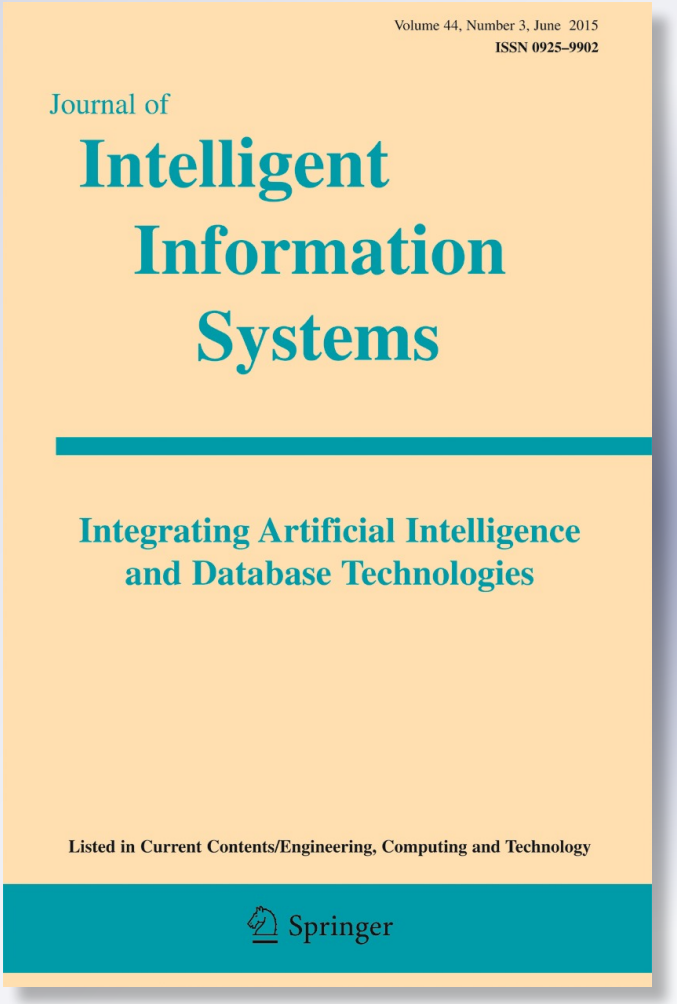
Semantic grounding of social annotations for enhancing resource classification in folksonomies

Antonela Tommasel & Daniela Godoy

Journal of Intelligent Information Systems
Integrating Artificial Intelligence and Database Technologies

ISSN 0925-9902
Volume 44
Number 3

J Intell Inf Syst (2015) 44:415-446
DOI 10.1007/s10844-014-0339-y



Springer

J Intell Inf Syst (2015) 44:415–446
DOI 10.1007/s10844-014-0339-y

Semantic grounding of social annotations for enhancing resource classification in folksonomies

Antonela Tommasel · Daniela Godoy

Received: 28 December 2013 / Revised: 3 October 2014 / Accepted: 7 October 2014 /
Published online: 4 November 2014

Abstract User-generated annotations in tagging or bookmarking sites such as *Flickr* or *Delicious* can provide a promising and interesting source of information for aiding tasks such as Web resource classification. However, the use of tags brings up some challenges. Since there are no constraints on the terms that can be used for tagging, noise and ambiguity are introduced when users annotate resources. Moreover, traditional bag-of-words representations ignore connections between terms and, thus, are affected by synonymity and hyponymia. Althougth tag-based representations are a valuable source for classifying resources, the problems associated with the unsupervised nature of tags may hinder classification results. This paper presents an approach for semantically analysing social annotations in order to attain enriched concept-based representations of Web resources. Representations are enriched with concepts extracted from *WordNet* and *Wikipedia* to overcome problems caused by natural language as well as enhancing the quality of information available for performing an effective classification of resources. Several strategies for tag pre-processing, concept disambiguation and incorporation of semantic entities to representations are discussed and evaluated in this paper. Experimental results showed that the strategies proposed to associate tags with conceptual entities allow improving resource classification results, outperforming traditional approaches based on bag-of-words representations.

Keywords Social tagging systems · Folksonomies · Semantic-based representations · Resource classification

A. Tommasel (✉) · D. Godoy
ISISTAN Research Institute, CONICET-UNCPBA, Paraje Arroyo Seco, Campus Universitario, Tandil, Buenos Aires, Argentina
e-mail: antonela.tommasel@isistan.unicen.edu.ar

1 Introduction

Social annotations allow people to freely describe resources by adding keywords, categories or metadata. Bookmarking sites such as *Delicious*[1] or *Flickr*[2] encourage users not only to tag resources such as Web pages, pictures or videos, but also to collaboratively enrich resources by sharing their annotations and categories. As no special knowledge or abilities are needed to use the unstructured terms or labels that represent tags, social tagging popularity has quickly grown. In this context, the evolution of new technologies and social annotation software led to the appearance of a new social phenomenon, folksonomies.

Folksonomies (Mathes 2004) are the result of free and non-hierarchical annotations of resources in a social environment that contrasts with taxonomies, which are traditionally associated with a systematic and hierarchical ordering aiming at categorising documents or Web pages (e.g. Web directories). While taxonomies are defined by small groups of experts and thus, have a limited scope and can become outdated, folksonomies motivate all Web users to assign tags. As described in Hotho et al. (2006), Solskinnsbakk et al. (2012) a folksonomy describes users, resources, tags and user-based assignment of tags to resources. Their most important feature is the reflection of user vocabulary, which enables serendipity and social connections, and aids the search and navigation of resources.

Recent works have started to explore and analyse the impact of social annotations and collective knowledge found in folksonomies in resource classification tasks. The capability of tags to replace the content of resources for classification (Aliakbary et al. 2009; Zubiaga et al. 2009, 2011b) as well as the analysis of tag distribution and the different motivations of users to annotate resources (Korner et al. 2010; Noll and Meinel 2007, 2008) have been the focus of different studies. Particularly, the potential of tags for classification has been explored in recent works. For example, Aliakbary et al. (2009) analyse the usefulness of tags in replacement of the content for predicting resource categories, Zubiaga et al. (2009) try to show how tags can be used as complementary data for Web site classification, Yin et al. (2009) propose to use tags for describing non-textual objects for classification, and Zubiaga et al. (2011a) analyse classification performance based on tags provided by users with different motivations.

Although tags are a valuable source of information, classification based only on syntactic characteristics of tags has some drawbacks (Lops et al. 2013). Since there are no constraints on the terms that can be used for tagging, users can freely choose the scope, sense or generality of a resource characterisation. In consequence, issues like synonymity (two different terms can convey the same meaning, for example 'animal' and 'creature'), homonymy or polysemy (a term can have several senses, for example 'house' can refer to an aristocratic family line, or to a building in which something is sheltered or located, among others), and morphological variations appear (a concept can have more than one spelling, for example 'organisation' and 'organization' refer to the same topic) negatively affecting the results of classification.

This work addresses 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. Concepts are more abstract than tags and, thus, they can encompass tags meaning and expose other relations beyond term matching. Enriching terms with concepts has two advantages: it

[1] http://delicious.com

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

allows to identify related topics, and solves semantic issues such as synonymity. For example, the tags 'animal' and 'creature' do not have any explicit relation. However, when they are associated to more abstract concepts such as 'fauna' and 'beast', a synonym relation is discovered. Concepts are extracted from two lexical databases: *WordNet* and *Wikipedia*. Both have different origin and characteristics. While *WordNet* is an English lexical corpus in which nouns, adjectives and adverbs are grouped into set of synonyms each expressing a distinct concept, *Wikipedia* is the biggest free encyclopaedia available in all languages. Several strategies for relating tags to concepts from these two sources are evaluated and compared with respect to simple tag-based representations.

The article is organised as follows. Section 2 summarises related research regarding classification based on social annotations. Section 3 presents the proposed approach for associating semantics to tags and, thus, enriching resource representations. Section 4 reports experimental evaluation carried out using different strategies for pre-processing tags and building semantically enriched representations. Finally, conclusions are stated and future lines of research are analysed in Section 5.

2 Related works

The problem of resource classification using social annotations has been addressed in numerous works (Aliakbary et al. 2009; Zubiaga et al. 2009; Yin et al. 2009). In Aliakbary et al. (2009) the authors tried to analyse the usefulness of tags in replacement of content for predicting the category of resources in the Open Directory Project (ODP) hierarchy. The proposed approach builds a representation both of resources and the ODP categories based on their tags, and then, classifies resources based on these representations. A three-step preprocessing of tags is applied to deal with ambiguity problems. First, all tags are converted to lower-case, non-informative tags manually selected by a group of experts are removed and, finally, a stemming algorithm is applied. To assign a category to each resource, the tag vector of the resource is compared to the vectors corresponding to each category, significantly increasing the computational complexity and thus the processing times. The authors concluded that social annotations are better representatives of a resource than its content. They also suggested that their approach could be used in a semi-automatic classification system where an expert chooses the category of resources based on the categories proposed by the algorithm.

Similarly to the previous work, in Zubiaga et al. (2009) the authors aimed to show how tags can be used as complementary data for classification of Web sites. For each selected URL, its tags, reviews and notes were retrieved. Several experiments were carried out to determine the benefit of the approach using a Support Vector Machines (SVM) classifier and each of the defined resource representations: the 10 most popular tags, binary weighted tags, ranking of tags and a combination of notes and reviews using a TF-IDF weighting. They also evaluated how the classification results evolved as the training data for the classifier increased. Results showed that as the training data increase, the superiority of the tag-based classification approach improves. They concluded that weighted tags are more useful than non-weighted tags for classification tasks excepting for small datasets.

Yin et al. (2009) proposed the use of tags for the semantic enrichment of non-textual Web objects like products, images or videos for further classification. In addition, the authors developed different strategies for weighting tags and stated the need of reducing the space of tags used aiming at lowering the computational cost associated to classification. They expressed the extra difficulty of classifying non-textual objects in relation to text document

classification due to three reasons. First, there is a lack of features since, unlike documents, non-textual resources cannot be easily represented in a meaningful text space. Second, there is a lack of interconnections as Web objects tend to exist in isolated settings, relations between them are both limited and untruthful. Finally, there is a lack of labels and training examples and, thus, creating large training sets of certain types of web objects is laborious and even infeasible, which affects classification performance. The authors suggested that social annotations could solve the mentioned problems. As tags use a free vocabulary, they are useful in all of the objects' domains, overcoming the last problem. In consequence, they proposed to use tags to semantically enrich non-textual objects as products, images and videos and, enabling the classification of objects to facilitate their searching and browsing, and to reveal connection patterns between them.

