Web Page Change Detection Using Data Mining Techniques and Algorithms

J.Rubana Priyanga[1*],M.sc.,(M.Phil)
Department of computer science
D.N.G.P Arts and Science College.
Coimbatore, India.
*rubanapriyangacbe@gmail.com

R.Kousalya,MCA.,M.Phil.,(Ph.D)
Department of computer Applications
Head/Assistant professor
D.N.G.P Arts and Science College
Coimbatore, India.
kousalyacbe@gmail.com

Abstract—**This paper describes web page detection for structural change detection which have to provide direct access to information on the web page. A new technique has been provided for detecting changes in Web page. The technique is a new method to measure the similarity of two pages that represent the actual and the previous version of the detected page and it has been effectively used to discover changes in selected portion of an original web page. The newly proposed CMW technique and X-Diff algorithm can be used to detect the changes on selected web page and has been implemented in the CMW system to provide detection of web page change service for indexing purposes and keeping web pages up-to-date, later it has been used by search engine. Web-crawling which is used to arrange the unstructured data to structured data. The focused crawler mechanism that only scans the pages by using general crawling policies. The proposed algorithm and technique extracts the changes very efficiently from the various web pages and increase the detection accuracy and implement the CMW technique to reduce time and improve the speed**

Keywords— *Web page change detection, CMW technique, X-Diff Algorithm, detected page, web page up-to-date, web crawling.*

I. INTRODUCTION

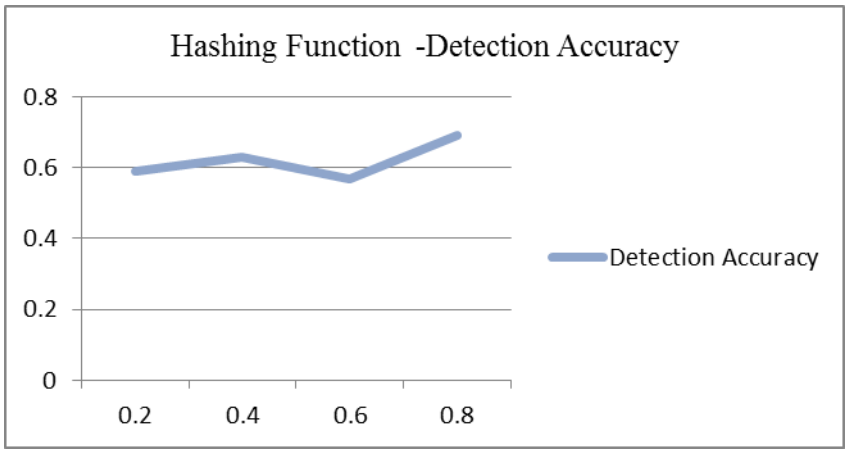
The users are interested in viewing the updated web page and newly gathered information on the web page each and every time they visit it. Changes are happening mostly in web pages as they are classified such as content changes, layout change, and attributes change. New pages are uploaded frequently to provide new and more information to the user. The paper also describes the architecture of a system, called CMW, which allows creating web update that changes the web page of interest and X-Diff algorithm is to be used to detect the changes. A web updated pages allow the CMW users to specify the type of changes they are interested in, and the actions that has to be performed when changes are detected. The proposed approach is to make efficient method to detect the web page changes.The types of changes which can occur in a web page and the architecture which can be used to detect the changes. A CMW technique is used to detect changes and increase the efficiency and accuracy.

II. EXISTING HASHING FUNCTION AND DETECTION ACCURACY

In existing approach the change detection are happen by two web page node having the same HTML tag type, and by hashing the web page in order to provide direct access to node information and assign hash value to each leaf node and tag value to the non leaf nodes. The performance is based on number of node similarity computations and the time consumed to complete the similarity coefficients.

Hashing

Alpha Values	Detection Accuracy
0.2	0.71
0.4	0.76
0.6	0.79
0.8	0.81



Proposed Approach:

To make efficient web page change detection using CMW technique. CMW system has been used to provide detection of web page change service for indexing purposes and keeping web pages up-to-date.X-Diff algorithm can be used to detect the changes on selected web page. Increase the detection accuracy, reduce time and improve the speed.

