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

by Aakash Sharma Kriti

Submission date: 26-May-2021 09:53AM (UTC+0530)
Submission ID: 1594355760
File name: KritiAakash-final-for-plagarism_1.pdf (893.57K)
Word count: 7245
Character count: 37443

CHAPTER 1

INTRODUCTION

1.1 GENERAL

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.

1.2 PURPOSE

1.2.1 Innovation Idea of the Project

The Project uses a ML algorithm to display ads to the user based on the user browsing history as well as website relevancy. It also tries to predict the free space in DOM where the ads can be placed such that there is no hindrance to the original website content and user experience is not disturbed.

1.2.2 Objective of Project

The following are the two major objectives of the project:

a. To better the current user experience by filtering ads based on browsing history and website relevancy.
b. To predict the fine space for placing ads such that user experience is enhanced.

1.2.3 Present System

Today, everyone uses browsers for anything they need. And ads are displayed to them, based on their browsing history.

This history is stored in the user's device itself which is then parsed by the browser to display the ad, as the user visits some particular website.

Now, this ad can be targeted anywhere and on any website, irrespective of its relevance, which gives a bad user experience.

1.2.4 Proposed System

The project proposes a system that is a combination of contextual and behavioural advertisement strategies.

The model that is being used today by most popular search engines like Google AdSense is contextual advertising which aims to generate ads based on page or keyword relevancy. There is a dire need to incorporate behavioural targeting in order to exploit the available user data.

Mining this gold called data we aim to enhance user experience and help companies boost their revenue by meeting their target pool.

1.3 SCOPE

The project displays the ideal behaviour of browsers and search engines. The project tries to make the user's browsing experience much better that what it is today.

Google handles 3.8 million searches every minute. So, one can imagine the number of ads flowing on the internet, which needs to be properly served to the right users.

The proposed idea is cheap and more beneficial to all the potential users (advertising companies)

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.

1.4 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 1.4.1: 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 1.4.2: 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.

