Getting the Most Out of Social Annotations for Web Page Classification

Arkaitz Zubiaga
NLP & IR Group at UNED
Dpto. Lenguajes y Sistemas Informáticos
ETSI Informática, UNED
azubiaga@lsi.uned.es

Raquel Martínez
NLP & IR Group at UNED
Dpto. Lenguajes y Sistemas Informáticos
ETSI Informática, UNED
raquel@lsi.uned.es

Víctor Fresno
NLP & IR Group at UNED
Dpto. Lenguajes y Sistemas Informáticos
ETSI Informática, UNED
vfresno@lsi.uned.es

ABSTRACT

User-generated annotations on social bookmarking sites can provide interesting and promising metadata for web document management tasks like web page classification. These user-generated annotations include diverse types of information, such as tags and comments. Nonetheless, each kind of annotation has a different nature and popularity level. In this work, we analyze and evaluate the usefulness of each of these social annotations to classify web pages over a taxonomy like that proposed by the Open Directory Project. We compare them separately to the content-based classification, and also combine the different types of data to augment performance. Our experiments show encouraging results with the use of social annotations for this purpose, and we found that combining these metadata with web page content improves even more the classifier's performance.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Clustering]

General Terms

Experimentation

Keywords

social bookmarking, social annotations, web page classification

1. INTRODUCTION

Social bookmarking is a Web 2.0 based phenomenon that allows users to describe web contents by annotating them with different kind of metadata in a collaborative and aggregated way. Websites such as Delicious[1], StumbleUpon[2] and Diigo[3], among others, allow their users to add information to a web page, collecting hundreds of thousands of annotations per day [5], easing its subsequent retrieval. The motivation to add metadata to shared bookmarks via tags is enhanced by social aspects such as recommendation and collaboration, two of the main features of this new version of the Web [12].

This user-generated data is added in several ways: trying to give a topical description of the document by means of a tag or a set of tags; offering subjective or personal assessments; adding free text to describe web pages or image contents; making personal valuations for web contents, etc. As a result, a global community of volunteer users creates a huge repository of described resources that can be used later, mainly to improve search and navigation across the web documents.

This invites researchers to take advantage of this metadata to improve information management and access. Most of the works found in the literature analyze web pages' textual contents, the traditional metadata headers added by authors for assisting the crawlers and web spiders, or the hyperlink and document structure [3, 16]. Anchor text -the text of and around inlinks- has helped in some IR tasks [2] and text classification [16].

The impact of social annotations has been explored in several contexts related to information retrieval and the web. The research on tagging has increased lately, and several studies have already analyzed the semantic aspects of tagging, as well as why it is so popular and successful in practice [11]. However, the community is still discovering and describing all the features of the information provided by these social annotations. Bao et al. observed that tags provide a multi-faceted summary of web documents [1], and Noll and Meinel [9] suggest that tags are better suited for classification tasks whereas anchor texts are better for augmenting web information retrieval. These authors suggest that tags are qualitatively different than page content. Other works have used tags in some clustering contexts [12]. All these papers addressed a main question: How can tagging data be used to improve web document clustering, classification and information retrieval systems? In this paper, we try to solve this issue with regard to classification tasks.

The amount of web documents is increasing in a very fast way in the last years, what makes more and more complicated its organization, management and retrieval. For this

[1]http://delicious.com
[2]http://www.stumbleupon.com
[3]http://www.diigo.com

reason, web page classification has gained importance as a task to ease and improve information management and access. Web page classification can be defined as the task of labelling and organizing web documents within a set of predefined categories, which is generally a multiclass problem, since it is usual to cover numerous topics or classes. Web page classification approaches differ from text classification approaches in that the former, in addition to the text content, can make use of web-specific features such as HTML mark-up, hyperlinks, visual analysis, etc. As it was stated before, social tagging systems allow to add a new kind of information - user generated annotations - to improve this task. Although social annotations are not always descriptive of the content of the resources, since they sometimes describe opinions and preferences, we believe they could result valuable to the web classification task if we are able to apply them in the right way separately or in combination with other information sources.

In this paper, we demonstrate how user-generated tags from large-scale social bookmarking websites can be used as a complementary data source to page text for improving automatic classification of web pages. We take as a baseline the classification by means of the web pages' text content, and then we explore the classification by means of different types of social annotations both in a separate way and combining them. We do not deal with HTML mark-up, hyperlink, or structural information; we only focus on the use of text content and social annotations. Our experiments show encouraging results for the use of social annotations in web page classification tasks. Moreover, combining data sources outperforms individual results.

The remainder of this paper is organized as follows. Section 2 presents a general overview on social annotations, summarizing the main types and their features. Previous works in the area, and their relation to ours is covered in Section 3. Next, Section 4 describes the process of generating of the dataset with which we carried out the experiments. In Section 5, we give some brief explanations on how Support Vector Machines work for classification tasks. We continue with the details of the set of experimentations we carried out, presenting its settings and analyzing the different representations for the classifiers in Section 6, whereas Section 7 shows how the classifiers can be combined and analyzes the results. Finally, we conclude with our thoughts and future work in Section 8.

2. SOCIAL ANNOTATIONS

Social bookmarking sites allow users to save and annotate their preferred web pages. Tagging in these social media systems has demonstrated to be a very suitable way for users to annotate resources of interest in the Web. These annotations are made in a collaborative way, and so makes possible a big number of metadata to be available for each web page.