III. RELATED STUDY

1. *Document Tree based Approach*: Document Tree based approach is used to comparing the nodes of the trees. It gives the relevancy to the web pages and notifies the user about detecting the changesThe algorithm defines the good comparison study for the different algorithms and provides simple method for detecting changes.So it's difficult to compare signature for each and every node.
2. *Optimized Hungarian algorithm:* The operations of the Hungarian by running time and accuracy analysis.The algorithm focuses on finding the most similar subtree, finding out of order tags or unclosed tags, edit scripting to find minimum edge weight monitoring for bipartite graph.Three measures for detecting changes are also considered which are intersect (percentage of similar words),typedist (position of elements),attdist(relative weight of similar attributes).Limitation of this algorithm is that running time may be large.
3. *XML TREE DIFF:* The algorithm presents support for change control in the context of the Xyleme project that is investigating dynamic warehouses capable of storing massive volume of data. This algorithm is efficient in speed and memory space. It uses operations such as change node, delete node and insert node.
4.*CH-DIFF and CX-DIFF*: The system automates the change detection and timely notification of HTML/XML pages based on user specified changes of interest.CX-DIFF algorithm consists of steps like object extraction and signature computation, filtering of unique inserts/deletes and finding the common order subsequence between the leaf nodes of the given trees.
5. *Level Order Traversal:* Breadth first traversal includes document tree construction, document tree encoding and tree matching, for the detection of structural changes and content changes.
6.*Copernic Tracker:*The software can track changes in the text and images and monitors for the presence of specific text.The Web page and does not provide a utility for monitoring a specific region of the web page. The product does not reveal performance of speed or accuracy.
7. *A Website-Watcher:*The ability to monitor password protected web pages. The system offers limited for selecting a zone to monitor and lacks a proper user interface to the changes.
8.WYSIGOT*:* The application which is a commercial used to detects changes between HTML pages. The system has to be installed on the local machine and the granularity of change detection is at page level.

IV. METHODOLOGY

We describe the evaluation process of web page in the CMW system that allows users to specify and execute web update queries using a visual interface. Change detection results shows when a web page is raised. The system is implemented in java and implementation is accomplished with the software stack that includes Ubuntu Linux, Apache, MySql and PHP.The architecture of modules consists of change monitoring service, the query engine, the change detection module, and the change presentation module. The system has composed of two main applications such as, a visual query editor, that handles query specification, and an active query engine, that evaluates web page.The system maintains an object store, where the objects describing the currently updated web page are serialized. Each query object maintains information about the list of target zones,and for each target zone the list of targets contained inside that zone.The behaviour of CMW by explaining how each module behaves.

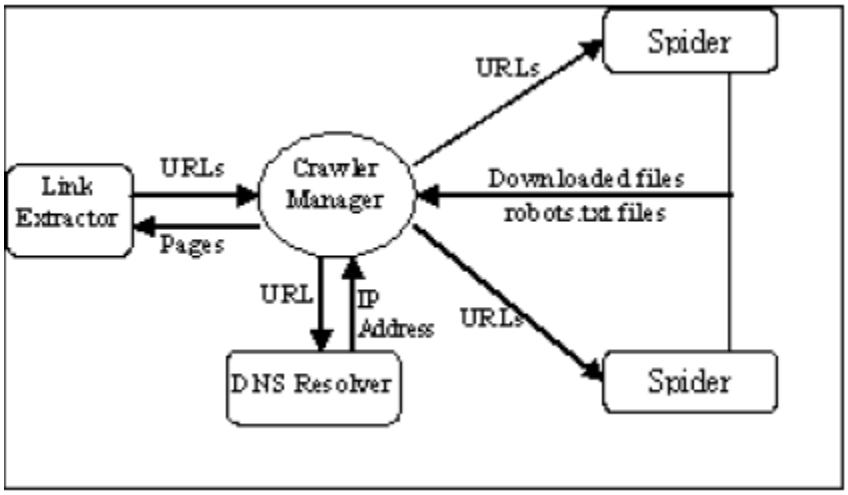
Focused Crawler: The basic idea of the crawler was to classify crawled pages with categories in topic taxonomy. To begin, the crawler requires topic taxonomy such as Yahoo. Through an interactive process, the user can correct the automatic classification, add new categories to the taxonomy and mark some of the categories as good (i.e., of interest to the user).The crawler uses the example URLs to build a Bayesian classifier that can find the probability (Pr(cjp)) that a crawled page p belongs to a category c in the taxonomy. The definition Pr(rjp) = 1 where r is the root category of the taxonomy.

$$\sum_{c \in \text{"good"}}^{n} \Pr(c \mid p)$$

A relevance score is associated with each crawled page. When the crawler as soft focused mode, it uses the relevance score of the crawled page to score the unvisited URLs extracted from it.The manner similar to the naive best URL crawler, it picks the best URL to crawl next. In the as hard focused mode, for a crawled page p, the classifier first finds the leaf node c* with maximum probability of including p. If any of the parents of c* are marked a good by the user, then URLs from the crawled page p are extracted and added to the frontier.

Crawling Policy : A Standard Crawlers policies to crawl the web pages,
1) A selection policy: that states which pages to download,
2) A re-visit policy: that states when to check for changes to the pages,
3) A politeness policy: that states how to avoid overloading Web sites, and
4) A parallelization policy: that states how to coordinate distributed web crawlers.

