Enhanced Smart Phone based Crawler using K-Means Document Clustering

Kuldeep Varshney[1], Nandini Sharma[2]
[1] P.G. Student, Department of Computer Science & Engineering, SRCEM, Palwal, Haryana, India
[2] Assistant Professor, Department of Computer Science & Engineering, SRCEM, Palwal, Haryana, India
(E-mail: varshneykuldeep056@gmail.com)

***Abstract*—** Now every single person acknowledges that the smartphones have been serving distinctive indexed lists to masses than computer or web for quite last few years. In present, the smartphone serves various outcomes dependent on the handset you are utilizing to seek. The distinctions are regularly unobtrusive, or concentrated on the request of universal results that are incorporated into the portable outcome set, however using machine learning algorithmically attempting to organize content that will function admirably on the smartphones that presented the inquiry, and give high need to content that probably will work as potential information therefore it will take a short time and rendering contextual information from the smartphones using K-Means document clustering that the data is downloaded from smartphone using mobile crawler. Consequently, under this scheme we proposed to produce the mobile crawler which will crawl the data from smart phones and stores into machine repository as the corpus, subsequently using K-Means clustering the document cluster will be formed and the index will be originated to find out relevant information at the event when of required to the user.

***Keywords*—** *Machine Learning; Sockets; Crawlers; K-Means; Euclidean Distance; Indexing;*

I. INTRODUCTION

The internet world is having trillion of web documents or site pages and seeking archives which are progressively explicit with the client's Requirement is progressively troublesome to get appropriate and prompt information. The mobile networks and smartphones underpin dynamic substance which is developing progressively including news, current issues, new innovation, budgetary data, showcasing, excitement, instruction become generally dispersed over a wide region of mobile or smartphone phenomena. The web crawler, for the most part, downloads just the pertinent or explicit website pages as indicated by the client prerequisites instead of downloading all pages like conventional web indexes. So the essential objective of web crawler is to choose and search out the website pages that satisfy the client's prerequisite. The connection investigation calculations like page positioning calculation and different measurements are used to organize the URLs dependent on their positioning and choice approaches for downloading most explicit site pages. In most social applications accessible today the information is held for a restricted period and is ordinarily in the free arrangement. With a beneficial use of overall hunt utilizing Mobile Crawler. There's a lot of chance 4.28 billion individuals utilize cell or smartphones in the world. 62% (approx world population) use cell phones. There is no current versatile application comprising of careful usefulness that we need the scheme to give to the end client by which the prompt and appropriate information can be rendered instantly. The objective of our scenario is to fabricate an application where a client can utilize an application's information and furthermore in the event that any user needs to look through a specific element they can seek it on the web outside or to the smartphone using mobile networks and its application thereafter utilizing a mobile crawler which will be embedded in the smartphones as the application to crawl from the repository of phone or smartphone and download or render the information to machine for pre-processing, clustering, indexing and querying for the information.

Sockets: To convey over Transmission Control Protocol, a user/consumer or the client program and a server program initially set up an association with each other, with each program restricting an attachment to its finish of the association. To impart, the client and the server every read from and keeps in touch with the attachment bound to the association (utilizing a byte stream). Regularly, there is a system of gadgets (physically unmistakable, yet associated). These gadgets can be anything (printers, fax machines, Monitors, satellites, versatile smartphone, etc.) Each such gadget is alluded to as a hub in the system. Hubs which are fit for running projects are alluded to as hosts. The machine an end-client is using is alluded to as a neighborhood hub (localhost); different hubs are alluded to as remote. Therefore we used the socket programming where the bots are responsible to connect to smartphones using TCP/HTTP protocol and crawl from document listed in the storage of smartphones and download the same on the client machine to generate the corpus for clustering, indexing and querying the information. The below example depicts the modus operandi of socket programming comprising client and server both.

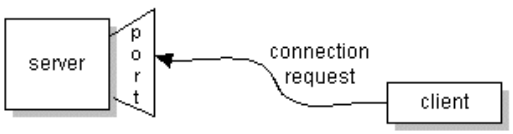

Figure 1 : To make a connection request, the client tries to rendezvous with the server on the server's machine and port.

Crawler: Mobile crawlers put into operation elegant crawling procedure to optimize their crawling process. Figure. 2 presents architecture of a mobile crawler. The main component of search engine, involved in the crawling process, is referred as crawling manager. The crawling manager is responsible for providing the details of documents on smartphones, which are targeted by mobile crawlers and monitors the crawled locations. The mobile crawler starts its operations by receiving a list of target smart phones using crawling manager.

