Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 189-196 recibido 03-05-2011 aceptado 23-05-2011

Augmenting Web Page Classifiers with Social Annotations

Enriqueciendo Clasificadores de Páginas Web con Anotaciones Sociales

Arkaitz Zubiaga, Raquel Martínez, Víctor Fresno
NLP & IR Group @ UNED
Madrid, Spain
{azubiaga,raquel,vfresno}@lsi.uned.es

Resumen: La falta de contenido textual representativo en muchas páginas web sugiere el estudio de metadatos adicionales para mejorar tareas de clasificación de páginas web. Los sitios de marcadores sociales proveen un medio accesible para aumentar en gran medida los metadatos disponibles con anotaciones dadas por usuarios. Aún no se ha explorado a fondo en este campo. En este trabajo, analizamos la utilidad de las anotaciones sociales para clasificación de páginas web. Evaluamos los resultados sobre dos niveles de categorización, así como su utilidad para páginas de entrada y profundas. Concluimos que las anotaciones sociales pueden mejorar los clasificadores de páginas web en múltiples casos, y presentamos un método para sacar el máximo partido mediante la combinación de clasificadores.
Palabras clave: etiquetado social, clasificación, folksonomías

Abstract: The lack of representative textual content in many web documents suggests the study of additional metadata to improve web page classification tasks. Social bookmarking sites provide an accessible way to increase available metadata in large amounts with user-provided annotations. This field remains relatively unexplored. In this work, we analyze the usefulness of social annotations for web page classification. We evaluate the results on two different categorization levels, and analyze their suitability for home and deeper pages. We conclude that social annotations could enhance web page classifiers in multiple cases, and we present a method to get the most out of them using classifier committees.
Keywords: social-tagging, classification, folksonomies

1 Introduction

Web page classification is the task of labeling web documents with their corresponding categories from a predefined taxonomy. To perform it automatically, document content is commonly used to represent web pages. However, the lack of representative content makes it insufficient in many cases (Qi and Davison, 2009). In this context, social bookmarking sites present a means 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.com, StumbleUpon.com and Diigo.com, among others, allow their users to bookmark web pages, collecting hundreds of thousands of annotations per day (Heymann, Koutrika, and Garcia-Molina, 2008). As a result, a global community of volunteer users creates a huge repository of annotated resources that can ease future retrieval.

Until now, most of the works in the field have shown the suitability of social tags for this kind of task. Nonetheless, the study of the optimal representation based on social tags, and the use of other social annotations, are still unexplored.

In this work, we analyze and study the use of metadata extracted from social bookmarking sites to classify a set of annotated web pages. The pages are labeled according to the taxonomy of the Open Directory Project[1] (ODP). We evaluate the results relying on the first and second levels of the categorization scheme, analyzing the reliability of social annotations in each case. We find two types of

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

ISSN 1135-5948

social annotations, tags and comments provided by end users, to be applicable and useful for web page classification, both in shallower and deeper levels of the categorization scheme. We also analyze their suitability for home and deep pages. We conclude proposing a way to represent each kind of annotation, and we present a method to outperform their results by combining different data using classifier committees.

Next, in Section 2, we describe the nature and types of social annotations. We present the related work in Section 3. After that, we detail the settings of the experiments in Section 4. We present and analyze the results in Section 5, and conclude the paper in Section 6.

2 Social Annotations

Social bookmarking sites allow users to save and annotate their favorite web pages, sharing annotations 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 web page. Going into further details on these metadata, different kinds of user-generated annotations can be defined: (1) **Tags** are keywords defining a web page. 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 $T_j = \{w_1t_1, ..., w_nt_n\}$, where $w_1, ..., w_n \leq k$. (2) **Notes** are free texts describing the content of a web page. By means of (3) **highlights**, users can select the most relevant part of a bookmarked web page. (4) **Reviews** are free texts valuating a web page. Even though this kind of annotations can initially look subjective, users tend to mix descriptive texts with opinions. (5) **Ratings** are valuations indicating the extent to which users like or dislike a web page, commonly by means of punctuations from one to five.

The use of tags was originally suggested to make easier future search and retrieval of relevant documents. In most social bookmarking systems, there are no constraints on the keywords that users can select as tags.

Among the annotations described above, it is obvious that *ratings* cannot contribute to topical web page classification, since they have nothing to do with categories. For this reason, we based our study on all the social annotations but *ratings*. Thus, we consider three families grouping the remaining annotations: *tags*, *notes & reviews* (grouped as *comments*), and *self-content & anchor texts* (grouped as *content*). In our experiments, *highlights* were not considered due to their low representativity over the web pages, as we point out later.

