User Clustering Algorithms In Online Advertising

Honey Vachharajani, Rajeev Gupta, Nikhlesh Pathik

Abstract: ***The Digital Advertising has emerged as one of the main source of revenue for major part of Internet economy. The audience communicates with the digital world by using search engines, social networking, online market- ing and banking sites and many more. To generate more income through advertisement the ad-publishers and the ad-networks need to be watchful about users interest when targeting them for their brand or product pro- motion through these channels. The placement of advertisement depends on the users interest as is it involves the higher probability of a click on the ad, which offers benefit to all the entities, involved. This paper surveys different clustering approaches proposed by various authors for user clustering. At the end of the paper, the various methods are compared based on phases, techniques and usages.***

Index Terms: ***Audience Clustering, Unsupervised Learning, User Profiling, Ad Targeting Clustering Algorithms.***

I. INTRODUCTION

In Today's technology dependent world user profiles are the virtual depiction of each user. A user is identified by various attributes possessed by him that includes personal interest, hobbies, age, sex, location, income and many more. From these attributes a user profiling process is done which is used by user service personalization sector. The main objective of such ser- vice sectors are to survey the users interest and prefer- ences to match with the service delivery. The progress of these services counts on how accurate these service providing sectors understand the user and how well this knowledge is implemented on by the provider on the service. If a user engaged in multiple activities related to a particular content of a certain topic, the user is likely to be interested in that subject. Hence, user tracking companies spends in user logging activities, examines the contents consumed by users. Based on this information they build models to construct a user representation which reflect online interests. The Fig 1 shows a way to track the users by using Ghostery plug-in we can see that there are 11 users on the homepage of Flipkart and 13 users on the particular section of the page in Fig 2.

Revised Manuscript Received on July 05, 2019.

Honey Vachharajani, Computer Engineering Dept, R.G.P.V/S.I.S.T.E.C Gandhinagar, Bhopal, India
Rajeev Gupta, Computer Engineering Dept, S.I.S.T.E.C Gandhinagar. Bhopal, India
Nikhlesh Pathik, Computer Engineering Dept, S.I.S.T.E.C Gandhinagar. Bhopal, India

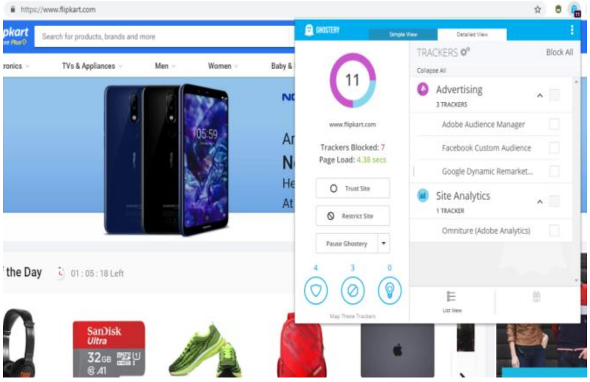

Fig1: Image of Google tracker showing the active users on the home page

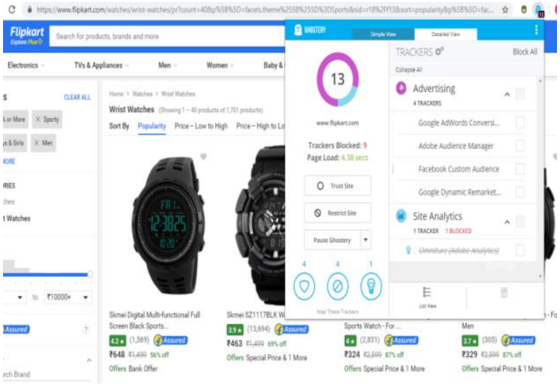

Fig 2: Image of Google tracker showing the active users on the section of a particular page

