Exploiting the Social Capital of Folksonomies for Web Page Classification

Daniela Godoy[1,2] and Analía Amandi[1,2]

[1] ISISTAN Research Institute, UNICEN University
Campus Universitario, Paraje Arroyo Seco, CP 7000,
Tandil, Buenos Aires, Argentina
[2] CONICET, Buenos Aires, Argentina
{dgodoy, amandi}@exa.unicen.edu.ar

Abstract. Collaborative tagging systems (CTSs), also known as folksonomies, have grown in popularity on the Web and social tagging has become an important feature of many Web 2.0 services. It has been argued that the power of tagging lies in the ability for people to freely determine the appropriate tags for resources without having to rely on a predefined lexicon or hierarchy. The free-form nature of tagging causes a number of problems in this social classification scheme, such as synonymy and morphological variety. However, social tagging can be a valuable source of information to help in the organization of Web resources. In this paper we present an empirical analysis carried out to determine the importance of social tagging in Web page classification. Experimental results showed that tag-based classification outperformed classifiers based on full-text of documents.

Keywords: Social Tagging Systems, Web Page Classification, Folksonomies.

1 Introduction

Collaborative tagging systems (CTSs) are one of the most popularized content sharing applications associated with Web 2.0 and an important feature of many Web 2.0 services. The practice of collectively creating and managing tags to annotate and categorize content, has achieved widespread success on the Web as it allows to easily browse and search huge volume of shared resources. In sites such as *Del.icio.us*[1], *Technorati*[2] or *Flickr*[3] users annotate a variety of resources (Web pages, blog posts or pictures) using a freely chosen set of keywords in order to improve search and retrieval of such information.

Folksonomies [7] are the primary structure of the novel social classification scheme introduced by tagging systems and are usually contrasted with traditional pre-defined taxonomies used on the Web. This scheme relies on the convergence of

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

tagging efforts of a large community of users to a common categorization system that can be effectively used to organize and navigate large information spaces. In fact, the term folksonomy is a blend of the words taxonomy and folk, and stands for conceptual structures created by the people [6].

In spite of the novel mechanisms for finding interesting resources in vast information spaces provided by collaborative tagging systems, problems associated with the completely unsupervised nature of social tagging (ambiguity, synonymy, noise, etc.) may reduce its efficiency in content indexing and searching, thereby hindering the task of users.

In order to effectively organize on-line information, however, distributed classification provided by folksonomies might become an essential and value resource. In this sense, social tagging can help to automatize or assist the time consuming and laborious task of manually classifying documents into a set of pre-defined categories. Hammond et al. [5] and Guy and Tonkin [12] agree that tagging can play a complimentary role alongside more formal types of organization.

In this paper we empirically evaluate social tagging information, consisting of collaboratively generated, open-ended tags, to categorize content such as Web pages. For this purpose, we carried out a number of experiments using a collection of pages categorized by experts in a Web directory and the tag assignments given by non-expert users in a folksonomy. Several representations of tag vectors as well as pre-processing operations over tags were also evaluated.

The rest of the paper is organized as follows. Section 2 presents the empirical analysis carried out to compare content-based with tag-based classification of Web pages. Section 3 briefly reviews related research. Finally, our findings are summarized in Section 4.

2 Empirical Evaluation

To investigate the role social tagging can play in Web page categorization we performed an empirical analysis using *Social-ODP-2k9* [13] dataset, which basically links Web pages with their corresponding categories in a Web directory such as the Open Directory Project (ODP)[4] and tags assigned to these pages by users in a popular social bookmarking site such as *Del.icio.us*.

The Open Directory is one of the largest, most comprehensive human-edited directory of the Web. It is maintained by an international community of volunteer editors who evaluate sites for inclusion in the directory and categorize each Web document into one or more predefined and hierarchically organized categories. In *Del.icio.us*, in contrast, users store, share, and discover bookmarks using a non-hierarchical classification system in which they tag their bookmarks with freely chosen index terms.

Social-ODP-2k9 [13] is a dataset created during December 2008 and January 2009 with data retrieved from *Del.icio.us*, *StumbleUpon*[5], the ODP and the Web. This

[4] http://www.dmoz.org/
[5] http://www.stumbleupon.com/

dataset contains 12.616 unique URLs, all of them annotated by at least 100 users to ensure Web page popularity. For each of these Web pages the dataset has the corresponding category from the ODP as well as the number of users that annotate them, the top 10 list of assigned tags, the full tag activity (FTA) and also notes and reviews about the page.

