FILTERING ADS BASED ON USER BROWSING DATA AND VISITING PLATFORMS USING DATA ANALYTICS AND MACHINE LEARNING

15CS496L – Zeroth Review Report

Submitted by

Aakash Sharma – RA1711003010039

Kriti – RA1711003011474

In partial fulfilment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR- 603 203

January 2021

BONAFIDE CERTIFICATE

Certified to be the project report titled **“FILTERING ADS BASED ON USER BROWSING DATA AND VISITING PLATFORMS USING DATA ANALYTICS AND MACHINE LEARNING”** is the bonafide work done by Aakash Sharma (Reg No: RA1711003010039) and Kriti (Reg No:RA1711003011474), of CSE B.Tech Degree course in the theory **15CS496L – Zeroth Review Report** during the academic year 2020-21.

	Sign of the MDD faculty	**Sign of Academic Advisor**
Mr.U.M.Prakash	**Mr. T. Balachander**	**Name**
Assistant Professor	**Asst.Prof (Sr.G)**	**Designation**
Computer Science	**Department**	**Department**

Date of Submission: 22nd January, 2021

ABSTRACT

Digital Advertising is data-driven strategy for target audience. It has gained popularity because of the revenue it generates for the advertising agencies. Also, digital advertisement can be considered as an upgrade to the traditional advertisement as it generates data that could be processed to give insights on the usage so that it can be altered to receive more audience.

Considering the user’s end, the experience can be unpleasable as the advertisements are shown to the user on every platform being browsed, irrespective of their relevance. This is because the advertisements are based solely on user’s history of browsing. For example, if a user has a browsing history of clothes from SiteX and is currently working on some educational portal, the clothes’ advertisements from SiteX would be shown to the user on this educational portal because of the browsing history. The advertisements will not only be a distraction to the user but will also leave a distaste for the portal that gives the irrelevant advertisements and hence create a bad user experience.

On the other hand, companies that are paying a large amount to the advertising agencies for their ads will be at a loss as there is a high possibility that the target audience is not reached or the ads are killing the target audience because they are leaving a distaste.

Through our project we aim to provide the user with an option to filter ads based on relevancy and browsing history so that the experience at a portal is maintained, It also ensures that the target audience of the companies are the people who are visiting or have visited some related sites. So the probability that the user will click their ad is high, and for the company the target audience is PURE.

With this project, we aim to build a solution prototype depicting ideal workflow of Ads in the Digital environment where the user experience is maintained and the advertising companies are benefited.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ADS	-	Advertisements

INTRODUCTION

The past decade has seen a huge growth in online advertising. Advertisers are very interested in precisely targeted ads, they want to spend the smallest amount of many and get the maximum targeted users i.e. the users which are most likely to buy their product. This is resolved by targeted advertising.

Digital advertising is data-driven strategy for target audience. But the user experience can become unpleasable due to irrelevant ads. Search engines store users' browsing data on the cloud. This data is then used to display ads to the users.

A particular website sets a number of ads to be displayed to the user which might or might not be relevant to the portal. We aim to alter this approach, to enhance user experience by filtering ads based on the user's history and relevancy of the current site. This will not only create a good user experience, but will also provide the advertising companies with the target audience.

The aim of this report is to implement other methods to achieve the targeted users to the advertising companies without disturbing the user experience on a particular portal or website.

RELATED WORKS

In earlier days of the internet, online advertising was mostly prohibited. Later, it started with online advertisements via emails and then expanded to other internet platforms. Advertising is usually done in the form of text, visual images, banners, animations etc. Theses advertisements frequently target users with particular traits to increase the ads effect.
The common notion of ads is to collect the user's data through the browsing device and use to target the ads to that user. And this essentially forces the user to see these ads on every website or platform they visit, irrespective of its relevancy.

Search engines rank websites on the basis of number of clicks on it, and these clicks add to the market value of the website or portal.

