Semantic enrichment of social annotations for Web resource classification

Antonela Tommasel and Daniela Godoy*

ISISTAN Research Institute, UNICEN University
Campus Universitario, B7001BBO, Tandil, Bs. As., Argentina
*Also at CONICET (National Council for Scientific and Technical Research)
tommantonela@gmail.com

Abstract Social annotations voluntarily provided by users in tagging or bookmarking sites such as Delicious or Flickr have been recognized as an interesting source of metadata for assisting tasks such as classification of Web resources. However, the open-ended nature of the tags employed to annotate resources leads to problems such as the introduction of noise and ambiguity that may hinder classification results. This paper presents an approach for semantically analyse social annotations in order to attain enriched, concept-based representations of Web resources. Experimental results showed that the strategies proposed to relate tags to conceptual entities allow to improve the results of resource classification.

1 Introduction

Social annotations, also known as tags, are a phenomenon that allows people to describe resources by adding metadata collaboratively. Web sites such as *Delicious*[1] or *Flickr*[2] allow users to add tags to resources (e.g. Web pages, pictures or videos) to ease their further search and retrieval. Social tagging has had an immediate success, mainly because no special knowledge or abilities are required to use tags as they are unstructured terms or labels. Thus, social annotations have leaded to the existence of a large amount of metadata for each Web resource.

As the new technologies and software for social annotation of resources evolved, a new phenomenon appeared on the Web. Folksonomies [10], are the result of free and non-hierarchical annotations of resources in a social environment that contrasts with taxonomies, traditionally associated with a systematic and hierarchical ordering aiming to categorize documents or Web pages (e.g. Web directories).

The impact of social annotations for classifying resources into pre-defined categories, taking advantage of the collective knowledge available in folksonomies, started to be addressed by recent works. The capability of tags to replace the content of resources for classification [1,18,19] as well as the analysis of the distribution of tags and the different motivations of users to annotate resources [8,12] have been considered in different studies. All studies aim to answer the following question: How can the resource classification be improved by using their metadata?

[1] http://delicious.com
[2] http://www.flickr.com/

Even though tags are a valuable source of information for classification, tag-based classification has some drawbacks. Since there are no constrains on the terms that can be used for tagging, users can freely choose the scope, sense or generality for a resource characterization. In consequence, issues like synonymia, hypernymy and morphological variations appear, negatively affecting the results of classification. This work address the problem of resource representation in social tagging systems from a semantic point of view. Semantic information is associated to tags in order to increase their descriptive power and solve issues stemming from natural language ambiguity, as the ones mentioned before. Several strategies for relating tags to concepts are evaluated and compared with tag-based representations.

The article is organized as follows. Section 2 summarizes related research. Section 3 presents the proposed approach for associating semantics to social annotations and, thus, enriching resource representations. Section 4 reports experimental results. Finally, conclusions are stated in Section 5.

2 Related Works

The problem of resource classification using social annotations has been addressed in numerous works. For example, [12], [18] and [1] studied metadata like tags and comments, and their usefulness for resource classification. Yin et al. [16] proposed the use of tags as a semantic enrichment of non-textual Web objects like products, images or videos for a further classification. In addition, the authors developed different strategies for weighing tags and stated the need of reducing their amount aiming to lower the computational cost associated to classification. Other studies [8,17,12] focused on understanding the underlying motivation of users behind tagging to infer which users lead to better predictions. In contrast with these approaches, that only perform a syntactic pre-processing of tags before classification (like stemming), in this work we associate semantic to tags in order to overcome the problems caused by natural language as well as decreasing the amount of information needed to effectively classify resources.

Two approaches are commonly used to identify semantics for tags. The first approach is based on using clustering techniques to distinguish related group of tags and, thus, expose their meaning [5,4]. The second approach associates semantic entities (concepts) to tags using ontologies and establish relations between them [6,7].

