Indian Journal of Science and Technology, Vol 9(37), DOI: 10.17485/ijst/2016/v9i37/101810, October 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Customized M-clustering Algorithm Comparison with Clustering Algorithms in Data Mining with the Case Study of Lead Generation Techniques

E. Manigandan[1], V. Shanthi[2] and Magesh Kasthuri[1]

[1]SCSVMV University, Enathur, Kanchipuram-631561, Tamil Nadu, India; ibmmani78@yahoo.com, magesh.kasthuri@wipro.com

[2]Department of MCA, St. Joseph's College of Engineering, Chennai – 600119, Tamil Nadu, India; drvshanthi@yahoo.co.in

Abstract

Objectives: Clustering algorithm is broadly used as spectral algorithm in social media, where a reference of contact is used and mined further for various combinations of suggested friends and lookups. This paper Identifies key lead generation techniques to be used in customer relationship management for sales marketing to decide data gathering. Also to define key merits and demerits of these techniques and to prepare a Matrix of comparison of these techniques to justify the data source and data set. M-Cluster algorithm is used for lead qualification i.e. training set preparation and data evaluation. **Methods:** Todefine a training set based on various attributes/fields in the data given for classification. This training set is used to run the data process and to produce expected result. This is assessed and accepted for definite data set processing or additional run for interim training data set preparation. **Findings:** This study is taken to customize clustering algorithm for data mining process in a customer relationship management field as the space of data is more and variant. Also proving the usability of customized clustering algorithm in data mining and the efficiency in processing mechanism compared to other methods used in current situation of data mining in customer relationship management is the part of this study. The customized algorithm developed as part of this study considers the two major areas of data mining using clusters. **Applications:** The results produced tremendous trends that the clustered algorithm suits to any data mining process when scaling and data classification are diversified and less in control.

Keywords: Classification, Clustering Algorithm, Data Mining, K-Means, Lead Generation, M-Clustering Algorithm

1. Introduction

Speed to market is precarious to companies that are focussed by sales in a competitive market. The past a prospective customer can be loomed in the decision making process of a purchase, the higher are the chances of changing that prospect into a customer. Old-fashioned approaches to find sales leads such as company surveys and direct marketing are manual, costly and not scalable.

Over the past decade the World Wide Web has grown into an information-mesh, with most significant facts being stated over Web sites. Numerous newspapers, press releases, trade journals, business magazines and other correlated sources are on-line. These sources could be used to find potential buyers automatically.

Lead generation is a technique used to invite or generate customer leads using technique like contextual advertising. It is an easy and effortless way of inviting people/users and humanising potential customers out of them. Pop-up advertising or subscription for free magazines are in a way gathers data about the users which is then filtered to produce leads. This process is called lead nurturing.

* Author for correspondence

2. Clustering Algorithms

In[1]Clustering algorithms are eye-catching for the task of class identification in spatial databases. Though, the application to huge spatial databases rise the subsequent necessities for clustering algorithms: nominal necessities of area information to decide the input constraints, finding of clusters with random shape and good efficacy on great databases. The well-known clustering techniques offer no result to the grouping of these requirements in Figure 1.

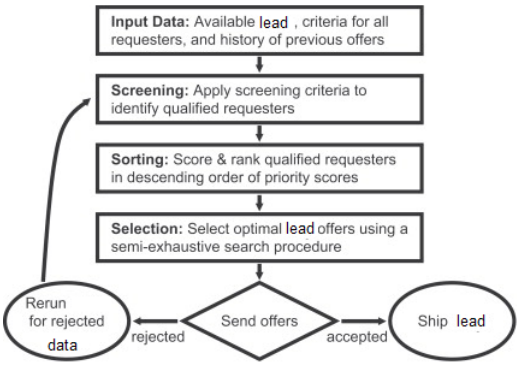

Figure 1. Flow diagram representing Lead generation and Data mining process[19].

In[2] Data mining includes in clustered material collected as 'raw data' from customers from several forums like social networking, tendencies in browsing pages, trends in search or pages visited. Studying such raw clusters of data which is huge in volume not only comprises several analytics techniques but also includes in filtering various necessary preference set which makes the base of data analysis.

Clustering Techniques supports in such a situation where on emphases analysis area and gather required subset of volumes of data collected from Lead generation and process them to filter the preference set and produce the essential results in terms of reports, diagrams, trend analysis and statistical data points[3].

