On the Role of Social Tags in Filtering Interesting Resources from Folksonomies

Daniela Godoy
ISISTAN Research Institute, UNICEN University
Campus Universitario, CP 7000, Tandil, Argentina
Also at CONICET, Argentina
dgodoy@conicet.gov.ar

Abstract

Social tagging systems allow users to easily create, organize and share collections of resources (e.g. Web pages, research papers, photos, etc.) in a collaborative fashion. The rise in popularity of these systems in recent years go along with an rapid increase in the amount of data contained in their underlying folksonomies, thereby hindering the user task of discovering interesting resources. 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 classification is evaluated as a means to learn how to identify relevant information based on positive examples exclusively, since it is assumed that users expressed their interest in resources by annotating them while there is not an straightforward method to collect non-interesting information. The results of using social tags for personal classification are compared with those achieved with traditional information sources about the user interests such as the textual content of Web documents. Finding interesting resources based on social tags is an important benefit of exploiting the collective knowledge generated by tagging activities. Experimental evaluation showed that tag-based classification outperformed classifiers learned using the full-text of documents as well as other content-related sources.

1 Introduction

Social tagging systems have grown in popularity on the Web in the last years on account of their simplicity to categorize and retrieve shared content using open-ended tags. In sites such as *Del.icio.us*[1], *Flickr*[2] or *CiteULike*[3], users annotate a variety of resources (Web pages, blog posts or pictures) using a freely chosen set of keywords, which facilitates later search and retrieval of such contents.

Folksonomies [Mathes, 2004] are the primary structure of the novel social classification scheme introduced by tagging systems, which relies on the convergence of the tagging efforts of a large community of users to a common categorization system that can be effectively used to organize and navigate a massive amount of freely accessible, user contributed and annotated Web resources.

In spite of the novel mechanisms for searching and retrieving resources provided by collaborative tagging systems, the rapid increase in size of communities using these systems as well as the large amount of shared content available make the discovery of relevant resources a time consuming and difficult task for users. This problem is aggravated by the completely unsupervised nature of social tags, resulting in ambiguity, noise, etc.; which may reduce their effectivity in content indexing and searching.

The goal of this paper is to study the utility of social tags as a source of information for filtering resources from folksonomies according to the user interests. In social tagging systems resources receive tag assignments by members of the community, describing their content in a collective sense. Thus, it can be assumed that users are likely to be interested in additional content annotated with similar tags to the ones collectively assigned to resources they showed interest in before.

Social tags associated to the resources annotated by the user can be used to build a user interest profile that, in turn, can be applied to filter further incoming information from tagging systems (e.g. RSS feeds). From a user perspective, social tags can be thought of as indicators of user awareness and potential interest in a given resource [Arakji *et al.*, 2009], allowing users to capitalize on the associations made by persons who have assigned similar tags to other resources.

In order to identify interesting resources, tag-based classifiers are learned using the resources users annotate and have in their personomies, the tag collection of a single user, as positive examples of their interests. This is a special case of classification in which it is necessary to determine whether an example (resource) belongs to a target class (*interesting*) when only examples of the target class are given, which is known as one-class classification.

The rest of the paper is organized as follows. Section 2 gives an overview of one-class classification using Support Vector Machines (SVM) classifiers. Section 3 describes the dataset used for experimentation, gathered from *Del.icio.us* bookmarking site. The empirical analysis carried out to compare content-based and tag-based classification of Web pages in personomies is presented in Sections 4 and 5, respectively. Section 6 reviews related research. Finally, empirical findings are summarized in Section 7.

2 One-class Classification

User actions of assigning tags to resources are a strong indication of relevance about its content. Consequently, positive examples of the user interests can be easily collected from folksonomies. On the contrary, it would be

[1] http://del.icio.us/
[2] http://www.flickr.com/
[3] http://www.citeulike.org/

hard to identify representative negative examples or non-interesting resources since users might not tag a potentially interesting resource because of multiple reasons, such as not knowing about the existence of the resource, lack of time to tagging or even reading it, etc.