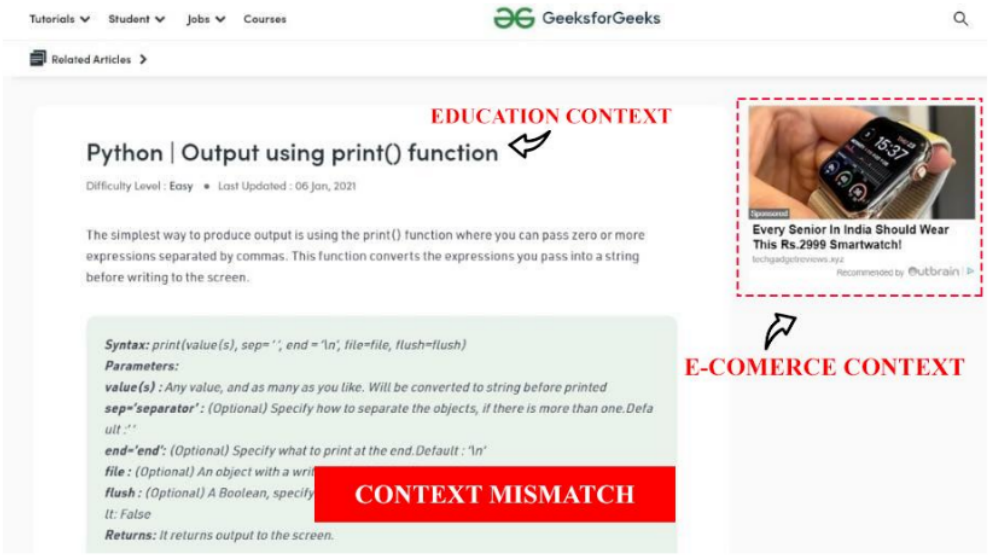

Fig. 1.4.1 Context Mismatch example-1

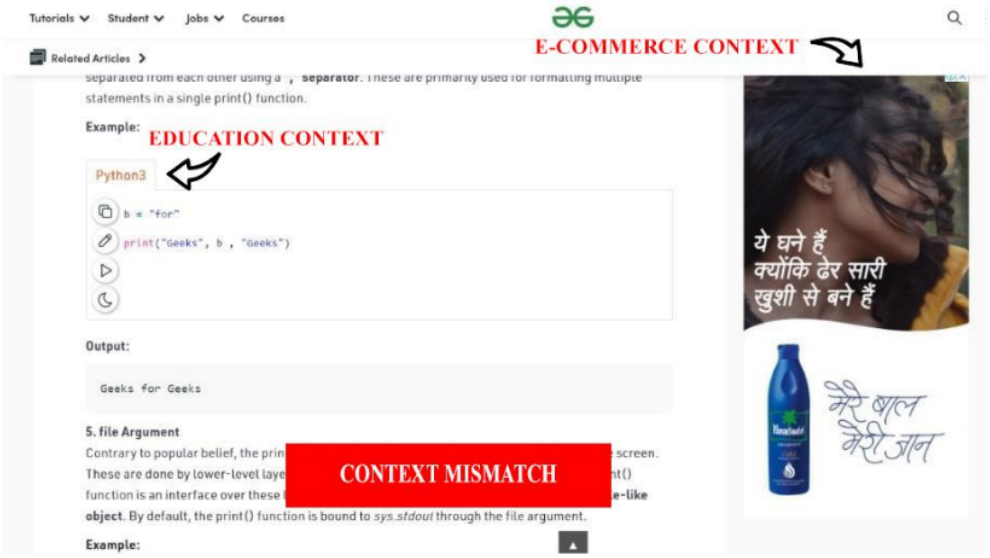

Fig 1.4.2 Context Mismatch example-2

CHAPTER 2

LITERATURE SURVEY

2.1 Tree-Based Real-Time Advertisement Recommendation System in Online Broadcasting

By Seongju Kang, Chaeeun Jeong and Kwangsue Chung

(1)The paper uses a tree model to categorise and filter ads based on the tree skeleton that is proposed. The author makes use of a sorted Hash Map to enable fast tree search.
The paper proposes to store user's information and traverse through trees to find the right category of ads.
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.
Though there is an extensive methodology proposed to categorise advertisements using the tree structure, the problem of getting unnecessary ads is still not resolved.

2.2 Contextual Internet Multimedia Advertising

By Tao Mei, & Xian-Sheng Hua

(2)The paper suggests a second generation of multimedia advertisement using contextual advertisement over behavioural one. The current model used by most of the advertisement generating/posting companies is that of contextual one. The most popular example can be Google's AdSense. But we wish to propose a structure composed of both kinds of advertisement strategies.
The paper proposes the model by showing the most relevant advertisement (both globally and locally (depending on the text and image surrounding)) at an appropriate position.
The authors make extensive use of computer vision and multimedia retrieval techniques.
They also suggest that active tagging be used to categorize images and videos for more success in terms of relevancy of ads. The position of the advertisements should be determined by using a behavioural model. 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.

2.3 A collaborative filtering approach to ad recommendation using the query-ad click graph

By Tasos Anastasakos, Dustin Hillard, Sanjay Kshetramade and Hema Raghavan

(3)This paper uses click-through data and ranks the relevancy of ads based on the click-graph formed using some collaborative technique.

The model implemented by the authors is compared to the three baselines that are supposed to be nearly ideal for what today's search engines employ. At the end of the study, it is concluded that the model so developed gives better results on an average than currently employed baselines.

It should also be remarked that the problem is described using a bipartite graph. In addition to the bipartite graph, a proposed Query-ad-click-graph has also been mentioned to be included in the future models,

It should be kept in mind that Click-data is not an ideal model as the data that is received has higher potential to be noisy or incorrect data due to the fact that the clicks might have been accidental or incorrect. Thus, while evaluating the results against models based not on click-data the dataset in consideration might be affected.

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.

2.4 Text Recognition using Image Processing

By Chowdhury Md Mizan, Tridib Chakraborty and Suparna Karmakar

(4)The paper aims to convert or recognise text from a hardcopy that has been printed. The text so recognised should be converted to the desired format of the user.

The authors proposed a method that converts the image to grayscale and pre-process it with noise removal and skew correction, and then finally extracts the text from the image after segmentation of lines (if present) from the text.

The only drawback is that the model becomes unsuccessful or cannot work with banners and other illustration/images based ads where the pixel density is very high and saturation is difficult. It is only able to extract text from the image, so another algorithm to classify that ad (based on text will be required).

Thus, for image based advertisements, we would suggest another algorithm.

2.5 A client side buffer management algorithm to improve QoE

By Waqas ur Rahman, Dooyeol Yun, and Kwangsue Chung

(5)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 of the proposed model is done by implementing the model in the DASH-IF player. DASH is a Javascript based player.

When the model was tested against predefined parameters, the paper discovered that the model yields better results than currently employed models by DASH-IF.

Since the parameters are predefined and might not work for delay or segment longer than 4sec as mentioned in the paper.
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.

