Chapter 6
Probabilistic Models for Social Media Mining

Flora S. Tsai
Nanyang Technological University, Singapore

ABSTRACT

This paper proposes probabilistic models for social media mining based on the multiple attributes of social media content, bloggers, and links. The authors present a unique social media classification framework that computes the normalized document-topic matrix. After comparing the results for social media classification on real-world data, the authors find that the model outperforms the other techniques in terms of overall precision and recall. The results demonstrate that additional information contained in social media attributes can improve classification and retrieval results.

INTRODUCTION

The rapid growth of technology has led to information overload from online such as blogs (Chen, Tsai, & Chan, 2007), social networks (Tsai, Han, Xu, & Chua, 2009), mobile information (Tsai et al., 2010), and Web services (Tsai et al., 2010). Novelty mining can help solve the problem of information overload by retrieving novel yet relevant information, based on a topic given by the user (Ng, Tsai, & Goh, 2007; Ong, Kwee, & Tsai, 2009), and can be used to solve many business problems, such as in corporate intelligence (Tsai, Chen, & Chan, 2007) and cyber security (Tsai, 2009; Tsai & Chan, 2007). Although users can retrieve all the novel documents, each document still needs to be read to find the novel sentences within these documents (Tsai & Chan, 2011). Therefore, to serve users better, later studies of novelty mining were performed at the sentence level (Kwee, Tsai, & Tang, 2009; Tang & Tsai, 2009; Tang, Tsai & Chen, 2010; Tsai, Tang, & Chan, 2010; Zhang & Tsai, 2009b). Furthermore, the Web is changing from a datacentric Web into Web of semantic data and Web of services (Yee, Tiong, Tsai, & Kanagasabai, 2009). The use of Web services has significance in the business domain, where they are used as means of communication or exchanging data between businesses and clients (Kwee & Tsai, 2009).

DOI: 10.4018/978-1-4666-2157-2.ch006

Previous studies on social media mining (Tsai, Chen, & Chan, 2008; Liang, Tsai, & Kwee, 2009) use existing Web and text mining techniques without consideration of the additional dimensions present in the social media. Because of this, the techniques are only able to analyze one or two dimensions of the blog data (Tsai & Chan, 2010). In this paper, we propose unsupervised probabilistic models for mining the multiple dimensions present in social media. The models are used in the novel social media classification framework, which categorizes social media according to their most likely topic.

Problem Definition

This paper addresses the problem of multidimensional social media mining, which is a big challenge in the data mining community. Although blogs may share many similarities to Web and text documents, existing techniques need to be reevaluated and adapted for the multidimensional representation of blog data, which exhibit attributes not present in traditional documents. The proposed techniques aim to leverage multiple blog dimensions of authors and links to improve the results of mining information from blog data and to address and solve the problem of mining information from blog data using multiple dimensions of social media.

RELATED WORK

Related work on social media mining include techniques that focus on sentiment or opinion mining, or judging whether a particular blog post is negative, positive, or neutral to a particular object. One of the main tasks in the Text Retrieval Conference (TREC) Blog Track was the Opinion Retrieval Task, which involved finding blog posts that express an opinion about a given topic (Ounis et al., 2006; Macdonald, Ounis, & Soboroff, 2007).

Other studies attempt to filter out spam blogs, or splogs, which can greatly misrepresent any estimations of the number of blogs posted. Previous work in splog detection include splog detection using self-similarity analysis on blog temporal dynamics (Lin et al., 2007) and Support Vector Machines (SVMs) to identify and splogs (Kolari, Finin, & Joshi, 2006).

Other related work in social media mining is topic distillation, which was the second main task in TREC Blog 2007 (Macdonald, Ounis, & Soboroff, 2007). The blog distillation, or feed search, task focuses on blog feeds, which are aggregates of blog posts. Blog distillation task searches for a blog feed with a principle, recurring interest in topic t. For a given topic t, systems should suggest feeds that are principally devoted to t over the timespan of the feed, and would be recommended to subscribe to as an interesting feed about t (Macdonald, Ounis, & Soboroff, 2007). This task has direct relevance to the problem of searching for blogs to which a user may wish to subscribe. As many blog posts are inherently noisy, finding the relevant feeds is not a trivial problem.

Other related studies include a joint probabilistic document model (PHITS) (Cohn & Hofmann, 2001) which modeled the contents and inter-connectivity of document collections. A mixed-membership model (Erosheva, Fienberg, & Lafferty, 2004) was developed in which PLSA was replaced by LDA as the generative model. The Topic-Link LDA model (Liu, Niculescu-Mizil, & Gryc, 2009) quantified the effect of topic similarity and community similarity to the formation of a link. The citation-topic (CT) model was proposed in (Guo et al., 2009) for modeling linked documents that explicitly considers the relations among documents.