Empirical Study of Soft Clustering Technique for Determining Click Through Rate in Online Advertising

Akshi Kumar[1], Anand Nayyar[2], Shubhangi Upasani[3], Arushi Arora[4*]

[1]Department of Computer Science & Engineering, Delhi Technological University (DTU), Delhi, India; akshi.kumar@gmail.com

[2]Graduate School, Duy Tan University, Đà Nẵng, Viet Nam; anandnayyar@duytan.edu.vn

[3]Department of Electronics and Communication Engineering, DTU, Delhi, India; shubhangi.upasani@gmail.com

[4*] Department of Electrical Engineering, DTU, Delhi, India; aroraarushi1997@gmail.com

Abstract: Online advertising is an industry with the potential of maximum revenue extraction. Displaying the ad which is more likely to be clicked plays a crucial role in generating maximum revenue. A high Click Through Rate (CTR) is an indication that the user finds the ad useful and relevant. For suitable placement of ads online and rich user experience, determining CTR has become indispensable. Accurate estimation of CTR helps in placement of advertisements in relevant locations which would result in more profits and return of investment for the advertisers and publishers. This paper presents the application of a soft clustering method namely fuzzy c-means (FCM) clustering for determining if a particular ad would be clicked by the user or not. This is done by classifying the ads in the dataset into broad clusters depending on whether they were actually clicked or not. This way the kind of advertisements that the user is interested in can be found out and subsequently more advertisements of the same kind can be recommended to him, thereby increasing the CTR of the displayed ads. Experimental results show that FCM outperforms k-means clustering (KMC) in determining CTR.

Keywords: advertising; Click Through Rate; clustering

1. Introduction

Online advertising has become an important source of revenue for a wide range of businesses. With the tremendous growth in online advertising each year, it has also taken over a major area of research. Most of the revenues generated by widely used search engines as well as prevalent websites come from advertisements. It is therefore important to display relevant advertisements (ads) to the users and avoid the advertisements that are often disliked by them.

Online advertising is more economical than traditional ways of advertising like mass markets and niche media. Internet ads have a wider audience and can be viewed for days and nights altogether, in contrast to ads on television and radios that last for shorter durations and are displayed with limited frequency. Market segmentation is much more effective over the Internet than on any other medium. Thorough study of markets, customer preferences and habits and segmenting consumers into cohesive groups can be done efficiently through online advertising.

Online advertising also offers small businesses numerous benefits like robust targeting, consumer insights and more effective return on investment.

Advertisements fall into two broad categories- sponsored search advertisements and contextual advertisements. Sponsored search ads are displayed on the same web pages that show results of search queries entered by users. The core purpose behind sponsored advertising is to enhance the advertiser's brand image as the ads displayed have the same form and qualities as the advertiser's original content. Contextual advertising, on the other hand uses automated systems that display ads relevant to the user's identity and website's content. Google AdSense is one of the many well- known examples of contextual advertising. Google robots display only those ads that the users find relevant and useful. When a user visits a website, features like ad size, ad placement etc. are extracted from the search query and sent to a server. Relevant ads are selected based on user's past history, CTR and other data. Increasing the number of ads is not a good idea as it will shoot up the earnings only for a short time before the user switches to other search engines due to poor user experience. To maximize revenue, precise placements of ads are therefore required.

CTR refers to the number of times the advertiser's ad has been clicked (clicks) divided by number of times the ad appears on the screen (impression). Relevant placement of ad is a precondition for increasing the CTR for it. Ad performance can also be measured using CTR. CTR determination has several issues associated with it. The advertisers need to pay every time an ad is clicked. It is on the basis of CTR that the search engine decides what ads are to be displayed and in what order of appearance. This is ensured by combining together the likelihood of an ad being clicked and the cost of the ad per click to create a display format that will yield maximum return. Many algorithms based on the supervised methodology have been designed to predict the CTR of advertisements like support vector machines, Decision Trees, Naïve Bayes etc.

