Chapter 7
Recommender Systems: Sources of Knowledge and Evaluation Metrics

Denis Parra and Shaghayegh Sahebi

Abstract. Recommender or Recommendation Systems (RS) aim to help users dealing with information overload: finding relevant items in a vast space of resources. Research on RS has been active since the development of the first recommender system in the early 1990s, Tapestry, and some articles and books that survey algorithms and application domains have been published recently. However, these surveys have not extensively covered the different types of information used in RS (sources of knowledge), and only a few of them have reviewed the different ways to assess the quality and performance of RS. In order to bridge this gap, in this chapter we present a classification of recommender systems, and then we focus on presenting the main sources of knowledge and evaluation metrics that have been described in the research literature.

7.1 Introduction

Recommender or Recommendation Systems (RS) aim to help a user or a group of users in a system to select items from a crowded item or information space [70]. In order to generate recommendations, a RS might try to match users’ characteristics with items’ characteristics by performing content filtering, or it might look at previous interactions of the user in the system to match users with similar patterns [53]. A typical domain where RS are useful is the World Wide Web (WWW): with its...
overwhelming growth of available information and the continuously growing number of different devices that can be used to access it RS have taken on an important role in the daily lives of people to find relevant resources, such as movies [41], books [56], music [13], tourism destinations [12], or cooking recipes [26].

The first recommender system, Tapestry [32], was introduced almost 20 years ago by Goldberg et al. to deal with the increasing amount of messages that users received by email. This early system—as well as GroupLens developed by Paul Resnick et al. [96] and Ringo by Shardanand and Maes [107]—made use of a technique called Collaborative Filtering (CF) to provide recommendations to a center user based on previous actions performed by herself and by like-minded users, denoted as nearest neighbors. All these systems make use of some form of deviance measure between a predicted and a real value of preference for evaluation. In their seminal paper, Herlocker et al. [42] survey different tasks and metrics for RS, introducing, among others, the concepts of serendipity and novelty. However, these concepts started to have a larger impact on the evaluation of RS after the Netflix prize.

The Netflix Prize was a contest created by the movie rental company Netflix in October of 2006 [11]. The Netflix Prize challenged the data mining, machine learning and computer science communities to improve the algorithm Cinematch by at least 10% in terms of predicting the ratings that users assigned to movies. The winners of this challenge would receive a $1 million dollar prize. Netflix released a dataset of 100 million anonymous movie ratings and the evaluation was based on Root Mean Square Error (RMSE), a metric that we explain in section 7.4.1. Although the community of researchers engaged in RS existed well before this contest, the Netflix Prize attracted a large amount of people from different areas. It might not be a coincidence that the ACM Recommender Systems conference, targeted specifically for RS, began in 2007. Despite the benefit of attracting a large community of researchers to the field, the Netflix Prize had the negative effect of focusing on accuracy in the active evaluation period, giving less importance to important characteristics of the recommendations such as coverage, novelty, or diversity. By the time the challenge was finished, the RS community started to show more interest in other quality metrics.

Some studies have gone beyond accuracy to evaluate RS such as recommendation diversification by Ziegler et al. in 2005 [128] and Zhou et al. in 2010 [125], serendipity by Murakami et al. in 2008 [80] and by Zhang et al. in 2011 [124], and coverage by Ge et al. in 2010 [29]. More recently Vargas and Castells try to combine accuracy and serendipity in a single evaluation framework [113]. These new trends in RS evaluation stem from several factors, among which we count:

- **Accuracy and user satisfaction are not always related:** Some articles showed that rating prediction accuracy is not always correlated with other metrics [95], and most importantly, not necessarily correlated with user satisfaction [39] [70]. This result supported the need for creating new evaluation measures that better

---

1. [http://www.netflixprize.com](http://www.netflixprize.com)
2. [http://www.netflix.com](http://www.netflix.com)