Group 14: Hybrid Advertising System Employs Big Data Analysis

Zijie Han
1155152516@link.cuhk.edu.hk
Group 14

Yanting Song
1155152477@link.cuhk.edu.hk
Group 14

Dingxi Jiang
1155149027@link.cuhk.edu.hk
Group 14

Yubin Ren
1155147507@link.cuhk.edu.hk
Group 14

1 MOTIVATION

People's increasing reliance on mobile apps has generated massive datas in many fields and has driven a huge ads market. Efficient advertising recommendation can benefit to advertisers, advertising vendors, and customers. However, there are still some difficulties for vendors to provide suitable ads to customers. Our project can help this from four aspects, as listed below:

- Help users find more favorite products and reduce the long tail effect
- Deepen understanding of users, provide customized services, achieve accurate delivery, and improve service quality, so that every user with different needs can get a complete user experience.
- Reduce information overload, filter low-value information, and improve information utilization.
- Increase the website visits and ad click rates.

2 MOST RELATED COURSES

Our project aims to utilize the most important Big Data concepts and try to make ourselves focus on the main topics in Big Data. Thus, our project mainly use three technologies related to the course:

- MapReduce /Frequent Itemsets
- Scalable Clustering
- Recommender systems

3 DELIVERABLE

- A tinny report explaining deep relations among ads and users.
- An ad recommendation system employs hybrid algorithm, pushing the best ads to users.
- A final report illustrating the algorithms, system architect and deliverables' superiority.

4 DATA SET ILLUSTRATION

Files: ad.csv, click.csv, user.csv
Size: 2GB
Information: According to official requirements, all advertising information and user information are anonymized and displayed in digital form. Thus, the report can be a little bit weird because of abstract instance. The contest will provide contestants with a set of user ad click history records in a 91-day (3 months) time window as a training data set. Each record contains the date (from 1 to 91), user information (age, gender), and information about the advertisement that was clicked (material id, advertisement id, product id, product category id, advertiser id, advertiser industry) id, etc.),

and the number of times the user clicked on the ad that day. The test data set will be the ad click history of another group of users. The test data set provided to participants will not contain the age and gender information of these users. This competition requires participants to predict the age and gender of the users appearing in the test data set.

5 TECHNOLOGY AND ALGORITHM

Generally our project uses **MapReduce** to preprocess the data, such as getting the click through rate of an advertisement. The **A-priori** algorithm can help us to find frequent itemsets and find some internal relations between the data. **BFR or CURE** have shown a good performance in clustering, accordingly we use BFR or CURE to cluster the advertising data set in this project. At last, Advertising System recommends advertisements to users according to the internal relationship between the data obtained before through **hybrid recommendation algorithm**.

6 RELATED WORK

MapReduce is used as a model for programming and it is also greatly used to process and generate large datasets that is amenable to various tasks in the real-world. Instead of doing distributed programming, programmers only need to focus on the computation of a map and a reduce function, and with the help of the underlying runtime system, the computation can be automatically parallelized by large-scale clusters of machines. The system can also handle machine failures, and schedule inter-machine communication to make efficient use of the network and disks. We will use MapReduce which is the essential technique to handler our massive datasets. Only if in this way, can we efficiently carry out our project[2].

To maximize the quantity of conversions and set cost-per-click (CPC) as a KPI constraint. Xun Yang et al. converted such a problem into a linear programming problem and leverage the primal-dual method to derive the optimal bidding strategy. They also proposed a feedback control-based solution and devise the multivariable control system to address the applicability issue. This algorithm gave us some inspiration which might be useful to choose the best ads for users.[1].

The explosive development of e-commerce and mobile Internet has made information search and selection more difficult. Users are overloaded by various options on the Internet but most of them don't have enough time or relevant knowledge to decide which one

to choose by themselves. Luckily, Recommender Systems Handbook has provided us lots of useful ways to cope with recommending and information overload[3].

There are many approaches have been widely used in the design of recommender systems, such as collaborative filtering, k-nearest neighbor (k-NN).

7 APPLICATION

Advertising System will use multiple algorithms to analyze different types of advertisements which depends on users' preferences, and finally recommend suitable advertisements to users according to the hybrid algorithm.

8 TIMELINE

Table 1: Rough Timeline

Mile Stone	Time
Dataset analysis, key algorithm discussion	10/11/2020
Pre-processing data implementation	24/11/2020
Key algorithm implementation	25/11/2020
Demo goes alive	1/12/2020
Parameters tuning, final deliverables	6/12/2020

REFERENCES

[1] Xun ang, Yasong Li, Hao Wang, Di Wu, Qing Tan, Jian Xu, and Kun Gai. [n.d.]. Bid Optimization by Multivariable Control in Display Advertising. *Association for Computing Machinery*, 1966–1974.

[2] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified Data Processing on Large Clusters. *Association for Computing Machinery* 51, 1 (Jan. 2008), 107–113. https://doi.org/10.1145/1327452.1327492

[3] Shapira B. DRicci F., Rokach L. 2010. *(2011) Introduction to Recommender Systems Handbook.* Springer, Boston, MA. https://doi.org/10.1007/978-0-387-85820-3_1