Exploiting Social Annotations for Resource Classification

Arkaitz Zubiaga, Víctor Fresno, Raquel Martínez

NLP & IR Group @ UNED

Madrid, Spain

ABSTRACT

The lack of representative textual content in many resources suggests the study of additional metadata to improve classification tasks. Social bookmarking and cataloging sites provide an accessible way to increase available metadata in large amounts with user-provided annotations. In this chapter, we study and analyze the usefulness of social annotations for resource classification. We consider as a resource anything that can be socially annotated online. We conclude that social annotations could enhance resource classifiers in many cases, and we present a method to get the most out of them using classifier committees.

INTRODUCTION

Resource classification is the task of labeling resources with their corresponding categories from a predefined taxonomy. Resource classification is of vital importance for information management and retrieval tasks, and for assisting the semi-automatic development of categorized directories. In the case of web pages, it is also essential to focused crawling, and to topic-specific web link analysis, among others. It can also help improve search results when it is applied to organizing ranked results.

To carry out this kind of tasks automatically, the textual content is commonly used to represent the resource to classify. Many times, the lack of representative content makes it insufficient, though (Qi & Davison, 2009). In this way, social bookmarking sites present an accessible way to get additional descriptive metadata.

Social bookmarking is a Web 2.0 based phenomenon that allows users to describe web contents by annotating them with different kinds of metadata in a collaborative and aggregated way. Websites like Delicious[i], StumbleUpon[ii], LibraryThing[iii] and Diigo[iv], among others, allow their users to add information to a web page, collecting hundreds of thousands of annotations per day (Heymann et al., 2008).

This user-generated data is added in several ways: trying to give a topical description of the resource by means of a set of tags; offering subjective or personal assessments; adding free texts as descriptions; making personal valuations of contents, etc. As a result, a global community of volunteer users creates a huge repository of described resources that can ease their subsequent retrieval.

Until now, the use of social annotations for resource classification tasks has remained relatively unexamined. A few works have shown the suitability of social tags for this kind of tasks. Nonetheless, the study of the optimal representation based on social tags, and the use of social annotations other than social tags, are still unexplored.

In this chapter, we study and analyze the use of metadata extracted from social bookmarking and cataloging sites to classify a set of annotated resources. We perform the experiments with two different types of resources: web pages and books. We find two types of social annotations to be applicable and useful for resource classification: tags and comments provided by end users. We conclude proposing a way to represent each kind of annotation, and we present a method to outperform their results by means of combining different data using classifier committees.

Next, in Section Background, we describe the nature of social annotations and the existing types. We continue in Section Related Work presenting the earlier works in the literature. After that, we detail the settings of our experiments as well as the datasets we used, to continue in Section Results with the analysis of the results. We discuss them in Section Discussion. Finally, we conclude with our thoughts and future work.

BACKGROUND

Social bookmarking and cataloging sites allow users to save and annotate their preferred resources, sharing them with the community. These annotations are made in a collaborative way, so that it makes possible a large amount of metadata to be available for each resource. Going into further details on these metadata, different kinds of user-generated annotations can be defined:

- **Tags:** keywords defining and characterizing a resource are known as tags. In collaborative tagging systems, each user u_i can post a resource r_j with a set of tags $T_{ij} = \{t_1,...,t_p\}$, with a variable number p of tags. After k users posted r_j, it is described with a weighted set of tags $T_j = \{w_1\ t_1,...,w_n\ t_n\}$, where $w_1,...,w_n \leq k$. The resulting organization from users' tagging activity is known as a folksonomy.
- **Notes or descriptions:** free text describing a resource is known as a 'note' or 'description'.
- **Reviews:** a review is a free text valuating a web page. Even though this kind of annotations can initially look subjective and non-descriptive, users tend to mix descriptive texts with opinions.
- **Ratings:** valuations indicating to what extent users like or dislike a resource, commonly by means of punctuations from one to five.
- **Highlights:** highlights are only applicable when the resources are web pages, by selecting the most relevant part or parts of the bookmarked web page.

In most of the social bookmarking systems, there are no constraints on the keywords users can set as tags.

The use of tags was originally suggested to make easier the later search and retrieval of relevant documents. Many of the works in this area focus on the study of dataset properties (Ramage et al., 2009), the analysis of usage patterns of tagging systems (Golder & Huberman, 2006), and the discovery of hidden semantics in tags (Yeung et al., 2008). Incorporating social annotations with document content and other sources of information is a natural idea (Zhou et al., 2008), especially when trying to improve information management tasks.

Finally, most of the annotations described above in this Section seem to be really interesting for topical resource tasks. Nonetheless, it is obvious that 'Ratings' cannot contribute to this kind of classification, since they do not provide topical information. For this reason, in the process of creation of our dataset we based on all the social annotations but ratings. Thus, we consider three families grouping the remaining annotations: 'Tags', 'Notes & Reviews' (grouped as 'Comments'), and 'Content'. In our experiments, 'highlights' were not considered due to their low representativity over the web pages, as we point out later, so that they are not listed above. Moreover, they are not available when the resources are books.