The common elements of such profiles include user demographics (age, gender), user locations, and a list of interests revealed by historical content consumption. The general approach used by online advertisers, content providers and ad networks is to imagine of "profiles" of target users based on their interests. User tracking companies can collect the user activity data and try to understand user behaviors. For example, if a user engage in multiple activities related to the content of a certain topic, the user is likely to be interested in that subject. When a user visit a particular website, the content provider shares users (hashed) cookie under its domain and the users current activity with its partner companies. Today cookie syncing is ubiquitous. Since the past decennium advertising has emerged as the main resource of income generation for many websites. The capability to track and model each user has inspired the creation of numerous applications like targeted advertising. In the initial stages of online advertising, advertisers arranged with content providers to display their ads on individual web pages to a certain number of visitors over a certain course of time. Nowadays, advertisers have much more flexibility in

Retrieval Number: B10060782S419/2019©BEIESP
DOI: 10.35940/ijrte.B1006.0782S419

targeting audiences.

Advertisers and ad platforms uses the concept of user profiling, which includes demographic and geographical information as well as fine granular interests. When advertisers set up ad campaigns, they can specify to target users interested in particular topics. Therefore, on a particular web page, instead of seeing the same ads, visitors will see different ads related to their past online activity. So users are likely to encounter ads based on their activity history across the web. One key application of user online profiling is ad campaign targeting. A typical user targeting method is based on the pre-defined user interest taxonomy created by advertising platforms. Advertisers must first analyze and understand the behaviors of their existing customers. They must then utilize the user representation taxonomy from a particular advertising platform to approximate the desired group of customers to target. In this paper different types of methods suggested and approached by different authors has been discussed and a comparison chart is based on the same. Here advertisers provide a list of past converters as "seed users" and system automatically determines users exhibiting behavior patterns similar to the seed and makes them available for advertisers to target. This paper surveys different clustering approaches proposed by various authors for user clustering. At the end of the paper, the various methods are compared based on phases, techniques and usages. Section II explains user profiling, the parties involved in it, the concept of online advertising. Section III surveys different clustering algorithms based on user clustering. Section IV compares the algorithms discussed in section III with respect to various aspects like dependence, phases, technique used and usages.

II. BACKGROUND

A. User Profiling

User profiles are the effect of the user profiling method and symbolize the users and their interests. The process creates an initial user profile whenever a new user comes and then continuously updates the profile as per the user's changing preferences, interests and needs.[7][10]. Any user provides a collection of information that will show the traits of a user based on personnel choices and behavior.

This collection of personal information can either be static data or dynamic. The matter and quantity of the facts in a user profile varies based on the application area.[7] The accuracy of the user profile depends on the user fact gathering process and how well the gathered information is processed. [9].

In the explicit process, the user provides the facts regarding the user's interest and preferences explicitly to the system. These methods provides the fixed traits of the user. Implicit information is collected dynamically by surveying the user's communication with the system automatically. It is called implicit or dynamic user profile. In case of a hybrid user profile, it uses both the ways. It starts with the collection of static facts followed by implicit methods to update the user profile and vice versa [6].

B. User Targeting

The main parties in the online advertising market are described below:

- Publishers: They are online content providers, including traditional ones like the New York Times, Yahoo Finance, etc., and new types of content providers such as social networks (e.g., Facebook) and mobile apps (e.g., game apps, fitness apps). The goal of publishers in the market is to monetize their user traffic (ad impression inventory) by displaying ads to visitors.
- Advertisers: the companies or individuals that seek to promote their products. These include traditional advertisers who want to sell products or brand their companies and the new group of advertisers in the mobile domain who want to get more people to install and engage with their apps. The goal of advertisers is to boost their brands or sell as many products as possible within their advertising budget.
- Ad-networks : the companies that connect publishers and advertisers, allowing advertisers to deliver ads to publisher web pages. The goals of ad-networks are two-fold, to help publishers maximize their inventory monetization and to help advertisers deliver ads to online users which maximize their return on advertising investment.
- Online users: the online content consumers, such as web page viewers or mobile app users. In the market, they are the creators of advertising inventory and are identified by anonymous IDs such as browser cookies. In the context of online advertising, the main parties conduct practices to make user targeting possible and drive advances in user targeting performance.

