. It is assumed that users express their interest on resources belonging to a folksonomy by assigning tags to them, whereas there is no straightforward method to collect uninterestingness judgments. Filtering interesting resources based on social tags is an important benefit of exploiting the collective knowledge generated by tagging activities of Web communities. In this paper, the results achieved with tag-based classification are compared with those obtained using more traditional information sources such as the full text of Web pages. Experimental evaluation showed that tag-based classifiers outperformed those learned using the text of documents as well as other content-related sources. Moreover, tag-based classification becomes essential for folksonomies in which no additional content is available because of the nature of resources being stored (e.g., tagging of photos or videos). Copyright © 2012 John Wiley & Sons, Ltd.

Received 12 October 2010; Revised 22 July 2011; Accepted 4 June 2012

KEY WORDS: social tagging systems; one-class classification; social media search; folksonomies

1. INTRODUCTION

Social tagging systems are a novel mechanism to share and classify resources, which has recently reached a great popularity over the Web. In sites such as *Del.icio.us*[‡], *Flickr*[§] or *CiteULike*[¶], users annotate a variety of Web resources, including pages, blog posts, videos, or pictures, using a freely chosen set of keywords or open-ended tags. In the simplicity for publishing and annotating content using an uncontrolled vocabulary lies the reason of the widespread success of social or collaborative tagging activity.

*Correspondence to: Daniela Godoy, UNICEN University, Campus Universitario, CP 7000, Tandil, Buenos Aires, Argentina.

†E-mail: daniela.godoy@isistan.unicen.edu.ar

‡http://del.icio.us/

§http://www.flickr.com/

¶http://www.citeulike.org/

From an end-user perspective, the massive amount of generated content collected in folksonomies, the underlying data structure of tagging systems, poses the challenge of effectively finding interesting resources to read as well as filtering information streams coming from social systems. Folksonomies [1] emerged from the collective behavior of users to organize and navigate the massive amount of freely accessible, user contributed, and annotated Web resources. In spite of the novel mechanisms for searching and retrieving resources provided by this social classification scheme, the rapid increase in size of communities using social sites, the volume of published content, and the unsupervised nature of tagging (leading to ambiguous and noisy tags among other problems) make the discovery of relevant resources a time-consuming and difficult task for users.

Recent works have focused on applying recommendation technologies to folksonomies to assist users in selecting appropriate tags to resources and, to a lesser extent, finding relevant information and locating like-minded users [2, 3]. Particularly, tag-based profiling approaches inferring the user preferences for certain tags have been applied to personalized recommendation of resources. In these approaches, vectors of weighted tags or tag networks capturing co-occurrence and semantic relations are used to represent the user interests. However, social tags are expected to capture the collective knowledge of the community about shared contents as oppose to the personal view given by the user's own tags.

In contrast to tag-based profiling approaches, this paper studies the utility of social tags as a source for modeling user interests. It is assumed that users are likely to be interested in additional content annotated with similar social tags to the ones collectively assigned to resources they showed interest in before. Thus, tags associated to resources annotated by a user can be used to build an interest profile that, in turn, can be applied to filter further incoming information from tagging systems (e.g., Really Simple Syndication feeds). In this approach, social tags can be thought of as indicators of user awareness and potential interest in a given resource [4], allowing users to capitalize on the associations made by persons who have assigned similar tags to other resources.

Tag-based classification is based on using the resources users annotate and have in their individual personomies (i.e., the restriction of the folksonomy to a single user) for training personalized classifiers that learn to recognize further potentially interesting resources. Because the tagged resource collection of a given user contains only positive examples of the user interests, this is a special case of classification known as one-class classification. This problem consist in determining whether an example (a Web resource in this case) belongs to a target class (interesting or relevant to the user interests) when only positive examples of the target class are given.

The rest of the paper is organized as follows. Section 2 reviews related research about personalization in social tagging systems as well as tag-based classification in general. Section 3 gives a brief overview of one-class classification using support vector machines (SVMs) classifiers. Section 4 summarizes the experimental setting including the dataset description, gathered from *Del.icio.us* bookmarking site, and the evaluation methodology. The empirical analysis carried out to compare content-based and tag-based classification of Web pages based on personomies is presented in Sections 5 and 6, respectively. Empirical findings are discussed in Section 7, and finally, concluding remarks are stated in Section 8.

2. RELATED WORKS

Recent works have focus on the application of recommendation technologies to social tagging systems to assist users in finding relevant information, selecting appropriate tags for resources, or locating like-minded users within a tagging community [2, 3]. Most of these works addressed the problem of tag recommendation as a mechanism to foster the convergence to a shared vocabulary using content-based methods [5, 6], collaborative filtering [7–9], text categorization [10–12], and other strategies [13, 14].

For personalized resource recommendation in folksonomies, tag-based profiling approaches emerged as an alternative to traditional content-based ones. In [15] and [16], a vector of weighted tags is obtained using their frequency of occurrence in the resources, a user annotate and the inverse user frequency, respectively. In a similar way, tags weighted according to the fraction of the user tagging actions that they covered are used in [17] to generate hot lists of recommendations by

DOI: 10.1002/cpe

combining tags and item overlapping in the construction of per-tag common interest networks. Firan *et al.* [18] investigated tag-based user profiles in contrast to more conventional profiles based on song and track usage in the music search portal *Last.fm*. The results showed that tag-based profiles significantly improve the quality of recommendations. Au Yeoung *et al*. [19] proposed an algorithm that performs graph-based clustering over the network of user tagged documents to identify interest topics and extract tag vectors starting from them. In [20], tag clustering is used to group tags with similar meanings as the basis for a personalization algorithm for recommendation in folksonomies. Both users and resources are modeled as weighted tag vectors, and tag clusters are intermediary between them. On the one side, similarity among a user vector and tag clusters allows to understand what tag cluster is relevant for this user. On the other side, vectors describing resources are used to detect relevant resources for a given cluster of tags. Then, the recommendation algorithm uses a tag, a user profile and tag clusters as inputs and produces an ordered set of resources.

