2009 International Conference on Computational Science and Engineering

Web Page Classification Using Social Tags*

Sadegh Aliakbary
Sharif University of Technology
aliakbary@ce.sharif.edu

Hassan Abolhassani
Sharif University of Technology
abolhassani@ sharif.edu

Hossein Rahmani
Sharif University of Technology,
Leiden Institute of Advanced Computer Science
hrahmani@liacs.nl

Behrooz Nobakht
Sharif University of Technology
behrooz@ce.sharif.edu

Abstract — **Social tagging is a process in which many users add metadata to a shared content. Through the past few years, the popularity of social tagging has grown on the web. In this paper we investigated the use of social tags for web page classification: adding new web pages to an existing web directory. A web directory is a general human-edited directory of web pages. It classifies a collection of pages into a wide range of hierarchical categories. The problem with manual construction and maintenance of web directories is the significant need of time and effort by human experts. Our proposed method is based on applying different automatic approaches of using social tags for extending web directories with new URLs.**

Keywords: Social Tagging, Web Directory, Classification.

I. INTRODUCTION

Exponential growth of data available on the web is ongoing. Since there is a vast variety in both information content and quality of web pages, organizing them is not at all an easy task. In order to extract and exploit the full potential of the massive information resources, some tasks such as describing and organizing are required. The need for web page organization has been on the desk for a long time and varied types of approaches have been proposed and applied for solving the problem. Earlier, domain experts did the classification manually. Some web directories such as yahoo [6], looksmart [5] and Open Directory Project (dmoz) [4] were created. Due to large number of resources available on the web and over-changing nature of web pages, the need for semi-automatic and automatic classification methods emerged very soon. According to [3], some of the approaches are text-based categorization on statistical and machine-learning algorithms like K-Nearest Neighbor approach [2], Bayesian probabilistic models [7][8], inductive rule learning [10], support vector machines [11] and neural networks [12]. Applying automatic or semi-automatic methods for web page classification will save a lot of time and human efforts. However, due to some existing problems related to web contents, such as their noisiness and lack of effective informative feature extraction algorithms, it seems that we should look for other sources of knowledge for improving the performance of web page classification. One of the main problems in classifying the web pages is to separate the noisy part from the informative part. Part of web pages such as copy right block, privacy and navigations parts are not related to web page content and we assume them as the noisy part of the web page. The rest of the web page's content which reflects the aim of web page are the informative parts. We should mention that separating the noisy part from the informative part is one of the major challenges in the web page classification area. We used social tags as a new source of information about web pages.

With the rise of Web 2.0 technologies, web users with different backgrounds are creating annotations for web pages for different purposes. Perhaps, adding semantic labels for increasing accessibility to the web page is the main target of web users. For example, the famous social bookmark web site del.icio.us [13] (henceforth referred to as "Delicious"), has more than two million registered users before its fourth birthday [14]. Social annotations are emergent useful information that can be used in various ways. However, to the best of our knowledge, there is no direct work exploiting this valuable information for web directory extensions.

In this Paper, we introduce our social annotation based method for web page classification. Del.icio.us is used as the source of social tags and DMOZ as the base hierarchy of web pages. In fact, for each new URL we try to predict its location in dmoz directory using its tags in del.icio.us. We proved that when we have enough taggers for a web page, the tags would be well informative descriptions for that web page. This paper proposes a classification method for web resources. We investigated new methods for enriching existing web directories with new URLs. A web directory has some predefined categories (topics) and some web pages. The experts manually assign each web page to the different categories. In our proposed method, our classifier assigns new web pages to existing categories. The key idea is to describe both the web page and the category in terms of associated tags, and then to assign the resource to the category with the most similar tag-space representation. The proposed method is evaluated by the Jaccard similarity between actual category in DMOZ and their predicted categories, and by precision. The proposed method can also be used as a semi-automatic approach for web page classification: An expert watches the output of the proposed method and decides whether the suggested topic is correct or needs correction.

The rest of the paper is organized as follows. Section 2 and 3 discuss motivation and related works, respectively. Section 4 describes the novel tag-based web page classification method in detail. The experimental results are dealt with in Section 5. We present our conclusions and avenues for future works in this direction in Section 6.