Other studies (Körner et al. 2010; Zubiaga et al. 2011a; Noll and Meinel 2007) focused on understanding the underlying motivation of users behind tagging to infer which users lead to better predictions. In Körner et al. (2010), Zubiaga et al. (2011a) the identification of two kinds of users was proposed, categorisers and describers, which are assumed to have different motivations to tag resources. Whereas describers assign tags to describe resources, categorisers prefer to assign categories to resources, which results in tags with different characteristics depending on who assigned them. Categorisers assign tags according to shared high-level characteristics aiming at constructing and maintain a navigational aid for later browsing. On the other hand, describers use tags to precisely and accurately describe resources for later retrieval. This distinction is important because tags assigned by describers are useful for information retrieval (because these tags focus on the content of resources), as opposed to tags assigned by categorisers, which are more useful for capturing different interpretations of a resource (because they focus on user-specific views of a resource). Zubiaga et al. (2011a) also analysed classification performance using tags provided by categorisers or describers. They concluded that tags provided by describers are more useful than tags provided by categorisers due to the fact that social tagging systems suggest tags to users according to their personomy, which motivates users to use a reduced vocabulary. In addition, as the authors considered that the behaviour of users is a key aspect in classification tasks, they suggested the identification of other kind of users aiming at improving classification results.

Noll and Meinel (2007) analysed people's motivation for providing tags from other point of view. The authors compared user metadata available in social tagging systems with metadata supplied by the authors and publishers of resources aiming at analysing how different they were. The experiments suggested that users tend to focus their tagging activities on popular pages and thus, the majority of tags are concentrated on a small proportion of resources. Regarding tag composition, they concluded that generally users tend to use broad terms included in the content of the resource rather than in the title or descriptions. In addition, the authors recommended applying pre-processing strategies to tags to identify synonyms and heteronyms and to separate composite tags.

In other study (Noll and Meinel 2008), the same authors analysed the characteristics of social annotations provided by users regarding their usefulness for resource classification. Specifically, they studied whether users tend to use tags to classify documents into broad or specific categories, tag popularity, and the matching between freely-chosen user tags and categories in the ODP assigned by expert editors. Their results showed that users tend to tag top-level pages in the Web hierarchy, which implies that their classification could be based on tags, whereas deeper pages classification might require using a content analysis or information derived from tags assigned to parent pages in the hierarchy. Additionally, they

found that tag popularity could help to identify tags that could provide the most relevant classification information and, finally, that users preferred broad terms rather than specific terms when tagging resources.

The described approaches focused on analysing user behaviour when tagging and its effect on classification, rather than the nature of tags themselves. Thus, tags are only syntactically pre-processed and their semantic is ignored.

In general, two approaches can be used to associate semantic to terms. The first approach uses clustering techniques to distinguish between related group of terms and, then, expose their meaning (Deerwester et al. 1990; Dattolo et al. 2011; Schütze and Silverstein 1997). The second approach associates concepts to terms and establishes relations between them using ontologies (Hotho et al. 2003; Lan 2011; Dagher and Fung 2013).

Term clustering uses statistical methods to compute similarity. In Deerwester et al. (1990), the authors presented a latent semantic analysis that uses matrices to build a semantic space where related terms and documents are linked aiming at predicting which terms are related to a resource, even though, no explicit association can be observed. If there is no correlation between the appearance of two terms, it would not be possible to use the terms in the matrix to compute their true association. The appearance of certain patterns of words can help to determine the likelihood of other words. As a result, terms can be associated to resources in which they do not actually appear. The mathematical technique used allows to explicitly represent both terms and documents and thus, to effectively capture the term-document relations. However, the approach presents drawbacks. As new terms cannot be efficiently added, the approach is restricted to a static set of relations. Unlike the approach presented in this work, the obtained relations do not generate concepts and, thus it is difficult to generalise their 'semantic space'. Since each term is represented only once, its weight comprises all the senses weights, negatively affecting the disambiguation of those cases in which the correct sense is not the most popular one. This can lead to distortions and low accuracy. Finally, as strategies for weighting, stemming or removing low-frequency terms are not defined, it is difficult to use the approach in a real environment. Additionally, Schütze and Silverstein (1997) confirmed the benefits of applying latent semantic analysis techniques to clustering tasks as they help to reduce the size of resource representations by removing their noise. While this technique does not affect clustering quality, it helps to reduce the processing time.

Lexical databases are used to extract concepts to include semantic information in tasks like resource classification. *Katoa* (Lan 2011) is a tool that adds semantic information to texts using *Wikipedia*[3] and *WordNet*[4] as semantic sources. This tool is based on the deficiencies of the bag-of-words model, which is considered to be ambiguous as it ignores the fact that different words could have the same meaning (i.e. synonym) while the same word might have different meanings in different contexts (i.e. polysemy) and considers that words are independent from each other. Huang et al. (2009) explored methods to improve the model deficiencies by using concepts instead of words as the resources content description. In addition, the work investigated how unambiguous concepts can be efficiently and effectively used to help text clustering.

Katoa aims to recognise any object that can occur in natural language texts. Objects could be material and concrete such as rivers, immaterial and abstract such as the error rate of an

[3] http://www.wikipedia.org/

[4] http://WordNet.princeton.edu/

algorithm, or imaginary such as a fictional character. The purpose of recognising objects is to group them into meaningful units, i.e. their corresponding concepts. *Katoa* is able to handle all kinds of texts written in English. Although the methods and lexical databases used are independent of any particular domain, topic-specific domains can affect the match between texts and concepts in the databases. Concepts are extracted from *Wikipedia* and *WordNet* and assigned to resources in order to build the enriched representations to be used for clustering resources. The process of indexing terms in *WordNet* is separated into two tasks, first the identification of candidate concepts, and then its disambiguation. In theory, two disambiguation strategies are defined. The first one, selects the most popular sense in English as defined by *WordNet*, and the second one is based on a context analysis. In the case of *Wikipedia*, the outgoing links (anchor text) sorted in order of relevance are used for text comparisons to find the correct sense of tags as performing comparisons between short links is more efficient than performing full-text comparisons. The likelihood of each possible sense is used in the text disambiguation process. The binary weighting obtained the best results when *WordNet* was used. However, *Wikipedia* outperformed *WordNet*, and using both lexical databases did not improve the clustering performance due to the introduced noise. Other weighting strategies only improved results when hierarchical clustering and *Wikipedia* were used.

The strategies used by *Katoa* have some shortcomings. First, the tool fails to produce results for most part of the input since it is not able to associate concepts to input terms, losing information for clustering. Oppositely, in this work strategies to benefit from the lexical resources aiming at widening their coverage are defined with the goal of improving classification results. For example, *Katoa* only implements disambiguation based on the most common sense, whereas multiple strategies are adopted in this work to define the representation of resources and to address ambiguity.

Hotho et al. (2003) defined strategies to add semantics to the full content of resources in order to improve text document clustering. The authors claimed that enriching term vectors with concepts has benefits. Mainly, it solves the synonymity problem and also introduces more general concepts which can help to identify related topics. In addition, various strategies are implemented for building the resource representations based on adding concepts from *WordNet*, replacing terms by concepts or deleting terms without concepts. To solve the ambiguity of terms, three strategies were defined: adding all the possible senses, adding only the most popular sense, and using a context based disambiguation (Agirre and Rigau 1996) which measured the proximity of two concepts as the length of the shortest path between concepts found in *WordNet*. In addition, strategies to add hypernyms to modify the levels of generality of concepts and the frequency of each term were proposed. Their results showed that combining terms with *WordNet* concepts disambiguated by the context-based strategy outperformed the performance of clustering based only in content. With the exception of the context disambiguation strategy, all the strategies described are applied in this work to tags.

Finally, there are also hybrid approaches for associating semantic to tags such as Dattolo et al. (2011), which is based on creating bi-graphs where nodes represent tags, and links represent the co-occurrence of such tags on different resources. In addition, the authors tried to detect synonyms and homonyms by using different heuristics based on distance metrics and *WordNet* synonym identification. They defined a folksonomy as a multi-graph with coloured edges and applied the disambiguation heuristics to reduce graphs aiming at decreasing the task complexity. The used heuristics included the Levenshtein distance, synonymy search in *WordNet*, translations and stemming. Unfortunately, the heuristics, with the exception of the one using *WordNet*, suffered the same problems described for the clustering approach.

For the homonym detection, clusters of related tags to be used in the disambiguation tasks are created. Although synonym detection strategies were not conclusive, the authors confirmed their hypothesis that synonyms belong to different contexts that can be identified by using several strategies.

There are also works focused in defining a context disambiguation strategy for terms using *Wikipedia* (Fogarolli 2009; Milne and Witten 2008b). Fogarolli (2009) presented a similarity measure based on counting the common elements between the bi-directional links of each of the possible term senses and the context of each term, i.e. the other terms in the vector that represents a resource. This approach assumes that only one term needs to be disambiguated. The bi-directional links are used due to the supposition that if there is a symmetric relationship between the articles it is because they are strongly related, and thus, it is possible to discard wrong senses with a high accuracy. In Milne and Witten (2008b) it was proposed an approach for computing the relatedness of two articles based on their outgoing links. The formula expressed for comparing only two articles, compares the number of outgoing links of each article, its minimum and maximum, the links intersection and the total number of articles in *Wikipedia*. The relatedness of a candidate sense was defined as the weighted average of the relatedness with each one of the context articles that are assumed to be non-ambiguous.

Unlike the described approaches, which only perform a syntactic pre-processing of tags, this work proposes to combine syntactic and semantic analysis of social annotations. The semantic analysis enriches terms with concepts extracted from lexical databases, such as *Wikipedia* and *WordNet* in order to overcome problems caused by natural language as well as enhancing the quality of information available for performing an effective classification of resources.

3 Associating semantic to social annotations

This work proposes a method aiming at improving the classification of resources belonging to a folksonomy by the semantic enrichment of tags users assigned to them. From tag-based representations, which exploit social annotations to describe resources, enhanced concept-based representations are gleaned by relating tags to concepts from *WordNet* and *Wikipedia*.

The tag-based representation of a resource is formally defined as $R = \{(t_r, w_r) | r = 1, ..., n_r\}$ where R is the resource being analysed, t_r represents a tag selected by users to annotate the resource, and w_r represents the weight associated to t_r according to its importance in the resource representation. Finally, the function relating lexical entries in databases such as *WordNet* or *Wikipedia* with their corresponding concepts is denoted $Ref_C(t)$.

For accomplishing this goal, the process of building semantic representations of resources was decomposed into several steps, which are shown in Fig. 1:

1. *Category and tag pre-processing.* (Section 3.1). Lexical structure of categories and tags is analysed. Morphological variations are reduced by using a stemming algorithm and the language of tags is identified.
2. *Tag Weighting.* (Section 3.2). Different weighting strategies are derived from the number of times a tag is used.
3. *Matching Tags with Concepts.* (Section 3.3). Communication with lexical databases is established and concepts are retrieved. Then, semantic entities are chosen according to the selected disambiguation strategy in order to solve polysemy.
4. *Final Resource Representation.* (Section 3.4). The combination of tags and concepts to be used for obtaining the final resource representation is chosen.