Graph representations were also proposed to model the relationships among tags in a user profile. Michlmayr *et al.* [21] compared tag-based profiles consisting of a single vector of weighted tags with graph representations, in which nodes correspond to tags and edges denote co-occurrence or other relationships among them. *Add-A-Tag* [22] algorithm extends this model to include temporal information by updating the weights of edges in the graph using an evaporation technique known from ant algorithms for discrete optimization. The idea of using semantic relationships among tags in tag-based profiles has also been explored in [23], in which the semantic distance between two tags is calculated based on co-occurrence statistics and common sense reasoning.

User profiles in these works model the user preferences in terms of the tags a user employed instead of using social tags attached to resources in the user personomy. In consequence, resource discovery is constrained by the degree of coincidence between the user tags and the tags assigned by other members of the community to these resources. In this research, a user profile models the type of resources a user is interested in exploiting the social tags attached to them, that is, using a collective description of the resource.

The idea of exploiting social tags for supporting individual users in social tagging systems has been explored in some recent works. To build a user profile, the static textual description of items is integrated with a more dynamic one given by tags in [24]. This approach uses *WordNet* synsets for defining a semantic indexing of resources and tags that are disambiguated using the textual content of the resource. By using this semantic representation of resources, a multivariate Poisson model for naïve Bayes text classification was used to evaluate whether tags improve classification of resources. The precision of content-based profiles was comparable with that of tag-based profiles, both social and personal ones, although results suggested that tags alone are not sufficient to provide accurate recommendations. Sen *et al.* [25] constructed implicit and explicit tag-based recommendation algorithms based on user inferred tag preferences. Inference is based on two direct signals of a user interest in a tag, if the user has applied the tag and if the user has searched for the tag, and a third implicit signal that is the tag quality, a user preference towards a tag may be correlated with its quality. Tag-based recommender systems, referred to as *tagommenders*, aim at filtering resources that appear relevant to target users according to their inferred preferences for tags. Vatturi *et al.* [26] create a personalized tag-based recommender for each user consisting of two Naïve Bayes (NB) classifiers trained over different time frame. One classifier predicts the user current interest based on a shorter time interval, and the other classifier predicts the user general interest in a bookmark considering a longer time interval. If any classifier predicts the bookmark as interesting, it is recommended. The user study results show that the tag-based recommender performs well with real data using tags from an enterprise social bookmarking system. In the same line, this paper presents the results of an empirical study conducted to determine the value of collective annotations for predicting resource interestingness for individual users and the impact of different representation preprocessing techniques tending to normalize tags.

Exploiting the collective knowledge encapsulated in social tags for classification of resources into general directories or hierarchical categories was studied in several works. Zubiaga *et al.* [27] explore the use of SVM to classify Web pages into their corresponding categories in the *Open Directory Project* (ODP) using the tags assigned to these pages in *Del.icio.us*. Besides tagging activity,

DOI: 10.1002/cpe

other metadata-like notes, comments, and reviews were also evaluated for classification, finding that tags in conjunction with comments achieved good results for Web page classification. Moreover, if the motivation for tagging is considered, it was found that users can be discriminated as categorizers or describers [28], having the tags assigned for the first type of users a greater utility for classification [29]. Noll and Meinel [30] compare three different annotations provided by readers of Web documents for classification, such as social tags, hyperlink anchor texts, and search queries of users trying to find pages. The results of this study suggest that tags seem to be better suited for classification of Web documents than anchor words or search keywords, whereas the latter ones are more useful for information retrieval. In a further study [31], the same authors analyzed at which hierarchy depth tag-based classifiers can predict a category using the ODP directory. It was concluded that tags may perform better for broad categorization of documents rather than for narrow categorization. Thus, classification of pages in categories at inferior hierarchical levels might require content analysis. In these studies tags demonstrate to be an important source of information for categorization, beyond the textual content of resources. However, this issue was analyzed for the problem of organizing resources in general taxonomies or directories, not from a personal perspective but from a social one.

3. ONE-CLASS CLASSIFICATION

User actions of assigning tags to resources are a strong indication of content relevance to the user interests. Consequently, positive examples of the user interests can be easily collected from folksonomies as users annotate resources. On the contrary, it would be hard to identify representative negative examples or noninteresting resources because users might not tag a potentially interesting resource because of multiple reasons, lack of time to tagging or even reading it, lack of motivation to tag, or simply not knowing about the existence of a given resource in a folksonomy.

The task of determining whether a document is interesting for a user using only positive examples for training can be seen as a one-class classification problem. One-class classification differs in one essential aspect from conventional classification as it assumes that only information of one of the classes, the target class, is available. The goal is to define a boundary between the two classes estimated from data belonging to the relevant class only, such that the classifier accepts as many of the target objects as possible while minimizing the chance of accepting outlier objects.

Support vector machines are a useful technique for data classification. It has been shown to be perhaps the most accurate algorithm for text classification and it is widely used in Web page classification. Schölkopf *et al.* [32] extended the SVM methodology to handle training using only positive information, and Manevitz *et al.* [33] applied this method to document classification and compare it with other one-class methods. In a previous work [34], SVM was compared with a prototype-based classifier such as Rocchio [35] and a density-based classifier such as the combined one-class classifier [36]. Both classifiers were outperformed by SVMs during experimentation.

