Expert Systems with Applications 39 (2012) 9723–9729

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Evaluating tag filtering techniques for web resource classification in folksonomies

Nicolás Tourné [a], Daniela Godoy [b,c,*]

[a] Faculty of Exact Sciences, UNICEN University, Campus Universitario, 7000 Tandil, Bs. As., Argentina
[b] ISISTAN Research Institute, UNICEN University, Campus Universitario, 7000 Tandil, Bs. As., Argentina
[c] CONICET, Bs. As., Argentina

ARTICLE INFO

Keywords:
Web resource classification
Social tagging systems
Folksonomies

ABSTRACT

Social or collaborative tagging systems emerged as a novel classification scheme on the Web based on the collective knowledge of people. In sites such as *Del.icio.us*, *Technorati* or *Flickr*, users annotate a variety of resources, including Web pages, blogs, pictures, videos or bibliographic references; using freely chosen textual labels or tags. Underlying collaborative tagging systems are ternary data structures known as folksonomies relating resources and users through tags, this information facilitate accessing and browsing massive repositories of resources. Collective annotations provided by people in the form of tags can also be exploited to organize resources on-line in a more formal classification scheme such as the ones provided by hierarchies or directories, alleviating the task of manual classification commonly required by systems like directories on the Web. In this paper we present an empirical study carried out to determine the value of tags in resource classification. Furthermore, the use of several filtering and pre-processing operations to reduce the ambiguity and noise in tags are analyzed to determine whether they allow to increase the quality of resource classification.

1. Introduction

Social tagging systems are one of the most popularized content sharing applications associated with the emergent Web 2.0. The practice of collectively create and manage tags to annotate and categorize content has achieved widespread success on the Web due to its simplicity. In sites such as *Del.icio.us*,[1] *Technorati*[2] or *Flickr*[3] users annotate heterogeneous resources, including Web pages, blog posts, pictures or videos, using a freely chosen set of keywords or open-ended tags.

The tripartite data structure underlying collaborative tagging systems is known as folksonomy (Mathes, 2004) and relates resources, tags and users. The social classification scheme proposed by folksonomies contrasts with traditional predefined taxonomies or directories found on the Web. Whereas a taxonomy offers a rigid scheme of hierarchical categories, commonly established and populated with the help of human experts, folksonomies relies on the convergence of 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.

Although having different conceptions, both classification schemes coexist on the Web. In fact, to effectively organize on-line information into categories, distributed classification provided by folksonomies might become an essential and valuable source of information. Thus, social tagging can help to automatize or assist the time consuming and laborious task of manually classifying resources into a set of predefined categories. Moreover, social tags might enable the classification of resources with not associated textual content such as pictures or videos. Hammond, Hannay, Lund, and Scott (2005) and Guy and Tonkin (2006) agree that tagging can plays a complimentary role alongside more formal types of organization like hierarchical catalogs.

We carried out an empirical evaluation of using collaboratively generated, open-ended tags, to categorize resources such as Web pages. Experiments were based on a collection of pages categorized by experts in a Web directory and the tag assignments given by non-expert users in *CABS120k08*[4] (Noll & Meinel, 2008b) folksonomy. Multiple meta-data available were compared for obtaining representations of resources and evaluated with different classifiers.

One of the major problems related to folksonomies is the completely unsupervised nature of tagging, leading to problems such as ambiguity and noise in textual labels or tags. Thus, syntactical variations are common and can be attributed to several reasons, for example the use of synonyms, typographical misspellings and grammatical variations. The existence of tag variations not only causes an increase in the number of features to be considered during learning, but also a reduction in the performance of classifiers that considers them as different, independent tags. Consequently,

* Corresponding author at: ISISTAN Research Institute, UNICEN University, Campus Universitario, 7000 Tandil, Bs. As., Argentina.
E-mail address: daniela.godoy@isistan.unicen.edu.ar (D. Godoy).

[1] http://del.icio.us/.
[2] http://technorati.com/.
[3] http://www.flickr.com/.
[4] http://www.michael-noll.com/cabs120k08/.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.088

other goal of this work is to evaluate tag pre-processing operations to minimize effects of syntactic variations in tags with the aim of increasing the quality of resource classification. For this purpose, several filtering methods were evaluated over tags, including the use of stemming, synonyms and misspelling correction.

