Chapter 15
Internet Forums:
What Knowledge can be Mined from Online Discussions

Mikołaj Morzy
Poznan University of Technology, Poland

ABSTRACT

An Internet forum is a web application for publishing user-generated content under the form of a discussion. Messages posted to the Internet forum form threads of discussion and contain textual and multimedia contents. An important feature of Internet forums is their social aspect. Internet forums attract dedicated users who build tight social communities. There is an abundance of Internet forums covering all aspects of human activities: politics, sports, entertainment, science, religion, leisure, hobbies, etc. With large user communities forming around popular Internet forums it is important to distinguish between knowledgeable users, who contribute high quality contents, and other types of users, such as casual users or Internet trolls. Therefore, social role discovery becomes an important issue in discovery of valuable knowledge from Internet forums. This chapter provides an overview of Internet forum technology. It discusses the architecture of Internet forums, presents an overview of data volumes involved and outlines technical challenges of scraping Internet forum data. A broad summary of all research conducted on mining and exploring Internet forums for social role discovery is presented. Next, a multi-tier model for Internet forum analysis (statistical analysis, index analysis, and network analysis) is introduced. Social roles are automatically attributed to Internet forum users based on egocentric graphs of user activity. The issues discussed in the chapter are illustrated with real-world examples. The chapter concludes with a brief summary and a future work agenda.

INTRODUCTION

In this section a brief introduction to the problem of mining Internet forums is presented. Introduction begins with defining what data mining is and what types of methods are commonly employed to discover knowledge in large repositories of data. Next, the description of Internet forums, a new technology enabling social conversations in the Web 2.0 era is presented.

DOI: 10.4018/978-1-60960-067-9.ch015

Mining Knowledge from Data

Contemporary information systems contain limitless volumes of data. Valuable knowledge is hidden in these data under the form of trends, regularities, correlations, and outliers. Traditional querying models utilized by database systems or data warehouses are not sufficient to extract this knowledge. The value of the data can be greatly increased by adding means to automatically discover useful knowledge from large volumes of gathered data. Recent advances in data capture and data harvesting further increase the amount of data which are continuously loaded into contemporary database systems. Unfortunately, the advances in data gathering techniques are not followed by the increased ability to process and utilize the data. The amount of data to be processed grows quicker than the ability to process it. Therefore, advanced systems are required to automatically process very large amounts of data and acquire useful knowledge from the data self-reliantly. Data mining is the discipline which aims at "… the discovery and extraction of useful, previously unknown, non-trivial, and ultimately understandable patterns from large databases and data warehouses" (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996). Also brings together databases, decision support systems, machine learning, artificial intelligence, statistics, data visualization, and several other disciplines. Data mining uses different models of knowledge to present patterns discovered in raw data. These models include, but are not limited to, association rules, cyclic rules, characteristic and discriminant rules, classifiers, decision trees, sequential patterns, clusters, time series, and outliers. In parallel, numerous algorithms have been developed to discover and maintain patterns.

Data mining methods can be generally divided into two classes: Predictive tasks and Descriptive tasks. Predictive tasks apply algorithms and techniques to discover hidden patterns in the data and, based on discovered regularities, to provide predictive information which can be used to infer unknown values of attributes or to forecast future behavior. An example of a predictive task is the identification of target customer groups, customer retention analysis, prediction of the future behavior of customers, etc. Descriptive tasks aim at the discovery of patterns which can be used to describe the existing data concisely and to capture general data properties. A typical example of a descriptive task is the discovery of similar customer groups, the discovery of groups of products often purchased together, or the identification of outliers in a dataset. A data mining technique used to discover the hidden knowledge in social structures formed in online Internet forum communities is presented in this chapter.

Internet Forums as New Means of Communication

An Internet forum is a web application for publishing user-generated content under the form of a discussion. Usually, the term *forum* refers to the entire community of users. Discussions considering particular subjects are called *topics* or *threads*. Internet forums (The Latin plural *fora* may also occasionally be used) are sometimes called web forums, discussion boards, message boards, discussion groups, or bulletin boards. Internet forums are not new to the network community. They are successors of tools such as Usenet Newsgroups and Bulletin Board Systems (BBS) that were popular before the advent of the World Wide Web. Messages posted to a forum can be displayed either chronologically, or using threads of discussion. Most forums are limited to textual messages with some multimedia content embedded (such as images or flash objects). Internet forum systems also provide sophisticated search tools that allow users to search for messages containing search criteria, to limit the search to particular threads or sub-forums, to search for messages posted by a particular user, to search within the subject or body of the post, etc.