ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Review on Mobile Phone Crawler using K-Means Clustering

Neetu Kumari, Gurpreet Singh Saini, Arko Bagchi

M.Tech Student, Dept. of CSE, Delhi College of Technology & Management, Palwal, Haryana, India

Asst. Professor, Dept. of CSE, Delhi College of Technology & Management, Palwal, Haryana, India

H.O.D, Dept. of CSE, Delhi College of Technology & Management, Palwal, Haryana, India

ABSTRACT: Over the last decade, the technology landscape all over the world has undergone a massive transformation. Leading the charge of evolution has been the advent of the Internet and the boom in the use of smart devices. The mobile platform has taken the world by storm with more than half of the world population currently using smart mobile devices. This phenomenon has opened up a world of lucrative opportunities for businesses worldwide. Companies can now use the mobile platform for all kinds of critical business operations like marketing, sales, lead acquisition, customer relationship management and a variety of other important procedures. Consequently, the realm of mobile web has also grown exponentially over the last few years. Companies are racing against time and competition, to feature mobile friendly or responsive websites to capture the large, connected smart phone user audience. As a result, many of the important business processes pertaining to the Internet at large have crossed over to the domain of mobile world. Mobile crawling or scraping the phones on the same heterogeneous networks for business intelligence is also something which is successfully transitioning to the mobile platform. Mobile crawling is currently a process that many companies are actively doing in order to take advantage of the vast and versatile mobile phones platform and extract business-critical, actionable data which can be used in formulating business strategies and plans and for myriad other important business processes.

Phone crawler program relies on portable crawlers. Portable crawler could be a substitutive approach for net crawler that crawls the information and conjointly acts as a server by that crawling process will simply get live image conjointly, and it's appropriate for the user. There is not any sort of handy machine that acts as a server, portable as a server works and crawl the information. In gift offered portable platforms is robot, Motion analysis, Apple iPhone OSand Symbian. robot mobile phones square measure the first immense OS threw 2016. UNIX system s/w utilized in robot that work on admittance the essential laws requested by the user. this analysis work is consummated that is price and time. I developed a portable computer program server. It crawl the information and conjointly send live image to the sender. Here portable is acts as a server. within the chapter is half-dozen showing the code in chapter five in conjunction with the block diagrams of every module. The system created during this analysis needs solely robot portable and Wi-Fi network affiliation. portable crawler labor under a filter and filter pages that not modified from the time once last crawl happened. portable crawler presents simplest and gifted looking out, that sort of crawler supported robot java surroundings. By using that sort of mobile crawler procedure, its smaller searches compare to different search engines. By this crawler system presentation improved, reason behind of this can be those pages that don't seem to be customized and not repossess, in conjunction with this close to photocopy recognition feature adds a lot of privilege to scale back unwanted downloads.

KEYWORDS*:* Mobile Crawler, Mobile Agent, Search Engine, Document Clustering.

DOI: 10.15680/IJIRCCE.2017. 0503334

I. INTRODUCTION

In the proposed scheme the below information give help For demonstrate of problems. Given this volatile growth, there are following problems which currently index in the web these problems are:

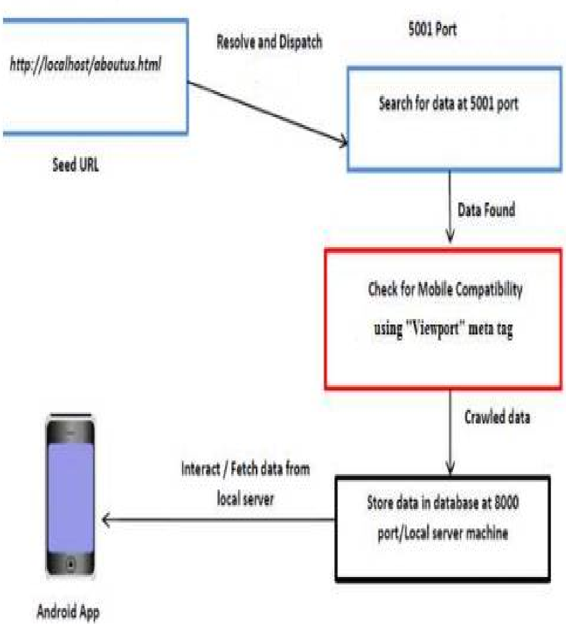
Effectiveness: At presented investigate machines add needless convey to the already overcrowded Internet. In progress of web one alternative process is Mobile phone search engine which build indexes. Many phases are presented with high efficiency for downloading. By this scenario particular mobile phone search engines and justify their usefulness.