These services provide easy-to-use user interfaces to annotate and describe web resources and, furthermore, to allow users to share annotations and freely-defined categories with the community. The best-known social bookmarking sites have reached this condition hosting large amount of online objects, building huge user communities, and by generating vast web traffics. Nowadays, these social bookmarking systems have reached an unimaginable success a short time ago.

Going into further details on these metadata, different kind of user-generated annotations can be defined:

- **Tags:** keywords trying to define and characterize a web page are known as tags.

  Referring to Van der Wal's definition[4], tags are added in a broad way in social bookmarking. Unlike for narrow tagging, where only a user adds tags to the resource (commonly its author), a broad tagging system aggregates the tags added by different users in a collaborative way. Formally, each user u_i can post an item i_j with a set of tags $T_{ij} = \{t_1, ..., t_p\}$, with a variable number p of tags. After k users posted i_j, it is described as a weighted set of tags $T_j = \{w_1 t_1, ..., w_n t_n\}$, where $w_1, ..., w_n \leq k$. Tags are the most common feature and are available in almost all the social bookmarking services.

  E.g., we found the following tags in the top of the weighted tags for flickr.com (out of a total of 61,143 users): *18,442 photos, 15,258 flickr, 12,969 photography, 12,072 photo, 8,449 sharing, 7,130 images.*

- **Notes:** free text describing the content of a web page is known as a 'note'. Users can write a note to describe what do a web page contain, and so they can later remember what they were about.

  In addition to tags, notes are commonly used to annotate web pages in social bookmarking sites. Delicious, for instance, offers the possibility to add notes besides tags.

  E.g., a note for flickr.com: *A photo and video sharing service site. Now I use it to manage my photos and videos and pulish them on my blog.*

- **Highlights:** by means of highlights, users can select the most relevant part of the bookmarked web page.

  With this kind of metadata, users specify the part of the web page they have specially interested in, or the part they have considered relevant. The best-known site using highlights is Diigo.

  E.g.: *the first paragraph of a web page.*

- **Reviews:** a review is a free text valuating a web page.

  Though this kind of annotations can initially look subjective, users tend to mix descriptive texts with opinions. StumbleUpon is a social bookmarking site relying on this kind of metadata.

  E.g.: a review for flickr.com: *flickr is a great place to share photos, learn more about photography and see the pictures of others..*

- **Ratings:** valuations indicating to what extent users like or dislike a web page, commonly by means of punctuations from one to five. As a result, an average value is usually offered. It is also available in StumbleUpon.

  E.g.: *5 stars.*

[4] http://personalinfocloud.com/2005/02/explaining_and_.html

Tags were originally created to make it easier for users to manage their own information, being easier to find and retrieve relevant documents later from these metadata. In most of the social bookmarking systems, there are no constraints on the keywords users can select as tags, and so it is usually free to choose the scope, generality, or the semantics desired to annotate a resource. In this context, users tend to provide tags with diverse levels of specificity.

A reason behind the success of these social bookmarking sites is that tags can be chosen by users arbitrarily without any restriction. However, although tags represent an open vocabulary, there is an evidence suggesting that, for a particular object type, users tend to use the most popular tags into the system [9]. Moreover, potential troubles can be found in several forms. First, a tag is usually generated in a different way than a term on the document content, representing a summary of the document from the perspective of a single user. Thus, some annotated resources in certain domains might be more trustworthy than others; not all the users have the same confidence in all the domains.

Much of the work in this context has been focused on the study of dataset properties [12], the analysis of usage patterns of tagging systems [4], and the discovery of hidden semantics in tags [17]. Incorporating social annotations with document content and other sources of information is a natural idea [18], specially if we are trying to improve information management tasks. Usually the number of tags is much smaller than the number of terms in the content of a document being tagged.

Golder and Huberman found that after about 100 users tagged a specific web page, the distribution of its top tags for it tends to converge [4]. This evidence suggests that objects tagged with similar tags across users could contain similar contents.

Finally, most of the annotations described above in this section seem to be really interesting for topical web page classification tasks. Nonetheless, it is obvious that 'Ratings' cannot contribute to this kind of web page classification, since they do not offer topical information. For this reason, in this work we based on all the social annotations but ratings. Besides, we consider three families grouping the remaining annotations: 'Tags', 'Notes & Reviews' (grouped as 'Comments'), and 'Content'. 'Highlights' are not considered in this classification due to its low representativity over the web pages, as we point out later. On the other hand, note that we do not handle spam and misbehavior detection in this work.

3. RELATED WORK

There are some works in the literature analyzing the usefulness of social annotations for web document management tasks. As far as information retrieval is concerned, some approaches have been proposed to improve web search. For instance, in [1], [6] and [15], authors studied the inclusion of data from social bookmarking systems to modify web search. On the other hand, recently, in [5], a series of experiments on the social bookmarking site Delicious were carried out to determine whether user generated metadata can be used to augment web search engines.

Bao et al. [1] propose two novel algorithms to incorporate social bookmarking data into page ranking. One of them, SocialSimRank, calculates the estimated similarity between social annotations of the web pages and web queries, whereas the other, SocialPageRank, captures the popularity of web pages by means of the amount of annotations assigned to a page. The algorithms were evaluated using two query sets within a dataset created from Delicious. The experiments showed that the former algorithm can successfully find the latent semantic relations among annotations, and the latter can provide the static ranking from the web annotator's perspective. The authors concluded that both algorithms improved the quality of search results significantly.