Term clustering uses statistics methods to compute tag similarity. In [5], the authors use matrices to build a semantic space where related terms and documents are placed together even when the terms do not appear in the same document. The obtained relations do not generate concepts as the ones defined by using a lexical database or ontology like we propose, in consequence it is difficult to generalise their “semantic space”. In addition, the approach fails to solve the homonym problem which in turn affects term disambiguation. Since each term is represented only once, its weight comprises all the senses weights, affecting negatively those cases in which the correct sense is not the most popular one, leading to distortions and low accuracy.

Lexical databases are used to extract concepts to include semantic information in tasks like resource classification. *Katoa* [7] is a tool that adds semantic information

to a text using *Wikipedia*[3] and *WordNet*[4] as semantic sources. However, the strategies used by *Katoa* have some shortcomings. First, the tool fails to produce results for most part of the input, negatively affecting the classification due to the loss of information. Oppositely, we define strategies to benefit from the lexical resources aiming to widen their coverage and, in turn, improve classification results. For example, *Katoa* only implements disambiguation based on the most common sense (the first returned by *WordNet*) whereas we adopt the strategies proposed by [6] to define the representation of resources and to address ambiguity.

Finally, there are also hybrid approaches like [4] based on creating bi-graphs where nodes represent tags and links represent the co-occurrence of such tags on the different resources. In addition, the authors try to detect synonyms and homonyms by using different heuristics based on distance metrics and *WordNet* synonym identification. Unfortunately, the heuristics with the exception of the one using *WordNet*, suffer the same problems described for the clustering approach.

3 Associating Semantic to Social Annotations

In this paper we propose a method aiming to improve the classification of resources belonging to a folksonomy by the semantic enrichment of tags assigned to them by users. From tag-based representations exploiting social annotations to describe resources, enhanced concept-based representations are gleaned by relating tags to concepts in *WordNet* dictionary. *WordNe*t [11] is a large lexical database of English language which groups words into sets of synonyms called synsets and describes various semantic relations between these synonym sets. For that purpose, different strategies for incorporating semantics to tag-based representations of resources (Section 3.1) and finding conceptual entities for tags (Section 3.2) are proposed.

3.1 Resource Representation Strategies

The tag-based representation of a resource is formally defined as $R = \{t_r\}$ where R is the resource being analysed and t_r is the set of tags that users have assigned to it or annotated the resource with. Each tag t_r has also an associated weight w_{tr} according to its importance in the resource representation. Finally, the function relating terms of lexical entries in *WordNet* with their corresponding concepts is denoted $Ref_C(t)$.

Aiming to improve the classification results using *WordNet*, the strategies presented in [6] are used. According to the authors, the enrichment of the term sets using the ontology proposed in *WordNet* has two benefits. First, it resolves synonyms. Second, it introduces more general concepts which can help with the identification of new related topics. To incorporate the information extracted from *WordNet*, three strategies are proposed based on adding or replacing tags by concepts.

[3] http://www.wikipedia.org/
[4] http://WordNet.princeton.edu/

Expanding the Tag Set The first strategy consists in the expansion of the tag set $\{t_r\}$ with the new entries for the set of concepts $\{c_t\}$ obtained from each one of the existing tags. The original set is replaced by the set containing the original tags and the *WordNet* concepts: $\{t_r\} \cup \{c_t\}$. Those tags that do not have a *WordNet* representation, continue to belong to the resulting set.

This strategy allows the existence of repeated terms. Each tag that has a *WordNet* entry, appears at least twice in the new representation, once as part of the former $\{t_r\}$ and at least once as part of $\{c_t\}$. Those situations required a modification on the weight associated with the concepts, which is calculated as an addition of the weights, excepting the case of the relative weighting where it is recalculated to adjust the results into the corresponding range of values, as it is explained in Section 4.2.

