International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

General Article

Social Tag Classification using SVM

Shilpa P Patil[Á*] and Devaraj Verma C[Ḃ]

[Á]Information Science & Engineering ,PESIT ,Bangalore ,India

Accepted 28 May 2014, Available online 01 June2014, **Vol.4, No.3 (June 2014)**

Abstract

In our daily lives, organizing resources like books or webpages into a set of categories to ease future access is a common task. The usual largeness of these collections requires a vast endeavour and an outrageous expense to organize manually. As an approach to effectively produce an automated classification of resources, consider the immense amounts of annotations provided by users on social tagging systems in the form of bookmarks. This project deal with the utilization of user provided tags to perform a social classification of resources. Those resources are accompanied by categorization data from sound expert-driven taxonomies. We analyse different representations using tags and compare to other data sources by using different settings of SVM classifiers. Finally, we explore combinations of different data sources with tags using classifier committees to best classify the resources.

Keywords: *Social Tagging, Social Annotations, Classification, SVM, Metadata.*

Introduction

The collective classification of resources into a commonly agreed structure. While libraries and librarians have performed the task of classification for centuries, the process of manually categorizing resources is expensive. The Library of Congress in the United States for example reported that the average cost of cataloging a bibliographic record by professionals was $94.58 in 2002. Given these costs, social classification systems and algorithms represent an interesting alternative. Social tagging systems like Delicious, LibraryThing or GoodReads have demonstrated their ability to quickly generate large amounts of metadata in the form of tags. These tags have been shown to be useful for, for example, information access and organization .Yet, little is known about the usefulness of social tagging data for classifying resources, or about the type of tagging behavior that yields the best classification result.

Most of the automated classifiers rely on the content of the resources, especially regarding webpage classification tasks. Nonetheless, the lack of representative data within many resources makes the classification task more complicated. In some cases, it may not be feasible to obtain enough data for certain kinds of resources such as books, where the full text is not available. Without sufficient data, representing the content becomes more challenging.

As a way to solve these issues, social tagging systems provide an easier and cheaper way to obtain metadata related to resources. Social tagging systems are a means to save, organize, and search resources, by annotating them with tags that the user provides. Systems like Delicious, LibraryThing, and GoodReads collect user annotations in the form of tags on their respective collections of resources. These user-generated tags give rise to meaningful data describing the content of the resources . User annotations can be useful to find out the aboutness of resources and to help infer the categorization. By providing tags, users are creating their own categorization system for a given resource. Given that a large number of users are providing their own annotations on each resource, our objective is focused on finding out an approach to amalgamate their contributions in such a way that resembles the categorization by professionals.

This work includes exploration of the social annotations provided by end users on social tagging systems as to performing a social classification of resources. This work focuses on the use of support vector machines (SVMs) as a state-of-the-art classification algorithm. We create three large-scale social tagging data sets including different kinds of resources, webpages and books. We analyze the characteristics of these data sets to understand how users tag, and how the nature of a social tagging system can affect the use of social tags to automatically classify resources and analyze the performance of classifying resources using social tags by comparing three different settings of classifiers.

Social Tagging

Tagging is an open way to assign tags to resources (e.g..webpages, movies, or books), enabling future retrieval in an easier way, by using tags as metadata related to resources. In addition, when a tagging system is social, tags by all the users are publicly accessible , and profitable for the community of users. The collection of tags defined by them creates a tag-based organization, so-

*Corresponding author **Shilpa P Patil** is a PG student and **Devaraj Verma C** is working as Asst Prof

called folksonomy. A folksonomy is also known as a community-based taxonomy, where the classification scheme is nonhierarchical, as opposed to a classical taxonomy-based categorization scheme. For instance, LibraryThing, GoodReads, and Delicious are social tagging systems, where each resource can be tagged by all the users who consider it interesting.

Annotations provided by users on social tagging systems have been widely deployed by researchers as metadata related to resources for tasks such as information retrieval, recommender system, discovery of emergent semantics and enhanced browsing and navigation through annotated resources.