*Scaling***:** Web data recovered by 55 Mbit per second whenever estimation for downloading was 80 TB of pages per day. Using this technology it was maintain the indexes. Using that estimates for growth of Web indices provided in 1998, a web search give focus on limitations of technology which have storage space and networks which communicate, phone crawler server can maintain the indices very effectively.

Quality of Index: For the query processor part of result of web searches are necessary and required. Quality of search results are not increases automatically because of increasing the size of web indexes. In present available search engines are maintain 110 million pages Error! Reference source not found. and find approx thousands matches related to search result. In view of the fact that in the web search, that cannot limit the pages. Accommodation of growing web, it is necessary for finding a way which improves the searching result. For search the data very easily and early necessary to develop the Mobile phone crawler in the near future. Using assortment of original knowledge today's phone search provide better search result which challenge devoted search engines by using data mining, graph theory areas. In addition, for improving the competence of data gathering, it necessary that a new mobile phone search engine is needed when user done searching in grouping search engine. Two main issues are associated with crawling system. Crawling policy is the main first issues associated with Mobile phone crawling system, this policy makes a decision for downloading the next page. Other second issues that necessary of a highly optimized organization structural design which work on download many pages per seconds healthy touching hurtle. In this search engine search engine firstly make indices of all documents which available in mobiles SD cards. That way by mobile phone using indices provide controlling search facilities. By using mobile phone crawler approach it pass through a filter of unchanged links from remote server without downloading the links and only download modified pages. Crawlers used mobile agent for building web indices that way crawler proposed crawling approach. This approach called Mobile phone crawler. Locally all data be located in the indices, by using data crawler it transported to the site of the source and filter out unwanted data locally and search the related query given by user. In this approach that mobile phone search engine acts as a server. There we don't want any type of network connection only router should be used for such crawling system. That way this approach done work very fast in locally and also perform the search very clear in globally. Today's positive reception of hardware industry is better compare to other previous devices. Which has quicker processors, more rapidly Internet connections, high quality sensors and also intelligent to host more demanding applications. Mobile phone device's applications are developed by using java Platform, android, .NET Compact Framework, Flash Lite and Java ME which provide highly purposeful mobile multimedia Applications. These technologies agree to the use of various technologies, like Java, Android etc. Regrettably, web server's quality is thinkable.

DOI: 10.15680/IJIRCCE.2017. 0503334

Figure 1: Depicts the scenario.


http://localhost/aboutus.html
Seed URL
Resolve and Dispatch
5001 Port
Search for data at 5001 port
Data Found
Check for Mobile Compatibility using "Viewport" meta tag
Crawled data
Store data in database at 8000 port/Local server machine
Interact / Fetch data from local server
Android App

K-mean clustering **:** K-mean document clustering is the part of Partitioning Clustering analysis which aims to form k groups from the n data points taken as an input. This partitioning happens due to the data point associating itself with the nearest mean. Clustering has been investigated extensively since traditional clustering algorithms often fail to detect meaningful clusters in high-dimensional data spaces, many recently proposed subspace clustering methods suffer from two severe problems: First, the algorithms typically scale exponentially with the data dimensionality and/or the subspace dimensionality of the clusters. Second, for performance reasons, many algorithms use a global density threshold for clustering, which is quite questionable since clusters in subspaces of significantly different dimensionality will most likely exhibit significantly varying densities. In this scheme, our framework is based on an ancient filter, refinement architecture that scales at most quadratic w.r.t. the data dimensionality and the dimensionality of the subspace clusters. It can be applied to any clustering notions including notions that are based on a local density threshold. A broad experimental evaluation on synthetic and real-world data empirically shows that our method achieves a significant gain of runtime and quality in comparison to state-of-the-art subspace K-means document clustering algorithms.

Figure 2: Depicts the same.

$$\mathrm{RSS}_k = \sum_{\vec{x}\in\omega_k} |\vec{x} - \vec{\mu}(\omega_k)|^2$$

$$\mathrm{RSS} = \sum_{k=1}^{K} \mathrm{RSS}_k$$

Where

ω_k	Document cluster k
$\vec{\mu}$	Mean or centroid of the documents in cluster ω_k
$\vec{x}$	Document vector in cluster k

II. EXISTING WORK

Flow of Web Crawler: Figure 3: Shows the flow of a basic consecutive crawler. The crawler maintains a list of unvisited URLs called the frontier.