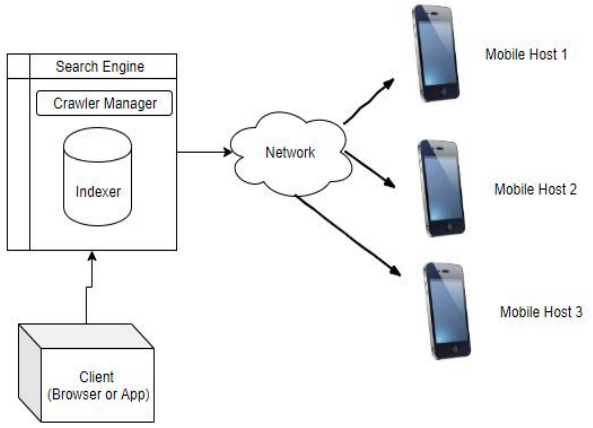

Figure 2: Architecture of Mobile Crawler

K-Mean Clustering: Information mining a particular region named content mining is utilized to order the tremendous semi-organized information needs legitimate grouping. Most extreme content reports include quick recovery of data, course of action of records, investigating of data from the archives. Declaration of content information and order of the archives is a mind-boggling process. The primary goal of this paper is to create a particular open source to class the clusters of indistinguishable records in the interrelated envelopes and to bring down the unpredictability of finding each archive. Calculations considered are difficulties for open research duties. Which portrays the record bunching process dependent on the grouping procedures, parceling grouping utilizing K-means and furthermore computes the centroids comparability and group similarity.

Process implicated for text clusters: Pre-processing of content involves expelling of undesirable clamor in the printed information utilizing Stop words calculation for each archive. Thereafter the features are produced by utilizing the base of the words on applying stemming calculations. After creating the words, each stem word is determined for their loads utilizing TF-IDF (Term Frequency-Inverse Document Frequency). In the wake of allocating the loads, decline the quantity of highlights and select highlights just that have greatest loads in the record. After getting the term-record network with loads applies dividing grouping calculation to amass the reports into K groups or clusters. Lastly estimations and investigation are done by the grouping/clustering technique using centroids. Beneath figure2 gives the stream graph chart for the whole procedure developed in text clustering.

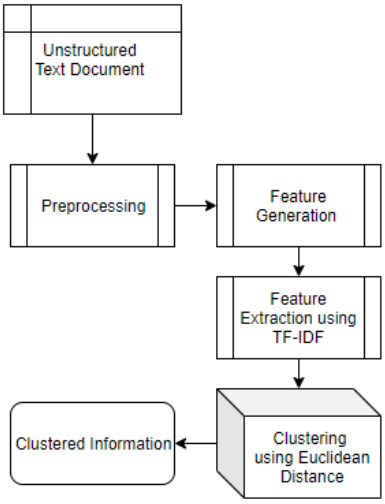

Figure 3: K-Means text Clustering Work Flow Model

Bitmap Indexing: The bitmap indexing is a substitute technique for the column ids indexes. It is easy to speak to, and utilizes less space-and CPU-effective than line ids when the quantity of unmistakable estimations of the listed segment is low. The files improve complex inquiry execution by applying minimal effort Boolean activities, for example, OR, AND, and NOT in the choice predicate on different files at one an opportunity to decrease look space before heading off to the essential source information. Numerous varieties of the Bitmap Index (Pure Bitmap Index, Encoded Bitmap, and so forth.) have been presented, expecting to decrease space necessity just as improve question execution.

		bitmap index			
RID	A	=0	=1	=2	=3
1	0	1	0	0	0
2	1	0	1	0	0
3	3	0	0	0	1
4	2	0	0	1	0
5	3	0	0	0	1
6	3	0	0	0	1
7	1	0	1	0	0
8	3	0	0	0	1
		b_1	b_2	b_3	b_4

Figure 4: An illustration of an equality-encoded bitmap index, whereas RID is the trace ID and A is an integer feature with principles in the assortment of 0 to 3.

II. LITERATURE REVIEW

Pavalam S. M., S. V. Kasmir Raja, Jawahar M., and Felix K. Akorli elaborated that, the mammoth upcoming of web innovation, information has detonated to an impressive sum. Huge volumes of information can be investigated effectively through web search tools, to remove profitable data. Web crawlers are a key piece of a web index, which is a program (continues with the hunt term) that can cross through the hyperlinks, lists them, parses the documents and include new connections into its line and the referenced procedure is completed a few times until pursuit term evaporates from those pages. The web crawler searches for refreshing the connections which have just been listed. This paper quickly audits the ideas of a web crawler, it's an engineering and its various sorts. It records the product utilized by different versatile frameworks and furthermore investigates the methods for use of web crawler in portable frameworks and uncovers the likelihood for further research.