In this paper we used the category assigned for ODP to Web documents as ground-truth or expert categorization, to compare with tagging of non-expert users provided by *Del.icio.us* folksonomy. For classification we used the 17 categories in the first level of the ODP taxonomy. In regards to social tagging information we used from the dataset both the list of the 10 most popular tags, including their number of occurrence, and the full tag activity (FTA) listing all tags assigned to a resource, which is limited to the latest 2000 users.

2.1 Content-based Classification

In order to establish the relative importance of content and tags in Web page classification, we first evaluated the performance of text classifiers over the full text of documents as well as their titles.

From the total 12.616 Web pages in the *Social-ODP-2k9* dataset, experiments reported in this paper were performed over the 9.611 English-written pages identified using the classification approach presented in [3]. The content of these Web pages was filtered using a standard stop-word list and the Porter stemming algorithm [11] was applied to the remaining words.

Table 1 summarizes the resulting Web pages collection and its main statistics. The notation used for reporting these statistics is based on the formal definition of folksonomies. These social structures 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. $Y \subseteq U \times T \times R$ [6]. In this folksonomy, $\prec$ is a user-specific sub-tag/super-tag-relation possible existing between tags, i.e. $\prec \subseteq U \times T \times T$.

Web pages were randomly partitioned into a training set of approximately 70% of the collection, 6.727 pages, and the remaining 30% for testing, 2.884 pages. In the experiments we compared the performance of three classifiers *k*NN (considering k=100), Naïve Bayes and SMO, a sequential minimal optimization algorithm for training a Support Vector Machines (SVM) classifier using a polynomial kernel [10]. For evaluating the classifiers we used the standard precision and recall summarized by F-measure as well as accuracy [2].

Figure 1 depicts the scores of F-measure achieved by the three classifiers per class and in average. SMO outperforms Naïve Bayes in all cases and *k*NN showed a very poor performance. As a result of the highly unbalanced distributions of examples in the 17 classes some categories obtained high scores, such as *Computers*, whereas classifiers failed to recognize examples in other categories, like *Adult*. Figure 2 shows the results when the classifiers were trained using the title of documents only. In this case, the results observed the same trends, although with inferior performance than full-text classification.

Table 1. Summary of statistics for Web pages in the dataset used for experimentation.

Category	|R|	# unique words in text content	# unique words in title	|T| in the top 10 lists	|T| in the FTA
Adult	25	2.841	94	25	2.920
Arts	920	46.006	1.657	1.524	62.081
Business	427	18.313	1.096	1.125	37.127
Computers	2.922	80.117	3.908	3.051	165.485
Games	232	18.953	544	566	15.571
Health	95	7.760	298	351	9.815
Home	216	15.750	536	529	19.773
Kids and Teens	311	13.383	651	778	30.839
News	115	11.812	265	305	16.857
Recreation	249	16.215	662	728	21.211
Reference	521	22.625	914	1.064	55.266
Regional	1.158	38.208	2.225	2.371	78.707
Science	662	31.248	1.214	1.333	58.376
Shopping	412	19.953	1.269	945	27.037
Society	842	37.335	1.572	1.654	65.275
Sports	84	12.039	271	269	4.901
World	420	27.616	1.248	1.271	45.160
All categories	9.611	199.968	10.865	8.932	426.086

2.2 Tag-based Classification

In social classification schemes, raw tags are noisy and inconsistent as they are not introduced according to a controlled vocabulary. In consequence, to determine whether tags are a valuable source of information for classification of Web page some filtering techniques were considered and compared.

Tags variations can be attributed to several factors [12,4]:

- compound words consisting of more than two words that are not grouped consistently. 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.

Other factors that might account for tag variations include misspelling and incorrect encodings.