In [6], the authors present a formal model for folksonomies and the FolkRank algorithm, which take into account the structure of folksonomies. With this ranking scheme they generate personalized rankings of the folksonomy items and recommend users, tags and resources. In addition to structuring the search results, the proposed algorithm is also applied to find communities within the folksonomy. Yanbe et al. [15] propose an approach to enhance search on the web combining the link-based ranking metric, with other metric derived from social bookmarking data by means of the number of users voting for a page. The authors also include user-created annotations and general statistics of user behaviour towards pages in their prototype. As a result, the prototype allows users to search for web pages through complex queries: by their content, associated metadata (tags), temporal aspects in social bookmarking systems (timestamps associated with bookmarks), user sentiment (sentiment tags) and other features, such as the use of comments in bookmarks. To test the prototype, they used Hatena Bookmark[5], a Japanese social bookmarking service. They conclude that page quality measure can be improved by incorporating the popularity statistics of pages in social bookmarking systems, as well as a hybrid and enhanced web search is possible and can provide advantages.

Heymann et al. [5] suggest that URLs annotated by social bookmarking systems are not numerous enough to impact the crawl ordering of a major search engine, and the tags produced are unlikely to be much more effective than a full text search emphasizing page titles. However, the authors also consider that if social bookmarking continues to grow as in the past several years, it could reach the scale of the current web. They also propose user interface features as a way of improving the quality of tags for search (recommending quality tags), or to have domain-specific sites where tags can result more specific.

Social annotations have also been used for web page organization tasks. In [12] the inclusion of tagging data improved the performance of two clustering algorithms when compared to clustering based on page content on its own. The authors first provide tagging data to the well known K-means clustering algorithm obtaining better results. Next, they proposed a new algorithm, Multi-Multinomial LDA, an extension of the Latent Dirichlet Allocation algorithm, which explicitly models text and tags, improving the results of the former. They also found that tagging data was even

[5]http://b.hatena.ne.jp

more effective for more specific collections than for a collection of general documents.

In [10] a very interesting study was carried out analyzing 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 Open Directory Project. They analyzed at which hierarchy depth matches occurred, concluding that tags may perform better for broad categorization of documents rather than for narrow categorization. The study also point 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 also observed that tag noise (the opposite of popular tags) provides helpful data for information retrieval and classification tasks in general. In a previous work [9], the same authors suggest that tags provide additional information about a web page, which is not directly contained within its content.

In [11], also Noll and Meinel studied three types of metadata about web documents: social annotations (tags), anchor texts of incoming hyperlinks, and search queries, provided by readers of web documents, authors of web documents, and users trying to find web documents, respectively. They conclude that tags are better suited for classification purposes than anchor text or search keywords, so that tags seem to be particularly helpful for identifying the "aboutness" of web documents.

Different from the above works, we carry out a web page classification task by using different types of social annotations and web document content. We use these types of data both separately and in combination to better analyze the impact of each source of information on a classification task.

4. SOCIAL-ODP-2K9 DATASET

In order to carry out our classification experiments, we aimed to retrieve and collect a set of popular annotated web pages. Being a page annotated at least 100 times, ensures a web page to be socially popular, and to have a converged tag set [4], so we can find these tag sets with fixed proportions of their tag weights. We generated our dataset by using a list of URLs taken from a recent feed on Delicious, limiting to those pages with at least 100 users annotating them.

We monitored the recent feed on Delicious during three weeks in December 2008 and January 2009. In this way, we got a list of 87,096 unique URLs, all of them with at least 100 users annotating them on Delicious.

On the other hand, we decided to use the Open Directory Project[6] (ODP) as the gold standard for web page classification. We found 12,616 URLs in our list matching the URL list in the ODP. For the classification scheme, we rely on the 17 categories in the first level of the ODP taxonomy. It is worth to note that there are a few web pages classified in more than one category. In these cases, we chose one of them randomly, to avoid a web page falling in more than one category.

Analyzing the nature of our dataset, we found that the distribution of the documents among the categories is not homogeneous. This means there are more documents for some classes than for others. The statistics for the distribution of documents in the collection is shown in Figure 1.

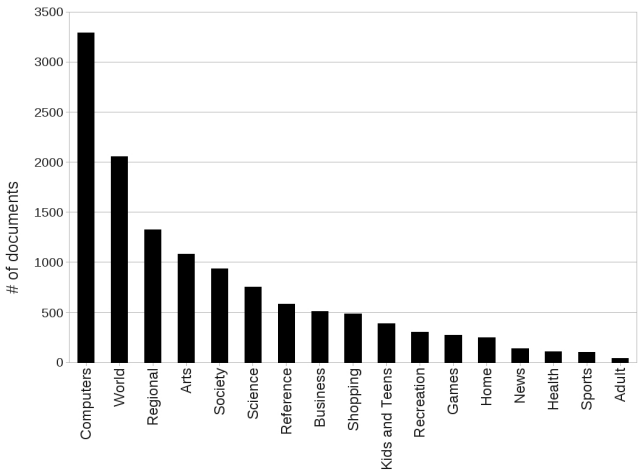

Figure 1: Document count for each class, in a descendant ordering

Once we had this list of classified URLs, we retrieved the following information for each URL from the considered social bookmarking sites:

