Chapter 2
A Recommendation System for Scientific Papers through Bayesian Nonparametric Hybrid Filtering

Abel Rodriguez
University of California – Santa Cruz, USA

Radhakrishna Vuppala
University of California – Santa Cruz, USA

ABSTRACT

Recommender systems have become an important area of research with numerous applications on e-commerce. This chapter introduces a joint statistical model for user preferences and item features that can serve as the basis for a recommendation about recently published scientific papers. The model is constructed using ideas from the literature on Bayesian nonparametric mixture modeling. More specifically, user preferences are modeled using an Infinite Relational Model (IRM) in which both users and items are independently partitioned into homogeneous groups, while item features are modeled using a topic model, which also partitions items into groups with homogenous features. Information is shared across both components of the model through a common partition of items. Hence, the model is a hybrid system that combines ideas from collaborative and content-based filtering. The chapter discusses three different computational strategies, including a Markov chain Monte Carlo algorithm for full posterior inference, an iterated conditional maximization algorithm, and a mean-field variational algorithm for point estimation and prediction in large datasets where Markov chain Monte Carlo approaches might not be practical. The model is illustrated through simulation studies and by analyzing data from CiteULike.

1. INTRODUCTION

Recommendation systems aim to predict user ratings for items such as books, movies or songs they have not yet considered, and use those ratings to suggest new items to the user. Recommendation systems are widely used in a variety of e-commerce applications, with Websites such as amazon.com, pandora.com and netflix.com being well known examples of platforms that incorporate them into their business models. Furthermore, the underlying models and algorithms behind recommendation
A Recommendation System for Scientific Papers through Bayesian Nonparametric Hybrid Filtering

systems have applications in a number of scientific areas as diverse as genomics and political science. For recent reviews, see Adomavicius and Tuzhilin (2005), Jannach et al. (2010) or Shapira (2011).

Collaborative filtering approaches to recommendation systems (Schafer et al., 2007; Herlocker et al., 2000; Su & Khoshgoftaar, 2009) leverage information about the behaviors, activities or preferences of multiple users to predict what others will like based on how similar their overall preferences are. The item-to-item filtering algorithm used by amazon.com (Sarwar et al., 2001; Linden et al., 2003), which makes suggestions based on what other items have been bought by users that acquired the item currently being browsed, is a well known example of a recommendation system based on collaborative filtering. Collaborative filtering methods might use clustering tools to find items that are close to those each other in terms of user preferences (Ungar & Foster, 1998; Linden et al., 2003; Hofmann, 2003), or might rely on matrix factorization algorithms such as singular value or non-negative matrix factorizations (Patererek, 2007; Koren, 2008), or probabilistic factor models (Salakhutdinov & Mnih, 2008; Agarwal & Chen, 2009). Breese et al. (1998) and Herlocker et al. (2004) present comparative evaluations of a number of collaborative filtering algorithms.

In contrast to collaborative filtering, content-based filtering methods (Pazzani & Billsus, 2007; Van Meteren & Van Someren, 2000; Mooney & Roy, 2000) use information about the characteristics of the item to make recommendations. This requires that each item be characterized in terms of a set of measurable features. The recommendation engine behind pandora.com, which uses song attributes extracted from the Music Genome Project (John, 2006) to make suggestions starting with a user-provided seed, is a well-known example of a content-based filter. Content-based recommenders rely on classifiers such as support vector machines (Cortes & Vapnik, 1995) or classification trees (Breiman, 1993, 1996, 2001).

Collaborative and content-based filters have complementary advantages and disadvantages. For example, collaborative filters do not rely on machine analyzable content and are therefore capable of accurately recommending complex items such as movies without requiring an “understanding” of the item itself. However, these systems often require a large number of users in order to make accurate recommendations. Also, collaborative filters cannot be used to make recommendations for “new” items that have not been previously rated by any user, and perform poorly when the ratings are very sparse. On the other hand, content-based filters typically only require moderate amounts of information about each user, and are capable of predicting ratings for unrated items. However, their performance is heavily dependent on the quality of the feature set used to describe the items. Also, content-based filters typically do not allow for “serendipity” (i.e., the possibility that a user might like a completely different type of item that she has not experienced yet), which is handled well by collaborative filters. Hybrid approaches to recommendation systems (Burke, 2007; Melville et al., 2002; Pazzani, 1999; Popescul et al., 2001; Yoshii et al., 2006; Salter & Antonopoulos, 2006) combine collaborative and content-based filters to create more robust systems that overcome these limitations.

This chapter focuses on the problem of making recommendations about recently published scientific papers to researchers and scientists using the text of the documents as well as the information contained in their own as well as other researcher’s libraries. The development of personalized online libraries such as CiteULike (http://www.citeulike.org) and Mendeley (http://www.mendeley.com) has opened the door to the development of this type of systems. Computer-based recommendation systems of this type have numerous advantages over traditional approaches used by scientist to find relevant papers, such as citation tracking and keyword searches. We pursue