II. MOTIVATIONS OF THE RESEARCH

The motivations for this research revolve around applying novel methods of using social tags for the problem of web page classification. If there are enough annotations (tags) for a web

* This paper is supported by Iran Telecommunication Research Center (ITRC).

978-0-7695-3823-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CSE.2009.411

page, these tags provide good, high-level, and less-noisy information about the content of the web page. In this situation we use the tags, instead of the content, to create text vector-space representation of the web pages.

Some of the benefits of using social tags over pure content of web pages are as follows:

- **Less noisy content**: Comparing to regular web page content, social tags have much less noisy content. Regular web pages have several kinds of intra-page noises, including site logos, decoration images, copyright notices, privacy statements and advertisements. There are several workarounds for noise removal [9], [23], and [29]. By using social tags, if there are enough annotations for a web page, the need for content cleaning algorithms will be eliminated.
- **High speed**: If we analyze the trajectory of web page classifiers, we will find out that there are huge number of classifiers which utilize summarization techniques for increasing both precision and speed of classification process. Social tags are good summarized representation of web pages. Using folksonomy tags for classification tasks is useful and leverages the information provided by user-generated content.
- **Stability**: Due to dynamic nature of internet, web page contents are changing constantly. Through monitoring the web page related tags, we will find out that big changes are rarely happened to the tag set of a specific web page. According to [1], usually after about one hundred bookmarks, each tag's frequency is a nearly fixed proportion of the total frequency of all tags used.

III. Related Works

A. *Research on Web Page Classification*

There is a variety of methods for web page classification with great diversity of success. The main classification techniques can be organized into the following broad categories:

1. Manual classification by domain specific experts.
2. Categorization using HTML META tags (which serves the purpose of document indexing).
3. Document content-based approaches.
4. Categorization based on a combination of document content and META tags.

Manual Classification: In this category of web page classification, a number of domain experts analyze the content of the web page and choose the best category for a specific web page. There are several problems related to these methods. The huge volume of web content rules out this approach. Dynamic environment and over-changing nature of web content make the manual classification of web pages highly difficult resulting in interests in automating the classification process.

Classification by HTML META tags: These classification techniques solely rely on content attributes of keywords and description html tags (i.e. *<META name=''Keywords''>* and *<META name=''description''>*). Sometimes, relying on these tags might give accurate results to a large extent. [19] shows that by introducing the "title" content to the similarity evaluation, recall and precision increases about 21.7%. However these approaches have their own problems. For example, there are many web pages without any value for these HTML tags or have some default values. There is also the possibility of noise in these HTML tags. For example, web page author may include keywords that don't reflect the content of the page, just to increase the hit-ratio of the page in search engine results.

Content-based and combined approaches: In these categories, text content of each page will be used through the classification process. Since text content of web pages is filled with lots of noises and irrelevant data, strong preprocessing is needed for extracting the relevant text content. In the most text-based approaches, the stop words are removed and some techniques such as style tree [20] might be used for removing the noisy blocks (i.e. copy right, privacy and navigation blocks). The next step of preprocessing is stemming. The remaining words are the stemmed keywords and can be used for classification. Building feature vectors from these keywords is the last phase of pre-processing. After pre-processing, each document is represented by a feature vector and the vectors are classified into an appropriate category using text classification methods such as the K-Nearest Neighbor classification algorithm.

B. *Research on Social Annotations*

Some work has been done on exploring the social annotations for different purposes. [15] has used social annotations for creating folksonomies. [17] looks for creating semantic web components. [18] and [28] used social tags for enriching the search process. [16], [26], and [27] used social tags for visualization. The term "Folksonomy", a combination of "folk" and "taxonomy", was first proposed by T. V. Wal in a mailing list [22]. It provides user-created metadata rather than the professional created and author created metadata [15]. A general introduction of folksonomy could be found in [25]. [1] analyzed the structure of collaborative tagging systems as well as their dynamical aspects. [24] proposed Adapted PageRank and FolkRank to find communities within the folksonomy. The problem of visualizing the evolution of tags is considered in [16]. They presented a new approach based on a characterization of the most interesting tags associated with a sliding time interval.