3 Related Work

Most of the research on social tagging systems focus on the study of dataset properties (Ramage et al., 2009), the analysis of usage patterns of tagging systems (Golder and Huberman, 2006), and the discovery of hidden semantics in tags (Yeung, Gibbins, and Shadbolt, 2008). Incorporating social annotations with document content and other sources of information is a natural idea (Zhou et al., 2008).

Little work has been done analyzing the usefulness of social tags for web page organization tasks. 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 and 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 the level of hierarchy in which 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. In (Noll and Meinel, 2008b), the same authors studied three types of metadata about web documents: tags, anchor texts of incoming hyperlinks, and search queries employed by users to access them. They conclude that tags are better suited for classification purposes than anchor texts or search keywords.

Beyond the above works, the study of different tag representations, and the use of so-

cial annotations other than tags requires further analysis.

In a previous work (Zubiaga, Martínez, and Fresno, 2009), 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. Social tags and comments showed high performance against text content. In present paper, we further analyze the application of social annotations to this task, getting in more depth both on more specific categorization over narrower categories, and on their suitability to home and deep pages.

4 Experiment Settings

4.1 The Social-ODP-2k9 Dataset

We use a collection of 12,616 web pages, made up by a list of popular URLs retrieved from the recent feed of Delicious during December 2008 and January 2009[2].

In addition to fetching the page content and the corresponding categorization for these pages, we gathered the following data from social bookmarking sites:

Bookmark data from Delicious: this includes, for each web page, the number of users bookmarking it, the top 10 tags, notes, and the Full Tagging Activity (FTA). The latter 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.

Reviews from StumbleUpon: 9,919 URLs in our dataset have review information.

Highlights from Diigo: only 1,920 of the documents in our dataset provide highlight information, so that we decided not to use this information in our study.

Finally, we fetched a set of anchor texts for each web page. We requested Google for up to 300 pages linking to each web page, and extracted the corresponding anchor texts, i.e., the texts within the links pointing to each page.

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

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

4.2 ODP Hierarchy

We rely on the hierarchical structure of the well-known ODP as the categorization scheme. Particularly, we experiment the classification by using both the top and second levels of the hierarchy. The top level is made up by 17 categories, whereas the documents belong to 390 categories in the second level. Nonetheless, the low representation for some of the categories made us reduce the taxonomy and the document set. We removed the categories with fewer than 5 documents, as well as the underlying documents, for the second level experiments. This turned into a taxonomy with 243 categories for 12,286 documents.

4.3 Support Vector Machines

We use multiclass Support Vector Machines (SVM) (Weston and Watkins, 1999) to perform web page classification tasks. We use the freely available "svm-multiclass"[3]. We set it up to the linear kernel and the default parameters.

To evaluate the performance, we randomly perform 6 different selections for each of the training sets ranging from 600 to 6,000 documents. 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 within the whole test set.

5 Classifying with Social Annotations

In our experiments, we first treat each group of data separately: *content*, *comments* and *tags*. This way, we study the performance of each data. Then, we combine them by means of classifier committees.

5.1 Content-based Classification

For the content-based baseline, we rely on the text content of web pages. We also consider the use of incoming anchor texts as a part of the document content. Note that, in this baseline, we do not deal with other web-specific characteristics like HTML structure.

In order to evaluate the classification by means of data from content and anchor texts, we propose the following two approaches: using only text content, and merging both content and anchor texts. For the vectorial representation of each of the approaches, first

[3]http://svmlight.joachims.org/

ODP Top level							
Training set size	600	1400	2200	3000	4000	5000	6000
content	**.518**	**.560**	**.579**	**.588**	**.595**	**.604**	**.610**
cont. + anchor	.461	.501	.519	.534	.542	.553	.563
ODP Second level							
Training set size	600	1400	2200	3000	4000	5000	6000
content	**.337**	**.394**	**.422**	**.437**	**.450**	**.464**	**.470**
cont. + anchor	.279	.331	.360	.376	.391	.404	.415

Table 1: Accuracy results of content-based classification.

we strip HTML tags in the web documents. After that, we carry out some linguistic processing: we remove the stopwords, and run the Porter stemmer. After removing terms with low Document Frequency (DF) values (those appearing only in one document), we weigh the terms using the Term Frequency-Inverse Document Frequency (TF-IDF) function. These values define the final vectors of the documents.