Motivated by this, the goal is to analyse and assess the application of un-supervised approach namely FCM for determining the CTR of an advertisement. This clustering algorithm divides the total ads in the dataset into broad clusters based on whether they have been clicked or not by the user. This classification of ads helps in assessing the kind of advertisements that the user is really interested in and hence predict more of these kinds of ads. This would in turn result in a rise in the CTR of the displayed ads because of a greater number of clicks. The results obtained from FCM have been contrasted with KMC which has been used as a baseline model. The two techniques are assessed based on metrics like Precision, Recall and Accuracy.

The following content is compiled as follows: Section 2 reviews work done by various researchers along with the brief idea of the algorithm used by them and their results. Section 3 elaborates the data set taken, methods employed for preparing the data before implementing the algorithm and the algorithms used. The scrutiny of the experimental outcomes is done in the Section 4. Section 5 culminates the results along with the suggestions for future research prospects.

2. Related Work

After a thorough assessment of various studies in the past, it was found that a lot of algorithms have been proposed for the prediction of relevant ad to be displayed to the user. The author Avila Clemenshia P. et al.[1] in 2016, proposed a CTR prediction model using Poisson's regression, linear regression and support vector regression algorithms and displayed the ads accordingly. The dataset was provided by a digital marketing agency. Their results stated that Support Vector Regression performed best among the three. Evaluation of the results was done on the basis of Root mean squared error (RMSE) and correlation coefficient.

Authors Thore Graepel et al. [2] in 2010 devised a new Bayesian CTR algorithm. The ad Predictor presented showed better outcomes when compared to the baseline Naïve Bayes in terms of relative information gain (RIG) and areas under the curve (AUC). Dustin Hillard et al. [3] in 2010, implemented a model for estimating ad relevance. They refined it by including indirect feedback after consolidating basic features of text overlap. In case of presence of adequate observations, click history was used. In case of no or few observations, a model was learnt that could also be used for unpredicted ads. The precision, recall and f score values were noted and improvement was observed with the new model.

A Logistic Regression approach was suggested by author Gouthami Kondakindi et al. [4] in 2014, to predict whether an ad will be clicked or not. The dataset used for this purpose was from Avazu provided as a part of Kaggle competition. They started off with simple Naive Bayes followed by Vowpal Wabbit and finally got the best scores with logistic regression together with proper data pre-processing. Lihui Shi et al. [5] in 2016, designed a framework for prediction of CTR and average cost per click (CPC) of a keyword using some machine learning algorithms. The performance data for the advertiser's keywords was gathered from Google Adwords. The author had applied different machine learning algorithms such as regression, random forest and gradient boosting to evaluate the prediction performance on both CTR and CPC. Results concluded that random forest transpires to be the best for both the metrics while gradient descent results are least reliable.

Chieh-Jen Wang et al. [6] in 2011, implemented a model for learning of user's click behaviors from advertisement search and click logs. Decision tree (DT), CRF, SVM and backpropagation neural networks (BPN) are the algorithms which were employed to carry out the imposition. The experimentation finally led to proving that CRF model outperformed the two baselines and SVM remarkably. Deepayan Chakrabarti et al. [7] in 2008, developed a model for Contextual Advertising in which the revenue accrued by the site publisher and the advertising network depend upon the suitability of the ads displayed. This was followed by mapping the model to standard cosine similarity matching.

Haibin Cheng et al. [8] in 2010, developed the model for customization of click models in sponsored search. The results demonstrated that the accuracy of CTR could significantly be improved by personalized models in sponsored advertising. Bora Edizel∗ et al. [9] in 2017, proposed the use of that deep convolutional neural networks for CTR prediction of an advertisement. One approach involved query-ad depiction being learnt at character while the second entail word level model by pre-trained word vectors. The conclusion signified better outcome than the standard machine learning algorithms trained with well-defined features.