The rest of the paper is organized as follows. Section 2 presents background as well as related works regarding classification of Web resources based on social tags. Section 3 presents the empirical analysis carried out to evaluate tag-based classification of Web resources. Section 4 explores different processing operations to be performed over tags in order to determine their utility to improve classification results. Finally, Section 5 summarizes our findings.

2. Background and related works

Folksonomies are the primary structure of a social classification scheme which relies on the convergence of 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. This classification scheme is usually contrasted with the use of pre-defined taxonomies. Indeed, the term *folksonomy* is a blend of the words *taxonomy* and *folk*, and stands for conceptual structures created by the people (Hotho, Jáschke, Schmitz, & Stumme, 2006).

Formally, a folksonomy 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). In this folksonomy, $\prec$ is a user-specific sub-tag/super-tag-relation possible existing between tags, i.e. $\prec \subseteq U \times T \times T$.

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_i the projection on the ith dimension. In social tagging systems, tags are used to organize information, which is also shared, within a personal information space. Thus, other users can access a user personomy by browsing and searching the entire folksonomy using the available tags.

In addition to facilitate searching and browsing heterogeneous resources in folksonomies, tags can provide valuable information for other tasks such as classification, clustering (Lu, Hu, & Park, 2011; Ramage, Heymann, Manning, & Garcia-Molina, 2009) and recommendation of resources (Carmagnola, Cena, Cortassa, Gena, & Torre, 2007; Sen, Vig, & Riedl, 2009; Symeonidis, Nanopoulos, & Manolopoulos, 2010; Tso-Sutter, Marinho, & Schmidt-Thieme, 2008; Zheng & Li, 2011). Particularly, we will address the problem of tag-based classification to determine the categories resources belong to in a standard (flat or hierarchical) classification scheme. The possibility of exploiting the collective knowledge encapsulated in social tags for classification of resources into general directories or hierarchical categories is a problem that has been recently addressed in several works.

Noll and Meinel (2008b) studied and compared three different annotations provided by readers of Web documents, social annotations, hyperlink anchor texts and search queries of users trying to find Web pages. *CABS120k08* dataset, also used in this work, was created for such study from sources that included *AOL500k*, the Open Directory Project,[5] (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. In a further study (Noll & Meinel, 2008a), the same authors analyzed at which hierarchy depth tag-based classifiers can predict a category using *DMOZ100k06*[6] 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.

Zubiaga, Martínez, and Fresno (2009) explore the use of support vector machines (SVM) in the *Social-ODP-2k9*[7] dataset created with data retrieved from *Del.icio.us, StumbleUpon*,[8] the ODP and the Web. In this work additional resource meta-data such as notes and reviews were evaluated in addition to tagging activity. Tags and comments obtained promising results in Web page classification. Moreover, if the motivation for tagging is considered it was found that users can be discriminated as categorizers or describers (Körner, Kern, Grahsl, & Strohmaier, 2010), having the tags assigned for the first type of users a greater utility for classification as was demonstrated in Zubiaga, Körner, and Strohmaier (2011). In Godoy and Amandi (2010) multiple classifiers as well as the impact of some pre-processing techniques over tags were analyzed over the same dataset showing the superiority of SVMs. Aliakbary, Abolhassani, Rahmani, and Nobakht (2009) proposed 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.

In these studies tags demonstrate to be an important source of information for categorization, beyond the textual content of resources. Other works address the same problem but from a personal point of view, using social tags to classify resources for a individual user instead of organizing resources in general taxonomies or directories. Vatturi, Geyer, Dugan, Muller, and Brownholtz (2008) created a personalized tag-based recommender for each user consisting of two naïve Bayes classifiers trained over different time frames. 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. The role of social tags in the identification of interesting resources for a given user was also studied in Godoy (2010) using one-class SVMs since they show better performance than other classifiers in the task stated (Godoy, 2012). In all the mentioned works, tag-based classification improves the results of content-based classification.

3. Tag-based classification of web resources

This section describes the empirical study carried out to evaluate tag-based classification of resources. The dataset employed for experimentation is described in Section 3.1, the different information sources considered to represent documents are detailed in Section 3.2 and the results of using different classifiers and Web resource representations are summarized in Section 3.3.

3.1. Dataset description

CABS120k08 (Noll & Meinel, 2008b) is a dataset for research in Web 2.0 consisting of 117,434 documents with associated meta-data collected from multiple sources. Meta-data consists of an intersection of the AOL Search query log corpus *AOL500k* and the

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

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

[7] http://nlp.uned.es/social-tagging/socialodp2k9/.