Abhijeet tawde , Jayesh Patil ,Priya kurandale, Sharique khan elaborated that, web clients and users are developing quickly. Nowadays cause extraordinary inconvenience and exertion in the utilization Side to get the page being looked, which is of concern and Relevant client prerequisites for the general client approach Search for pages from a substantial number of accessible idea chains of command Use a question to peruse the web from an accessible web index And get results dependent on the pursuit design, a couple of them The outcomes are identified with the inquiry, and most are definitely not. Web crawlers assume a significant job in the web index Consider the key factors in execution.

Asst. Prof. Snehal Mane, Asst. Prof. Poonam Gholap , Asst. Prof. Rakhee Kundu depicts that, to recover data from the web we use Google, Yahoo, and MSN which are increasingly renowned web search tools. The internet searcher is one instrument to find data on the www. It looks for and distinguishes things in the database with reference to catchphrases entered by the client (where we get applicable information additionally which isn't actually what we which is tedious). For web creeping, we utilize concentrated crawler which dependent on philosophy engineering. Centered crawler look for site pages having more page rank for client prerequisite. Where cosmology is a particular about the space. It is a piece of man-made reasoning. Likewise, site pages nearby are preferred with metaphysics structure. In the plan, we utilize Focused Crawler and Ontology for the accurate administration related site.

Sachin Shinde , Bharat Tidke depicts that clustering is one of the unsupervised learning strategies in which a lot of fundamentals is isolated into uniform gatherings. The k-implies strategy is a standout amongst the most generally utilized grouping procedures for different applications. For the Searching just as perusing research papers clients need additional time or clients go through a few hours for seeking or perusing single papers, so this is an all the more expending procedure, so it is necessitated that utilization upgraded internet searcher which depends on the quickest perusing calculation which gives best yield or results. So we are proposed Enhanced design with improved K-implies calculation, which proposes a strategy for making the calculation increasingly powerful and effective, to show signs of improvement grouping with diminished unpredictability. It will look through the base catchphrase of the substance from the learning database. Proposed work utilizes the web index dependent on bunching and content mining.

III. PROPOSED WORK

In the above scheme we proposed amalgamated techniques using socket, crawling, clustering and indexing which is responsible from extracting the files from the mobile phones means with the help of sockets the data will be fetched from mobile phones, then it's fairly easy to avoid bot-traps of the infinite loop kind to crawl into smart phone's and download the documents into machines from the smart phone's .Then onwards the k-mean cluster will be formed and subsequently, the help of bitmap index the documents or information will be delivered to clients. Below are the workflow steps for proposed scheme:-

1. With The help of Socket Connection will be established with smart phones.

2. Mobile Crawler will crawl the document available on smart phone and download the same on downstream to computer or node.

3. With the help of preprocessing technique the noise will be removed from the documents

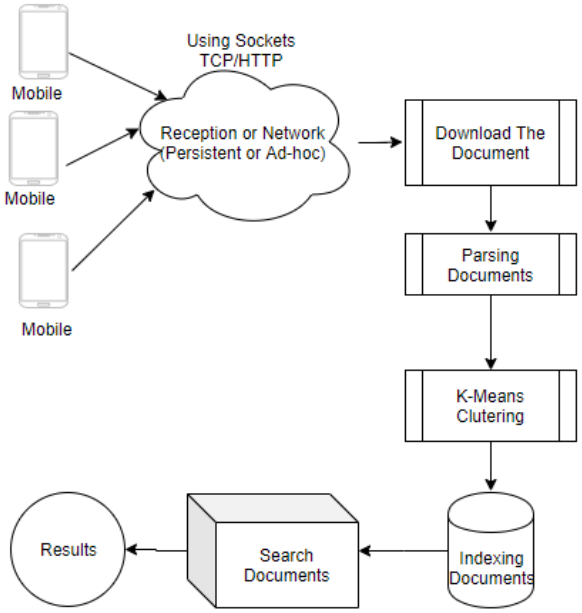

Figure 5 Workflow or Implementation Diagram of Proposed Scheme

Below are the steps elaborated to achieve the proposed scheme:-

Step 1. Socket to Connect Smart Phones

a) Server Model

Create socket

Bind socket to a specific port where clients can contact you

Loop
(Receive Stateless/Statefull Message from client x)+
(Send Stateless/Statefull Reply to client x)*
Close Socket
Create socket
Bind socket to a specific port where clients can contact you
Loop
(Receive Stateless/Statefull Message from client x)+
(Send Stateless/Statefull Reply to client x)*
Close Socket

b) Client Model
Create socket
Loop
(Send Message To Well-known port of server)+
(Receive Message From Server)
Close Socket

Step 2 : Crawling Algorithm in mobile phones

1. migrate to smart phones;
2. Iterate from the phones repository;
3. for all documents ∈ document_list do begin
4. load page;
5. extract page schema;
6. download the page on down stream at node

Step 3 : K-Means

1. Clusters the data into k groups where k is predefined.
2. Select k points at random as cluster centers.
3. Assign objects to their closest cluster center according to the Euclidean distance function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

Step 4: Index Search Model.