The task of determining whether a document is interesting for a user basing training only on positive examples 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 idea is to define a boundary between the two classes estimated from data belonging to the relevant class, such that it accepts as much of the target objects as possible while minimizes the chance of accepting outlier objects.

SVMs (Support Vector Machines) are a useful technique for data classification, which has been shown that is perhaps the most accurate algorithm for text classification, it is also widely used in Web page classification. Schölkopf et al. [Schölkopf *et al.*, 2001] extended the SVM methodology to handle training using only positive information and Manevitz et al. [Manevitz and Yousef, 2002] apply this method to document classification and compare it with other one-class methods.

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 [Schölkopf *et al.*, 2001], 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 $\mathcal{R}_{\mathcal{X}} = \{x \in \mathbb{R}^n \text{ with } f_{\mathcal{X}}(x) \geq 0\}$ while the volume of $\mathcal{R}_{\mathcal{X}}$ is minimal. This problem is termed minimum volume set (MVS) 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 MVS estimation by first mapping the data into a feature space $\mathcal{H}$ using an appropriate kernel function $\phi : \mathbb{R}^n \rightarrow \mathcal{H}$ which transforms training examples to another space. Here, the Gaussian RBF kernel is used, formulated as $\exp\left[-\gamma \left\|x_i - x_j\right\|^2\right]$, where $\|.\|$ denotes the Euclidean norm in $\mathbb{R}^n$. For training, a certain number of data points 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.

In order 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

$$\boldsymbol{w}^T\phi(x_i) \geq \rho - \xi_i$$

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

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 verify the dual optimization problem:

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

subject to

$$0 \leq \alpha_i \leq 1/(\upsilon l), i = 1, \ldots, l$$

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

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

In this work we used LibSVM[4] [Chang and Lin, 2001] library which solves a scaled version of 2 as follows:

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

subject to

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

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

Finally, the decision function is:

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

In order to adjust the kernel for optimal results, the parameter γ need to be tuned to control the smoothness of the boundary, i.e. large values of γ lead to flat decision boundaries. The setting of this parameter is initially set to $\gamma = 0$, variations of this value are then discussed in Section 5.

3 Dataset Description

Emerging social structures in tagging systems, also known as folksonomies, can be defined as a tuple $\mathbb{F} := (U, T, R, Y, \prec)$ which describes the users U, resources R, and tags T, and the user-based assignment of tags to resources by a ternary relation between them, i.e. $Y \subseteq U \times T \times R$ [Hotho *et al.*, 2006]. The collection of all tag assignments of a single user constitute a personomy, i.e. the personomy $\mathbb{P}_u$ of a given user $u \in U$ is the restriction of $\mathbb{F}$ to u, i. e., $\mathbb{P}_u := (T_u, R_u, I_u, \prec_u)$ with $I_u := \{(t, r) \in T \times R \,|\, (u, t, r) \in Y\}$, $T_u := \pi_1(I_u)$, $R_u := \pi_2(I_u)$, and $\prec_u := \{(t_1, t_2) \in T \times T \,|\, (u, t_1, t_2) \in \prec\}$, where π_1 the projection on the *i*th dimension.

Empirical evaluation was carried out using data collected from *Del.icio.us*[5] social bookmarking system. From this site 50 complete personomies were gathered from different users appearing in the main page. Each personomy includes all of the user bookmarks and the corresponding tag assignments. In this collection of personomies there are users with as few as 10 and as much as 2521 bookmarks. For each Web page, in turn, 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 [Cavnar and Trenkle, 1994]. This allows to apply language-dependent pre-processing

