MINING ADVERTISEMENT USING BIGDATA ANALYTICS

K.Sathiyamurthy

Associate Professor, Department of Computer Science and Engineering, Pondicherry Engineering College, India
sathiyamurthyk@pec.edu

P.Dhivya

Student, Department of Computer Science and Engineering, Pondicherry Engineering College, India
dhivyap519@pec.edu

Abstract- Technology advancement have drastically changed online advertising, most notably in the ability to measure advertising outcomes and in displaying targeted advertisements. Online publishers and online advertisers are now screening increasing interest in using targeted online advertising for the purpose of tracking users across web sites in order to infer their interests and preferences. Thereby, deploying optimized advertisements using their preferences and interests.For mining advertisement, predicting web-browsing behavior of each individualon the internet is more important. This prediction helps any advertisers or the publishers to successfully interact with users in providing relevant advertisement, many intelligent interfaces requires a technique for analyzing, recognizing and predicting user behavior actions. Initially we need to collect a log of datasets with user behavior attributes. Hidden Markov Model is used to derive a pattern for predicting the various behavior of the users on the web. Also, it helps in analyzing the performance of user behavior profile and estimating the advertising cost to perform Mapreducejobs based on user's search query.Using click through rate the effectiveness of advertisement can be evaluated here.With allthese collected information, optimized advertisement is deployed to the user using Bigquery mechanism.

Keywords*: Click through rate, Mapreduce, Advertisement mining, Hidden markov model, Bigquery.*

I. INTRODUCTION

Online advertising hasbecame the potential areas of research nowadays. They have aggravated not only studies and research in different fields, but also introduced many new challenges. It is a generalform of controlled communication that attempts to persuade users, using various strategies and appeals,inorder to buy or use a particular product. This advertising world has now became incredibly complex with many involved parties playing different roles. The major roles involved in onlineadvertising includes Advertiser, Publisher, Ad-network. Ad targeting takes many variables into account: the advertiser's budget for presenting ads, the content of the page on which the ad is to be placed (called contextual targeting), and the user viewing the ad (called behavioral targeting)[1].However,the main challenge in this type of advertising includes how to extract the user's interest from web pages in a diffuse context and this problem can be resolved by adapting behavioral targeting technique that offers the ability to create a personalized experience for each user. Moreover, meeting the requirements of real time application with the huge data is quite difficult to handle. Hence the system should be able to deal with real time by serving users with appropriate advertisement using bigdata analytics.

Big data analytic solutions and services is indispensable to support variousbusiness people to perform business market analysis and to craft intelligent decisions. Big data includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, visualize and process the data within a tolerable elapsed time[1][2]. Big data denotes combinations of data sets whose volume, variability

and velocity makes them difficult that can be captured, managed, processed or analyzed by relational databases, within the particular time necessary to make them more useful.Optimization of relevant advertisements to users can be incorporated by adopting suitable stochastic model.

This project also reports about the experimental work done on big data analytic services in mining advertisement and provides an optimal solution using Hadoop Distributed File System (HDFS) for the storage purpose and thereby processing such large data sets using the Map Reduce programming framework. Analyzing the performance of user behavior profile and estimating the value for advertising cost to perform Mapreduce jobsbased on user's search query can be implemented by incorporating Hidden Markov Modeltechnique[3]. Click through rate is the ad factor used to estimate the advertising cost and to increase the effectiveness of advertisement.After such estimation and analysis, finally retrieval of relevant advertisement is deployed to the user using Bigquery.The next section discusses about the existing work carried out in mining advertisement and Bigdata analytics.

II. RELATED WORK

Behavioral targeting of online advertisements is a relatively new phenomenon, Beales (2010)[4] uses data collected from online advertising networks to find prices and conversion rates (i.e., the likelihood of a click eventually leading to a sale) for behaviorally targeted advertisements. In their work they complement these empirical findings for prices by analytically studying the effect of behavioral targeting on advertisers' payoffs, as well as on publishers' revenues. Jun Yan(2009) [14], address how behavioral targeting enhances the online advertisement. In their they found three significant conclusion on Behavioral targeting advertisement.