Nature of Social Annotations

Every annotation becomes social when it is available either on a public Internet website or a private network where a community is involved. Any user of the website can then access these annotations. Not all the annotations are provided in the same way, though. The type of social media the user interacts with may define some constraints, mainly by setting who is able to annotate each item. In this sense, two kinds of annotations can be distinguished (Smith, 2008):

- **Simple annotations:** users describe their own resources or items, such as photos on Flickr.com, news on Digg.com or videos on Youtube.com, but nobody else annotates others' resources. Usually, the author of the resource is who annotates it. This means no more than one user tags a resource.
- **Collaborative annotations:** many users annotate the same resource, and every person can tag and describe it with their own tags and reviews in their own vocabulary. The collection of tags assigned by a single user creates a smaller folksonomy, also known as personomy. As a result, several users tend to post the same resource. For instance, CiteULike.org, LibraryThing.com and Delicious are based on collaborative annotations, where each resource (papers, books and URLs, respectively) can be annotated and tagged by all the users who considered it interesting.

These terms are specially applied to tagging systems. Tags present high likelihood to coincide across users, making the aggregated tags of collaborative tagging systems especially strong rather than simple tagging systems. Notwithstanding this, the collaborative idea can also be applied to any other annotations, as different users use to provide their own annotations (e.g., reviews) to the same resource.

In this chapter, we rely on collaborative annotations in social bookmarking and cataloging systems, taking advantage of the strength of the aggregated annotations provided by large amounts of users.

Related Work

There are a few works in the literature analyzing the usefulness of social tags specifically for web page organization tasks, but no enough attention has been paid to other kind of social annotations. In Ramage et al. (2009) the inclusion of tagging data improved the performance of two clustering algorithms when compared to content-based clustering. They found that tagging data was more effective for specific collections than for a collection of general documents.

Noll & Meinel (2008a) present a study of the characteristics of social annotations provided by end users, in order to determine their usefulness for web page classification. The authors matched user-supplied tags of a page against its categorization by the expert editors of the ODP. They analyzed at which hierarchy depth matches occurred, concluding that tags may perform better for broad categorization of documents rather than for more specific categorization. The study also points out that since users tend to bookmark and tag top level web documents, this type of metadata will target classification of the entry pages of websites, whereas classification of deeper pages might require more direct content analysis. They observed that tag noise (the opposite of popular tags) provides helpful data for information retrieval and classification tasks in general. In a previous work, the same authors (Noll & Meinel, 2007) suggested that tags provide additional information about a web page, which is not directly contained within its content.

Also, Noll & Meinel (2008b) studied three types of metadata about web documents: social annotations (tags), anchor texts of incoming hyperlinks, and search queries to access them. They concluded that tags are better suited for classification purposes than anchor texts or search keywords.

The above works had shown the usefulness of social tags for web page organization tasks but, to the best of our knowledge, different tag representations, and social annotations other than tags have not been applied.

In a previous work (Zubiaga et al., 2009a), we presented a preliminary study on the use of social annotations for web page classification, applied to the top level of the ODP categorization scheme, where social tags and comments show high performance against textual content. In this chapter, we extend this work by analyzing the usefulness of social annotations for other kinds of resources.

MAIN FOCUS OF THE CHAPTER

Raw Data

Our experiments require large collections of resources with both a considerable number of annotations and categorization within standard and consolidated taxonomies. Golder and Huberman found that after about 100 users tagged a specific web page, the distribution of its top tags tends to converge (Golder & Huberman, 2006). Thus, we considered 100 users as a threshold for a bookmarked resource to be popular.

Next, we introduce the two datasets we use in our experiments, made up by sets of categorized web pages and books, which include social annotations provided by users of well-known social tagging sites. An important feature of these categorized resources is that all of them show a marked imbalancement in the distribution across the categories.

Web Pages Dataset

As a starting point, we monitored the recent feed of Delicious looking for popular bookmarked URLs during December 2008 and January 2009. This resulted in an initial list of 94,130 unique URLs. As we decided to consider the Open Directory Project (ODP) as a categorization scheme for the web pages, we looked for all the available matches between this list and URLs making up the ODP. We found that 12,616 URLs had a category assigned in ODP. As a few of them had two categories assigned, we randomly selected one of them.