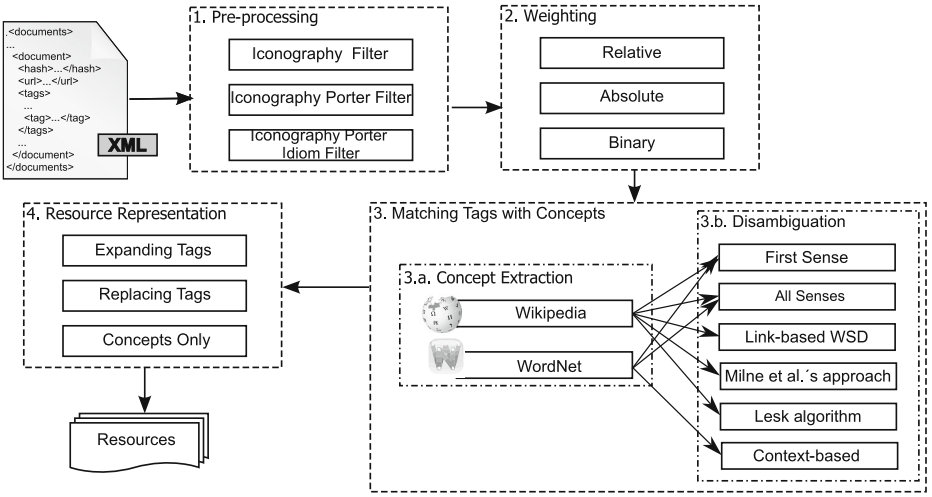


Fig. 1 Steps for building a semantic resource representation

3.1 Tag pre-processing

This step tries to reduce the syntactic variations of tags in order to improve the probability of finding their associated concepts. Since the used lexical databases contain only English terms, several pre-processing strategies are developed to deal with non-English tags. The first pre-processing strategy ignores tags composed by non-English characters, for example Kanjis or Russian characters. Also, if the name of the category has any of those characters, the example is discarded. The second alternative carries out a language detection based on TextCat[5], a Perl implementation of the algorithm presented in Cavnar and Trenkle (1994). The last alternative implements the Porter Stemmer algorithm (Porter 1997) that removes automatically word suffixes, and then performs language detection.

The language detection algorithm used in this work is based on the simultaneous definition and analysis of the continuous sequences of characters belonging to a longer sequence (N-grams) and Zipf's law (Zipf 1935), which implies that each language has a dominant set of words in terms of the frequency of use. The detection process consists in comparing the N-grams appearing in the new document to classify with the N-grams profile of known languages. The language of a new document is defined as the nearest existing language profile.

The used language detection algorithm does not have a perfect accuracy, possible failings include a list with several languages alternatives or no language at all. Due to the fact that the identification is better as the length of the text increases, to avoid mistakes during the language detection, all the tags of each resource are given to the language detection algorithm. If English is not in the top 3 languages of the output, the resource is classified as a non-English one and treated accordingly.

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

3.2 Tag weighting

Attribute weighting consists in assigning a numeric value to a term according to its relevance for describing a resource. This is done through a weighting function that attempts to estimate the term relevance. A successful weighting function scores relevant attributes with high values while the irrelevant ones with close to zero values.

Although choosing a weighting function is important, it has not usually been considered as important as attribute selection in information retrieval systems (Buckley 1993; Salton and Buckley 1988). The importance of attribute weighting in improving classification performance has been analysed by numerous works (Jankowski and Usowicz 2011; Kohavi et al. 1997). Several authors (Leopold and Kindermann 2002; Lan et al. 2005) have also stated that the selection of an adequate weighting function is even more important than the parameterisation of the kernel in SVM algorithms. The weighting function can also help to remove irrelevant attributes by defining a low weight threshold.

There are different approaches to define the weighting function (Lan et al. 2005; Salton and Buckley 1988) such as the statistical occurrence of attributes or the historical usefulness of the attribute for classification. In many cases, the initial weights must be based on statistical information since historical information is not available.

In this work, three alternatives are considered for weighting tags. The first alternative 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 alternative, tags are weighted according to the number of times that users have assigned the tag to the resource, i.e. how many users annotate the resource with a given tag. The last alternative uses a relative weighting for tags, i.e. the number of times that users have assigned the tag divided by the total number of times that each tag was assigned.

3.3 Matching tags with concepts

Finding external semantic entities or concepts 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. In this situation, if concepts associated with a wrong sense are chosen, noise can be added to the resource representation. As a consequence, strategies for selecting the correct sense are needed. Different disambiguation strategies for selecting an appropriate sense for a concept are considered in this work to solve this problem. Some of them can be used with both lexical databases and others are specific to *WordNet* or *Wikipedia*.

3.3.1 Both lexical databases

All Senses This disambiguation strategy does nothing to solve ambiguity, it simply considers all the extracted concepts from the lexical databases and adds them to the final set or resource representation. The concept set associated to each tag is defined as $\{c_t\} = \bigcup \{Ref\ (t_n) : t_n \in T\}$.

First Sense This disambiguation strategy takes advantage of lexical databases' 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 $\{c_t\} = \bigcup \{first\,(Ref\,(t_n)) : t_n \in T\}$.

3.3.2 Wikipedia

Link-based WSD This disambiguation strategy presents an approach to text disambiguation based on the linking structure of *Wikipedia* (Fogarolli 2009). The process starts with the definition of the term vector associated to a resource by choosing its 50 most important terms. In this work, such vector is composed by the resource tags. The author defines the concept of 'strong links' as a bidirectional connection between two articles. Those links between pages mean that they are semantically related and belong to the same context. In consequence, symmetric links represent strong relations between articles. A comparison between the strong links of the ambiguous term and its context is carried out and the sense with the maximum number of elements in common with the context terms is chosen as the correct one.

In this work, the ambiguity analysis is performed for each of the ambiguous tags by comparing the outgoing links with the rest of the tags. Aiming at reducing the computational complexity of this strategy, the bidirectional link condition is not considered.

Table 1 presents an example of application of this disambiguation strategy. First, Table 1a shows the tags associated with a resource, in this case the ambiguous one is *collection*. Its possible senses and their associated links are shown in Table 1b. Finally, Table 1c shows the matching between senses and tags. The sense *computing* is chosen as the correct one as it has four elements in common with the tags.

Milne et al.'s approach In Milne and Witten (2008a, b) an approach to disambiguate two terms based on *Wikipedia*'s link structure rather than the category hierarchy or the content of their associated articles is presented. Theoretically, this approach offers a low computational complexity alternative as the textual content of articles can be ignored. Additionally, it is more accurate than similar approaches found in literature (Strube and Ponzetto 2006; Gabrilovich 2007) since it is closely tied to the manually defined semantic of articles. Originally, the approach considered both incoming and outgoing links. However, due to

Table 1 Example of context-based disambiguation in *Wikipedia*

$T = \{set, map, array, list, java, computer, collection, casting\}$

(a) Resource to disambiguate

$Collection(computing) = \{object\text{-}oriented, class, map, tree, set, array, list\}$

$Collection(museum) = \{curation, curator\}$

(b) Ambiguous tag senses

$T \cap Collection(computing) = \{map, array, list, set\}$ $\quad |T \cap Collection(computing)| = 4$

$T \cap Collection(museum) = \{\}$ $\quad |T \cap Collection(computing)| = 0$

(c) Matching between senses and tags

its computational complexity, the authors decided to only consider the incoming links. Formally, a relatedness measure is defined as:

$$Relatedness(a,b) = \frac{log(max(|A|,|B|)) - log(|A \cap B|)}{log(|W|) - log(min(|A|,|B|))} \quad (1)$$

where a represents an article from an ambiguous term, b represents an article from the unambiguous context, A and B represent the incoming links to articles a and b respectively, and W represents the number of articles in *Wikipedia.*

The relatedness of a candidate sense is defined as the weighted average of its relatedness to each non-ambiguous context article. For example, to decide among the senses a_1, a_2 and a_3 of term A that is used in a context composed by the articles b, c and d; $Relatedness(a_1, b)$, $Relatedness(a_1, c)$ and $Relatedness(a_1, d)$ should be computed. Likewise, relatedness of a_2 and a_3 needs to be computed. Finally, the sense with the lowest weighted average or the highest inverse weighted average relatedness is selected for each of the terms.

Since context terms are not equally useful for sense disambiguation, comparisons are weighted accordingly. For example, the term "the" is unambiguous since it can only be used as a definite article, but it has no value to disambiguate other terms. Therefore, $Relatedness$ results are combined with the link probability of each term. In summary, the final value is computed considering the sum of weights associated to each context term, the number of terms involved, how terms relate to each other and their link commonness.

Milne et al.'s approach assumes that context articles are unambiguous and there is only one ambiguous term, i.e. there are always going to be unambiguous articles to compare to. When using tags, that assumption cannot be guarantee. Furthermore, it is not possible to guarantee that the most related senses would be chosen by computing the $Relatedness$ between pairs of articles.

In this work, the measure is modified in order to simultaneously disambiguate sets of ambiguous tags by considering tuples of articles belonging to each ambiguous tag and their different combinations. Once measurements are carried out, the maximum value and its associated senses are obtained, thus disambiguating all tags. The modified measure can be defined as:

$$Relatedness(\langle t_{11}, ..., t_{nm}\rangle) = \frac{log(max(|L_{t_{11}}|,...,|L_{t_{nm}}|)) - log(|L_{t_{11}} \cap ... \cap L_{t_{nm}}|)}{log(|W|) - log(min(|L_{t_{11}}|,...,|L_{t_{nm}}|))} \quad (2)$$

where $< t_{11}, ..., t_{nm} > \in T_1 \times ... \times T_n$, T_n represents each of the tags to disambiguate, $\{t_{n1}, ..., t_{nm}\}$ is the set of senses belonging to T_n, $L_{t_{nm}}$ represents the set of incoming links to t_{nm}, and W represents the number of articles in *Wikipedia*.

As an example, consider the disambiguation of the following terms: *Flash*, *Animation* and *Superman*. Table 2a shows senses associated to them, and Table 2b shows senses incoming links. All senses and links are extracted from *Wikipedia*. Finally, Table 3 summarises the relatedness between the different 3-tuple sense combinations. Results show that the most related senses are *Flash (comics)*, *Animation* and *Superman,* which are selected for disambiguating the original terms.

On the other hand, Table 4 summarises the relatedness between the pair senses based on the original formula presented in Milne and Witten (2008a), for the same example. The most closely related senses are *Flash (comics)*, *Animation* and *Superman (comic book)*. The

Table 2 Example of Milne et al. context-based disambiguation input

Flash	**Adobe Flash**- Adobe Flash (formerly SmartSketch FutureSplash, FutureSplash Animator and Macromedia Flash)is a multimedia platform used to add animation, video, and interactivity to web pages. **Flash (comics)** - The Flash is a name shared by several fictional comic book superheroes from the DC Comics universe. **Flash memory** - Flash memory is a non-volatile computer storage chip that can be electrically erased and reprogrammed.
Animation	**Animation** - Animation is the rapid display of a sequence of images of 2-D or 3-D artwork or model positions in order to create an illusion of movement. **Computer animation** - Computer animation is the process used for generating animated images by using computer graphics.
Superman	**Superman** - Superman is a fictional character, a comic book superhero appearing in publications by DC Comics, widely considered to be an American cultural icon. **Superman (comic book) -** Superman is an ongoing comic book series featuring the DC Comics hero of the same name.