For example, let suppose a resource annotated with the tags *business*, *ruby* and *web2.0*. Table 1 shows the set of *WordNet* concepts for each of these tags senses.

business	1. business, concern, business concern, business organization, business organisation – (a commercial or industrial enterprise and the people who constitute it) 2. commercial enterprise, business enterprise, business – (the activity of providing goods and services involving financial and commercial and industrial aspects) 3. occupation, business, job, line of work, line – (the principal activity in your life that you do to earn money) 4. business, business sector – (business concerns collectively) 5. clientele, patronage, business – (customers collectively) 6. business, stage business, byplay – (incidental activity performed by an actor for dramatic effect)
ruby	1. ruby – (a transparent piece of ruby that has been cut and polished and is valued as a precious gem) 2. ruby – (a transparent deep red variety of corundum; used as a gemstone and in lasers) 3. crimson, ruby, deep red – (a deep and vivid red color) 1. red, reddish, ruddy, blood-red, carmine, cerise, cherry, cherry-red, crimson, ruby, ruby-red, scarlet – (of a color at the end of the color spectrum (next to orange); resembling the color of blood or cherries or tomatoes or rubies)
web2.0	

Table 1: *WordNet* entries for the tags

Table 2 shows the concept set associated to each tag according to the concepts in Table 1 and without any disambiguation strategy. The first tag has repeated concepts in its different senses, but in the final representation of the resource, each concepts appears only once. The repetition shown here affects weighting, as it was previously stated.

Replacing Tags with Concepts The second strategy is similar to the first one. The only difference is that when a tag has an entry in *WordNet* it is removed from $\{t_r\}$ and replaced by the *WordNet* concepts. Those tags that do not have an entry on *WordNet* remain in the result set without changes. The resulting set is defined by $\{t_r\} \cup \{c_t\} - t_1 \in$

business, business, concern, business concern, business organization, business organisation, commercial enterprise, business enterprise, business, occupation, business, job, line, line of work, line, business, business, business, business sector, clientele, patronage, business, business, stage business
ruby, ruby, ruby, crimson, ruby, deep red, red, reddish, ruddy, blood-red, carmine, cerise, cherry, cherry-red, crimson, ruby, ruby-red, scarlet
web2.0

Table 2: Example of the resul set using the expanding strategy

$T : \exists Ref(t_1)$ where $Ref(t_1)$ represents the set of *WordNet* concepts for the tag $t_1$and T is the tag set for the resource.

Considering the same tags as in the previous example and using the concepts in Table 1, Table 3 shows the resulting set. In this case, just one of the original tags is part of the final set.

business, concern, business concern, business organization, business organisation, commercial enterprise, business enterprise, business, occupation, business, job, line of work, line, business, business, business sector, clientele, patronage, business, business, stage business
ruby, ruby, crimson, ruby, deep red, red, reddish, ruddy, blood-red, carmine, cerise, cherry, cherry-red, crimson, ruby, ruby-red, scarlet
web2.0

Table 3: Example of the result set using the replace strategy

Concept Set Only The last strategy totally replace the original tag set with the representations from *WordNet*. Those tags with no entry on *WordNet*, do not appear in the final representation. The resource set is then defined as $\{c_t\}$.

Considering the tags from the first example and the concepts associated with them shown in Table 1, Table 4 shows the resulting concept set. As it can be seen, the tag that does not have an entry on *WordNet* is not included in the resource representation.

business, concern, business concern, business organization, business organisation, commercial enterprise, business enterprise, business, occupation, business, job, line of work, line, business, business, business sector, clientele, patronage, business, business, stage business
ruby, ruby, crimson, ruby, deep red, red, reddish, ruddy, blood-red, carmine, cerise, cherry, cherry-red, crimson, ruby, ruby-red, scarlet

Table 4: Example of the result set using the concept set only strategy

3.2 Matching Tags with Concepts

Finding external semantic entities or concepts within *WordNet* a tag is referring to involves first the disambiguation of possibly polysemous tags, i.e. terms having multiple meanings. For example the term *business* has different senses, it can refer to the volume of commercial activity or to an immediate objective, among others. As a consequence, adding or replacing terms can add noise to the resource representation. Three alternatives for selecting an appropriate sense for a concept are considered in this work to solve this problem.