2.6 MetaFlow: A Scalable Metadata Lookup Service for Distributed File Systems in Data Centres

By Peng Sun, Yonggang Wen, Duong Nguyen Binh Ta and Haiyong Xie

(6)The paper suggests a new look-up service model for the current DHT or distributed hash table based metadata management systems. 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. The drawbacks could be that it only works for large data storage in data centres. Since it is not tested against smaller data storage, the accuracy or results may be depreciated.

The thus proposed model will increase the response time if the operations are less, which makes the ads loading slowly. Due to the slower rate of loading of advertisements, the browsing speed and experience may be affected.

2.7 Factors Affecting Online Advertising Recall: A Study of Students

By Peter J. Danaher and Guy W. Mullarkey

(7)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 methodology proposed by the paper 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. But it is known that click-through advertisements can generate noisy data due to the fact that the clicks can be unintentional or faulty.

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.

2.8 Understanding Online Interruption-Based Advertising: Impacts of Exposure Timing, Advertising Intent, and Brand Image

By Jason C. F. Chan, Zhenhui Jiang and Bernard C. Y. Tan
(8)This paper inspects the three design factors that constitute the Interruption-based advertisement strategy for online services. The design factors so tested are 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 the ad.

The authors make use of persuasive technology, particularly Fogg's principle. The most crucial limitation is that only a pop-up method of advertisement was used in the experiment. Advertisement forms like pop-under remained unexplored. Since, only one type of interrupting advertisement was used, that is, pop-up form. This can mislead the results for other forms not under study

It should also be remarked that ANOVA has been 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 of advertisements.

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. Different industries may use different standards for the image they use for ads.

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.

2.9 A Survey on Web Tracking: Mechanisms, Implications, and Defences

By Tomasz Bujlow, Valentín Carela-Español, Josep Solé-Pareta, and Pere Barlet-Ros
(9)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 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. The paper references 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 scripts 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 be easily accessible to third-party trackers. Also use of this type of method might be illegal in many countries. The paper suggests the informed use of Opt-out cookies but they are hardly implemented or are hardly actually used by sites. Do Not Track requests are not implemented by all the browsers so, it fails the purpose.

Author also suggests Tor browser whose working gets affected due to limited bandwidth and causes jitters and delays.

2.10 A Clickthrough Rate Prediction Algorithm Based on Users' Behaviours

By Xi Xiong, Chuan Xie, Rongmei Zhao Yuanyuan Li, Shenggen Ju and Ming Jin
(10)The paper explains the different features of advertisement click log files. 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 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. It is to be noted that the dataset used is Tencent SOSO 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 the GBDT model makes the model slower for large amounts of data. This is because of the fact that 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.

2.11 Opinion Mining, Sentiment Analysis and Emotion Understanding in Advertising: A Bibliometric Analysis

By Sanchez-Nunez, P., Cobo, M. J., de las Heras-Pedrosa, C., Pelaez, J. I., and Herrera-Viedma

(11)The paper aims to fill the gap between the old advertisement strategies and newly developed neuroscience based strategies. Thus, they aim to analyse the relationship between the AI based methods that are being implemented to get relevant advertisements on a large scale. The roles of Opinion mining, sentiment analysis and emotion understanding with respect to advertisement are explore. This article analyses those works that address the relationship between sentiment analysis, opinion mining, and emotion understanding in advertising. The roles of these AI domains are defined and their significance is established in the paper.

WoS or 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.

2.12 Advertising Strategies for Mobile Platforms With "Apps"

By Wang, R., Gou, Q., Choi, T.-M., and Liang, L.

(12)The paper explores the potential of advertisements using mobile based applications. The authors propose the use of a game-theoretical model to bridge the gap between the user and provider. The authors also mention the negative impact of the aggressive advertisement approach. They deem the interaction result and experience to be negative when the users are bombarded with forceful advertisements. To make the users participate and interact more with the ads, they suggest a more playful approach to keep themenggaed.

The methods analyse the use case of the current flow of agents involved in advertisement through the app, its advantages and limitations. It also proposes a new method to involve the advertising agency directly to put forward the advertisement. Security issues are reflected with the new proposed method. No organisation would want the advertising agencies to be involved with their application directly.

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.

2.13 In-Depth Survey of Digital Advertising Technologies

By Chen, G., Cox, J. H., Uluagac, A. S., and Copeland, J. A
(13)This paper explores the digital advertising ecosystem and the relationships across its subcategories. It also explores the major issues with the complex platforms the advertisements are represented in and their impact on crucial real life events like economics, finance, politics and much more.

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. But the primary focus remains on in-app advertisement and online ads and the complex system they are posted in.

