ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 8, August 2017

Improve Performance of Crawler Using K-means Clustering

[1] Swati G. Bhoi,[2] Prof. Ujwala M. Patil
[1] M.E. Scholar, [2] Associate Professor
[1][2] Department of Computer Engineering, SES's R. C. Patel Institute of Technology, Shirpur, India

***Abstract*— Nowadays the Internet is part of life because of any information is easily available on the Internet. It has a large size of information; hence the high efficiency and get relevant information are challenging issue due to the changing nature of the deep web. As crawler plays important role in such cases. So we proposed such crawler which provides efficient and extracts relevant information from web. The smart crawler contains two-phases as site locating and in-site exploring. We developed smart crawler using K-means clustering methods. Clustering makes a group of similar data items known as clusters. Here we describe K-means clustering techniques. The most famous clustering method is K-means methods which divide data items in K clusters and provide better result with high efficiency. Also we compare the result of existing system and smart crawler using K-means provide an efficient harvesting rate of deep websites within the least amount of time.**

***Index Terms*— Deep Sites, SCDI, ACHE, Crawler, K-means, URLs.**

1. INTRODUCTION

The web crawler fetches relevant information from large content which is present on the web. The crawler is the program or tool, we can also call it as a spider [1]. The available information on the internet is divided by crawler as per its relevancy, but some problem occurred during this process. The problem is that various types of data presents on the Internet so the complication occurs when data will be extracted and further processing done on these data. In clustering, data can be partitioned and form set of similar data objects. The clustering is an unsupervised method for data analysis [2]. Clustering can parse a data and makes its groups by referring information related to data that computed objects and relationships between them. The some clustering methods discussed below: [3] -[4].

A. Partitioning methods

In this partition, identify data objects and all clusters computed once after that the cluster divide into distinguishable cluster with precautions that compulsory each data object is present at least in the single cluster.

B. Hierarchical methods

Hierarchical method divides into two types as agglomerative and divisive. There is a collection of data objects and for that it can create hierarchical decomposition. The agglomerative method is a bottom-up approach in which it can merge the clusters as per their similarity, this process is repeated till the termination stage and at final it has singleton cluster. The cluster have some data objects, from that object top-down approach starts and split all the clusters into smaller cluster. Hence this process is known as divisive method which is the opposite of agglomerative method. The final output of this method is a single data object.

C. Density-based methods

This method encountered difficulty to describe an arbitrary shape cluster so it finds only, round or spherical shapes clusters.

D. K-Means Clustering

K-means algorithm is based on clustering methods and it is also the concept of data mining. The clustering means it can partition data and groups this data object. The collection of similar data objects, forms, group which is known as a cluster.

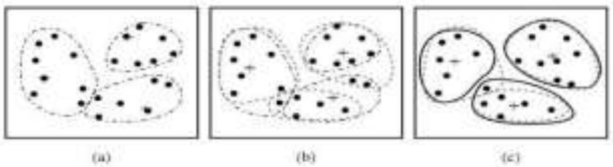

Fig 1: Example of K-means Algorithm [3].

In most of classification methods the resulting cluster may know before execution of the clustering algorithm, but K-means refer unsupervised learning method that means when the clustering algorithm is executed then and then the resulting cluster can be generated, before execution the result clusters are not known [3].

II. RELATED WORK

Olston et al. used batch crawling and incremental crawling [1]. In batch crawling, after the specific timestamp the crawling process can be paused for some time and resume as a way to obtain more recent snapshots of previously crawled pages. It does not contain duplicate copy of any page. Whereas in incremental crawling, Web pages may appear number of times in the crawl order and crawling is an uninterrupted process that conceptually never aborted. The incremental crawling is more powerful than the batch crawling so most modern crawlers perform incremental crawling and also incremental crawling allows re-visitation of pages at different rates [1].

F. Zhao et al. implemented the two stage framework smart crawler. As the deep web grows very quickly, hence the complicated task is getting appropriate information from the web. To solve such problems author implements smart crawler, this has two stage frameworks. The first stage is site locating which extract the relevant information from the web; it can rank the website as per related topic. The second stage is in-site exploring, it can prioritize the link. The purpose of implementation of author is used to get relevant information from Internet within the least amount of time [2].

