Combining Graphs and Big Data to Recommend Apps

Luis F. Chiroque[1], Héctor Cordobés[1], Antonio Fernández Anta[1], Rafael A. García Leiva[1], Philippe Morere[1], Lorenzo Ornella[2], Fernando Pérez[3], and Agustín Santos[1]

[1] Institute IMDEA Networks, Madrid, Spain
[2] Politecnico di Torino, Turin, Italy
[3] [3] U-tad, Madrid, Spain

Abstract. Recommendation engines (RE) are becoming highly popular, e.g., in the area of e-commerce. A RE offers new items (products or content) to users based on their profile and historical data. The most popular algorithms used in RE are based on collaborative filtering. This technique makes recommendations based on the past behavior of other users and the similarity between users and items. Metrics used for the computation of similarity include Euclidean distance, cosine distance, and correlation based distances. We have examined alternative similarity definitions based on the properties of the networks formed by users and items. The evaluated similarity metrics use graph theoretic concepts like the degree, several centrality measures, and flow maximization.
In this paper we present how the techniques proposed have been evaluated in a real environment for the recommendation of applications to smartphone users. Training the RE required the pre-processing of a large dataset consisting of around 1 billion records. A big data environment, based on Hadoop/Elastic Map Reduce, HBase, and Pig was set up for building and processing the application and user graphs. The big data environment reduced the processing time from more than one week in a single machine, to a couple of hours in the Hadoop cluster. Hence, the application of big data techniques allows a near real-time re-training of the RE.

1 Introduction

Motivation It is becoming very common in online platforms (shopping websites, online newspapers, online social networks, smartphone apps, etc.) to recommend items to the users that will (hopefully) be of her interest. The items to recommend are selected by a recommendation engine (RE), that typically uses the user profile and context, and historical data. The RE typically has a catalog of items from which to choose its recommendation, and there are spaces in the online platform viewing area in which the recommended product is presented. The context of the user typically includes its past navigation history, including the current viewing context, which may involve a product (e.g., in a shopping website), a piece of news (e.g., in an online newspaper), a user profile (e.g., in an online social network), or the application that is being executed (e.g., in a smartphone).

Recently, the most popular algorithms used in RE are based on collaborative filtering [1], . This technique makes recommendations based on the historical data of all the users and the estimated similarity between them. Metrics used for the computation of customers' similarity include Euclidean distance, cosine distance, and correlation-based distances.

Contributions We have developed RE based on collaborative filtering to promote an ecosystem of smartphone apps. In this ecosystem, the users of the apps get banners advertising other apps that they have not installed. The RE wants to maximize the click-through rate (CTR) of users in these banners, which hopefully implies maximizing the installation of new apps. In this work we have evaluated different algorithms for the RE, some of them new, based on the available data on users, banners, and apps. The new RE define networks formed by users and apps, and uses graph theoretic concepts like the degree, several centrality measures, and flow maximization.

In order to train the RE it has been required the pre-processing of a large dataset consisting of around 1 billion records, which contains activity of several million users. A big data environment, based on Apache Hadoop [2], Amazon Elastic Map Reduce [3], HBase [4], and Pig [5] was set up for this preprocessing, which involved cleaning the data and building the tables and networks used by the RE.

The big data environment reduced the preprocesssssing time from more than one week in a single machine, to a couple of hours in the Hadoop cluster. Hence, the application of big data techniques allows a near real-time re-training of the RE.

It is worth to mention that the different RE developed were tested with real users and apps for about a week, which allowed to identify some techniques and algorithms that have not been explored in the literature and gave the largest CTR.

References

1. Koren, Y., Bell, R.M.: Advances in collaborative filtering. In Ricci, F., Rokach, L., Shapira, B., Kantor, P.B., eds.: Recommender Systems Handbook. Springer (2011) 145–186
2. The Apache Software Foundation: The Apache Hadoop project (2014) [Online http://hadoop.apache.org/; accessed 6-August-2014].
3. Amazon Web Services, Inc: Amazon Elastic MapReduce (2014) [Online http://aws.amazon.com/elasticmapreduce/; accessed 6-August-2014].
4. The Apache Software Foundation: Apache HBase (2014) [Online http://hbase.apache.org/; accessed 6-August-2014].
5. The Apache Software Foundation: Apache Pig (2014) [Online http://pig.apache.org/; accessed 6-August-2014].