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

[5] http://del.icio.us/

tasks to both texts and social tags. The resulting folksonomy counts with $|U| = 50$ users, $|T| = 233.997$ tags and $|R| = 49.265$ bookmarks or Web pages, related by a total of $|Y| = 128.642.112$ tag assignments. Table 1 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 of anchors and titles. 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 full tagging activity. The effect produced in these numbers by two tag filters explained in Section 5 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 66% for training and a 34% for testing. In order 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. This is, each personomy was divided into a training set used to learn the classifier and a testing set used to assess its validity. Since 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 user and an equivalent number of Web pages gathered from a different personomy in the collection. This second personomy was randomly chosen among those presenting no resource intersection with the current user. In other words, it is assumed that two users having no common resources in their personomies do not shared interests, so that one user resources will be uninteresting to the other one. Although this is not strictly true, it can be considered as an approximation to obtain a negative set for testing. For evaluating the classifiers, the standard precision and recall summarized by F-measure as well as accuracy were used and error-bars indicate standard deviations [Baeza-Yates and Ribeiro-Neto, 1999].

4 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. In order 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 used as baseline for comparing the performance of tag-based classifiers.

Web page texts were filtered using a standard stop-word list and the Porter stemming algorithm [Porter, 1980] 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 anchor text attached to hyperlinks (i.e. the visible, clickable text in a hyperlink) belonging to the page and the page title. Each of these elements is extracted from pages belonging to a user personomy to learn a classifier for such user. 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 the text from anchors. The title of resources alone, however, did not result to be a good source for filtering interesting information. Naturally, the relatively low scores of F-measure is caused by the absence of negative information during learning. In addition, in the negative testing set might be some interesting pages, due to possible violations of the assumption that users do not shared interests if their personomies do not intersect each other, for which the prediction is correct but taken as an error. Nevertheless, text classifiers were still able of recognizing part of the user interests and are a valuable source for filtering a stream of incoming information (e.g. a RSS feed).

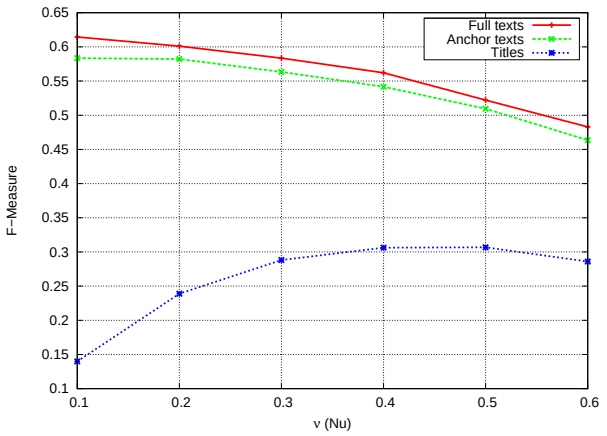

Figure 1: F-measure scores using different textual sources from the content of resources for classification

5 Social-based Classification

Social tagging systems on the Web own their success to the opportunity of freely determining a set of tags for a resource without the constraint of a controlled vocabulary, lexicon or pre-defined hierarchy [Mathes, 2004]. However, the free-form nature of tagging also leads to a number of vocabulary problems. Among the reasons producing tags variations are [Golder and Huberman, 2006; Tonkin and Guy, 2006; Echarte *et al.*, 2008]:

- inconsistently 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 #, -, +, /, :, _, &, ! are frequently used at the beginning of tags to cause some incidental effect such as forcing the interface to list some 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 due to these reasons 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 this study the overall top 10 tags associated to resources in the folksonomy, this is the 10 more frequent tags per resource, were evaluated as a source for classification and compared with the use of the complete set of tags assigned for users to such resources, also known as the full