Essentially, one-class SVM algorithm consists in learning the minimum volume contour that encloses most of the data, and it was proposed for estimating the support of a high-dimensional distribution [32], given a set of training vectors $\mathcal{X} = \{x_1, \ldots, x_l\}$ in $\mathbb{R}^n$. The aim of SVM is to train a function $f_{\mathcal{X}} : \mathbb{R}^n \rightarrow \mathbb{R}$ such that most of the data in $\mathcal{X}$ belong to the set $\mathcal{R}_{\mathcal{X}} = \{x \in \mathbb{R}^n \text{ with } f_{\mathcal{X}}(x) \geqslant 0\}$ while the volume of $\mathcal{R}_{\mathcal{X}}$ is minimal. This problem is termed minimum volume set estimation, and the membership of x to $\mathcal{R}_{\mathcal{X}}$ indicates whether this data point is overall similar to $\mathcal{X}$.

One-class SVM solves minimum volume set estimation by first mapping the data into a feature space $\mathcal{H}$ using an appropriate kernel function $\phi : \mathbb{R}^n \rightarrow \mathcal{H}$ that transforms training examples into another space. Here, a sigmoid kernel is used, formulated as $\tanh\left[\gamma x_i^T x_j + r\right]$, where r is a shifting parameter that controls the threshold of mapping. For training, a certain number of examples of the positive class are treated as if they belong to the negative class. SVM approach proceeds in $\mathcal{H}$ by determining the hyperplane $\mathcal{W}$ that separates most of the data from the hypersphere origin, separating a certain percentage of outliers from the rest of the data points.

DOI: 10.1002/cpe

To separate the data points from the origin, the following quadratic programming problem needs to be solved:

$$\min_{w,\xi,\rho} \frac{1}{2}\boldsymbol{w}^T\mathbf{w} - \rho + \frac{1}{\nu l}\sum_{i=1}^{l}\xi_i$$

subject to

$$\boldsymbol{w}^T\phi\left(x_i\right) \geqslant \rho - \xi_i$$

$$\text{and } \xi_i \geqslant 0,\ i = 1, 2, \ldots, l$$

where ξ_i are so-called slack variables and υ (Nu) tunes the fraction of data that are allowed to be on the wrong side of $\mathcal{W}$; this parameter defines the trade-off between the percentage of data points treated as belonging to the positive and negative classes. Then, a solution is such that α_i verifies the dual optimization problem:

$$\min_{\alpha} \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} \tag{1}$$

subject to

$$0 \leqslant \alpha_i \leqslant 1/\left(\upsilon l\right), i = 1, \ldots, l$$

$$\boldsymbol{e}^T\boldsymbol{\alpha} = 1$$

where $Q_{ij} = K\left(x_i, x_j\right) \equiv \phi\left(x_i\right)^T\phi\left(x_j\right)$.

In this work, we used LibSVM [‖] [37] library that solves a scaled version of (2) as follows:

$$\min \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} \tag{2}$$

subject to

$$0 \leqslant \alpha_i \leqslant 1, i = 1, \ldots, l$$

$$\boldsymbol{e}^T\boldsymbol{\alpha} = \nu l$$

Finally, the decision function is

$$\operatorname{sgn}\left(\sum_{i=1}^{l}\alpha_i K\left(x_i, x\right) - \rho\right)$$

To adjust the kernel for obtaining optimal results, the parameter γ needs to be tuned to control the smoothness of the boundary. In the sigmoid kernel, gamma serves as an inner product coefficient in the hyperbolic tangent function. Lower gamma values lead to smoother functions, which prevents overfitting of classifiers to training data. Higher values of γ result in flat decision boundaries, which are useful to reproduce highly irregular decision boundaries. This parameter is initially set to the default $\gamma = 0$; variations of this value are then discussed in Section 7.

[‖]http://www.csie.ntu.edu.tw/~cjlin/libsvm/

DOI: 10.1002/cpe

Table I. Summary of Web page statistics per personomy in the dataset used for experimentation.

	Minimum	Maximum	Average	±	SD	Total/unique
No. of full-text terms	13,670	100,664	51,119.80	±	29,763.54	255,599
No. of anchor text terms	9161	56,489	29,702.80	±	16,088.61	148,514
No. of title terms	700	4490	2126.20	±	1364.61	10,631
No. of tags in the top 10 lists	771	3837	2100.40	±	1203.04	10,502
No. of tags in the top 10 lists after filtering symbols	752	3710	2046.60	±	1165.55	10,233
No. of tags in the top 10 lists after stemming	692	3397	1895.40	±	1058.56	9477
No. of tags in the FTA	1867	11,236	5963.20	±	3365.33	29,816
No. of tags in the FTA after filtering symbols	1678	10,001	5377.20	±	3006.29	26,886
No. of tags in the FTA after stemming	1468	8683	4659.00	±	2613.69	23,295

SD, standard deviation; FTA, full tagging activity.

4. EXPERIMENTAL SETTING

Empirical evaluation was carried out using data collected from *Del.icio.us* social bookmarking system. From this site, 200 complete personomies were gathered from randomly chosen users.[**] Each personomy includes all of the user bookmarks and their corresponding tag assignments. In this collection, the size of personomies range from as few as 11 and as many as 5610 bookmarks with an average of 1260.02 resources per personomy. For each Web page, all tags assigned by other members of the community were also extracted from *Del.icio.us*, obtaining the full tagging activity (FTA) or annotations related to each resource.

