Vol 8. No. 1 – March, 2015

African Journal of Computing & ICT

© 2015 Afr J Comp & ICT – All Rights Reserved - ISSN 2006-1781
www.ajocict.net

Modified Page Ranking System

L.N. Onyejegbu & O.D. Egbe
Department of Computer Science
University of Port Harcourt
Port Harcourt, River State, Nigeria
E-mail: nneka2k@yahoo.com; oladavidegbe@gmail.com.
Phone: +2348036748634, +2347036106989

ABSTRACT

The need to find information more efficiently on the World Wide Web has become the concern of researchers and developers over decades, and the most efficient way to get the best from the web is when the page ranking system can conveniently rank millions of pages in the shortest time possible. To achieve that the web is looked at, as a directed graph called webgraph with pages on the web corresponding to the nodes or vertices of a directed graph and the hyperlinks between pages represented as edges or arcs. In this graph, each node is a web page and each edge is directed, representing a link from one page to another using the HTML hyperlink notation. The page rank algorithm is what determines the quality of information retrieved to the user. In this paper, we developed an efficient rank algorithm that can retrieve quality information that meet the information need of the user, and the fuzzy c-means clustering was used to accelerate the processing time of the page rank algorithm while retrieving information. Object oriented methodology was adopted for the analysis and the implementation was done using Java programming language.

Keywords: Page rank, clustering, fuzzy c-means, webgraph, and crawler.

African Journal of Computing & ICT Reference Format:
L.N. Onyejegbu & O.D. Egbe (2015 Modified Page Ranking System. Afr J. of Comp & ICTs. Vol 8, No. 1. Pp 205-212. .

1. INTRODUCTION

The World-Wide Web has spawned a sharing and dissemination of information on an unprecedented scale. The existence of an abundance of dynamic and heterogeneous information on the Web has offered many new opportunities for users to advance their knowledge discovery. As the amount of information on the Web has increased substantially in the past decade, it is difficult for users to find information through a simple sequential inspection of web pages or recall previously accessed URLs. Consequently, the service from a ranking system becomes indispensable for users to navigate around the Web in an effective and more precise manner. But if the search engine is not designed properly with the ability to rank pages that are relevant to a particular search topic, the whole exercise becomes unfruitful and one end up wasting precious time chasing down useless URLs.

Search Engines essentially act as filters for the wealth of information available on the Internet. They allow users to quickly and easily find information that is of genuine interest or value to them, without the need to wade through numerous irrelevant web pages. There is a lot of filtering to do in 2004 the number of pages in Google's index exceeded the number of people of the planet, reaching the staggering figure of over 8 billion, and in 2013 Google says that the web now has 30 trillion unique individual pages.

That up an astonishing 30 times in five years: Google reported in 2008 that the web had just one trillion pages. Google says that it stores information about those 30 trillion pages in the Google Index, which is now at 100 million gigabytes. That's about a thousand terabytes, and you'd need over three million 32GB USB thumb drives to store all that data [1]. Since the Web is a huge repository of information that has been growing exponentially over the years. This rapid growing nature of the web has led to the invention of various techniques for managing the vast amount of content available online in order to realize its potential as a useful information resource [8].

With that much content out there, the Internet would be essentially unworkable without an efficient page ranking system, with Internet users drowning in sea of irrelevant information and shrill marketing messages. For a page ranking system to satisfy the need of user's trying to retrieve information, web mining techniques have to be employed by the search engines to extract relevant documents from the web database documents and provide the necessary and required information to the users. The search engines become very successful and popular if they use efficient ranking mechanisms.

