On-line advertisement click prediction

SAHITI KORRAPATI[1,*]

[1] *School of Informatics and Computing, Bloomington, IN 47408, U.S.A.*
[*] *Corresponding authors: sakorrap@iu.edu, S17-IR-2013*

March 13, 2017

This project aims at predicting the most suitable advertisements to be displayed on the web pages based on relevance by ranking each ad based on the liklihood of clicking. Data is obtained as CSV files from Kaggle Datasets and is stored in Hadoop Data File system(HDFS). In this project, various Bigdata tools and softwares are used to carry out the analytical computations efficiently. And Ansible is used to deploy all the necessary softwares on a cloud.

Keywords: Hadoop, Ad click Prediction, BigData, Ansible, Chameleon cloud

https://github.com/sakorrap/sp17-i524/tree/master/project/S17-IR-2013/report.pdf

1. INTRODUCTION

It has been analyzed that an average American spends about 23 hours per week surfing on-line [1]. This on-line user activity is being captured by companies to perform analysis for advertisements or recommendations or any other purpose. This has given rise to the field of "Web Analytics" and one such application is Ad Click prediction. Many measures are available to assess the ad performance. One such measure to assess the immediate ad response is click-through rate (CTR) of the advertisement [2] which is defined as the ratio of a number of clicks on an ad to the number of times the ad is shown, expressed as a percentage [3]. The user activity data that is used for prediction is enormous. To handle such large volumes of data Big data technologies come handy. The dataset that we are dealing with in this project is released by Outbrain which is 2 Billion page views and 16,900,000 clicks of 700 Million unique users, across 560 sites [4]. The data is anonymized and is in CSV file format. Paraquet compressing along with impala/drill to be decided to query and analyze the data compressed in files that are stored in HDFS. Programming language for analyzing the data is yet to be decided between JAVA and Python. Ansible is used to deploy the software and chameleon cloud for running virtual machines.

2. BACKGROUND

The dataset contains a sample of users' page views and clicks, as observed on multiple publisher sites in the United States between 14-June-2016 and 28-June-2016. Each viewed page or clicked recommendation is further accompanied by some semantic attributes of those documents [4]. The dataset contains numerous sets of content recommendations served to a specific user in a specific context. Each context (i.e. a set of recommendations) is given a display_id. In each such set, the user has clicked on at least one recommendation. Our task is to rank the recommendations in each group by decreasing predicted likelihood of being clicked [4].

2.1. Data fields description

Each user in the dataset is represented by a unique id (uuid). A person can view a document (document_id), which is simply a web page with content (e.g. a news article). On each document, a set of ads (ad_id) are displayed. Each ad belongs to a campaign (campaign_id) run by an advertiser (advertiser_id). Figure 1 shows the fields in our dataset. Metadata about the document is also provided, such as which entities are mentioned, a taxonomy of categories, the topics mentioned, and the publisher [4].

3. SETUP AND CONFIGURATION

TBD

4. WORK FLOW

TBD

5. EXPERIMENTS AND RESULTS

TBD

6. LICENSING

TBD

7. CONCLUSION

TBD

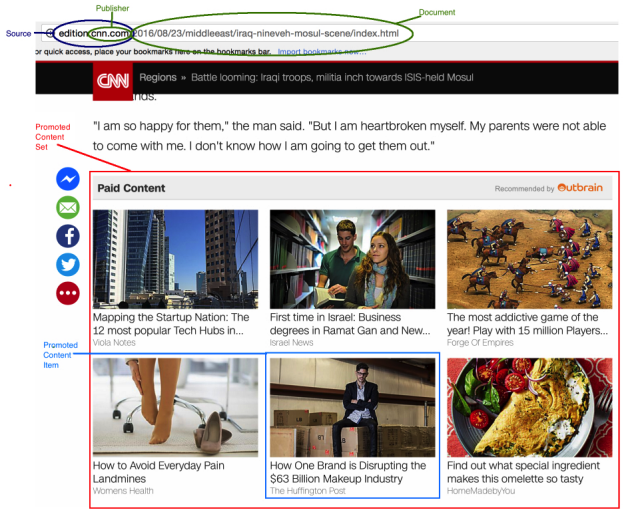

Fig. 1. Displaying Source, Publisher, Document, Promoted content set and items [4]

8. EXECUTION SUMMARY

1. Feb 23 - Mar 6, 2017 Exploring Python, Ansible and Chameleon cloud
2. Mar 10 - Mar 13, 2017 Exploring the datasets available and come up with the project proposal
3. Mar 14 - Mar 20, 2017 Decide on the architecture and workflow
4. Mar 21 - Mar 24, 2017 Configure and setup the workbench
5. Mar 25 - Mar 30, 2017 Prepare high level design
6. Mar 31 - Apr 6, 2017 Implementation and Testing
7. Apr 7 - Apr 11, 2017 Automate the deployment process using Ansible
8. Apr 12 - Apr 17, 2017 Optimizing and work on benchmarking
9. Apr 18 - Apr 22, 2017 Project review and completing any pending work
10. Apr 23 - Apr 24 Complete the report and commit the code

ACKNOWLEDGEMENTS

The authors thank Professor Gregor Von Laszewski and all the AIs of big data class for the guidance and technical support.

REFERENCES

[1] D. Mielach, "Americans spend 23 hours per week online, texting," Web-page, accessed mar-13-2017. [Online]. Available: http://www.businessnewsdaily.com/4718-weekly-online-social-media-time.html
[2] marketingterms.com, "Clickthrough rate definition," Web-page, accessed mar-13-2017. [Online]. Available: http://www.marketingterms.com/dictionary/clickthrough_rate/
[3] Wikipedia, "Click-through rate," Webpage, accessed Mar-13-2017. [Online]. Available: https://en.wikipedia.org/wiki/Click-through_rate
[4] Outbrain, "Data introduction," Webpage, accessed Mar-13-2017. [Online]. Available: https://www.kaggle.com/c/outbrain-click-prediction/data

AUTHOR BIOGRAPHIES

Sahiti Korrapati is pursuing her MSc in Data Science from Indiana University Bloomington