Chapter VII
Social Network Mining from the Web

Yutaka Matsuo
Japanese Institute of Advanced Industrial Science and Technology, Japan

Junichiro Mori
DFKI GmbH, Germany

Mitsuru Ishizuka
University of Tokyo, Japan

ABSTRACT

This chapter describes social network mining from the Web. Since the end of the 1990s, several attempts have been made to mine social network information from e-mail messages, message boards, Web linkage structure, and Web content. In this chapter, we specifically examine the social network extraction from the Web using a search engine. The Web is a huge source of information about relations among persons. Therefore, we can build a social network by merging the information distributed on the Web. The growth of information on the Web, in addition to the development of a search engine, opens new possibilities to process the vast amounts of relevant information and mine important structures and knowledge.

INTRODUCTION

Social networks play important roles in our daily lives. People conduct communications and share information through social relations with others such as friends, family, colleagues, collaborators, and business partners. Social networks profoundly influence our lives without our knowledge of the implications. Potential applications of social networks in information systems are presented in Staab, Domingos, Mika, Golbeck, Ding, and Finin (2005). Examples include viral marketing through social networks (see also Leskovec, Adamic, & Huberman, 2005) and e-mail filtering based on
A social network is a social structure comprising nodes, which generally represent individuals or organizations. The structure reflects the ways in which they are connected through various social familiarities ranging from casual acquaintances to close familial bonds. Social network analysis (SNA) is a technique in sociology, by which a node is called an actor and an edge is called a tie. From the 1930’s, social network analysis has been applied to various kinds of relational data, which relate one agent to another; such data cannot be reduced to the properties of individual agents themselves (Scott, 2000). In contrast to the long history of SNA in sociology, research on complex networks has received much attention since 1998, led by researchers from statistical physics and computer science fields: D. Watts, A. Barabasi, and A. Newman.

Social networks have become familiar recently because of the increasing use and development of social networking services (SNSs). As a kind of online application, SNSs are useful to register personal information including a user’s friends and acquaintances; the systems promote information exchange such as sending messages and reading Weblogs. Friendster and Orkut are among the earliest and most successful SNSs. Increasingly, SNSs target focused communities such as music, medical, and business communities. In Japan, one large SNS has more than 7 million users, followed by more than 70 SNSs that have specific characteristics for niche communities. Information sharing on SNSs is a promising application of SNSs (Goecks & Mynatt, 2004; Mori, Ishizuka, Sugiyama, & Matsuo, 2005) because large amounts of information such as private photos, diaries, and research notes are neither completely open nor closed: they can be shared loosely among a user’s friends, colleagues, and acquaintances. Several commercial services such as Imeem and Yahoo! 360° provide file sharing with elaborate access control.

In the context of the Semantic Web studies, social networks are crucial to realize a web of trust, which enables the estimation of information credibility and trustworthiness (Golbeck & Hendler, 2004). Because anyone can say anything on the Web, the web of trust helps humans and machines to discern which contents are credible, and to determine which information is reliably useful. Ontology construction is also related to a social network. For example, if numerous people share two concepts, the two concepts might be related (Mika, 2005). In addition, when mapping one ontology to another, persons who are between the two communities, or those who participate in both, play an important role. Social networks enable us to detect such persons with high betweenness.

Several means exist to demarcate social networks. One approach is to compel users to describe relations to others. In studies of the social sciences, network questionnaire surveys are often performed to obtain social networks, for example, asking “Please indicate which persons you would regard as your friend.” Current SNSs realize such procedures online. However, the obtained relations are sometimes inconsistent: users do not name some of their friends merely because they are not in the SNS or perhaps the user has merely forgotten them. Some name hundreds of friends, but others name only a few. Therefore, deliberate control of sampling and inquiry are necessary to obtain high-quality social networks on SNSs.

In contrast, automatic detection of relations is possible from various sources of information such as e-mail archives, schedule data, and Web citation information (Adamic, Buyukkokten, & Adar, 2003; Miki, Nomura, & Ishida, 2005; Tyler, Wilkinson, & Huberman, 2003). Especially in some studies, social networks are extracted by measuring the co-occurrence of names on the Web. Pioneering work was done in that area by H. Kautz: the system is called Referral Web (Kautz, Selman, & Shah, 1997). Several researchers have used that technique to extract social networks, as described in the next section.

This chapter presents an overview of social