© 2015 Afr J Comp & ICT – All Rights Reserved - ISSN 2006-1781

Page ranking algorithms are used by the search engines to present the search results by considering the relevance, importance and content score and web mining techniques to order them according to the user interest. Some ranking algorithms depend only on the link structure of the documents i.e. their popularity scores (web structure mining), whereas others look for the actual content in the documents (web content mining). In this study we will analysed few algorithms which uses link structure or web structure mining and few algorithms which uses web content mining for calculating the page rank value of webpages and proposed one algorithm which uses both web structure mining as well as web content mining for calculating the page rank value of webpages using the webgraph with the application fuzzy C-Means clustering to facilitate accelerated retrieval of the ranked information. This gives better and efficient results as compare to other and overcome some limitations of web structure mining based algorithms.

2. REVIEW OF RELATED WORK

The *World Wide Web* is the collection of information resources on the Internet that are using the Hypertext Transfer Protocol. It is a repository of many interlinked hypertext documents, accessed via the Internet. Web may contain text, images, video and other multimedia data. In order to analyse such data, some techniques called web mining techniques are used by various web applications. Web mining is the application of data mining techniques to discover patterns from the Web, it is the extraction of interesting and potentially useful patterns and implicit information from artefacts or activity related to the World Wide Web. [9].

There are three different knowledge discovery domains in web mining: Web Content Mining, Web Usage and Web Structure Mining. Most page ranking algorithms are mostly based on web structure mining they exploit only the link structure of the webpages in ranking.

2.1 Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a webpage was designed to convey to the users. Web content mining is related but is different from data mining and text mining. It is related to data mining because many web documents contains data and data mining techniques can be applied in web content mining [7]. It is related to text mining because much of the web content is text. Web data contents may involve the different types of data. These are: Text, Image, Audio, Video, Metadata and Hyperlinks.

2.2 Web Usage Mining

Web usage mining is the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of Web based applications. It involves the mining of web logs to discover access patterns of the pages accessed by the user [10]. Analysing regularities in web log records can help to identify potential customers for e-commerce, help in customization of web pages, improving server performance. Web server saves all entries of pages accessed in web logs.

2.3 Web structure mining

It is the process of retrieving the information from World Wide Web into more structured forms and indexing the information to retrieve it quickly [6]. Web structure mining, is a tool used to identify the relationship between Web pages linked by information or direct link connection. The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks,

Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.
Web structure mining can also have another direction -- discovering the structure of Web document itself. This type of structure mining can be used to reveal the structure (schema) of Web pages, this would be good for navigation purpose and make it possible to compare/integrate Web page schemes. This type of structure mining will facilitate introducing database techniques for accessing information in Web pages by providing a reference schema. [4].

2.4 Comparative Study of Page Ranking Systems

There are several algorithms proposed based on link analysis, these are algorithms that exploit the hyperlinks nature of the web in page ranking. The most important algorithms are PageRank, HITS (Hyper-link Induced Topic Search) and Weighted PageRank which is a content based algorithm, they are comparatively discussed below.

2.4.1 PageRank Algorithm

Page Rank algorithm is the most commonly used algorithm for ranking the various pages. Working of the Page Rank algorithm depends upon link structure of the web pages. The Page Rank algorithm is based on the concepts that if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages. The Page Rank considers the back link in deciding the rank score. If the addition of all the ranks of the back links is large then the page is provided a large rank. A simplified version of PageRank is given by:

$$PR(b) = d \sum_{a \in B(b)} \frac{PR(a)}{L(a)} \quad \text{..... 2.1}$$

Where the PageRank value for a web page b is dependent on the PageRank values for each web page v out of the set M_u (this set contains all pages linking to web page b), divided by the number L(a) of links from page a. [5].

© 2015 Afr J Comp & ICT – All Rights Reserved - ISSN 2006-1781

In a nutshell the assumption in Page and Brin's theory is that the most important pages on the Internet are the pages with the most links leading to them. PageRank thinks of links as votes, where a page linking to another page is casting a vote. This makes sense, because people *do* tend to link to relevant content, and pages with more links to them are usually better resources than pages that nobody links. PageRank doesn't stop there. It also looks at the importance of the page that contains the link. Pages with higher PageRank have more weight in "voting" with their links than pages with lower PageRank. It also looks at the number of links on the page casting the vote. Pages with more links have less weight.