Authors suggest increasing 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, 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 in the research thus the complexity can vary for other mobile systems.

2.14 A Context-aware Recommendation System Based on Latent Factor Model

By Zhenling Zhang, Yingyuan Xiao, Wenxin Zhu, Xu Jiao, Ke Zhu, Huafeng Deng and Yan Shen
(14)The paper cites the benefits of a recommendation system for contextual advertisement. They propose a new model called C-LFM that serves as a recommendation model. This model aims to add contextual information to Latent Factor Model (LFM) to enhance results obtained by recommendations. They test the model against various factors to test the effectiveness and differentiate them from the previous models based on similar concepts.

The main performance metric the model is tested against RMSE which is largely dependent on the correlation coefficient.

The hardware requirement of the above experiment is limited to the Dell PC with 8G memory and 64 bit win10 operating system. This leaves out a majority of systems and the results remain unexplored for smaller devices like mobiles.

The number of latent factors is also limited to 10. The contextual media is considered as a subpart of Latent Factor Model in order to reduce the dimensional complexities. LFM also uses the stochastic gradient descent method. This is done in order to optimize the ranking loss. The proposed method is found to be working effectively for the current system and it is proposed without any validation that it may work with accuracy for other systems too.

2.15 Some Effective Techniques for Naive Bayes Text Classification

By Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng
(15)This paper proposes the importance of Naive Bayes in Data mining tasks. Naive Bayes is deemed ineffective when tested against automatic text in classification problems. The discrepancies in the result are attributed to estimation process parameters.

The paper thus proposes two heuristics: per-document text normalization and feature weighting method. The dataset used is Reuters215784 and 20 Newsgroups. This paper proposes a Poisson Naive Bayes text classification model with weight-enhancing methods. The underlying assumption is that a document is generated by a multivariate Poisson model.

To compensate and rectify errors for lack of training data for the minor categories, they suggest enhancing weights. The result seems successful in building probabilistic text classifiers. But the cost of time and space is increased and can lead to slower environments built based on this model.

No exceptional results are seen when the heuristic term frequency transforming method, and pivoted length normalization, is used. The proposed system is deemed effective in case of a spam-filtering system or an adaptive news-alert system. For future prospects, adaptive filtering and relevance feedback should be included.

2.16 Inference

The above working papers of literature survey show that though contextual advertisement form is a prevalent and rather preferred form of advertisement strategy currently, there is a dire need to shift to behavioural advertisements. With the new developments in the fields of Artificial Intelligence, domains like Opinion Mining and Sentiment Analysis can be implemented to reach better or pure audiences on a larger scale.

We have also discovered that older methodologies for determining the target pool for advertisement need a revising. On-click advertisements, if implemented, need a more refined approach.

The security issues with the digital advertising ecosystems are studied at length. We consider all the aspects and conclude that though ad-blocking might be more effective for the users, uninformed use of ad-blockers, opt-out cookies, permission policies will still lead to data leak and misused. Ad-blockers in turn also negatively impact the advertising industries that depend largely on the revenues generated by the advertisements they post on various platforms. Therefore we suggest an alternative new method that improves the target pool and user experience at the same time by combining the perks of Contextual and Behavioural advertising strategies.

For the behavioural part, we plan to implement text categorization for text based advertisements using NLP techniques. Sentiment Analysis can also be explored for different advertisements. Text recognition using Image recognition can be implemented to convert the text of image based advertisements and then the text thus obtained can be used to categorise

the advertisement. The image based advertisements can be categorised using Image Recognition and be stored using Hash Maps.

The ads that are categorised can be then posted to relevant websites belonging to the same category, thus implementing the contextual advertisement too.

This hybrid model will be successful for future technologies and has the ability to change the advertising field drastically.

CHAPTER 3

SYSTEM REQUIREMENTS

3.1 Hardware Requirements

A system with following minimum requirements:

1. Intel Core i7-8700k CPU with 6 cores
2. NVIDIA Ge-force Titan X pascal GPU

3.2 Software Requirements

1. JavaScript
2. HTML
3. CSS
4. NodeJS
5. ReactJS
6. TypeScript
7. Express
8. Google API
9. Mongo Database
10. Google Collab
11. Python
12. Numpy
13. Pandas
14. Skitlearn
15. MatPlotLib

CHAPTER 4

MODULES DESCRIPTION

4.1 Javascript

Javascript is a browser scripting language for web. Using it one can manage how web pages behave on user's interactions and perform tasks internally in the browser. It acts as the backbone of web pages. From sending network request to receiving data, everything on a web page is done using javascript.