3. Comparison with Prior Algorithms

K-Means is a classification algorithm that requires the information to be standardized (k)[4] so that no specific inconstant or subset of variables leads the analysis which is used as a Dimension-reduction tool when a data set has many characteristics. But it doesn't use training set for data model preparation and K-Means Clusters should have small within-cluster dissimilarity compared to between clusters Dissimilarity. Performance and error rate accuracy is improved in M-Cluster algorithm. This is historical research methodology whereas M-Cluster algorithm follows field research methodology.

Complete linkage algorithm[5] pursues to reduce the space amidst the records in two clusters that are outermost from each further and usages cluster for classification. As each cluster comprises a single record only, there is no dissimilarity among single linkage and complete linkage[6,7]. Uniformity in data analysis in this algorithm can result in dead-ended result if wrong 'means' is chosen. This algorithm is combination of supervised and historical methodology whereas M-Cluster is unsupervised/field research in terms of training set selection. Even, random sampling can be used for proper data set selection.

Predict churn is a data processing algorithm used for data analysis in churns. The churn is similar to the training set concept used in M-Cluster algorithm. The churn is selected randomly or based on historical data analysis results.

This means, the algorithm would be efficient by age. This is a supervised modelling and hence can go iterative process of refining the churn. M-cluster algorithm is an unsupervised training methodology[8].

4. Lead generation

In [9,10] Marketing approach in Online Business has evolved for the past few decades from product centric to a customer centric. A great looking website can build a brand image for your business but definitely the design or look-and-feel alone will not be fruitful to increase your sales. Now-a-days more businesses are shifting their services from local store (Offline sales) to online. In[11,12] Designing or developing a website is not too difficult, the problem begins while turning that same website into an excellent platform of marketing. Lead-gen is such marketing concept to increase business by generating leads in a B2B websites.

Partitioning and Hierarchical algorithms these two techniques doesn't fit for real-time data mining in a Customer Relationship Management system as they have following problem as per our study:

- In[13,14] Data set selection should be either random or based on historical data processing results. Hence, this works in a supervised methodology.
- The data storage (data model used to store the data set for filtering and processing) is not as efficient as they are not indexed on a matrix representation of multi-folded[15] field representation.
- Algorithm customization is not so generic or flexible solution as they re-create the processing cycle from the beginning. Hence it is a costlier process to run various iteration/cycles of the processing to refine the result ratio or reduce error rate.

To overcome all these, M-Cluster is customized in such a way that:

- Store the data model in clustering representation with a matrix format correlating each distinctive field in an indexed searchable manner.
- In[16,17] Runs in unsupervised approach by stowing different fields in indexed model and based on the result necessary, the field assortment and its equivalent test set choice would take place.

Clustering algorithm can be customized(M-Cluster) in such a way that it can run the training set cycle several times (and henceforward cost effective as the succession runs in a random selected Lead generation is a method used to invite or produce customer leads using lead generation methods like contextual advertising (target ads). It is an informal and simple way of inviting people/ users and nurturing potential customers out of them. The above examples of Pop-up advertising or subscription for free magazines are in a way collects data about the users which is then filtered to generate leads. This process is called lead nurturing.

Earlier, marketing people will take survey in roads, buses and then in road shows which is then matured to stalls in symposium, shows and tele-marketing and chain marketing. The advantages of these techniques are numerous but with the advent of online shopping and internet usage, these techniques have gone to its next dimension called Lead generation.

By concept, Lead generation is simple to understand but difficult and swirls when tries to implement it practically. When we understand the types of lead generation and the techniques that is popularly used for lead generation, we will get a fair idea of Lead generation concepts and methodologies in Figure 2.

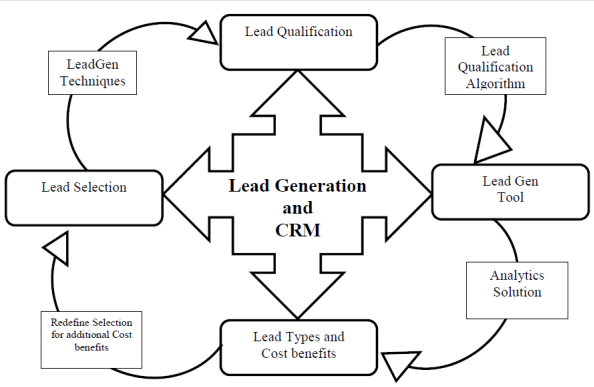

Figure 2. Lead generation cycle in a nutshell[15].