C. Online Advertising

Online Advertising is a type of marketing or advertising which uses internet to promote the products to the customers. Different types of models are required to represent users and their behavior. It provides advertisers a unique platform to compute real profit on investment for advertising campaigns. Every advisement or marketing agencies have different categories for campaigning or promoting themselves. The target audiences plays an important role in it as the campaigns are divided in sub campaigns depending on the audience categories.

Traditionally advertiser use to face the problem of risking his investments as there was no awareness of user choices or preferences that leaded to lots of money wastage. In today's digital world the advertiser are aware of user preferences s that it becomes easy for them to decide the amount to be invested on sub campaigns which reduces the risk of monetary loss [5].

D. Roles Of Parties In Online Advertising

- Publishers embed user tracking companies user tracking code in their web pages, such that all their users detailed content consumption activities are logged by user tracking companies.
- Ad-networks typically do user tracking by themselves but may also purchase user data from other user tracking companies to get a complete view of online user behavior. Adnetworks build models to create a

profile of online users, including demographics, geo location, interests, etc. These user representations provide advertisers with the flexibility to target their desired groups of audiences.

- Advertisers are the consumers of user online profile modeling results. Given limited budgets, advertisers cannot explore the whole user space to gain potential customers. Instead they use their business insights to design ad campaigns targeting particular groups of users, which are expected to bring a higher return on advertising investment.
- Users often show some consistency in their online content consumption patterns. These patterns are captured by user tracking companies to compose online user profiles, which are updated over time. User profiles not only benefit advertisers but also expose users to more relevant ads and improve the online experience. [3]

The Fig 3 shows how user profiling is used in the marketing world to place an advertisement as per the users interest.

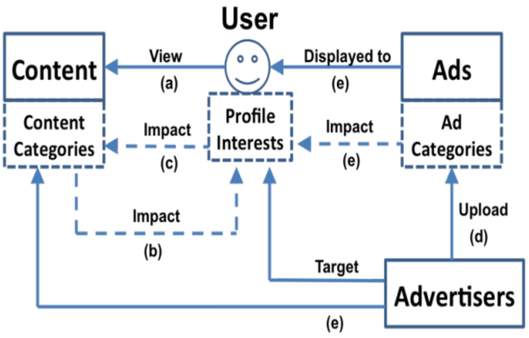

Fig 3: Role of Advertiser, User, Publisher and Ad networks

- Users view the content of interest to them;
- Contents viewed impact user profiles;
- Profiles of users of a content page potentially impact categories of the content;
- Advertisers upload ads which also have categories;
- Advertisers target content and user categories, which affects which ads users are shown, and so on.

III. Classification Of Audience Clustering Methods

For user clustering various methods are proposed and used. In this section, the methods surveyed by us are described with the suggested approaches.

A. Location -Based Targeting In Online Advertising

In this approach, users are clustered based on the geographic area or the location. Advertisers may be given a polygonal area of the map and then the advertisement is placed to the users of that area[8]. The users here are clustered based on search related to that area or staying in that area.. Since advertisers often offer local products or services, users will be mostly interested in ads that refer to products and services in their geographic location. There is less interest in advertising a grocery shop in Mumbai if the user lives in Delhi.

The current location of the user is traced by his IP address which is then matched with the polygonal map of the area provided by the advertiser. This model works by assigning geographic span to the web pages and then classifies the target pages based on the geographically interested span or not. It then forms a cluster based on similar searches and interests of the user. The advertisers then targets the clusters for a particular type of ad based on his interest. The drawback to this approach for a particular user's current location when considering the geographic span of an ad, the geographic context of the Web page he is visiting can also have an important role to play [11]. For instance, a user currently located in Mumbai, but visiting a page for a vacation resort in Singapore is likely to be interested in restaurants in Singapore, rather than Mumbai

B. Similarity–Based Look Alike Audience

The comparison is done between the pairs of seed users and all the existing users based on features. It is done by measuring distance between them. The similarity is done between 2 users that looks like the seeds [2].

$$\mathrm{sim}_{\mathrm{cosine}}\,(\mathrm{fi},\mathrm{fj}) = \frac{\mathrm{fi}\cdot\mathrm{fj}}{||\mathrm{fi}||||\mathrm{fj}||} \qquad (1)$$

$$\mathrm{sim}_{\mathrm{jacard}}(\mathrm{fi},\mathrm{fj}) = \frac{\sum_{g=1}^{k}\min(\mathrm{fi}\;\cdot\mathrm{fj})}{\sum_{g=1}^{k}\max(\mathrm{fi}\;\cdot\mathrm{fj})} \qquad (2)$$

It is popularly used in small scale.

