Chapter 8
Bimodal Cross-Validation Approach for Recommender Systems Diagnostics

Dmitry I. Ignatov
National Research University Higher School of Economics, Russia

Jonas Poelmans
Katholieke Universiteit Leuven, Belgium

ABSTRACT

Recommender systems are becoming an inseparable part of many modern Internet web sites and web shops. The quality of recommendations made may significantly influence the browsing experience of the user and revenues made by web site owners. Developers can choose between a variety of recommender algorithms; unfortunately no general scheme exists for evaluation of their recall and precision. In this chapter, the authors propose a method based on cross-validation for diagnosing the strengths and weaknesses of recommender algorithms. The method not only splits initial data into a training and test subsets, but also splits the attribute set into a hidden and visible part. Experiments were performed on a user-based and item-based recommender algorithm. These algorithms were applied to the MovieLens dataset, and the authors found classical user-based methods perform better in terms of recall and precision.

INTRODUCTION

A modern Internet user rather frequently faces recommender systems. Recommender systems are defined by the ACM Recommender Systems conference as "software applications that aim to support users in their decision-making while interacting with large information spaces. They recommend items of interest to users based on preferences they have expressed, either explicitly or implicitly" RecSys (2011). The paper by Adomavicius et al. (2005) presented a survey on the state of the art of recommendation algorithms and grouped them in three main categories: content-based (also referred to as item-based), collaborative (also referred to as user-based), and hybrid recommendation approaches. An example of a web site where recommender systems are frequently

DOI: 10.4018/978-1-4666-1900-5.ch008

used is an online bookshop. If a user buys book X in an online book shop she also gets recommendations in the form ‘’other customers who bought book X also bought books Y and Z’’. There are also a lot of web systems which can recommend potentially interesting web sites to a particular user; they are called social bookmarking systems (e.g. http://del.ici.ou.us). Other examples include the websites http://facebook.com/ and http://twitter.com/ and for Russian companies, the websites http://imhonet.ru/ and http://www.ozon.ru/.

Besides the Internet the most popular and non-technological way to get recommendation is still friends’ suggestions. However, if a user wants more items to buy (to watch, to read etc.) the task is getting harder, because there may be a lot of different options of the choice, her friends may not be informed about latest items in the field or just have different tastes. To cope with these difficulties she can use so-called collaborative filtering Goldberg et al. (1992). Recommender algorithms based on collaborative filtering techniques utilize a fairly simple scheme. They find users of the system who have similar to her tastes or preferences, then compose the list of items the users selected and rank these items, and as a result she gets Top-N items of the list. Herlocker et al. (2004) presented in depth research on evaluating the quality of collaborative filtering approaches. Another less evident but interesting application is recommending key phrases in web advertising systems, where firms buy advertising phrases from web search engines to show advertisement by a user’s request Ignatov et al. (2008), Ignatov et al. (2008). This approach made use of Galois operators to obtain morphological association rules.

RECOMMENDER ALGORITHMS

In this paper without loss of generality we consider only two groups of recommender techniques, which can be called the classical ones, mainly user-based and item-based approaches Badrul et al. (2000), Deshpande et al. (2004). A key notion for these techniques is similarity, which can be expressed as Jacquard measure, Pearson correlation coefficient, cosine similarity etc. Initial data are usually represented by an object-attribute matrix, where the rows describe objects (users) and the columns represent attributes (items). A particular cell of the matrix can be either 1 or 0, which stands for the fact that the item was purchased or not respectively. Also the values can be rates or marks of items, for example, film’s rates given by users.

User-Based Recommendations

User-based methods find similarity between a target user u0 and other users of the recommender system. As a result the target user has n most frequently bought items by k most similar to u_0 users (customers). Let u_0 be a target user, u_0^1 be items that she evaluated, $sim(u_0, u)$ be a similarity between the target user u_0 and another user u. In this research we use Pearson correlation coefficient as a similarity measure. Define the set of nearest neighbors (neighborhood) for the target user by the formula:

$$N(u_0) = \left\{ u \mid sim(u_0, u) \le \Theta \right\}.$$

However, it is appropriate to obtain Top-k nearest neighbors, that is Top-k defines the threshold Θ. Hence the set of nearest neighbors includes k users which have similarity with u_0 higher than a certain threshold. After ordering the users by decreasing similarity, one should select not only Top-k of them, but also check the similarity value of (k+1)-th user in the list. If this similarity value is equal to the preceding one than one should add (k+1)-th user to the neighborhood $N(u_0)$. One should repeat the procedure until the next similarity value changes. Since we predict the rate of an item i by a specific target user u_0 we are interesting only those users from the neighborhood who have evaluated i:

Related Content

A Deep Neural Network Model for Cross-Domain Sentiment Analysis
Suman Kumari, Basant Agarwaland Mamta Mittal (2022). *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines (pp. 157-173).*
www.irma-international.org/chapter/a-deep-neural-network-model-for-cross-domain-sentiment-analysis/308485

Multi-Agent System Based on Data Mining Algorithms to Detect Breast Cancer
Imane Chakour, Yousef El Mourabitand Mohamed Baslam (2020). *Critical Approaches to Information Retrieval Research (pp. 210-224).*
www.irma-international.org/chapter/multi-agent-system-based-on-data-mining-algorithms-to-detect-breast-cancer/237647

E-Commerce and Data Mining: Integration Issues and Challenges
Parviz Partow-Navidand Ludwig Slusky (2004). *Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance (pp. 321-333).*
www.irma-international.org/chapter/commerce-data-mining/27924

LF-LDA: A Supervised Topic Model for Multi-Label Documents Classification
Yongjun Zhang, Zijian Wang, Yongtao Yu, Bolun Chen, Jialin Maand Liang Shi (2018). *International Journal of Data Warehousing and Mining (pp. 18-36).*
www.irma-international.org/article/lf-lda/202996

User Behaviour Pattern Mining from Weblog
Vishnu Priyaand A. Vadivel (2012). *International Journal of Data Warehousing and Mining (pp. 1-22).*
www.irma-international.org/article/user-behaviour-pattern-mining-weblog/65571