The topic of interest for this research is to study lead generation techniques and define a more effective technique for Data mining and generate Leads for CRM. Techtarget (www.techtarget) is one of the major lead-gen supplier in Information Technology industry. It has as many as 41 inter websites (theserverside, search SOA etc.,) which put out matters liberally to the users. *Registered* users can freely browse and download attracted articles from this site. The word 'Registered' is tinted in previous sentence as lead-gen starts from there in Techtarget where certain information through user is taken and traded to publicists based on the curiosity (lead nurturing).

Since Lead-gen is not a de-facto standard or owned by a single company, there are many techniques conceptualized for lead-gen by different practitioners. Techtarget is using as well as recommends the following popular lead-gen techniques.

- ***Banner Advertising*** – Advertising with banners on top, bottom and sides of the webpage
- ***E-Newsletters*** – Periodical newsletter to registered users with advertisements or lead-generating sources (ex: newsletter from theserverside.com)
- ***List Rentals*** – Rental or free space (subsects) in websites where sponsored advertisements will be displayed (esnips, blog space etc.,)
- ***Mediacasts*** – Any kind of mediacasts (video, podcast). A sample can be accessible freely and interested users can contact the concerned party for full version of media. Media based tutorials are example for this kind.
- ***White Paper Sponsorship Program*** – Articles to circulate freely ***only*** after sign-up with certain data about users (which is used for generating leads)

- ***Web Site Advertising*** – Any kind of advertisements in sites (Pop-up, adwares)
- ***Web Site Media Kits*** – Any free media kits like sample CD, audio clip etc.,

There are more lead-gen techniques which are popular but can be used only in specific areas and not considered as generic ideas. They are listed and discussed below:

Free special reports, articles – Provide a sample report (Credit Score reports in US), Technical and Management articles (Whitepaper, Sample chapters, Excerpts from books). This will induce the user to purchase the full version from the vendor.

Networking or chain referrals – Friend Referrals in chain (like introduce 10 friends during sign-up of torrents or social networking sites). Twitter, Facebook, MSN, Google chat uses this technique to broaden its user space.

Targeted email – Sending group announcement mails to specific group of targeted audience with brochures, discount details and invitation to join a network. ***Speaking engagements(Seminars, workshops, continuing education)***- Conducting seminars, Boot camp, training sessions and training lectures will attract specific audience to participate and go to the next step of trade enquiries.

Free consultations – Provide free consultation (Online or face-to-face) with limited areas thereby increase the business with the sample free consultation. Consultation for Tax payment, Medical check-up etc., comes under this category.

Targeted direct mail letters – Sending newsletter or discount mails to specific targeting audience by personally addressing their needs. Thorough study has to be done to gauge the user needs to frame the targeted email. This is different from Targeted email in that, this is intended for one-to-one correspondence whereas the former is meant for group broadcasts.

Social bookmarking – Use Social bookmarking sites like Digg, which can put your website on the first page of Google search result the very next day of adding your bookmark. Once you have submitted your bookmark in Digg, from next day onwards when you do a search for the keyword (which you bookmarked in Digg) in Google, you can get your webpage as the first result of search. There are also other social bookmarking sites like Reddit to try such concept.

5. Contextual Advertisement

Contextual advertising is a form of directedpromotionfor adsseeming on websites like adspresented on top and right of yahoo mail service. The ads themselves are carefully chosen and served by automaticschemes based on the content shown to the user and thus called ***contextual***. Sponsored advertisement and shopping advertisement based on searches from search engine are classic example of contextual advertisement. With the usage of rich media (streaming audio/video) these advertisements are much more attractive and powerful in cultivating leads.

Yahoo advertisement network, Google adsense and advertisement from one-click websites are fine examples of ready-to-use (plug) contextual advertisement servers. In simple, if we register our website with one-click and provide space for them, they will display sponsored advertisement in their space based on the area of surfing in our websites from our customer.

6. Business with Lead-gen

Lead nurturing is the vital part of lead-gen which concentrates required information to the advertisers from the leads generated.