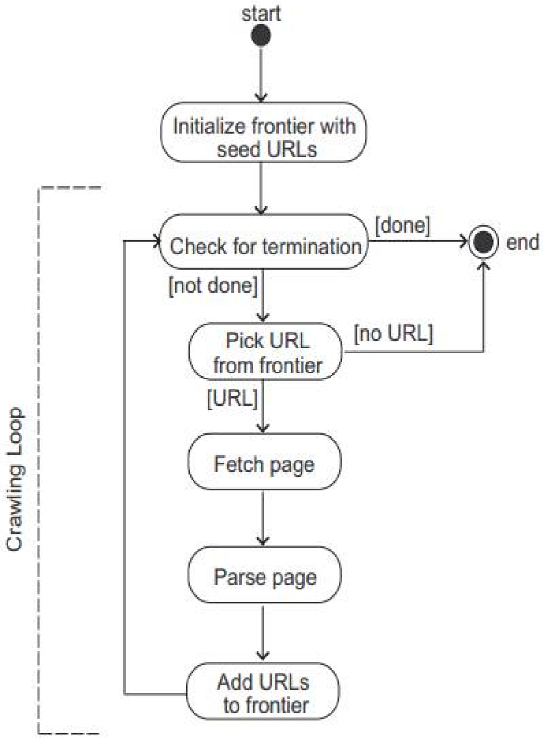


Figure 3

The list is initialized with seed URLs, which might be provided by a user or another program. Each travel loop involves choosing the next address to crawl from the frontier, fetching the page corresponding to the address through hypertext transfer protocol, parsing the retrieved page to extract the URLs and application-specific information, and finally adding the unvisited URLs to the frontier. Before the address is superimposed to the frontier they might be appointed a score that represents the calculable good thing about visiting the page comparable to the URL. The crawling method might be terminated once a precise range of pages are crawled. If the crawler is ready to crawl another page and therefore the frontier is empty, the situation signals a dead-end for the crawler. The crawler has no new page to fetch, and hence it stops. Crawling will be viewed as a graph search downside. The Web is seen as an oversized graph with pages at its nodes and hyperlinks as its edges. A crawler starts at a few of the nodes (seeds) then follows the perimeters to succeed in other nodes. The process of taking a page and extracting the links inside it's analogous to increasing a node in graph search. A topical crawler tries to follow edges that square measure expected to lead to parts of the graph that are relevant to a subject.

III. LITERATURE REVIEW

M. Theobald, R. Schenkel, and G. Weikum [1]Despite the great advances in XML data management and querying, the currently prevalent X Path or XQuery-centric approaches face severe limitations when applied to XML documents in large intranets, digital libraries, federations of scientific data repositories, and ultimately the Web. In such environments, data has much more diverse structure and annotations than in a business-data setting and there is virtually no hope for a common schema or DTD that all the data complies with. Without a schema, however, database style querying would often produce either empty result sets, namely, when queries are overly specific, or way too many results, namely, when search predicates are overly broad, the latter being the result of the user not knowing enough about the structure and annotations of the data. An important IR technique is automatic classification for organizing documents into topic directories based on statistical learning techniques. Once data is labeled with topics, combinations of declarative search, browsing, and mining-style analysis is the most promising approach to find relevant information, for example, when a scientist searches for existing results on some rare and highly specific issue. The anticipated benefit is a more explicit, topic-based organization of the information which in turn can be leveraged for more effective searching. The main problem that we address towards this goal is to understand which kinds of features of XML data can be used for high-accuracy classification and how these feature spaces should be managed by an XML search tool with user-acceptable responsiveness. This work explores the design space outlined above by investigating features for XML classification that capture annotations (i.e., tag-term pairs), structure (i.e., twigs and tag paths), and ontological background information (i.e., mapping words onto word senses). With respect to the tree structure of XML documents, we study XML twigs and tag paths as extended features that can be combined with text term occurrences in XML elements.