All Senses This strategy does nothing to solve the ambiguity, it simply considers all the extracted concepts from *WordNet* adding them to the final set or resource representation. The concept set associated to each tag is defined as $\{c_t\} = \bigcup\{Ref\,(t_n) : t_n \in T\}$. When this strategy is used, all the concepts in Table 1 for each tag are part of the final representation of the resource.

First Sense This strategy takes advantage of *WordNet* output that offers an ordered list of concepts associated to senses reflecting how common is the sense in the English language. Most common senses are listed before least common ones. When using this strategy, only the concepts from the first sense are added to the resulting set which is defined as $\{c_t\} = \bigcup\{first\,(Ref\,(t_n)) : t_n \in T\}$. In this case, not all the concepts presented in Table 1 remain in the final set, just the ones shown in the first position.

Context-based Disambiguation The last strategy performs a context-based disambiguation by implementing the Lesk algorithm [9]. The algorithm disambiguates terms appearing in small text fragments surrounding them, in the case of tags, the context is given by the other tags assigned to the resource. The definition of each tag sense is compared against all the senses of the other tags. The sense chosen is the one with more words in common with the other tag senses. For implementing this strategy it is necessary to detect the part of the speech (Noun, Verb, Adjective, Adverb) of each tag. The part of the speech is defined as the part of the speech of the first sense of the *WordNet* entry, thus, limiting the algorithm input to those tags with a *WordNet* entry.

4 Experimental Evaluation

4.1 Dataset Description

Social-ODP-2k9[5] [18] dataset was used for experiments, this dataset was created between December 2008 and January 2009 with data obtained from bookmarking sites like *Delicious*, *StumbleUpon*[6], the Open Directory Project and the Web.

The Open Directory Project, also known as DMoz, is the biggest directory edited by human beings, built and maintained by a global community of volunteers. Tags were

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

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

obtained from *Delicious*, a service that allows the storage of favourite Web sites, their categorization using tags, and sharing the bookmarks with other users.

The collection contains data of 12.616 URLs as well as their additional metadata. This includes the top 10 tags, which are the 10 most popular tags for each URL weighted according to the amount of users that have assigned the tag. Other metadata not used in these experiments are notes from *Delicious* and reviews from *StumbleUpon*. The total number of tags in the collection is 12.116, out of which 53,8% percent are unique, and each class has an average of 1.339 tags assigned to their resources.

For the selection of the URLs included in the collection, the authors have taken a list from *Delicious*, restricting the URLs to those sites that have been tagged by at least 100 users to guarantee the popularity of each Web site. The URL category was taken from the Open Directory Project (ODP), corresponding to one of the 17 categories on the first level of the taxonomy. In some cases, the URL contained more than one category, situation solved by selecting one of them randomly. Categories are not uniformly distributed in the collection. Figure 1 shows the distribution of the top level categories.