	Min	Max	Average	±	SD	Total
# full-text terms	1.997	115.585	47.138,14	±	29.063,53	2.356.907
# anchor text terms	681	62.521	24.632,82	±	16.268,12	1.231.641
# title terms	31	4.581	1.806,44	±	1.236,02	90.322
# tags in the top 10 lists	57	4.117	1.739,76	±	1.097,52	86.988
# tags in the top 10 lists after filtering symbols	57	3.882	1.686,14	±	1.055,23	84.307
# tags in the top 10 lists after stemming	55	3.462	1.495,68	±	934,83	74.784
# tags in the FTA	150	10.902	4.679,94	±	3.053,78	233.997
# tags in the FTA after filtering symbols	141	9.872	4.328,46	±	2.791,12	216.423
# tags in the FTA after stemming	122	8.678	3.757,00	±	2.426,63	187.856

Table 1: Summary of Web page statistics per personomy in the dataset used for experimentation

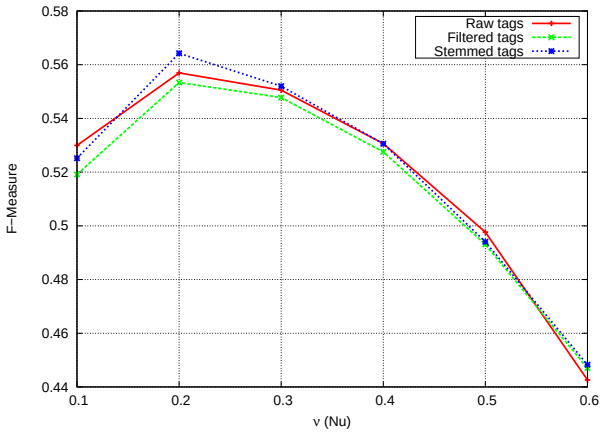

Figure 2: F-measure scores using frequency-based representations of the top 10 tags associated to resources for classification

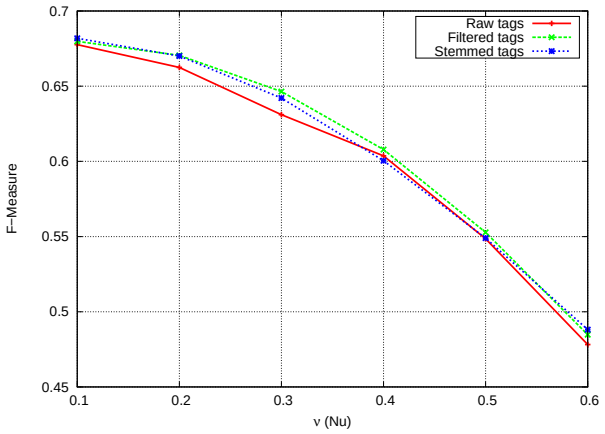

Figure 3: F-measure scores using binary representations of the top 10 tags associated to resources for classification

tagging activity of the resource. Frequency-based and binary representations of the resulting tag vectors were also considered and compared. Binary vectors were constructed to indicate the occurrence or non-occurrence 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 the resource; this is f_{ij} is the frequency of usage of tag *i* for the resource *j,* these vectors are normalized according to their length.

5.1 Results using top-10 tags

Figure 2 depicts F-measure scores achieved with one-class SVM classifiers leaned using the top 10 list of tags. In the figure, results are shown for raw tags as well as tags resulting of applying the mentioned filtering strategies, first symbol removal and then stemming. In regards to the tag filtering operations it can be deduced according to these results that removing symbols and joining compound words slightly diminish the performance of classifiers, whereas stemming improves it. Note that the filters were applied in the previously mentioned order, so that stemming can potentially achieve better results if applied over raw tags directly.

In general terms, the results of using binary representation for tag vectors, which are shown in Figure 3, provides a significant improvement over normalized frequency vectors. In this representation scheme, removing symbols and joining compound words reduce the noise of tags resulting in an improvement of F-measure scores. However, the use of stemming does not lead to further improvements, even damaging the performance of classifiers.

5.2 Results using full tagging activity

Figures 4 and 5 depict the results using the same configuration of experiments but applied to vectors resulting of the full tagging activity attached to resources.