Some applications based on social annotations have also been explored in [21]. They proposed a tripartite model of actors, concepts and instances for semantic emergence. [17] explored machine understandable semantics from social annotations in a statistical way and applied the derived emergent semantics to discover and search shared web bookmarks. [18] lightened the limitation of the amount and quality of anchor text by using user annotations to improve the quality of intranet search. [26] provided a possible way to navigate through the tagging space by understanding the semantics of tagging. They use Self-Organizing Map (SOM) to visualize multi-dimensional data onto a 2D map. SOM provides a graphical map that reveals important information in tagging space for the users of the collaborative tagging systems. [27] has developed an online visualization tool called Cloudalicious33 that gives insight into how folksonomies are developed over time for a given URL. [28] experimented with an exploratory system using folksonomy tags to enhance searching. They integrated Google's search functionality and the URL check functionality provided by del.icio.us to provide adaptive guidance to users. As the user uses the exploratory system interface to input his/her keywords, the system sends the keywords to Google and also to del.icio.us to extract the corresponding tags. Then the keywords contained in the tags are displayed as clickable hints. Their preliminary evaluation showed that their technique has increased the accuracy of search.

Different from the above mentioned researches, we investigate the capability of social annotations in improving the quality of web page classification

IV. PROPOSED METHOD

In this section we introduce our social annotation based method for web page classification. Del.icio.us is used as the source of social tags and DMOZ as the base hierarchy of web pages. In fact, for each new URL we try to predict its location in dmoz directory using its tags in del.icio.us. When we have enough taggers for a web page, the tags would be good informative descriptions for that web page [28]. So, if we can effectively eliminate the impact of noise from the tags, they will be concise representatives of web pages.

We first describe the preprocessing phase and then the novel algorithms for predicting the class of URLs are explained.

A. Preprocessing

Sometimes tags have some kind of noise, ambiguity or similarity. We tried to decrease the effect of these noises in the preprocessing phase. As described in Figure.1, preprocessing phase contains three stages of "converting to lowercase", "bad tags removal" and "stemming".

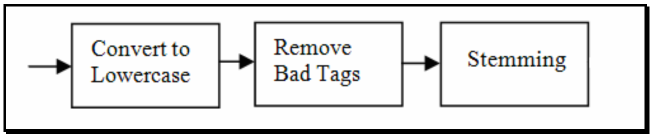

Figure 1. The stages of preprocessing phase.

In the first stage, all tags are converted to lowercase. Then in the "bad tags removal" stage, by using a blacklist, we remove non-informative tags. There is no automatic method for detecting these terms and non-informative terms are detected manually by experts. Some of these useless tags are generated automatically by tagging tools (e.g. "imported_ie_favorites") and some others are informative only for the tagger (e.g. "cool", "toRead" and "myStuff"). These tags are categorized as "self reference" and "task organizing" in [1]. Inspecting popular tags in the dataset, we have manually filled and managed the blacklist. Some tags of our blacklist are: «*Imported, Interesting, TODO, firefox:toolbar, safari_export, Toread, Importediefavorites, Imported_ie_favorites, .imported, Toolbarfavorites, toolbar_favorites, Fromsafari*»

In the last stage of preprocessing, the English tags are stemmed using Porter algorithm. The process of stemming unifies different forms of a common word (e.g. "shop" and "shopping").

B. Exhaustive Search

In our proposed method, each web page is represented by its "tag vector". A tag vector is represented in text vector space model and contains all tags and their weights for the web page. For example tag vector for http://java.sun.com is something like (*<java, 377>, <program, 147>, <sun, 118>, <development, 102>, …*). The tags and their weights are extracted from a dataset containing tag information of various web pages. This dataset is constructed through crawling Delicious site.