This also makes a certain amount of sense. Pages that are important are probably better authorities in leading web surfers to better sources, and pages that have more links are likely to be less discriminating on where they're linking. The main advantage of the Google's PageRank measure is that it is independent of the query posed by user, this means that it can be pre computed and then used to optimize the layout of the inverted index structure accordingly. However, computing the Page-Rank requires implementing an iterative process on a massive graph corresponding to billions of Web pages and hyperlinks. The main disadvantage of this algorithm is that it favours older pages, because a new page, even a very good one, will not have many links unless it is part of an existing web site.

2.4.2 HITS Algorithm

HITS (**Hyperlink-Induced Topic Search HITS**; also known as **hubs and authorities)** is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It was a precursor to PageRank, it ranks the web page by processing in links and out links of the web pages. In this algorithm a web page is named as authority if the web page is pointed by many hyperlinks and a web page is named as HUB if the page point to various hyperlinks. An Illustration of HUB and authority are shown in figure1.

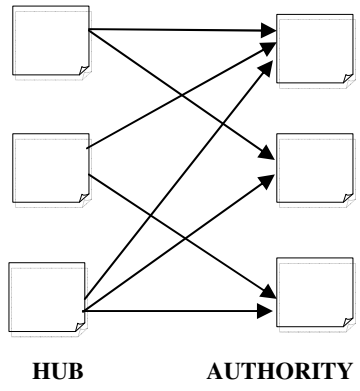

Figure 1: Description of Hit Algorithm

HITS is technically, a link based algorithm. In HITS algorithm, ranking of the web page is decided by analysing their textual contents against a given query. After collection of the web pages, the HITS algorithm concentrates on the websites and future work include to calculate the rank score by utilizing more than one level of reference page list and increasing the number of human user to classify the web pages (Kleinberg, 1998).

To begin the ranking $\forall p, auth(p) = 1$ and $\forall p, hub(p) = 1$ two types of updates are considered Authority Update Rule and Hub Update Rule. In order to calculate the hub and authority scores of each node, repeated iterations of the Authority Update Rule and the Hub Update Rule are applied. A k-step application of the Hub-Authority algorithm entails applying for k times first the Authority Update Rule and then the Hub Update Rule.

For authority update $\forall p,\ auth\ (p)$ is updated to be the summation

$$Auth\ (p) = \sum_{i=1}^{n} hub\ (i) \quad \text{.....2.2}$$

Where n is the total number of pages connected to p and i is a page connected to p. That is, the Authority score of a page is the sum of all the Hub scores of pages that point to it.

For, hub update $\forall p,\ hub\ (p)$ is the summation:

$$Hub\ (p) = \sum_{i=1}^{n} auth(i) \quad \text{.....2.3}$$

Where n is the total number of pages p connects to and i is a page which p connects to. Thus a page's Hub score is the sum of the Authority scores of all its linking page. HITS, like Page and Brin's PageRank, is an iterative algorithm based on the linkage of the documents on the web. However it does have some major differences:

It is query dependent, that is, the (Hubs and Authority) scores resulting from the link analysis are influenced by the search terms. As a corollary, it is executed at query time, not at indexing time, with the associated hit on performance that accompanies query-time processing. It is not commonly used by search engines. (Though a similar algorithm was said to be used by Teoma, which was acquired by Ask Jeeves/Ask.com).

It computes two scores per document, hub and authority, as opposed to a single score. It is processed on a small subset of relevant documents (a focused subgraph or base set), not all documents as was the case with PageRank.

The HITS algorithm has some obvious limitation which are discussed below: HITS is a purely link-based algorithm. It is used to rank pages that are retrieved from the Web, based on their textual contents to a given query. Once these pages have been assembled, the HITS algorithm ignores textual content and focuses itself on the structure of the Web only.