The analysis of the background work clearly connotes that most of the algorithm used in past are supervised classification algorithms that involves two classes namely clicked and not clicked- demonstrating whether advertisement was clicked or not. The study of unsupervised clustering is still not much discovered in this domain to the extent of our understanding. This research paper is an endeavor to compare FCM and KMC algorithm using the dataset.

3. Data Characteristics

The following figure 1 demonstrates the system architecture of the research undertaken. Each block of the diagram has been explained in the following sections.

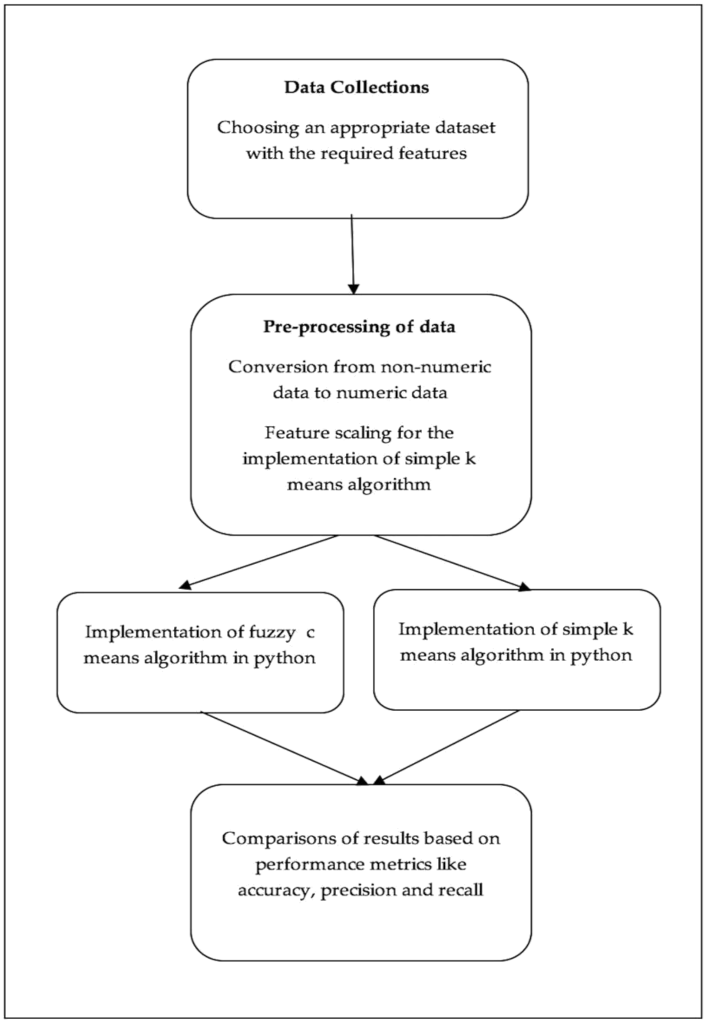

Data Collections
Choosing an appropriate dataset with the required features
Pre-processing of data
Conversion from non-numeric data to numeric data
Feature scaling for the implementation of simple k means algorithm
Implementation of fuzzy c means algorithm in python
Implementation of simple k means algorithm in python
Comparisons of results based on performance metrics like accuracy, precision and recall

Figure 1. System Architecture

3.1. Data Collection

The dataset is acquired from Avazu (Kaggle) for purpose of writing this paper. It contains 11 days' worth of data in order to build and test prediction models using various machine learning algorithms. As the given data is approximately 6GB, the data taken was 10hrs of data for training and 2hrs of data for testing. The data is ordered chronologically and the clicks and non-clicks are sampled according to different strategies.

Following features are included in the dataset: -

- id (unique id to identify advertisement)
- click (0 for not clicked and 1 for clicked)
- hour (in format of YYMMDDHH ; it refers to the date and time the data is recorded)
- banner_pos (0 for top and 1 for bottom)
- site_id (refers to the website id)
- site_domain (website domain id)
- site_category (category to which website belongs)
- app_id (app id)
- app_domain (app domain)
- app_category (class to which app belongs)
- device_id (id of device from which ad was clicked
- device_ip (ip address of the device)
- device_model (model number of device)
- device_type (mobile/laptop/desktop)
- device_conn_type(internet connection type-wifi/mobile data)
- C1 and C14-C21 (anonymized categorical variables).