It carries all the important information the seeds in the user feature area carries.

Some drawbacks of the system are

- The computational complexity is in the order of O(kMN), where N are the candidate users and M are the seeds.A user can have k features on an average. The drawback is there can be trillions of candidate users and each user can have on an average hundreds of thousands of features in real online advertising world.

All the features of the seed user are treated equally without any priority. It does not have any provision of identifying the most similar candidate user.

C. Segment Approximation-based Look-alike System

In this type of system user are categorized based on various interest groups. The categorization is called segments [1]. For example, A user is said to belong to an art segment user is interested in paintings of MF Hussain or in the classical dance segment if she repeatedly visits the webpages of Bharatnatyam. In this method user feature based pre-built user segments are used. The goal is to evaluate those segments which are most frequently shared by as many seed users as possible.

Fig 4 shows the users segments based on the age and the location. The solid circles represent the seed circles and the hollow circles are the candidate users. The ranking of the features are done based on the criteria as per the demand. Based on this ranking top features are picked. The users that match those features are selected as lookalike audience.

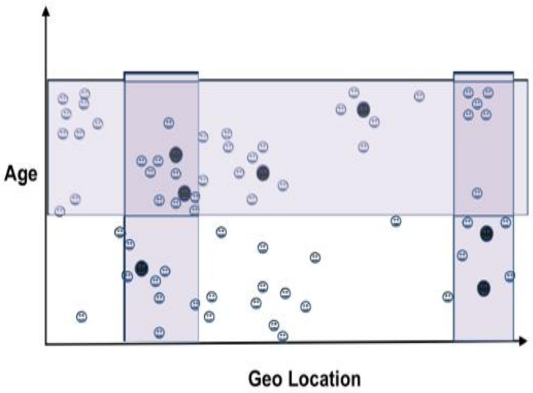

Fig 4: Shows the users segments based on the age and the location

For this method to work accurately it should have good inclusion over user interest with a immense pre-built segments quality. Its is effective for Big Brands as they have lots of coverage and funds for the same but not for small brands. A user's attribute has the feature field with humongous binary features having intense variety.

D. Score Alike Audience

Here a target set of audiences is provided by the advertiser initially. These target audiences may be previous consumers. Algorithms are use to discover the expanded similar users having same observable patterns like the target audience set. Advertisers are using this method to market their product to those audiences whose purchasing pattern is same like the ones who have already used their product. Once the algorithm is trained the advertiser don't have to explicitly give the target user set. The system will trace the user behaviour and automatically find the user bindings. Here the features of the user are not treated equally rather score is assigned to each user depending upon the criteria specified by the advertiser.A score Function is created by using techniques like LSH along with the kNN search. The higher the value of the score function suggests the more promising. The LSH techinique is used in two steps:

- Model building time: U the universe represents the set of all the users of advertisement system which are associated into hashed space of a lower range.
- Prediction time: The expanded user set is retrieved by associating the seed users into the same space.

Drawback of look-alike models is the goals of the advertiser may vary in wide range with respect to the working in different areas. So it is difficult assign score and one has to explicitly provide the promising value of a user to an advertiser.

E. Graph Constraint Look Alike System

This method is based on the simple similarity and regression-based methods. It is categorized into two phases:

- A user-2-user similarity graph is built at gobal level, It limits those candidate users to the nearest neighbors of seeds which are similar.
- The ranking of the candidate users are done on their ad campaigns specific feature. Fig 5 illustrates the graph-constraint look-alike model.