(a) Tags and their associated senses

Adobe Flash	Adobe, FrameMaker, Adobe Systems, Arcade game, Animated Cartoon, Computer animation, E-learning, Ingrid Bergman, History of video games, Macromedia, Multiple-image Network Graphics
Flash (comics)	Comic book, Batman, DC Comics, DC Universe, Andy Kubert, Justice Society of America, Justice League, Superman, Super Friends, Super Hero
Flash memory	Computer data storage, Commodore 64, Computer memory, Data storage device, Database, Digital camara, Hard disk drive, Firmware, Memory card, Memory stick
Animation	Adobe Flash, Animation, Anime, Cartoon, Comic book, Computer animation, DC Comics, Super Hero, Lex Luthor, Multimedia
Computer animation	Adobe Flash, Animated Cartoon, Animation, Animator, Anime, Computer-aided design, Motion capture, Tron (film), Stop motion, Image
Superman	Comic Book Guy, Comic book, DC Comics, Daily Planet, Lex Luthor, Martian, Marvel Comics, Super Hero, Kryptonite, Flash (Comic)
Superman (comic book)	Adventures of Superman, Adventure Comics, Batman, Superman, Supergirl, Comic book, Doctor Mid-Nite, Animated Cartoon, Kryptonite, Legends (comics)

(b) Senses incoming links

selected senses differ from the ones obtained by the new proposed approach, which choose the most context-accurate sense for *Superman*.

Lesk algorithm This disambiguation strategy performs a context-based disambiguation by means of Lesk algorithm (Lesk 1986). This algorithm uses a low computational complexity technique to find the correct sense for ambiguous tags based on the assumption that the

Table 3 Example of Milne et al. context-based disambiguation

$Relatedness(Adobe\ Flash,\ Animation,\ Superman) = 0$
$Relatedness(Flash\ (comics),\ Animation,\ Superman) = 0,3378$
$Relatedness(Flash\ memory,\ Animation,\ Superman) = 0$

$Relatedness(Adobe\ Flash,\ Computer\ Animation,\ Superman) = 0$
$Relatedness(Flash\ (comics),\ Computer\ Animation,\ Superman) = 0$
$Relatedness(Flash\ memory,\ Computer\ Animation,\ Superman) = 0$

$Relatedness(Adobe\ Flash,\ Animation,\ Superman\ (comic\ book)) = 0$
$Relatedness(Flash\ (comics),\ Animation,\ Superman\ (comic\ book)) = 0,1766$
$Relatedness(Flash\ memory,\ Animation,\ Superman\ (comic\ book)) = 0$

$Relatedness(Adobe\ Flash,\ Computer\ Animation,\ Superman\ (comic\ book)) = 0,1766$
$Relatedness(Flash\ (comics),\ Computer\ Animation,\ Superman\ (comic\ book)) = 0$
$Relatedness(Flash\ memory, Computer\ Animation,\ Superman\ (comic\ book)) = 0$

same topic is shared by terms in a particular context. Formally, this strategy can be defined as Navigli (2009):

$$score_{Lesk}(S_1, S_2) = |gloss(S_1) \cap gloss(S_2)| \tag{3}$$

where w represents the terms to disambiguate, $Senses(w_j)$ represents the set of senses associated with each term, $S_i \in Senses(w_j)$ and $gloss(S_i)$ represents the words contained in the textual definition of S_i.

The pair of sense definitions with most coincidences is chosen as the correct pair of senses. A drawback of this approach is its sensibility to the exact terms that compose a definition as results can radically change by the presence or absence of a single word. Additionally, the original algorithm analyses the overlapping between dictionary definitions, which tend to be short, thus not providing enough vocabulary to accurately differentiate senses.

Whereas the original algorithm disambiguates pairs of terms, an extended version to simultaneously disambiguate all ambiguous tags is proposed in this work. The modification tries to avoid choosing mistaken senses derived of comparing only pairs of them as in Milne et al.'s approach example. Additionally, dictionary definitions are replaced by outgoing links.

3.3.3 WordNet

Context-based Disambiguation This strategy performs a context-based disambiguation by means of Lesk algorithm (Lesk 1986). 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 chosen sense is the one with more words in common with the other tag senses.

Formally, this strategy can be defined as $\{c_t\} = \bigcup\{Lesk\,(Ref\,(t_n)) : t_n \in T\}$ where $Lesk = \{max(\cap(Ref\,(t_1) \times ... \times Ref\,(t_n))) : t_1, ..., t_n \in T\}$. For implementing this strategy it is necessary to detect the part of the speech (Noun, Verb, Adjective, Adverb) of each tag.

Table 4 Example of Milne et al. context-based disambiguation (original definition)

$Relatedness(Adobe\ Flash,\ Animation) = 0,1766$	0,1324
$Relatedness(Adobe\ Flash,\ Computer\ Animation) = 0,1766$	
$Relatedness(Adobe\ Flash,\ Superman) = 0$	
$Relatedness(Adobe\ Flash,\ Superman\ (comic\ book)) = 0,1766$	
$Relatedness(Flash\ (comics),\ Animation) = 0,3378$	0,2533
$Relatedness(Flash\ (comics),\ Computer\ Animation) = 0$	
$Relatedness(Flash\ (comics),\ Superman) = 0,3378$	
$Relatedness(Flash\ (comics),\ Superman\ (comic\ book)) = 0,3378$	
$Relatedness(Flash\ memory,\ Animation) = 0$	0
$Relatedness(Flash\ memory,\ Computer\ Animation) = 0$	
$Relatedness(Flash\ memory,\ Superman) = 0$	
$Relatedness(Flash\ memory,\ Superman\ (comic\ book)) = 0$	
$Relatedness(Animation,\ Adobe\ Flash) = 0,1766$	0,2270
$Relatedness(Animation,\ Flash\ (comics)) = 0,3378$	
$Relatedness(Animation,\ Flash\ memory) = 0$	
$Relatedness(Animation,\ Superman) = 0,4439$	
$Relatedness(Animation,\ Superman\ (comic\ book)) = 0,1766$	
$Relatedness(Computer\ animation,\ Adobe\ Flash) = 0,1766$	0,0706
$Relatedness(Computer\ animation,\ Flash\ (comics)) = 0$	
$Relatedness(Computer\ animation,\ Flash\ memory) = 0$	
$Relatedness(Computer\ animation,\ Superman) = 0$	
$Relatedness(Computer\ animation,\ Superman\ (comic\ book)) = 0.1766$	
$Relatedness(Superman,\ Adobe\ Flash) = 0$	0,1563
$Relatedness(Superman,\ Flash\ (comics)) = 0,3378$	
$Relatedness(Superman,\ Flash\ memory) = 0$	
$Relatedness(Superman,\ Animation) = 0,4439$	
$Relatedness(Superman,\ Computer\ animation) = 0$	
$Relatedness(Superman\ (comic\ book),\ Adobe\ Flash) = 0,1766$	0,1735
$Relatedness(Superman\ (comic\ book),\ Flash\ (comics)) = 0,3378$	
$Relatedness(Superman\ (comic\ book),\ Flash\ memory) = 0$	
$Relatedness(Superman\ (comic\ book),\ Animation) = 0,1766$	
$Relatedness(Superman\ (comic\ book),\ Computer\ animation) = 0,1766$	

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.

3.4 Resource representation strategies

Automatic classification techniques are generally based on words, which is why the most popular model for resource representation is called bag-of-words (Rijsbergen 1979). Each resource is represented by a vector whose dimensions correspond to a weighted term according to its frequency without considering the position or order of the occurrences. However, this kind of representation has some shortcomings (Lan 2011). Firstly, it ignores the fact that different words can have the same meaning. Secondly, it does not consider the different meanings a word can have depending on the context. Thirdly, all connection between words are ignored, they are assumed to be independent from each other, and thus shallow representations of resources are provided. Finally, the model is not robust to add new resources as they are not going to be correctly classified unless they include the same words as the training examples.

Novel resource representation strategies that take into account semantic relationship between tags and concepts extracted from *WordNet* or *Wikipedia* are important to alleviate the problems mentioned before. In Hotho et al. (2003) several alternatives are presented in order to semantically enrich resource representations.

According to the authors, the enrichment of term sets using the ontology proposed in *WordNet* has two benefits: it solves synonymity and introduces more general concepts that can help with the identification of new related topics. Even though the strategies defined by the authors are exclusively design to be used with *WordNet*, in this work they are extended to be used with *Wikipedia*. The proposed resource representation strategies add or replace tags with concepts to incorporate the information from the lexical databases.

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

This strategy allows the existence of repeated terms. Each tag that has an entry on the used lexical database appears at least twice in the new representation, once as part of the former set $\{t_r\}$ 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.

Replacing tags with concepts The second resource representation strategy is similar to the first one. The only difference is that when a tag has an entry in a lexical database it is removed from $\{t_r\}$ and replaced by its concepts. Those tags that do not have an entry on the lexical databases remain in the result set without changes. The resulting set is defined as $\{t_r\} \cup \{c_t\} - t_1 \in T : \exists Ref(t_1)$ where $Ref(t_1)$ represents the set of concepts extracted from the lexical database for the tag t_1 and T is the tag set for the resource.

Concept set only The last resource representation strategy replaces the complete original tag set with the representations from the lexical databases. Those tags with no entry on the lexical databases do not appear in the final representation. The resource set is then defined as $\{c_t\}$.

4 Experimental evaluation

4.1 Dataset description

Social-ODP-2k9[6] (Zubiaga et al. 2009) dataset was used for experiments. This dataset was created between December 2008 and January 2009 with data obtained from several sources, including *Delicious*, *StumbleUpon*[7], the Open Directory Project (ODP) and the Web.

Tags were obtained from *Delicious*, a service that allows the storage of favourite Web sites, their categorisation using tags, and sharing the bookmarks with other users. On the other hand, categories were obtained from the ODP, also known as DMoz, which is the biggest directory edited by human beings, built and maintained by a global community of volunteers.

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 collection comprises 12,116 tags, out of which 53.8 % are unique. Each top-level class or category has an average of 1,339 tags assigned to their resources.

For selecting the URLs included in the collection, the authors took a list from *Delicious*, restricting the URLs to those sites that have been tagged by at least 100 users to guarantee each Web site popularity. The URL or resource category was taken from the ODP. In some cases, the URL belonged to more than one category, situation solved by selecting one of them randomly. Categories are not uniformly distributed in the collection. Figure 2a shows the distribution of the top level categories. There are 17 top-level classes and the one with the highest amount of examples accounts for the 26 % of the resources. Figure 2b shows the distribution of the 20 most popular categories out of the 4,621 low-level categories, which account for only 17 % of the resources.