The results of the content-based classification upon both levels are shown in Table 1. The *training set size* row shows the number of instances used for training the classifiers, whereas their quality is calculated by means of the accuracy measure. It can be seen that the accuracy drops about 5% when anchor texts are included, so using them seems to be harmful for this task. The main majority of the anchor texts in our collection seem to provide entity-related information, e.g., *Association for Computing Machinery* or *ACM* for the web page in *http://portal.acm.org*. This shows that many anchor texts do not provide topic-related information and, in consequence, they are not useful for thematic classification. (Fisher and Everson, 2003) drew a similar conclusion, stating that anchor texts are useful for classification only when there is a sufficiently high link density and links are of sufficiently high quality.

5.2 Tag-based Classification

Previous works suggest the use of tags to classify web documents, and show encouraging results (Noll and Meinel, 2008a) (Aliakbary et al., 2009). Going further, we focus on the following issues: which is the best way to exploit them? Do they outperform the content-based classification even when classifying into narrower categories? 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 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 keeps 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 web page and the top list of tags, it is possible to define the fraction of users assigned each tag. A tag would have been annotated by the 100% of the users when its weight matches the user count of a web page, 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, the tag i in a document 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 is the occurrence or non-occurrence of a tag in the top 10 list or the FTA of a web page, depending on whether we rely on the Top 10 of tags or the FTA, respectively. These approaches ignore weights of tags, 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 web page ($\{w_1, ..., w_n\}$, as described above) is considered as it is in these two approaches, relying on the Top 10 list of tags and the FTA, respectively. In this case, by definition, the weights of the tags are kept, although the amount of users bookmarking a web page is ignored. Note that different orders of magnitude are mixed up now, since the count of bookmarking users ranges from 100 to ∼61K depending on the URL.

For the FTA-based approaches, we removed the tags appearing only in a document, in order to relax the computational cost while keeping the representativity.

The results in Table 2 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, all the approaches relying on the FTA perform better than their equivalent approaches relying on the Top 10 tag list. Thus, we infer that relying on the FTA of a web page, considering even tags in the tail, yields the best results. Other approaches we tried based

ODP Top level							
Training set size	600	1400	2200	3000	4000	5000	6000
ranked	.466	.477	.490	.496	.496	.501	.488
fractions	.456	.470	.473	.475	.474	.474	.476
unweighted (10)	.503	.515	.519	.522	.527	.520	.524
unweighted (FTA)	.523	.552	.557	.563	.561	.566	.569
weighted (10)	.510	.574	.604	.620	.634	.641	.652
weighted (FTA)	**.526**	**.590**	**.616**	**.636**	**.645**	**.654**	**.665**
ODP Second level							
Training set size	600	1400	2200	3000	4000	5000	6000
ranked	.296	.329	.337	.340	.345	.356	.338
fractions	.296	.323	.330	.329	.328	.330	.330
unweighted (10)	.319	.352	.347	.361	.365	.371	.368
unweighted (FTA)	**.403**	**.470**	**.489**	.493	.506	.521	.513
weighted (10)	.342	.429	.470	.492	.511	.525	.536
weighted (FTA)	.347	.439	.478	**.504**	**.521**	**.534**	**.547**

Table 2: Accuracy results of tag-based classification.

on intermediate *Top N* approaches, where *N* was higher than 10 and lower than the number of tags in the FTA (e.g., N = 50, and N = 100), produced intermediate accuracy results. This suggests considering even tags in the tail. Hence, annotations of users differing from common behaviors are also helpful. Other ideas like a possible removal of useless or harmful tags set by misbehaving users remain as open issues.

Nonetheless, among the FTA-based approaches, the weighted and the unweighted have different results depending on the categorization level we work with. The weighted performs better for all the training set sizes with the top level categorization, whereas the unweighted outperforms with the smallest training sets for the second level categorization. However, the weighted approach outperforms the unweighted one when the training set has 3,000 documents or more. Note that the second level is made up by 243 categories, thus the weighted approach, relying on more scattered values, may require larger training sets than the unweighted approach to benefit the classifier. Thus, we conclude that even though the weighted approach requires more documents for the training phase, it is the best approach to represent tags.