[8] http://www.stumbleupon.com/.

Table 1
Summary of *CABS120k08* corpus (Noll & Meinel, 2008b).

Overview	Total	Comment
Total documents	117,434	
Total categories	84,663	
Total searches	2,617,326	
Total anchor texts	2,242,621	
Total users	3,383,571	
Total bookmarks	1,289,563	Unique: 9.1%
Total tags	3,383,571	Unique: 26.3%
Categorized documents	117,434	100.0%
Searched documents	117,434	100.0%
Anchored documents	95,230	81.1%
Bookmarked documents	59,126	50.3%
Tagged documents	56,457	48.1%

Open Directory Project (ODP), self-defined as the largest, most comprehensive human-edited directory of the Web.

Meta-data associated to documents obtained from different sources help to gain more knowledge about them. Particularly, the dataset compile several views of the documents: social annotations provided by readers of Web pages, hyperlink anchor text provided by authors of these documents, and search queries of users trying to find them on the Web. In addition, documents have been categorized in one or more categories from ODP, offering the hierarchical paths within the directory.

AOL500k corpus is one of the largest publicly available collections of search queries today (Pass, Chowdhury, & Torgeson, 2006). It consists of 20 million web queries collected from 650,000 users on AOL Search over three months in 2006. *CABS120k08* was created by an intersection of *AOL500k* and the Open Directory. Thus, only documents that were both searched for and subsequently visited (AOL500k) as well as categorized (Open Directory) were included. Table 1 summarizes *CABS120k08* statistical characteristics.

It can be observed in the table that all documents have an ODP category, whereas 50.3% have been found in *Del.icio.us* and a few of them have not tag assigned. To sum up, the meta-data used in this study are:

- *Tags*: include the full history of a social bookmark, this for each document in the dataset its full bookmarking history from HTML Web pages crawled from *Del.icio.us*.
- *Anchor texts*: defined as the text that appears within the bounds of an *HTML* $\langle a \rangle$ tag. In the dataset, up to 100 referring pages per document were processed.
- *Queries*: refers to all queries used in *AOL500k* corpus in which a given Web pages was present in the result set.

3.2. Web resource representation

The three different information sources were evaluated as a means to represent documents both alone (tags, query terms and anchor texts), all of the sources combined (queries + anchortexts + tags) and three other combinations (queries + anchortexts, queries + tags, anchortexts + tags). In the resulting datasets, stopwords were removed using a list of 600 words of English language, Porter stemming algorithm (Porter, 1980) was applied (the use of stemming is discussed in Section 4) and binary weights were assigned to terms.

In addition, the performance of two classifiers was compared for the classification in this task, naïve Bayes and SMO from *Weka*[9] library of machine learning algorithms. SMO is a sequential minimal optimization algorithm for training support vector machines (SVMs) classifier using a polynomial kernel (PolyKernel) or a radial basis function (RBFKernel) kernel.

For evaluating the classifiers we used the standard precision and recall, summarized by *F*-measure, and accuracy (Baeza-Yates & Ribeiro-Neto, 1999). Accuracy measures the proportion of correct decisions made by classifiers. Precision is the number of correct classified examples divided by the number of examples classified as belonging to the class and recall is the number of correct classified examples divided by the number of examples belonging to the class. In all experiments, the results of 10-fold cross-validation are reported.

3.3. Tag-based classification results

Figs. 1 and 2 show the results achieved using the mentioned Web page representations and classifiers in terms of accuracy and *F*-measure, respectively. With respect to the meta-data used to represent the Web pages, it can be observed in both figures that tag-based representations obtained better results that anchor-texts and queries, the last ones showing the poorest performance. Consequently, these element also affects negatively the performance of those combinations in which queries are included (queries + anchortexts, queries + tags and queries + anchortexts + tags). In most cases the combination of anchor-texts and tags outperforms the remaining ones. Among the classifiers, naïve Bayes performance was inferior to the two SMO variants, being PolyKernel the one reaching the highest accuracy and *F*-measure scores.

Fig. 3 depicts the evolution of accuracy as the size of the training size increases for the SMO PolyKernel classifier. Confirming previous results, the use of anchor-texts and tags is the one obtaining the best results, closely followed by the use of tags alone. We considered the results obtained with the anchortexts + tags representations as baseline for evaluating tag processing operations in the following section.