© 2015 Afr J Comp & ICT – All Rights Reserved - ISSN 2006-1781

HITS algorithm has some problems which are which are discussed here like, high rank value is given to some popular website that is not highly relevant to the given query. And drift of the topic occurs when the hub has multiple topics as equivalent weights are given to all of the out-links of a hub page.

2.4.3 Weighted PageRank (WPR)

Weighted Page Rank Algorithm is proposed by Wenpu Xing and Ali Ghorbani. Weighted page rank algorithm (WPR) is the modification of the original page rank algorithm. WPR decides the rank score based on the popularity of the pages by taking into consideration the importance of both the in-links and out-links of the pages. This algorithm provides high value of rank to the more popular pages and does not equally divide the rank of a page among its' out-link pages. Every out-link page is given a rank value based on its popularity. Popularity of a page is decided by observing its number of in-links and out-links. Simulation of WPR is done using the Website of Saint Thomas University and simulation results show that WPR algorithm finds larger number of relevant pages compared to standard page rank algorithm [2].

The more popular webpages are, the more linkages that other webpages tend to have to them or are linked to by them. It extended PageRank algorithm, this algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing link gets a value proportional to its importance i.e. each out-link page gets a value proportional to its popularity (its number of in-links and out-links). The popularity from the number of in-links and out-links is recorded as $W^{in}_{(u,v)}$ and $W^{(out)}_{(u,v)}$ respectively.

$W^{in}_{(u,v)}$ is the weight of link (v, u) calculated based on the number of in-links of page u and the number of in-links of all reference pages of page v.

$$W^{in}_{(u,v)} = \frac{I_u}{\sum_{p \in R(v)} I_p} \quad \text{.....2.4}$$

Where

- I_u and I_p represent the number of in-links of page u and page p, respectively. R (v) denotes the reference page list of page v.
- $W^{(out)}_{(u,v)}$ Is the weight of link (v,u) calculated based on the number of outlinks of page u and the number of out- links of all reference pages of page v.

$$W^{(out)}_{(u,v)} = \frac{O_u}{\sum_{p \in R(v)} I_p} \quad \text{.....2.5}$$

Where

- O_u and O_p represent the number of out-links of page u and page p, respectively. R(v) denotes the reference page list of page v.

The weighted page rank algorithm has some advantages and limitation which are discussed below: The Quality of pages returned by this algorithm is high as compared to PageRank algorithm. And It is more efficient than PageRank because the rank value of a page is divided among its out-link pages according to importance of that page.

As this algorithm considers only the link structure of the pages on the web not the content of the page, it returns less relevant pages to the user query.

3. DISADVANTAGES OF THE EXISTING SYSTEM

1. The algorithm relies mainly on the in-links of a page for ranking, without putting into consideration the out-links, content and usage of a webpage.
2. It takes a lot of time to calculate page rank for large URLs.
3. The algorithm of the existing system is link analysis based, which is prone to retrieve junks which results in false positive ranking.
4. A page irrelevant to the query still receives a high priority because of its many in-links.

3.1 The Proposed System

The proposed new system combines the features of PageRank algorithm with an improved algorithm that attached weight to pages to determine the most relevant pages, weight of web page is calculated on the basis of input and outgoing links and on the basis of weight, the importance of page is decided. The proposed algorithm is called content-link weighted page rank algorithm it is a hybrid page ranking algorithm that is aimed at overcoming the limitation inherent in the original page rank and weighted rank algorithm. It employ Web structure mining as well as Web Content mining technique to enable the users get the required relevant documents easily on the top few pages.

To make the rank algorithm scalable, to enable it handle large hyperlinks of webpages, fuzzy logic c-means clustering was applied to it since it can find a structure in a collection of unlabelled data by not matching the query exactly it improves the efficiency, effectiveness and speed of ranked pages retrieval.