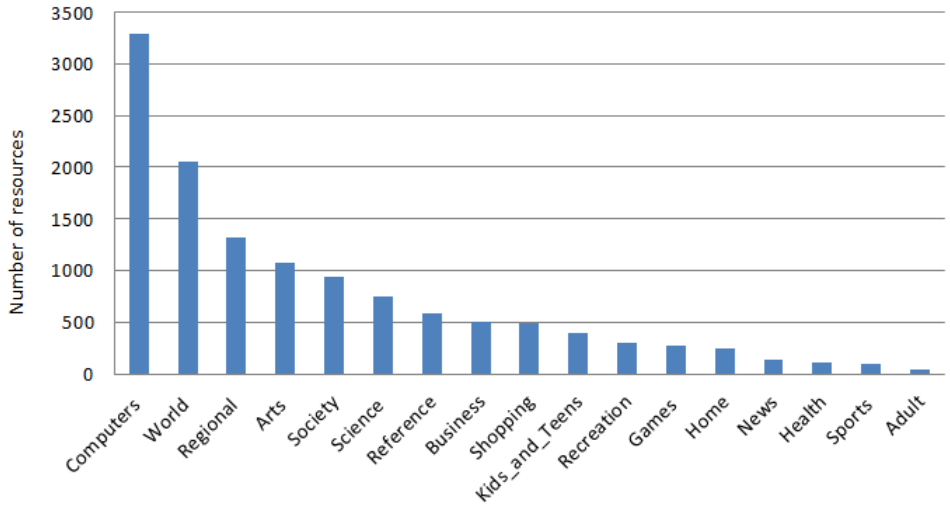


Figure 1: Category Distribution in the *Social-ODP-2k9* collection

4.2 Pre-processing

Each resource in the collection is represented by the tags assigned by users and its corresponding category taken from the highest level of the ODP taxonomy. For tags weighting, three alternatives were considered. The first uses a binary weighting in which a value of 1 indicates that the tag is used to annotate the resource and a value of 0 indicates that the tag is not used to annotate the resource. In the second, tags are weighed

according to the amount of times that users have assigned the tag to the resource, i.e. how many users annotate the resource with a given tag. Finally, the last alternative, uses a relative weighting for tags, i.e. the amount of times that users have assigned the tag divided by the total number of times that each tag was assigned.

Since the *WordNet* version used only contains English terms, those tags composed by characters that do not belong to the English alphabet, for example Kanjis or Russian alphabet, were deleted. If the name of the category had any of those characters, the example was discarded. In addition, two other alternatives were tested to deal with non-English tags. The first consisted in carrying out an idiom detection before continuing with the process of removing tags. The second alternative implements the Porter Stemmer algorithm [14] that removes automatically the word suffixes, and then performs idiom detection. The idiom detection was based on TextCat[7], a Perl implementation of the algorithm presented in [3] that recognises 69 idioms.

Idiom detection does not have a perfect accuracy, possible failings include a list with different several idiom alternatives or no idiom at all. Due to the fact that the identification is better as the length of the text increases, to avoid mistakes during the idiom detection, all the tags of each resource were given to the tool. If English is not in the top 3 idioms of the output, the document is removed from the analysis.

4.3 Methodology

A Sequential Minimal Optimization (SMO) [13] classifier, which is an optimization of the Support Vector Machines (SVMs) [15], was used to classify resources. SVMs are characterised for being a model that represents the sample points in the space, separating the classes with the widest possible margin. An accurate classification is defined by a wide separation between classes, in consequence, SVMs establish an optimal separation of points from different classes by creating hyper-planes. New instances are classified by means of the proximity to the points in the model. The SMO represents an alternative to SVM method as it allows an optimization in the computation of the solution space by analytical methods avoiding the generation of quadratic problems that introduce more computations, slowing down the execution. The WEKA[8] implementation of the algorithm was used in these experiments.

For evaluating the classifiers, the standard accuracy, precision and recall, summarized by F-measure, were employed [2]. In all cases classifiers were evaluated using a classical 10-fold cross-validation strategy.

First, the top-10 tags with the amount of times that users have assigned each of them and the category from the ODP were retrieved for each resource. Before semantic enrichment, the three alternatives of pre-processing described in Section 4.2 were evaluated: the iconography filter (Icon), the iconography filter with idiom recognition (IconIdioma) and the same filter using Porter stemming algorithm (IconPorterIdioma).

After the pre-processing tasks and the semantic enrichment of tags with concepts extracted from *WordNet*, different datasets were constructed for each possible com-

[7] http://odur.let.rug.nl/vannoord/TextCat/

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

bination of pre-processing, resource representation and disambiguation strategies. The classification algorithm was evaluated for the different types of attribute weighting described in Section 4.2. The baseline for comparing and evaluating the strategies are the results of resource classification based only in tags without considering any semantic information.

4.4 Experimental Results

The baseline results of resource classification are presented in Figure 2, these results allowed to evaluate the quality of the different pre-processing strategies when tag-based representations are considered. In this figure it can be observed that tag weighting using the tags absolute frequency is the worst performing. Figure 3 shows a comparison between the different pre-processing strategies for concept-based representation using the first sense disambiguation and the addition of concepts strategies during their construction using relative weighting in Fig. 3(a) and binary weighting in Fig. 3(b).