4. Evaluating tag processing approaches

In social classification schemes, tags are noisy and inconsistent as they are not introduced according to a controlled vocabulary. Variations in tags can be attributed to several factors (Echarte, Astrain, Córdoba, & Villadangos, 2008; Guy & Tonkin, 2006):

- 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;*

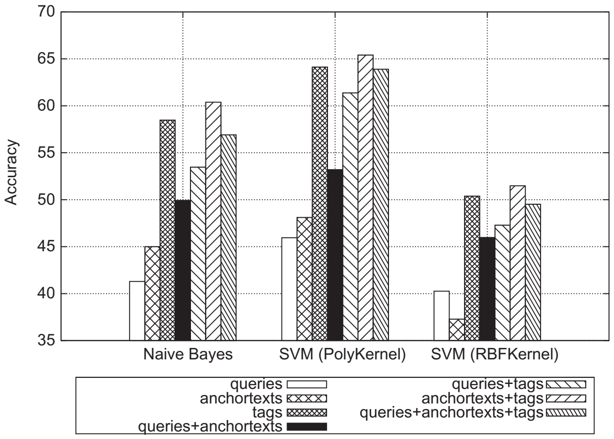

Fig. 1. Accuracy obtained using different representations and classifiers.

[9] http://www.cs.waikato.ac.nz/ml/weka/.

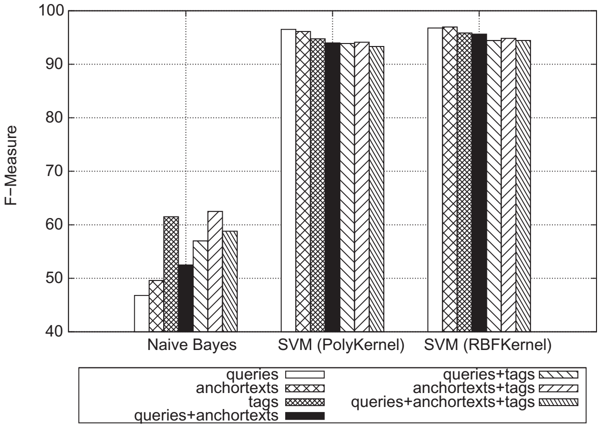

Fig. 2. *F*-Measure scores obtained using different representations and classifiers.

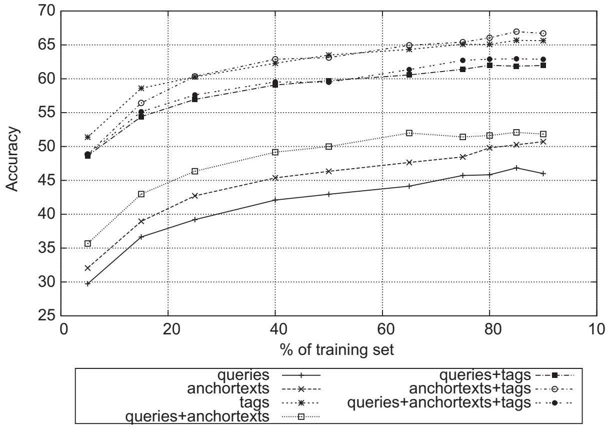

Fig. 3. Evolution of accuracy as the number of training examples increases.

- 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;
- The grammatical number used (singular or plural) and verbal times (gerunds, past and other forms), for example, *blog*, *blogs* and *blogging;*
- Typographical misspellings during the tagging process for example, *semntic Web* and *semntic Web*;
- Synonyms are different words used to express a same concept involved in an annotation or tag.

The reduction of these syntactic tag variations might help to improve the quality of folksonomies and, in turn, the classification of resources. Since tags shown to be a valuable source of information for Web page classification, several filtering techniques were considered and compared in this work in order to determine whether they can help to improve classification results.

Experimental evaluation with a dataset extracted from a widely used folksonomy, such as *Del.icio.us*, was carried out to determine the effect of different processing operations over tags tending to normalize them and avoid the mentioned problems. Initially, raw tags were filtered to remove the symbols enumerated before as well as join compound words. Then, three operations were considered as it is depicted in Fig. 4. Misspelling correction to fix typing errors, synonyms to considered other words with the same meaning and stemming to correct morphological variants caused by grammatical number and verbal forms.