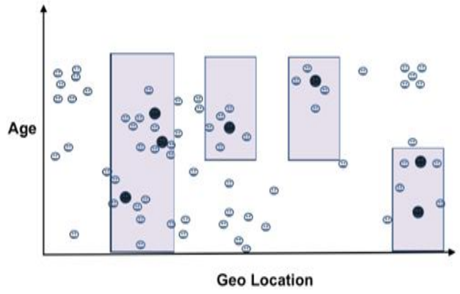

Fig 5: Shows the audience cluster with respect to age and geo location. Solid dots are the seed users and hollow dots are candidate users.

The vertical bars represent the critical features of a specific campaign. Solid dots are the seed users and hollow dots are candidate users inside the bar matches the advertisers criteria [1].The similar audiences are derived in two phases global graph construction and campaign specific modeling.

- Phase I: Global Graph Construction- It generates hash functions to refine the large set of user feature vector into confined amount of values of users. This restricted set of features is called signature. The user clustering is done based on signature. On arrival of the new user the mapping is done to find out the cluster it belongs to by hashing the user. There is no pair wise similarity search rather than individual user is hashed and respective cluster is associated with it [2][4]
- Phase II: Campaign Specific Modeling: In this phase the audience clusters from Phase I are cultured by adding scores based on feature weights specified by the campaign. The users with the highest ranks are the final audiences [1].

Different ad campaigns may care about different user features. A cake shops ad campaign may focus more about users' location not about their qualification. An TV series ad campaign focuses more on the users age and sex rather than location. An appropriate weight matrix for an ad campaign neglects the trivial features and discovers the similar audience which gives return on-investment for ad expenditure amount.

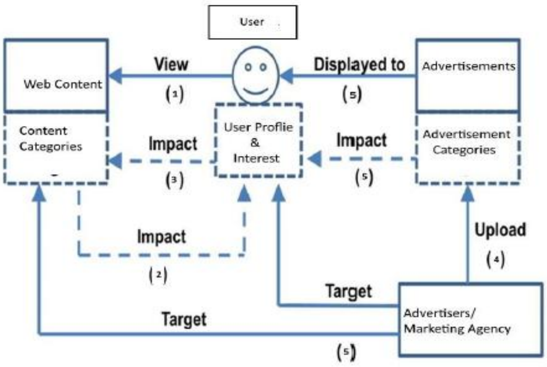

Fig 6: System Architecture with pipelining

F. Audience clustering using kmeans Algorithm

When objects can be categorized on different groups based on some criteria clusters are formed. It is the division of a data set into subgroups called as clusters, so that the data in the each cluster share some common trait. In online clustering the domain of users or objects classified or grouped into k numbers clusters based on attributes or features [18].

Here k symbolizes the number of clusters to be formed and it should be positive. k means uses the sum of square of distance between the data point and the centroid of the cluster and minimizes it. It is the simplest unsupervised learning algorithms. The main idea is to define k centers, one for each cluster. The centers should be placed as much far as possible because different result have different locations. For every point in the sample space of data or real world data, the association is done with the nearest center. After the last an initial cluster is derived and When no point is pending, initial cluster is formed and re-evaluate the value of k for the new centroids and the process repeats untill the centroids doesn't change further.

$$E = \sum_{i=1}^{k}(||x_i - v_i||)^2 \qquad (3)$$

Here c is randomly selected clusters. X is the collection of data points from x_1 to x_n and V is the collection of centers. It groups according to the features like geographic location, age, hobbies, webpage viewed, browser type, ad clicks etc.

G. Direct Interest-Aware Audience Selection

Whenever a user searches on the website or search engine a query log file or history is maintained. This method uses this log file to derive the user interests. In digital world, there are multiple users and each users have numerous histories. Every interest category is given a tag and the model is trained that assigns interest tags to users. The tags are implicitly created from the histories or log files. The advertisers takes advantages of this as the tags are available for each interest category with user association with it.

The approach has two phases:

- The modeling phase: History generated by various user queries creates large dictionary of interest. This dictionary is taken to train the model for mapping the tags of the queries and the inference phase.
- The inference phase,applies the tags to user query histories and obtains categorization of the users from interest dictionary. Fig 7 shows the modeling phase and the query phase.

