Chapter XIV
Privacy Concerns when Modeling Users in Collaborative Filtering Recommender Systems

Sylvain Castagnos
LORIA—Université Nancy 2, Campus Scientifique, France

Anne Boyer
LORIA—Université Nancy 2, Campus Scientifique, France

ABSTRACT

This chapter investigates ways to deal with privacy rules when modeling preferences of users in recommender systems based on collaborative filtering. It argues that it is possible to find a good compromise between quality of predictions and protection of personal data. Thus, it proposes a methodology that fulfills with strictest privacy laws for both centralized and distributed architectures. The authors hope that their attempts to provide a unified vision of privacy rules through the related works and a generic privacy-enhancing procedure will help researchers and practitioners to better take into account the ethical and juridical constraints as regards privacy protection when designing information systems.

INTRODUCTION

Do you remember the satirical paper from Zaslow (2002) in the Wall Street Journal? The problem was the following: a man suspects that his digital videorecorder named TiVo thought he was gay. Indeed, it inexplicably recorded programs with gay themes. This man decided to modify TiVo's gay fixation by recording war movies. Then the machine started giving him documentaries on Joseph Goebbels and Adolf Eichmann. He has overcompensated and the machine stopped thinking he was gay and decided he was a fan of the Third Reich. The general principle of TiVo is to record for its owner some programs it just assumes he will like, based on shows he has chosen to record. The recommendation process used what he did to predict what he likes. A

major aspect related to recommender systems is to collect pertinent data about what you do in order to determine what you are. Such systems are very popular in many contexts, as for example e-commerce, papers online, or Internet access. Recommenders individualize their prediction which each user. Therefore, they need to collect and to utilize personal data. Fundamental issues arise such as how to ensure that user privacy will be guaranteed, particularly when individuals can be identifiable?

We have to keep in mind that many consumers appreciate having computers able to anticipate what they like, what they want to do or to read. Web personalization has been shown to be advantageous for both online customers and vendors. But for consumers shopping on the Internet, privacy is a major issue. Almost three-quarters of Internet users are concerned about having control over the release of their private information when shopping online (Source: U.S. Census Data on http://www.bbbonline.org/privacy/). This is also true in the Internet context of information retrieval. As the amount of data available on Internet is so huge, it becomes mandatory to assist the active user when searching or accessing Internet resources. Furthermore, the number of available resources is still exponentially growing: for example, the number of pages referenced by Google has increased from 1 to 8 trillion between June 2000 and August 2005.

Traditional search engines use to provide the active user with too many results to ensure that he/she will identify the most relevant items in a reasonable time. For instance, Google returns 5.4 billion links when the user asks for "news." There are still 768 million sites about news related to New York City. Moreover, searches may never end since new resources constantly appear. Confronted with this overload of data, the rationality of the active user is bounded to the set of choices that can be considered by human understanding. He/she tends to stop the search at the first choice which seems satisfying (Simon, 1982). This is the reason why the relevancy of results is no longer guarantee in most of existing information providers. Furthermore, searching information by using keywords and logical operators seems not easy enough for the general audience. As a result, the scientific community is rethinking the existing services of search and access to information, under the designation "Web 2.0" (White, 2006).

There are several possible approaches to assist the active user: adaptive interfaces to facilitate the exploration and the searches on the Web, systems relying on social navigation, sites providing personalized content, statistical tools suggesting keywords for improving searches, and so forth. Another solution consists in providing each user with items likely to interest him/her. Contrary to the personalized content, this solution does not require to adapt resources to the potential readers. Each item has to be proposed to concerned persons by using push-and-pull techniques.

To supply the active user with his/her concerns, we first have to build his/her model of preferences by collecting data about his/her activities. This approach is based on an analysis of usage. Nevertheless, it is not always possible to collect quickly enough data about the active user. Collaborative filtering techniques (Goldberg, 1992) are a good way to cope with this difficulty. They amount to identifying the active user to a set of persons having the same tastes, based on his/her preferences and his/her past actions. This kind of algorithms considers that users who liked the same items have the same topics of interest. Thus, it becomes possible to predict the relevancy of data for the active user by taking advantage of experiences of a similar population.

There are several fundamental problems when implementing a collaborative filtering algorithm. Beyond technical questions such as quality of service or cold start, are ethical aspects such as intimacy preservation, privacy, or reglementary aspects. These questions are crucial since they are related to human rights and freedom and consequently will impact development and generalisa-