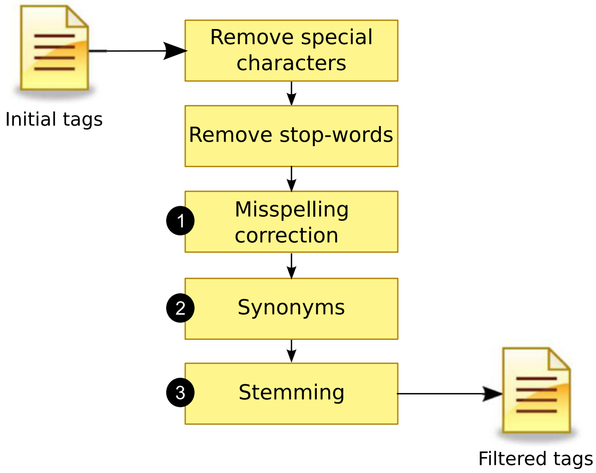

Fig. 4. Filtering process to improve Web page representations.

The capacity of each of these operations for improving Web page classification was evaluated separately. The results shown in Fig. 3 for the anchortexts + tags representation of resources were established as baseline. Then, each pre-processing operation was applied and the classification results compared against the baseline. The following sub-sections explain each of these operations and the results obtained.

4.1. Term stemming

In most languages, words have many morphological variants with similar semantic interpretations which can be treated as equivalents for information retrieval, as opposed to linguistic applications. For example, the words *computer*, *computers*, *compute*, *computes*, *computed*, *computational*, *computationally* and *computable* would be reduced to the single word stem *comput*. Thus, the dimensionality of the feature space can be reduced by mapping morphologically similar words onto their word stem.

This task is performed by stemming or conflation algorithms which are defined as processes of linguistic normalization in which morphological variants of words are reduced to their root form, called stem (Porter, 1980). The most common form of stemming is the elimination of suffixes and/or prefixes, such as the one used in Porter algorithm (Porter, 1980).

In spite of the benefits stemming algorithms can provide, stemmers can lead to a number of errors. These errors correspond to words with different meanings that are conflated to the same stem, which is known as over-stemming error. Also, stemming errors are caused by words with similar meanings that are converted into two different stems, which is known as under-stemming error.

To evaluate whether stemming improves the result of tag-based Web page classification, the use of stemming was compared in the same data without stemming. The results shown in the previous section were obtained using stemming. Fig. 5 compares the results obtained without using stemming with the results achieve before as the size of the training set increases. Clearly, the use of stemming significantly improves the results of Web resource classification.

4.2. Synonyms inclusion

In addition to syntactic variations of a same word, several word synonyms can be used by different users to annotate a resource. *WordNet*[10] (Miller, 1995), a large lexical database of English

[10] http://wordnet.princeton.edu/.

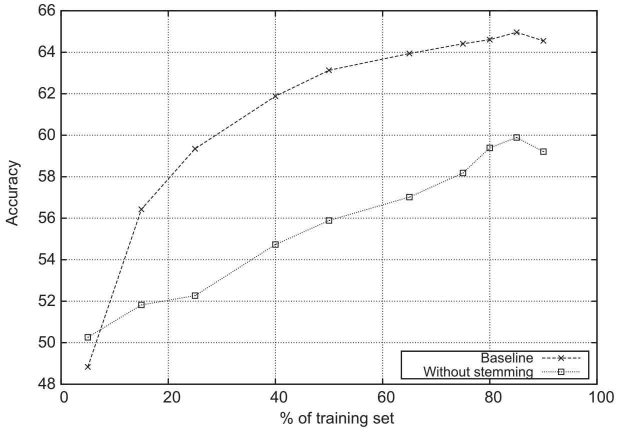

Fig. 5. Accuracy of classification using stemming.

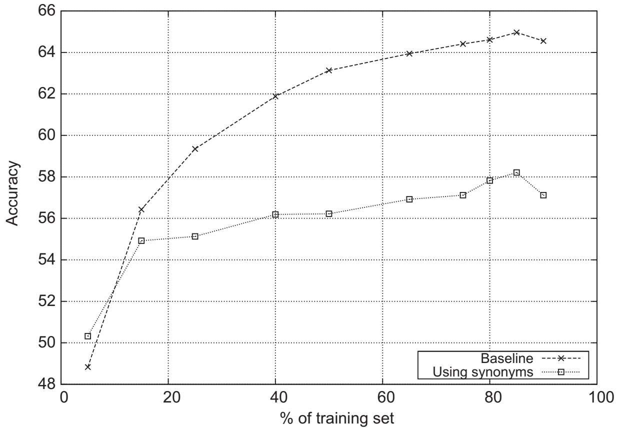

Fig. 6. Accuracy of classification using synonyms.

language, was used to obtained tag synonyms. In *WordNet* English words are grouped into sets of synonyms called synsets, belonging to different categories (nouns, verbs, adjectives and adverbs), and it records various semantic relations between these synonym sets.