5.3 Comment-based Classification

With regard to comments, two kinds of metadata are stored in our dataset: notes and reviews. Both are free texts describing or referring to a web page. The *a priori* difference among them is in the objectivity. Due to web interfaces, notes on Delicious tend to be objective descriptions, whereas reviews on StumbleUpon many times have a subjective idea in them. However, since objective terms may also be found in reviews, we consider studying their contribution to this task. As we stated above, there is a number of web pages without any review, so that reviews would not be able to classify web pages by themselves. However, this information could be useful combined with notes. We have tested the following two approaches:

ODP Top level							
Training set size	600	1400	2200	3000	4000	5000	6000
notes	.515	.571	.596	.612	.623	.632	.640
notes + reviews	**.520**	**.578**	**.602**	**.618**	**.630**	**.639**	**.646**
ODP Second level							
Training set size	600	1400	2200	3000	4000	5000	6000
notes	.346	.419	.454	.473	.490	.505	.516
notes + reviews	**.349**	**.423**	**.459**	**.478**	**.497**	**.511**	**.524**

Table 3: Accuracy results of comment-based classification.

Only notes: all the notes annotated to each page are merged into one. After merging, the vectorial representation is obtained for each web page. To achieve this, we based on the TF-IDF function, and removed terms appearing only in one document.

Merging notes and reviews: reviews are also taken into account for this approach. Similar to notes, we also merged notes and reviews. TF-IDF term weighting scheme was also applied to obtain the vectorial representation, with the same reduction.

Table 3 shows the results for these two approaches. Combining reviews with notes provides slightly better results than using only notes. The reason for this small contribution may be determined by the lack of reviews for many of the web pages. Even though reviews initially may have subjective nature by definition, they have shown to be slightly helpful for this task. From these results we could infer that, as the availability of reviews increases, they will become more beneficial.

5.4 Comparing Data: Content vs Annotations

With the three experiments above, we came up with the best approach for each kind of data. Once we had these results, we compared their usefulness, relying on the best approaches. Thus, we compare the following approaches: *self-content*, *comments* including *notes* and *reviews*, and *FTA-based*

ODP Top level							
Training set size	600	1400	2200	3000	4000	5000	6000
content	.518	.560	.579	.588	.595	.604	.610
comments	.520	.578	.602	.618	.630	.639	.646
tags	**.526**	**.590**	**.616**	**.636**	**.645**	**.654**	**.665**
ODP Second level							
Training set size	600	1400	2200	3000	4000	5000	6000
content	.337	.394	.422	.437	.450	.464	.470
comments	**.349**	.423	.459	.478	.497	.511	.524
tags	.347	**.439**	**.478**	**.504**	**.521**	**.534**	**.547**

Table 4: Comparison of accuracy results by content, comment and tag-based classifiers.

weighted approach for the tags.

Table 4 shows the results of the comparison of the three approaches. The results show that both social annotations improve the content-based baseline in either classification schemes. Moreover, when the size of the training set increases the content-based approach is left behind in the top level classification, with an increasing gap among the content and tag-based approaches.

Comparing the social annotations, the results show higher performance for the approach using tagging data, with an increasing gap for both top and second level classification schemes as the training set size grows. Also, the gap over the content-based approach is much higher than the gap over the comments in most cases. This makes social annotations really powerful as against the content in both categorization levels.

5.5 Using Classifier Committees

Even though tags outperform the other two approaches, all of them seem to be good enough to combine them trying to improve performance of classifiers. What if a classifier is getting right while the others are missing? Could we combine the results to best use them? An interesting approach to combine classifiers is the use of classifier committees (Sun et al., 2004), which combine predictions of different classifiers. A decision function defines the way predictions are merged.

A SVM classifier outputs a margin for each document over each class in the taxonomy, providing the reliability to belong to that class. The class with the largest positive margin for a document is then selected as the prediction of the classifier. The combination of predictions of SVM classifiers can be done by adding up their margins for each class. Each document will then have a new reliability value (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: $m'_{ijc} = m_{ijc}/max(m_i)$, where m_{ijc} is the margin by the classifier i between the document j and the hyperplane for the class c, and m'_{ijc} is its value after normalizing it.