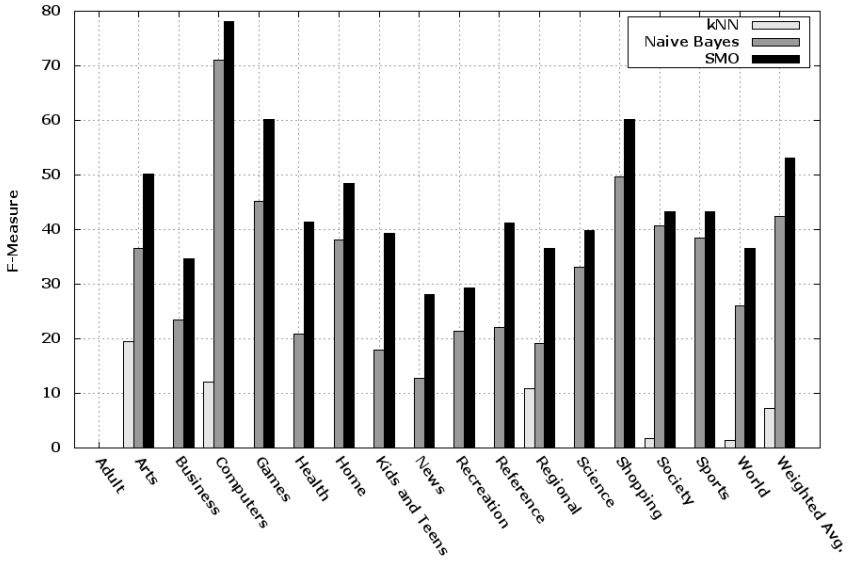

Fig. 1. F-measure scores per class for several text classifiers applied to Web page contents.

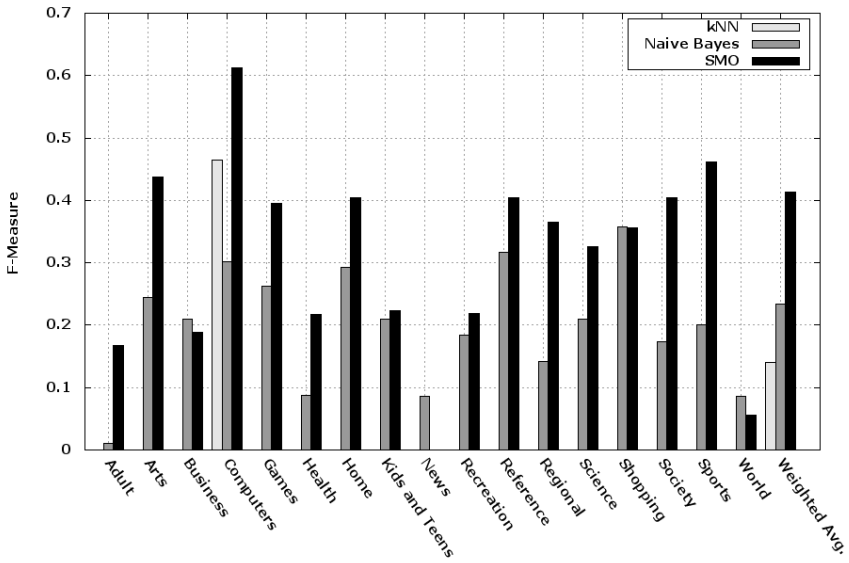

Fig. 2. F-measure scores per class for several text classifiers applied to the titles of Web pages.

Experimental evaluation was performed to determine the effect of two processing operations over tags tending to normalize them. The original raw tags were filtered to remove the symbols mentioned before as well as join compound words in a first processing considered. Then, tags were stemmed to their morphological roots using Porter stemmer algorithm.

In these experiments we also compared different weighting schemes for the resulting tags. Binary vectors were constructed to indicate the occurrence or non-occurrence of a tag in list of tags a Web page is annotated with. Tag frequency vectors indicate the number of users that used a given tag to annotate the resource, this is $(\log(1+f_{ij}))$ where f_{ij} is the frequency of the tag i in the document j, and these vectors can be normalized according to their average length.

Figure 3 shows the accuracy of classifiers leaned over the top 10 list of tags using SMO for the three representations of tag vectors considered after applying the mentioned processing operations. Tag frequency vectors performed slightly better than binary ones as well as normalized vectors, although differences are not significant. Regarding tag processing operations it can be deduced from these results that removing symbols and joining compound words reduces the noise of tags resulting in an improvement in accuracy. However, the use of stemming does not lead to a further improvement and can even reduce the precision of classifiers.