- **Number of users bookmarked it on Delicious:** a count of the users who had bookmarked the URL on Delicious. This does not mean the number of users who added tags or notes to the URL, as it is not mandatory to do it.
- **Top 10 tag list on Delicious:** as a summary on the most popular tags for a URL, Delicious provides the top 10 weighted list of tags. This includes the most annotated tags and their number of occurrences.
- **Notes on Delicious:** a list of the notes added by users to a URL can also be retrieved from Delicious. Though it is limited to the latest 2,000 notes, we consider it is enough for our purposes.
- **Full tag activity on Delicious:** the site provides a list of the user activity for a URL, including the tags they have added to it. This list is also limited to the latest 2,000 users.
- **Reviews on StumbleUpon:** we gathered the whole review information provided by StumbleUpon. It is worth to note that 2,697 URLs in our dataset had no review information. This suggests reviews on its own to be unable to perform a good classification task, but could be an interesting alternative to combine with other information.

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

- **Highlights on Diigo:** we got the full highlight information in Diigo for each URL in our dataset. However, only 1,920 of the documents in our dataset provide highlight information. The high number of web pages without highlight information makes it to be insufficient. For this reason, we decided not to use highlight information in our study.

Summarizing, our dataset is composed by 12,616 unique URLs with their corresponding ODP first-level category and a set of social annotations including tags and notes from Delicious, and reviews from StumbleUpon.

The dataset is available for research purposes[7].

5. SUPPORT VECTOR MACHINES

In this work, we use Support Vector Machines (SVM) [7] to perform web page classification tasks. This technique uses the vector space model for the documents' representation, and assumes that documents in the same class should fall into separable spaces. Upon this, it looks for a hyperplane that separates the classes; therefore, this hyperplane should maximize the distance between it and the nearest documents, what is called the margin. The optimization function for SVM classification results:

$$\min \left[\frac{1}{2} ||w||^2 + C \sum_{i=1}^{l} \xi_i^d \right]$$

$$\text{Subject to:} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

where C is the penalty parameter, ξ_i is a stack variable for the i^{th} document, and l is the number of labeled documents.

Though the traditional SVM approach performs binary classification tasks, multiclass classification approaches have also been proposed and used in the literature [14]. A multiclass SVM classifier for k classes defines that many hyperplanes in the training phase, so that each of them separates the documents in a class from the rest. In the test phase, when making predictions for each new document, the classifier is able to establish a margin over each of the hyperplanes. 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.

6. FEEDING WEB PAGE CLASSIFIERS WITH SOCIAL ANNOTATIONS

In order to explore the suitability of social annotations for web page classification and to discover the optimal way to exploit them, we carried out several experiments with different document representations. Among these experiments, the form in which the vectorial representation of the web pages is created changes in order to consider the different data sources, though the general settings for the classifier remain unchanged.

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

To perform the SVM-based multiclass classification, we have used SVMmulticlass [8], a software created for research purposes. All the experiments have been run with the default parameters recommended by the author, using a polynomial kernel. Although these settings could be optimized, it is not the aim of this work, since we want to evaluate the relative performance among the different representations. The accuracy metric, defining the percent of correctly predicted documents, has been used for evaluation.

For each of the representations we worked with, we did various random selections of training and test sets of documents (web pages). On the one hand, different sizes had been set for the training sets, ranging from 200 to 3000 documents. This allows to evaluate how does a representation evolve when the amount of training data increases. On the other hand, 6 different executions were done for each training set size. The variations among these executions were the documents selected for the training set. Shifting the documents selected for the training set allows to extract more real results. When plotting the results into the graphs, we relied on the average accuracy for the 6 executions, and we drew a line showing the results for the different training sets' sizes.

Next, we analyze and evaluate different proposals for representations based on tags and comments, separately. Finally, we compare the best approach of each type of metadata with the content-based classification, which is the baseline.

6.1 Classifying with Tags

Previous works suggest that tags could be used to classify web content but, which is the best way to exploit this metadata? Resorting to the data available in our dataset, the following three refer to tagging: the number of users bookmarking a URL, the top 10 weighted list of tags and the full tag activity (FTA). Based on this, several approaches could be proposed, since all this information could be treated in different ways. Next, we evaluate and compare the following approaches for tag-based representation:

- **Unweighted Tags:** the only feature considered for this representation is the occurrence or non-occurrence of a tag in the top 10 list of a web page. This approach ignores tags' weight, and assigns a binary value to each feature in the vector. This means all the values of each vector will be 0, but 10 of them will have a 1 as value.
- **Ranked Tags:** tags corresponding to the top 10 list of a web page 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:** the number of users who bookmarked a web page is taken into account in this approach besides the top 10 list of tags. Using these two values, it is possible to establish the fraction of users assigned each tag in the top 10. A tag would have been annotated by the 100% of the users when its weight matches the user count for a web page, getting the value of 1 as the fraction. According to this, a value from 0 to 1 is set to each tag in the top 10, corresponding to the

fraction of users annotated them. The weights of the tags are somehow considered, though they are normalized and so the order of magnitude for each web page is lost, and only a relative value is considered.

- **Weighted Tags (Top 10):** the weight for each of the tags in the top 10 list for a web page is considered as it is. Now, by definition, the weights of the tags are fully respected, but the amount of users bookmarking a web page is ignored. Note that different orders of magnitudes are mixed up now, since the count of bookmarking users range from 100 in the least-popular web pages to ∼61K in most-popular web page.