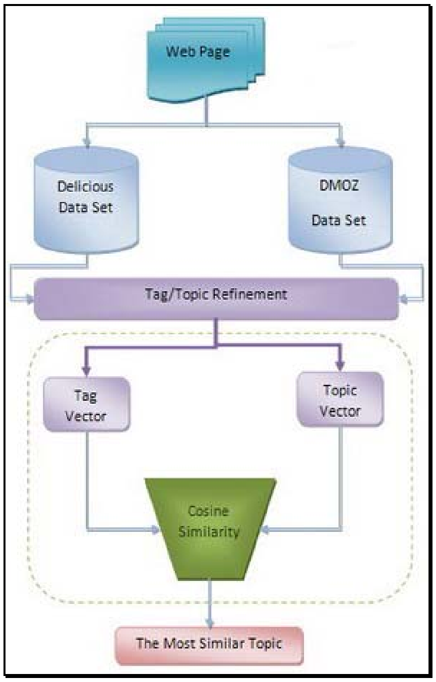

Figure 2. Exhaustive Search

We also construct a "topic vector" for each dmoz topic, which is represented in text vector space model. Each topic vector contains the information about all tags of the URLs in the topic and its subtopics. In fact, each topic and its subtopics have many web pages, and by combining the tags of these web pages, we construct tag vectors for the topics. In order to decrease the time complexity, we only considered the top 40 tags of each topic.

Now we have a vector corresponding to each category called "topic vector" and a vector corresponding to each new URL called "tag vector". These vectors have been refined in preprocessing phase as described in section 3.1. For each new URL, its tag vector is compared with all the topic vectors to find the most similar topic using cosine similarity measure, which is described in formula (1).

$$\cos(\ a,b) = \frac{\vec{a} \bullet \vec{b}}{\left|\vec{a}\right| \cdot \left|\vec{b}\right|} = \frac{\sum_{i=1}^{t} (a_i \cdot b_i)}{\sqrt{\sum_{i=1}^{t} a_i^{\ 2} \cdot \sum_{i=1}^{t} b_i^{\ 2}}} \tag{1}$$

Considering cosine similarity formula, we will find out that two vectors are normalized based on their weights. For instance, if 5 out of 5 users used the tag "linux" in a URL and 10 out of 1000 used the tag "linux" in another URL. By this kind of normalization, the first URL would be more related to “linux”.

The most similar topic is returned as the best topic for new URL. This method has a good precision, but the time performance can be better: each tag vector is compared with all topic vectors. In dmoz hierarchy, we have about 400,000 topics and this comparison is not necessary for many topics. We can filter most of the irrelevant topics to speed up the process of finding the best topic. In the next section we describe a simple rule for filtering irrelevant topics. Exhaustive search is illustrated in Figure 2.

C. Pruned Search

If we compare a tag vector with all the topic vectors, we may consider many topics, which are not relevant to the web page. So we filter the topics and only consider some of them.

For each web page, we watch its tag vector and find the most important tag. This is the tag with the most frequency, for example in the tag vector corresponding to *http://java.sun.com* (*<java,377>, <programming,147>, <sun,118>, <development,102>,* ...), the most important tag is java, because its frequency (377) is more than frequencies of other tags. Now, we only consider the topics which have this tag in their topic vectors and filter out other topics. By maintaining an index from tags to topics, finding these topics is done very fast. Therefore, many topics are omitted and for the remaining topics, the comparison between the tag vector and topic vectors is done by cosine similarity, and the most similar topic is returned as the suggested topic for the web page. Figure 3 shows the process of pruned search.

V. Evaluation of the Proposed Method

The accuracy of the method is calculated using both Jaccard measure and total precision. Jaccard measure simply calculates similarity of two sets **A** and **B** as formula 2.

$$\text{Jaccard Similarity (A, B)} = \frac{\left|A \cap B\right|}{\left|A \cup B\right|} \tag{2}$$

In our approach, each URL with the exact category “E” in the web directory is assigned the predicted category “P” by the proposed method. The similarity between P and E is calculated using Jaccard measure. For example if “**Computers/Programming**/Languages/Java” is the exact category and “**Computers/Programming**/Resources” is the predicted category, then Jaccard-similarity(P,E) will be 2/5: $A \cap B$ = { Computers, Programming} and $A \cup B$ = { Computers, Programming, Languages, Java, Resources} . The intersection is calculated only from the beginning of the topic path.