4.2 HTML

Hyper Text Markup Language (HTML) is a markup language designed especially for web pages. It has tags like xml which defines the elements of a web page. All the text, buttons, images on a webpage are displayed through HTML.

4.3 CSS

Cascading Style Sheets (CSS) defines the UI of a web page. It tells the properties of an HTML element. These revolve around their size, color, positioning etc. It has an object like structure with key value pairs (usually strings) that define the property of the element to which the style is targeted at.

4.4 NodeJS

It is a javascript runtime environment based on chrome v8 engine. It is designed to handle network requests at scale. It is a single threaded environment which means there is no need to worry about deadlock conditions on the server. It supports both synchronous and asynchronous programming.

4.5 ReactJS

It is a javascript library which extends the features of vanilla javascript by re-rendering only components which are changed. Hence optimising the DOM rendering and loading the content faster. It is declarative and component based.

4.6 Typescript

It is a language built on top of javascript to ensure type errors that leads to performance issues with javascript. Typescript ensures type checking between the object, hence it becomes easier to debug errors and hence saves a lot of time in development and testing.

4.7 Express

It is a javascript framework that works along with NodeJS server to handle network calls efficiently. It is unopinionated and minimalistic hence ensures faster handling and minimising server load. It provides easy functions to handle each type of requests and to manage data accordingly from the network call.

4.8 Google API

It provides various functions as an extension to the browser capabilities to perform certain actions in terms of an extension that can be installed in the browser as plugin to consume its features. It allows to manage chrome data and handle user actions as required to ensure better experience.

4.9 Google Collab

Google collab is an environment used to run executable code. Its major advantage is that it runs entirely on cloud and thus provides a faster way to run codes related to ML and Data analysis.

4.10 Python

Python is a high-level programming language. Its advantage is in terms of indentation and the readability of code. It is an interpreted language. It has a large support and community for almost every technology.

4.11 Numpy

Numpy is an open-source Python library that is used to perform operations on large scale muti-dimensional arrays. It is also used for matrices.

4.12 Pandas

Pandas is an open-source library written for data manipulation of text, excel and csv files for Python Programming language. It is very easy to implement and is fast and efficient which makes it a popular choice among programmers.

4.13 Skitlearn

20

Scikit-learn is an open-source machine learning library based on the Python programming language. It is majorly used for predictive data analysis. It consists of a variety of ready-made algorithms for classification, regression and clustering. It supports SVM too. The fact that it is interoperable with Scipy and Numpy, makes it a prime choice for anyone looking to implement machine learning algorithms.

CHAPTER 5

METHODOLOGY

The contextual advertisement uses the content of the current page to predict and display advertisements. The placement of the advertisement itself seems to bear no significance, according to the recent studies. The advertisements thus showed bear relevancy to some extent. But this advertisement method has the potential to improve massively.

In this report, behavioural targeting is integrated to achieve preferable results. To simplify our procedure, we have divided it into four major parts.

5.1 ARCHITECTURE DIAGRAM

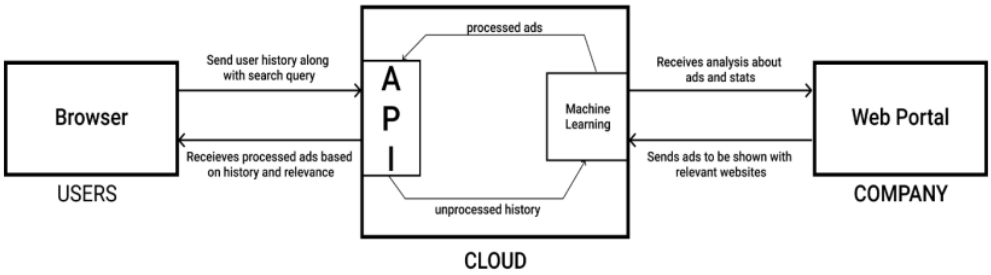

Fig. 5.1.1 Architecture diagram

The architecture diagram in our project consists of these key players: The User, Cloud and Company. The basic flow of the data is from 1-Ads from the Company to the Web Portal, from the Web portal to the ML model, from the ML model to the API. 2- Metadata from the user's browsing history to the API, then from the API to the ML model for prediction.
The Web Portal that is developed for the company serves as an input as it receives data from the company in the form of advertisements. The advertisements consist of an image and a description along with relevant tags. The data description is thus sent to the Machine Learning (more precisely a Text Classification model based on NLP algorithms) model in the Cloud Database. The data is refined and ads are processed and sorted into predefined categories. Thus, the advertisements from the company are processed and sorted.
When the user browses a website, its metadata or more precisely description is fetched and sent to the API. This unprocessed description is processed by the API and sent to the Machine Learning model to be categorized. The category of the website is predicted. Therefore now, we have the predicted category of advertisement from the company stored in our database as well as the predicted category of the website being browsed currently. The category of the website is matched with the category from the database. If a match is found, a relevant advertisement is sent to the API to be sent to the user's screen/website to be displayed. Thus, maintaining relevance and enhancing user experience.