  Note that the 4 approaches above generate sparse vectors with 12,116 dimensions, where only 10 of them have a value different from 0 for each web page.

- **Weighted Tags (FTA):** like the weighted tags approach above, the weight of the tags is used for this approach. Nonetheless, the full tag activity is used to set tag weights, not only the top 10. This way, a wider list of tags can be extracted, and a further representation obtained. To reduce vectors' dimensionality, relaxing the computational cost and maintaining the representativity, we removed all the tags appearing only in one web page. This resulted 175,728 dimensions.

  However, the full tag activity is limited by the system to the 2K last users, whereas the top 10 list provides data linked to all the users, even when more than 2K bookmarked it. Note that, in our dataset, 957 web pages were saved by more than 2K users, with an average user count of 5,329.

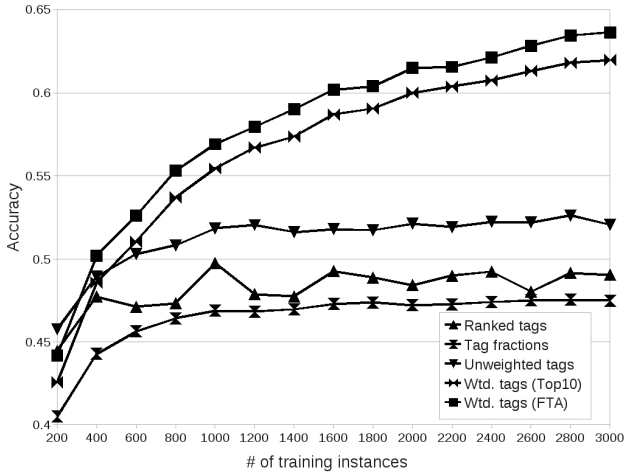

Figure 2: Results on tag-based classification

The results showing the accuracy for each of the approaches for web page classification using only tags are shown in Figure 2. It stands out the superiority of the approaches considering tag weights, as the number of training instances rises. On the other hand, the representations relying on rank positions, tag fractions and unweighted values are very far from the best performance. It seems clear that considering tag weights is a promising approach to improve classification rates, since it seems to provide useful information. Using a training set with only 200 documents, the unweighted approach improves the performance of the weighted representation. Moreover, when the training set increases, the accuracy of the weighted approaches continues growing, whereas it remains stable for the other ones, as an asymptotic behaviour.

Our conjecture is that learning with only 200 documents is much less representative for the weighted approaches, since their values are much more sparse than for the other approaches. This makes the learning phase for the weighted approaches to be less representative than for the rest, since the latter only use values from 0 to 1, having a higher representation of the possible values with only 200 values. Anyway, increasing the number of documents for the training set reverses this behaviour. Moreover, it is noteworthy that both weighted approaches perform better when the training set increases, whereas the other approaches reach its best performance with a training set of 1,000 documents, approximately.

Comparing the two approaches using weights, the more tags are considered in the representation, the higher performance gets the classifier. The approach using the full tag activity always outperforms the one relying on the top 10 tag list. The additional information used by the former improves the accuracy, showing a constant performance gap over the latter when the training set size changes.

We conclude that tag weights provide useful information for topical classification, much better when it is used without considering the full user count. Hence, it is more important to rely on how people has annotated a web page, but not so much on how many users did it.

6.2 Classifying with Comments

As far as comments are concerned, we can use two kinds of metadata stored in our dataset: notes and reviews. Both of them are free texts describing, defining or somehow referring to a web page. As we stated above, there is a significant number of web pages without any review, and so reviews would not be able to classify web pages on their own with enough accuracy. However, this information could be useful to combine with notes, offering additional information. In this way, we have tested two approaches to evaluate the use of comments for web page classification:

- **Only notes:** each web page is represented by the notes users annotated to it. First, all the notes are merged as if they were only one. Once a merged note has been created for each web page, their vectorial representation is obtained. To achieve this, we based on the Term Frequency-Inverse Document Frequency (TF-IDF) function, and we removed the least-frequent terms by their document frequency (df). As a result, we got vectors with 83,959 dimensions, where each dimension refers to a term, with a tfidf-weighted value.

- **Merging notes and reviews:** reviews are also taken into account for this approach. As well as all the notes for a web page are merged in one for the other approach, we made the same with notes and reviews. In

the same way, TF-IDF was applied to obtain the vectorial representation. The resulting vectors are composed by 95,149 dimensions.

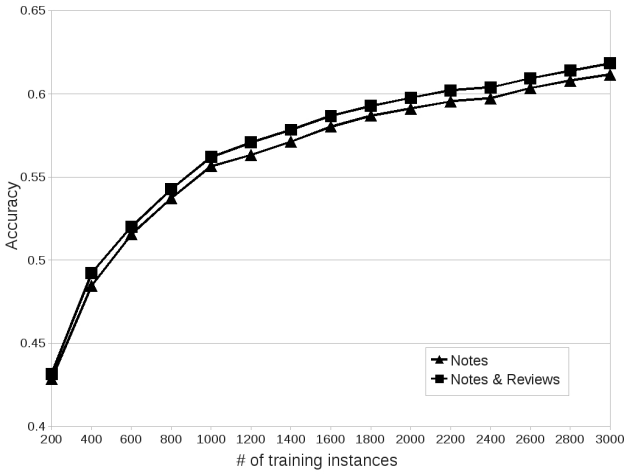

Figure 3: Results on comment-based classification

Figure 3 shows the results for these two comment-based approaches. Although both of them offer similar accuracy, the graph shows that considering reviews besides notes results slightly beneficial. Though reviews are allegedly subjective and could initially look harmful for topical web page classification, our results show that they can be useful to combine with objective notes and improve performance.