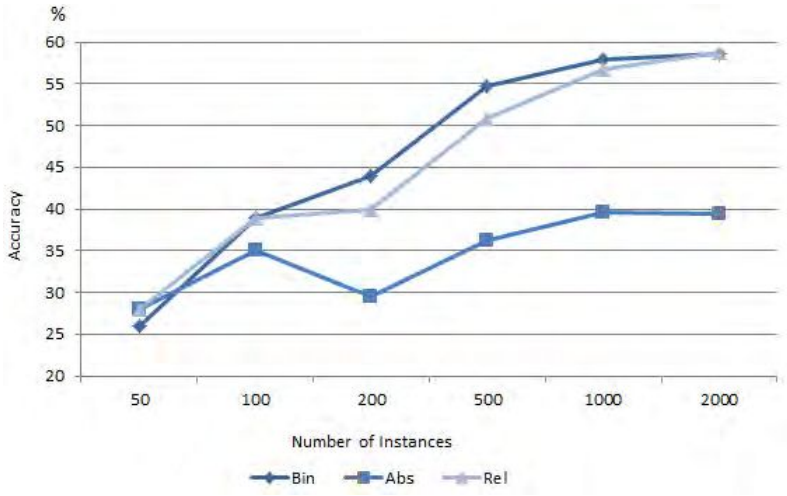
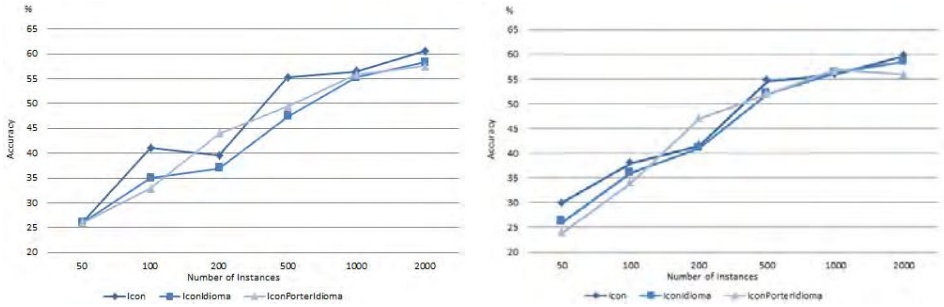


Figure 2: Classification results using tag-based representations (baseline)

Observing the three figures it can be concluded that the pre-processing strategies improved the baseline results in all cases. The simplest strategy, the iconography filter, held the best results. Using the idiom recognition or the Porter algorithm did not improve significantly classification results in spite of being more computationally expensive alternatives. The use of restrictive strategies seems to constrain the amount of available information causing difficulties in the semantic enrichment from *WordNet*. It can also be observed in the figures that as the number of training instances grows, the classifier increases its predictive knowledge on the provided data, allowing an improvement on class estimations and, thus, enhancing classification results.



(a) Relative weighting (b) Binary weighting

Figure 3: Classification results using concept-based representations and different pre-processing strategies

The performance of SVM classifiers for every combination of weighting, disambiguation and document representation strategies in semantic representations was evaluated using the best pre-processing strategy (iconography filter), Figure 4 shows the obtained results. The absolute weighting held the worst results, up to a 20% lower than when using the other weightings approaches. On the other hand, relative weighting outperforms the simple binary weighting, although having a difference lower than a 2% in all cases.

Considering a relative weighting of tags, Figure 5 allows to analyse the performance of the three disambiguation strategies. The best results were obtained with the context-based disambiguation with a 2% improvement over the other alternatives. In contrast, the worst results were obtained when all the tags senses are included in the final representations, proving that an indiscriminate semantic enrichment is not useful for improving resource classification. Also, this strategy maximises the attribute number for each resource which negatively affects the computational complexity of classification. Regarding the addition, replacement and deleting of tags for building representations, the results did not allow to determinate the superiority of any of the strategies as the best performing varied according with the disambiguation strategy used. The maximum accuracy was obtained by adding concepts to the original tag-based representation after context disambiguation.