Then, with this final list of URLs with their corresponding ODP category data, we fetched the page content for each of the URLs. In addition to the page content and the corresponding categorization of this set of pages, we gathered the following data from social bookmarking sites:

- **Bookmark data from Delicious:** Delicious is a social bookmarking site where each user can annotate with tags and describe with notes their preferred web pages. When gathering information from this site, we saved the following data for each web page:
    - **Number of users** bookmarking it, what is always equal or higher than 100.
    - **Top 10 list of tags** annotated by users, along with their corresponding weights, referring to the number of users.
    - **Notes** provided by users. This is optional, though, and not all the users fill in this field. Within our dataset, we found that roughly 10% of the bookmarks have a note attached to it.
    - The **Full Tag Activity (FTA)**. This includes an exhaustive list of users bookmarking each page, with the tags provided by each of them, so that a list of top tags larger than 10 can be inferred. We refer to each of the annotations made by different users for a URL as a bookmark. Note that the FTA is limited by the system to the 2,000 last users, whereas the top 10 list provides data corresponding to all the users, even when more than 2,000 bookmarked it. In our dataset, 957 web pages were saved by more than 2,000 users, with an average user count of 5,329.
- **Reviews** from StumbleUpon: StumbleUpon is a social bookmarking site intended for helping users discover new web pages. By navigating over the pages suggested by the site, it allows users to rate and describe them. On this site, we looked for reviews provided by users to our list of URLs, and we found that 9,919 of them have review information.
- **Highlights** from Diigo: Diigo is another social bookmarking site, with a new feature that

allows users to highlight and add sticky notes to web pages. We looked in this site for highlight data available for our URLs, but only 1,920 of the documents in our dataset were provided highlight information, so that we decided not to use this information in our study due to its low availability.

Summarizing, our final dataset is composed by 12,616 unique URLs with their corresponding ODP categorization, page content and incoming anchor texts, along with a set of social annotations including tagging data, notes and reviews.

We rely on the hierarchical structure of the Open Directory Project, a human-edited web directory, as the categorization scheme. Particularly, we experiment the classification by using the top level of the hierarchy, which is made up by 17 categories.

For this dataset, this is how we group the available data:

- **Content:** we consider the textual content of a web page as we crawled it from the Web.
- **Comments:** we merge notes from Delicious and reviews from StumbleUpon.
- **Tags:** we use tags annotated by users on Delicious.

The dataset is available as a benchmark[v].

Books Dataset

For the second dataset, which is made up by books, we started by gathering a set of popular works from LibraryThing. In this process, we found a set of 65,929 popular books. In the next step, we looked for classification labels assigned by experts to these books. We fetched their classification for both the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC) systems. The former is a classical taxonomy that is still widely used in libraries, whereas the latter is used by most research and academic libraries. We found that 27,299 books were categorized on DDC, and 24,861 books have an LCC category assigned to it. In total, there are 38,149 books with category data from either one or both category schemes.

Then, we fetched the following data from social cataloging sites for each of the books:

- **Bookmark data from LibraryThing**:
  - **Number of users** bookmarking it, what is always equal or higher than 100.
  - The **Full Tag Activity (FTA)**, which is similar to Delicious.
  - **Reviews:** a free text with user comments on the book.
- **Reviews from GoodReads:** similar to the reviews on LibraryThing.
- **Product descriptions and reviews from Amazon**. Product descriptions are usually formal descriptions and reviews provided by editorials. Reviews, on the other hand, are provided by end users.
- **Synopses from Barnes&Noble**. The synopsis of each book includes a brief summary of the content of the book.

For this dataset, this is how we group the available data:

- **Content:** as we do not have the textual content of the books, we consider the product reviews and the synopses as a summary of their content.
- **Comments:** we merge user reviews from LibraryThing, GoodReads and Amazon for the comments.
- **Tags:** we use tags annotated by users on LibraryThing.

Support Vector Machines

Regarding the algorithm we use for the classification tasks, we rely on our study in Zubiaga et al. (2009b). We analyzed the suitability of several variants of Support Vector Machines (SVM) (Joachims, 1998) to topical web page classification tasks, considering them as multiclass problems. We concluded that supervised approaches outperform semi-supervised ones, and that considering the task as a single multiclass problem instead of several smaller binary problems performs better. Thus, we use supervised multiclass SVM for our experiments. Even though the traditional SVM approach only works for binary classification tasks, multiclass classification approaches have also been proposed and used in the literature (Weston & Watkins, 1999; Hsu & Lin, 2002). We use the freely available and well-known "svm-light"[vi] in its adapted multiclass version named "svm-multiclass", with the linear kernel and the default parameters suggested by the author for text classification tasks.

A multiclass SVM classifier for *k* classes defines a model with a set of hyperplanes in the training phase, so that they separate the documents in a class from the rest (Crammer & Singer, 2002). In the test phase, when making predictions for each new document, the classifier is able to establish a margin for each class. These margins refer to the reliability of the document to belong to each of the classes. The bigger is the margin, the more likely is the document to belong to the class. As a result, the class maximizing the margin value will be predicted by the classifier.

To evaluate the resource classification performance using different training set sizes, we randomly select 6 different runs for each of the training sets. We present the accuracy based on the average of the 6 runs, in order to get more realistic results. The accuracy represents the proportion of correct predictions among the whole test set.