Extracting relevant information from a huge data dump has been described byJyotiNandimath 2013**.**In their work they have proposed one of the best open source tools to harness the distributed architecture in order to solve the data processing problems is Apache Hadoop. Various Apache Hadoop's components such as map-reduce algorithms and distributed processing, resolve various location-based data problems and provide the relevant information backinto the system and also increasing the user experienceThe application performs the operation on big data like counting average ratings, total recommendations, unique tags etc, in optimal time and producing an output withminimum utilization of resources. The data analysis and processing is used in a social networking application. Thus providing the necessary information to the application users with least effort**[**5**].** The process of revealing secret correlation and pattern using Hadoop MapReduce was described by Seref Sagiroglu (2013) [15].

Displaying ad for users based on prior searches and page reviews has been described byBadrishchandramouli and Jonathan Goldstein(2012).In their work, they proposed display advertising using behavioral targeting to scale well for big data using map-reduce clusters.Their contributions are twofold. First, they propose a novel framework called TiMR (pronounced *timer*), that combines a time-oriented data processing system with a M-R framework. A new cost-based query fragmentation and temporal partitioning schemes have been proposed for improving efficiency with TiMR. Second, they showed that the feasibility of this approach for BT, with new temporal algorithms that exploit new targeting opportunities. Experiments using real advertising data show that TiMR is efficient and incurs orders-of-magnitude lower development effort. Their BT solution is easy and succinct, and performs up to several times better than current schemes in terms of memory, learning time, and click-through-rate/coverage[6].

Introduction to Hadoop HDFS and MapReduce for storing large number of files and retrieve information from these files has been described by Amrit pal and kunal Jain(2014).The tera bytes size file can be easily stored on the HDFS and can be analysed with MapReduce. Thereby analysing the behaviour of the map method and the reduce method for increasing number of files and the number of bytes written and read by these tasksThey have analyzed the performance of the map reduce task with the increase

number of files. The output shows that the Bytes written do not increase in the same proportion as compared to the number of files increase. It can also be used for analyzing the sensor's output which are the number of files generated by the different sensing devices[7].

Jianqing Chen, Jan Stallaert(2012) have analysed the economic implications when an online publisher engages in behavioral targeting to present users with relevant advertisements based on their past browsing and search behavior and other available information (e.g., hobbies registered on a website).They have also defined that the revenue can double in some circumstances for online publisher using behavioral targeting. On the other hand, increased revenue for the publisher is not guaranteed in some cases and hence the publisher's revenue can be lowered, depending on the degree of competition and the advertisers' valuations[8].

Behavior targeting is one of the maketing method in online to by collecting the data of browser activities of the customer to find more target online advertisement to the customer[17]. Ahmed, Amr, et al. (2011) have discussed the previous user profile generation to predict the user behavior in online for targeting advertisement and social personalization. In their work they used a scalable distributed inference algorithm to handle thousands users and proposed a model to improve the behavioral targeting of displaying advertisement[16].

Chun-Jung Lin (2012) has applied the Hidden Markov Model in predicting the behavior of the users on the web. In their work, they have collected the log of web servers, clean the data and patch the paths that the users pass by. Based on the HMM, they constructed a specific model for the web browsing that can predict whether the users have the intention to purchase in real time. The related measures, such as speeding up the operation, kindly guide and other comfortable operations, can take effects when a user is in a purchasing mode. The simulation shows that their model can predict the purchase intention of uses with a high accuracy[9].

III. PROPOSED METHOD

3.1 Architecture diagram

The below figure demonstrates datasets collection (structured & unstructured) for mining advertisement, uploading those collected datasets into appengine, performing MapReduce programming. Data analysis for mining advertisement is done by building HMM model, thereby generating Log-likelihood values for predicting pattern and examining the efficiency of advertising cost in HDFS framework. Finally retrieving relevant advertisement by executing Bigquery.

3.2 Modules

The modules included in proposed approach are as follows.