From the total set of resources gathered from *Del.icio.us* site, experiments reported in this paper were performed over English-written pages, identified using the classification approach presented in [38]. Hence, language-dependent preprocessing tasks were evaluated over both texts and social tags. Given the formal definition of a folksonomy as the tuple $\mathbb{F} := (U, T, R, Y, \prec)$ that describes the users U, resources R, and tags T, and the user-based assignment of tags to resources by a ternary relation between them, that is, $Y \subseteq U \times T \times R$ [39], the resulting folksonomy counts with $|U| = 200$ users and $|R| = 252,104$ bookmarks or Web pages. The users in U create a total of $|T| = 172,154$ unique tags, related to the resources in R by a total of $|Y| = 907,480$ tag assignments. In addition, the resources in R received a total of 4,735,729 tag assignments with 1,031,287 different tags from the entire user community.

Table I summarizes the main statistics of this collection of Web pages averaged by personomy. It includes the number of unique terms in the full text of resources belonging to the different personomies as well as in the text from links and titles. It also contains the number of tags assigned by members of the community to the resources of each user, considering the overall top 10 tags and the FTA. The effect produced in these numbers by two tag filters explained in Section 6 is also detailed. The average numbers of each element correspond to the number of features classifiers have to deal with during learning.

In all experiments reported in this paper, evaluation was carried out using a holdout strategy that split data into a 70% for training and a 30% for testing. To make the results less dependent of the data splitting, in all experiments, the average and standard deviation of 10 runs for each user is reported. In other words, each personomy was divided into a training set used to learn the classifier and a testing set used to assess its validity. Because this testing set only contains interesting examples, uninteresting pages were extracted from the personomies of other users to evaluate the algorithm capacity of distinguishing uninteresting resources. This is, the testing set was created using the test set from the target user plus an equivalent number of Web pages gathered from a different personomy in the collection. This second personomy was randomly chosen among those presenting

[**] http://users.exa.unicen.edu.ar/~dgodoy/perso200delicious.html

DOI: 10.1002/cpe

neither resource nor tag intersection with the current user. In other words, it is assumed that two users having no common resources in their personomies and using completely different tags do not share interests, so that one user resources will be uninteresting to the other one. Although this may be not strictly true, this assumption allows the creation of a negative set for testing, and its violation can only lead to a underestimation in the precision of classifiers (i.e., some pages in the test set may be relevant in spite of being considered irrelevant, and then a correct prediction will be counted as a false positive when it is in fact a true positive).

For evaluating classifiers, precision and recall metrics were used [40], whereas error bars indicate standard deviations. Precision is defined as the number of relevant selected resources divided by the number of selected resources during filtering. Recall is defined as the number of relevant selected resources divided by the total number of relevant resources available. F-measure, denoted F_β, is a combination of precision and recall, in which β sets the relative degree of importance attributed to both metrics. In all experiments reported in this paper, F_1 score was used, which can be interpreted as a weighted average of precision and recall.

5. CONTENT-BASED CLASSIFICATION

Content is one of the main sources of information for determining the relevance of Web pages for users. It is assumed that similar contents to those previously seen by the user will be also interesting. To establish the relative importance of content and social tags in personal Web page classification, the performance of one-class classification over textual elements obtained from documents was first evaluated so that it can be use as baseline for comparing the performance of tag-based classifiers.

Web page texts were filtered using a standard stop-word list, and Porter stemming algorithm [41] was applied to the remaining terms. Figure 1 shows the results of training classifiers for identifying interesting Web pages using different textual sources such as the full text of documents, the text attached to links (i.e., the visible, clickable text in a hyperlink), and the page title. Each of these elements is extracted from pages belonging to a user personomy to learn a classifier or user profile. F-measure scores achieved with different values of ν (Nu) parameter of one-class classifiers are showed in the figure.

Classification using full text obtained the best results, closely followed by results obtained using the text from links. The titles of resources alone, however, did not turn out to be a good source of information for filtering interesting resources. Naturally, this is caused by the relatively small set of features or terms involved in learning and prediction when only titles are considered instead of the full text of pages.

The relatively low scores of F-measure is caused by the absence of negative information during learning. In addition, the negative testing set might have some interesting pages because of possible

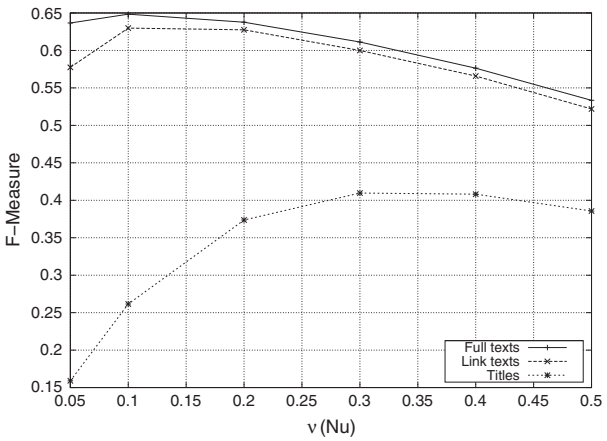

Figure 1. F-measure scores using different content-related sources for classification.

DOI: 10.1002/cpe

violations of the assumption that users do not share interests if their personomies do not intersect each other. Nevertheless, text classifiers were capable of recognizing part of the user interests and are a valuable source for filtering massive amounts of information from incoming data streams.

6. SOCIAL-BASED CLASSIFICATION

Social tagging systems on the Web owe their success to the opportunity of freely determining a set of tags for a resource without the constraint of a controlled vocabulary, lexicon, or predefined hierarchy [1]. However, the free-form nature of tagging also leads to a number of vocabulary problems. Multiple reasons can caused tag variations [42–44]:

- inconsistent grouping of compound words consisting of more than two words. Often, users insert punctuation to separate the words, for example, *ancient-egypt*, *ancient_egypt*, and *ancientgypt*;
- use of symbols in tags; symbols such as #, -, +, /, :, _, &, and ! are frequently used at the beginning of tags to cause some incidental effect such as forcing the interface to list a tag at the top of an alphabetical listing;
- morphological problems given by the use of singular, plural, or other derived forms of words. For example, *blog*, *blogs*, and *blogging.*

To prevent syntactic mismatches because of the aforementioned factors, the effect of different filtering strategies for tags was evaluated. First, original raw tags were filtered to remove the symbols mentioned before, allowing to join compound words at the same time. Then, the remaining tags were stemmed to their morphological roots using Porter stemming algorithm.