ODP Top level							
Training set size	600	1400	2200	3000	4000	5000	6000
tags	.526	.590	.616	.636	.645	.654	.665
cont. + comm.	.554	.604	.627	.643	.651	.660	.670
cont. + tags	.549	.611	.636	.654	.664	.673	.684
comm. + tags	.546	.612	.639	.657	.668	.678	.687
cont. + comm. + tags	**.563**	**.626**	**.651**	**.669**	**.679**	**.688**	**.699**
ODP Second level							
Training set size	600	1400	2200	3000	4000	5000	6000
tags	.347	.439	.478	.504	.521	.534	.547
cont. + comm.	.399	.466	.502	.522	.540	.558	.569
cont. + tags	.396	.483	.521	.547	.564	.579	.595
comm. + tags	.385	.480	.519	.544	.562	.577	.592
cont. + comm. + tags	**.409**	**.498**	**.535**	**.560**	**.578**	**.592**	**.606**

Table 5: Accuracy results of classifier committees (best simple classifier, tags, shown to enable comparison).

The class maximizing this sum will be predicted by the committees. Then, the sum of margins among the class c and the document j using a committee with n classifiers could be defined as $S_{jc} = \sum_i^n m'_{ijc}$. 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 combination experiments by using the best approaches for tags, comments and content, as described above in Section 5.4.

The results of the experiments using classifier committees are shown in Table 5. Note that the table includes the results of the tag-based classifier, enabling the comparison of the results by classifier committees to the best of the single classifiers. When different classifiers are combined, the errors of a classifier can be corrected by the others, as these results show. It is worthwhile noting 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 good at committees.

Making different combinations among the classifiers outperformed the best non-

combining approach in all cases. Either of the committees performs better than using only tags in both the top and the second level taxonomies. 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 yields the highest performance, outperforming any of the combinations where only two kinds of metadata are considered. The outperformance of the triple-committee over the tag-based classifier is remarkable, with a gap of at least .033 for the top level, and at least .056 for the second. Among double-committees, the performance is higher when tags are considered; this means that tags are also the most helpful for committees, not only as a single classifier. Finally, comments and content perform similarly for committees, as shown by the results.

5.6 Appropriateness for Home and Deep Pages

In a deeper analysis on the use of social annotations for web page classifications, we studied how correctly they perform over home and deep pages. We set a web page as home page if it only has the part of the domain in the URL[4], whereas we set it as a deep page if it also has a path besides the domain[5].

Using classifier committees, deep pages are classified with higher accuracy than entry or home pages for both the top and the second level (see Table 6). There is an exception for the second level when the training set has 6,000 documents, where the classifier gets similar results for both home and deep pages. From these results, it can be seen that both are classifiable this way.

Regarding the results of the non-combining classifiers (see Table 7), we show that the classification by means of tags outperforms that by content for both home and deep pages in either of the classification levels, except for the home pages using the smallest training set with 600 documents, where they perform similarly. Hence, our experiments contradict the hypothesis in (Noll and Meinel, 2008a) that more direct content analysis should be needed instead of relying on tags for deep pages. Note that, in our dataset, we have 10,153 home pages, whereas 2,463 are deep pages. Even though

[4]e.g., http://www.flickr.com/

[5]e.g., http://www.flickr.com/photos/

ODP Top level							
Training set size	600	1400	2200	3000	4000	5000	6000
home pages	.548	.613	.640	.657	.668	.678	.690
global	.563	.626	.651	.669	.679	.688	.699
deep pages	**.627**	**.679**	**.699**	**.715**	**.723**	**.729**	**.734**
ODP Second level							
Training set size	600	1400	2200	3000	4000	5000	6000
home pages	.401	.493	.532	.557	.576	.589	**.606**
global	.409	.498	.535	.560	.578	.592	.606
deep pages	**.440**	**.518**	**.550**	**.570**	**.589**	**.601**	.605

Table 6: Accuracy results of home pages vs. deep pages, using the classifier committees.

ODP Top level							
Home pages							
Training set size	600	1400	2200	3000	4000	5000	6000
content	.508	.552	.570	.580	.586	.594	.600
comments	**.509**	.568	.592	.608	.620	.629	.636
tags	.508	**.575**	**.603**	**.624**	**.633**	**.644**	**.656**
Deep pages							
Training set size	600	1400	2200	3000	4000	5000	6000
content	.561	.597	.613	.625	.634	.645	.655
comments	.565	.620	.641	.661	.670	.678	.688
tags	**.597**	**.650**	**.671**	**.689**	**.697**	**.700**	**.704**
ODP Second level							
Home pages							
Training set size	600	1400	2200	3000	4000	5000	6000
content	.334	.392	.419	.434	.447	.460	.468
comments	**.345**	.418	.454	.472	.490	.504	.516
tags	.332	**.426**	**.466**	**.494**	**.511**	**.524**	**.539**
Deep pages							
Training set size	600	1400	2200	3000	4000	5000	6000
content	.350	.401	.432	.446	.464	.479	.479
comments	.367	.443	.483	.504	.526	.541	.558
tags	**.412**	**.493**	**.528**	**.543**	**.563**	**.577**	**.580**

Table 7: Comparison of data sources to classify home and deep pages.

we have many more home than deep pages in our collection, we cannot conclude whether users tend to annotate more in this kind of pages, since our set of pages is conditioned by their matching between the ODP and the URLs we retrieved from Delicious.