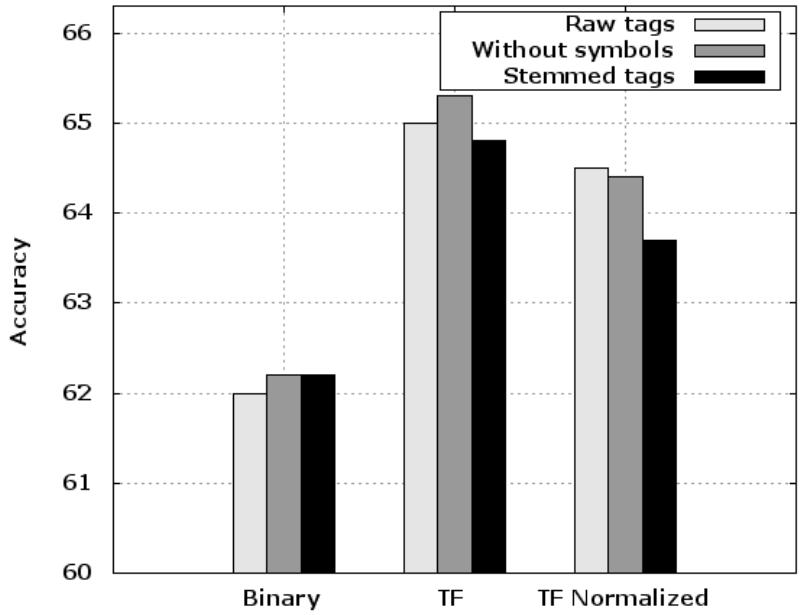

Fig. 3. Classification accuracy using tags in the top 10 list.

Figure 4 depicts the results using the same configuration of experiments but applied to tags in the full tag activity (FTA). FTA leads to a larger list of tags in spite of being limited to the 2000 last users. In consequence, normalization of TF vectors improves the performance of classifiers. In contrast to the previous experiments with the top 10 list of tags, raw tags in these experiments reach better accuracy levels than other representations, although differences are not significant. Stemming of tags in this situation seems to improve the accuracy of classification.

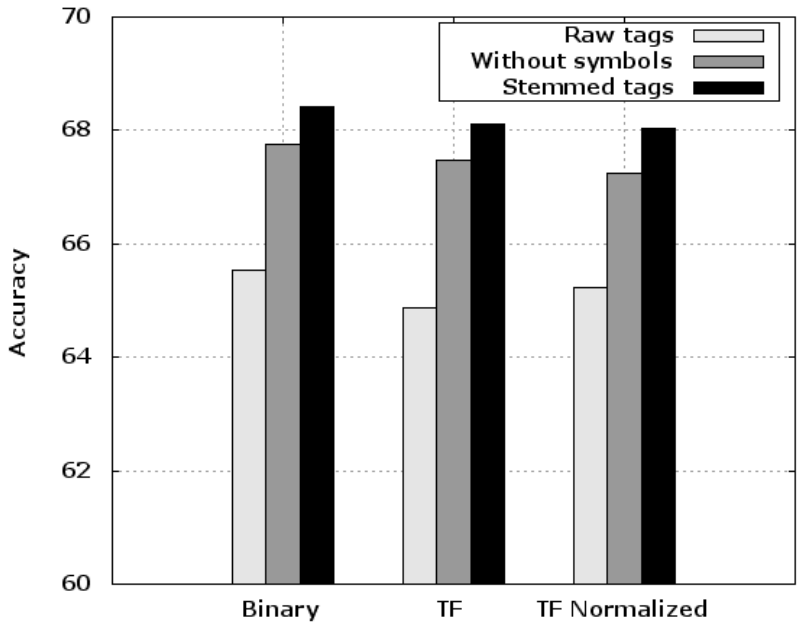

Fig. 4. Classification accuracy using tags in the FTA.

Figure 5 summarizes accuracy of the best performing classifiers learned based on the full-text of documents and their title. In both cases the best classification accuracy was achieved by SMO classifiers. Likewise, results using social tagging are showed for the top 10 list of tags associated to resources and the detailed tagging activity. These results correspond to SMO classifiers learned starting from tags. In one case, the best result was achieved filtering symbols and unifying compound words, while in the second case stemmed tags achieved better results.

Web page classification using collaboratively assigned tags outperforms both content-based classifiers in all scenarios. It is worth noticing that in addition to the improved performance, tag-based classifiers extracted from the top 10 list are learned in a smaller dimensional space than full-text classifiers. If the FTA tag is considered the dimensionality is in the same order of full-text of Web documents. Thus, it can be concluded that collective knowledge gathered in folksonomies becomes a valuable source of information for automatic classification of Web resources.

