Chapter 7.14
Measuring Similarity of Interests for Clustering Taggers and Resources

Christo Dichev
Winston-Salem State University, USA

Jinsheng Xu
NC A&T, USA

Darina Dicheva
Winston-Salem State University, USA

Jinghua Zhang
Winston-Salem State University, USA

ABSTRACT

Collaborative tagging systems demonstrate the potential to generate collectively built organization structures forming the basis for social navigation and shared knowledge. The effectiveness of these systems for finding and re-finding information depends not only on the created tag structures but also on the ability to identify similar users. In this article, we present our study on measuring user similarity based on shared interests, utilizing data from del.icio.us. The authors propose several methods for measuring similarities aimed at clustering tags and users. They also report our initial results related to implicit grouping of tags and users.

INTRODUCTION

The term Web 2.0 refers to the latest generation of community-centered systems that focus on collaboration and/or integration of multiple disparate sources of information into aggregate views. Among the most popular Web 2.0 or social software systems are the collaborative tagging sites (eg. O'reilly, 2005). As a result of the recent popularity of collaborative tagging systems, we are witnessing a steady growth of the tagging communities and of the volume of user-generated content and metadata (Guy &Tonkin, 2006;, Donkar, 2007). Thus these systems are setting a technological infrastructure for harvesting social knowledge. From a practical perspective, it is

important to know how this trend will evolve and how it will scale, when its usage base goes beyond the early adopters. Recent studies suggest that as tagging communities grow, the added content and metadata become harder to manage due to the increased content diversity. Using entropy as their metric, (Chi and Mytcowicz, 2007) conclude that: (i) social tagging is becoming less efficient; (ii) systems based on social tagging are becoming harder to navigate; (iii) the users of social tagging systems are becoming more diverse. The authors interpret this phenomenon as a sign that tags are becoming less effective in describing documents. One intuitive explanation is that users start using more (and more specific) tags in order to cope with the information overflow. This leads to decreasing tags' descriptive precision: previously useful tags become too generic and users respond to this tendency by using ever more specific tags.

How can we cope with entropy in collaborative tagging context? The collaborative nature of tagging relies on the fact that an individual's interests normally overlap with the interests of some other users, which implies use of similar resources and tags. This fact can be exploited further in two aspects: for identifying communities with shared interests and as a filtering mechanism for reducing irrelevant resources and tags. To this end we propose partitioning of tagging communities into explicit groups of shared interests. Dividing a community into groups would result in splitting up their bookmarks (resources). The benefit of the latter would be that users can navigate a smaller set of resources (those belonging to a particular group). For example, instead of exploring all resources tagged with "memory", a user can explore only those tagged with "memory" within their group (e.g. computer, memory, cache). The feasibility of this approach obviously depends on the premise that we can identify groups with different interests within a tagging community.

To test our hypothesis that tagging communities cluster around common tags and resources, we studied tags and bookmarks extracted from del.icio.us. In this article we present our study on measuring user similarity based on shared interests, utilizing the extracted data. We propose several methods for measuring similarities aimed at clustering tags and users and report the results of the study on grouping users with similar interests utilizing tags similarity.

Social Tagging and User Communities

Because of their lack of predefined taxonomic categories, social tagging systems rely on shared social, conceptual, and vocabulary structures evolving in parallel with the emerging user communities. In a collaborative tagging system users can find tags, resources, and taggers related to their interests thus becoming involved in and benefiting from the community.

Recognized downfalls of collaborative tagging include ambiguity, inability to handle hierarchies, meta-noise due to inappropriate tags assigned to resources, or tags that are assigned with a specific context in mind (Guy & Tonkin, 2006).. For instance, when looking for a description of a "golden delicious apple", a search for resources tagged with "apple" could return descriptions of different "apple fruits" as well as news from "Apple Macintosh". Further, with the lack of a hierarchical organizing structure, tags can grow to disorderly proportions even for a single user. One possible way to deal with these problems is to provide grouping based on common features or contextual factors. When a given collection of users, resources, and tags grows to a sufficient level, similar users can be grouped together for recommendation purposes, as well as for studying emerging communities. For example, analyses of bookmark co-occurrence may enable identifying groups of users with similar interests based on the assumption that users with significant overlaps in their bookmark collections have overlapping interests.