3.2. Pre-processing of data

The class attribute is removed owing to the fact that unsupervised clustering is the technique applied which will learn from the data and classify it into two clusters. The dataset contains features in both integer and string format. In order to carry out the implementation, conversion of strings into integer is done. Pandas package in Python is used to load the csv file i.e. the raw dataset into memory, identify the columns with string values and convert them into integer values using python standard hash function.

Feature scaling is done with the help of scikit-learn library in python in the case of KMC. It is done in order to ensure fast convergence to the optimal solution and normalize the range of variables.

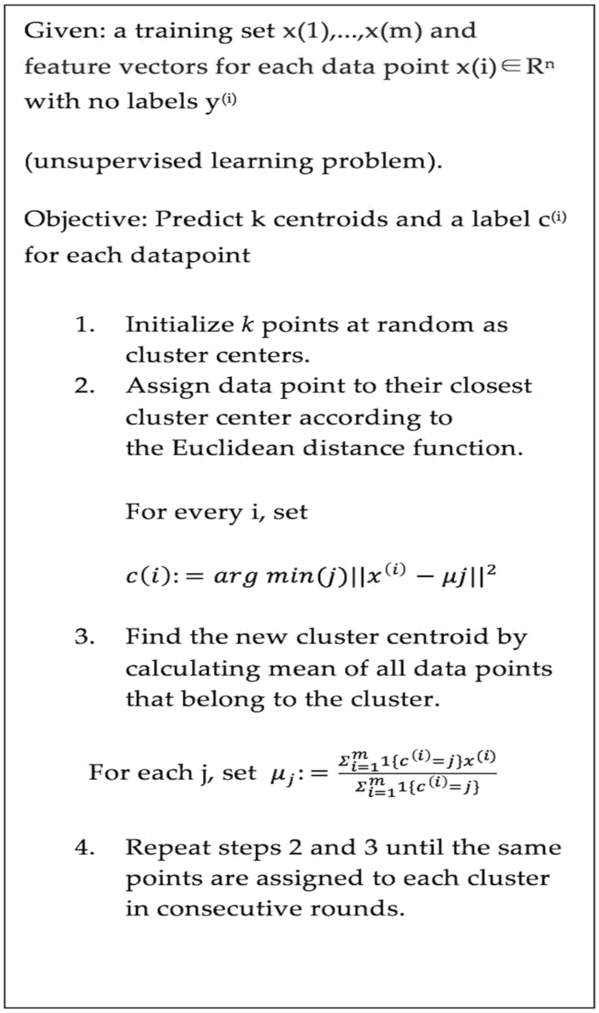
3.3. Implementation of machine learning algorithms

The following section explains the clustering algorithm used.

3.3.1. K-means Clustering:

It is one of the most straightforward learning algorithms available for unsupervised learning [10]. The procedure basically follows two main steps to classify the data set into k clusters. The numbers of clusters are fixed apriori. The algorithm starts by defining k cluster centers. Each value in the dataset is then associated to its nearest cluster. This is followed by a recalculation of the k cluster centers. A new binding is done between discrete units of information in the dataset and the new cluster centers and these steps are repeated till all units have been assigned to their respective clusters.