1. Data Collection
2. Data storage
3. Data Analysis
4. Data Retrieval

- Data collection

The Dataset is collected from various resources and websites that are related to advertisement in the format of the csv (comma separated values) as structured, unstructured and it constitutes the size of 1 Gigabytes. The dataset includes various attributes such as Item, session, Location, Category, Click through rate, Cost per click, Cost per impressions, Conversion rate, Cost per action. These collected datasets classified into

structured and unstructured data then stored in Appengine and is processed using MapReduce framework to find the pattern of the user behavior.

- Data storage

In MapReduce programming framework for mining advertisement, all these collected data along with user behavioral patterns were processed in isolation by mappers and reducers tasks. This MapReduce framework execute in parallel to find the user pattern on browser (for an instance the user browsing to purchase a mobile, how the user searches for his desired mobile). The obtained resultant pattern is used to generate the user behavior profile is imported to the datastore (Hadoop distributed File System) and Appengine. This user behavior profile helps to improve the effectiveness of the advertisement. The images allied to each ad related keywords are stored in the data store as a blobkey. Blob key indicates reference to the file on the blob store. Then the images for theeach keywords are imported to the blob store using blob key.

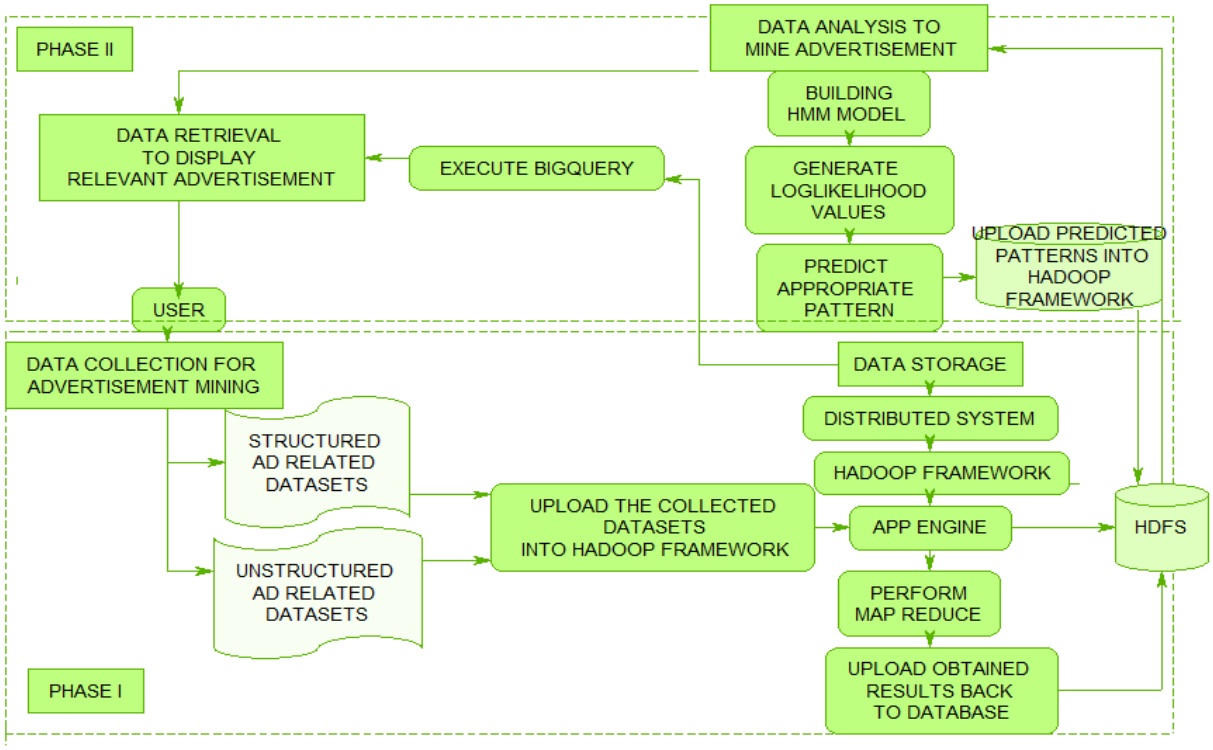

Figure 1. Proposed System Framework

- Data analysis

After importing collected datasets into appengine, next step is to analyze these datasets in order to increase the effectiveness of their advertising using behavioral targeting technique. Thereby, displaying optimized advertisement by utilizing this above information. Later, applying this click stream data into hidden markov model to uncover consumer browsing patterns and serving targeted advertisements matched to an individual.The problem of predicting a user's behavior on any web-site has acquired more importance due to the fast growth of technology and the need of personalizing and influencing a user's browsing experience. Hidden Markov models and their variations have been found well suited for addressing this problem.
Its aim is to:

- build a HMM with given advertisement related parameters
- generate sequences of observations using this HMM
- learn the parameters of a HMM using those sequences

It involves the following steps, constructing HMM model with given advertisement related datasets, generating sequences of observations and symbols using the above constructed HMM and learning these parameters of a HMM using Log-likelihood values.The necessary knowledge of user browsing behaviour history is helpful to predict the future sequences which are likely to be visited by the user in future. There is a huge scope for advertisers to design prediction model based on user's browsing page sequences. The main objective of these prediction models is to attain more prediction accuracy.
The below figure illustrates the model for mining advertisement using,

Visible states={Item,Location,Session,Cost,CTR,CPM,CR,CPC}

Hidden states={Low,Medium,High}

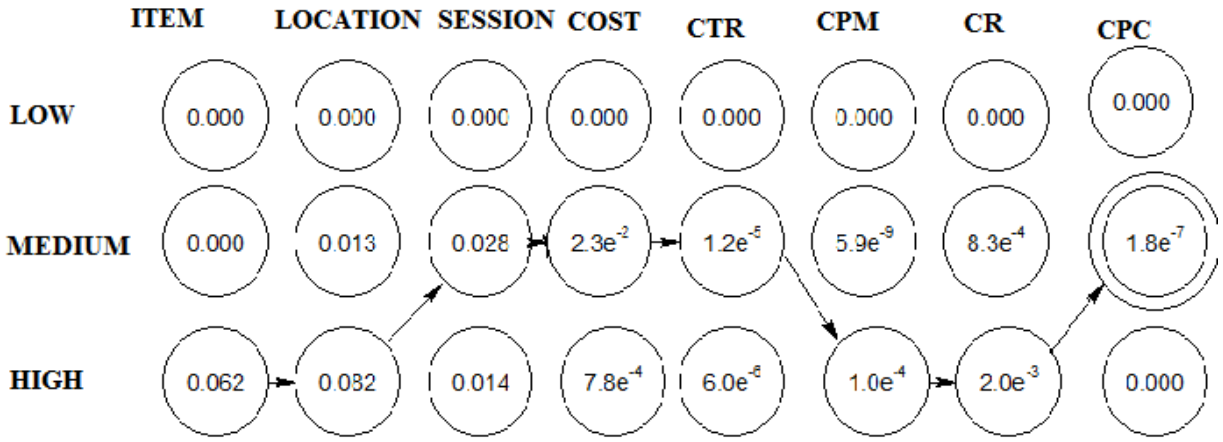

Figure 2:HMM Modelfor MiningAdvertisement

After building HMM model, log-likelihood values have to be estimated using Baum-Welch (forward-backward) algorithm inorder to derive the pattern for retrieving appropriate advertisement. Thereby, deriving an efficient procedure for estimating the model parameters from unlabeled data using observation sequences.Initialization: set (λ=A, B,π) with random initial conditions[10]. The algorithm updates the parameters of λ iteratively until convergence, following the rules below:

$$\overline{\pi}_i = \gamma_i(1)$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)}$$

$$\overline{b}_i(k) = \frac{\sum_{t=1}^{T} \delta_{O_t O_k} \gamma_i(t)}{\sum_{t=1}^{T} \gamma_i(t)}$$

Based upon these modeling parameters, estimating a pattern in relevant to existing user behavior profile repository[11]. The algorithm 1 shows the function of HMM to discover hidden states and finding the target advertisement to the user based on the maximum probability of the model along with user behavior pattern obtained from the MapReduce framework.