Savita et al. studied the clustering methods for implementing indexing of search engines. The search engine indexing can provide information easily to the user because it can take data from web and store it in web repositories. Search engine indexing can collect the data after than partition data and store this data. The clustering, loading, processing and storing are four phases of the proposed method and output of each phase is working as input of another phase. That means the output of clustering is input of loading phase, the output of loading phase as input of processing phase and the output of the processing phase acts as an input of storing phase respectively. The last phase that is storing phase is responsible to provide results of user query to users [3].

Bh. Bangoria et al. described different clustering methods. The first is partition methods, in this method at the beginning, it can compute all present clusters at once and after that data object can be divided into non-overlapping clusters. In hierarchical methods, each object forming a separate group. The sphere-shaped clusters find in density-based method. In Grid-based methods, the grid structure is made by the space of object can be transform into a fix number of cells [4].

A. Anitha et al. described the web content mining and web data clustering. The extraction of information or data from the web is known as web mining. The agent based and database content mining is two types of web content mining. Author also studied hierarchical algorithm. The bottom-up and top down are two approaches of hierarchical algorithm. In the bottom-up approach at the starting it can treat all singleton clusters after that it can merge the several similar clusters and forms one complete cluster. In the top-down approach divides the one single cluster into multiple clusters[5].

G. J. Kim et al. proposed Incremental Clustering Crawler that means ICC. In this all web pages are collected by the crawler and after that clustering done on that collected web pages incrementally. It can maintain the quality of the cluster when it generated the final cluster because the fact that not all sites are present during the clustering. The incremental clustering crawler downloaded as many web pages and can be handled 50 or more than 50 pages at once. The cluster finder search clusters and chooses the web pages which are crawled by internal crawler. The four parts of the internal crawler are site manager, web page collector, link extractor and URL manager. The all information about each site is managed by site manager and it also checks all termination conditions of crawling. The web page collector stored or collected that are downloaded all web pages from giving URLs. Link extractor extracts the links which are present in downloaded web pages [6].

P. Dubey et al. proposed the K-means algorithm. K-means based on clustering schemes. Clustering can divide the dataset into a number of data objects. The data objects in one cluster can follow the same property that means it has similar in nature, but they all are different from data objects which presents in an another cluster. Each cluster in K-means has a one cluster representative also known as centroid. The initial means acts as an input in case of clustering and it produced a final means as an output. If the n clusters are present, then there is the presence of n initial and final means compulsory. The centroid of clusters is updated due to its iterative nature [7].

Y. Thakare et al. implemented the K-means clustering algorithm. Nowadays clustering analysis highly adopted in biology, psychology, software engineering and most widely used in computer science. Generally, many researchers can follow K-means clustering algorithm because of its simple nature and high quality result. K-means forms different clusters and the data items in one cluster should be similar as possible and the data items in other clusters should be dissimilar. The basic concepts of K-means are simple. At the starting it can form a K cluster and after that it can compute the centroid of each cluster [8].

U. R. Raval et. al. studied different clustering techniques and from these techniques she described the K-means clustering. In present the computer application may have large volume dataset so such data store in digital media. So such data stored in together to partition this data clustering is an effective method. Clustering divides data points in each subset of a dataset into the similar cluster. The K-means is an iterative process of assigning each data point to its nearest centroid. The purpose of this paper is it can compute initial centroid and assigning the data points to its nearest clusters. K-means clustering is a centroid based fundamental clustering technique which is widely used due to its simplicity and robustness. This technique required user defines parameters such as the number of cluster k, cluster initialization and cluster metric. First, it can initialize clusters to make groups of initial points from centroid inside the dataset, these subsets are known as a cluster. After that, the new centroid allocated to new data points which compute by the mean value for each cluster. This iterative process goes until centroid does not change [9].