The pseudo code is given as follows:



```
Given: a training set x(1),...,x(m) and
feature vectors for each data point x(i)∈Rⁿ
with no labels $y^{(i)}$

(unsupervised learning problem).

Objective: Predict k centroids and a label $c^{(i)}$
for each datapoint

    1.  Initialize $k$ points at random as
        cluster centers.
    2.  Assign data point to their closest
        cluster center according to
        the Euclidean distance function.

        For every i, set

        $c(i) := arg\ min(j)||x^{(i)} - \mu j||^2$

    3.  Find the new cluster centroid by
        calculating mean of all data points
        that belong to the cluster.

      For each j, set  $\mu_j := \frac{\Sigma_{i=1}^{m} 1\{c^{(i)}=j\}x^{(i)}}{\Sigma_{i=1}^{m} 1\{c^{(i)}=j\}}$

    4.  Repeat steps 2 and 3 until the same
        points are assigned to each cluster
        in consecutive rounds.
```

Figure 2. Pseudo Code for k means clustering

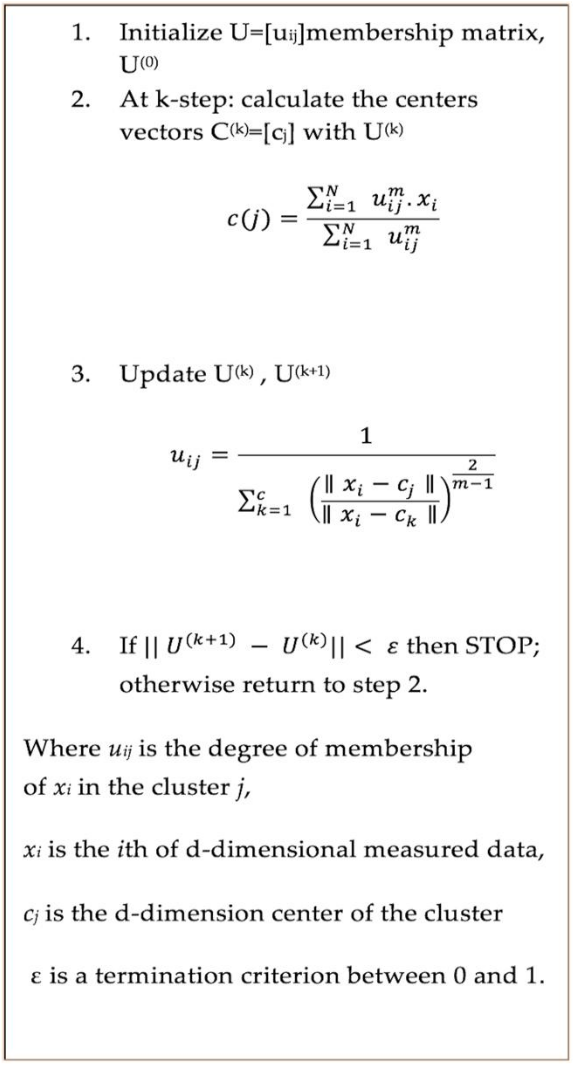
X={x1, x2……….xn} be the collection of observations and u={u1, u2…….uc} are the cluster centroids. We arbitrarily select 'c' cluster centers (1). Then, the distance between each observation in dataset and center is calculated. Following this, each observation is assigned to that cluster for which the distance between the observation and the center is least from all the available centers (2). This assignment step is followed by a recalculation of the cluster centers(3) by executing step 1 again . If all observations belong to the clusters calculated in the previous iteration, then the algorithm is terminated.

3.3.2. Fuzzy C-Means Clustering:

FCM is also referred to as soft clustering [11]. Each observation can be a part of more than one cluster at the same time. The distance between each cluster center and value in the dataset is evaluated and based on this distance, degree of membership is assigned to each observation. A higher degree of membership towards a cluster means that the observation is closer to that cluster center than compared to the rest of the clusters. The summation of degree of membership of each point in dataset equals one. Data points belong to distinct clusters in hard or non-fuzzy clustering.

The degree of membership assigned to each observation plays a pivotal role here. This denotes the extent to which an observation belongs to each cluster. As we move from center to the boundary of the cluster, the degree of belongingness of the observation decrease.