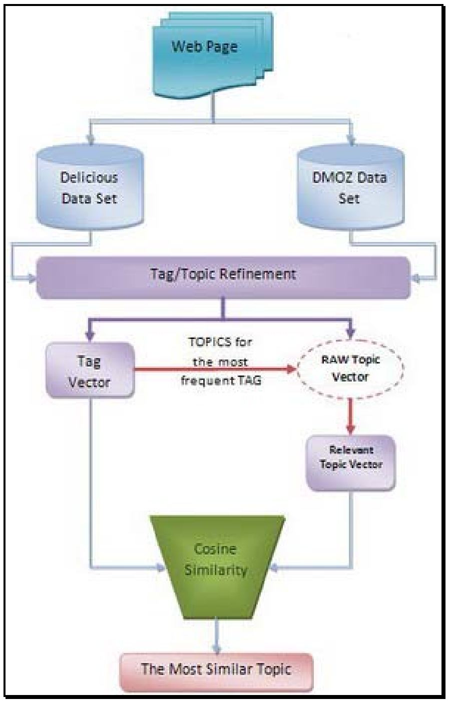

Figure 3. Pruned Search

Instead of Jaccard similarity, we can also calculate the total precision of the method. This measure will show the percentage of URLs assigned to their exact category.

$$\text{Precision} = \frac{\#\mathit{Correct}\,\mathrm{Pr}\,\mathit{edictions}}{\#\mathit{total}\,\mathrm{Pr}\,\mathit{edictions}} \tag{3}$$

In many cases, Jaccard similarity is more communicative than the precision measure. For example if the suggested topic is the parent of the exact topic (for example Computers/Programming/Languages instead of Computers/Programming/Languages/Java) the precision of this prediction would be zero, but this is not a bad predication.

A. Dataset

The dataset contains dmoz directory, dmoz URLs and social tags for these URLs. Dmoz directory and URLs are downloaded from its official site [4] and imported into the database. Since we just considered dmoz topic hierarchy (i.e. we did not analyze the dmoz text description), our proposed methods are applicable on any predefined web page taxonomy. For each URL in dmoz hierarchy, we retrieved the corresponding tags from delicious web site.

Our dataset contains all URLs in “Computers” category of dmoz hierarchy and their corresponding social tags. The dataset is restricted to this topic because constructing a complete dataset for all dmoz topics (with more than four million URLs) is a time-consuming process.

Our tag database contains more than 140,000 URLs, which almost half of them are tagged on delicious site. For each URL, we maintain its tags and also frequency of each tag. The frequency of a tag is the number of users utilizing the tag for bookmarking that URL. The dataset contains about 0.5 million triples of <URL, tag, frequency>.

Each topic in dmoz directory has a depth, describing the deepness of the topic in the directory. For example the depth of "Computers/Internet/Searching/Search_Engines/Google" is 5. The average depth of topics of URLs in dmoz hierarchy is 4.04 in our dataset.

B. Experimental Results

We partitioned the data set to Training Set (95% of the data set) and Test Set (5% of the data set). The Test Set is to examine the proposed method. The URLs of the test-set are not included in constructing the topic vectors.

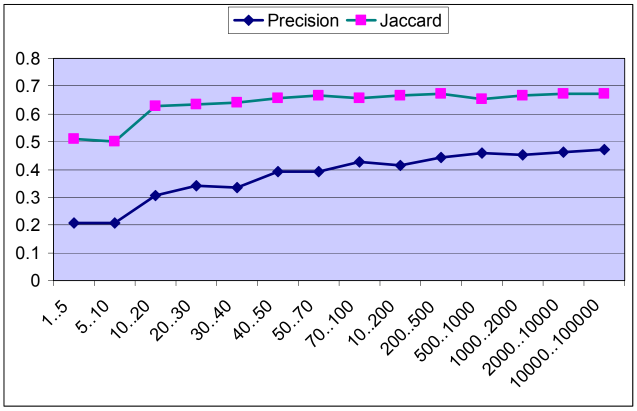

Figure 4. Evaluation of the proposed method.

Based on Jaccard similarity and precision measures, the evaluation of our proposed method is shown in Figure.2.

In Figure.2 the Y axis shows the Jaccard similarity or precision and X axis shows the number of users tagged a URL. For example, the value corresponding to "10..20" is the average accuracy of predicted topics for URLs having 10 to 20 taggers. As the Figure.2 shows, the Jaccard similarity rapidly exceeds 0.6 as number of taggers exceeds 10, and precision slowly approaches to 0.5.

As the diagram shows, when we have more than 1000 taggers for a web page, near half of the predictions of the proposed method is exactly correct. But the Jaccard similarity is converged fast and even with 10 taggers, the average Jaccard similarity is always more than 0.6.