Lead-gen is mainly conceptualized for better ROI (return of investment) through easy and meaningful way of marketing. Revenue can be generated from lead-gen in four broad categories as follows:

Cost per Ad – Costs or revenue to generate from the number of advertisement displayed. For example, advertisement displayed in Online magazine and e-newspaper like Chronicle, BBC, IBN.

Cost per Impression – Costs or revenue to generate from the number of impressions or attracted users from the site. For example, leads or enquiry from certain advertisements (Dental care, weight loss tool) in sites like Yahoo, AOL (free mail sites).

Cost per click – Costs or revenue that get generated from contextual advertisement sites like search engines (Google, Bing etc.,). For example, when we search for the word "IPhone 8GB" in Google, it gives shopping results and sponsored links from online shoppers like buy.com and amazon.com which is actually used to generate business by these online shoppers.

Cost per Lead – Costs or revenue generated out of leads or impressive usage in sites like Reply.com, Yahoo answers.

7. Lead Qualification

Once the leads are gathered from one or some of the above techniques, the raw leads ready to be processed and dispersed to publicists. They can be handledby hand or using CRM tools. When process the raw leads manually, it involves lot of experience to the processing team to carefully analyse and generate leads out of it. In[18] When the volume of leads is more or when the information to be processed in each lead is more and complex (meshed information between different data), then manual processing will be a cumbersome job and can result into failure results of preparing wrong leads or missed leads.

One can use Customer Relationship Management tools for this kind of lead processing. In market, a variety of Customer Relationship Management tools are available which can process the leads and even broadcast them online to required advertisers based on the demand of request. One such robust tool is Siebel CRM OnDemand which is offered to everybody on the Web, permitting organizations to produce more sales leads, enhance lead qualification (analyse the accuracy of information provided) and management, improve sales predicting (relative report with historical reports), less sales cycles, and givebetter customer service. With in-built analytics, all sales, marketing, and service professionals can instantly act on relevant business, customer, and market events.

Key steps and stages in the research involved are listed below:

- Identify key Lead generation techniques to be used in CRM for sales marketing
- Define key merits and demerits of these techniques and prepare a Matrix of comparison of these techniques[9]
- Generate an efficient Algorithm/pseudo steps for Lead qualification in Contextual Advertisement
- Prepare Case studies using this Algorithm to qualify Leads for Search engine based CRM
- Define steps in Cost effectness in Lead qualification (cost per lead) in search engine
- Define an Analytics solution for Drupal based CRM for Lead qualification

8. M-clustering Algorithm

Step-1: From a given Input set (s), prepare each set of elements for a group (of likely group from predictive column) into a cluster of elements

Step-1.1: Get user input on error rate expected € [tolerance level] and input base lining T (threshold)

Step-2: Identify the size of training set(T_S) based on T and € [$T_S = I * € / T$] where I is input volume (size).

Step-2.1: From each cluster, pick a set of elements (having relation) and examine if it can be used for predictive results (identification of training set) based on €

Step-3: If the training set is not effective, repeat step-2.1 until a predictive training set is identified

Step-4: For each cluster (Q)

Step-4.1: Delete the top element of Q (say u) and merge it with its closest cluster u.closest (say v) and calculate the novelillustrative points for the compound cluster w.

Step-4.2: Also eliminate U and v from T and Q.

Step-4.3: For all the clusters x in Q, update x.closest and relocate x (this is to update the address reference for relative child/node element in the cluster)

Step-4.4: Insert w into Q

Step-4.5: Repeat Step 4

Step-5: Traverse each cluster for the mean elements and prepare results based on the means of evaluation (M-Cluster evaluation).

9. Experimental Setup

Customized Clustering algorithm (M-Cluster) is using ranking based samples as compared to K-Means or othergenerally used data mining algorithms, there is a definite benefit in Customized ClusteringAlgorithm (M-Cluster) when used with continuous growing set of data and also in improvised time taken for processing the records.

M-Cluster is implemented in an experimental evaluation setup as a WEKA Attribute Selection class and Cluster classifier class to validate the processing time/ results from M-Cluster evaluated data.

For example, when a growing set of records (in both dimensions of fields and records) is fed to M-Cluster algorithm, we found the consistency in processing rate and time execution as tabulated below: Table 1.