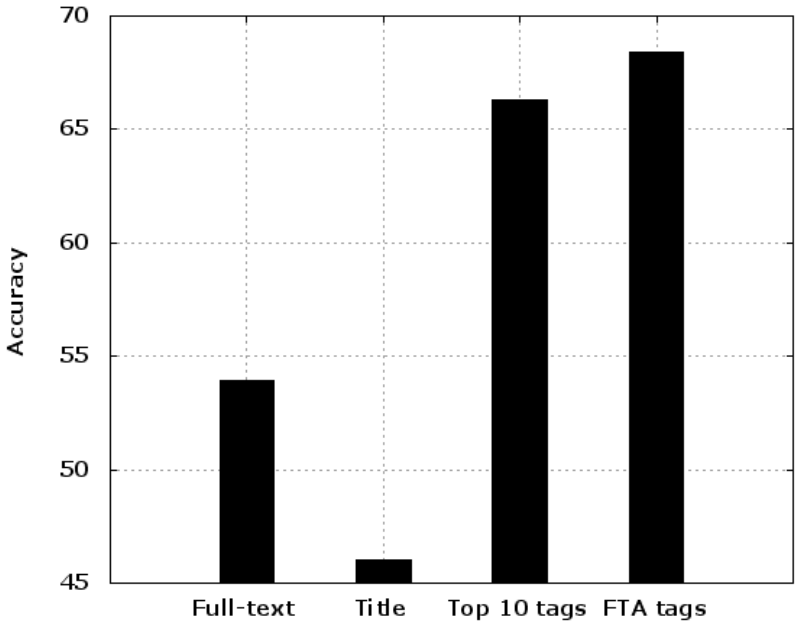

Fig. 5. Comparison of classification accuracy of content-based and tag-based classifiers.

3 Related Works

Zubiaga et al. [13] explore the use of Support Vector Machines (SVM) in the *Social-ODP-2k9* dataset. In this work additional resource metadata such as notes and reviews were evaluated in addition to tagging activity. Promising results were obtained using tags and comments for Web page classification. In contrast to this work, we concentrate the evaluation in the tagging activity of users within a folksonomy and evaluate the impact of some pre-processing techniques over tags tending to normalize them.

Noll and Meinel [9] study and compare three different annotations provided by readers of Web documents, social annotations, hyperlink anchor text and search queries of users trying to find Web pages. *CABS120k08*[6] dataset was created for such study from sources such as *AOL500k*, ODP, *Del.icio.us* and *Google* in general. 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.

[6] http://www.michael-noll.com/wiki/CABS120k08/

In a further study [8], the same authors analyzed at which hierarchy depth tag-based classifiers can predict a category using *DMOZ100k06*[7] dataset with information from ODP and *Del.icio.us*. 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.

Aliakbary et al. [1] propose a method for describing both Web pages and categories in terms of associated tags, and then to assign the resource to the category with the most similar tag-space representation. Experiments carried out with a set of Web pages from the Computers category of ODP showed that the method behave better than content-based classification.

4 Conclusions

In this paper we presented an empirical analysis to establish the ways in which social tagging can contribute to automatic Web document classification, then, helping to bridge the gap between the strict structure of taxonomies and the completely open nature of folksonomies. Experimental results obtained with a collection of Web pages categorized by experts in a Web directory not only showed that tag-based classifiers outperformed content-based ones in broad level categories, but also that pre-processing tags operations such as removal of symbols, compound words and reduction of morphological variants have a discrete impact on classification performance. The first operations improve classification accuracy in most experiments. However, stemming demonstrate to reduce accuracy when fewer tags were used (the top 10 tags assigned to resources), whereas it showed to be useful when more tags were involved in classification. Other operations over tags that we will investigate in future works are the correction of misspelled words and enconding variations. Even tough Web page organization in directories is an interesting application for tag-based classification, other prominent application in which social tagging can be exploiting is personalized Web classification (for example, interesting/uninteresting Web pages).

Acknowledgements

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

[7] http://www.michael-noll.com/wiki/DMOZ100k06

References

1. S. Aliakbary, H. Abolhassani, H. Rahmani, and B. Nobakht. Web page classification using social tags. In *Proceedings of the 2009 International Conference on Computational Science and Engineering (CSE '09)*, pages 588–593, 2009.
2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing, 1999.
3. 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.
4. 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)*, pages 557–564. Springer-Verlag, 2008.
5. T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11, 2005.
6. 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.
7. A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. Computer Mediated Communication, 2004.
8. 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, 2008.
9. M. G. Noll and C. Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 1:640–647, 2008.
10. J. Platt. Fast training of support vector machines using sequential minimal optimization. In C. Burges B. Schoelkopf and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.
11. M. Porter. An algorithm for suffix stripping program. *Program*, 14(3):130–137, 1980.
12. E. Tonkin and M. Guy. Folksonomies: Tidying up tags? *D-Lib*, 12(1), 2006.
13. 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.