© 2015 Afr J Comp & ICT – All Rights Reserved - ISSN 2006-1781

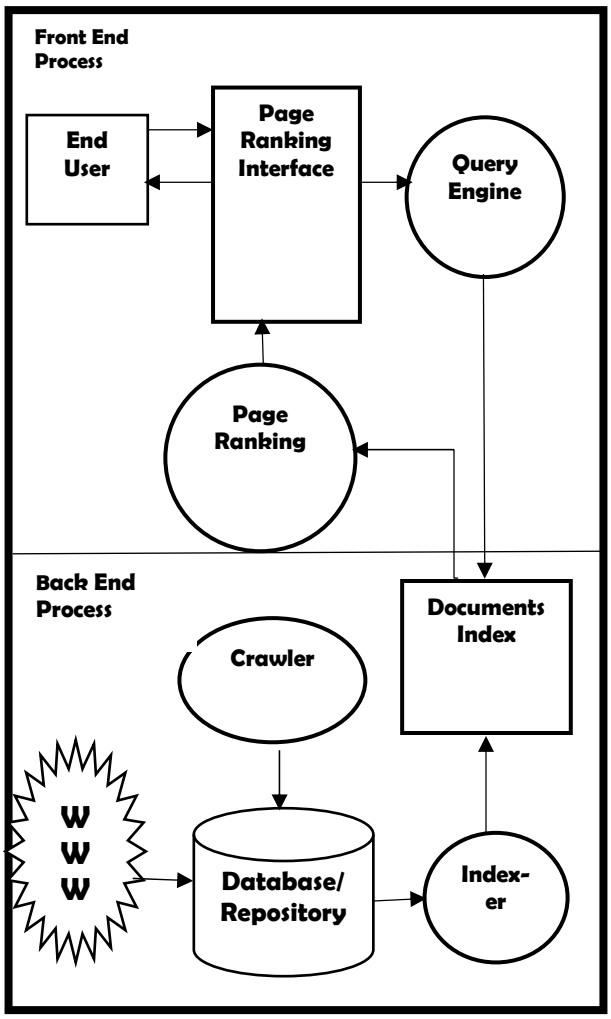

Figure 2 Architecture of the proposed system.

3.2 Content-Link Weighted Pagerank Algorithm for the Proposed System

This algorithm exploits the hybrid approach for calculating page rank value as it uses both web structure mining as well as web content mining. In this algorithm the importance and relevance of the webpages is calculated by taking into account weight of in links, weight of out links and number of visit to the link by users and by taking new parameter content weight of the web pages with respect to the query terms Wc

Consider equation $W^{in}_{(u,v)} = \frac{Iu}{\sum_{p \in R(v)} I_p}$ ---------- (3.1)

And equation $W^{(out)}_{(u,v)} = \frac{Ou}{\sum_{p \in R(v)} I_p}$ ---------- (3.2)

$W^{in}_{(u,v)}$= weight of the in-links of the pages

$W^{(out)}_{(u,v)}$=Weight of out links of the page.

$TL\ (v)$= Total number of visits of all links present on v.

Wc =Content weight of the web pages with respect to the query terms.

Step 1: Take the link structure of the retrieved webpages from the crawler.

Step 2: Obtain the web graph from the link structure of the retrieved webpages.

Step 3: Give initial page rank value to the all webpages as one.

Step 4: Using equation number (3.1) and (3.2), Calculate the weights of in links and out links and also calculate total no. of visits of all links by using client side script.

Step 5: Calculate the Content weight from the equation (3.5).

Step 6: Apply the proposed algorithm as in following equation

$$PR(u) = (1-d) + d * Wc \sum_{v \in B(u)} \frac{(V_u * .7 * W^{in}_{(u,v)} + .3 * W^{(out)}_{(u,v)}) PR(v)}{TL(v)}$$

…… (3.3)

Where,

- PR (u) and PR (v) are ranking of the Webpages u and v respectively.
- d is the dampening factor,
- V_u is the number of visits of link which points from **v** to u.
- $TL(v)$ is the total number of visits of all links present on v.
- $B(u)$ is the pages which points to webpage u.
- $W^{in}_{(u,v)}$ Is the weight of in links of connecting page *v* and *u.*
- $W^{(out)}_{(u,v)}$Is the weight of out links of connecting page *v* and *u.*
- Wc Is content weight of the web pages with respect to the query terms

Step 7: Iteratively repeat process until ranks of all Webpages are stable i.e. same in two consecutive iteration.

