Chapter 4
Community Discovery:
From Web Pages to Social Networks

Damien Leprovost
University of Bourgogne, France

Lylia Abrouk
University of Bourgogne, France

David Gross-Amblard
University of Bourgogne, France

ABSTRACT

This chapter presents a state of the art of research on the discovery of Web communities, in a general sense. For this purpose, the authors discuss various notions of communities and their related assumptions: hypertextual communities, tag communities and semantic-based communities.

COMMUNITY DISCOVERY: FROM WEB PAGES TO SOCIAL NETWORKS

During the past ten years, the Web has turned into an open collaborative system, often called Web 2.0, where anyone can provide information to others using publishing tools such as forums, blogs and wikis. The Web 2.0 also contains social web sites like Myspace, Facebook or Flickr, where people can annotate third-party information with tags. Various kinds of people use these collaborative systems, ranging from simple visitors to experts of a discussion topic.

From the huge amount of the resulting web pages one can observe emerging structures, forming communities of topics. Similarly, social networks enable to build communities of people, according to their friendship or common interests. A natural challenge for the next decade is to discover and exploit these communities.

This survey presents a state of the art of Web communities emergence, and is organized as follows. The first section discusses the concept of communities and their related assumptions, and presents our analysis method. The subsequent

DOI: 10.4018/978-1-61350-513-7.ch004

sections support the resulting classification: hypertextual communities, tag-based communities, social networks and semantic communities. The last section concludes.

Communities and Social Networks

What is a community? Indeed, community is an ambiguous term with over 120 definitions noted by Poplin (1979). According to the Wikipedia entry[1] on communities, "*A community is a group of interacting organisms sharing a populated environment. In human communities, intent, belief, resources, preferences, needs, risks, and a number of other conditions may be present and common, affecting the identity of the participants and their degree of cohesiveness"*. In this survey, we consider that a community is a virtual group of interacting entities (a web page, a user, etc.) sharing something or having something in common, each entity being identified in some way (by an URL or a UID). From there, we can distinguish between web pages communities and social networks. More specifically, a social network is a social structure made up of individuals (or organizations) called nodes, which are connected by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, emotional relationships, knowledge, and so on. Hence, a social network is a specific type of community, where relations are explicit, declared (or assumed) by its members.

First Movement: From Implicit to Explicit Communities

As illustrated by the social network example, there is a historical and natural movement on the Internet from implicit communities to explicit ones. This operation of community discovery is the hard goal of many organizations. We can illustrate this proposition by two scenarios:

- **Web pages community discovery:** due to its openness, the Web allowed a huge number of users for producing content, according to a agreed format, HTML. Hence users or organizations first produced content, regardless of its classification or accessibility. Since group of web pages addressing the same topics were hard to find, a first step toward expliciting communities was registration: that is the effort of declaring a web site to an authority, like DMOZ[2], Lycos[3], or Yahoo! Directory[4]. A second step was the use of automatic content analysis to identify topics of interest. Finally, a great success was obtained by combining this content analysis with the examination of explicit web links between pages (Kleinberg,1998; Brin & Page, 1998).
- **Social community discovery:** as a communication artifact, the Internet naturally hosts social communications. Being part of a thematic mailing list is an example of an explicit community. Similarly, several systems ask users to fulfill detailed profiles to find their matching partners. These approaches, not covered in the present survey, suffer from the hindrance of profile typing, often neglected by users. More recently, registering on a social network, applying for a group and making its friendship relations explicit are other examples. But systems like Facebook[5] are actively analyzing this explicit social graph to discover and suggest new friends to a given user. This way, the social network is expanding from a hidden, implicit state to an explicit one.

The evolution of these notions of communities therefore illustrates a movement from implicit communities to explicit ones. It is not a transformation, but rather the continuous inclusion of unknown or not understood preexisting elements.