Experiments

Tag-based Classification

Previous works suggest that tags could be used to classify web documents, and show encouraging results while using them (Noll & Meinel, 2008a; Aliakbary et al., 2009; Zubiaga et al., 2009a). Going further, we would like to resolve the following issues: which is the best way to exploit these metadata? And do they outperform the content-based classification even when classifying into narrower categories? Are they also useful for classifying resources other than web pages? Next, we propose, evaluate and compare several approaches for tag-based representation relying on these data:

- **Ranked Tags (Top 10):** tags corresponding to the top 10 list of a resource are assigned a value in a rank-based way. The first-ranked tag is always set the value 1, 0.9 for the second, 0.8 for the third, and so on. This approach respects the position of each tag in the top 10, but the different gaps among tag weights are ignored.
- **Tag Fractions (Top 10):** taking into account both the number of users who bookmarked a resource and the top list of tags, it is possible to define the fraction of users assigning each tag. A tag would have been annotated by the 100% of the users when its weight matches the user count of a resource, getting a value of 1 as the fraction. According to this, a value from 0 to 1 is set to each tag in the top 10. Thus, for the tag *i* in a resource annotated by *p* users, the value would be defined as w_i / p.
- **Unweighted Tags (Top 10 and FTA):** the only feature considered for these two representations are the occurrence or non-occurrence of a tag in the top 10 list or the full tag activity of a resource, depending on whether we rely on the top 10 of tags or the FTA, respectively. These approaches ignore tags' weights, and assign a binary value to each feature in the vector.

- **Weighted Tags (Top 10 and FTA):** the weight for each of the tags of a resource ($\{w_1,...,w_n\}$, as described above) is considered as it is in these two approaches, relying on the top 10 list of tags and the FTA, respectively. Now, by definition, the weights of the tags are fully respected, although the amount of users bookmarking a resource is ignored. Note that different orders of magnitude are mixed up now, since the count of bookmarking users range from 100 to higher values.

Note that for the approaches above relying on the FTA, the dimensionality of the vectors is reduced, in order to relax the computational cost while maintaining the representativity. The reduction consists of tags appearing only in a document.

Next, we present the results of the experiments for all the datasets and classification schemes.

Web Page Classification - ODP							
	Training Set Size						
	600	**1400**	**2200**	**3000**	**4000**	**5000**	**6000**
Tag Fractions	0.456	0.470	0.473	0.475	0.474	0.474	0.476
Tag Ranks	0.466	0.477	0.490	0.496	0.496	0.501	0.488
Unweighted Tags (Top 10)	0.503	0.515	0.519	0.522	0.527	0.520	0.524
Unweighted Tags (FTA)	0.523	0.552	0.557	0.563	0.561	0.566	0.569
Weighted Tags (Top 10)	0.510	0.574	0.604	0.620	0.634	0.641	0.652
Weighted Tags (FTA)	**0.526**	**0.590**	**0.616**	**0.636**	**0.645**	**0.654**	**0.665**

Table 1: Accuracy results for tag-based web page classification

Book Classification - DDC							
	Training Set Size						
	3000	**6000**	**9000**	**12000**	**15000**	**18000**	**21000**
Tag Fractions	0.719	0.717	0.720	0.721	0.727	0.721	0.724
Tag Ranks	0.791	0.783	0.778	0.782	0.788	0.787	0.797
Unweighted Tags (Top 10)	0.756	0.763	0.753	0.766	0.759	0.759	0.758
Unweighted Tags (FTA)	0.624	0.622	0.628	0.629	0.629	0.628	0.624
Weighted Tags (Top 10)	0.858	0.861	0.862	0.865	0.866	0.866	0.864
Weighted Tags (FTA)	**0.861**	**0.864**	**0.864**	**0.867**	**0.869**	**0.869**	**0.868**

Table 2: Accuracy results for tag-based book classification (DDC)

Book Classification - LCC							
	Training Set Size						
	3000	**6000**	**9000**	**12000**	**15000**	**18000**	**21000**
Tag Fractions	0.739	0.740	0.741	0.743	0.741	0.738	0.746

Book Classification - LCC							
Tag Ranks	0.783	0.790	0.788	0.783	0.789	0.795	0.790
Unweighted Tags (Top 10)	0.759	0.772	0.764	0.771	0.763	0.770	0.763
Unweighted Tags (FTA)	0.654	0.660	0.661	0.661	0.658	0.655	0.661
Weighted Tags (Top 10)	0.852	0.854	**0.856**	0.858	0.858	0.855	0.858
Weighted Tags (FTA)	**0.853**	**0.857**	**0.856**	**0.861**	**0.861**	**0.857**	**0.861**

Table 3: Accuracy results for tag-based book classification (LCC)

The results in Table 1, Table 2, and Table 3 show the marked inferiority of the ranked and fraction-based approaches. These two representations do not seem to be a good way to carry out a topical classification task.

On the other hand, it is clear that the approaches considering weights are far better than the unweighted approaches, what shows the relevance of considering the agreement between users, rather than just whether or not a tag appears annotated in a resource.