In addition to the mentioned preprocessing strategies for tags, different representations of the resulting tag vectors describing resources were empirically evaluated. Social tags were considered in two forms: (i) the overall top 10 tags associated each resource in the folksonomy, that is, the more popular tags assigned to a resource; and (ii) the FTA of a resource, that is, the complete set of tags assigned for users to such resources.

Frequency-based and binary representations of the resulting tag vectors were also considered and compared for both FTA and top 10 descriptions. Binary vectors were constructed to indicate the occurrence or nonoccurrence of a given tag in the list of tags a Web page is annotated with. Frequency vectors indicate the number of users that employ a given tag to annotate such resource; that is, f_{ij} is the frequency of usage of tag i for the resource j; these vectors were normalized according to their length.

6.1. Results using the full tagging activity

Figure 2 depicts F-measure scores achieved with one-class SVM classifiers learned using the FTA associated to resources, that is, all tags assigned to each resource by members of community. Figure 2(a) and (b) shows the results for frequency-based and binary representations of the resulting tag vectors, respectively.

In both figures, results are shown for raw tags as well as tags resulting from applying the mentioned filtering strategies, first symbol removal and then stemming. In regard to the tag filtering operations, it can be deduced according to the results in Figure 2(a) that removing symbols and joining compound words as well as performing stemming did not improve the performance of classifiers, even harming their behavior in some cases. Binary representations outperformed frequency-based ones in terms of classifiers effectiveness, and in this case, filters attain small performance enhancements over raw tags.

6.2. Results using the top-10 tags

Figure 3 depicts the results using the same configuration of experiments but applied to vectors obtained using the top 10 tags attached to resources, that is, the most popular tags assigned to each resource in the entire folksonomy.

DOI: 10.1002/cpe

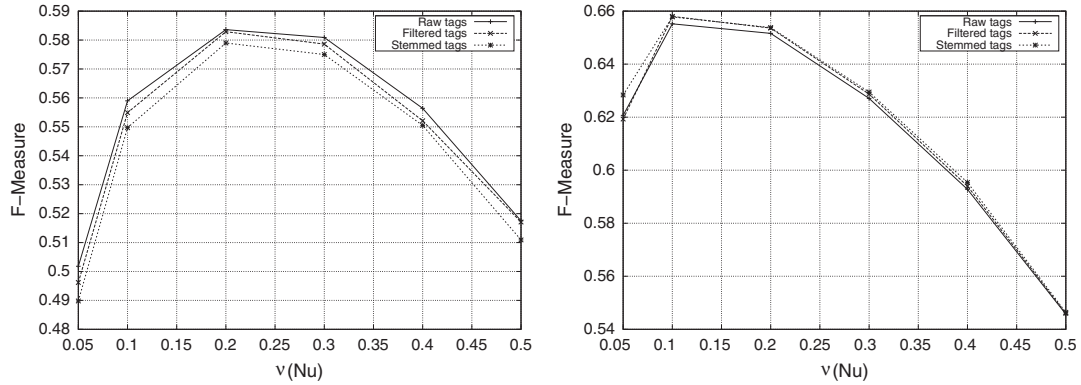

Figure 2. F-measure scores achieved with the full tagging activity associated to resources. (a) Frequency-based representation of tag vectors and (b) binary representation of tag vectors.

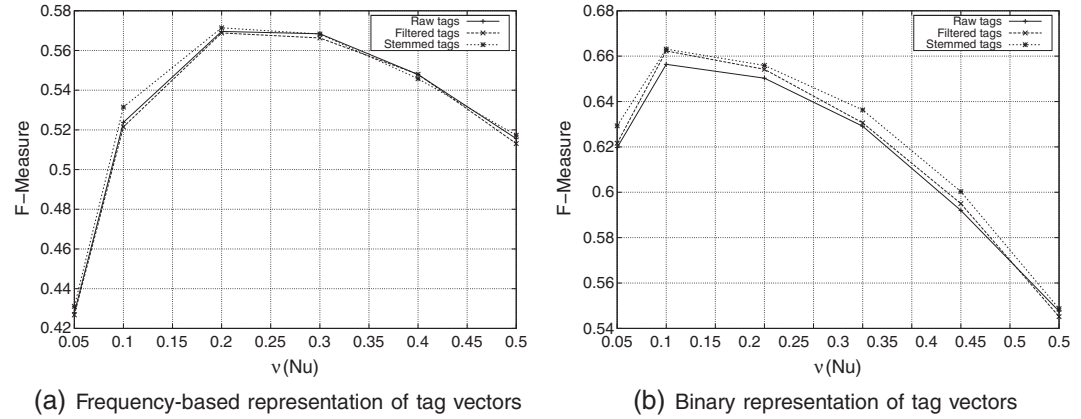

Figure 3. F-measure scores achieved with the top 10 tags associated to resources. (a) Frequency-based representation of tag vectors and (b) binary representation of tag vectors.

In general terms, the results of using binary representation for tag vectors, which are shown in Figure 3(b), provide a significant improvement over normalized frequency vectors. Similar to the previous experiments, in this representation scheme, removing symbols and joining compound words lead to small improvements. However, the use of stemming seems to have a positive effect over raw tags.

7. SUMMARY OF RESULTS

Figure 4 summarizes the results obtained for full-text classification of Web resources and tag-based classification using both the top 10 tags associated to each resource and the FTA in both their frequency-based and binary representations.