This algorithm reduce the problem of theme drift which is present on every link structure based algorithms as it uses the new parameter content weight from web content mining. Wc parameter takes user's query in to account and because of this the results retrieved are efficient and relevant as per user's query.

© 2015 Afr J Comp & ICT – All Rights Reserved - ISSN 2006-1781

3.3 Application of fuzzy logic C- means clustering

Fuzzy logic c- means clustering is used to accelerate the retrieval of ranked pages.

Step 1: Initially enter a key word to search.

Step 2: Apply the Content-link Weighted Rank Algorithm and calculate the rank. (Content-link Weighted Rank is a numerical value to represent the rank of a web page.)

Step 3: Input: Page P, Inlink and Outlink Weights of all back links of P, Query Q, d (damping factor)

Step 4: Output: Rank score

Step 5: Relevance calculation:
(a) Find all meaningful word string of Q (say N)
(b) Find whether the N strings are occurring in P or not? Z = Sum of frequencies of all N strings.
(c) S = Set of the maximum possible strings occurring in P.
(d) X = Sum of frequencies of strings in S.
(e) Content Weight (CW) = X/Z ….. (3.5)
(f) C = No. of query terms in P
(g) D = No. of all query terms of Q While ignoring stop words
(h) Probability Weight (P/W) = C/D ……. (3.6)

Step 6: If the Keyword has already calculate the rank by using the Content-link Weighted Page Rank in such case it simple use fuzzy algorithm formula to calculate cluster value.
The Cluster value is

$$K \approx \sqrt{\frac{n}{2}} \quad \text{…… (3.7)}$$

Where n is the number of data points
Based on the cluster value then that page will be looked for in that cluster and display on the screen.

Step 7: Apply FCM algorithm to cluster the scattered data. FCM is based on minimization of the objective function below.

$$F(u, v) = \sum_{i=1}^{N}\sum_{j=1}^{C}(\mu_{ij})^m \| x_i - v_j \|^2$$

…….. (3.8)

Let X = {x_1, x_2, x_3 ..., x_n} be the set of data points and V = {v_1, v_2, v_3 ..., v_c} be the set of centres.
$'||x_i - v_j||'$ is the Euclidean distance between i^{th} data and j^{th} cluster center
Initialize the Initial membership matrix U= $[\mu_{ij}]$, which is U(0)
(|i| = number of members, |j| = number of clusters)
Membership matrix, U, shows how much a document belongs to a cluster

Step 8: At k-step calculate centroids for each cluster using the equation below, the centres vectors V(k) = [vj] with U(k) if k ≠ 0. (If k=0, initial centroids location by random)
K is the iteration step. The membership values are calculated w.r.t the new centres. Belongingness of the document to the cluster is calculated using Euclidian distance between the centre and the data point.

$$V_j = \frac{\sum_{i=1}^{N}(\mu_{ij})^m x_i}{\sum_{i=1}^{N}(\mu_{ij})^m} \quad \text{…….. (3.9)}$$

Step 9: Calculate the degree of membership μij is the degree of membership of *xi* in the cluster *j*, *xi* is the *i*th of d-dimensional measured data, *Vj* is the dimension centre of the cluster.
Update U(k), U(k+1)

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - v_j\|^2}{\|x_i - v_j\|^2}\right)^{2/m-1}} \text{… (3.10)}$$

Step 10: If ||U(k+1)- U(k) ||< ε where, ε < 1 is the termination criterion. The usual choice of ε is 0.001 then STOP otherwise, return to step 2