Among the weighted approaches, the difference is not so clear, but the use of the FTA instead of just the top 10 list of tags gets slightly better results. This suggests that the tags in the top 10 are the most relevant for the classification task, as they have been annotated by a bigger number of users, but the tags in the tail only considered in the FTA also provide some useful information.

This suggests considering the full tag activity, so that annotations of users differing from the most common behaviors may also be helpful. When the computational cost matters, though, it could be enough to consider just the top 10 tags to reduce it.

Other ideas like a possible removal of useless or harmful tags set by misbehaving users remain as an open issue, though.

Comparing Data: Content vs Comments vs Annotations

With the experiments above we found the best approach to represent resources using tags for classification tasks. Once we had these results, we compared the usefulness of tags as against to the other data inputs. Thus, we compare the results of tags to comments and content.

Web Page Classification - ODP							
	Training Set Size						
	600	**1400**	**2200**	**3000**	**4000**	**5000**	**6000**
Content	0.518	0.561	0.579	0.588	0.595	0.604	0.610
Comments	0.520	0.578	0.602	0.618	0.630	0.639	0.646
Weighted Tags (FTA)	**0.526**	**0.590**	**0.616**	**0.636**	**0.645**	**0.654**	**0.665**

Table 4: Accuracy results for different data inputs on web page classification

Book Classification - DDC							
	Training Set Size						
	3000	**6000**	**9000**	**12000**	**15000**	**18000**	**21000**
Content	0.767	0.792	0.802	0.809	0.809	0.815	0.817

Book Classification - DDC							
Comments	0.777	0.808	0.820	0.831	0.833	0.839	0.840
Weighted Tags (FTA)	**0.861**	**0.864**	**0.864**	**0.867**	**0.869**	**0.869**	**0.868**

Table 5: Accuracy results for different data inputs on book classification (DDC)

Book Classification - LCC							
	Training Set Size						
	3000	**6000**	**9000**	**12000**	**15000**	**18000**	**21000**
Content	0.767	0.789	0.798	0.803	0.806	0.807	0.810
Comments	0.780	0.803	0.816	0.823	0.827	0.828	0.833
Weighted Tags (FTA)	**0.853**	**0.857**	**0.856**	**0.861**	**0.861**	**0.857**	**0.861**

Table 6: Accuracy results for different data inputs on book classification (LCC)

Table 4, Table 5, and Table 6 show the results of the comparison of the three approaches. Our results show that both social annotations, tags and comments, improve the content-based baseline in either classification schemes.

Comparing the behavior of the social annotations, the results show higher performance for the approach using tag information than using comments. The use of comments outperforms the content-based classification, though.

Using Classifier Committees

Even though the tag-based representation outperforms the other two approaches, all of them offer encouraging results and look good enough to try to combine them and improve even more the classifier's performance. Though, what if a classifier is getting right while the others are making a mistake? Could we combine the results to get the most out of them?

An interesting approach to combine classifiers is known as classifier committees (Sun et al., 2004). Classifier committees rely on the predictions of various classifiers, and combine them by means of a decision function, which serves to define the weight and/or relevance of each classifier in the final prediction.

An SVM classifier outputs a margin for each resource over each class in the taxonomy, meaning the reliability to belong to that class. The class with the largest positive margin for each resource is then selected as the classifier's prediction. Thus, combining SVM classifiers' predictions could be done by means of adding up their margins or reliability values for each class. Each resource will then have a new reliability value (i.e., the sum of margins) for each class. Nonetheless, in this case, since each of the three classifiers work with different type of data, the range of the margins they output differ. To solve this, we propose the normalization of the margins based on the maximum margin value outputted by each classifier (*max(m$_i$)*):

$$m'_{ijc} = m_{ijc} / \max(m_i)$$

where m_{ijc} is the margin by the classifier i between the resource j and the hyperplane for the class c, and m'_{ijc} is its value after normalizing it.

The class maximizing this sum will be predicted by the classifier. Then, the sum of margins between the class c and the resource j using a committee with n classifiers could be defined as:

$$S_{jc} = \sum_{i=1}^{n} m_{ijc}$$

If the classifiers are working over *k* classes, then the predicted class for the resource *j* would be defined as follows:

$$C_j^* = \arg\max_{i=1..k} S_{ji}$$

In our study, we performed the combining experiments by using the best approach for tags, comments and content.