For each tag, synonyms were extracted from *WordNet* and added to the Web page representation, so that the semantic meaning of tags is enriched. Fig. 6 shows the comparison of Web page classification results using synonyms with respect to the baseline. The incorporation of synonyms in the representation of examples causes a degradation of classification performance. The inferior performance of classifiers using synonyms can be attributed to the lack of context to disambiguate tag meaning and the consequent incorporation of noise to tagging activity. Other semantic operations over tags should be analyzed to be considered in the context of resource classification for gleaning semantically richer resource representations (Garcia-Silva, Corcho, Alani, & Gomez-Perez, 2012).

4.3. Misspelling correction

In this work, spell-checking is performed using three libraries based on different algorithms and dictionaries: *Tumba!*,[11] *JaSpell!*[12] and *Hunspell*.[13] The spell-checker was applied to each tag and those wrongly written were replaced by correctly spelled ones suggested by each algorithm. If there is not suggested word to replace a misspelled tag, likely because the tag does not exist in the spell-checker dictionary, the tag is discarded.

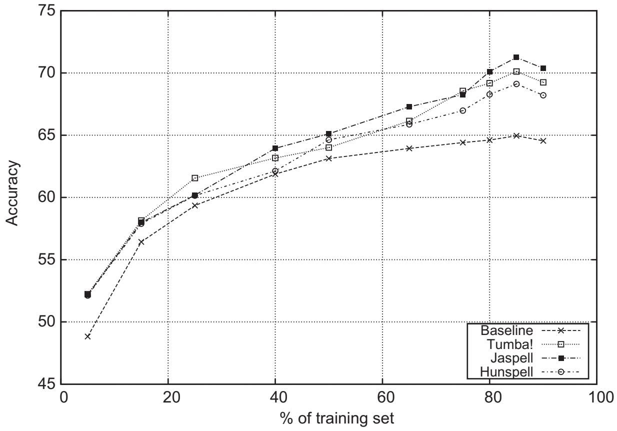

Fig. 7. Accuracy of classification after checking spelling.

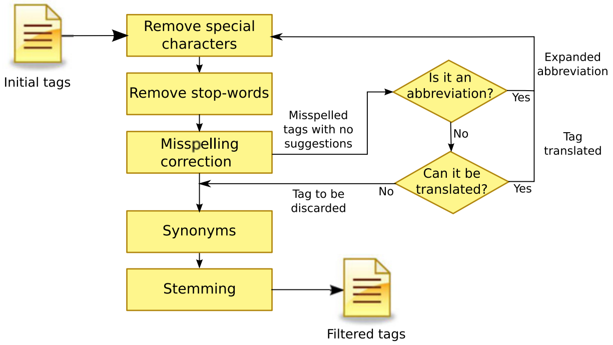

Fig. 8. Filtering process to improve Web page representations with enhanced misspelling correction.

Fig. 7 shows the result of Web page classification using the spell-checkers. Clearly, the use of any of the three algorithms results in an improvement of classification accuracy. *JaSpell* seems to have a slight advantage with respect to the other two spell-checkers in this regard.

The previous approach for processing misspelled words imply a loss of information as many tags were discarded (approximately 12%) when no suggestion was found to replaced them. However, a better treatment of such tags can lead to a further improvement in classification results. We observed that most of the eliminated tags corresponds to either abbreviations or non-English words. Then, both cases were considered to define an enhanced misspelling correction method.

Fig. 8 illustrates the resulting enhanced method for misspelling correction. Misspelled tags for which the spell-checker does not have any suggestion to offer, are first checked against a list to see if they correspond to an abbreviation. The Oxford English Dictionary: List of Abbreviations[14] was used for these experiments. For terms not found in this list of abbreviations, available translations are looked for. The *Google API Translate Java*[15] was employed for this task. Translated tags as well as expanded abbreviations are back to the step of removing characters and stop-words that might have

[11] http://xldb.fc.ul.pt/wiki/Tumba!.