5.2 WEB PORTAL

The interactive web portal consists of an upload option and an insight option. The company can upload the image advertisement and select/define tags for the same. The company's representative will also receive regular insights for the advertisement. To implement the insight generation, the paper proposes the use of Data Analytics to depict monthly views, clicks and buys through the advertisement.

Fig. 5.2.1 Components of the Web Portal

5.3 TOPIC CLASSIFICATION

Topic Classification is a part of Natural Language Processing that uses Deep Neural Networks (or any suitable algorithm that yields the highest accuracy) to predict or classify the text. It is a supervised learning method. Topic Classification in this paper aims to predict the category of the advertisement. Initially, Topic Modelling, an unsupervised learning method, was opted, with predefined tags, but the accuracy of the model was less than unexpected. To improvise, Topic Classification was chosen as a better alternative. To achieve topic classification, multiple NLP algorithms have been implemented and tested on the 'Advertisement Transcripts from Various Industries' Dataset. Our aim is to define the category of an advertisement. For Example, if the advertisement contains phrases like 'cloth,' 'sale,' 'discount' - the same will be classified as an 'E-commerce ad.' This tag helps the model pick this advertisement and display it on any E-commerce site to increase relevancy.

5.4 DATASET

The dataset used for this paper is 'Advertisement Transcripts from Various Industries' from Kaggle. It consists of approximately 2000 advertisement descriptions along with their category and corresponding companies.

The classes are completely mutually exclusive. Consequently, there is no overlap between batches.

5.5 PRE-PROCESSING

The dataset consisted of two essential and two inessential columns. The inessential columns ('Advertiser' and 'Product or spot) are dropped during the pre-processing step.

The dataset is visibly highly imbalanced (Fig 5.4.1). The category in the majority being 'Automotive' and the minority category being 'HealthCare'. To remove the inconsistencies in the data, and to achieve oversampling without over fitting, SMOTE is used at a later stage.

The textual data present in the 'Ad copy' column is cleaned by removing the stop words and bad characters/symbols (some applications of NLP).

TFIDF is used on the target class to provide the relevancy of the words with respect to the entire document thus, can be interpreted by the Machine.

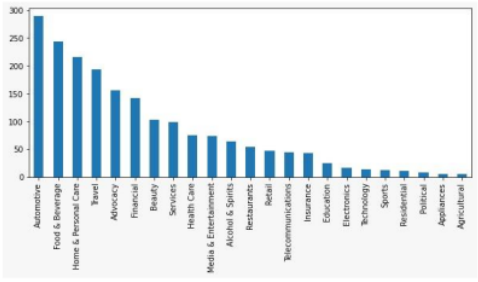

Fig. 5.5.1 Imbalanced Categories in the dataset

5.6 MODELS

Preliminary to testing the model, chi-squared test using unigrams and bigrams are used to find the most relevant/correlated words per category.
The data needs Multiclass Classification. Multinomial Naive Bayes Classifier is used as it uses a multinomial distribution for each of the features.

The paper tests different parameters to yield the best accuracy for the predicted models while maintaining the minimum amount of loss suffered (Fig 5.6.1). The data is tested against these four models-

A. Logistic Regression
B. (Multinomial) Naive Bayes
C. Linear Support Vector Machine
D. Random Forest

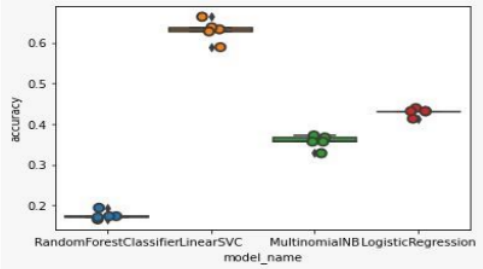

Fig. 5.6.1 Accuracy of different models against the test data

The result of the accuracies is given in figure 5.6.2

```
model_name
LinearSVC                0.630491
LogisticRegression       0.429457
MultinomialNB            0.356072
RandomForestClassifier   0.175194
Name: accuracy, dtype: float64
```

Fig. 5.6.2 Accuracy results of different models

The best result is yielded by Linear Support Vector Machine. A model is built and is cross-validated by a confusion matrix and indicated using a heat map. (Fig 5.6.3)