F-measure scores of one-class SVM classifiers learned using frequency vectors, depicted in Figure 4, shows that the filter used to remove symbols and join compound words does not improve significantly the performance of classifiers whereas stemming obtains slightly enhancements. In turn, binary representations again outperformed frequency-based ones and filters attain small performance enhancements over raw tags.

5.3 Summary of results

Figure 6 summarizes the results obtained for full-text classification of Web resources and tag-based classification using both the top 10 tags of each resource and its full tagging activity in their frequency-based and binary representations.

It can be observed that the results of using the main source of information about the content of a resource, which is the text of the resource itself, is consistently outperformed by the use of social tags when a binary representation of tag vectors is applied. Classification based on frequency-based representations of the full tagging activity of resources also reached better performance than full-text classification. In contrast, the use of the 10 more frequent tags used to annotate resources as a means for classification exhibit inferior performance in identifying interesting Web pages for users.

Figure 7 shows the results obtained with the same classification sources by setting ν to 0.1, the point at which

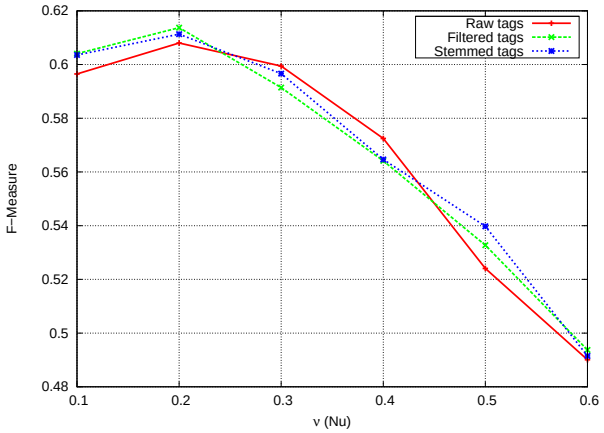

Figure 4: F-measure scores using frequency-based representations of the full tagging activity for classification

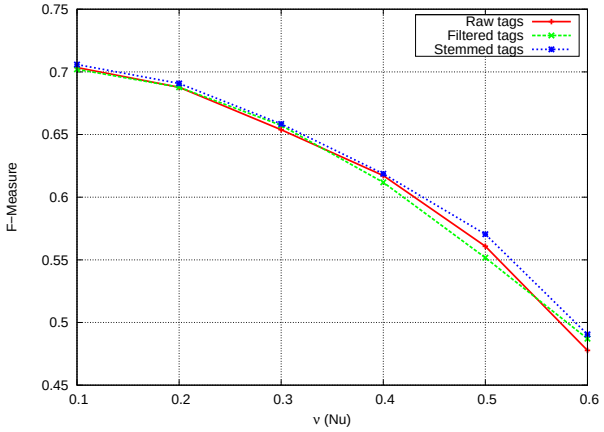

Figure 5: F-measure scores using binary representations of the full tagging activity for classification

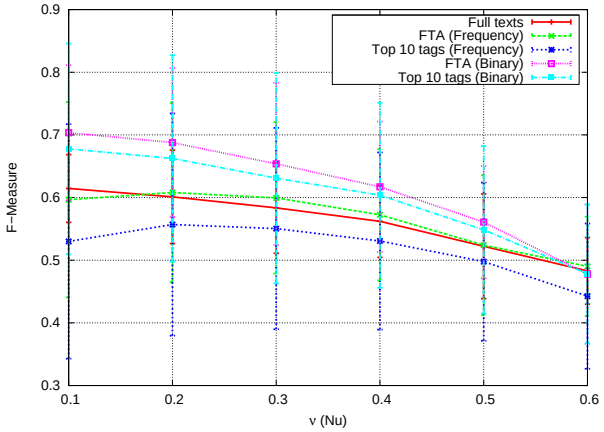

Figure 6: F-measure scores considering content and social tags as sources for classification

