Revealing Learner Interests through Topic Mining from Question-Answering Data

Yijie Dun, Mathematics and Computer Institute, Northwest University for Nationalities, Lanzhou, China
Na Wang, School of Electronics and Communication Engineering, Zhengzhou University of Aeronautics, Zhengzhou, China
Min Wang, School of Informatics, Guangdong University of Foreign Studies, Guangzhou, China
Tianyong Hao, School of Informatics and Collaborative Innovation Center for 21st-Century Maritime Silk Road Studies, Guangdong University of Foreign Studies, Guangzhou, China

ABSTRACT

In a question-answering system, learner generated content including asked and answered questions is a meaningful resource to capture learning interests. This paper proposes an approach based on question topic mining for revealing learners’ concerned topics in real community question-answering systems. The authors’ approach firstly preprocesses all questions associated with learners. Afterwards, it analyzes each question with text features and generates a weight feature matrix using a revised TF/IDF method. In order to decrease the sparsity issue of data distribution, the authors employ three concept-mapping strategies including named entity recognition, synonym extension, and hyponym replacement. Applying an SVM classifier, their approach categorizes user questions into representative topics. Three experiments are conducted based on a TREC dataset and an actual dataset containing 1,120 questions posted by learners from a commercial question-answering community. Results demonstrate the effectiveness of the method compared with conventional classifiers as baselines.

KEYWORDS
Classification, Learner Interest, Question Answering, Topic Mining

1. INTRODUCTION

User interest mining has attracted more and more attention in recent years in the fields of data mining, particularly in topic mining, information retrieval, text classification, citation analysis and social network analysis (Zhang & Sun, 2012). Different from an information retrieval system, acquiring a relevant answer from a Question-Answering (QA) system is regarded as requiring more complex Natural Language Processing (NLP) techniques rather than solely document retrieval. Therefore, it is regarded as the next step beyond search engines. As a typical QA system, Community Question-Answering (CQA) system, also as User-Interactive Question-Answering (UIQA) system, has drawn a lot attention in recent decades (Hao et al., 2011; Hao & Agichtein, 2012).

In such CQA systems, a typical scenario is that a user may ask concerned questions and post answers to his/her interested questions. These posted questions or answered questions can potentially reflect users’ interests (Ren et al., 2015) since most questions are closely related to his/her career, hobbies, business, daily life, etc. Michelson et al. (2010) addressed that discovering the topics of interest for a particular Twitter user can be conducted by generating “topic profiles” of Twitter users based upon what they tweet about. White et al. (2009) reported that user interests could be predicted
by local context information and topical interest may be less dynamic. Therefore, the question topics can be potentially used as the representations of users’ interests.

It is a meaningful task to mine those question topics over time to capture users’ interests. There is much research on leveraging user queries to model user interests for improving personalized search (Teevan, 2005; Liu, 2002; Dou et al., 2007; Sun et al., 2005). Specifically, by utilizing user interests, a CQA system can direct users to the topics that they are likely interested in for enhancing user experience. On the other hand, it helps identify potential problem-solvers or recommend questions to users who have expertise or are active in answering questions on these topics by matching questions with interested topics. These traits can potentially help speed up the flow of the question-answering process.

Particularly, most users are learners in certain types of CQA systems. One example is Hujiang1 - a language learning community that has more than 90 million registered learners. Mining users' concerned topics is substantially helpful in understanding their interested content for enhancing learning experience. This also helps solve their facing problems or difficulties during their learning process.

In this paper, we propose a new approach empowered by preference distribution to capture learners’ interests by analyzing the topics from their asked questions or answered questions. Regarding each question as a short text, the research problem can be converted into a text topic classification problem. To that end, instead of mapping questions to certain learners directly, we firstly classify all questions into topics using multi-class SVM enhanced by our proposing weight matrix generation and three concept mapping strategies. Afterwards, we collect questions posted by the learners to find out what topics the questions belong to. These questions are further grouped by their topics forming a ranked topic list to reveal the preference distribution of the learner. With the topic preferences of all learners, a QA system is thus able to understand each learner’s interests and allocate incoming questions to a specific group of learners more wisely.

The paper is organized as follows: In section 2, we briefly survey the research of user interest mining as necessary background knowledge. We then present the framework of our approach, weight matrix generation, concept mapping strategies, and SVM classification in section 3. The experiments, results and discussions are presented in section 4, while section 5 summarizes the work.

2. RELATED WORK

With the popularization of search engines, Question answering (QA) is an emerging information service aiming to find most relevant answers with respect to user queries. Compared with search engine, QA is preferable since direct and exact answers can be expected, instead of a long list of ranked, possibly relevant but usually long documents (Hao & Qu, 2016). Among them, Community Question-Answering (CQA) systems such as Yahoo! Answers2, are becoming popular online information services. With more and more users participating, CQA systems are serving on many aspects from general knowledge acquisition to specific software learning. For example, Baidu Knows3 has millions of active users every day and solved 3.58 billion user questions as of February 2015. Their posted questions and answers in turn could be good resources for understanding users’ interests.

In such kinds of personalized information service systems, there are several types of user interest modeling methods. Demonstration modeling is a method that relies on users to provide the demonstrations of relevant interests. In practice, this type of method requires users to mark or annotate data in order to obtain corresponding demonstrations for automated modeling. Later on, there is a user model constructed automatically according to a user’s usage/browsing behaviors, which can
Distance Education in Small Island Nations
www.igi-global.com/chapter/distance-education-small-island-nations/11821?camid=4v1a