One-Class Classification for Finding Interesting Resources in Social Bookmarking Systems

Daniela Godoy

ISISTAN Research Institute, UNICEN University
Campus Universitario, CP 7000, Tandil, Argentina
Also at CONICET, Argentina
dgodoy@conicet.gov.ar

1 Introduction

Social tagging systems have grown in popularity on the Web in the last years on account of their simplicity to categorize and retrieve shared content using open-ended tags. In sites such as Del.icio.us, Flickr or CiteULike, users annotate a variety of resources (Web pages, blog posts or pictures) using a freely chosen set of keywords rather than pre-defined categories. The rapid increase in size of communities using social sites as well as the large amount of shared content available make the discovery of relevant resources a time consuming and difficult task for users.

To alleviate this problem, the content as well as social tags associated to the resources annotated by a user can be used to build a user interest profile that, in turn, can be applied to filter further incoming information from tagging systems (e.g. RSS feeds). In social tagging systems resources receive tag assignments by members of the community, describing their content in a collective sense. Thus, it can be assumed that users are likely to be interested in additional content annotated with similar tags to the ones assigned to resources they showed interest in before.

In this work we evaluate two one-class classification algorithms as a means to identify interesting resources for users. Tag-based classifiers are learned using the Web resources users annotate and have in their personomies as positive examples of their interests. This is a special case of classification in which it is necessary to determine whether an example (resource) belongs to a target class (interesting) when only examples of the target class are given, which is known as one-class classification.
2 Dataset Description and Pre-Processing

Empirical evaluation was carried out using data collected from Del.icio.us\(^1\) social bookmarking system. From this site 50 complete personomies were gathered containing all of the user bookmarks and the corresponding tag assignments. In this collection of personomies there are users with as few as 10 and as much as 2521 bookmarks. For each Web page, all tags assigned by other members of the community were also extracted from Del.icio.us.

From the total set of resources gathered, experiments reported in this paper were performed over English-written pages. The resulting folksonomy contains with \(|U| = 50\) users, \(|T| = 233,997\) tags and \(|R| = 49,265\) bookmarks or Web pages, related by a total of \(|Y| = 128,642,112\) tag assignments. In all experiments, evaluation was carried out using a holdout strategy that split each personomy into a 66% for training and a 34% for testing. In order to make the results less dependent of the data splitting, average and standard deviation of 10 runs for each user is reported.

Since this testing set only contains interesting examples, uninteresting pages were extracted from the personomies of other users to evaluate the algorithm capacity of distinguishing uninteresting resources. This is, the testing set was created using the test set from the user and an equivalent number of Web pages gathered from a different personomy in the collection. This second personomy was randomly chosen among those presenting no resource intersection with the current user. In other words, it is assumed that two users having no common resources in their personomies do not shared interests, so that one user resources will be uninteresting to the other one.

In these experiments three different information sources were evaluated to create one-class classifiers: the full-text of resources, the full tagging activity associated to resources (all tags assigned by member of the community) and the overall top 10 tags associated to resources (the 10 more frequent tags per resource). Web page texts were filtered using a standard stop-word list and Porter stemming algorithm. To prevent syntactic mismatches due to these reasons the effect of different filtering strategies for tags was evaluated. First, original raw tags were filtered to remove the symbols (such as #, -, +, /, :, _, &, !), which allows joining compound words. Then, the remaining tags were stemmed to their morphological roots. Both frequency-based (number of users that employ a given tag) and binary representations (occurrence or non-occurrence of a given tag) of the resulting tag vectors were considered in the experiments.

3 One-class Classification

User actions of assigning tags to resources are a strong indication of relevance about its content. Consequently, positive examples of the user interests can be easily collected from folksonomies. On the contrary, identify representative negative examples or non-interesting resources is more complex since users might not tag a potentially interesting resource because of multiple reasons, such as lack of time, etc. The task of determining whether a document is interesting for a user basing training only on positive examples can be seen as a one-class classification problem. One-class classification differs in one

\(^1\) http://del.icio.us/
essential aspect from conventional classification as it assumes that only information of one of the classes, the target class, is available.