Algorithm 1: HMM to predict the target advertisement
Input: Click stream data
Output: max[P(O|M)]
begin:
1: Load click stream data
2: Build HMM Model M(A,B,π)
3: For 1 to k
4: Generate sequences of obv (o_1,o_2, ……, o_k)
5: Find Q: o_1,o_2, ……, o_k //sequence state
6: Calculate max[P(Q, o_1,o_2, ……, o_k)]
7: Compute $\sum_k \xi k(i,j)$ and $\sum_k \gamma k(i)$
8:Iterate:
9: Aij= $\frac{\sum_k \xi k(i,j)}{\sum_k \gamma k(i)}$

10: Bij= $\frac{\sum_k \xi k(i,j)}{\sum_{k,Ok=vm} \gamma k(i)}$

11: until max[P(O|M)]
12: return max [P(O|M)]
end

- Data retrieval

Targeted advertisement can be portrayed to the browser using Bigquery from HDFS after performing MapReduce tasks. The obtained pattern for user behavior is imported to the datastore, then the images for the particular keywords are imported to the blobstore using map reduce. After performing this process, next step is to execute Bigquery mechanism. In order to apply behavioral targeting technique and to ensure the working of mechanism, the keyword is fed to the query processor and the query is instigated. MapReduce takes the request and initiates the job process. Thereby, appropriate advertisement is mined from the database related to the particular keyword and is deployed to the user.

IV. EXPERIMENTAL SETUP

In our experiment we evaluate the efficiency of advertisement for datasets of user behavioral profile and also measured click through rate and fitness values using HMM .

System Setup: We use a machine with 6GB of main memory, 1 TB hard disk

Dataset: Total size of dataset is 1GB.

4.1 Comparison of maximum loglikelihood estimation values between hmm and naivebayes theorem

Optimization of advertisement can be determined by means of Naive bayes theorem and Hidden Markov Model and thereby estimating which mechanism yields better results in terms of the maximum likelihood estimation (MLE) of data. HMM models are trained using the transition and emission

probabilities. Then, the performance of each model is evaluated. Models are compared according to the average log likelihood values for both the probabilities.Our implementations shown in below table shows the testing result for Maximum Log Likehood values for dataset using Naive bayes theorem and Hidden Markov model.

Table 1: Maximum likelihood values comparison

Parameters	Naivebayes theorem (MLE)	Hidden Markov model (MLE)
Item	0.0824	0.1934
Session	0.1159	0.2251
Location	0.1411	0.2495
CTR	0.1613	0.2592
CPC	0.1817	0.3012
CPM	0.1932	0.3253
CR	0.2043	0.3707
CPA	0.2642	0.4193