[12] http://jaspell.sourceforge.net/.

[13] http://hunspell.sourceforge.net/.

[14] http://www.indiana.edu/~letrs/help-services/QuickGuides/oed-abbr.html.

[15] http://code.google.com/p/google-api-translate-java/.

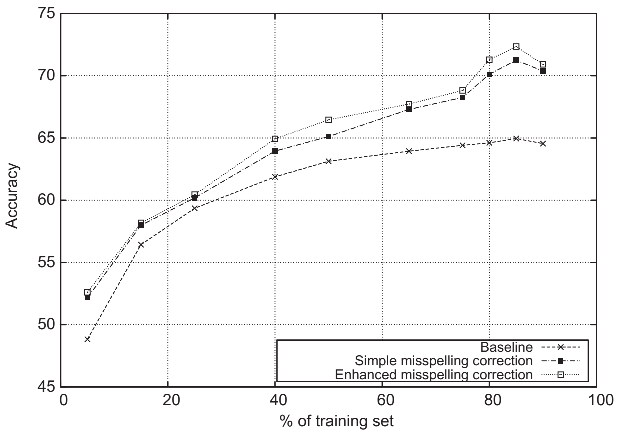

Fig. 9. Accuracy of classification using misspelling correction.

Table 2
Summary of classification results.

Pre-processing operation	Precision	Recall	*F*-Measure
Baseline	64.8	64.3	64.5
Without stemming	59.6	57.6	58.6
Synonyms	58.8	56.5	57.6
Simple misspelling correction (JaSpell)	76.0	73.5	74.7
Enhanced misspelling correction	**76.8**	**74.2**	**75.5**

been incorporated in this process. If no translation is found, the tag is finally discarded.

The results of Web resource classification using the enhanced method for misspelling correction are depicted in Fig. 9. Classification accuracy improves with the expansion of abbreviations and translation of non-English words since approximately 1.7% terms were recovered using this method. Table 2 summarizes the results of all of the proposed filtering or pre-processing operations evaluated for tags in terms of precision, recall and *F*-measure. Each row shows the results with respect to the baseline in the first row of the table. In bold, the best results were obtained with the enhanced method for misspelling correction is applied. It is worth noticing that each processing operation was evaluated separately, so that applied together are expected to lead to even better performance in classification.

5. Conclusions

Social tags consist in collective knowledge stored in the folksonomies collaborative tagging systems are based on, which is mostly used to easy access and browse shared resources. However, the use of tags for classifying resources can also help to bridge the gap between the strict structure of taxonomies and the completely open nature of folksonomies. Organizing on-line resources into a set of flat or hierarchical categories based on the tags assigned to them by users is a valuable tool for reducing the effort required to create catalogs, such as Web directories, commonly done by human experts. Moreover, it becomes essential for the classification of non-textual resources for which content-based classification is not possible (e.g. pictures or videos). Other prominent application in which social tagging can be exploiting is personalized Web classification, for identifying example, interesting/uninteresting Web pages.

Experiments were carried out using a standard dataset in the area providing multiple meta-data information about Web resources such as query terms, anchor-texts and tags. First, baseline results were obtained by comparing the scores achieved using the mentioned meta-data alone and combined with each other to represent resources and several classification algorithms. Second, preprocessing operations were evaluated to improve the quality of classification by reducing ambiguity and noise in tags. The use of stemming to reduce morphological variations have a positive impact in classification as well as a simple misspelling correction method, achieving the best results with a enhanced misspelling correction approach that includes the expansion of abbreviations and the translation of non-English tags. In contrast, the simple incorporation of synonyms to resource representations degrades the performance of classifiers. Other methods to add semantics to tags need to be explored in this direction and are planned to be investigated in future works.

Acknowledgments

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

References

Aliakbary, S., Abolhassani, H., Rahmani, H., & Nobakht, B. (2009). Web page classification using social tags. In *Proceedings of the 2009 international conference on computational science and engineering (CSE '09)* (pp. 588–593).

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing.

Carmagnola, F., Cena, F., Cortassa, O., Gena, C., & Torre, I. (2007). Towards a tag-based user model: How can user model benefit from tags? In *Proceedings of the 11th international conference on user modeling (UM2007). LNCS* (4511, pp. 445–449).