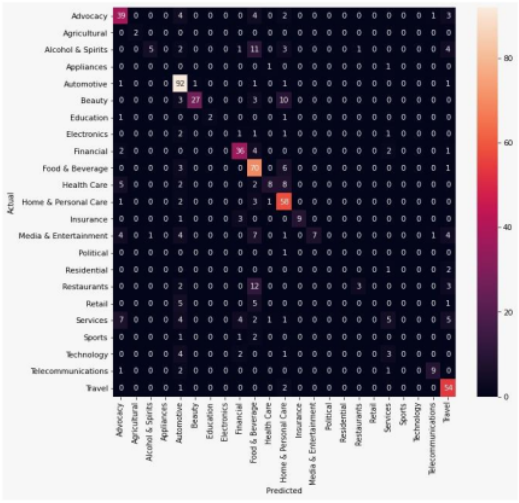

Fig. 5.6.3 Linear SVM matrix heat map

The model evaluation score per category displayed in Fig 5.6.4. The metrics the model is evaluated against are precision, recall, and F1 score. These results are yielded without the application of SMOTE. To further improve the results, SMOTE is applied and the model is again developed and evaluated against the same metrics.

Accuracy of up to 84 percent is achievable using the above method.

	precision	recall	f1-score	support
Advocacy	0.64	0.74	0.68	53
Agricultural	1.00	1.00	1.00	2
Alcohol & Spirits	0.83	0.19	0.30	27
Appliances	0.00	0.00	0.00	2
Automotive	0.69	0.95	0.80	97
Beauty	0.96	0.63	0.76	43
Education	1.00	0.50	0.67	4
Electronics	0.00	0.00	0.00	6
Financial	0.75	0.80	0.77	45
Food & Beverage	0.55	0.88	0.68	80
Health Care	0.73	0.32	0.44	25
Home & Personal Care	0.60	0.89	0.72	65
Insurance	1.00	0.69	0.82	13
Media & Entertainment	1.00	0.24	0.39	29
Political	0.00	0.00	0.00	1
Residential	0.00	0.00	0.00	3
Restaurants	0.75	0.15	0.25	20
Retail	0.00	0.00	0.00	11
Services	0.36	0.17	0.23	29
Sports	0.00	0.00	0.00	4
Technology	0.00	0.00	0.00	10
Telecommunications	0.82	0.69	0.75	13
Travel	0.68	0.95	0.79	57
accuracy			0.67	639

Fig. 5.6.4 Model Evaluation Score per Category

5.7 PROCESSING USERS DATA

Users' data is a pivotal point in behavioural targeting. With the availability of users' browsing data, the model will select the pertinent advertisement from our database. For Example, if the user has visited an E-commerce site, say, 'X' before and has browsed the site for clothes, now when the same user visits any other E-commerce site 'Y' with the same intent, the advertisement from the previously browsed E-commerce site 'X' will be displayed on site 'Y.'

Thus, the user experience is maintained, and the probability of companies reaching their target pool increases exponentially.

To tag the current site, a separate model will be used to classify the site into different topics according to the site's content.

5.8 DISPLAYING PROCESSED ADS

Now that the data is processed, the machine learning algorithm will send the best suited Ad as a result, this Ad will be then picked up by the service plugin from the AWS S3 bucket and will be injected in the DOM of the website user is currently in.

The chrome extension which is been setup in the browser will hence pick this image from the bucket, and inject it onto the webpage the user is currently in. In this way the correct ad will be displayed to the end user.

CHAPTER 6

RESULT

In the first iteration, we try to build the architecture design as shown in Fig.V.1, the figure represents an end to end flow. This gives an insight to the micro-service interactions following the model-view-controller design pattern. This will hence enable the service scalable to a large scale without the intervention of other services or dependency upgrades. Here, the three major components are browser-plugin, the cloud service API and the Web Portal for Advertising Company. The diagram represents the flow of interaction of one micro-service to another. The browser sends the user's data

In the second iteration, we try to automate the UI flow through which ads can be stored directly into the database pool with its tags and description. We followed no-sql database for fast query fetching and traffic look ahead. This ensures quick iteration by the machine learning model to read the tags on advertisements and cluster them accordingly. The UI also provides analytics to the Advertising Company consuming libraries like D3 and others. The images are stored in the Cloud storage to ensure large bandwidth.

In the third iteration, we have connected our UI to the backend that performs CRUD operations to the users and posts data. The backend makes a connection between the Mongo DB client and AWS S3 for cloud storage. It listens to various incoming requests from the browser and process them accordingly. It also manage user authentication, to avoid unnecessary creation of ads.