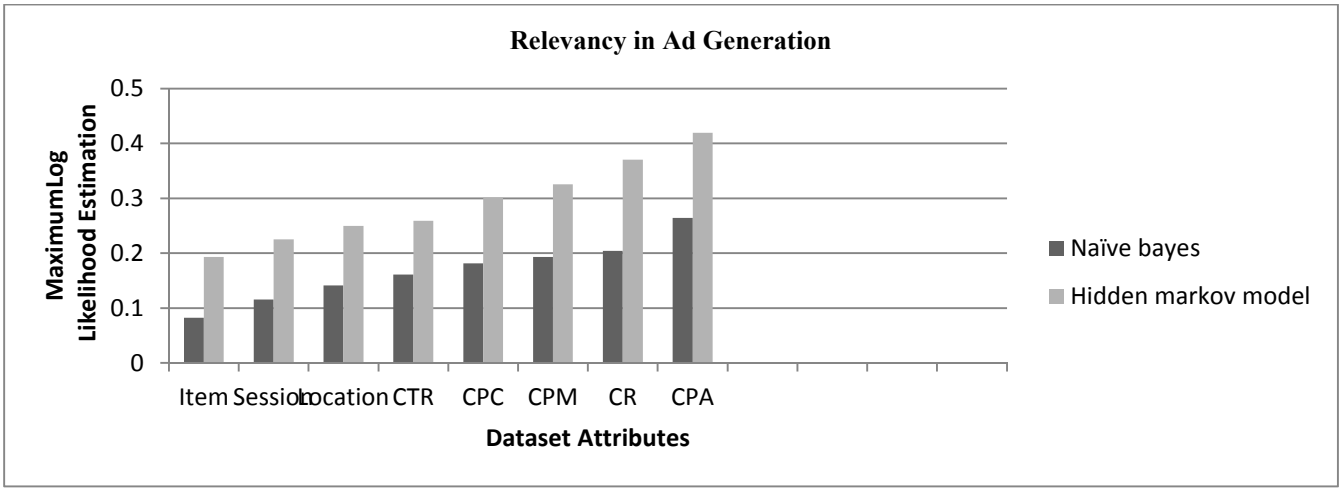

Figure 3 Relevancy in Ad Generation

On observation of the graph, we notice that Hidden markov model using Baum Welch algorithm produce better results in deploying optimized advertisements than using Naive bayes theorem.

4.2 Hidden markov model performance analysis using fitness value

The performance of the Hidden Markov Model was found to depend strongly on the fitness value used[12]. The fitness value usually incorporates a balance factor using likelihood values that is inferred from Hidden Markov Model. The below Table shows the testing result for the Fitness value comparison for Hadoop framework with and without using HMM model.

Table 2:Fitness value Analysis

	Hadoop framework using HMM	Hadoop framework without using HMM
	Fitness value	Fitness value
Iteration 1	0.3247	0.1148
Iteration 2	0.4862	0.1961
Iteration3	0.5261	0.2245
Iteration 4	0.7082	0.3486
Iteartion 5	0.7806	0.4932

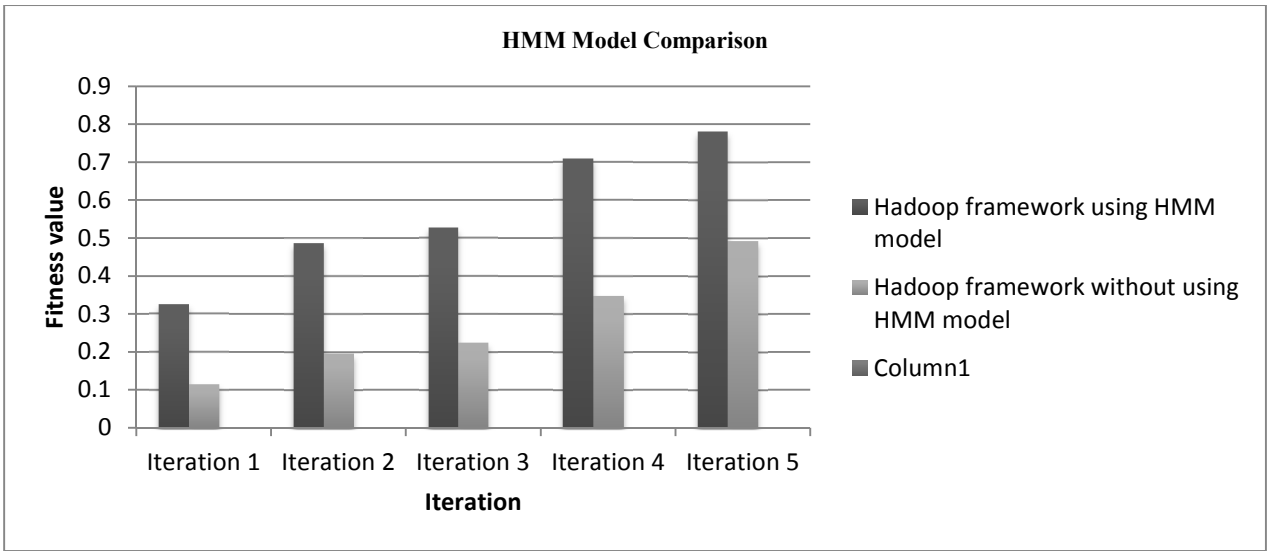

Figure 4: Performance Analysis of Hidden markov model

The graph depicts the fitness value measure for each iteration corresponding to Log likelihood values with and without using HMM model in hadoop framework and their resulting values emphazises that the fitness value enhances while using HMM model in hadoop framework rather, it gradually decreases while not considering HMM model.