6.3 Comparison of Approaches: Content vs Annotations

Once we had tested web page classification using tags and comments, and we obtained which is the best of our approaches for each kind of metadata, we aimed to compare them to our baseline, the content-based web page classification. The main goal of this comparison is to discover which kind of metadata performs better, and whether using social annotations to classify web pages outperforms content-based classical methods. This comparison involves:

- **Content:** we consider as the content of a web page the plain text extracted from it after removing HTML tags. The content of the web pages is then represented into vectors by means of the TF-IDF algorithm. This leads to vectors with 156,028 dimensions.
- **Comments:** we directly base on the winning approach for comments, combining notes and reviews.
- **Tags:** the winning approach is also used for tags, using the tag weights for the full tag activity.

Figure 4 shows the results for this comparison of approaches. Our results show that both social annotations improve our baseline, the content-based web page classification. Again, an exception occurs when only 200 documents are considered

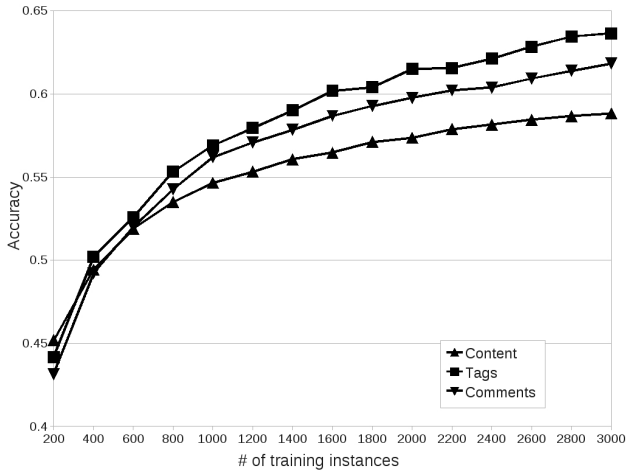

Figure 4: Results on tag, comments and content-based classification

in the training set, where the content-based representation outperforms the rest. However, when the size of the training set increases the content-based approach is left behind.

Comparing the behavior of the social annotations, the results show higher performance for the approach using tag information. Analyzing the case with a training set with 3,000 documents, the accuracy using tags is 8% higher than the one by using content, and 3% higher than the one by comments.

Going into further details, an analysis on the misclassified documents can also be done. Note that since an SVM classifier outputs a set of margins for each document, that is, a margin for each class, a list of prediction ranks can be inferred. For our experiments over 17 categories, the best rank for the right category of a misclassified document is second, whereas the worst is 17th. Figure 5 shows the number of documents for each rank, corresponding to the training set with 3,000 documents. Note that documents ranked first correspond to correct guesses, that is, correctly classified documents.

In the same way, Figure 6 shows the average rank for all the predictions by each classifier. These graphs show, again, the superiority of tags with respect to comments and content. Moreover, since most of the misclassifications are well-ranked (second or third), and only a few of them fall into the tail, the classifiers show great ability to predict documents' class.

7. COMBINING CLASSIFIERS TO IMPROVE PERFORMANCE

Though the tag-based representation outperforms the other two approaches, all of them offer encouraging results and look good enough so that we can try to combine them and improve even more the classifier's performance. An interesting approach to combine classifiers is known as classifier committees [13]. Classifier committees rely on the predictions of various classifiers, and combine them by means of a

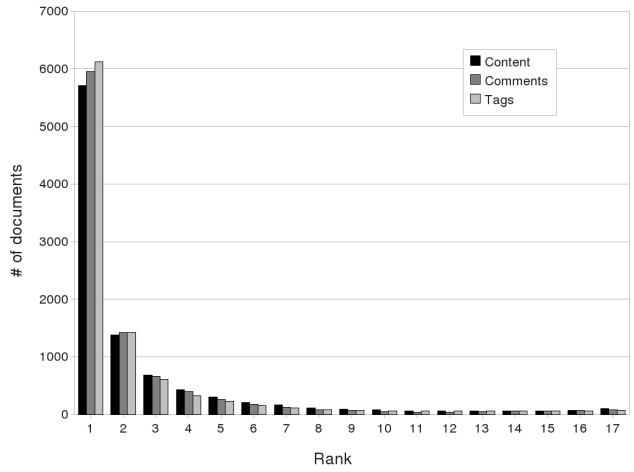

Figure 5: Number of documents per prediction rank with 3,000 training documents

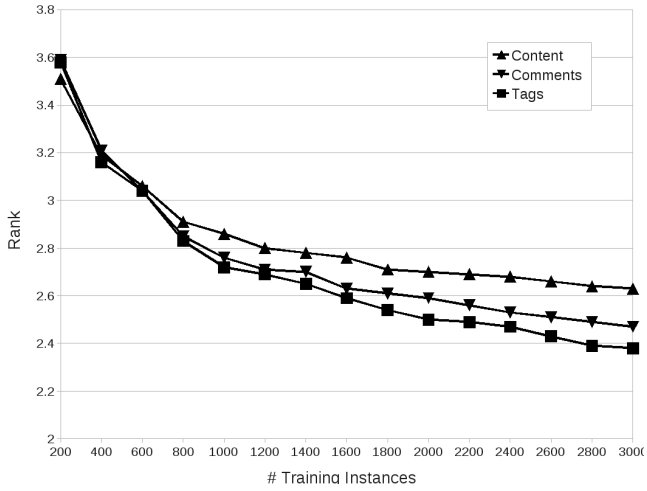

Figure 6: Average prediction rank for classifiers' outputs

decision function, which serves to define the weight and/or relevance of each classifier.