Web Page Classification - ODP							
	Training Set Size						
	600	**1400**	**2200**	**3000**	**4000**	**5000**	**6000**
Tags	0.526	0.590	0.616	0.636	0.645	0.654	0.665
Content + Comments	0.554	0.604	0.627	0.643	0.651	0.660	0.670
Content + Tags	0.560	0.623	0.651	0.669	0.681	0.690	0.700
Comments + Tags	0.555	0.621	0.650	0.671	0.681	0.691	0.702
Content + Comments + Tags	**0.572**	**0.634**	**0.661**	**0.679**	**0.690**	**0.700**	**0.709**

Table 7: Accuracy results of classifier committees for the web page classification

Book Classification - DDC							
	Training Set Size						
	3000	**6000**	**9000**	**12000**	**15000**	**18000**	**21000**
Tags	**0.861**	0.864	0.864	0.867	0.869	0.869	0.868
Content + Comments	0.778	0.803	0.814	0.821	0.823	0.827	0.830
Content + Tags	0.823	0.842	0.845	0.849	0.851	0.852	0.852
Comments + Tags	0.857	**0.866**	**0.868**	**0.872**	**0.875**	**0.876**	**0.876**
Content + Comments + Tags	0.824	0.843	0.847	0.852	0.855	0.856	0.856

Table 8: Accuracy results of classifier committees for the book classification (DDC)

Book Classification - LCC							
	Training Set Size						
	3000	**6000**	**9000**	**12000**	**15000**	**18000**	**21000**
Tags	**0.853**	0.857	0.856	0.861	0.861	0.857	0.861
Content + Comments	0.800	0.823	0.834	0.844	0.854	0.865	0.892
Content + Tags	0.808	0.828	0.840	0.848	0.859	0.868	0.896
Comments + Tags	0.848	**0.863**	**0.873**	**0.879**	**0.888**	**0.896**	**0.917**

Book Classification - LCC							
Content + Comments + Tags	0.812	0.833	0.844	0.854	0.864	0.872	0.900

Table 9: Accuracy results of classifier committees for the book classification (LCC)

The results of the experiments using classifier committees are shown in Table 7, Table 8, and Table 9. Note that the table also includes the tag-based classifier's results, enabling the comparison of the results by the classifier committees against the best of the simple classifiers. In most of the cases, when different classifiers are combined, the errors of a classifier can be corrected by the rest, as these results show. It is worth to note that a classifier with the highest accuracy does not have to be the best on committees. The gaps among the margins outputted for the ideal class and the rest are also relevant for a classifier to perform well at committees.

Making different combinations among the classifiers has outperformed the best non-combining approach in all cases for the web page classification task. Either of the committees performs better than using only tags in this case. Among the committees, the best results are always for the one that includes the three kinds of metadata. Merging the outputs of the classifiers based on tags, comments and content resulted the highest performance, outperforming any of the combinations where only two kinds of metadata are considered. Among the double-committees, the performance is higher when tags are considered; this means that tags seem to be the most helpful for committees, and not only as a single classifier.

Regarding the book classification task, the classifier committees require bigger training sets to outperform the best non-combining approach. This does not happen for the smallest training sets, but there is a clear outperformance of the combining approaches in the largest training sets. As opposed to the web page classification, it is slightly better to combine the outputs of only comments and tags, without considering content. It obtained better results rather than the triple-committee. In this case, content does not seem to be as helpful as for web pages when it comes to combining classifiers. The underperformance of content as against to tags in much bigger for books than for web pages, so that it is likely that the content-based approach is not providing good margin values.

Discussion

Our analysis on the use of social annotations for automated classification shows that both social tags and comments are representative enough to perform the task. Nonetheless, other type of social annotations like highlights and ratings do not seem to be useful at present; highlights are not popular enough yet, as we showed that most of the web documents remain unannotated and cannot be represented, whereas ratings do not provide useful information, at least for topical classification. Both using social tags and comments have shown outperforming results against the content-based approach. Among these two types of social annotations, tags show the best results. In this case, relying on a detailed representation considering the full tag activity seems to be the optimal approach.

Moreover, we conclude that none of the three kinds of data is refusable, since all of them may provide positive results when dealing with classifier committees, particularly with web pages. Combining the outputs outperforms the non-combining approaches in most cases. Thus, we conclude that tags are the annotations that best fit the expert-based categorization scheme, as well as the best contributors for classifier committees.

FUTURE RESEARCH DIRECTIONS

It has been shown the high usefulness of social tags when it comes to automated classification, and it also has been shown that considering the weights of tags as users' consensus is also really helpful. Though, all the tags have been considered in the same way, without any semantic or linguistic processing. Grouping together synonymous tags, and detecting the usefulness of each tag, being able to rule out useless tags, should help improve and relax this kind of tasks.

CONCLUSIONS

In this chapter, we have studied and analyzed the application of different annotations from social bookmarking and cataloging sites to the automated classification task. As a Gold Standard, we rely on the top level category schemes of the Open Directory Project for web pages, and the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC) for books. We found social tags and comments, among the existing social annotations, to be representative enough to perform this kind of task. Our experiments show encouraging results for the use of social annotations, outperforming the use of textual content. Finally, we conclude with social tags as the best representation approach, especially by relying on aggregated tagging data for a resource. In our analysis, they have shown their appropriateness for classifying different kind of resources like web pages and books.