4.3 Click through rate comparison for mobile dataset

The performance of online advertising in Hadoop framework is commonly measured by ads Click-Through Rate (CTR)from various users. This work, also describes how much behavior of a user can improve ads CTR through the segmentation of users into a number of small groups for delivering the targeted ads[13].The CTR of advertisement is defined as the number of users who clicked it over the number of users who either clicked it or only displayed it.The below Table shows the testing result for click Through rate comparison for mobile dataset in commercial website and education website

Table 3: Click through Rate Comparison

	COMMERCIAL WEBSITE			EDUCATION WEBSITE		
	CPC	CPM	CTR	CPC	CPM	CTR
Dataset 1	28	66	44	19	60	25
Dataset 2	32	73	46	22	65	32
Dataset 3	41	79	65	28	72	39
Dataset 4	52	84	70	31	78	58
Dataset 5	56	88	79	37	81	66
Dataset 6	62	91	82	43	84	69

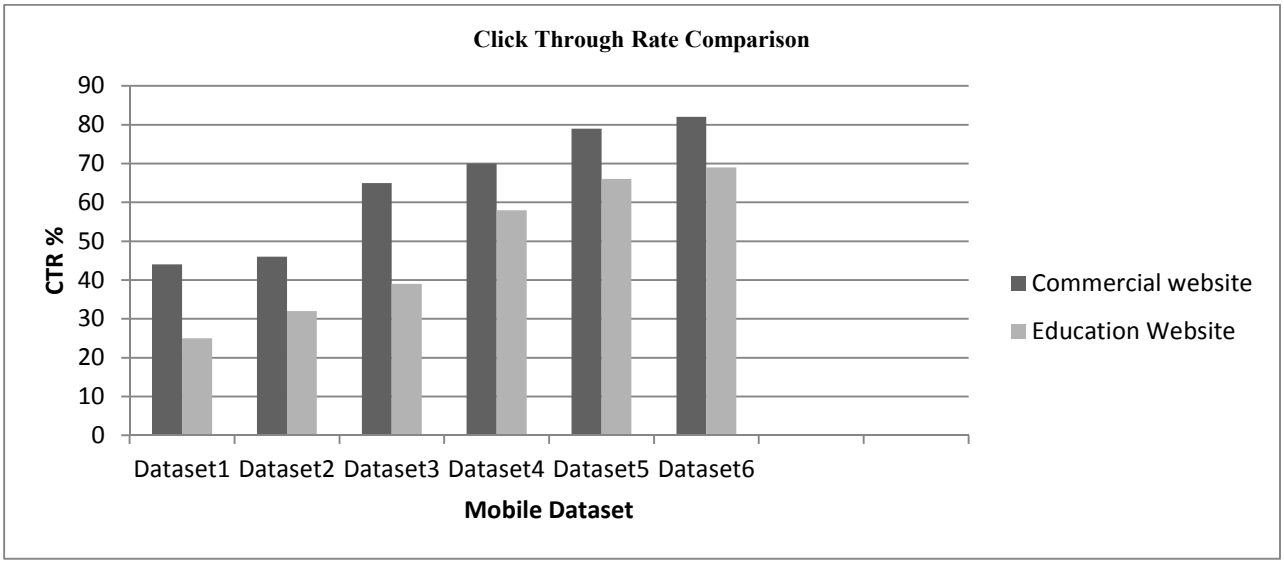

Figure 5 Click through Rate Comparison

The obtained results signifies that CTR % has been drastically increases in commercial website for mobile dataset rather than education website.

