Potential Topics Discovery from Topic Frequency Transition with Semi-supervised Learning

Yoshiaki Yasumura, Hiroyoshi Takahashi, and Kuniaki Uehara

Shibaura Institute of Technology, Japan
Kobe University, Japan
yasumura@shibaura-it.ac.jp,
{takahashi,uehara}@ai.cs.kobe-u.ac.jp

Abstract. This paper presents a method for potential topic discovery from blogsphere. A potential topic is defined as an unpopular phrase that has potential to spread through many blogs. To discover potential topics, this method learns from topic frequency transitions in blog articles. Though this learning requires sufficient amount of labeled data, labeled data is costly and time consuming. Therefore this method employs a semi-supervised learning to reduce labeling cost. First, this method extracts candidates of potential topics from categorized blog articles. To detect potential topics from the candidates, a classifier is built from topic frequency transition data. Experimental results with real world data show the effectiveness of the proposed method.

Keywords: Web mining, potential topic, semi-supervised learning, topic frequency transition.

1 Introduction

Blog is a media that individual can easily provide commentary and news on a particular subject. Since the number of bloggers increases rapidly, a lot of blog articles are updated daily. Thus, blog articles reflect the trend of the real world. This fact enables us to analyze market trend by monitoring information on blogs[1,2,3].

So far, a lot of methods are proposed for monitoring and analyzing information on blogs. One of the most popular methods is detecting a burst of a word in a document stream of blogs [4,5,6,7,8,9]. Since burst words are viewed as hot topics in the blogsphere, detecting burst provides market analyzer the trend in the blogsphere. However, most burst topics are not valuable information from the view point of marketing because they are already popular in the blogsphere. Valuable topics are described in a few blogs and have a potential to spread through many blogs. We call such topics "potential topics". The system that can discover potential topics as early as possible is required for marketing.

For discovering potential topics, Kosaka et al. proposed a method that extracts potential topics from blogsphere by learning from topic frequency transition[10]. However, this method adopts supervised learning that requires sufficient amount

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 477–486, 2012.

of labeled data. Labeled data is costly and time consuming. In addition, it is difficult to label topics because we cannot know when the learner can recognize the topic is a potential topic.

In this paper, we develop a system for discovering potential topics from the blogsphere by semi-supervised learning. Semi-supervised learning requires a few labeled data and a lot of unlabeled data. This system first extracts candidates of potential topics by filtering general phrases. Next a predictor for detecting potential topics is built by semi-supervised learning from the data of topic frequency transition in the blogsphere.

2 Potential Topics Discovery

In this section, we present a method for discovering potential topics. First, we describe potential topics and their usefulness. Second, we present a method for categorizing blog articles and building a predictor for detecting potential topics.

2.1 Potential Topics

Valuable information for marketing can create or capture new demand. The system that can detect such information as early as possible is required for marketing. One of the systems is topic extraction by detecting a burst of a word. A burst of a word is defined as sharp increase in frequency of the word. However, most burst topics are not valuable information because they are already popular. Fig. 1 shows an example of burst detection. This graph charts the topic frequency transition of the phrase "subprime loan problem" in the blogsphere. From the graph, this phrase is first described in blogs in March, 2007. After that, the phrase increased in frequency gradually, and burst in October, 2007. If burst of the phrase is detected at that time, it is not valuable information. This is because subprime loan problem was already reported on TV and newspaper, and stock price began to fall at that time. However, the phrase "subprime loan problem" is valuable information if it was detected before bursting. In order to detect the topics earlier, we try to extract potentiality of the phrase by analyzing the topic frequency transition before bursting. To do this, we built a predictor for detecting potential topics. The predictor learns from the topic frequency transition of the old data of the bursted topics. For creating the predictor, this method learns from topic frequency transitions in blog articles. Though this learning requires sufficient amount of labeled data, labeled data is costly and time consuming. Besides it is difficult to label topics because we cannot know when the learner can recognize the topic is a potential topic. So this method employs a semi-supervised learning to reduce labeling cost.

In order to discover potential topics as early as possible, we extract specialists who can describe potential topics of the category earlier in their blogs. To extract them, we classify blog articles into some category. By analyzing the topic frequency transition in each category, we can detect potential topics earlier. Fig. 2 shows an example of topic frequency transition of the phrase "subprime loan