4.2 Lexical databases

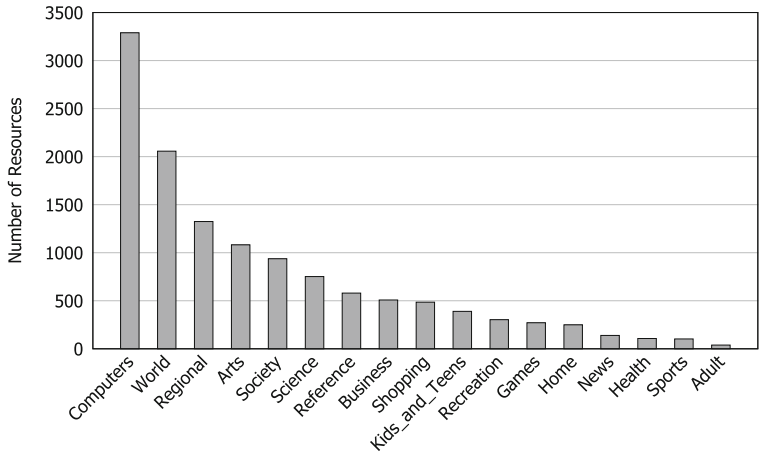
Two lexical databases were used in this work to semantically enrich resources, *WordNet* and *Wikipedia*.

WordNet (Fellbaum 2005) is a lexical database of English. Nouns, adjectives and adverbs are grouped into sets of cognitive synonyms also known as synsets. Each synset represents a distinct concept, which are interlinked by semantic and lexical relations. The resulting network of related words and concepts can be searched and it is useful for natural language processing tasks. *WordNet* resembles a thesaurus, as it groups words together based on their meanings. However, there are important differences. Firstly, *WordNet* interlinks not just word forms based on their syntactical composition but specific senses of words. As a result, words that are close to one another in the network are semantically disambiguated. Secondly, *WordNet* labels the semantic relations among words, whereas thesauri do not follow any explicit pattern other than meaning similarity to group the words. The *Java API for WordNet Searching (JAWS)*[8] was used to extract concepts associated to a tag from *WordNet*.

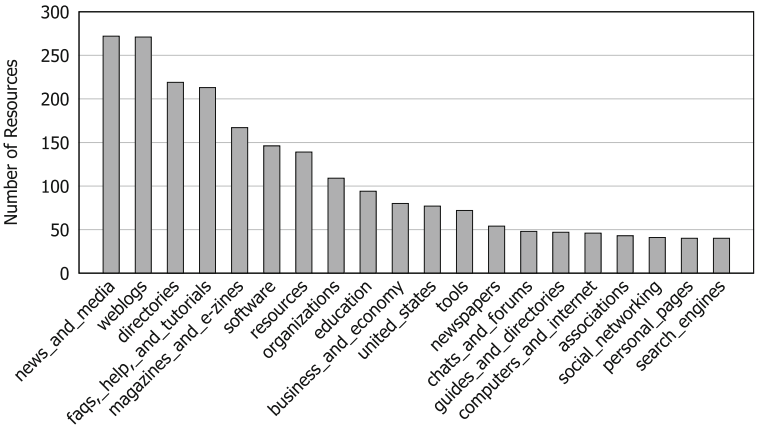
[6] http://nlp.uned.es/social-tagging/socialodp2k9/

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

[8] http://lyle.smu.edu/~tspell/jaws/index.html/



(a) Top-level categories



(b) Low-level categories

Fig. 2 Category distribution in the *Social-ODP-2k9* collection

Wikipedia was created in 2001 with the goal of building a free encyclopaedia in all languages. At present, it is the largest and most widely used encyclopaedia (Medelyan et al. 2009). Traditional paper encyclopaedias consist of articles alphabetically arranged with links to other articles, external references to other sources of information, and a general index of topics. Those structures have been adapted by *Wikipedia* for the online environment (Medelyan et al. 2009). *Wikipedia* is an interesting example of large-scale collaboration, although it may be risky to use it as a source of information as it has several problems associated with misinformation, lack of accuracy within articles and bias of coverage across them due to the lack of formal procedures of expert revision and open editing policy (Medelyan et al. 2009). However, online encyclopaedias have proved to be a promising source of semantic information (Maree and Belkhatir 2013).

To extract concepts from *Wikipedia*, *Wikipedia Miner* was used[9] (Milne and Witten 2009). This tool's purpose is to facilitate the exploration and content extraction from *Wikipedia* by providing an easy access to its structure and content. It also allows a semantic comparison between terms and concepts, and topic detection of given documents. The toolkit indexes pages so that they can be efficiently searched. The most common scenario for searching is to return an article or set of articles that could refer to the queried term. A common approach for page searching is to analyse page titles and resolve differences between different type of pages (articles, redirecting and disambiguation pages) using the links existing between them. According to this scheme, matching articles are directly used, redirects are resolved to their target articles and disambiguation pages are searched for the different senses they list. By default, anchor texts are processed without modifications as they already encode many variations such as: syntax, letter case, pluralism and punctuation.

4.3 Methodology

A Sequential Minimal Optimisation (SMO) (Platt 1999) classifier, which is an optimisation of Support Vector Machines (SVMs) (Vapnik 1995), was used to classify resources. SVMs are characterised for being a model that represents the sample points in the space, separating the classes by 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 according to their proximity to the training points in the model. SMO represents an alternative to SVM method as it allows an optimisation 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[10] implementation of the algorithm was used in these experiments.

For evaluating the classifiers, the standard accuracy, precision and recall measures, summarised by F-measure, were employed (Baeza-Yates and Ribeiro-Neto 1999). In all cases classifiers were evaluated using a classical 10-fold cross-validation strategy.

In the reported experiments, top-10 tags with the number of times that users have assigned them and the category from the ODP were retrieved for each resource. To improve training information quality for the classifier, only categories with at least two resources on the category level being analysed (top or low-level ODP categories) were considered.

Before semantic enrichment, the three pre-processing strategies described in Section 3.1 were evaluated: the iconography filter (Icon), the iconography filter with language detection (IconIdioma) and the same filter using Porter stemming algorithm (IconPorterIdioma). After the pre-processing tasks and the semantic enrichment, different datasets were constructed for each possible combination of strategies.

WordNet was the first lexical database analysed. In this case, the concepts associated with tags were *WordNet* synsets. Experiments were carried out combining all the pre-processing, disambiguation and weighting strategies previously described (Section 3). In summary, classification trials were performed for all the combinations of the variables presented in Table 5.

To build the datasets based on *Wikipedia*, the results obtained with *WordNet* were considered, thus applying only the pre-processing strategies showing the best results. Once the

[9] http://wikipedia-miner.cms.waikato.ac.nz/

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

Table 5 Strategy combinations for *WordNet* experiments

Pre-processing Strategies	Resource Representation Strategies	Disambiguation Strategies	Weighting Strategies	Category Level
• Icon	• Adding	• All	• Binary	• Top-Level
• IconPorter	• Replacing	• First	• Absolute	• Low-Level
• IconPorterIdioma	• Deleting	• Context	• Relative	

correct sense was chosen for each of the tags belonging to a resource, the dataset construction considered the alternative labels (Labels), i.e. all terms that are used to reference the selected article, the outgoing links (Links_Out) and the incoming links (Links_In).

In summary, to begin the analysis, datasets for all the combinations of the variables in Table 6 were built, totalling 1,620 datasets. While for *WordNet* the lowest level of the category was considered for all the combinations of strategies, in the case of *Wikipedia* the lowest level of the category was only considered for the combination of strategies that held the best results for the top-level categories.

The baseline for comparing and evaluating the enhancements introduced by the semantic enrichment approach presented as well as determining the best performing strategies is the results of resource classification based only in the original tags, without any semantic information.

4.4 Experimental results

The developed semantic approach heavily relies on finding relations between the lexical databases and the social annotations. Table 7 summarises the proportion of tags that matched with semantic entities in both lexical databases for each of the pre-processing strategies defined. Filtering tags increments the number of semantically enriched tags. Using the Porter algorithm before performing the language detection decreases the proportion of semantically enriched tags. This can be explained by the fact that it is not possible to find the stems in *WordNet*. *Wikipedia* allows the semantic enrichment of a bigger number of tags than *WordNet* with differences up to a 40 % depending on the pre-processing strategy used. In both cases, the number of semantically enriched tags is maximised by combining the iconography filtering and the language detection covering a 57 % of tags in *WordNet* and a 84 % in *Wikipedia*.

Table 6 Strategy combinations for *Wikipedia* experiments

Pre-processing Strategies	Resource Representation Strategies	Disambiguation Strategies	Weighting Strategies	Origin of Concepts Strategies	Category Level
• Icon	• Adding	• All	• Binary	• Labels	• Top-Level
• IconIdioma	• Replacing	• First	• Absolute	• Links_In	
	• Deleting	• Link Based WSD	• Relative	• Links_Out	
		• Milne et al.'s approach			
		• Lesk Links			

Table 7 Tags coverage in both lexical databases

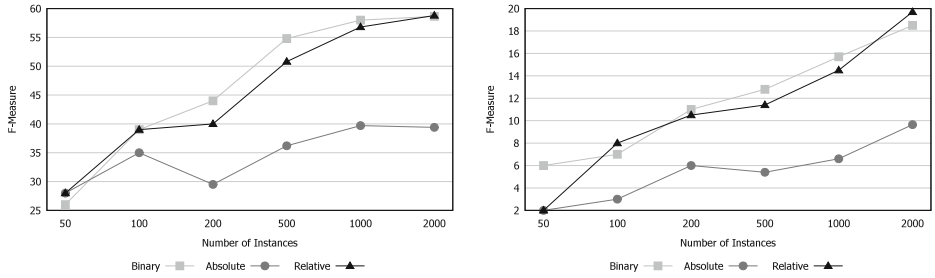
Strategy	# of tags	# matching tags	% matching tags
(a) WordNet			
Without pre-processing	12,116	5,533	45.6668
Icon	9,849	5,151	52.2997
IconIdioma	6,494	3,728	57.4068
IconPorterIdioma	2,208	948	42.9347
(b) Wikipedia			
Without pre-processing	12,116	8,933	73.7289
Icon	9,849	8,073	81.9677
IconIdioma	6,494	5,475	84.3085
IconPorterIdioma	2,208	1,402	63.4963

The baseline results of resource classification are presented in Fig. 3. These results allowed to evaluate the quality of the different pre-processing strategies when tag-based representations were considered. In this figure it can be observed that tag weighting using absolute frequency is the worst performing function, with differences up to a 4 % with respect to the relative weighting. Roughly speaking, the results for the low-level classes are up to a 40 % lower than the results for the top-level classes. These differences can be explained in part by the difference in the number of top-level and low-level classes as well as the number of examples by class. There are only 17 top-level classes and 4,621 low-level classes, thus, the training material available for each of the corresponding low-level classes is reduced.

4.4.1 WordNet results

Figure 4 shows a comparison between the different pre-processing strategies for concept-based representation using the first sense disambiguation strategy and adding concepts to the resource representation using relative, binary and absolute weighting.