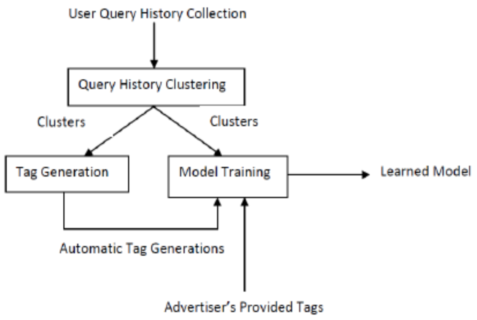

Fig 8: Illustration of two phases of direct Interest Aware Audience Selection

H. K-mediods Clustering Method Based On Map Reduce

K is the number of mediods to be determined. Based on this the number of clusters are formed it allocates n data points into k clusters which has k mediods. The main aim is to take care of the highest intra-cluster cohesion and lowest intercluster cohesion among objects. The unlikeliness between the n mediods and all the other data points in their respective clusters is reduced. Every cluster is recognized with a medoid. These mediods are considered as point of reference for respective clusters. The KMedoids algorithm is more powerful as it detects and handles outliers and noise more effectively as it minimizes a sum of pair wise differences. In step 1, k numbers of clusters are selected. In step 2, from the set of all objects the algorithm selects k number of objects randomly which symbolizes the mediods for the k clusters initially. The step 3, The remaining points are associate to cluster with either maximum similarity or minimum dissimilarity to each of the data points. In step 4, the previous medoids are re-identified by recalculating mean of all the points in their respective clusters. Last two steps are repeated untill there is no data point remaining [12].

The MapReduce is a way of processing extensive data in a parallel fashion which is beneficial in distributed and parallel computing platforms. Since it works in parallel and is able to perform faster and effective in processing of data[12] Hadoop and HDFS (Hadoop Distributed File System) is use to store large amount of data and further execution of it.

I. Direct Interest-Aware Audience Selection

The clustering techniques used is Self Organizing Map (SOM) followed by K-Means clustering. SOM is a Neural Network that is unsupervised machine learning. The network is build with the number of neurons to represent the classes with the inputs as features. The k means clustering is applied on each class [13].

J. Direct Interest-Aware Audience Selection

It treats and detects outliers in the online phase of the data streaming [14]. When there are cases of the new stream objects that cannot be inserted in any micro-cluster due to insufficient similarities, it uses an auxiliary memory to store them. From time to time, these objects stored in auxiliary memory are checked in micro-cluster and if any similarities are found they are inserted in the respective clusters. All the remaining objects are kept in the auxiliary memory until they are inserted or of no use .Then, residual objects are removed from it. It also uses CluStreamOD, to deal with outliers

K. Direct Interest-Aware Audience Selection

It first evaluates the audience value function. It finds the actual distribution features of the audience based on the division of the audience. Gradually it concludes to the factors responsible for the distribution of the audience. The value of the current audience serves as the index from which final clustering occurs. These values may be on the features like audience views, clicks, conversion rate, recent visits and total exposure time five attributes.

Then K Means parallel Algorithm is applied by dividing the audiences according the number of processors and each processor runs the cluster algorithm to local dataset and then is combined for all followed by dynamic load balancing [15].

L. Direct Interest-Aware Audience Selection

It can generate comparable clusters with monimum variation within two or more clusters. The main benefit of this methods is it eliminates dead centers. The algorithm works as follows: It uses Mahalanobis distance to compute interrelationship of the pixel with its parent and neighboring clusters. It is called statistical distance [16].

Based on variance selection criteria the elements with minimum correlation with their base cluster is selected .The selected element is then transferred to the appropriate nearby clusters. Since the elements having the minimum cluster to cluster are transferred the resulting cluster will have the low intra-cluster variance and homogeneity[16].

