Chapter 8
Exploiting Social Annotations for Resource Classification

Arkaitz Zubiaga
NLP & IR Group, UNED, Spain

Víctor Fresno
NLP & IR Group, UNED, Spain

Raquel Martínez
NLP & IR Group, UNED, Spain

ABSTRACT

The lack of representative textual content in many resources suggests the study of additional metadata to improve classification tasks. Social bookmarking and cataloging sites provide an accessible way to increase available metadata in large amounts with user-provided annotations. In this chapter, the authors study and analyze the usefulness of social annotations for resource classification. They show that social annotations outperform classical content-based approaches, and that the aggregation of user annotations creates a great deal of meaningful metadata for this task. The authors also present a method to get the most out of the studied data sources using classifier committees.

INTRODUCTION

Resource classification is the task of labeling resources with their corresponding categories from a predefined taxonomy. Resource classification is of vital importance for information management and retrieval tasks, and for assisting the semi-automatic development of categorized directories.

DOI: 10.4018/978-1-61350-513-7.ch008

In the case of web pages, it is also essential to focused crawling, and to topic-specific web link analysis, among others. It can also help improve search results when it is applied to organizing ranked results.

To carry out this kind of tasks automatically, the textual content is commonly used to represent the resource to classify. Many times, the lack of representative content makes it insufficient, though (Qi & Davison, 2009). In this way, social

bookmarking sites present an accessible way 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[1], StumbleUpon[2], LibraryThing[3] and Diigo[4], among others, allow their users to add information to a web page, collecting hundreds of thousands of annotations per day (Heymann et al., 2008). As a result, a global community of volunteer users creates a huge repository of described resources that can ease their subsequent retrieval. Until now, the use of social annotations for resource classification tasks has remained relatively unexamined. The little work performed so far has shown the suitability of social tags for this kind of tasks. Nonetheless, the study of the optimal representation based on social tags, and the use of social annotations other than social tags, are still unexplored.

In this chapter, we study and analyze the use of metadata extracted from social bookmarking and cataloging sites to classify a set of annotated resources. We perform the experiments with two different types of resources: web pages and books. We find two types of social annotations to be applicable and useful for resource classification: tags and comments provided by end users. We propose a way to represent each kind of annotation, and we present a method to outperform their results by means of combining different data using classifier committees.

Next, in Section Background, we describe the nature of social annotations and the existing types. We continue in Section Related Work presenting earlier research in the literature. After that, we detail the settings of our experiments as well as the datasets we used, to continue in Section Results with the analysis of the results. We discuss them in Section Discussion. Finally, we conclude with our thoughts and future work.

BACKGROUND

Social bookmarking and cataloging sites allow users to save and annotate their favorite resources, sharing them with the community. These annotations are provided in a collaborative way, so that it makes possible a large amount of metadata to be available for each resource. Going into further details on these metadata, different kinds of user-generated annotations can be defined:

- **Tags:** Keywords defining and characterizing a resource are known as tags. 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 $T_j = \{w_1\ t_1,...,w_n\ t_n\}$, where $w_1,...,w_n \leq k$. The resulting organization from users' tagging activity is known as a folksonomy. In most of the social bookmarking systems, there are no constraints on the keywords users can set as tags. The use of tags was originally suggested to make easier the later search and retrieval of relevant documents. Most of the research in this field has focused on the study of dataset properties (Ramage et al., 2009), the analysis of usage patterns of tagging systems (Golder & Huberman, 2006), and the discovery of hidden semantics in tags (Yeung et al., 2008). Incorporating social annotations with document content and other sources of information is a natural idea (Zhou et al., 2008), especially when trying to improve information management tasks.
- **Notes or descriptions:** Free text describing a resource is known as a note or description.
- **Reviews:** A review is a free text valuating a web page. Even though this kind of annotations can initially look subjective and non-descriptive, users tend to mix descriptive texts with opinions.