Echarte, F., Astrain, J., Córdoba, A., & Villadangos, J. (2008). In *Pattern matching techniques to identify syntactic variations of tags in folksonomies. Proceedings of the 1st world summit on the knowledge society (WSKS '08)* (pp. 557–564). Athens, Greece: Springer-Verlag.

Garcia-Silva, A., Corcho, O., Alani, H., & Gomez-Perez, A. (2012). Review of the state of the art: Discovering and associating semantics to tags in folksonomies. *Knowledge Engineering Review, 27*(1), 57–85.

Godoy, D. (2010). On the role of social tags in filtering interesting resources from folksonomies. In *Proceedings of the 18th international workshop on personalization and recommendation on the web and beyond (ABIS 2010)* (pp. 295–301) Kassel, Germany.

Godoy, D. (2012). Comparing one-class classification algorithms for finding interesting resources in social bookmarking systems. In *Resource Discovery. LNCS* (6799, pp. 88–103). Springer.

Godoy, D., & Amandi, A. (2010). In *Exploiting the social capital of folksonomies for Web page classification. Proceeding of the 10th IFIP WG 6.11 conference on e-business, e-services, and e-society (I3E 2010)* (pp. 151–160). Buenos Aires, Argentina: Springer.

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

Hammond, T., Hannay, T., Lund, B., & Scott, J. (2005). Social bookmarking tools (I): A general review. *D-Lib Magazine*, 11.

Hotho, A., Jáschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In *The semantic web: Research and applications, 3rd european semantic web conference (ESWC 2006). LNCS* (4011, pp. 411–426). Springer.

Körner, C., Kern, R., Grahsl, H. -P., & Strohmaier, M. (2010). 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)* (pp. 157–166) Toronto, Ontario, Canada.

Lu, C., Hu, X., & Park, J. (2011). Exploiting the social tagging network for Web clustering. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 41*(5), 840–852.

Mathes, A. (2004). Folksonomies – cooperative classification and communication through shared metadata. Computer Mediated Communication. URL <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Noll, M. G., & Meinel, C. (2008a). Exploring social annotations for Web document classification. In *Proceedings of the 2008 ACM symposium on applied computing (SAC '08)* (pp. 2315–2320) Fortaleza, Ceara, Brazil.

Noll, M. G., & Meinel, C. (2008b). The metadata triumvirate: Social annotations, anchor texts and search queries. In *Proceedings of the IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology* (pp. 640–647), Los Alamitos, CA, USA.

Pass, G., Chowdhury, A., & Torgeson, C. (2006). A picture of search. In *Proceedings of the 1st international conference on scalable information systems (InfoScale '06)* Hong Kong.

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

Ramage, D., Heymann, P., Manning, C. D., & Garcia-Molina, H. (2009). Clustering the tagged Web. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09) (pp. 54–63) Barcelona, Spain.

Sen, S., Vig, J., & Riedl, J. (2009). Tagommenders: Connecting users to items through tags. In *Proceedings of the 18th international conference on world wide web (WWW'09)* (pp. 671–680) Madrid, Spain.

Symeonidis, P., Nanopoulos, A., & Manolopoulos, Y. (2010). A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis. *IEEE Transactions on Knowledge and Data Engineering, 22*(2), 179–192.

Tso-Sutter, K. H. L., Marinho, L. B., & Schmidt-Thieme, L. (2008). Tag-aware recommender systems by fusion of collaborative filtering algorithms. In *Proceedings of the 2008 ACM symposium on applied computing (SAC '08)* (pp. 1995–1999).

Vatturi, P. K., Geyer, W., Dugan, C., Muller, M., & Brownholtz, B. (2008). Tag-based filtering for personalized bookmark recommendations. In *Proceeding of the 17th ACM conference on information and knowledge management (CIKM'08)* (pp. 1395–1396) Napa Valley, California, USA.

Zheng, N., & Li, Q. (2011). A recommender system based on tag and time information for social tagging systems. *Expert Systems with Applications, 38*(4), 4575–4587.

Zubiaga, A., Martínez, R., & Fresno, V. (2009). Getting the most out of social annotations for Web page classification. In *Proceedings of the 9th ACM symposium on document engineering (DocEng' 2009)* (pp. 74–83) Munich, Germany.

Zubiaga, A., Körner, C., & Strohmaier, M. (2011). Tags vs shelves: From social tagging to social classification. In *Proceedings of the 22st ACM conference on hypertext and hypermedia (HT'11)* (pp. 93–102), Eindhoven, Netherlands.