Chapter 6

A Semi-Supervised Algorithm to Manage Communities of Interests

Pascal Francq

Paul Otlet Institute, Belgium & Université Catholique de Louvain, Belgium

ABSTRACT

This chapter presents a genetic algorithm, called the Similarity-based Clustering Genetic Algorithm (SCGA), used to group users’ profiles. This algorithm is integrated in an approach which allows to share documents among users browsing a collection of documents. The users are described in terms of profiles, with each profile corresponding to one area of interest. While browsing through the collection of documents, users’ profiles are computed. These profiles are then grouped into communities of interests using the SCGA which is based on the Grouping Genetic Algorithm (GGA). In fact, the SCGA can solve other similar problems under certain circumstances. The approach is part of a more generic model to manage information called the GALILEI Framework. This framework, which provides promising results, has been developed in a software platform available under the GNU GPL license.

INTRODUCTION

With the wide spread of document-oriented systems, such as the Internet, the amount of information available through electronic documents has exploded (Lawrence & Giles, 1999). Because of this expansion, information retrieval systems that facilitate the information extraction process by structuring and retrieving it are most welcome, since finding all the relevant information is still a crucial problem (Fogarty & Bahls, 2002). As this book claims, communities of interests are one of the solutions to share interesting knowledge between users. As explained in the first chapter, managing communities of interests is a difficult task, in particular detecting which users belongs to which community.

Several approaches exist to build communities of interests. One of them is to “profile” the users interests based on the content of the documents they read, and then grouping them regarding these profiles. I call this approach “social brows-
ing” (Francq, 2007) since the idea is to gather information to regroup similar users while they are browsing documents on the Internet or in intranets. The GALILEI platform implements this approach (Mathyhs, et al., 1998; Francq, 2003). It is an open source platform that proposes a complete implementation of a framework to manage numeric information (called the GALILEI Framework) which allows the building and the maintaining of communities of interests. One of the main issues of the framework is to group users into such communities, in particular because the number of communities (clusters) is unknown and the method has to be run frequently (typically communities are weekly updated).

This chapter presents the Similarity-based Clustering Genetic Algorithm (SCGA) that tackles this latest problem. Section 2 presents the general approach used to build communities of interests in the GALILEI Framework, and section 3 provides a technical overview on how users’ interests are computed. Section 4 studies the grouping problems in general, while section 5 describes the SCGA in particular. Section 6 proposes a validation methodology for the proposed approach which is then used to provide the results discussed in section 7. Finally, section 8 proposes some future research directions, section 9 presents a case study where communities of researchers are detected, and section 10 draws the conclusions of this chapter.

THE GALILEI FRAMEWORK

The main purpose of the GALILEI Framework is to propose an integrated model to manage digital information (mostly electronic documents). A complete description of the framework is outside the scope of this chapter, but an important feature is to identify users’ interests as precisely as possible and consequently grouping them. The approach integrated in the GALILEI Framework, and called social browsing, proposes (a) to model the multiple interests of a given user as separate profiles (each profile corresponds to a particular interest), (b) to automatically compute the descriptions of these profiles based on documents assessments done by the corresponding users and (c) to group these profiles into communities of interests (one user belonging to as many communities as the number of his or her profiles).

In order to compute the profiles descriptions, the GALILEI framework uses the relevance assessments of documents read by the users for a particular profile and the analysis of their content (Technical Overview describes the document analysis process). Currently, three different assessments have been adopted:

1. The document is relevant. A Web page about the Beatles is, for example, a relevant document for a “Beatles profile” representing a fan of this group.
2. The document is partially relevant (fuzzy relevant), but does not fall exactly within the scope of the domain. A Web page about the Wings may be considered by a “Beatles profile” as fuzzy relevant since there are connections between the two groups (Paul McCartney plays in both).
3. The document is outside the scope of the interest (irrelevant). The homepage of Steve Jobs is (probably) completely irrelevant for a “Beatles profile” (Apple, the name of the company he founded, is not related with the record label of the Beatles while having the same name).

Of course, other types of assessments are possible such as rating the documents on a scale (for example between 0 and 5). But, it is always difficult for users to determine the “right rating” corresponding to their perception of the relevance of a given document. Limiting the choice to three well-defined categories of relevance makes the approach more intuitive for the users.

The users’ profiles are then clustered based on their descriptions: similar profiles are grouped