Table 1. Table showing processing rate and time execution consistency[19]

Fields	Records	Total Data (Fields X Records)	Time taken with M-Cluster algorithm	
			To build Clusters (in ms)	Attribute selection (in ms)
5	14	70	62	78
5	24	120	39	15
5	150	750	16	234
17	57	969	23	110
7	209	1463	47	131
8	209	1672	63	167
36	683	24588	453	406
20	810	16200	437	188
9	768	6912	78	125
21	1000	21000	250	25
20	1500	30000	1093	328
9	12960	116640	1094	151
15	32561	488415	1753	6188

10. Conclusion

Clustering algorithms are most popular in Social networking area to build and extend customer focus areas which we propose to be used for any Data mining area to focus on customer needs and trend analysis.

Lead-gen is the future of current marketing trends and many websites developed now-a-days are developed facilitated with sophisticated lead-gen techniques to attract customers and to chain advertisers to generate business. As long as, the lead-gen techniques are not annoying to customers (too many adwares) and as long as customers give true information as leads, lead-gen can be proved to be most successful marketing trend. The challenge remains in the ways or ideas of attracting customers and filter fake and unwanted information from the lead generated.

11. References

1. Robles-Kelly A, Edwin R. Hancock, graph matching using spectral seriation and string edit distance. Lecture Notes in Computer Science. 2003; 2726:154–65
2. Kamble A. Incremental clustering in data mining using genetic algorithm. International Journal of Computer Theory and Engineering. 2010 Jun; 2(3):326–8.
3. Chapman H. Data clustering: algorithms and applications. CRC. 1 edition. 2013 Aug; 21:13–19
4. Chaudhari B, Parikh M. A comparative study of clustering algorithms using Weka tools. International Journal of Application or Innovation in Engineering and Management. 2012; 1(2):154–8.
5. Kumar S, Suseendran G. Incremental quality based reverse ranking for spatial data. Indian Journal of Science and Technology. 2016 Jan; 9(1):1–10.
6. Mirkin B, Nascimento S. Additive spectral method for fuzzy cluster analysis of similarity data including community structure and affinity matrices. Information Sciences. 2012; 183:16–34.
7. Davidson I, Ravi SS, Ester M. Efficient incremental constrained clustering. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007); 2007.p. 240–9.
8. Park H-H, Park J, Kwon Y-B. Topic clustering from selected area papers. Indian Journal of Science and Technology. 2015 Oct; 8(26):1–7.
9. Robles-Kelly A, Edwin R. Hancock, graph edit distance from spectral seriation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27(3):365–78.
10. Kyoo-Sung N, Doo-Sik L. Bigdata platform design and implementation model. Indian Journal of Science and Technology. 2015 Aug; 8(18):1–8.
11. Venkataraman S, Sivakumar S, Selvaraj R. A novel clustering based feature subset selection framework for effective data classification. Indian Journal of Science and Technology. 2016 Jan; 9(4):1–9.
12. Bikku T, Rao NS, Akepogu AR. Hadoop based feature selection and decision making models on big data. Indian Journal of Science and Technology. 2016 Mar; 9(10):1–6.
13. Karamizadeh F, Zolfagharifar SA. Using the clustering algorithms and rule-based of data mining to identify affecting factors in the profit and loss of third party insurance, insur-

ance company auto. Indian Journal of Science and Technology. 2016 Feb; 9(7):1–9.
14. Hariharan R, Mahesh C, Prasenna P, Kumar RV. Enhancing privacy preservation in data mining using cluster based greedy method in hierarchical approach. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–8.
15. Kamyab M, Delafrooz N. Investigating the effect of personality traits, subjective norms and perceptions of customers on using internet banking. Indian Journal of Science and Technology. 2016 Jan; 9(1):1–8.
16. Suganthi R, Kamalakannan P. Exceptional patterns with clustering items in multiple databases. Indian Journal of Science and Technology. 2015 Nov; 8(31):1–10.
17. Karthick N, Kalarani KA. An improved method for handling and extracting useful information from big data. Indian Journal of Science and Technology. 2015 Dec; 8(33):1–7.
18. Sajana T, Rani CMS, Narayana KV. A survey on clustering techniques for big data mining. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–12.
19. Delafrooz N, Farzanfar E. Determining the customer lifetime value based on the benefit clustering in the insurance industry. Indian Journal of Science and Technology. 2016 Jan; 9(1):1–8.