3.1 One-class SVM

SVMs (Support Vector Machines) are a useful technique for data classification, which has been shown that is perhaps the most accurate algorithm for text classification, it is also widely used in Web page classification. Schölkopf et al. [6] extended the SVM methodology to handle training using only positive information. Essentially, one-class SVM algorithm consists in learning the minimum volume contour that encloses most of the data and it was proposed for estimating the support of a high-dimensional distribution [6], given a set of training vectors $X = \{x_1, \ldots, x_l\}$ in $\mathbb{R}^n$. The aim of SVM is to train a function $f_X : \mathbb{R}^n \rightarrow \mathbb{R}$ such that most of the data in $X$ belong to the set $\mathcal{R}_X = \{x \in \mathbb{R}^n \text{ with } f_X(x) \geq 0\}$ while the volume of $\mathcal{R}_X$ is minimal. This problem is termed minimum volume set (MVS) estimation, and the membership of $x$ to $\mathcal{R}_X$ indicates whether this data point is overall similar to $X$. In this work we used LibSVM² library implementing one-class SVM.

In order to adjust the kernel for optimal results, the parameter $\gamma$ need to be tuned to control the smoothness of the boundary, i.e. large values of $\gamma$ lead to flat decision boundaries. Figure 1(a) shows the results obtained with the same classification sources by setting $\gamma$ to 0.1, the point at which the best results were achieved, and varying the value of $\gamma$ of one-class SVM classifiers [1]. The figure not only shows how higher values of $\gamma$ lead to small increases in F-measure scores but, more importantly, with values of $\gamma > 0.4$ any form of representation of social tags outperforms full-text classification. Furthermore, binary representations of top-10 tags associated to resources become the best performing among the social classification schemes.

3.2 Rocchio

Prototype-based classifiers represent each class in terms of a prototype vector in the same dimensional space as documents, making it feasible to estimate the similarity between documents and prototypes of classes. A materialization of such prototype-based classifier approach is the Rocchio algorithm which was originally developed as a method for relevance feedback in information retrieval [5]. In that context, it is applied to automatically optimize queries initially formulated by a user on the basis of relevance judgments of the retrieved documents.

The basic idea of Rocchio algorithm is to represent each document as a vector in the vector space model so that documents with similar content have similar vectors. Rocchio algorithm learns a class model by combining document vectors belonging belonging to the same class into a single prototype or centroid vector. Then, classification is done by judging the angular distance of the instance to be classified from the prototype vector of the target class. In one-class classification, if the instance similarity to the centroid of the target class is above a certain threshold, the document is considered

²http://www.csie.ntu.edu.tw/~cjlin/libsvm/
3.3 Result Comparison

Figure 2(a) summarize the performance of content and tags-based classifiers in terms of accuracy, showing that when the classifiers capability of making correct decisions is considered, tags-based classifiers outperformed full-text ones. Also, among tag-based classifiers those using for training the top 10 tags assigned to each resource were the ones with superior performance. Thus, top 10 tags offers good accuracy levels and, at the same time, an important reduction in leaning and prediction complexity given the smaller size of the dimensional space during learning. In term of F-Measure, Figure 2(b), SVM and Rocchio different results for binary and frequency representacions. Rocchio outperforms SVM when frequency vectors are involved, more likely because is a method coming from IR area, whereas SVM improves Rocchio results in binary vectors.

4 Related Works

Tag recommendation in social tagging systems had been approached in several works. However, the problem of filtering resources according to the user interest had received less attention and, to the best of our knowledge, no approaches have addressed this problem using one-class classification over social tags. Tag-based profiling consisting on tag vectors in which tag weights are given by their frequency of occurrence in the resources a user tagged had been proposed in [4]. In [3], profiles are represented by graphs in which nodes correspond to tags and edges denote relationships between them. The idea of using semantic relationships among tags in tag-based profiles has also been explored in [2]. Vatturi et al. [7] create a personalized tag-based recommender for each user consisting of two NB classifiers trained over different time-frame. One classifier
predicts the user current interest based on a shorter time interval and the other classifier predicts the user general interest in a bookmark considering a longer time interval. If any classifier predicts the bookmark as interesting, it is recommended. The user study results show that the tag-based recommender performs well with real data using tags from an enterprise social bookmarking system. In the work presented in this paper, one-class classifiers can be seen as tag-based profiles for users.

5 Conclusions

In this paper experiments evaluating two one-class classification approaches for identification of interesting resource in social tagging systems were reported. One-class SVM and Rocchio classifiers were used to learn the user interests from sources such as the full-text of resources and social tags associated to them. Experimental results obtained with a set of personomies extracted from Del.icio.us site showed that tag-based classifiers outperformed full-text classification, particularly whit binary representations, and SVM reached better performance values than Rocchio.

References