Following are the common targeted methods used to target these advertisements –

1. Demographic Targeting
2. Property Targeting

3. Behavioural Targeting

Advertisers and publishers uses a wide range of payment calculated methods, some of them are –

1. Cost per mille (CPM)
2. Cost per click (CPC)
3. Cost per engagement (CPE)
4. Cost per view (CPV)
5. Cost per install (CPI)

The following data represents the time spend by users on various internet platforms:
[Table 2.1: Time spend by users on different internet platforms]

Social Media	33%
Online TV & Streaming	16%
Music Streaming	16%
Online Press	13%
Others	22%

The following data represents comparison between various online media used for ads:
[Table 2.2: Comparison between various online media used for ads]

Social Networks	37%
Individual retailer websites	34%
Price comparison websites	32%
Multi brand websites	21%
Visual social networks	20%
Travel review websites	16%
Emails from brand/retailers	14%
Deal of the day websites	12%
Mobile Apps	11%
Blogs	11%
Digital press and magazines	6%

We can clearly see the impact of social media, websites and other internet platforms, and one can imagine the type of data that is accessed through these portals. In the name of storing just the browser history, personal and individual details, including name, emails, etc. are also stored. And that is shared through different partners which is a whole market in itself.

Businesses make an average of $2 in revenue for every $1 they spend on Ads. The average click through rate across all industries is 3.17% for the search network and 0.46% on the display network. Every 4 in 10 internet users say that they follow their favourite brands on social media. Over 37% of online shoppers use social media for their product advertisement.

The other work is highlighted in the table below:

[Table 2.3: Literature Survey]

Ref. paper Number	**Objective**	**Remarks**	**Limitations**
	Contextual Internet Multimedia Advertising	Paper suggests a second generation of multimedia advertisement using contextual advertisement over behavioural one. This is done by showing the most relevant advertisement (both globally and locally (depending on the text and image surrounding)) at an appropriate position. Used computer vision and multimedia retrieval techniques. Active tagging to be used to categorize images and videos for more success in terms of relevancy of ads. Should use a behavioural model to position the ads.	The advertisements following contextual models should have videos/images categorized according to their content rather than pre-provided labels that have limited success rate. Relevancy of advertisement needs improvement. There is scope of applying psychological elements or studies conducted on user behaviour to make ads more relevant and increase revenue.
	Tree-Based Real-Time Advertisement Recommendation System in Online Broadcasting	The paper uses a tree model to categorise and filter ads based on the tree skeleton that is proposed. The paper proposes to store user's information and traverse through trees to	The predefined tree model makes it unable to introduce new categories of ads and thus it is not apt to handle the emerging categories of new ads and digital marketing.

		find the right category of ads.	Problem of getting unnecessary ads is still not resolved.
	A collaborative filtering approach to ad recommendation using the query-ad click graph.	Paper uses click-through data and ranks the relevancy of ads based on the click-graph formed using some collaborative technique. Model implemented is compared to the three baselines that are supposed to be nearly ideal for what today's search engines employ. The model gives better results on an average than currently employed baselines. Problem describes using a bipartite graph. Proposed Query-ad-click-graph.	Click-data is not an ideal model as the data that is received has higher potential to be a noisy or incorrect data due to the fact that the clicks might have been accidental or incorrect. If the rate of position-bias is proved to be more, then the accuracy of the model will reduce even though they have implemented a normalised CTR based on position of the click.
	Text Recognition using Image Processing	The method converts the image to grayscale and pre-process it with noise removal and skew correction, and finally extract the text from the image after segmentation of lines (if present) from the text.	Cannot work with banners and other illustration/images based ads where the pixel density is very high and saturation is difficult. It's only able to extract text from the image, so another algorithm to classify that ad (based on text will be required).