By observing the figure it can be concluded that using pre-processing strategies improved baseline results. The simplest pre-processing strategy, the iconography filter, held the best results. Using the language detection or the Porter algorithm did not significantly improve classification results in spite of being more computationally expensive alternatives. The use



(a) Top-level categories (b) Low-level categories

Fig. 3 Classification results using tag-based representations (*baseline*)

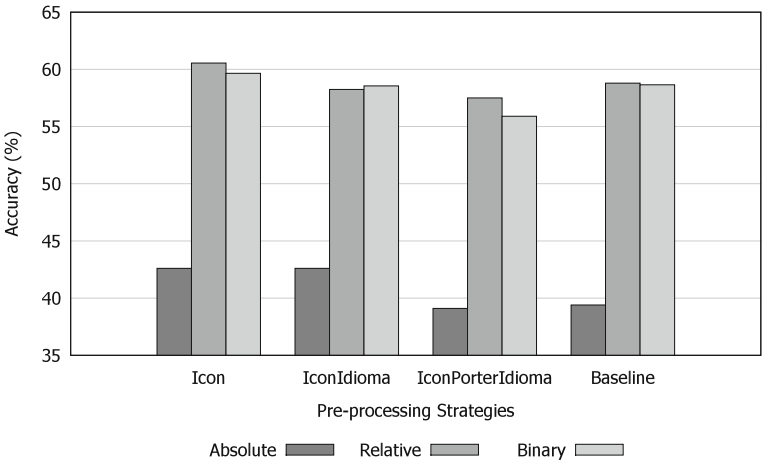


Fig. 4 Classification results using concept-based representations, *WordNet* and different pre-processing strategies (top-level categories)

of restrictive pre-processing strategies seemed to limit the amount of available information causing difficulties in the semantic enrichment based on *WordNet*.

After evaluating the effects of the pre-processing strategies, experiments were focused on the performance of semantically enriched representations using *WordNet*. The performance of SVM classifiers for every combination of weighting, disambiguation and document representation strategies was evaluated using the best performing pre-processing strategy (iconography filter). Figure 5 shows the obtained results. The absolute weighting held the worst results, whereas the relative weighting strategy slightly outperformed the simple binary weighting.

Considering a relative weighting of tags, Fig. 6 shows the performance of the three disambiguation strategies explained in Section 3.3.3. The best results were obtained using the

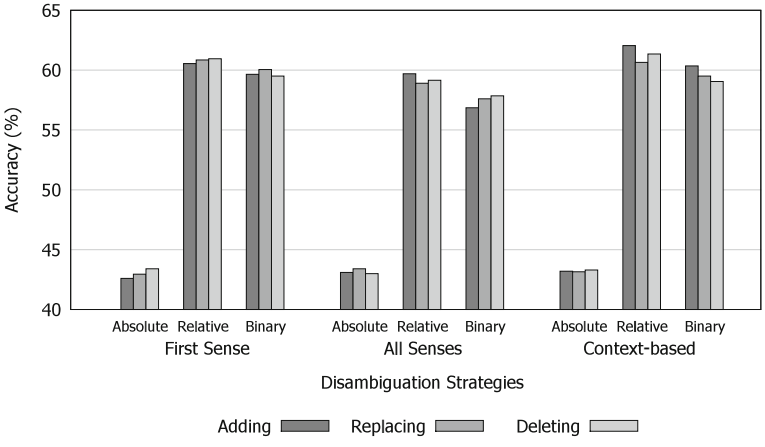


Fig. 5 Evaluation of disambiguation, weighting and resource representation strategies based on *WordNet* (top-level categories)

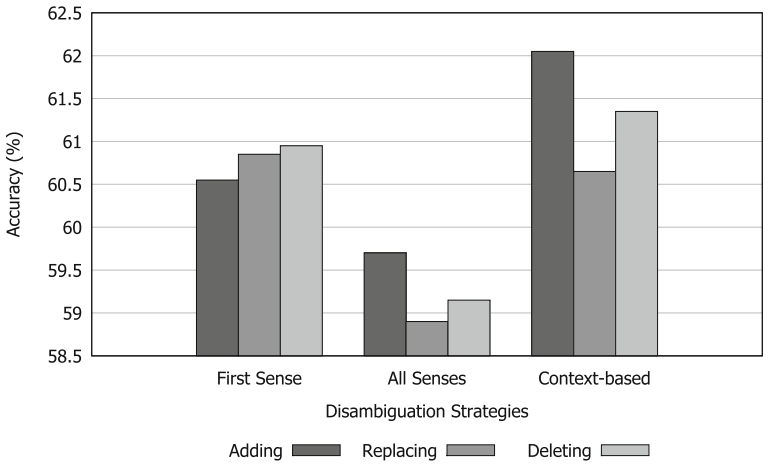


Fig. 6 Evaluation of disambiguation and resource representation strategies based on *WordNet* (top-level categories)

context-based disambiguation. As expected, the worst results were obtained when all the tags senses were included in the final representations, proving that an indiscriminate semantic enrichment is not useful for resource classification. Also, that strategy maximises the number of attributes belonging to each resource, thus increasing computational complexity of classifications. Regarding the resource representation strategies, results did not allow to determine the superiority of any of them as the best performing varied according to the disambiguation strategy used. The maximum accuracy was obtained by adding concepts to the original tag-based representation after context disambiguation. As the figure shows, all the resource representation strategies had a similar behaviour as the number of considered instances increased. In consequence, a dominant strategy could not be chosen. Results improved as the classifier had more available training information.

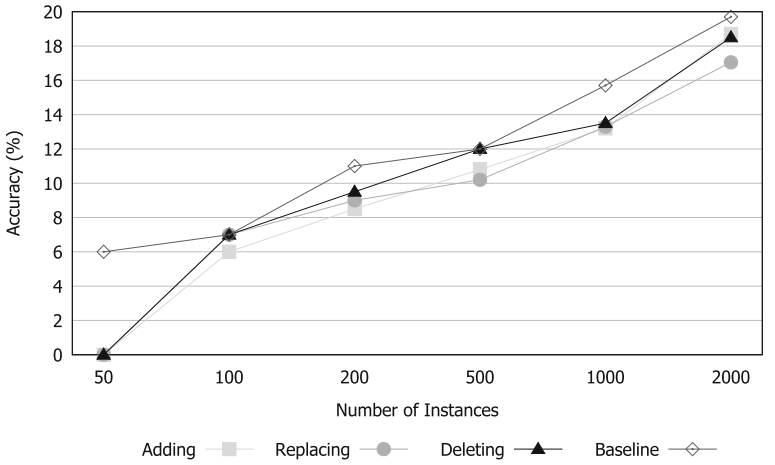


Fig. 7 Best Top-level categories strategy applied to Low-level categories (*WordNet*)

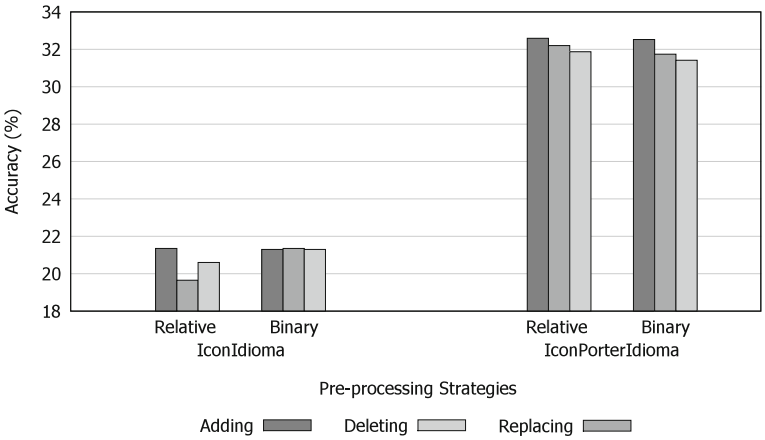


Fig. 8 Evaluation of pre-processing, weighting and resource representation strategies based on *WordNet* (low-level categories)

In relation to classification exclusively based on tags for low-level categories, using the best combination of strategies obtained for the top-level categories did not significantly improve baseline results, as shown in Fig. 7. The findings showed that the best performing semantic enrichment for classification based on the top-level classes is not effective when low-level classes are considered, suggesting that categories at different levels have different semantic enrichment needs. In this case, for choosing the correct sense of a tag, the first sense disambiguation strategy resulted more effective than the one based on the context (other tags assigned to the resource). By selecting other pre-processing strategies, classification results improved the baseline in increasing percentages as the number of evaluated instances grew.

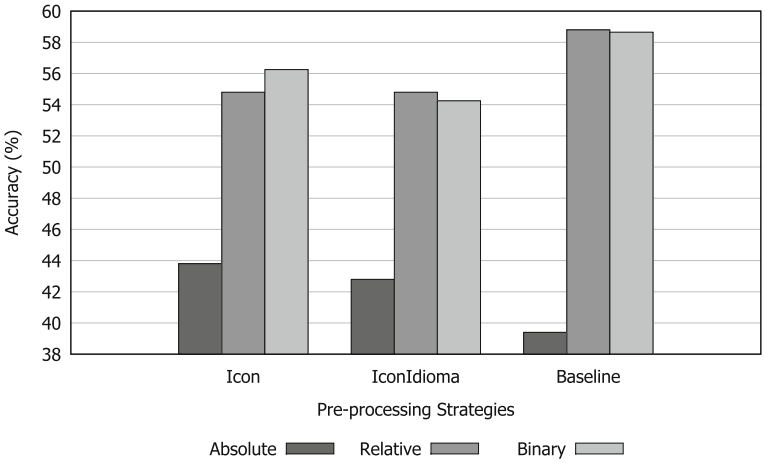


Fig. 9 Classification results using concept-based representations, *Wikipedia* and different pre-processing strategies (top-level categories)

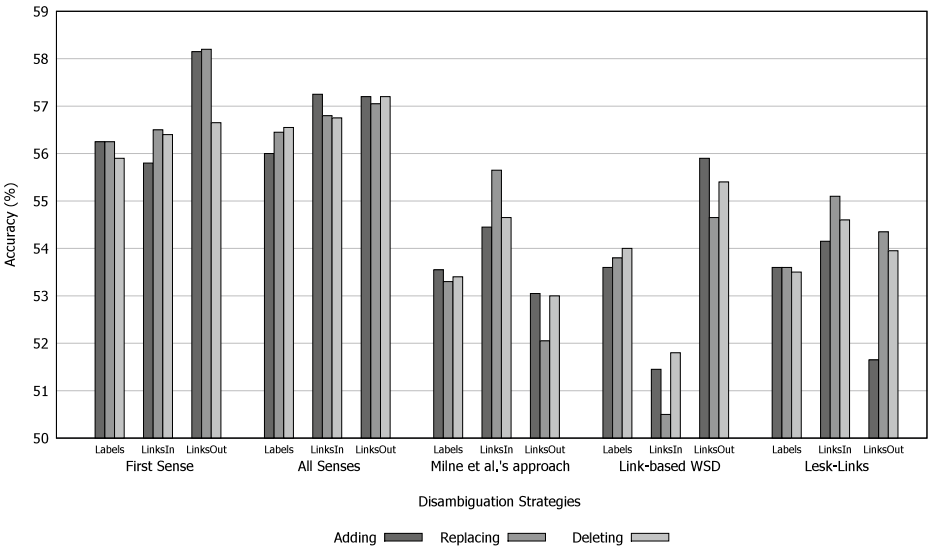


Fig. 10 Evaluation of disambiguation, origin of concepts and resource representation strategies based on *Wikipedia* (top-level categories)

Figure 8 presents classification results for the other pre-processing strategies, first sense disambiguation and weighting strategies. All combinations improved baseline results. The iconography filter followed by the Porter stemmer algorithm and the language detection improved baseline results by a 12 %, whereas the iconography filter followed by the language detection improved results by an 8 %. It was not possible to observe significant differences between the various resource representation or the weighting strategies, suggesting that the least computationally complex strategies can be used.