Finally, Figures 6 (a) and (b) summarize the improvement in resource classification achieved using semantically enriched representations of resources over the original tag-based representations in terms of F-measure and accuracy, respectively. In both figures, it can be seen that the bigger the dataset the more important semantic information becomes for finding the correct category of resources. This can be attributed to the wider tag space, possibly introducing noise and increasing ambiguity during classification, cause for the existence of more resources.

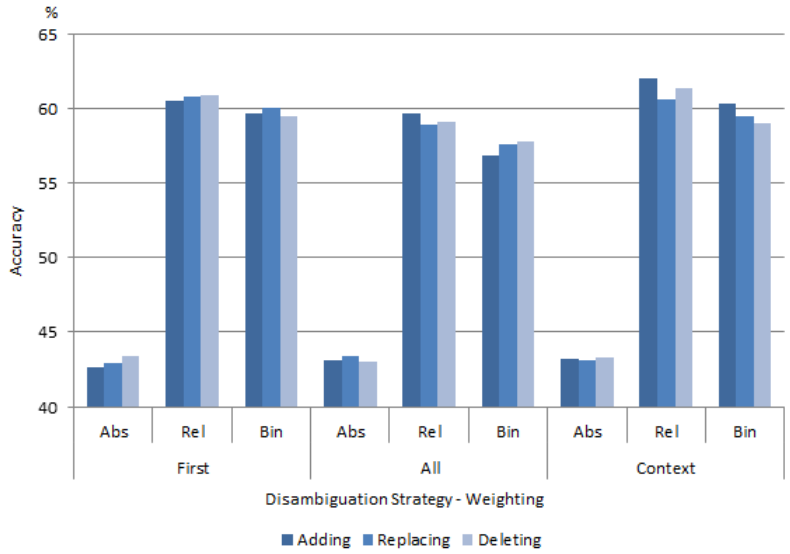


Figure 4: Evaluation of strategies in the construction of concept-based representations

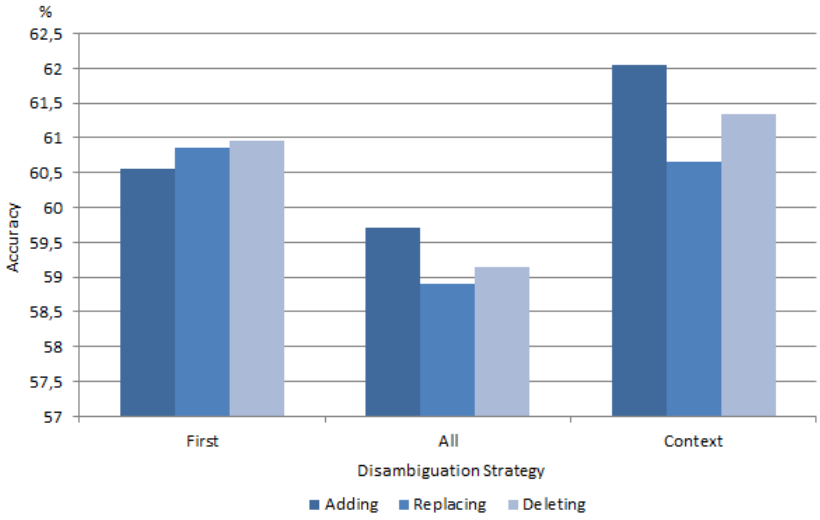
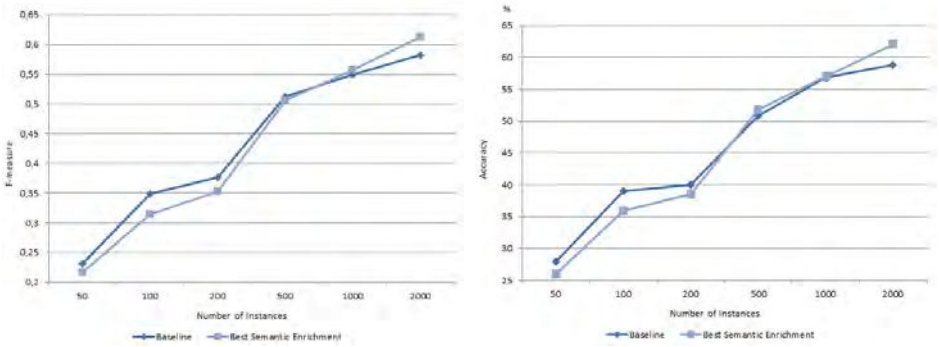