Searching for hidden web resources is the major problem so it leads to develop a smart crawler which contains a two-stage framework. The candidate frontier extracts the links from these pages, and to prioritize them it can ranking link using link ranker. Site URLs, is the site's database contains a set of the entire site URL. Link ranker extracts the appropriate forms which are related to extracting URLs, hence link ranker responsible to improve performance of the adaptive link learner. The crawler can work in this fashion and its result improves by using clustering techniques which are K-means.

III. IMPLEMENTATION

A crawler is a tool, which helps to provide relevant information to the user. It can visit the web and downloading the bulk of pages which presents on the web. Smart crawler has a two phases, first is site locating and the second is in-site exploring as shown in figure 2 [2]. The first phase can search such sites which have data related to search topic and the prioritize assign to sites in second stage. Site locating stage is started from seed sites. Reverse searching searches the highly rank pages and site database takes these pages.
Site frontier collects all unvisited URLs which are aligned by site frontier and site classifier classify this site as per its relevancy. Link frontier stores such links which leads to other pages which present on web. The pages lead by links are fetch by page fetcher and similarly the form classifier categorizes forms as per relevancy and the result is searchable forms are stored in form database. The link ranker ranks the links.

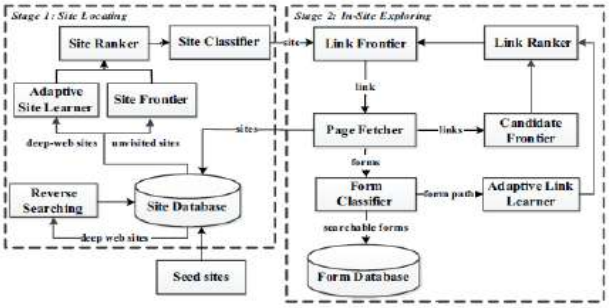

Fig 2: Architecture of Smart Crawler [2]

The URL in link ranker searches relevant forms in adaptive link leaner which improves the accuracy of link ranker. The appropriate site searched in site locating included with site collecting, site ranking, and site classification. In site collecting, it can visit the various links from visiting web pages, but it may not be enough for the site frontier. So for some parse domain the size of site frontier may reduce to zero. Hence the reverse searching and incremental two-level site prioritizing are two crawling strategies use to address this problem, to search more sites [2].

Site Locating
Site locating start from seed site and it fetch topic specific data. This stage includes crawling strategies which are reverse searching and incremental site prioritizing.

• *Reverse Searching Algorithm*

In reverse Searching, the page is relevant when two conditions are true which are : –If the topic related form are present in the page and also if the set threshold is lower than the size of seed sites then the page has relevancy.

Input for System: seed sites and harvested deep websites
Output from System: relevant sites

```
while number of candidates sites less than a threshold value do
    // pick a deep website
    site = get Deep Web Site (site Database, seed Sites)
    result Page = make reverse Search (site)
    links = extract Links from (result Page)
    for each link in links do
        page = download Page following (link)
        relevant = classify Respective (page)
        if relevant then
            Most relevant Sites = extract Unvisited Site (page)
            Output most relevant Sites
        end
    end
end
```

• *Incremental Site Prioritizing*

In incremental site prioritizing previous data can be initialized to the link ranker and website ranker. Then, website frontier stores all unvisited sites and website Ranker arranged all these sites, and optionally visited websites can be fetched site list.

Input: site Frontier
Output: searchable forms and out –of site links

```
Queue=Site Frontier. Create Queue (High Priority)
Queue =Site Frontier. Create Queue (Low Priority)
while site Frontier is not empty do
    if Queue is empty then
        H Queue. Add All(Queue)
        L Queue. clear ()
    end
    Site = H Queue. poll ()
    Relevant = classify Site (site)
    if relevant then
        Perform In Site Exploring (site)
        Output forms and Out Of Site Links
        Site Ranker. rank (Out Of Site Links)
        if forms is not empty then
            H Queue. add (Out Of Site Links)
end
```

A. In-site Exploring

To perform fast searching, site exploring the priorities assigns to links. It contains the link ranker and form classifier. The smart crawler gives more accurate and better results than other crawlers and it can try to provide more accuracy than other crawler. But when the sites are increasing for crawling, then the process becomes slow that means the crawling time required more. The search engine provides more relevant information to the crawler. The maintaining relevancy in the extraction of information from the web is might be problematic. Hence, to overcome this problem we used clustering methods. Cluster forms the set of similar data objects is formed. It can divide the dataset into a number of groups. Center all address text. If there is n number of addresses, then it uses n centered tab, and so on. We use the k-means clustering method which collects same data object in one cluster.