Combining SVM classifiers' predictions is commonly done by means of adding up their margins (or their reliabilities) for each class. Each document will then have a new sum for each class. The class maximizing this sum will be predicted by the classifier. Then, the sum of margins among the class c and the document j using a committee with n classifiers could be defined as:

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

where m_{ijc} is the margin by the classifier i between the document j and the hyperplane for the class c.

If the classifiers are working over k classes, then the predicted class for the document j would be:

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

In our study, we performed the combining experiments by using the best approaches for tags, comments and content. The combinations we have tested in our study include all the possibilities: tags + content, tags + comments, comments + content, and tags + comments + content.

The results for the experiments using classifier committees are shown in Figure 7. Note that the graph also includes the line for the tag-based representation, which serves to compare the results by the classifier committees against the best of the simple classifiers.

When different classifiers are combined, the errors of a classifiers can be corrected by the rest, as these results show. It

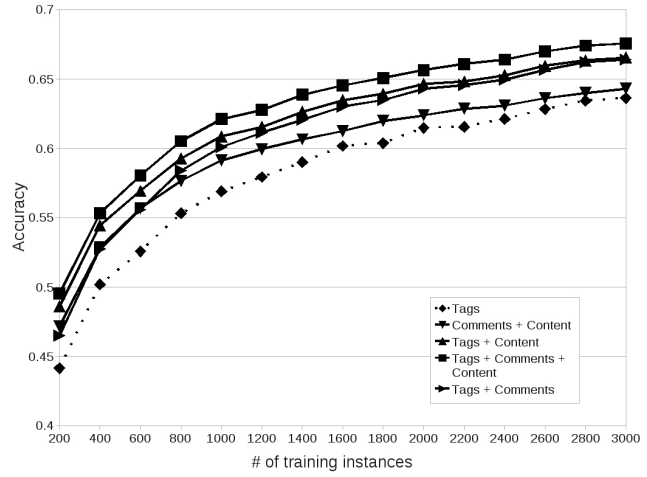

Figure 7: Results on combining classifiers

can be seen that making different combinations among the classifiers has always outperformed the best non-combining approach.

Among the combinations, the best results are for that including the three kinds of metadata. Merging the outputs of the classifiers based on tags, comments and content has offered the highest performance, outperforming any of the combinations where only two kinds of metadata are considered. On average, the accuracy by the triple-combination is 2% higher than the accuracy by the best double-combination. For those combinations including only two kinds of metadata, the performance is higher when the information on tags is considered, showing again that it is the best metadata for this task.

Merging tags and content outperforms combining tags and comments, whereas the latter performs better than the com-

bination of content and comments. This infers that tags perform best for combinations, followed by content and comments, in this order.

We also analyzed how each simple classifier contributes to the final performance. Figure 8 shows the broke down accuracy. It shows that the main difference on the final accuracy for each training set size is determined, mainly, by the number of documents correctly predicted by all the classifiers. While the number of correctly guessed predictions remains stable for the cases when 2 or less classifiers have predicted right, the correct guesses for the documents correctly predicted by the 3 classifiers goes up. Note that the final prediction will always be right when all the classifiers on its own have correctly guessed, what occurs more frequently as long as the training set increases.

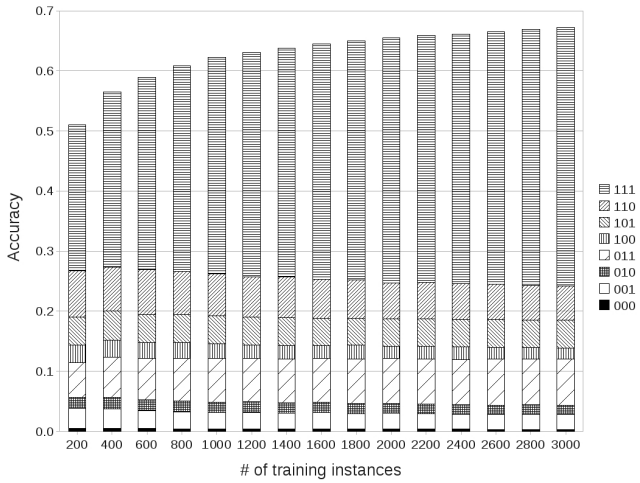

Figure 8: Contribution of each classifier to the triple-combination (The legend shows three binary digits corresponding to content, comment and tag-based classifiers, respectively; a 1 means a correct guess, whereas a 0 represents a misclassification)

Finally, breaking down the accuracy of the system, and retrieving the particular accuracy of each of the 17 categories (shown in Figure 9), the accuracy value does not correspond to the number of documents in the class. It looks much easier to correctly guess documents for *Computers*, *World*, *Arts*, *Shopping*, *Games*, and *Sports* categories, than for *Business*, *Kids and Teens*, *Recreation*, *News* and *Adult* categories.