	A client side buffer management algorithm to improve QoE	The paper proposes different algorithms called rate algorithms to enhance user experience by increasing video quality by managing resources of the network and resolving playback buffers. Evaluation is done by implementing the model in the DASH-IF player. Model yields better results than currently employed models by DASH-IF.	The parameters are predefined and might not work for delay or segment longer than 4sec. There are not enough comparisons done with various lengths of segments >4. Thus though the model may seem to work better than current models for parameters less than or equal to the selected parameters, the model might fail for larger parameters. The choice of encoding services (like VBR) chosen by the streaming services might fail the model.
	MetaFlow: A Scalable Metadata Lookup Service for Distributed File Systems in Data Centres	The paper uses a metadata lookup service to distribute lookup workload and increase continuous and large file operations in the file system. It forwards the metadata requests by using the switches to leverage the usual B-tree architecture.	Works for large data storage in data centres. It will increase the response time if the operations are less, which makes the ads loading slow.
	Factors Affecting Online Advertising Recall: A Study of Students	The paper studies the impact of different factors on the effectiveness of an advertisement. The paper particularly is concerned with the relation between page exposure duration and its effect on advertisement.	The paper seems to be more effective on advertisements that are based on memory response rather than instantaneous response. For most of the cases, click-through advertisements will

			prove to be more useful. The ad recall depends on a threshold value that they have concluded. This value is highly dependent on the fact whether the user is on the site skimming for information or is there on the site for a longer duration. Thus, categorizing and labelling such sites should be done manually and by machine learning classification as well.
	Understanding Online Interruption-Based Advertising: Impacts of Exposure Timing, Advertising Intent, and Brand Image	The paper focuses on advertisements based on interruptions. Three parameters are primarily explored are the time taken by the ad to load, the content of the ad and the brand of ad. Authors use persuasive technology, particularly Fogg's principle. Only a pop-up method of advertisement was used in the experiment. ANOVA used for testing of the formed Hypotheses.	Only one product (airplane ticket) was under study. Paper deals with only one type of website- online retail store. Therefore the results may differ for other categories. Only one type of interrupting advertisement was used, that is, pop-up form. This can mislead the results for other forms not under study. The study so conducted deals only with one website, thus the image of the website becomes a key factor in distorting the results that are so obtained by the study.

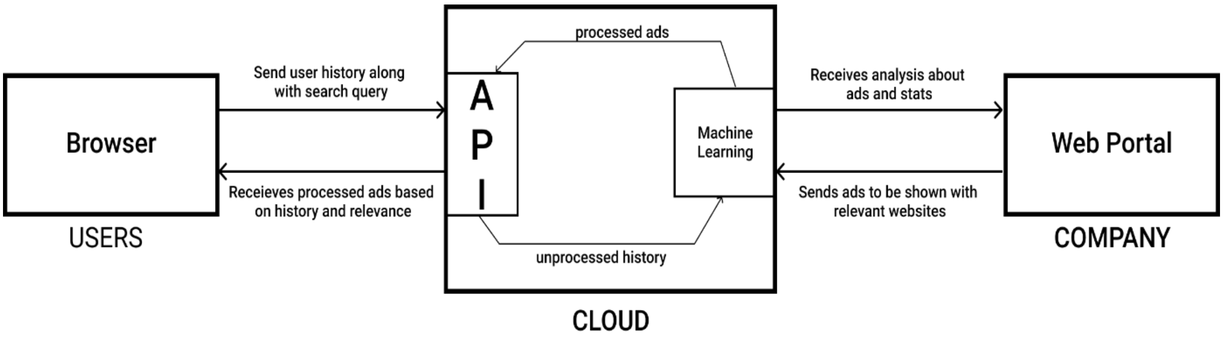
			The study was done on a small group of 180 people from the same environment, thus the actual result will definitely vary for a larger group. A small reward was given to the participants which could be an indication of a reward based system. Thus, in the real world, the user may or may not participate in surveys regarding advertisements if there is no reward for their actions.
	A Survey on Web Tracking: Mechanisms, Implications, and Defences	The paper deals with tracking methods and their implications on privacy of the user. The authors state the underlying threats like those of discrimination and exploitation behind the smokescreen of collecting data for targeted advertisement. The author elaborates on the use of newer technologies like JavaScript, Flash and Java to initiate undesirable transfer/ stealing of data without information of the user. The paper wishes to benefit the user as well as companies that rely on advertisements solely for	The paper reference materials with no scientific backing to draw conclusions about tracking and handling data. The paper suggests the use of different mechanisms and tools to avoid being tracked. One of them is to install an extension that blocks the execution of script which may lead to failure of many websites. The paper also suggests use of VPNs to hide the IP address of the user. Using VPNs may lead to slower services over the internet. The data from VPNs might again