The pseudo-code for FCM is as follows:



```
1.  Initialize U=[$u_{ij}$]membership matrix, $U^{(0)}$
2.  At k-step: calculate the centers vectors $C^{(k)}$=[$c_j$] with $U^{(k)}$

        $c(j) = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$

3.  Update $U^{(k)}$ , $U^{(k+1)}$

        $u_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{\| x_i - c_j \|}{\| x_i - c_k \|} \right)^{\frac{2}{m-1}}}$

4.  If $|| U^{(k+1)} - U^{(k)} || < \varepsilon$ then STOP;
    otherwise return to step 2.

Where $u_{ij}$ is the degree of membership
of $x_i$ in the cluster $j$,

$x_i$ is the $i$th of d-dimensional measured data,

$c_j$ is the d-dimension center of the cluster

ε is a termination criterion between 0 and 1.
```

Figure 3. Pseudo Code for fuzzy c means clustering

Let X = {x1, x2, x3 ..., xn} be the collection of discrete units in dataset and V = {c1, c2, c3 ..., ck} be the centers. Select ‘k’ cluster centers and initialize the membership matrix step (1). Compute the centers 'cj' step(2). Calculate the membership 'µij' in the membership matrix using step(3). Repeat step 2) and 3) until the smallest value of J is obtained.

4. Results & Analysis

The results were analyzed using the Accuracy, Precision and Recall [12] as an efficacy criterion. Accuracy is the measure of the closeness of the predicted observations to the actual value. It is the ratio of rightly predicted inspections to the total inspections made. Higher value of accuracy indicates more true positives and true negatives. Precision refers to the correctness of a model. It is simply the ratio of all the precisely predicted positives to the total number of positives predicted. More the number of true positives implies more precision. Recall is a measure of responsiveness or sensitivity of a machine model. It is the ratio of the correctly predicted positives to the count of values that belong to the class 'yes'. Recall and precision are inversely related. The following Table 1 depicts the performance analysis of FCM and KMC.

	k- means		**fuzzy c-means**	
	Cluster 0	**Cluster 1**	**Cluster 0**	**Cluster 1**
Accuracy	43.85438544	43.85438544	76.61767177	76.61767177
Precision	15.62280084	81.0428737	17.12049388	83.51293103
Recall	52.05182649	42.16809357	91.03165299	9.345230918

Table 1. Comparison of results for k means and fuzzy c means clustering

As the table 1 shows, the accuracy, recall and precision values obtained for the two clusters i.e. clicks/non-clicks predictions for the users is higher for FCM than for KMC. This means that the rightly predicted true positives and true negatives (accuracy), the degree of exactness (precision) and degree of completeness (recall) are relatively higher for FCM than for KMC.

Both the algorithms divide the data into two clusters, Cluster 0 and Cluster 1, which in turn represent a division of the data set into two categories. These categories are based on click/non-click classification for the ads i.e. one cluster has all the ads clicked by the user. The other cluster has all ads which were not clicked by him. This gives a clear idea of the type of ads the user is inquisitive about and willing to click on.

After the classification of the ads in the dataset into categories- 'Click' and 'Non-Click', we get a fair estimate of the kind of advertisements the user really wants to see. We can thereby predict more of such ads, hence increasing the CTR for the same. The number of 'clicks' divided by the total number of times the ad is displayed i.e. 'impressions' gives the CTR.