8. CONCLUSIONS AND OUTLOOK

In this work, we have analyzed and evaluated the use of social annotations for web page classification over the Open Directory Project. Although social bookmarking sites provide several types of annotations, some of them are not popular enough to consider them at present. We have found tags and comments to have enough spreading for its use in web page classification tasks. Classifying with tags and comments has shown encouraging results, improving our content-based baseline. The more documents are used in the training set, the bigger becomes this superiority for tags and comments over page content. Moreover, we have shown that combining both social annotations and page content improves even more the performance of the classifier.

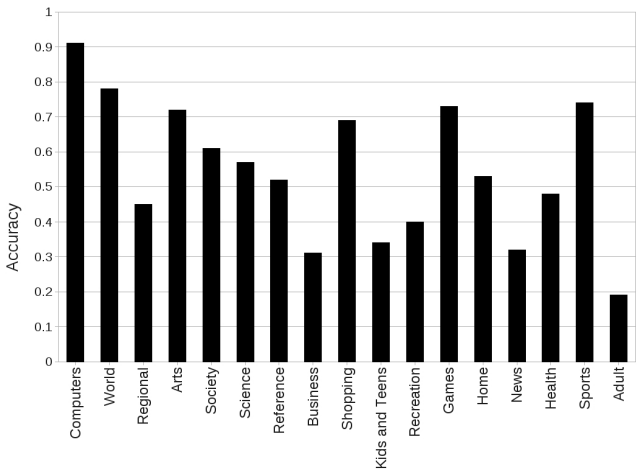

Figure 9: Accuracy for each class (in a descendant ordering by document count, as in Figure 1)

Our experiments corroborate the suggestions and conclusions of Noll and Meinel [9, 10] with regard to the usefulness of tag data to improve web document classification. Particularly, when performing a broad categorization task, like that represented by the first level of the ODP hierarchy.

Future work will include evaluating de use of social annotations with more specific categorization. The use of highlights as an input for the web page classifier remains unstudied, due to its insufficient popularity at present. On the other hand, a deeper study could be done on the tag-based representation, e.g., filtering subjective tags, resolving tag synonymy and polysemy, etc. as well as on the comment-based representation, e.g., filtering noisy comments. This study could probably throw light on the difference on accuracy rates of the classifier over the different first-level ODP categories.

9. ACKNOWLEDGMENTS

We wish to thank the anonymous reviewers for their helpful and instructive comments. This work has been partially supported by the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267), the Regional Ministry of Education of the Community of Madrid and by the Spanish Ministry of Science and Innovation project QEAVis-Catiex (TIN2007-67581-C02-01).

10. REFERENCES

[1] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proceedings of the 16th international conference on World Wide Web*, pages 501–510, Banff, Alberta, Canada, 2007. ACM.

[2] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and*

development in informaion retrieval, pages 459–460, Toronto, Canada, 2003. ACM.

[3] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using web structure for classifying and describing web pages. In *Proceedings of the 11th international conference on World Wide Web*, pages 562–569, Honolulu, Hawaii, USA, 2002. ACM.

[4] S. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science, 32(2)*, pages 198–208, 2006.

[5] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the international conference on Web search and web data mining*, pages 195–206, Palo Alto, California, USA, 2008. ACM.

[6] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *The Semantic Web: Research and Applications*, pages 411–426, 2006.

[7] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*, pages 137–142, Berlin, 1998. Springer.

[8] T. Joachims. Making large-scale support vector machine learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in kernel methods: support vector learning*, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[9] M. G. Noll and C. Meinel. Authors vs. readers: A comparative study of document metadata and content in the www. In *Proceedings of the 2007 ACM symposium on Document engineering*, pages 177–186, Winnipeg, Manitoba, Canada, 2007. ACM.

[10] M. G. Noll and C. Meinel. Exploring social annotations for web document classification. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 2315–2320, Fortaleza, Ceara, Brazil, 2008. ACM.

[11] M. G. Noll and C. Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, volume 1, pages 640–647, 2008.

[12] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 54–63, Barcelona, Spain, 2009. ACM.

[13] B.-Y. Sun, D.-S. Huang, L. Guo, and Z.-Q. Zhao. Support vector machine committee for classification. In F. Yin, J. Wang, and C. Guo, editors, *Advances in Neural Networks - ISNN 2004*, volume 3173 of *Lecture Notes in Computer Science*, pages 648–653. Springer, 2004.

[14] J. Weston and C. Watkins. Multi-class support vector machines. In *Proceedings of the 1999 European Symposium on Artificial Neural Networks*, 1999.

[15] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 107–116, Vancouver, BC, Canada, 2007. ACM.

[16] Y. Yang, S. Slattery, and R. Ghani. A Study of Approaches to Hypertext Categorization. *J. Intell. Inf. Syst.*, 18(2-3):219–241, 2002.

[17] C. M. A. Yeung, N. Gibbins, and N. Shadbolt. Web Search Disambiguation by Collaborative Tagging. In *Proceedings of the Workshop on Exploring Semantic Innotations in Information Retrieval at ECIR'08*, pages 48–61, Mar. 2008.

[18] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. Exploring Social Annotations for Information Retrieval. In *Proceedings of the 17th international conference on World Wide Web*, pages 715–724, Beijing, China, 2008. ACM.