6 Conclusion

We have analyzed and studied the use of social annotations for web page classification over the top and second levels of the ODP. We show that tags and comments are representative enough to perform the task. Using tags or comments have shown outperformance against the content-based approach, both when classifying into broader and narrower categories. Among these social annotations, tags show the best results, for which using the FTA is 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. Combining all of them outperforms any of the other combinations as well as the non-combining approaches. 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.

As opposed to the hypotheses in the previous work by (Noll and Meinel, 2008a), we show that social annotations perform better than content both for (a) broader categories within a higher level or narrower categories within a deeper taxonomy level, and for (b) home or deep pages. On the one hand, social annotations outperform content for either classification levels in the taxonomy. On the other hand, the classification of both home and deep pages is predicted better with social tags. Moreover, social annotations provided for deep pages seem to fit better the taxonomy, whereas home pages generally get lower results. Our conjecture is that deep pages tend to be annotated by more specific tags, being more specific pages, therefore they get a more precise tag set, better fitting the classification scheme.

As a future work, we plan the application of these experiments to other collections of tagged resources to generalize the conclusions.

References

Aliakbary, Sadegh, Hassan Abolhassani, Hossein Rahmani, and Behrooz Nobakht. 2009. Web page classification using social tags. *IEEE Intl. Conf. on Computational Science and Engineering*, 4:588–593.

Fisher, Michelle and Richard Everson. 2003. When are links useful? experiments in text classification. In Fabrizio Sebastiani, editor, *Advances in Information Retrieval*, volume 2633 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, pages 547–547.

Golder, Scott and Bernardo A. Huberman. 2006. The structure of collaborative tagging systems. *Journal of Information Science, 32(2)*, pages 198–208.

Heymann, Paul, Georgia Koutrika, and Hector Garcia-Molina. 2008. Can social bookmarking improve web search? In *WSDM '08*, pages 195–206, New York, NY, USA. ACM.

Noll, Michael G. and Christoph Meinel. 2008a. Exploring social annotations for web document classification. In *Proc. of the 2008 ACM Symposium on Applied Computing*, pages 2315–2320, Fortaleza, Ceara, Brazil. ACM.

Noll, Michael G. and Christoph Meinel. 2008b. 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.

Qi, Xiaoguang and Brian D. Davison. 2009. Web page classification: Features and algorithms. *ACM Computing Surveys*, 41:12:1–12:31, February.

Ramage, Daniel, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. 2009. Clustering the tagged web. In *Proc. of the Second ACM Intl. Conference on Web Search and Data Mining*, pages 54–63, Barcelona, Spain. ACM.

Sun, Bing-Yu, De-Shuang Huang, Lin Guo, and Zhong-Qiu Zhao. 2004. Support vector machine committee for classification. In *Advances in Neural Networks - ISNN 2004*, pages 648–653.

Weston, J. and C. Watkins. 1999. Multiclass support vector machines. In *Proc. of the 1999 European Symposium on Artificial Neural Networks*.

Yeung, Ching Man Au, Nicholas Gibbins, and Nigel Shadbolt. 2008. Web search disambiguation by collaborative tagging. In *Proc. of the Workshop on Exploring Semantic Innotations in Information Retrieval at ECIR'08*, pages 48–61, March.

Zhou, Ding, Jiang Bian, Shuyi Zheng, Hongyuan Zha, and C. Lee Giles. 2008. Exploring social annotations for information retrieval. In *Proc. of the 17th international conference on World Wide Web*, pages 715–724, Beijing, China. ACM.

Zubiaga, Arkaitz, Raquel Martínez, and Víctor Fresno. 2009. Getting the most out of social annotations for web page classification. In *DocEng '09: Proc. of the 9th ACM symposium on Document Engineering*, pages 74–83, New York, NY, USA. ACM.