In our fourth iteration, we tried to figure out and classify our data according to accuracy of different models. We try to balance our dataset and then put it to train the model to obtain high accuracy. In this iteration, we also try to figure out different ways to classify the images from the Advertisement post.

In this last iteration, we try to extract user's data from web crawlers and send it to the flask API to process it and give back the recommended ad. This Ad we then parse to the user's website DOM, and hence the ad is displayed to the user. The API consumes Tensorflow.js and has direct access to cloud storage from where it sends the ads to the user. Enabling this not only provide excess of data but the bandwidth is also increased to a greater extent.

CHAPTER 7

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.

In this paper, we have demonstrated the advertising structure that uses both contextual and behavioural advertising strategies. However, existing system suffer from large efforts being wasted to find the right audience for the right advertisement. Today, a system is established where it's all about reaching the target goals for posting the ads, which is not profitable for both the entities, but for the advertising platforms. These platforms generate a large revenue just by posting ads, distributing it, and collecting insights. But the advertisement company are still not able to reach their target audience, at the same time it decreases the experience of every other user visiting the platform.

We have established the necessity for a combination methodology to enhance user experience. We have used Machine Learning and NLP algorithms extensively throughout this paper to classify text or topic and to classify the sites as well. The accuracy of 84 per cent has been achieved following the methodology so proposed.

For the future prospects, we also suggest an image recognition system to define tags for an image advertisement. This paper will prove beneficial to those who are looking to incorporate artificial intelligence and big data in the advertisement field.

CHAPTER 8

FUTURE ENHANCEMENT

The project has provided a basic prototype of an ideal browser behaviour, which caters to user experience as well as the revenue generating companies. Advertisements giants like Google AdSense or Facebook Ads have a complete business model through which the target is completed. It means that their service will make sure to display the ads irrespective of different categories of platforms just to meet the target. But in order to display the ads, in such a manner, they neglect the user experience as the ads may be displayed on sites irrelevant. The proposed system here uses text classification to categorise ads based on the context. This can be further improved drastically after the implementation of image classifier. Hence the requirement to manually add categories will no longer be required. It will also act as a fairer means to categorise ads, and human errors will be minimised.

The business model incorporates both contextual and behavioural advertising, so it encounters a large number of computations in terms of processing the request and the training the machine learning model and then deploying. Scaling it and adding load-balancer will give some relief to the servers and make them more optimised.

It can also incorporate advertisements statistics and analytics which can be beneficial for the advertisement companies. These can be implanted from fetching the add data and displaying them using graphs from popular libraries like Power Bi etc.

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

ORIGINALITY REPORT

7%	6%	4%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ir.kaist.ac.kr Internet Source	1%
2	researchr.org Internet Source	1%
3	Zhenling Zhang, Yingyuan Xiao, Wenxin Zhu, Xu Jiao, Ke Zhu, Huafeng Deng, Yan Shen. "A context-aware recommendation system based on latent factor model", 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2017 Publication	<1%
4	Peng Sun, Yonggang Wen, Duong Nguyen Binh Ta, Haiyong Xie. "MetaFlow: A Scalable Metadata Lookup Service for Distributed File Systems in Data Centers", IEEE Transactions on Big Data, 2018 Publication	<1%
	en.wikipedia.org	

5	Internet Source	<1%
6	www.researchgate.net Internet Source	<1%
7	Submitted to University of Western Sydney Student Paper	<1%
8	www.semanticscholar.org Internet Source	<1%
9	www.smartinsights.com Internet Source	<1%
10	Submitted to sgscol Student Paper	<1%
11	www.coursehero.com Internet Source	<1%
12	Seongju Kang, Chaeeun Jeong, Kwangsue Chung. "Tree-Based Real-Time Advertisement Recommendation System in Online Broadcasting", IEEE Access, 2020 Publication	<1%
13	Waqas Ur Rahman, Dooyeol Yun, Kwangsue Chung. "A client side buffer management algorithm to improve QoE", IEEE Transactions on Consumer Electronics, 2016 Publication	<1%
14	journals.cambridge.org Internet Source	

		<1%
15	export.arxiv.org Internet Source	<1%
16	www.wordstream.com Internet Source	<1%
17	Submitted to Edith Cowan University Student Paper	<1%
18	scholars.cityu.edu.hk Internet Source	<1%
19	www.jtsinstitute.biz Internet Source	<1%
20	Submitted to The University of Manchester Student Paper	<1%
21	99firms.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		