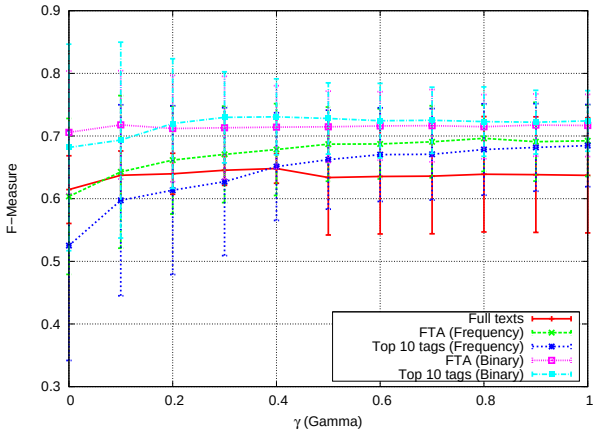

Figure 7: F-measure scores for variations of γ (gamma) parameter of one-class SVM classifiers

the best results were achieved, and varying the value of γ of one-class SVM classifiers. The figure not only shows how higher values of γ lead to small increases in F-measure scores but, more importantly, with values of $\gamma > 0.4$ any form of representation of social tags outperforms full-text classification. Furthermore, binary representations of top-10 tags associated to resources become the best performing among the social classification schemes.

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, etc. 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.

Learning classifiers using collaboratively assigned tags also impacts on the dimensionality of the classification problem. In fact, tag-based classifiers extracted from the top 10 list of tags are learned in a smaller dimensional space than full-text classifiers and yet are better predictors as can be observed in the last results reported. Table 1 summarized the number of unique features, terms or tags according to the case, the classification problem have to deal with.

Figure 8 summarizes the performance of content and tags-based classifiers in terms of accuracy for $\upsilon = 0.1$, the parameter setting leading to the best results in most experiments. Confirming previous results, if the classifiers capability of making correct decisions is considered, tags-based classifiers outperformed full-text ones. Also, among tag-based classifiers those using for training the top 10 tags assigned to each resource were the ones of superior performance. Thus, top 10 tags offers good accuracy levels and, at the same time, an important reduction in leaning 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 9. For studying this aspect of classification, the 50 users were divided in five groups according to the amount of resources in their personomies. In the first group users having less than 300 annotated resources were placed, then users having from 300 to 600 resources, 600 to 1000, 1000 to 2000, and more than 2000 resources.