C. Li, L. Zhi-shu, Y. Zhong-hua, and H. Guo-hui, T. K. Shih [2,4]Finding information on WWW is difficult and challenging task because of the extremely large volume of the WWW. Search engine can be used to facilitate this task, but it is still difficult to cover all the webpages on the WWW and also to provide good results for all types of users and in all contexts. Focused crawling concept has been developed to overcome these difficulties. There are several approaches for developing a focused crawler. Classification-based approaches use classifiers in relevance estimation. Semantic-based approaches use ontologies for domain or topic representation and in relevance estimation. Link analysis approaches use text and link structure information in relevance estimation. The main differences between these approaches are: what policy is taken for crawling, how to represent the topic of interest, and how to estimate the relevance of webpages visited during crawling. We present in this report a modular architecture for focused crawling. We separated the design of the main components of focused crawling into modules to facilitate the exchange and integration of different modules. We will present here a classification-based focused crawler prototype based on our modular architecture.

H. Liu, E. Milios, and J. Janssen,S. Chakrabarti, M. Van den Berg, and B. Dom[3,5]Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The patterns, associations, or relationships among all collected data can provide information which can be converted into knowledge about historical patterns and future trends. A Web crawler may also be Web search engines and some other sites use Web crawling software to update their web content or indexes of others sites web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly.

G. Pant and P. Srinivasan[7]Context of a hyperlink or link context is defined as the terms that appear in the text around a hyperlink within a Web page. Link contexts have been applied to a variety of Web information retrieval and categorization tasks. Topical or focused Web crawlers have a special reliance on link contexts. These crawlers automatically navigate the hyperlinked structure of the Web while using link contexts to predict the benefit of following the corresponding hyperlinks with respect to some initiating topic or theme. Using topical crawlers that are guided by a Support Vector Machine, we investigate the effects of various definitions of link contexts on the crawling performance. We find that a crawler that exploits words both in the immediate vicinity of a hyperlink as well as the entire parent page performs significantly better than a crawler that depends on just one of those cues. Also, we find that a crawler that uses the tag tree hierarchy within Web pages provides effective coverage. We analyze our results along various dimensions such as link context quality, topic difficulty, length of crawl, training data, and topic domain. The study was done using multiple crawls over 100 topics covering millions of pages allowing us to derive statistically strong results.

IV. PROPOSED WORK

In the scheme we will produce the mobile crawler using mobile agent which will establish the socket connection with mobile phone and will crawl the mobile phone data storages and over a downstream will download the data in document repository and will form a document cluster by which the search will be formed and document will be rendered to targeted clients, below figure.4 depicts the scenario.

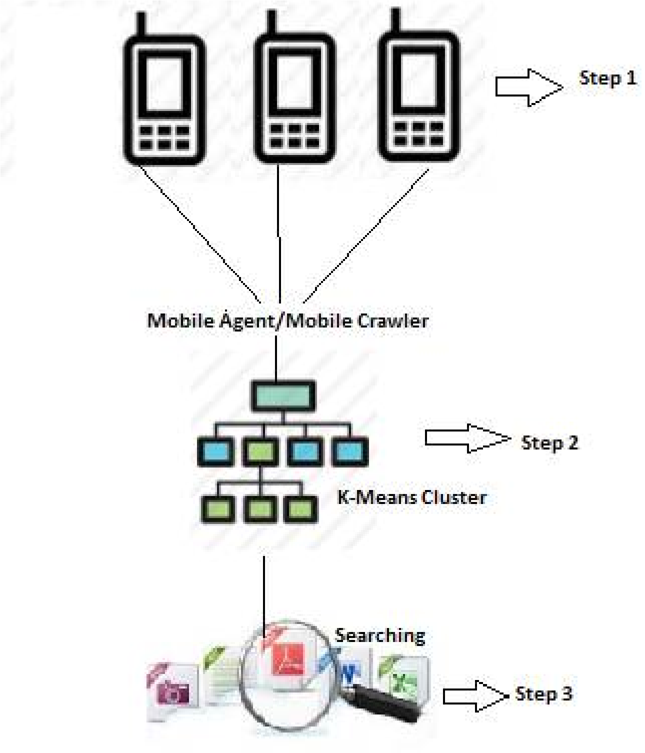


Figure 4.

REFERENCES

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

[2] 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.

[3] H. Liu, E. Milios, and J. Janssen, "Focused Crawling by Learning HMM from User's Topic-specific Browsing," presented at the Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI), Beijing, China, 2004.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] F. Menczer, G. Pant, and P. Srinivasan, "Topic-driven crawlers: Machine learning issues," *ACM TOIT*, 2002.