In the figure, it can be observed that the results of using frequency-based representations of tag vectors (both FTA and top-10) are the ones showing the poorest performance among the different classifiers. The main source of information about the content of a resource, which is the text of the resource itself, leads to a medium performance. Both frequency-based and content-based classifiers are consistently outperformed by the use of social tags when a binary representation of tag vectors is applied. Saliently, the top 10 vectors behaves slightly better than the FTA vectors in their binary representations.

DOI: 10.1002/cpe

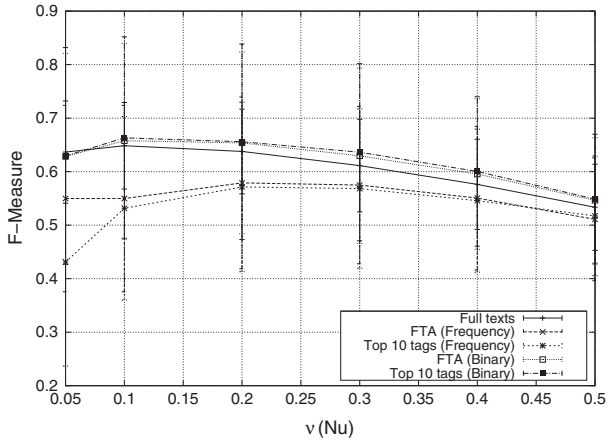

Figure 4. Summary of F-measure scores of content and tag-based classifiers.

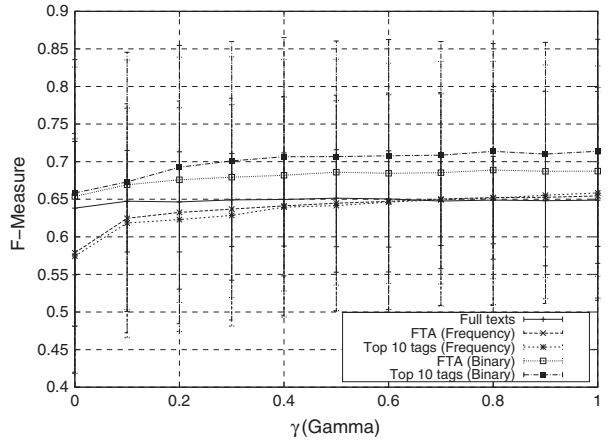

Figure 5. F-measure scores for variations of γ (gamma) parameter of one-class support vector machine classifiers.

Classification based on binary-based representations using the FTA of resources as well as the top 10 tags reached best performance among tag-based classifiers. In contrast, the use of frequency-based tag vectors exhibits inferior performance in identifying interesting Web pages for users. An important result is that richer representation obtained from the FTA of resources leads to similar or even inferior classification results than the reduced representation using only the top 10 tags.

Figure 5 shows the results of using the same information sources for classification setting ν to 0.2, the point at which the best results were achieved and varying the value of γ during learning of one-class SVM classifiers. Higher values of γ lead to small increases in F-measure scores, so the best scores are achieved in the interval $0.8 < \gamma < 1$. More importantly, for gamma values $\gamma > 0.6$, any form of representation of social tags, even frequency-based, outperforms full-text classification. Similar to previous results, binary representations are best performing. More importantly, top 10 tag vectors improve the results of FTA vectors by a wider margin after $\gamma > 0.2$

The last observation becomes important as it impacts directly on learning complexity. Indeed, tag-based classifiers extracted from the top 10 list of tags are learned in a smaller dimensional space than full-text classifiers, and yet, they are better predictors. Likewise, top 10 tag vectors level the results of classifiers considering the FTA of resources, which are also learned in a higher dimensional space. Table I summarizes the number of unique features, terms or tags according to the case, and the classification problem they have to deal with in each case to illustrate this issue.

DOI: 10.1002/cpe

It is worth mentioning that full text is used in these experiments as baseline for comparison, but this source of information is not always available in social tagging systems, in which resources can be a variety of things, such as images, music, bibliographic references, and so on. In these situations, classification must entirely rely on social tags. Thus, it can be concluded that collective knowledge lying in folksonomies becomes a valuable source of information for automatic, personal classification of Web resources.

Figure 6 summarizes the performance of content and tags-based classifiers in terms of accuracy for $\nu = 0.2$ and $\gamma = 1$, the parameter setting leading to the best results in the reported experiments. Confirming previous results, if the classifiers capability of making correct decisions is considered, tag-based classifiers outperformed full text ones. Also, among tag-based classifiers those based on binary representations demonstrate a superior performance. For both representation alternatives, vectors used for training the top 10 tags assigned to each resource were the ones of superior performance. Thus, top 10 tags offer good accuracy levels and, at the same time, an important reduction in learning and prediction complexity given the smaller size of the dimensional space.

Finally, the incidence of the different sources of information for filtering Web pages, content, or social tags is analyzed according to the size of personomies in Figure 7. For studying this aspect of classification, the 200 users were divided into five groups according to the amount of resources in their personomies. In the first group, users having less than 300 annotated resources were placed, and then users having from 300 to 600 resources, 600 to 1000, 1000 to 2000, and more than 2000 resources.

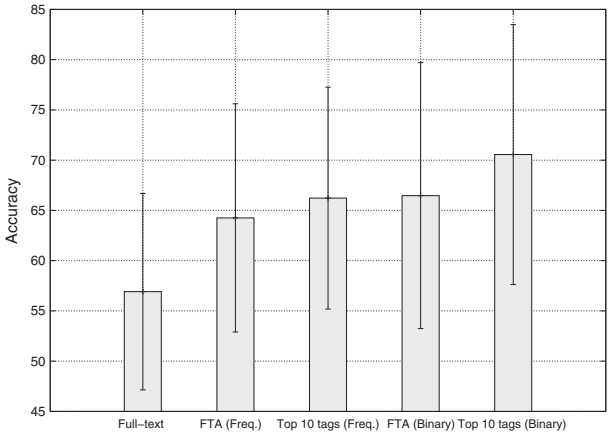

Figure 6. Summary of content and tag-based classification accuracy.