REFERENCES

Aliakbary, S., Abolhassani, H., Rahmani, H., & Nobakht, B. (2009). Web page classification using social tags. *IEEE International Conference on Computational Science and Engineering, vol. 4* (pp. 588-593. Vancouver, BC, Canada. IEEE Computer Society 2009.

Crammer, K., & Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research, vol. 2*, 265-292.

Golder, S., & Huberman, B. A. (2006). The structure of collaborative tagging systems. *Journal of Information Science, 32*(2), 198-208.

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can social bookmarking improve web search?. *WSDM '08: Proceedings of the international conference on Web search and web data mining* (pp. 195-206). New York, NY, USA: ACM.

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks, 13*(2), 415-425.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European Conference on Machine Learning* (pp. 137–142). Berlin: Springer.

Noll, M. G., & Meinel, C. (2007). Authors vs. readers: A comparative study of document metadata and content in the WWW. *Proceedings of the 2007 ACM symposium on Document engineering* (pp. 177-186). Winnipeg, Manitoba, Canada: ACM.

Noll, M. G., & Meinel, C. (2008a). Exploring social annotations for web document classification. *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 2315-2320). Fortaleza, Ceara, Brazil: ACM.

Noll, M. G., & Meinel, C. (2008b). The metadata triumvirate: Social annotations, anchor texts and search queries. *Web Intelligence and Intelligent Agent Technology* (pp. 640-647). Sydney, Australia: IEEE CS Press.

Qi, X., & Davison, B.D. (2009). Web page classification: Features and algorithms. *ACM Computing Surveys, 41*(2). (pp. 1-31).

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

Smith, G. (2008). *Tagging: people-powered metadata for the social web*. Berkeley, California, United States: New Riders.

Sun, B.-Y, Huang, D.-S., Guo, L., & Zhao, Z.-Q. (2004). Support vector machine committee for classification. *Advances in Neural Networks – ISNN 2004* (pp. 648-653). Dalian, China: Springer.

Weston, J., & Watkins, C. (1999). Multi-class support vector machines. Proceedings of the 1999 European Symposium on Artificial Neural Networks (pp. 219-224). Bruges, Belgium: D-Facto.

Yeung, C. M. A., Gibbins, N., & Shadbolt, N. (2008). Web search disambiguation by collaborative tagging. *Proceedings of the Workshop on Exploring Semantic Innotations in Information Retrieval at ECIR'08* (pp. 48-61).

Zubiaga, A., Martínez, R., & Fresno, V. (2009a). Getting the most out of social annotations for web page classification. *DocEng'09: Proceedings of the 9th ACM symposium on Document engineering* (pp. 74-83). New York, NY, USA: ACM.

Zubiaga, A., Fresno, V., & Martínez, R. (2009b). Is unlabeled data suitable for multiclass SVM-based web page classification?. *SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing* (pp. 28-36). Morristown, NJ, USA: ACL.

ADDITIONAL READING

Babinec, M. S., & Mercer, H. (2009). *Metadata and open access repositories*. Philadelphia, PA: Taylor & Francis.

Baca, M., & Getty Research Institute. (2008). *Introduction to metadata*. Los Angeles, CA: Getty Research Institute.

Bonino, S. (2009). *Social tagging as a classification and search strategy: A smart way to label and find web resources*. Saarbrücken, Germany: VDM Verlag Dr. Müll.

Caplan, P. (2009). *Metadata fundamentals for all librarians*. New Delhi: Indiana Pub. House.

Foulonneau, M., & Riley, J. (2008). *Metadata for digital resources: Implementation, systems design and interoperability*. Chandos information professional series. Oxford: Chandos.

Gartner, R., L'Hours, H., & Young, G. (2008). *Metadata for digital libraries: State of the art and future directions*. Bristol: JISC.

Golder, S., & Huberman, B. A. (2006). The structure of collaborative tagging systems. *Journal of Information Science, 32*(2), 198-208.

Granitzer, M., Lux, M., & Spaniol, M. (2008). *Multimedia semantics: The role of metadata*. Berlin: Springer.

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can social bookmarking improve web search?. *WSDM '08: Proceedings of the international conference on Web search and web data mining* (pp. 195-206). New York, NY, USA: ACM.

Hider, P. (2009). *Information resource description: Creating and managing metadata*. London: Facet.

Hillmann, D. I., Guenther, R., Hayes, A., Library of Congress., & Association for Library Collections & Technical Services. (2008). *Metadata standards & applications*. Washington, D.C: Library of Congress.

International Conference on Metadata and Semantics Research, Sartori, F., Sicilia, M.-A., & Manouselis, N. (2009). *Metadata and semantic research: Third international conference*, MTSR 2009, Milan, Italy, October 1-2, 2009 : proceedings. Berlin: Springer.

Lanius, L., & Vermont. (2009). *Embracing Metadata: Understanding MARC and Dublin Core [workshop]*. Montpelier, Vt: Vermont Dept. of Libraries.

