Post Recommendation in Social Web Site

Long Wang, Justus Bross, and Christoph Meinel
Hasso Plattner Institute, University Potsdam
14482, Potsdam, Germany
{long.wang,justus.bross,christoph.meinel}@hpi.uni-potsdam.de

Abstract. Web 2.0 applications attract more and more people to express their opinions on the Web in various ways. However, the explosively increasing information in social web sites requires an effective mechanism to timely filter and summarize social common interest, and the moderator needs this mechanism as well to recommend the proper posts and guide public discussions. In this paper, we discuss the problem of recommending post in online communities: we firstly cluster the posts in groups based on their semantic relations, then filter the potential clusters by computing the cluster’s support, and finally select the recommended posts as content representatives considering global and local support from each clusters. We compare different feature selections between tags, keywords and topics on cluster formation, and discuss their differences. The human judgement in our experiment shows that the recommendation based on marked tags is much more effective and concise than those on keywords and hidden topics.

Keywords: Post recommendation, Clustering, Global Support, Local Support.

1 Introduction

Web 2.0 technologies supply richer interaction possibilities for end users to compose their posts and discuss common topics in online communities. The users publish or annotate personal content without much central governance in various forms, from plain text to rich multimedia, which could generate vast amount of noisy information. Moreover, the content on social communities tends to have a shorter lifespan because much of it focuses on an ongoing real-world event or a current ”hot” topic, and public interest in such content and subsides rapidly over time. A post is a basic entry in a social web site on a particular subject from textual commentary, description, to graphics or videos. Thus several ways are used to facilitate users to find the interesting posts: using an internally embedded searching engine, selecting the posts in reverse chronological order in front page, clicking the ones listed in ”most favorite” or ”top N rated”, or making a choice by entering the target categories or tag cloud.

In this paper, we propose a method on post recommendation in online communities. Using a case study of IT-Gipfelblog (http://it-gipfelblog.hpi-web.de), which is a weblog to discuss ICT (Information & Communications Technology)
related topics. The process of discovering recommended posts is briefly described in the followings: we firstly cluster the posts based on their contents; then only the clusters having a support over a defined threshold are treated as the potentials for supplying recommended posts; and finally from each potential clusters, the posts having global and local support are ranked as part of recommendation list. The clustering step guarantees the clear conceptual distance between recommended posts and the global or local support ensures the confidence on the support from users. We try using different features such as tags, keywords and topics in clustering posts and evaluate their differences on the final recommended post lists.

The rest of the paper is structured as follows: in section 2 we describe the features related with a post in social web site. Section 3 provides the methods of clustering and representative selection based on support. After that, we discuss our experiment results in section 4. Finally, conclusion and future work are given in section 5.

2 Features Related for a Post

Selecting proper features to represent every post is the necessary work in post recommendation. Here, we illustrate the related features for a post. Generally, the features are classified into two categories: content-related and usage-related.

2.1 Content-Related: Keywords, Tags and Topics

The basic methodology proposed in text modeling is to reducing each document in the corpus to a vector of real numbers, each of which represents the importance of a content feature to that document. In online communities, the features related to the content of a post can be tags, keywords or topics. Most of previous work on text modeling is keyword based. Keywords are stemmed and filtered with the weights showing their importance to a post, and the weight is usually measured by $tf$ or $tf \times idf$ approaches.

Tag marking is a prominent feature in social web sites, by which the users annotate their personal understandings and preferences on a post. Recently, discovering tag-based social interest attracted much attraction [1,2]. [2] found user-generated tags are condensed and consistent with the web content and more concise and closer to human understanding. Tags are especially indispensable to represent semantics for the posts with only videos. However, tags depend highly on the activities of users, though they have a higher-level abstraction on the content. Compared with tags, the keywords related with a post could have a large scale, but each of them owns a measurable weight.

Different from tags and keywords which are observable, the topics or contexts are latent. Latent semantic analysis (LSA) was discussed to discover the topics hidden in corpora and proven to be efficient for text modeling [3,4]. Recently LSA is used for detecting online reviews or opinions [5,6]. Topic feature has two advantages compared with tags and keywords: one is that topics are