1. Bitmap--Search(x, k) returns (y, i) such that keyi[z][y]
2. keyi[z][y] = k or nil
3. i ← 1
4. while i ≤ n[x] and k > keyi[z][y]
5. do i ← i + 1
6. if i ≤ n[x] and k = keyi[z][y]
7. then return (x, i)
8. if leaf[x]
9. then return nil
10. else Disk-Read(ci[x])
11. return B-Map-Search(ci[x], k[z][y])

IV. SIMULATION RESULTS

The above-proposed scheme is been evaluated using different stages and the test error and is less an accurate potential information retrieval is very high to perform the above-mentioned models and measures we implemented the eco-system based on windows and android phones comprising of Android framework for communication-based on sockets using stateless connectors based on hypertext transfer protocol on transmission control protocol. Thereafter the pre-processing, clustering and indexing is performed using K-Means and for indexing and querying and searching the Bitmap index is responsible. Therefore the below chart with data labels depicts the time frame utilized to evaluate the time using the various resource with different infrastructure.

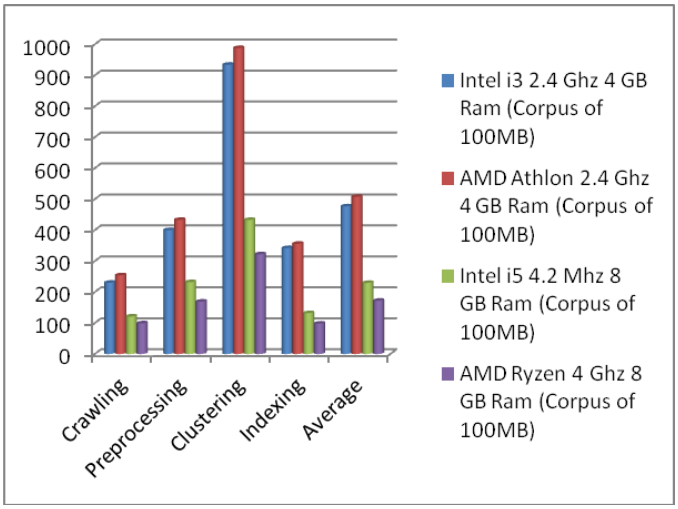

Figure 6: Time in seconds performed on various hardware resources

CONCLUSION AND FUTURE SCOPE

The above proposed scheme holds the enhanced measure to establish the connectors with stateless sockets using http for less time and quick communication with smart phones thereafter crawlers iterates to storages of the Smartphone's and thereafter originates the data stream downloads the data using serialization form with byte stream models for fast downloads. Herein, The preprocessing techniques are based on corpus or noise cleaning using application programming interface omits the noise and generate the text documents for further processing. Therein the K-means using Euclidean method find and evaluate the accurate distance and last not the least the bitmap index is responsible to originate the clustered information into indices for quick and prompt retrieval.

For the future scope the firmware can be created as listener program over the smart phones regardless to their make, ecosystem and configuration whereas using the Global System for Mobile Communication either using Wireless Transfer Agents or Satellite the communication can be established and data from the smart phones can be downloaded into central data repositories like big-data vide cloud so the optimum solution can be created to access to appropriate and accurate information using K-Means and Bitmap indexes promptly.

REFERENCES

[1] Pavalam S. M., S. V. Kasmir Raja, Jawahar M., and Felix K. Akorli , Web Crawler in Mobile Systems, International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012

[2] Abhijeet tawde , Jayesh Patil ,Priya kurandale, Sharique khan, Implementing a web crawler in a smart phone mobile

application, Emerging Technologies in the Computer World", January -2017, International Journal of Advance Engineering and Research Development e-ISSN (O): 2348-4470, p-ISSN (P): 2348-6406

[3] Asst. Prof. Snehal Mane , Asst. Prof. Poonam Gholap , Asst. Prof. Rakhee Kundu, Web Focused Crawling based on Ontology, International Advanced Research Journal in Science, Engineering and Technology Vol. 2, Issue 12, December 2015

[4] Sachin Shinde , Bharat Tidke, Improved K-means Algorithm for Searching Research Papers Sachin Shinde et al , International Journal of Computer Science & Communication Networks,Vol 4(6),197-202

[5] M. Theobald, R. Schenkel, and G. Weikum, "Classification and focused crawling for semistructured data," Intelligent Search on XML Data, pp. 145-157, 2003.