B. K-means Clustering

In 1967 the Mr. J. MacQueen introduced the K means clustering algorithm and after that in 1975 Mr. J. A. Hartigan and Mr.M. A. studied further in this algorithm deeply [3]. The cluster analysis can divide the data into k clusters which has a collection of data items in each cluster.

• *Algorithm:*

In K-means, if it wants to produce K-clusters then it takes K-initial means and k-final means. Hence the name of the algorithm is K-means [7]. The result is that it can produce k-final means. When the algorithm is terminated that time the every object in dataset present in one cluster. To find nearest mean to object, the all over means are searching by determining the cluster. This multiple clusters are formed in this algorithm; it can group the item present in the dataset. It is an iterative process, so after the each iteration it goes closer to final mean by updating new computed mean. The crawler improves its efficiency due to the use of K-means because K-means clustering is simple, vigorous and fast and also easily understands. When the datasets are separate from each other then the K-means gives best results. This algorithm also used in the medical field like tumor detection [10].

Input for System: Document Vectors DV.
Output from System: K Clusters.

1. The any set of K instances chooses as centers of the clusters.
2. Next, each instance assigns to the cluster, which is closest.
3. The cluster centroids are recalculated.
4. This process is iterated until there is not much change in the cluster centroids.

IV. EVALUATION

We evaluate the efficiency of our proposed solution over real web data in 8 representative domains. The goal of our evaluation is computing efficiency of smart crawler by getting relevant deep web sites and searchable forms. We used JAVA to developed smart crawler.

A. Experimental Setup

The smart crawler involves one of the components is form classifier which is trained by TEL-8 dataset. This TEL-8 dataset contains 8 representative domains, which form 3 groups "TEL"- means Travel group, Entertainment group and Living group [2].

Table 1: Eight domains for experiments [2]

Domain	Description
Airfares	Airfare Search
Automobiles	Used Cars Search
Books	Books Search
Car Rentals	Car Search
Hotels	Hotel Search
Jobs	Job Search
Movies	Movie Titles And Dvds Search
Music Records	Music Cds Search

The existing system is smart crawler which has two phases as first is site locating and the second is in-site exploring. The first phase fetches relevant sites and second phase prioritize these sites. But the drawback of the existing system is that Crawling large amount of data causes time consumption. To overcome these drawbacks we propose a smart crawler using K-means classifier. Our proposed solution provides faster processing and consumes less amount of time for large data. Also, it can provide more accurate results than existing systems. We compare crawler efficiency of common sites, SCDI, Smart crawler using naive bayes classifier and smart crawler using K-means classifier by fetching several pages from different domains. Fig 3 and Fig 4 shows the compare result of common sites, SCDI, Smart crawler using naïve bayes and smart crawler using K-means classifier. We compute the effectiveness graph and coverage graph. The fig 3 and fig 4 retrieved relevant searchable forms and deep websites. The effectiveness graph shows the result of a number of relevant forms from all available forms and the coverage graph shows result of total retrieves deep websites from total available deep websites.

Due to two stage framework smart crawler using Naïve Bayes gives better results than SCDI but smart crawler using K-means gives better result than all other methods because of the clustering method.