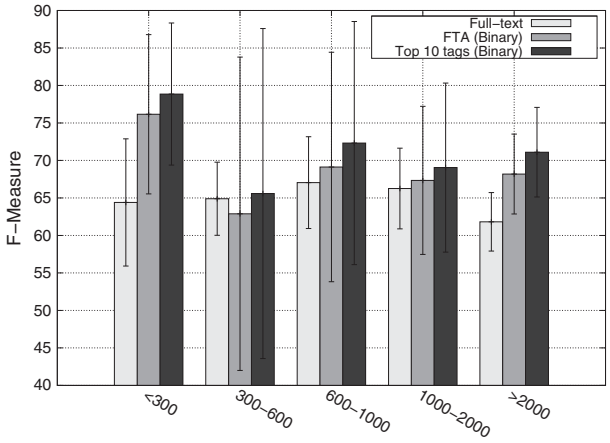

Figure 7. F-measure scores according to personomy sizes.

DOI: 10.1002/cpe

Along the different personomy sizes, one-class classifiers showed similar F-measure scores. The difference between social tags and full text as source for classification is more noticeable in smaller personomies, in which a few documents are available to learn a classifier, than in medium-size and large personomies, in which more text is available for content-based classification. For these personomies, taking advantage of the collective knowledge from folksonomies results crucial for accurately filtering interesting results.

8. CONCLUSIONS

In this paper, the value of social tags in filtering resources from social tagging systems according to the interests of individual users was empirically assessed. One-class classification was used to learn user interests from diverse content-related sources, including full text, text from links and titles, as well as social tagging sources, including the top 10 list of all tags associated to resources and their FTA. Then, the extend to which each source can contribute to automatic, personal Web document classification was empirically evaluated and compared.

Experimental results obtained with a set of personomies extracted from *Del.icio.us* bookmarking system showed that tag-based classifiers outperformed content-based ones. Some tag filters such as removal of symbols, joint of compound words, and reduction of morphological variants have a discrete impact on classification performance. It was observed that binary representations of tag vectors were best performing in terms of learning and prediction. In addition, tag-based classifiers learned using only the overall top 10 tags associated to resources reached superior performance levels than those learned using the FTA even though they are learned in a considerably smaller dimensionality space.

ACKNOWLEDGEMENTS

This research was supported by the National Scientific and Technical Research Council (CONICET) under grant PIP No 114-200901-00381.

REFERENCES

1. Mathes A. Folksonomies - cooperative classification and communication through shared metadata, 2004. Computer Mediated Communication.
2. Dattolo A, Ferrara F, Tasso C. The role of tags for recommendation: a survey. In *Proceedings of the 3rd International Conference on Human System Interaction (HSI'2010)*, Rzeszow, Poland, 2010; 548–555.
3. Milicevic AK, Nanopoulos A, Ivanovic M. Social tagging in recommender systems: A survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review* 2010; **33**(3):187–209.
4. Arakji R, Benbunan-Fich R, Koufaris M. Exploring contributions of public resources in social bookmarking systems. *Decision Support Systems* 2009; **47**(3):245–253.
5. Lipczak M. Tag recommendation for folksonomies oriented towards individual users. In *Proceedings of ECML PKDD Discovery Challenge (RSDC08)*, Antwerp, Belgium, 2008; 84–95.
6. Lu Y-T, Yu S-I, Chang T-C, Hsu JY. A content-based method to enhance tag recommendation. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence (IJCAI'09)*, Pasadena, California, USA, 2009; 2064–2069.
7. Symeonidis P, Nanopoulos A, Manolopoulos Y. A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis. *IEEE Transactions on Knowledge and Data Engineering* 2010; **22**(2):179–192.
8. Jäschke R, Marinho L, Hotho A, Schmidt-Thieme L, Stumme G. Tag recommendations in folksonomies. In *Knowledge Discovery in Databases (PKDD 2007)*, volume 4702 of LNCS, 2007; 506–514.
9. Tso-Sutter KHL, Marinho LB, Schmidt-Thieme L. Tag-aware recommender systems by fusion of collaborative filtering algorithms. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC '08)*, Fortaleza, Ceará, Brazil, 2008; 1995–1999.
10. Sood S, Owsley S, Hammond K, Birnbaum L. TagAssist: Automatic tag suggestion for blog posts. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'2007)*, Boulder, Colorado, USA, 2007.
11. Basile P, Gendarmi D, Lanubile F, Semeraro G. Recommending smart tags in a social bookmarking system. In *Bridging the Gep Between Semantic Web and Web 2.0 (SemNet 2007)*, Innsbruck, Austria, 2007; 22–29.
12. Illig J, Hotho A, Jäschke R, Stumme G. A comparison of content-based tag recommendations in folksonomy systems. In *Postproceedings of the International Conference on Knowledge Processing in Practice (KPP 2007)*, Novosibirsk, Russia, 2007; 136–149.