C. Comparison to Content-based Approaches

As described in related works section, several researchers have tried to categorize web pages based on their contents. In content-based approaches, a computer program finds the best category for a web page according to its content. Noisy content of web pages reduces the performance of the content based approach. Moreover, analyzing huge amount of web page contents is also a time consuming process. In order to compare our proposed method with content-based approaches, we implemented a content-based classifier for categorizing dmoz web pages.

In this implementation, the content of the web page is extracted using Contextor framework (described later), and then this content is tokenized and converted to vector space representation. Now, for each web page (URL), we have a "content vector". For each topic of dmoz we also create a topic vector using text description of the topic and the descriptions of its URLs in the dmoz directory: In dmoz directory, each topic and URL has a title and a description. We have constructed the topic vectors using these text descriptions.

The topic vectors and content vectors are preprocessed as described in section 3.1. Each refined content vector is compared with dmoz topic vectors using cosine similarity measure. The most similar topic vector specifies the suggested topic for the URL. As you see, the process of finding the best topic is similar to our proposed method, despite the fact that in content-based approach the vectors are constructed using page contents, but in our proposed method, the vectors are constructed using social tags.

This algorithm is examined on the test-set. The test-set is the same set of the tag-based approach. In average, Jaccard measure for this method is 0.501 and the precision measure is 0.15. As the results show, tag-based method is categorizing web pages much better than the content based methods. These results confirm our assumption about tag-space representation of web pages: If we have enough tags for a web page, social tags are better representatives of web pages comparing to their content only. In addition, tag-based method is much faster than content-based method. Our experiments show that tag-based method is about 10 times faster than content-based method[†]. This is mainly because content-based approach requires much work on parsing and preprocessing of web page contents.

D. Contextor Framework

We have implemented a framework for extracting visible content of web pages. This framework, called Contextor[‡] (Content Extractor), reads web page HTML contents and returns visible text content of the web page. Normalization of the text, removal of stop words, stemming and all processing stages are implemented in this framework and also is pluggable. This framework is an extensible general purpose web mining software and can be used in other researches in the area. As Contextor takes advantage of Dependency Injection concepts along with Spring Framework, it is highly pluggable and extensible in the ways it needs to deliver its mission. Contextor is consisted of, as its major architectural elements: **(1)** Content Loader/Reader: A loading mechanism to load the raw data as "term vector" that is required to be processed through different sources such as web, file system, and database. **(2)** Term Filter: A filtering mechanism to apply a chain of data processing filters to a term vector using a set of

[†] For each URL, our tag-based method suggests a topic in less than a second, on an ordinary P4 PC.

[‡] http://contextor.sourceforge.net

"term deciders". **(3)** Term Decider: A term decision platform on which various and distinct "term decider's" are defined to decide whether a specific term is acceptable in a known process. **(4)** Document Object Model: an object model to describe the resource that has been loaded and filtered (and decided). **(5)** Statistical Descriptive Models: through the use of document object model, a number of statistical, comparative, or descriptive models can be built for various goals of data mining processing such as matching documents and comparing the similarities of two documents.

VI. CONCLUSION

In this paper, we studied the novel approach of integrating social annotations into the problem of web directory extension. We observed that the fast emerging annotations provided a multi-faceted summary of the web pages. A novel method is proposed to capture the social annotations' capability on determining the category of web pages. As the results show, the precision of proposed method increase with the number of users tagged a URL: more tagged URLs are better classified. We have also showed that social annotations are better representatives of web pages than their content, and the precision of proposed method is much better than content-based approaches.

Nowadays, the use of social tagging systems is growing rapidly. Web pages are tagged in various tagging services and the social annotations provide good information about web pages. So, our tag-based approach with no doubt will be helpful. The proposed method can also be used in a semi-automatic classification system. In such a system, the suggested topics are declared to an expert editor, and the editor can accept or change the suggested topic. Even if the suggested topic is not exactly correct, it shows the similar topics and helps the editor to find the exact topic faster and more accurately.