		their income. The author wishes the rules and regulations around the unwanted transfer of data to be more developed and accepted by all the bodies involved in the field of advertisement. Stealing data using cookies, explicit sign-in(form), DOM structure, HTTP cookies, HTTP combined with cookies, cookies being passed to others, Flash cookies , Cache, Fingerprinting, Metadata, Super cookies. Only Firefox prohibits third-party trackers from accessing user data by default. Implications like discrimination of prices, identity theft, and financial situation evaluation were also listed and discussed in depth.	be easily accessible to third-party trackers. Also use of this type of method might be illegal in many countries.\ Opt-out cookies are hardly implemented or are hardly actually used by sites. Do Not Track requests are not implemented by all the browsers. Author also suggests Tor browser whose working gets affected due to limited bandwidth and causes jitters and delays.
	A Clickthrough Rate Prediction Algorithm Based on Users' Behaviours	The paper explains the different features of advertisement click log file. Some of the features make it difficult to draw conclusions on the rate and predict it. The features from click log file are extracted and are converted to numerical data to increase their relevancy and boost potential use. The conversion thus enables the authors to reduce redundancy and scarcity of usable data.	The data under study is taken only by a particular browser. This may affect the model so developed and suggested and may lead to lower accuracy of the same when exposed to datasets from different browsers. The use of GBDT model makes the model slower for large amounts of data. This is because of the fact that

		The K-means model is used for classification of bigger samples. Heuristic methods are used to draw out usable features from the classified parameters. Gradient Boosting Decision Tree model is used for making the already extracted features easier to present. Logistic Regression is used for data with higher dimensions. Tencent SOSO data is used.	the trees in this model are built sequentially and not all at once which delays the procedure. In the paper, the RMSE under different features extracted from logistic regression are closer to the original features than the preferred GBDT model.
	Opinion Mining, Sentiment Analysis and Emotion Understanding in Advertising: A Bibliometric Analysis	Web of Science database was used as the primary source of information. VOS viewer was used for various tasks including network clustering and NLP techniques. ScIMAT was utilised in this paper to majorly study the evolution of the themes that predicts the user behaviour. Classification clusters prove to sustain both the sub categories of periods.	The recent developments in the field of science which deduce the use of intelligence/emotions in advertisements are not considered for this paper. Technologies, trends or patterns that are associated with user demands are not in the scope of this paper.
	Advertising Strategies for Mobile Platforms With “Apps”	The methods analysis the use case of current flow of agents involved in advertisement through the app, its advantages and limitations. It also proposes a new method to involve the	Security issues are reflected with the new proposed method. No organisation would want the advertising agencies to be involved with their application directly.

		advertising agency directly to put forward the advertisement.	The proposed strategy only focuses the problem from the business side view and to make maximum profit, but nothing has been touched upon improving the user experience.
	In-Depth Survey of Digital Advertising Technologies	Different factors like cost of ads, revenue generated from ads and privacy concerns are discussed in this paper. The effect of Social networking on advertisement is studied at length. Primary focus is on in-app advertisement and online ads.	Authors suggest to increase the importance and attention given to libraries and permissions on devices like mobile as their security concerns are higher. But a model should be implemented to handle these tasks and better inform the user as the user may fail to recognise the errors while giving permissions. There is a dire need for applications providing software's, especially like Google Play, to update their permission policies and introduce opt-out permissions according to the needs of the users. Ecosystems for mobiles are limited to iOS and Android.