DOI: 10.1002/cpe

13. Zhang Z-K, Zhou T, Zhang Y-C. Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs. *Physica A: Statistical Mechanics and its Applications* 2010; **389**(1):179–186.
14. Musto C, Narducci F, de Gemmis M, Lops P, Semeraro G. An IR-based approach for tag recommendation. In *Proceedings of the 1st Italian Information Retrieval Workshop (IIR 2010)*, Padua, Italy, 2010; 65–69.
15. Noll MG, Meinel C. Web search personalization via social bookmarking and tagging. In *Proceedings of 6th International Semantic Web Conference (ISWC) and 2nd Asian Semantic Web Conference (ASWC)*, Vol. 4825, LNCS, 2007; 367–380.
16. Diederich J, Iofciu T. Finding communities of practice from user profiles based on folksonomies. In *Proceedings of the 1st International Workshop on Building Technology Enhanced Learning Solutions for Communities of Practice (TEL-CoPS'06)*, Crete, Greece, 2006; 288–297.
17. Stoyanovich J, Yahia SA, Marlow C, Yu C. Leveraging tagging to model user interests in del.icio.us. In *AAAI Spring Symposium on Social Information Processing (AAAI-SIP)*, Stanford University, California, USA, 2008; 104–109.
18. Firan C, Nejdl W, Paiu R. The benefit of using tag-based profiles. In *Proceedings of the 2007 Latin American Web Conference (LA-WEB 2007)*, Santiago de Chile, Chile, 2007; 32–41.
19. Yeung CMA, Gibbins N, Shadbolt N. A study of user profile generation from folksonomies. In *Social Web and Knowledge Management, Social Web 2008 Workshop at WWW'2008*, Beijing, China, 2008.
20. Shepitsen A, Gemmell J, Mobasher B, Burke R. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys'08)*, Lausanne, Switzerland, 2008; 259–266.
21. Michlmayr E, Cayzer S. Learning user profiles from tagging data and leveraging them for personal(ized) information access. In *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization*, Banff, Alberta, Canada, 2007.
22. Michlmayr E, Cayzer S, Shabajee P. Add-A-Tag: Learning adaptive user profiles from bookmark collections. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'07)*, Boulder, Colorado, USA, 2007.
23. Huang Y-C, Hung C-C, Hsu JY-J. You are what you tag. In *AAAI Spring Symposium on Social Information Processing (AAAI-SIP)*, Stanford University, California, USA, 2008; 36–41.
24. de Gemmis M, Lops P, Semeraro G, Basile P. Integrating tags in a semantic content-based recommender. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys'08)*, Lausanne, Switzerland, 2008; 163–170.
25. Sen S, Vig J, Riedl J. Tagommenders: connecting users to items through tags. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*, Madrid, Spain, 2009; 671–680.
26. Vatturi PK, Geyer W, Dugan C, Muller M, Brownholtz B. Tag-based filtering for personalized bookmark recommendations. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*, Napa Valley, California, USA, 2008; 1395–1396.
27. Zubiaga A, Martínez R, Fresn V. Getting the most out of social annotations for Web page classification. In *Proceedings of the 9th ACM Symposium on Document Engineering (DocEng'2009)*, Munich, Germany, 2009; 74–83.
28. Körner C, Kern R, Grahsl H-P, Strohmaier M. Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT'10)*, Toronto, Ontario, Canada, 2010; 157–166.
29. Zubiaga A, Körner C, Strohmaier M. Tags vs shelves: From social tagging to social classification. In *Proceedings of the 22st ACM Conference on Hypertext and Hypermedia (HT'11)*, Eindhoven, Netherlands, 2011; 93–102.
30. Noll MG, Meinel C. The metadata triumvirate: Social annotations, anchor texts and search queries. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Sydney, Australia, 2008; 640–647.
31. Noll MG, Meinel C. Exploring social annotations for Web document classification. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC'08)*, Fortaleza, Ceará, Brazil, 2008; 2315–2320.
32. Schölkopf B, Platt JC, Shawe-Taylor JC, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. *Neural Computation* 2001; **13**(7):1443–1471.
33. Manevitz LM, Yousef M. One-class SVMs for document classification. *Journal of Machine Learning Research* 2002; **2**:139–154.
34. Godoy D. Comparing one-class classification algorithms for finding interesting resources in social bookmarking systems. In *Resource Discovery*, Vol. 6799, LNCS. Springer, 2012; 88–103.
35. Rocchio J. Relevance feedback in information retrieval. In *The Smart Retrieval System*, Salton G (ed.). Prentice Hall: Upper Saddle River, NJ, USA, 1971; 313–323.
36. Hempstalk K, Frank E, Witten IH. One-class classification by combining density and class probability estimation. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, Vol. 5211, LNCS. Springer: Berlin, Heidelberg, 2008; 505–519.
37. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm [Accessed on 10 October 2010].
38. Cavnar W, Trenkle J. N-gram-based text categorization. In *Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, USA, 1994; 161–175.

DOI: 10.1002/cpe

39. Hotho A, Jáschke R, Schmitz C, Stumme G. Information retrieval in folksonomies: Search and ranking. In *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference (ESWC'2006)*, Vol. 4011, LNCS. Springer: Berlin, Heidelberg, 2006; 411–426.
40. Baeza-Yates R, Ribeiro-Neto B. *Modern Information Retrieval*. Addison-Wesley Longman Publishing: Boston, MA, USA, 1999.
41. Porter M. An algorithm for suffix stripping program. *Program* 1980; **14**(3):130–137.
42. Golder S, Huberman B. Usage patterns of collaborative tagging systems. *Journal of Information Science* 2006; **32**(2):198–208.
43. Tonkin E, Guy M. Folksonomies: Tidying up tags?. *D-Lib Magazine, The Magazine of Digital Library Research* 2006; **12**(1).
44. Echarte F, Astrain J, Villadangos J. Pattern matching techniques to identify syntactic variations of tags in folksonomies. In *Proceedings of the 1st World Summit on the Knowledge Society (WSKS'08)*, Vol. 5288, LNCS. Springer-Verlag: Athens, Greece, 2008; 557–564.

DOI: 10.1002/cpe