REFERENCES

1. Golder, S., Huberman, B.A.: The Structure of Collaborative Tagging Systems. Technical report, Information Dynamics Lab, HP Labs, 2005.
2. Yang, Y., Lui, X.: A Reexamination of Text Categorization Methods. In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99), pp. 42-49, University of California, Berkeley, USA 1999.
3. Kwon, O.-W., Lee, J.-H.: Text Categorization Based on k-nearest Neighbor Approach for Web Site Classification. Information Processing and Management 39, pp. 25–44, 2003.
4. Open Directory Project, http://www.dmoz.org/
5. Looksmart, http://www.looksmart.com/
6. Yahoo!, http://www.yahoo.com
7. McCallum, A., Nigam, K.: A Comparision of Event Models for Naïve Bayes Text Classification. In AAAI-98 Workshop on Learning for Text Categorization, 1998.
8. Koller, D., Sahami, M.: Hierarchically Classifying Documents Using very few Words. In Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp.170-178, 1997.
9. Bar-Yossef, Z., Rajagopalan, S.: Template Detection via Data Mining and its Applications. In Proceedings of the 11th International World Wide Web Conference (WWW2002), 2002.
10. Apte, C., Damerau, F.: Automated Learning of Decision rules for Text Categorization. ACM Transactions on Information Systems, Vol 12, No.3, pp.233-251, 1994.
11. Dumais, S.T., Platt, J., Heckerman, D., Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization. In Proceedings of the Seventh International Conference on Information and Knowledge Management (CIKM'98), pp.148-155, 1998.
12. Weigend, A.S., Weiner, E.D., Peterson, J.O.: Exploiting Hierarchy on Text Categorization, Information Retrieval, I(3), pp.193-216, 1999.
13. Delicious, http://del.icio.us
14. Delicious Weblog, http://blog.del.icio.us
15. Mathes, A.: Folksonomies Cooperative Classification and Communication through Shared Metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html, December 2004.
16. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, R., Tomkins, A.: Visualizing Tags over Time. In Proceedings of WWW2006, pp. 193-202, May 23.26, 2006.
17. Wu, X., Zhang, L., Yu, Y.: Exploring Social Annotations for the Semantic Web. In Proceedings of WWW2006, pp. 417-426, May 23.26, 2006
18. Dmitriev, P.A., Eiron, N., Fontoura, M., Shekita, E.: Using Annotations in Enterprise Search. In Proceedings of WWW 2006 , pp. 811-817, May 23.26, 2006.
19. Hovy, E., Lin, C.Y.: Advances in Automatic Text Summarization, 1999 - acl.ldc.upenn.edu
20. Xiong, H., Pandey, G., Steinbach, M., Kumar, V.: Enhancing Data Analysis with Noise Removal. IEEE Transactions on Knowledge and Data Engineering, Volume 18, Issue 3, pp. 304-319, March 2006.
21. Mika, P.: Ontologies are us: a Unified Model of Social Networks and Semantics. In Proceedings of ISWC 2005. pp. 522-536, Nov. 2005.
22. Smith, G.: Folksonomy: social classification. http://atomiq.org/archives/2004/08/folksonom-y_social_classification.html, Aug 3, 2004.
23. Kushmerick, N. Learning to Remove Internet Advertisements. In Proceedings of the 3rd International Conference on Autonomous Agents, pp. 175–181, Seattle, Washington, 1999.
24. Hotho, A., Jaschke, R., Schmitz, C., Stumme, G.: Information Retrieval in Folksonomies: Search and Ranking. In Proceedings of ESWC 2006, pp.411-426, 2006.
25. Quintarelli, E.: Folksonomies: power to the People..http://www.iskoi.org/doc/folksono-mies.htm, June 2005.
26. Choy, S.-O., Lui, A.K.: Web Information Retrieval in Collaborative Tagging Systems. IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006), Hong Kong,pp. 352-355, IEEE Computer Society.
27. Russell, T.: Cloudalicious: Folksonomy over Time. In Proceedings of the 6th ACM/IEEE-CS joint Conference on Digital Libraries, Chapel Hill, NC, USA,pp. 364-364, ACM Press, 2006.
28. Han, P., Wang, Z., Li, Z., Kramer, B., Yang, F.: Substitution or Complement: An Empirical Analysis on the Impact of Collaborative Tagging on Web Search. IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006), Hong Kong,pp. 757-760, IEEE Computer Society, 2006.
29. Lin, S.-H., Ho, J.-M.: Discovering Informative Content Blocks from Web Documents. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.