Figure 5: Evaluation of disambiguation and concept incorporation strategies



(a) F-measure (b) Accuracy

Figure 6: Comparison of tag-based and concept-based representations

5 Conclusions

This work analysed and evaluated strategies to incorporate semantics to representation of resources in social tagging systems. This semantic approach is intended to solve ambiguity and other problems related to the free nature of social annotations or tags when used to categorize resources. Several strategies for tag pre-processing, concept disambiguation and incorporation of semantic entities to representations have been discussed.

Experiments carried out using a standard dataset of the area as *Social-ODP-2k9*, have shown that semantic enrichment of tags has a positive effect on resource classification, improving its results. It was also observed that the more instances are used to train the classifiers, the higher the superiority of semantic representations in comparison with simple tag-based representations. However, the results have also shown that the indiscriminate semantic enrichment it is not useful as it negatively affects the results and increases computational complexity. The best classification was performed using a context-based disambiguation of tags and a combination of tags and concepts in the final resource representation.

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 Co., Inc., Boston, MA, USA, 1999.
3. W. Cavnar and J. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.
4. A. Dattolo, D. Eynard, and L. Mazzola. An integrated approach to discover tag semantics. In *Proceedings of the 2011 ACM Symposium on Applied Computing (SAC '11)*, pages 814–820, New York, NY, USA, 2011.

5. S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.
6. A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In *Proceedings of the Semantic Web Workshop of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), August 1, 2003, Toronto Canada*, 2003.
7. L. Huang. *Concept-based text clustering*. PhD thesis, University of Waikato, New Zealand, 2011.
8. C. Körner, R. Kern, H-P. Grahsl, and M. Strohmaier. 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)*, pages 157–166, Toronto, Canada, 2010.
9. M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC '86)*, pages 24–26, New York, NY, USA, 1986.
10. A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. Computer Mediated Communication, 2004.
11. G. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
12. M. Noll and C. Meinel. Authors vs. readers: A comparative study of document metadata and content in the WWW. In *Proceedings of the 2007 ACM Symposium on Document Engineering (DocEng '07)*, pages 177–186, Winnipeg, Canada, 2007.
13. J. Platt. Advances in kernel methods. chapter Fast training of support vector machines using sequential minimal optimization, pages 185–208. 1999.
14. M. Porter. Readings in information retrieval. chapter An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.
15. V. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
16. Z. Yin, R. Li, Q. Mei, and J. Han. Exploring social tagging graph for Web object classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, pages 957–966, Paris, France, 2009.
17. A. Zubiaga, C. Körner, and M. Strohmaier. Tags vs shelves: From social tagging to social classification. In *Proceedings of the 22st ACM Conference on Hypertext and Hypermedia (HT'11)*, Eindhoven, Netherlands, 2011.
18. A. Zubiaga, R. Martínez, and V. Fresno. Getting the most out of social annotations for web page classification. In *Proceedings of the 9th ACM Symposium on Document Engineering (DocEng '09)*, pages 74–83, New York, NY, USA, 2009.
19. A. Zubiaga, R. Martínez, and V. Fresno. Analyzing tag distributions in folksonomies for resource classification. In *Proceedings of the 5th International Conference on Knowledge Science, Engineering and Management (KSEM'2011)*, 2011.