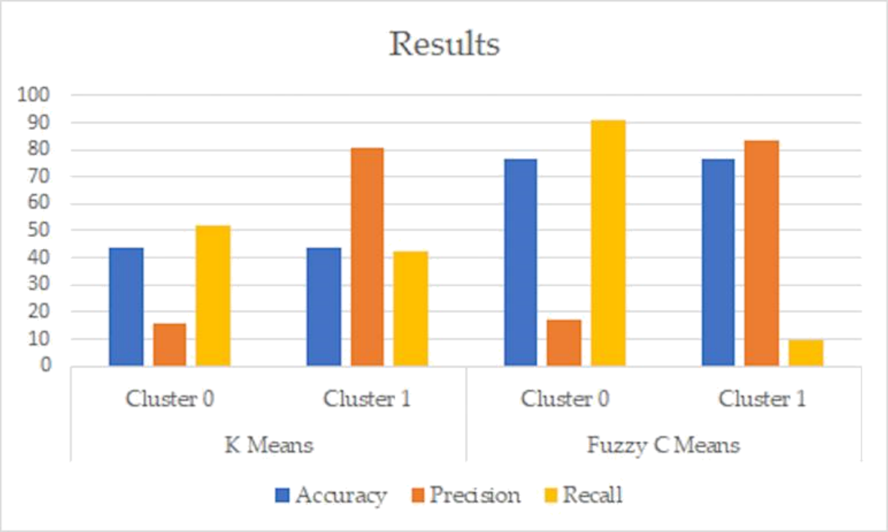


Figure 4. Results

6. Conclusion

This paper compares the CTR as assessed by KMC and FCM algorithms. The outcome shows that FCM achieves better values of performance metrics than KMC. The results are better because it considers the fact that each data point can lie in more than one cluster and involves complex calculation of membership matrix. Conventional KMC, on the other hand, relies on hard clustering and definitively assigns each data point to the clusters.

The accuracy of the model can further be improved by assessing the shape of the clusters, including other more refined attribute selection and extraction methods that could assist in better modelling of the present system. Attribute selection is concerned with selecting a subgroup of valid features for model selection whereas attribute extraction means deriving attributes from the already existing ones for subsequent learning. Superior results could be obtained by enhancing fuzziness coefficient. The degree of overlap between clusters is determined by the fuzziness coefficient. Higher value of m means larger overlapping between clusters. There exists a vast scope of application of other soft computing methodologies also, like swarm optimization etc. that can be applied for determining CTR for other datasets as well.

References

1. Avila Clemenshia P.; Vijaya M.S. Click Through Rate Prediction for Display Advertisement, International Journal of Computer Applications (0975-8887., IJCA, February 2016, Volume136-No.1.)
2. Thore Graepel; Joaquin Candela; Thomas Borchert; Ralf Herbrich. Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. IJCA 2010.
3. Dustin Hillard; Stefan Schroedl; ErenManavoglu; Hema Raghavan ; Chris Leggetter. Improving Ad Relevance in Sponsored Search. ACM 2010.F
4. GouthamiKondakindi; Satakshi Rana; Aswin Rajkumar; Sai Kaushik Ponnekanti; Vinit Parakh. A logistic Regression Approach to Ad Click Prediction, 2014.

5. Lihui Shi; Bo Li. Predict the Click Through Rate and Average Cost Per Click for Keywords Using Machine Learning Methodologies.IEOM, 2016.
6. Chieh-Jen Wang and Hsin-His Chen. Learning User Behaviors for Advertisements Click Predictions. ACM,2011.
7. Deepayan Chakrabarti; Deepak Agarwal; VanjaJosifovski. Contextual Advertising by Combining Relevance with Click Feedback.WWW, 2008.
8. HaibinCheng ; Erick Cantú-Paz. Personalized Click Prediction in Sponsored Search. ACM, 2010.
9. Bora Edizel; Amin Mantrach ; Xiao Bai. Deep Character-Level Click-Through Rate Prediction for Sponsored Search. Stat.ml,2017.
10. Jyoti Yadav and Monika Sharma, “A Review of K Means Algorithm”, Published in International Journal of Engineering Trends and Technology, Volume 4, Issue 7, July 2013.
11. Chen Yanyun, QiuJianlin, Gu Xiang, Chen Jianping, Ji Dan and Chen Li, “Advances in Research of Fuzzy C Means Algorithm”, Published in International Conference on Network Computing and Information Security, May 2011.
12. MPS. Bhatia, A. Kumar, Information Retrieval & Machine Learning: Supporting Technologies for Web Mining Research & Practice, Webology, Vol. 5, No. 2.