V. CONCLUSION

Hadoop framework using hidden markov model resolves the problem of extracting relevant advertisement from a huge data dump in a distributed environment.It has been proved that Hidden Markov Model is an exceptional method in the field of recognizing pattern for user behavior. We have utilized log datasets with user browsing behavior characteristics as training data to construct our model. HMM model is used to guess the browsers intentions in online. In addition, this model helps to analyze the behavior of user to increase

CPC,CPM,CPC rates. Finally, after estimating the individuals browsing pattern using log likelihood values, relevant advertisement can be deployed to the user using Bigquery mechanism. Also, it finds its applications in various fields such as mobile advertising, Geotargeting, E-governance, Netnography, social media optimization etc.,

VI. FUTURE ENHANCEMENT

In future, the effectiveness of the advertisement can be further enhanced by offering customer with more considerate utilities by providing relative sales information with a preferential price to retain the customer and increase better sales revenue. This model with more enhancements can be used to predict accurate CPC,CTR,CPM in real-time using a large set of training data distributed over the time frame. This will empower users to predict click rate values based upon social data and also the time during the day when users are most active. This mechanism can be further extended by adding some additional features and can be applied to audio and video data as well as semi structured data.

REFERENCES

[1]K.Sathiyamurthy, P.Dhivya, "Hadoop MapReduce for Advertisement mining using Bigdata Analytics", inProceedingsof *International Conference on Bigdata and Analytics on Business*, December 2014,New Delhi, India

[2]K.Sathiyamurthy, P.Dhivya, DJ.Panimalar,"Advertisement mining using Hidden Markov Model", in *International Journal of Advanced Research in Computer and Communication Engineering*Vol. 4, Issue 3, March 2015

[3]Paul M. Baggenstoss,"A Modified Baum–Welch Algorithm for HiddenMarkov Models with Multiple Observation Spaces"in *Proceedings of IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, May 2001.

[4]Online advertising behavioral targeting available athttp://www.networkadvertising.org/pdfs/NAI_Beales_Release.pdf

[5]JyotiNandimath, AnkurPatil, Ekata Banerjee, Saumitra Vaidya(Aug 2013)," Big Data Analysis Using Apache Hadoop" *IEEE IRI 2013, San Francisco, California, USA.*

[6]BadrishChandramouli, Jonathan Goldstein, SongyunDuan(2012),"Temporal analytics on big data for web advertising" *IEEE 28th International Conference on Data Engineering*.

[7] Amrit Pal, Pinki Agrawal, Kunal Jain, Sanjay Agrawal(2014),"A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data Using Hadoop*" IEEE Fourth International Conference on Communication Systems and Network Technologies*

[8] Jianqing Chen, Jan Stallaert," An Economic Analysis of Online Advertising Using behavioral targeting", in *American Society for Engineering Education*, vol. 6(4), pp .19-22,1996.

[9] Chun-Jung Lin , Fan Wu , I-Han Chiu ,"Using Hidden Markov Model to Predict the Surfing User's Intention of Cyber Purchase on the Web " in *Proceedings of the International Conference on Communication, Information & Computing Technology (ICCICT)*, pp.19-20,October,2012

[10] M. Awad and I. Khalil, Prediction of Users Web-Browsing Behavior: Application of Markov Model, *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 42, no. 4, pp.11311142, Aug. 2012.

[11]Jianqing Chen, Jan Stallaert," An Economic Analysis of Online Advertising Using behavioral targeting",*in American Society for Engineering Education*, vol. 6(4), pp .19-22,1996.

[12] Matthew Richardson,EwaDominowska, Robert Ragno,"Predicting Clicks: Estimating the Click Through Rate for New Ads",in *Proceedings of the International World Wide Web Conference Committee (IW3C2),*May,2007

[13] Paul M. Baggenstoss,"A Modified Baum–Welch Algorithm for HiddenMarkov Models with Multiple Observation Spaces"in *Proceedings of IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, May 2001.

[15] Sagiroglu, Seref, and Duygu sinanc, "Big data: A review." in *Proceedings of International Conference on Collaboration Technologies and Systems (CTS)* 2013. Vol. 3, pp. 42-47,2013.

[16] Ahmed, Amr, et al. "Scalable distributed inference of dynamic user interests for behavioral targeting." *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011.*series KDD'11 ,pp 114-122,2011.

[17] Dwyer, Catherine Ann. "Behavioral targeting: A case study of consumer tracking on levis. com." (2009). *Proceedings of the Fifteenth Americas Conference on Information Systems, San Francisco, California,2009.*