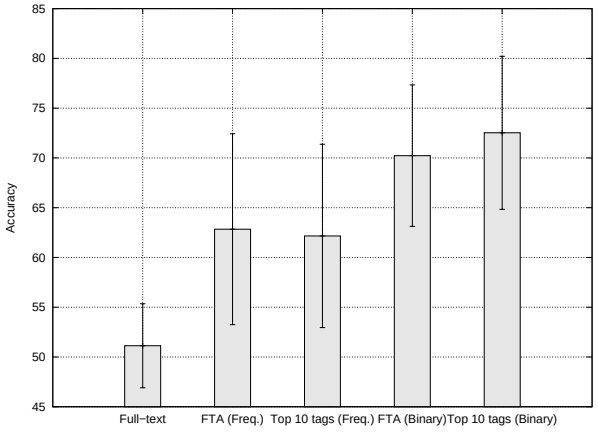

Figure 8: Accuracy of content and social tags classifiers

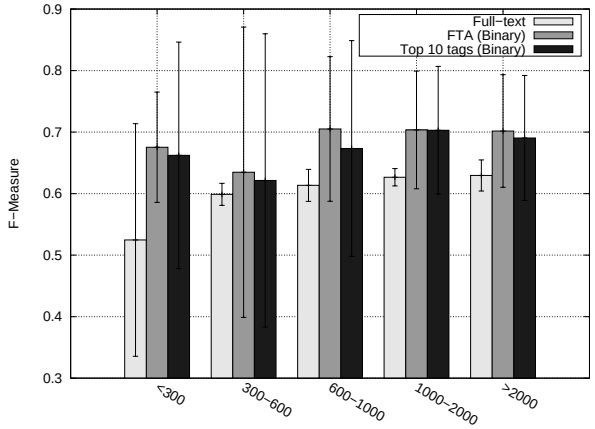

Figure 9: F-measure scores obtained for different personomy sizes

Naturally, as the number of annotated resources grows classifiers becomes increasingly better in filtering interesting Web pages as more information about the user interests is available during learning. However, the difference between social tags and full-text classifiers is more noticeable in smaller personomies, in which tags outperform text by a wider margin.

6 Related Works

Many works had approached the problem of tag recommendation in social tagging systems [Lipczak, 2008; Symeonidis *et al.*, 2008; Jäschke *et al.*, 2007; Milicevic *et al.*, 2010], however the problem of filtering resources according to the user interest had received less attention. To the best of our knowledge, no approaches have addressed this problem using one-class classification over social tags.

Vatturi et al. [Vatturi *et al.*, 2008] create a personalized ed tag-based recommender for each user consisting of two 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 [Ammari and Zharkova, 2009] an approach for filtering the blog posts that search engines retrieve is presented. SVM is used to train and build a predictive model for the targeted user; the retrieved posts are analyzed and classified by the predictive model. Finally, only the posts that are scored as relevant by the model are sent back to the user.

Tag-based profiling consisting on tag vectors in which tag weights are given by their frequency of occurrence in the resources a user tagged had been proposed in [Noll and Meinel, 2007]. In [Michlmayr and Cayzer, 2007], profiles are represented by graphs in which nodes correspond to tags and edges denote relationships between them. The idea of using semantic relationships among tags in tag-based profiles has also been explored in [Huang *et al.*, 2008]. In the work presented in this paper, one-class SVM classifiers can be seen as tag-based profiles for users.

The value of the collective knowledge encapsulated in social tags for classification of resources in general directories or categories was studied in several works, not from a personal perspective as in this work, but from a social point of view.

Zubiaga et al. [Zubiaga *et al.*, 2009] explore the use of Support Vector Machines (SVM) in the *Social-ODP-2k9* dataset, which links Web pages and tags assigned to them in *Del.icio.us* with their corresponding categories in a Web directory such as the *Open Directory Project* (ODP)[6]. In this work additional resource meta-data such as notes and reviews were also evaluated besides the tagging activity. Tags in conjunction with comments achieved good results for Web page classification.

Noll and Meinel [Noll and Meinel, 2008b] study and compare three different annotations provided by readers of Web documents, such as social annotations, hyperlink anchor texts and search queries of users trying to find Web pages, for classification. Coincidentally with our finding in the context of personal Web page classification, 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 last ones are more useful for information retrieval. In a further study [Noll and Meinel, 2008a], 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.

7 Conclusions

In this paper the role of social tags in filtering resources from folksonomies according to the interests of individual users was empirically analyzed. One-class classification was used to learn the user interests from diverse content sources (such as the full-text, anchor texts and titles) and social tagging sources (top 10 list of all tags associated to resources and their full tagging activity). Then, the extend to which each source can contribute to automatic, personal Web document classification was 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

[6] http://www.dmoz.org/

a discrete impact on classification performance. Interesting results were obtained using binary representations of tag vectors for learning and prediction. In this case, tag-based classifiers significantly improved the performance in filtering interesting results, even considering the top 10 tags assigned to resources in a quite smaller dimensionality space.

Acknowledgments

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

References

[Ammari and Zharkova, 2009] A. N. Ammari and V. V. Zharkova. Combining tag cloud learning with SVM classification to achieve intelligent search for relevant blog articles. In *Proceedings of the 1st International Workshop on Mining Social Media*, Sevilla, Spain, 2009.