© 2015 Afr J Comp & ICT – All Rights Reserved - ISSN 2006-1781

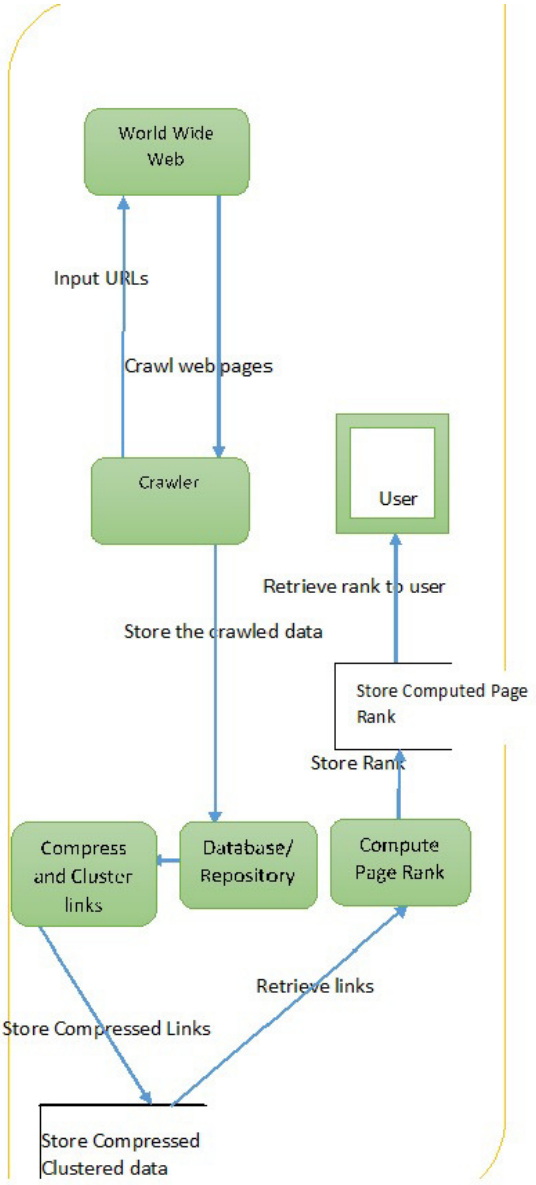

Figure 3: Data Flow Diagram of the Proposed System

4. RESULT ANALYSIS

The existing system took 3 hours to rank and retrieve result for 300Kb URLs. And as the number of hyperlinks increases the processing time increases exponentially, that is a considerable amount of time. But the modified page ranking system took just milliseconds to compute ranking for URLs and retrieve results.

Table 1.0: **Modified Page Rank Experimental Results**

S. No.	Number of URL Retrieved	Time taken in Milliseconds for the algorithm to calculate and retrieve rank
1	5	43
2	10	59
3	22	70
4	25	75

The time taken to retrieve results largely depend on the amount of system's resources available to process data like the CPU, RAM and disk space, this results may vary from time to time and from one system to the other.