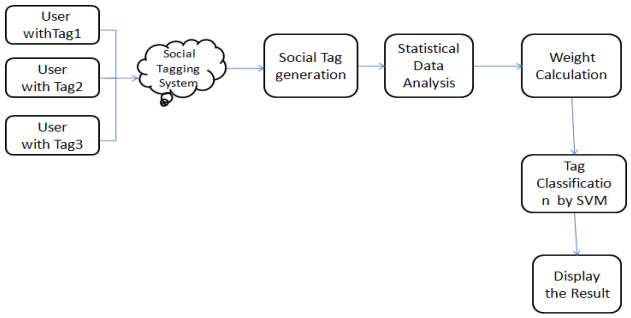
In a study of the characteristics of social annotations provided by end users, to determine usefulness of Tags for webpage classification ,weight of the tags are analyzed by normalizing the number of users annotating them. The least popular tag is given a value of 0, whereas the most popular is given a value of 1. Attention is not paid at whether this representation approach was appropriate to carry out the task. The authors matched user supplied tags of a page against its categorization by the expert editors of the ODP, even though they did not perform actual classification experiments. As per observation the power law curve was formed by the popularity of social tags, not only popular tags, but also the tags in the tail provide helpful data for information retrieval and classification tasks in general.

Resource Classification

Resource classification can be defined as the task of labeling and organizing resources within a set of predefined categories. This type of classification relies on previously categorized or labeled training sets of resources. The classifier uses these sets of resources to gather knowledge which, in turn, is used to classify new unknown resources. This work includes the analysis of several classification approaches using SVM, with the aim of analyzing their suitability to these tasks. These include different approaches to solving multiclass problems.

Annotations gathered together on social tagging systems can be harnessed for resource classification. Specifically, this work focuses on the study of several resource representation approaches using social tags. The evaluation of such representations by measuring their similarity to classifications by experts is carried out in this work .As per the classification provided by experts as a ground truth for the evaluation process, we perform the classification experiments by using a state-of-the-art classification method, so-called Support Vector Machines.

The main goal of this work is to shed new light on the appropriate use of the great deal of data gathered on social tagging systems. Given the interest of classifying resources, and the lack of representative data in many cases, Aim is at analyzing the extent to which and how social tags can enhance a resource classification task. Regarding the classification algorithm, we rely on Support Vector Machines (SVM) as a state-of-the-art classification method.



Representing Resources Using Tags

The tagging activity of a community of users creates an aggregated list of tags on each resource. A resource annotated by p users will have a list of n different tags, where each tag could have been utilized by at least 1 user and p users at most. The number of users who utilized a certain tag wt defines a value that allows to infer an ordered list of tags for a resource. Given that in this work, we rely on the vector space model to represent resources, this aggregation of annotations performed by different users could be represented in several ways, especially when it comes to assigning weights to tags.

Local Tag Weighting:

1. Fraction based tag weighting: The weight is computed according to the fraction of users who utilize a tag, wt=p, i.e., the number of users utilizing a tag on a resource, divided by the total number of users who annotated the resource.
2. Binary tag weighting : In a binary way, the presence of a tag represents a value of 1, and its absence a value of 0.
3. Frequency-based tag weighting (term frequency (TF)): It considers the number of users assigning the tag as a weight. The weight for each of the tags of a resource (w1….wn) is considered as it is in this approach.

Global Tag Weighting:

1.Term frequency-inverse document frequency (TF-IDF): TF-IDF is a term weighting function that serves as a statistical measure that defines the importance of a word to a document in a collection. When computing the TF-IDF value for the term i within the document j as a part of a document collection D, it comprises two underlying measures: 1) the TF, i.e., the number of appearances of the term i within the document j, and 2) the inverse document frequency (IDF), i.e., the logarithm of the number of documents in the whole set (D) divided by the number of documents in which the term i occurs, which refers to the general importance of the term i in the collection. The product of these two measures defines the TF-IDF weight of term i in the document j.

$$ibf_i = \log \frac{|B|}{|\{b : t_i \in B\}|}.$$

Classification Algorithm

This algorithm looks for a hyperplane that separates the

classes in a vector space model; this hyperplane should maximize the distance between it and the nearest resources, which is called the margin. several settings can be used in an SVM. Even though the SVM only solves binary classification problem by default, different approaches have been proposed to work with multiclass problems.