[Arakji *et al.*, 2009] R. Arakji, R. Benbunan-Fich, and M. Koufaris. Exploring contributions of public resources in social bookmarking systems. *Decision Support Systems*, 47(3):245–253, 2009.

[Baeza-Yates and Ribeiro-Neto, 1999] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing, 1999.

[Cavnar and Trenkle, 1994] W. Cavnar and J. Trenkle. N-gram-based text categorization. In *Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, USA, 1994.

[Chang and Lin, 2001] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Echarte *et al.*, 2008] F. Echarte, J. Astrain, A. Córdoba, and J. Villadangos. Pattern matching techniques to identify syntactic variations of tags in folksonomies. In *Proceedings of the 1st World Summit on The Knowledge Society (WSKS '08)*, volume 5288 of *LNCS*, pages 557–564. Springer-Verlag, 2008.

[Golder and Huberman, 2006] S. Golder and B. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[Hotho *et al.*, 2006] A. Hotho, R. Jáschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference, ESWC 2006*, volume 4011 of *LNCS*, pages 411–426. Springer, 2006.

[Huang *et al.*, 2008] Y-C. Huang, C-C. Hung, and J. Yung-Jen Hsu. You are what you tag. In *AAAI Spring Symposium on Social Information Processing (AAAI-SIP)*, pages 36–41, 2008.

[Jäschke *et al.*, 2007] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *Knowledge Discovery in Databases: PKDD 2007*, volume 4702 of *LNCS*, pages 506–514, 2007.

[Lipczak, 2008] M. Lipczak. Tag recommendation for folksonomies oriented towards individual users. In *Proceedings of ECML PKDD Discovery Challenge (RSDC08)*, pages 84–95, Antwerp, Belgium, 2008.

[Manevitz and Yousef, 2002] L. M. Manevitz and M. Yousef. One-class SVMs for document classification. *Journal of Machine Learning Research*, 2:139–154, 2002.

[Mathes, 2004] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. Computer Mediated Communication, 2004.

[Michlmayr and Cayzer, 2007] E. Michlmayr and S. Cayzer. 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.

[Milicevic *et al.*, 2010] A. K. Milicevic, A. Nanopoulos, and M. Ivanovic. Social tagging in recommender systems: A survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review*, 33(3):187–209, 2010.

[Noll and Meinel, 2007] M. G. Noll and C. Meinel. Web search personalization via social bookmarking and tagging. In *Proceedings of 6th International Semantic Web Conference (ISWC) and 2nd Asian Semantic Web Conference (ASWC)*, volume 4825 of *LNCS*, pages 367–380, 2007.

[Noll and Meinel, 2008a] M. G. Noll and C. Meinel. Exploring social annotations for Web document classification. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC '08)*, pages 2315–2320, Fortaleza, Ceará, Brazil, 2008.

[Noll and Meinel, 2008b] M. G. Noll and C. Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 640–647, Sydney, Australia, 2008.

[Porter, 1980] M. Porter. An algorithm for suffix stripping program. *Program*, 14(3):130–137, 1980.

[Schölkopf *et al.*, 2001] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[Symeonidis *et al.*, 2008] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. Tag recommendations based on tensor dimensionality reduction. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08)*, pages 43–50, Lausanne, Switzerland, 2008.

[Tonkin and Guy, 2006] E. Tonkin and M. Guy. Folksonomies: Tidying up tags? *D-Lib*, 12(1), 2006.

[Vatturi *et al.*, 2008] P. K. Vatturi, W. Geyer, C. Dugan, M. Muller, and B. Brownholtz. Tag-based filtering for personalized bookmark recommendations. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*, pages 1395–1396, Napa Valley, California, USA, 2008.

[Zubiaga *et al.*, 2009] A. Zubiaga, R. Martínez, and V. Fresn. Getting the most out of social annotations for Web page classification. In *Proceedings of the 9th ACM Symposium on Document Engineering (DocEng' 2009)*, pages 74–83, Munich, Germany, 2009.