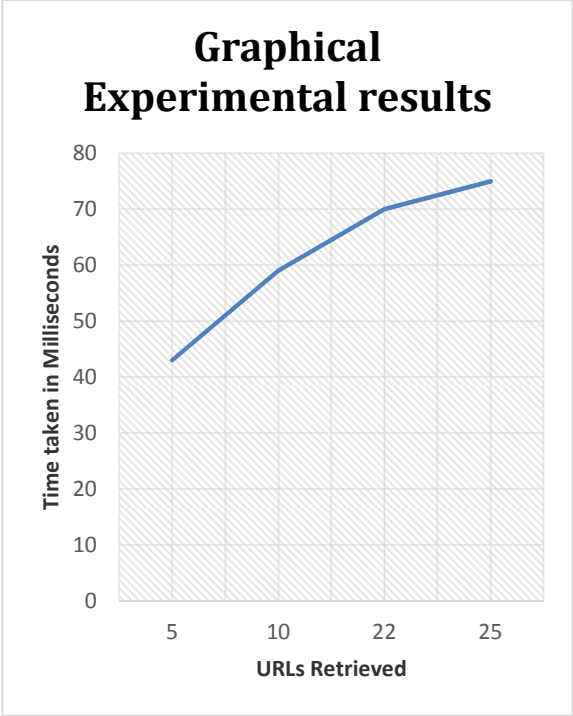

Figure 4. Experimental result in graph

The results above has proven that developing a page ranking system without a facility to speed up the rank algorithm processing time for large web links will make searching activity very boring even for large domain network, when it comes to the web where Google says that the web now has 30 trillion unique individual pages by 2013, ranking of pages will take hours to deliver query result to the user. Ranking algorithm no matter how efficient, is mostly designed to calculate the page rank that is why to facilitate fast retrieval of the calculated rank value, and enable the system maintain scalability fuzzy c-means clustering is used. In this the time taken to process the Web Pages is so small. As the number of web links is increased, the time is not increased. So, this helps to improve the processing time of the Web links.

© 2015 Afr J Comp & ICT – All Rights Reserved - ISSN 2006-1781

Fig 5: Sample output of the implementation

4.1 Implementation details

The system is implemented with an improved algorithm that takes into consideration both the in-links and the out-links of pages, and in addition it considers the page usage (number of visit to a page) and the content of a page to deliver quality query result.

5. CONCLUSION

Page-Rank computation time depends on the URL link structure and also the algorithm used for the computation, even though the page rank algorithm is highly efficient without the fuzzy c-means that retrieve the cluster that contain the required rank value calculated by the rank algorithm, information retrieval becomes time consuming. Fuzzy c-means clustering success in clustering the web documents to yield the best result in classification and fast retrieval is a milestone in this research work.

REFERENCE

[1] Donald, A. and Koetsier, J. (2014). Measuring the exponential growth of information on the internet today. *International Journal of Internet Computing (IJIC).* Vol. 4, Issue 7. Pp. 87-98.

[2] Ghorbani, A., Wenpu X. (2004). Weighted PageRank Algorithm. *Proceedings of the Second Annual Conference on Communication Networks and Services Research,* The computer society. Pp. 456-489.

[3] Kleinberg, J. M. (1998).Authoritative sources in a hyperlinked environment. *9th Proceedings of the ACM-SIAM Symposium on Discrete Algorithms.* Pp. 49-65

[4] Madria, S.and Rhowmich, S, and Lim F. (1999). Research issues in Web data mining. *In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference.* pp. 303-312.

[5] Page, L. Brin, S. Motwani, R. and Winograd, T. (1998). The PageRank Citation Ranking: *Bringing Order to the Web, Technical report, Stanford Digital Library Technologies Project*. Retrieved from http://www.cs.huji.ac.il/~csip/1998-66.pdf. 30/11/2014

[6] Rekha, J. and Purohit, G. N.(2011). Page Ranking Algorithms for Web Mining *International Journal of Computer Applications (IJCA).* Vol. 13, Issue 5. pp. 22-25.

[7] Raymond, K. and Hendrick, B. (2000). Web Mining Research a Survey. *Journal of Association for Computing Machinery (ACM).* Vol. 2, Issue 1. Pp. 1-15.

[8] Sonali, G., Manchanda, P. and Bhatia, K. (2012). The Automated Classification of Web Pages Using Artificial Neural Network. *Journal of Computer Engineering* (IOSRJCE), Vol. 4, Issue 1, pp. 20-25.

[9] Sonia, S. (2012). Web Mining. *International Journal of Emerging Technology and Advanced Engineering.* (ISSN 2250-2459) Vol. 2, Issue 4. Pp. 269-271.

[10] Verma, P. and Keswani, N. (2013). Web Usage Mining: Identification of Trends Followed by the user through Neural Network. *International Journal of Information and Computation Technology.* Vol. 3, No 7, pp. 617-624.