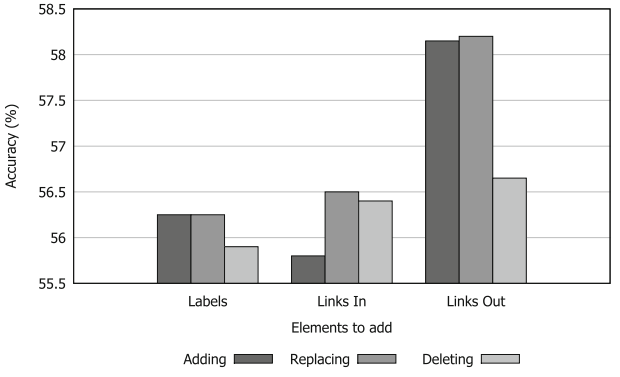


Fig. 11 Evaluation of origin of concepts and resource representation strategies based on *Wikipedia* (top-level categories)

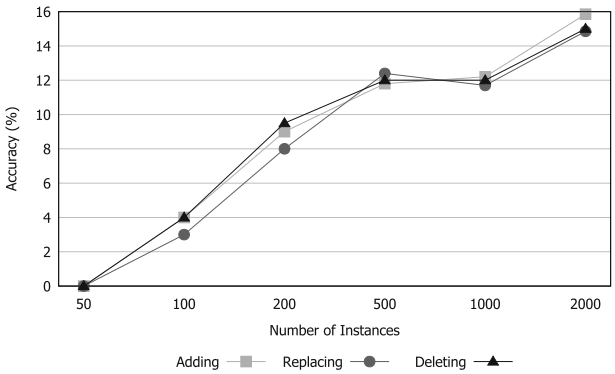


Fig. 12 Best Top-level categories strategy applied to Low-level categories (*Wikipedia*)

4.4.2 Wikipedia results

In order to evaluate semantic enrichment using *Wikipedia*, Fig. 9 shows a comparison between the Iconography filter (Icon) and the Iconography filter combined with language detection (IconIdioma) pre-processing strategies together with the alternative labels for each identified concept by using the first sense disambiguation strategy and relative, binary and absolute weighting strategies. These results were consistent with those obtained by using *WordNet*. Absolute weighting was the worst performing weighting strategy with differences up to a 13 % regarding the best results obtained by using binary weighting. Also, as the number of examples in the dataset increased, the classifier increased its predictive capacity. Regarding the pre-processing strategies analysed, the simplest one obtained the best results whereas the most computationally complex strategy did not significantly improve the baseline results.

Regarding weighting strategies, as well as when analysing *WordNet*, the best results were obtained for the simplest weighing strategy, i.e. the binary, which outperformed the relative weighing strategy by a 2 % or 3 % depending on how tags were enriched, i.e. alternative labels, incoming links or outgoing links. The absolute weighting strategy obtained the worst results up to a 15 % lower than the other alternatives.

Considering a binary weighting strategy, Fig. 10 shows the performance of the five disambiguation strategies presented in Section 3.3.1. The best results were obtained by selecting the most common sense associated to a tag, which outperformed the results of the other disambiguation strategies. On the contrary, the context disambiguation strategies' results were even lower than those of the all senses strategy. Oppositely to results achieved with *WordNet* concept-based representations, these results suggested that adding all possible tag senses is preferable than context-based disambiguation. These results may be explained by the nature and size of the context used by the disambiguation strategies in *Wikipedia* that considerably increased the number of attributes taken into account during the learning process.

Using first sense disambiguation, Fig. 11 shows the results of combining the origin of the concepts (Labels, LinksIn, LinksOut) and the different resource representation strategies. As regards the last issue, the worst results were obtained by using only concepts. On the other hand, the resource representation strategies based on adding or replacing concepts had similar results. In summary, the best results were obtained by combining concepts with

Table 8 Summary of evaluated strategies and results

Category Level	Preprocessing Strategies	Resource Representation Strategies	Disambiguation Strategies	Weighting Strategies	
(a) WordNet					
Top-Level	Icon	None of them showed a consistent superiority.	Context-based followed by first sense disambiguation.	Relative weighting.	
Low-Level	IconPorter-Idioma	None of them showed a consistent superiority.	First sense disambiguation.	None of them showed a consistent superiority.	
Category	Pre-processing	Resource	Disambiguation	Weighting	Origin of Concepts
(b) Wikipedia					
Top-Level	Icon	Replacing followed by deleting.	First sense disambiguation outperformed context-based strategies up to a 6 %.	Binary followed by relative weighting.	Outgoing links outperformed the other strategies by a 2 %.
Low-Level	Icon	For few instances deleting, in other cases replacing.	First sense disambiguation.	Binary.	Outgoing links.

tags having no associated concepts, i.e. not covered by the lexical database. There was a significant difference between the enrichment strategies presented. Considering the origin of concepts, alternative labels (i.e. tags synonyms) failed to obtain as accurate results as using outgoing links.

In summary, experiments showed that the maximum accuracy can be obtained by using a binary weighting strategy combined with a first sense disambiguation and replacing the

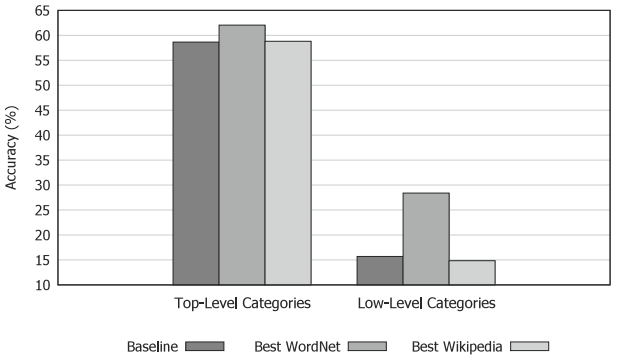


Fig. 13 Summary of classification results

Table 9 Best strategies accuracy results comparison

	50	100	200	500	1000	2000
Tags - Top-Level category	26	39	44	54.8	58	58.65
WordNet - Top-Level category	26	36	38.5	51.8	57	62.05
Wikipedia - Top-Level category	32	39	42	53.4	55.5	58.8
Tags - Low-Level category	6	7	11	12.8	15.7	18.5
WordNet - Low-Level category	4	2	15.5	23.4	28.4	-
Wikipedia - Low-Level category	0	3	8	12.4	11.7	14.85

tags with the outgoing links associated to their concepts. That combination outperformed baseline results.

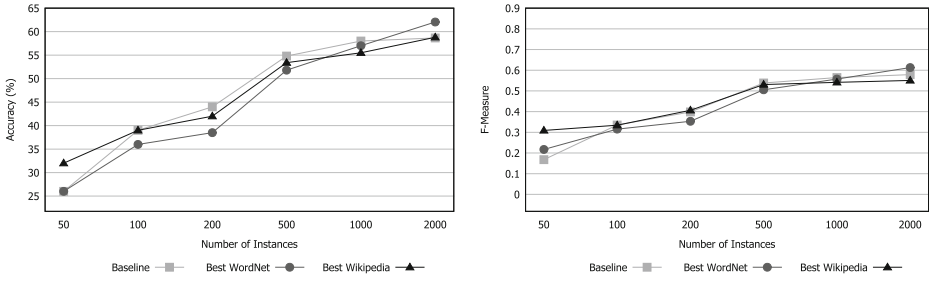
The best combination of strategies for classifying resources based on the top-level categories was evaluated for classifying the low-level categories. Figure 12 shows the results obtained for all datasets. The obtained results did not improve the results of using *WordNet* as the lexical database nor the baseline. These findings suggest that *Wikipedia* might not be adequate as a lexical database to semantically enrich resources when the low-level categories are considered due to its computational complexity and the fact that the classification results were worsened.

4.4.3 Summary of results

Table 8 presents a summary of the evaluated strategies and the achieved results. Figure 13 shows the results obtained for the best combination of strategies for top and low-level categories.

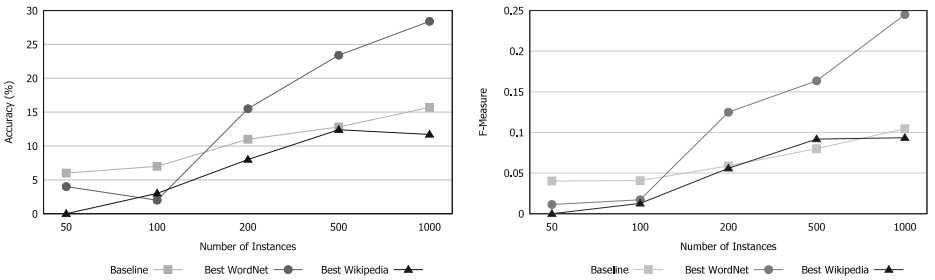
The best combination of strategies obtained for the top-level category was evaluated for the low-level categories of the ODP hierarchy. Figure 7 presents the results for the iconography filtering, relative weighting, context disambiguation, and the three resource representation strategies since previous results were inconclusive regarding the method to incorporate concepts to the original tag-based representations.

Table 9, and Figs. 14 and 15 summarise the improvements in terms of F-Measure and accuracy of resource classification results achieved by using the enriched representations of resources over tag-based representations. In both figures it can be observed that, the bigger the dataset the more important semantic information becomes for finding the correct



(a) Accuracy (b) F-Measure

Fig. 14 Best Top-level categories strategies comparison



(a) Accuracy (b) F-Measure

Fig. 15 Best Low-level categories strategies comparison

category of resources, and thus, for improving baseline results. This can be caused by the wider tag space resulting of including more resources, which introduces noise and increases ambiguity during classification. When the top-level categories were considered, *Wikipedia* was more effective than *WordNet* when the dataset had few instances, whereas for bigger datasets *WordNet* obtained better results. When the low-level categories were considered, baseline results were significantly outperformed by using *WordNet* as the source of the semantic enrichment. On the contrary, when *Wikipedia* was used as the source of the semantic enrichment, the results of experiments did not improve baseline results. In consequence, *WordNet* was more effective for semantic enrichment of tags.

Finally, an important issue to analyse is how the semantic enrichment affected the space of attributes for learning and classification. Table 10 and Fig. 16 show the original number of attributes when only tags were considered, as well as the number of attributes for different combinations of strategies. For each lexical database the disambiguation strategy that held the best results is included in the comparison. On the other hand, the chosen resource representation strategy is the one that maximises the number of attributes, i.e. the combination of tags and concepts.