Fitness function is calculated by

$$F(c_j) = \sum_{i=c_j} ||v_i - c_j||^2 \qquad (4)$$

M. Direct Interest-Aware Audience Selection

Every website has parent page, grand parent page, child page and sibling page. This algorithm focuses on the value of information from sibling pages as it believes that the pages pointing to the same types of other pages gives maximum similar information. This can be useful for web page clustering and for improving clustering quality [17]. The algorithm is strengthened by two types of edge weighting techniques namely simple and normalized. The results of the experiments conducted concludes that (1) Sibling pages Information drastically improves clustering quality; (2) Sibling pages are more beneficial than parent and child pages (3) The clustering quality is not improved by the weighting and pruning sibling links [17].

N. Direct Interest-Aware Audience Selection

This algorithm proposes the improvements over kmeans++ algorithm in terms of accuracy. Here triangular inequality property is used. it states that if x is a datapoint and y and z are two clusters then

if the distance

dis(y,z) ≤ 2 * dis(x,y) then dis(x,z) ≥ dist(x,y)

and hence dis(x,z) can be avoided. It can improve the efficiency by reducing the avoidable distance calculations. This will improve the speed of kmeans++ algorithm.

IV. COMPARISION

Table 1 gives the comparison of various methods described by different authors. Comparison is done on the basis various factors like methods used, phases and techniques used, and the application or the usage area of each method. This methods are proposed by different authors given in the references and comparison is given in the table. The survey on different algorithms concluded that an advertiser can use any of the methods depending on the budget for the techniques and its usage area. as each methods has its own advantages and disadvantages.

V. CONCLUSION

The ad targeting in online advertising involves plenty of investment in it. The advertisement agencies prefers the clustering algorithm based on the requirements and targets the similar users for the placements on ads based on their interests. It increases the probabilities of viewing the ad. This paper surveyed different user clustering algorithm and compared the various techniques based on different criteria and suggested the usages based on the algorithms.

REFERENCES

[1] Q. Ma, E. Wagh, J. Wen, Z. Xia, R. Ormandi, and D. Chen, "Scorelook-alike audiences," in2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Dec 2016, pp. 647–654. https://doi.org/10.1109/ICDMW.2016.0097

[2] R. Battiti, "Using mutual information for selecting features in supervisedneural net learning,"IEEE Transactions on Neural Networks, vol. 5,no. 4, pp. 537–550, July 1994. https://doi.org/10.1109/72.298224

[3] S. Ioffe, "Improved consistent sampling, weighted minhash and l1sketching," in2010 IEEE International Conference on Data Mining,Dec 2010, pp. 246–255 .https://doi.org/10.1109/ICDM.2010.80

[4] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey ofweb information extraction systems,"IEEE Transactions on Knowledgeand Data Engineering, vol. 18, no. 10, pp. 1411–1428, Oct 2006. https://doi.org/10.1109/TKDE.2006.152

[5] G. Chatzopoulou, C. Sheng, and M. Faloutsos, "A first step towardsunderstanding popularity in youtube," in2010 INFOCOM IEEE Con-ference on Computer Communications Workshops, March 2010, pp. 1–6. https://doi.org/10.1109/INFCOMW.2010.5466701

[6] F. Duarte, F. Benevenuto, V. Almeida, and J. Almeida, "Geographicalcharacterization of youtube: a latin american view," in2007 LatinAmerican Web Conference (LA-WEB 2007), Oct 2007, pp. 13–21. https://doi.org/10.1109/LA-Web.2007.17

[7] S. Rallapalli, Q. Ma, H. H. Song, M. Baldi, S. Muthukrishnan, andL. Qiu, "Modeling the value of information granularity in targetedadvertising,"SIGMETRICS Perform. Eval. Rev., vol. 41, no. 4, Apr.2014.

