Chapter 3
Detecting Trends in Social Bookmarking Systems: A del.icio.us Endeavor

Robert Wetzker
Technische Universität Berlin, Germany

Carsten Zimmermann
University of San Diego, USA

Christian Bauckhage
University of Bonn and Fraunhofer IAIS, Germany

ABSTRACT
The authors present and evaluate an approach to trend detection in social bookmarking systems using a probabilistic generative model in combination with smoothing techniques. Social bookmarking systems are gaining major interest among researchers in the areas of data mining and Web intelligence, since they provide a large amount of user-generated annotations and reflect the interest of millions of people. Based on a vast corpus of approximately 150 million bookmarks found at del.icio.us, the authors analyze bookmarking and tagging patterns and discuss evidence that social bookmarking systems are vulnerable to spamming. They present a method to limit the impact of spam on a trend detector and provide conclusions as well as directions for future research.

INTRODUCTION
Social bookmarking systems, such as del.icio.us, StumbleUpon or CiteULike, have been very successful in the recent past. Their success originated from members’ ability to centrally store bookmarks on the web. However, with the coming of age of these services, the perceived value of social bookmarking systems shifted towards the underlying social aspects, such as trend indication, advanced web search or recommendation functionality. These services are an invaluable source of information, since they provide a vast amount of user-generated annotations, such as
Detecting Trends in Social Bookmarking Systems

tags, and reflect the interests of millions of users. One social aspect of these systems derives from the fact that resources, in general web pages, are tagged by the community and not by the creator of content alone, as in other services like Flickr or YouTube (Marlow, Naaman, Boyd, & Davis, 2006). This characteristic, called collaborative tagging, was shown to provide relevant metadata (Heymann, Koutrika, & Garcia-Molina, 2008) and is expected to boost the semantic quality of labels (Surowiecki, 2004).

One of the first and most popular social bookmarking systems is del.icio.us. Because of its early acceptance in the market, the vast growth over the past five years and easy data accessibility, del.icio.us represents a suitable case for analyzing the characteristics of social bookmarking communities. Figure 1 shows the del.icio.us main page that lists the currently popular bookmarks and tags. Though this article mainly examines del.icio.us, we conjecture that the results presented here also apply for other social bookmarking services and collaborative tagging systems in general.

As shown by Heymann, Koutrika, and Garcia-Molina (2008), trends within the del.icio.us tagging and bookmarking behavior strongly correlate with real world events. This characteristic makes bookmarking services a valuable source for trend detection and creates new opportunities in areas such as product tracking or marketing. We will investigate the nature of trends within social bookmarking services and present a probabilistic method for the automated detection of trends within the del.icio.us community. In order to succeed in the trend detection task, we further present a method which limits the influence of spam users as their frequent appearance in the data and their anomalous behavior would, otherwise, strongly interfere with any trend patterns.

Figure 1. The del.icio.us main page (September 2008)