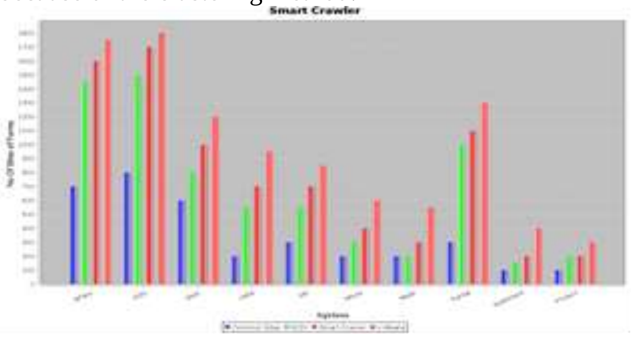

Fig 3: The numbers of relevant forms harvested by ACHE, SCDI, Smart Crawler and K-means.

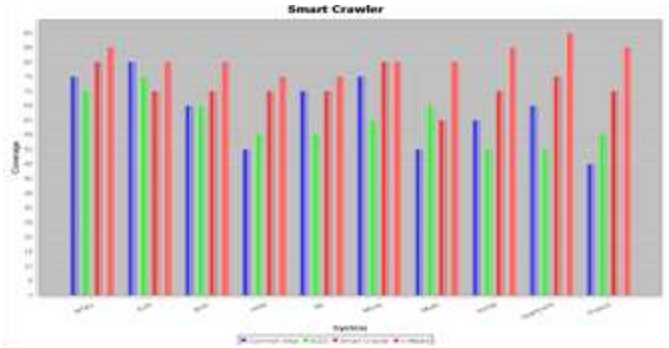

Fig 4: The numbers of relevant deep websites harvested by ACHE, SCDI, Smart Crawler and K-means.

CONCLUSIONS AND FUTURE WORK

The different types of data available on the Internet, so to extracts such data efficiency of the web we proposed smart crawler which have two stage frameworks that is first stage is site locating and the second is in-site exploring. The first phase can fetch relevant sites and the second phase is prioritizing the sites. The existing smart crawler implemented using naïve Bayes classifier, but we proposed it by using K-means classifier. Our proposed method is less time consuming and result also provided by smart crawler using K-means is more efficient. We can also improve results by combining pre-query and post-query approach in future.

REFERENCES

[1] Olston and M. Najork, "Web Crawling," Foundations and Trends in Information Retrieval, . 4, pp. 175- 246, 2010.

[2] F. Zhao, J. Zhou, C. Nie, and HaiJin, "SmartCrawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces," IEEE Transactions on Services Computing, vol. 99, pp. 1-14, 2015.

[3] Savita and Sachin Shrivastava, "Search Engine Indexing Using K-means Clustering Techniques," International Journal of Advance Research in Science and Engineering, vol. 5, pp. 218-227, 2016.

[4] Bh.Bangoria, N. Mankad and V. Pambhar, "Enhanced k-means clustring algorithm to reduce time complexity for numeric valuess," International Journal of Advance Engineering and Research Development, vol. 1, pp. 1-9, 2014.

[5] A. Anitha, "An Efficient Agglomerative Clustering Algorithm for Web Navigation Pattern Identification," In Scientific Research Publishing, vol. 7, pp. 2349-2356, 2016.

[6] G.H. Kim, Kyu-Young Whang and Min-Soo Kim, "Incremental Clustering Crawler for Community-Limited Search," Applications of Digital Information and Web Technologies, pp. 438-445, 2009.

[7] P. Dubey and A. Rajavat, "Implementation aspect of k-means algorithm for Improving performance" Proceedings of 28th IRF International Conference. vol. 10, pp. 96-102, 2015.

[8] Y. Thakare and S. Bagal, "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics," International Journal of Computer Application, vol. 110, pp. 12-16, 2015.

[9] Unnati R. Raval and Chaita Jani, "Implementing and Improvisation of K-means Clustering," International Journal of Computer Science and Mobile Computing, vol. 4, pp. 72-76, 2015.

[10] Sara Sandabad, Achraf Benba, Yassine Sayd Tahri and Ahmed Hammouch, "New method of tumor detection using K-means classifier and thresholding process," IJCSI International Journal of Computer Science, vol. 12, pp. 132-136, 2015.