ARCHITECTURE DIAGRAM


Browser
USERS
Send user history along with search query
Receieves processed ads based on history and relevance
A P I
processed ads
Machine Learning
unprocessed history
CLOUD
Receives analysis about ads and stats
Sends ads to be shown with relevant websites
Web Portal
COMPANY

[Fig. 3.1: Architecture diagram of complete flow]

CONCLUSION

The paper proposes to alter the usual notion of advertisement, in order to increase the user experience and provide more targeted user to the advertising company. This will enable proper and unbiased flow of ads marketing from small start-ups to big companies. Through this paper we are planning to create a novel methodology that will cease the transfer of the user's data without his knowledge to the visiting platform. This will uphold the privacy of a user and avoid the unnecessary transfer of data to other companies that are paying and relying on third-party trackers to get data from the unsuspecting user to improve their advertisements.

REFERENCES

1. Anastasakos, T., Hillard, D., Kshetramade, S. and Raghavan, H., 2009, November. A collaborative filtering approach to ad recommendation using the query-ad click graph. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1927-1930).
2. Kang, S., Jeong, C. and Chung, K., 2020. Tree-Based Real-Time Advertisement Recommendation System in Online Broadcasting. IEEE Access, 8, pp.192693-192702.
3. Tao Mei; Xian-Sheng Hua (2010). Contextual Internet Multimedia Advertising. , 98(8), 1416–1433.
4. Mizan, C.M., Chakraborty, T. and Karmakar, S., 2017. Text Recognition using Image Processing. International Journal of Advanced Research in Computer Science, 8(5).
5. Rahman, W.U., Yun, D. and Chung, K., 2016. A client side buffer management algorithm to improve QoE. IEEE Transactions on Consumer Electronics, 62(4), pp.371-379.

6. Sun, P., Wen, Y., Ta, D.N.B. and Xie, H., 2016. Metaflow: a scalable metadata lookup service for distributed file systems in data centers. IEEE Transactions on Big Data, 4(2), pp.203-216.
7. Chen, G., Cox, J.H., Uluagac, A.S. and Copeland, J.A., 2016. In-depth survey of digital advertising technologies. IEEE Communications Surveys & Tutorials, 18(3), pp.2124-2148.
8. Danaher, P.J. and Mullarkey, G.W., 2003. Factors affecting online advertising recall: A study of students. Journal of advertising research, 43(3), pp.252-267.
9. Chan, J.C., Jiang, Z. and Tan, B.C., 2009. Understanding online interruption-based advertising: Impacts of exposure timing, advertising intent, and brand image. IEEE Transactions on Engineering Management, 57(3), pp.365-379.
10. Bujlow, T., Carela-Español, V., Sole-Pareta, J. and Barlet-Ros, P., 2017. A survey on web tracking: Mechanisms, implications, and defenses. Proceedings of the IEEE, 105(8), pp.1476-1510.
11. Xiong, X., Xie, C., Zhao, R., Li, Y., Ju, S. and Jin, M., 2019. A Clickthrough Rate Prediction Algorithm Based on Users' Behaviors. IEEE Access, 7, pp.174782-174792
12. Sánchez-Núñez, P., Cobo, M.J., De Las Heras-Pedrosa, C., Peláez, J.I. and Herrera-Viedma, E., 2020. Opinion Mining, Sentiment Analysis and Emotion Understanding in Advertising: A Bibliometric Analysis. IEEE Access, 8, pp.134563-134576.
13. Wang, R., Gou, Q., Choi, T.M. and Liang, L., 2016. Advertising strategies for mobile platforms with "Apps". IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48(5), pp.767-778.