WordNet caused the minor increment in the size of the attribute space. On the contrary, *Wikipedia* increased the size of the attribute space up to 300 times when the incoming links were considered, and up to 40 times when the outgoing links were considered. As a result, the resource representation is likely to include noise, degrading resource classification results. It is important to highlight that, as the size of the resource representation increased, the temporal complexity of classification also increased. In summary, due to the fact that *WordNet* optimised the relation between classification results and the attribute representation size, it was the best database to carry out the semantic enrichment of tags.

Table 10 Attribute space size comparison

	50	100	200	500	1000	2000
Tags	311	512	877	1,692	2,641	4,109
Wikipedia-Labels	12,089	17,268	28,003	46,189	64,068	87,745
Wikipedia-Links In	464,333	581,511	859,106	1,009,612	1,127,591	1,237,994
Wikipedia-Links Out	29,534	41,482	61,713	91,189	120,942	160,422
WordNet	621	1,022	1,681	3,021	4,466	6,648

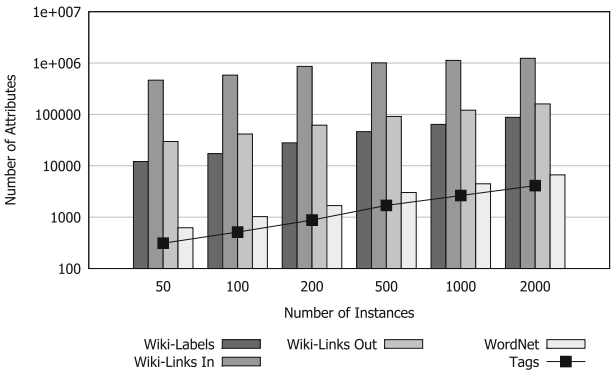


Fig. 16 Comparison of attribute space of different representation strategies

5 Conclusions

This work analysed and evaluated strategies to incorporate semantics to representation of resources from 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 categorise resources. Semantic enrichment can be also applied to other tasks, such as the creation of semantic tag-based profiles to provide personalised social tag recommendations (Hsu 2013). 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, *Social-ODP-2k9*, showed that the semantic enrichment of tags has a positive effect on resource classification. It was also observed that the more instances were used to train the classifiers, the higher the superiority of semantic representations in comparison with simple tag-based representations.

The best classification results when using *WordNet* as the source of the semantic enrichment for the top-level categories were obtained by combining an iconography filtering of tags, a context disambiguation strategy and a mixing of tags and concepts in the final representation of resources. However, when the same combination of strategies was evaluated for the low-level categories, it did not obtain the best results, exposing the different needs of information enrichment and disambiguation at each category level. The findings suggest that when low-level categories are considered, less noise is allowed and, thus, strategies that prevent its increment are preferable. However, those strategies can decrease the size and quality of the context needed for performing a disambiguation, which affects accuracy. In this context, choosing the most common sense of tags (i.e. first sense strategy) in combination with an iconography filter, the stemming algorithm and language detection reported the best results, which significantly outperformed baseline results. Since it was not possible to observe significant differences between the various resource representations nor the weighting strategies, the least computationally expensive ones can be chosen.

When *Wikipedia* was used as the source of the semantic enrichment of representations, the outgoing links combined to concepts selected by a first sense disambiguation strategy achieved the best results. This combination of strategies was more effective than the best combination of strategies selected for *WordNet* for classifying few instances. On the contrary, as the number of instances increased, *WordNet* significantly outperformed the results

obtained by *Wikipedia*. Context-based disambiguation strategies were not effective as their results were lower than the baseline ones. These findings suggests that the context provided to the strategies was neither useful nor enough for the strategy to choose the correct sense of tags. Furthermore, *Wikipedia* did not prove to be useful as a source of semantic enrichment of resources in low-level or more specific categories.

Acknowledgments This work has been partially funded by ANPCyT (Argentina) under grant PICT-2011-0366 and by CONICET (Argentina) under grant PIP No. 112-201201-00185.

References

Agirre, E., & Rigau, G. (1996). Word sense disambiguation using conceptual density. In *Proceedings of the 16th conference on computational linguistics - Volume 1, ACL, (COLING '96)* (pp. 16–22). Copenhagen, Denmark.

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.A., & Ribeiro-Neto, B.A. (1999). *Modern information retrieval*. Boston: Addison-Wesley Longman Publishing Co. Inc.

Buckley, C. (1993). The importance of proper weighting methods. In *Proceedings of the workshop on human language technology, association for computational linguistics, (HLT '93)* (pp. 349–352). Princeton, New Jersey.

Cavnar, W.B., & Trenkle, J.M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval* (pp. 161–175).

Dagher, G.G., & Fung, B.C.M. (2013). Subject-based semantic document clustering for digital forensic investigations. *Data & Knowledge Engineering (DKE)*, *86*, 224–241.

Dattolo, A., Eynard, D., Mazzola, L. (2011). An integrated approach to discover tag semantics. In *Proceedings of the 2011 ACM symposium on applied computing, ACM, (SAC '11)* (pp. 814–820). TaiChung, Taiwan.

Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Fellbaum, C. (2005). Wordnet and wordnets. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (pp. 665–670). Oxford: Elsevier.

Fogarolli, A. (2009). Word sense disambiguation based on wikipedia link structure. In *Proceedings of the 2009 IEEE international conference on semantic computing, IEEE Computer Society, (ICSC '09)* (pp. 77–82). Washington, DC.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on artifical intelligence, (IJCAI'07)* (pp. 1606–1611). Hyderabad: Morgan Kaufmann Publishers Inc.

Hotho, A., Staab, S., Stumme, G. (2003). 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 informaion retrieval, (SIGIR 2003)*. Toronto Canada.

Hotho, A., Jäschke, R., Schmitz, C., Stumme, G. (2006). Bibsonomy: a social bookmark and publication sharing system. In A. de Moor, S. Polovina, H. Delugach (Eds.), *Proceedings of the conceptual structures tool interoperability workshop at the 14th international conference on conceptual structures*. Aalborg: Aalborg University Press.

Hsu, I.C. (2013). Integrating ontology technology with folksonomies for personalized social tag recommendation. *Applied Soft Computing*, *13*(8), 3745–3750. doi:10.1016/j.asoc.2013.03.004, http://www.sciencedirect.com/science/article/pii/S1568494613001087.

Huang, A., Milne, D., Frank, E., Witten, I.H. (2009). Clustering documents using a wikipedia-based concept representation. In *Proceedings of the 13th Pacific-Asia conference on advances in knowledge discovery and data mining, (PAKDD '09)* (pp. 628–636). Bangkok: Springer-Verlag.

Jankowski, N., & Usowicz, K. (2011). Analysis of feature weighting methods based on feature ranking methods for classification. In *Proceedings of the 18th international conference on neural information processing, (ICONIP'11)* (pp. 238–247). Shanghai: Springer-Verlag.

Kohavi, R., Langley, P., Yun, Y. (1997). The utility of feature weighting in nearest-neighbor algorithms. In *Proceedings of the 9th European conference on machine learning* (pp. 85–92). Springer-Verlag.

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: ACM.

Lan, H. (2011). *Concept-based text clustering*. PhD thesis, University of Waikato, New Zealand.

Lan, M., Tan, C.L., Low, H.B., Sung, S.Y. (2005). A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *Special interest tracks and posters of the 14th international conference on world wide web, ACM, (WWW '05)* (pp. 1032–1033). Chiba, Japan.

Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. How to represent texts in input space? *Machine Learning*, *46*(1–3), 423–444.

Lesk, M. (1986). 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, ACM, (SIGDOC '86)* (pp. 24–26). Toronto, Canada.

Lops, P., de Gemmis, M., Semeraro, G., Musto, C., Narducci, F. (2013). Content-based and collaborative techniques for tag recommendation: an empirical evaluation. *Journal of Intelligent Information Systems*, *40*(1), 41–61. doi:10.1007/s10844-012-0215-6.

Maree, M., & Belkhatir, M. (2013). Coupling semantic and statistical techniques for dynamically enriching web ontologies. *Journal of Intelligent Information Systems*, *40*(3), 455–478. doi:10.1007/s10844-012-0233-4.

Mathes, A. (2004). *Folksonomies - cooperative classification and communication through shared metadata.* Computer Mediated Communication.

Medelyan, O., Milne, D., Legg, C., Witten, I.H. (2009). Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, *67*(9), 716–754.

Milne, D., & Witten, I.H. (2008a). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI workshop on wikipedia and artificial intelligence: an evolving synergy* (pp. 25–30). AAAI Press.

Milne, D., & Witten, I.H. (2008b). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on information and knowledge management, ACM, (CIKM '08)* (pp. 509–518). Napa Valley, California.

Milne, D., & Witten, I.H. (2009). An open-source toolkit for mining Wikipedia. In *Proceedings of the New Zealand computer science research student conference, (NZCSRSC'09)*(Vol. 9).

Navigli, R. (2009). Word sense disambiguation: a survey. *ACM Computing Surveys*, *41*(2), 1–69.

Noll, M.G., & Meinel, C. (2007). Authors vs. readers: a comparative study of document metadata and content in the www. In *Proceedings of the 2007 ACM symposium on document engineering, ACM, (DocEng '07)* (pp. 177–186). Winnipeg, Manitoba, Canada.

Noll, M.G., & Meinel, C. (2008). Exploring social annotations for web document classification. In *Proceedings of the 2008 ACM symposium on applied computing, SAC '08* (pp. 2315–2320). New York: ACM.

Platt, J.C. (1999). *Advances in kernel methods.* MIT Press, Cambridge, MA, USA, chap Fast training of support vector machines using sequential minimal optimization, (pp. 185-208).

Porter, M. (1997). *Readings in information retrieval.* Morgan Kaufmann Publishers Inc., CA, USA, chap An algorithm for suffix stripping, (pp. 313–316).

Rijsbergen, C.Jv. (1979). *Information retrieval*, 2nd edn. Newton: Butterworth-Heinemann.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Schütze, H., & Silverstein, C. (1997). Projections for efficient document clustering. In *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval, (SIGIR '97)* (pp. 74–81). Philadelphia: ACM.

Solskinnsbakk, G., Gulla, J.A., Haderlein, V., Myrseth, P., Cerrato, O. (2012). Quality of hierarchies in ontologies and folksonomies. *Data & Knowledge Engineering*, *74*, 13–25.

Strube M, & Ponzetto SP (2006). Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st national conference on artificial intelligence, (AAAI'06)* (pp. 1419–1424). MA: AAAI Press.

Vapnik, V.N. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

Yin, Z., Li, R., Mei, Q., Han, J. (2009). 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)* (pp. 957–966). Paris: ACM.

Zipf, G.K. (1935). *The Psychobiology of Language.* Houghton-Mifflin.

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, ACM, (DocEng '09)* (pp. 74–83). Munich, Germany.

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

Zubiaga, A., Martínez, R., Fresno, V. (2011b). Analyzing tag distributions in folksonomies for resource classification. In *Proceedings of the 5th international conference on knowledge science, engineering and management, (KSEM'11)* (pp. 91–102). Irvine: Springer-Verlag.