We use the most popular setting for supervised multiclass SVM i,e; native multiclass approach. This native multiclass approach considers the task with a single classifier, and thus, it learns a model for all the classes at the same time. The native multiclass approach we use in our experiment has been implemented by using svm-multiclass, a mSVM classifier by Joachims .

$$\min\left[\frac{1}{2}\sum_{m=1}^{k}\|\mathbf{w_m}\|^2 + C\sum_{i=1}^{l}\sum_{m\neq y_i}\xi_i^m\right]$$

$$\text{Subject to : } \mathbf{w_{y_i}}\cdot\mathbf{x_i} + b_{y_i} \geq \mathbf{w_m}\cdot\mathbf{x_i} + b_m + 2 - \xi_i^m, \xi_i^m \geq 0.$$

Experimental Result

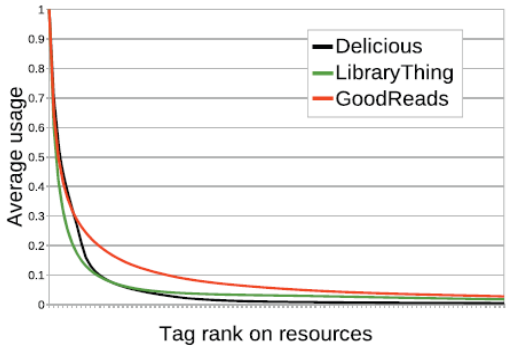


Fig shows the usage percents of tags, ordered by their usage rank (note the logarithmic scale). The three lines represent the usage of tags by users, on resources, or onbookmarks. The X-axis refers to the percent of the tag rank, whereas the Y -axis represents the percent of appearances in resources, users and bookmarks. For instance, if the tagranked first had been used on the half of the resources, the value for the top-ranked tag on resources would be 50 percent. Thus, these graphs enable us to analyze how popular are the tags in the top as compared to the tags in the tail on each site. On the other hand,it shows the average usage of tags in a given rank for resources for each data set. That is, we give a value of 1 to the tag used the most on a resource, hence ranked first for that resource. The second tag is given the value according to the fraction of users utilizing it as compared to the first one. And so on for tags ranked third, fourth, etc., on resources. Finally, we compute the average of tags ranked on each position, which is shown in the graph. It helps infer the popularity gap between top tags on resources and tags ranked lower.

The results show that the use of social tags almost always outperforms the other data sources. The exception is the mSVM for GoodReads. In general, GoodReads is the system that shows the worst performance of tags as compared to the other data sources.

This happens because GoodReads does not encourage users to attach tags to books. GoodReads requires users to add tags by following a two-step process, what makes the task less accessible.

Consequently, fewer users provide tags, and books tend to remain annotated with fewer tags. This makes tags from GoodReads not to be sufficient to yield an outperformance as Delicious and LibraryThing do. Tags from these two systems clearly outperform classification using content or reviews. Between these two data sources, user reviews usually outperform content, but not even reviews are enough to reach the performance of social tags.

Comparing the local weighting representations of social tags-Fractions, Binary, and TF there is also a clear difference among classifiers. TF is clearly the best solution when an mSVM classifier is used. However, TF performs worse or similar to Fractions and Binary approaches when combinations of binary classifiers oaaSVM and oaoSVM are used. This suggests that a native multiclass classifier as mSVM rather uses detailed weightings where the relevance of each tag is explicitly defined with the number of annotators. However, in the case of combinations of binary classifiers oaaSVM and oaoSVM—where only two classes are considered at a time, it is enough to rely on simpler weightings.

Conclusion

The results of our experiments show the great potential of social tags not only as a single classifier, but also to combine with other data sources. These results are best when a native multiclass classifier is used as the SVM setting. For the selection of an appropriate representation using social tags, the settings of the studied social tagging system should be taken into account. Among settings, we have shown that systems providing resource-based tag suggestions greatly alter folksonomies, and condition the success of certain representations.

References

Arkaitz Zubiaga (2009) Getting the most out of social annotations for web page classification

Arkaitz Zubiaga Mendialdua (Aug 2013), Harnessing Folksonomies for Resource Classification,

Mohamed Aly November (2005), Survey on Multiclass Classification Methods,

Arkaitz Zubiaga, Raquel Mart´ınez, and V´ıctor Fresno (23 Feb 2012)., Analyzing Tag Distributions in Folksonomies for Resource Classification, NLP & IR Group @ UNED

S. Aliakbary, H. Abolhassani, H. Rahmani, and B. Nobakht (2009), Web Page Classification Using Social Tags, Proc. IEEE Int'l Conf. Computational Science and Eng., vol. 4, pp. 588-593.

U. Kreßel (1999), Pairwise Classification and Support Vector Machines,Advances in Kernel Methods, pp. 255-268, MIT Press.

S. Golder and B.A. Huberman (2006), The Structure of Collaborative Tagging Systems, J. Information Science, vol. 32, no. 2, pp. 198-208.