Mittal, A. C. (2009). *Metadata management*. Delhi, India: Vista International Pub. House.

Noll, M. G., & Meinel, C. (2008a). Exploring social annotations for web document classification. *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 2315-2320). Fortaleza, Ceara, Brazil: ACM.

Peters, I. (2009). *Folksonomies. Indexing and Retrieval in Web 2.0 (Knowledge & Information: Studies in Information Science)*. Berlin, Germany: De Gruyter/Saur.

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

Smith, G. (2008). *Tagging: people-powered metadata for the social web*. Berkeley, California, United States: New Riders.

Taylor, A. G., & Joudrey, D. N. (2009). *The organization of information*. Westport, Conn: Libraries Unlimited.

Turrell, A. (2008). *Augmenting Classifications and Search with Tags to Create Usable Content- and Product-Based Websites*. Baltimore, MD: University of Baltimore.

Wu, X., Zhang, L., & Yu, Y. (2006). Exploring social annotations for the semantic web. *WWW '06: Proceedings of the 15th international conference on World Wide Web* (pp. 417-426). New York, NY, USA: ACM.

Zhou, D., Bian, J., Zheng, S., Zha, H., & Giles, C.L. (2008). Exploring social annotations for information retrieval. *Proceedings of the 17th international conference on World Wide Web* (pp. 715-724). Beijing, China: ACM.

Zubiaga, A., Martínez, R., & Fresno, V. (2009a). Getting the most out of social annotations for web page classification. *DocEng'09: Proceedings of the 9th ACM symposium on Document engineering* (pp. 74-83). New York, NY, USA: ACM.

KEY TERMS & DEFINITIONS

- **Tagging:** Tagging is an open way to assign tags or keywords to resources or items (e.g., web pages, movies or books), in order to describe them. This enables the later retrieval of the resources in an easier way, using tags as resource metadata. As opposed to a classical taxonomy-based categorization system, they are usually non-hierarchical, and the vocabulary is open, so it tends to grow indefinitely. For instance, a user could tag this chapter as social-tagging, research and chapter, whereas another user could use web2.0, social-bookmarking and tagging tags to annotate it.

- **Social tagging:** A tagging system becomes social when its tag annotations are publicly visible, and so profitable for anyone. The fact of a tagging system being social implies that a user could take advantage of tags defined by others to retrieve a resource.

- **Social bookmarking:** Delicious, StumbleUpon and Diigo, amongst others, are known as social bookmarking sites. They provide a social means to save web pages (or other online resources like images or videos) as bookmarks, in order to retrieve them later on. In contrast to saving bookmarks in user's local browser, posting them to social bookmarking sites allows the community to discover others' links and, besides, to access the bookmarks from any computer to the user itself. In these systems, bookmarks represent references to web resources, and do not attach a copy of them, but just a link. Note that social bookmarking sites do not always rely on social tags to organize resources, e.g., Reddit is a social bookmarking approach to add comments on web pages instead of tags. The use of social tags in social bookmarking systems is a common approach, though.

- **Social cataloging:** They are quite similar to social bookmarking sites in that resources are socially shared but, in this case, offline resources like music, books or movies are saved. For instance, LibraryThing allows to save the books you like, Hulu does it for movies and TV series, and Last.fm for music-related resources. As in social bookmarking sites, tags are the most common way to annotate resources in social cataloging sites.
- **Folksonomy:** As a result of a community tagging resources, the collection of tags defined by them creates a tag-based organization, so-called folksonomy. A folksonomy is also known as a community-based taxonomy, where the classification scheme is plain, there are no predefined tags, and therefore users can freely choose new words as tags. A folksonomy is basically known as weighted set of tags, and may refer to a whole collection/site, a resource or a user. A summary of a folksonomy is usually presented in the form of a tag cloud.
- **Personomy:** Personomy is a neologism created from the term folksonomy, and it refers to the weighted set of tags of a single user/person. It summarizes the topics a user tags about.
- **Simple tagging:** users describe their own resources or items, such as photos on Flickr, news on Digg or videos on Youtube, but nobody else tags another user's resources. Usually, the author of the resource is who tags it. This means no more than one user tags an item. In many cases, like in Flickr and Youtube, simple tagging systems include an attachment to the resource, and not just a reference to it.
- **Collaborative tagging:** many users tag the same item, and every person can tag it with their own tags in their own vocabulary. The collection of tags assigned by a single user creates a smaller folksonomy, also known as personomy. As a result, several users tend to post the same item. For instance, CiteULike, LibraryThing and Delicious are based on collaborative tagging, where each resource (papers, books and URLs, respectively) could be tagged (therefore annotated) by all the users who considered it interesting.

[i] http://delicious.com
[ii] http://www.stumbleupon.com
[iii] http://www.librarything.com
[iv] http://www.diigo.com
[v] http://nlp.uned.es/social-tagging/socialodp2k9/
[vi] http://svmlight.joachims.org