Harnessing Folksonomies for Resource Classification

Arkaitz Zubiaga
City University of New York

Arkaitz Zubiaga is a post-doctoral researcher at the Queens College of the City University of New York. Arkaitz earned his Ph.D. in Computer Science in July 2011 under the supervision of Prof. Raquel Martínez and Prof. Víctor Fresno at the NLP&IR Group of the UNED University in Madrid, Spain. He has been involved in the SigWeb community by contributing with publications and participating in several conferences including DocEng, Hypertext and CIKM. His main research interests lie in the fields of machine learning, social media mining, natural language processing and information retrieval. In his post-doctoral research, he has continued to extend his research in mining social media. Specifically, he has explored the power of real-time social streams as a means to characterize and discover information about current events and trends in social media.

DOI: 10.1145/2246063.2246067 http://doi.acm.org/10.1145/2246063.2246067

In our daily lives, organizing resources into a set of categories is a common task. Organizing resources into categories makes searching through those resources easier by limiting the focus to a specific category. Limiting the focus significantly reduces the amount of information one must search. Categorization becomes more useful as the collection of resources increases, when managing resources becomes more and more difficult if they are not organized appropriately. Large collections like those made up by books, movies, and web pages, for instance, are usually cataloged in libraries, organized in databases and classified in directories, respectively. However, the usual largeness of these collections requires a vast endeavor and an outrageous expense to organize manually.

Recent research is moving towards developing automated classifiers that reduce the increasing costs and effort of the task. Most of the research in this field has focused on self-content, where the publisher is the only author, as a data source to discover the aboutness of the resource. Self-content presents the problem that it is not always representative enough, and sometimes it is difficult to access depending on the type of resource. Little work has been done analyzing the appropriateness of and exploring how to harness the annotations provided by users on social tagging systems as a data source. Users on these systems save resources as bookmarks in a social environment by attaching annotations in the form of tags. It has been shown that these tags facilitate retrieval of resources not only for the annotators themselves but also for the whole community. Likewise, these tags provide meaningful metadata that refers to the content of the resources.

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.com, LibraryThing.com and GoodReads.com collect user annotations in the form of tags on their respective collections of resources. User-provided annotations can be useful as a data source by providing meaningful information that can help infer the categorization of the resources. Our hypothesis is that these large collections of annotations can enhance the automated resource classification task in a noticeable manner. By providing tags, users are creating their own categorization system for the given resource. The aggregation of users in an active community can create many bookmarks, tags, and therefore annotated resources. With more users contributing bookmarks and tags to these systems, the more accurately these resources can be annotated. 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. In this context, where users are providing large amounts of metadata, our challenge lies in making the most of them in order to enhance resource categorization tasks.

In this thesis, we deal with the utilization of these user-provided tags in search of the most accurate classification of resources as compared to expert-driven categorizations. After performing a set of experiments to choose a suitable classifier for this kind of task, we explore social annotations looking for a way to best use them. For this purpose, we have created three large-scale datasets including tagging data for resources from well-known social tagging systems: Delicious, LibraryThing, and GoodReads. Those resources are accompanied by categorization data from sound and consolidated expert-driven taxonomies. From these resources the appropriateness of social tags for predicting categories can be evaluated.

Specifically, we conduct three groups of experiments to find out how to make best use of social tags for classification. Firstly, we study several ways of representing the massive number of social tags by amalgamating the contributions of large communities of users. We analyze their suitability for the classification task, upon both broader top level categories and narrower deep level categories. Secondly, we explore the nature, characteristics, and distributions of tags in folksonomies, in order to determine how the settings of each system affect the tagging behavior and the usefulness of tags for the classification task. We go deeper into tag distributions by analyzing the usefulness of weighting schemes based on inverse frequency values. And thirdly, using state-of-the-art user behavior detection processes, we identify users on social tagging systems who better fit the classification task.

To the best of our knowledge, this thesis presents the first research work performing actual classification experiments utilizing social tags. By exploring the characteristics and nature of these systems and the underlying folksonomies, this thesis sheds new light on the way of getting the most out of social tags for the sake of automated resource classification tasks. Therefore, we believe that the contributions in this work are of utmost interest for researchers in the field, as well as for the scientific community in order to better understand these systems and further utilize the knowledge garnered from social tags.