[6] C. Li, L. Zhi-shu, Y. Zhong-hua, and H. Guo-hui, "Classifier-guided topical crawler: a novel method of automatically labeling the positive URLs," presented at the Proceedings of the 5th International Conference on Semantics, Knowledge and Grid (SKG), Zhuhai, China, 2009.

[7] H. Liu, E. Milios, and J. Janssen, "Focused Crawling by Learning HMM from User's Topic-specific Browsing," presented at the Proceedings of the

[8] IEEE/WIC/ACM International Conference on Web Intelligence (WI), Beijing, China, 2004.

[9] T. K. Shih, "Focused crawling for information gathering using hidden markov model," Master's thesis, Computer Science and Information Engineering, National Central University, Taiwan, 2007.

[10] S. Chakrabarti, M. Van den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," Computer Networks, vol. 31, pp. 1623-1640, 1999.

[11] Y. Ye, F. Ma, Y. Lu, M. Chiu, and J. Z. Huang, "iSurfer: A focused web crawler based on incremental learning from positive samples," presented at the Advanced Web Technologies and Applications, 2004.

[12] G. Pant and P. Srinivasan, "Link contexts in classifier-guided topical crawlers," IEEE Transactions on Knowledge and Data Engineering, vol. 18, pp. 107-122, 2006.

[13] I. Partalas, G. Paliouras, and I. Vlahavas, "Reinforcement learning with classifier selection for focused crawling," presented at the Proceedings of the 18th European Conference on Artificial Intelligence (ECAI) Amsterdam, The Netherlands, 2008.

[14] H. Zhang and J. Lu, "SCTWC: An online semi-supervised clustering approach to topical web crawlers," Applied Soft Computing, vol. 10, pp. 490-495, 2010.

[15] Ch. Makris, E. Theodoridis, I. Panagis, A. Perdikouri and E. Christopoulou. Retrieving information

[16] D. Boswell. Distributed High-Performance Web Crawlers: A Survey of the State of the Art, December 10, 2003.

[17] A. Heydon and M. Najork. Mercator: A Scalable, Extensible Web Crawler, Compaq Systems Research Center 130 Lytton Ave., Palo Alto, CA 94301.

[18] J. Cho and H. Garcia-Molina. The Evolution of the Web and Implications for an Incremental Crawler. Department of Computer Science, Stanford, CA 94305, December 2, 1999.

[19] WebCrawler Timeline

[20] Pavalam S. M., S. V. Kasmir Raja, Jawahar M., and Felix K. Akorli, Web Crawler in Mobile Systems, International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012

[21] Hao Li [17],The System Design of Mobile Application Crawler and The Implementation of Some Key Technologies,Examensarbete 30 hp June 2016

[22] Md. Abu Kausar and V. S. Dhaka,An Effective Parallel Web Crawler based on Mobile Agent and Incremental Crawling, Journal of Industrial and Intelligent Information Vol. 1, No. 2, June 2013

[23] The Web Robots Pages. Http://info.webcrawler.com/mak/projects/robots/robots.html

[24] David Eichmann. The RBSE Spider - Balancing Effective Search Against Web Load. In Proceedings of the First International World Wide Web Conference, pages 113--120, 1994.

[25] Oliver A. McBryan. GENVL and WWW: Tools for Taming the Web. In Proceedings of the First International World Wide Web Conference, pages 79--90, 1994.

[26] Brian Pinkerton. Finding What People Want: Experiences with WebCrawler. In Proceedings of the Second International World Wide Web Conference, 1994.

[27] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the Seventh International World Wide Web Conference, pages 107--117, April 1998.

[28] Google! Search Engine Http://google.stanford.edu/

[29] Mike Burner. Crawling towards Eternity: Building an archive of the World Wide Web. Web Techniques Magazine, 2 (5), May 1997.

[30] The Internet Archive. Http://www.archive.org/

[31] Web crawler. Http://en.wikipedia.org/wiki/Web_crawler

[32] Larbin Multi-purpose web crawler Http://larbin.sourceforge.net/index-eng.html

[33] WebSPHINX: A Personal, Customizable Web Crawler Http://www.cs.cmu.edu/~rcm/websphinx