[8] M. O. Shafiq and E. Torunski, "A parallel k-medoids algorithm forclustering based on mapreduce," in2016 15th IEEE InternationalConference on Machine Learning and Applications (ICMLA), Dec 2016,pp. 502–507. https://doi.org/10.1109/ICMLA.2016.0089

[9] C. Tsai, Y. Ding, M. Chiang, and C. Yang, "A novel clustering algorithmbased on searched experiences," in2017 IEEE International Conferenceon Systems, Man, and Cybernetics (SMC), Oct 2017, pp. 804–808 https://doi.org//10.1109/SMC.2017.8122707

[10] C. Lu, X. Zhang, J. Park, X. Hu, and T. He, "Web clustering based on theinformation of sibling pages," in2008 IEEE International Conferenceon Granular Computing, Aug 2008, pp. 480–485. https://doi.org//10.1109/SMC.2017.8122707

[11] Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan,"Adscape: Harvesting and analyzing online display ads," inProceedingsof the 23rd International Conference on World Wide Web, ser. WWW'14, 2014, pp. 597–608

[12] A. Pereira, E. R. d. F. Paiva, and M. C. Naldi, "Online detectionof outliers in clusters of continuous data streaming," in2017 BrazilianConference on Intelligent Systems (BRACIS), Oct 2017, pp. 324–329. https://doi.org//10.1109/BRACIS.2017.54

[13] Liu, "Design and implementation of network advertising precisemarketing system based on parallel k-means algorithm," in2014 IEEEWorkshop on Advanced Research and Technology in Industry Applica-tions (WARTIA), Sep. 2014, pp. 122–124 https://doi.org//10.1109/WARTIA.2014.6976207

[14] V. Vijay, V. P. Raghunath, A. Singh, and S. N. Omkar, "Variance basedmoving k-means algorithm," in2017

IEEE 7th International AdvanceComputing Conference (IACC), Jan 2017, pp. 841–847. https://doi.org//10.1109/IACC.2017.0173

[15] C. Lu, X. Zhang, J. Park, X. Hu, and T. He, "Web clustering based on theinformation of sibling pages," in2008 IEEE International Conferenceon Granular Computing, Aug 2008, pp. 480–485 https://doi.org//10.1109/GRC.2008.4664743

[16] C. Tsai, Y. Ding, M. Chiang, and C. Yang, "A novel clustering algorithmbased on searched experiences," in2017 IEEE International Conferenceon Systems, Man, and Cybernetics (SMC), Oct 2017, pp. 804–808 https://doi.org//10.1109/SMC.2017.8122707

[17] Q. Ma, M. Wen, Z. Xia, and D. Chen, "A sub-linear, massive-scale look-alike audience extension system a massive-scale look-alike audienceextension," inProceedings of the 5th International Workshop on BigData, Streams and Heterogeneous Source Mining: Algorithms, Systems,Programming Models and Applications at KDD 2016, ser. Proceedingsof Machine Learning Research, W. Fan, A. Bifet, J. Read, Q. Yang, andP. S. Yu, Eds., vol. 53.San Francisco, California, USA: PMLR, 14Aug 2016, pp. 51–67

[18] P. Tamilselvi and K. A. Kumar, "Unsupervised machine learning forclustering the infected leaves based on the leaf-colours," in2017 ThirdInternational Conference on Science Technology Engineering Manage-ment (ICONSTEM), March 2017, pp. 106–110. https://doi.org//10.1109/ICONSTEM.2017.8261265

AUTHORS PROFILE

Honey Vachharajani , M.E Computers, Teaching field 13 years experience, 5 national paper and 1 international paper, with ISSN 2320-3420, ISBN: 978-93-84659-91-2,, ISBN: 978-93-84659-91-2 ,and an international paper ISBN: 978-81-929339-1-7, life time member of I.S.T.E students and faculty chapter.

Dr. Rajeev Kumar Gupta earned his M.Tech and Ph.D. from MANIT, Bhopal. Prior to joining SISTec, he was associated with NIT, Bhopal as Assistant Professor. He has more than seven years of teaching experience in various organisations for PG & UG courses of Computer Science & IT. Published more than 25 research papers in various International Journals and Conferences including SCI and Scopus. Member of Coordination and Advisory Committee under TEQIP-III at Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV), Bhopal.

Prof Nikhlesh Pathik joined Sagar Group of Institutions in July 2015. Presently he is working as Associate Professor and Academic Coordinator. 5 papers published in National/International journals and conference.Active member of IEEE, SCRC.