Architecture:



X Diff algorithm:

This algorithm can be used to detect the changes in the web pages. Edit script is used to convert one pages into another. Edit script is basically a sequence of edit operations .X diff algorithm uses three basic edit operations that are Insert, Delete and Update. X diff is used to find the webpage change detection based on web crawler.Various steps used in x diff are as follows:

1. Parsing and Hashing: It parses the two XML document into the corresponding tree structure and assign a hash value to each and every node in the tree.
2. Matching: This algorithm star with the root node and compare the hash values of the node in the two trees. If the hash value is same then the two trees are considered to be identical and if these values are different then minimum cost of matching is calculated for the two trees.
3. Generating minimum cost edit script: It generates minimum cost of edit script based on the minimum cost of matching which is being generated.

V. SIMILARITY MEASURE

To effective change detection is based on selected portion of a web page, we first have to retrieve this portion of the document in the new document version. Since the new version of the web page text can be added or removed before and after the portion of the document is selected. We have to find the portion of the new document that is the most similar to the old one. We have to define a similarity measure between document (sub) trees in a way that it should be possible to compute it efficiently, and the measure must be normalized, allowing the comparison of different pairs of trees and the selection of the most similar one. In order to define the similarity between sub-trees, we first associate each element of the selected sub-tree to its current version in the new sub-tree, and then consider the similarity degree of the two sub-trees. So, we first have to define a measure of similarity between single elements and then use it to define a similarity measure between whole (sub)trees. Given a document tree T =<N, p, r, l, t, a> and an element r1of N, the characteristic of r1 (wΨr1) is a triple h<type (r1, a(r'),w(r'). The similarity measure of two elements is defined on the basis of the similarity between each component of the characteristics of the elements. Given two trees T1 and T2, and two nodes r1 and T1 and r2 and T2 to define,

$$\text{intersect}\,(w(r_1), w(r_2)) = \frac{|w(r_1) \cap w(r_2)|}{|w(r_1) \cup w(r_2)|}$$

$$\text{attdist}\,(a(r_1), a(r_2)) = \frac{\sum_{a_i \in \{a(r_1) \cap a(r_2)\}} \text{Weight}(a_i)}{\sum_{a_i \in \{a(r_1) \cup a(r_2)\}} \text{Weight}(a_i)}$$

$$\text{typedist}(\text{type}\,(r_1), \text{type}\,(r_2)) = \frac{\prod_{i=0}^{\text{suf}} (2^{\max - i})}{\prod_{i=0}^{\max} (2^i)}$$

the similarity of r1 and r2 (CSr1, r2) is defined as,

$$\text{CS}\,(r_1, r_2) = -1 + 2 \times (\alpha * \text{typedist}(\text{type}\,(r_1), \text{type}\,(r_2)) + \beta * \text{attdist}\,(a(r_1), a(r_2)) + \gamma * \text{intersect}\,(w(r_1), w(r_2)))$$

where $\alpha + \beta + \gamma = 1$.

The value of α,β,γ can be selected on the basis of the type of changes to detect.Clearly the similarity coefficient takes values from the interval (-1,1 where)1 corresponds to the maximum difference and 1 to the

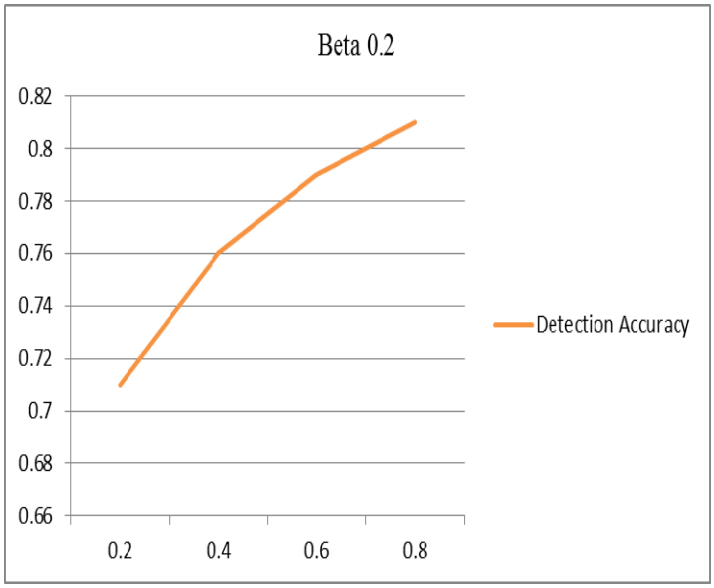
maximum similarity. An element that is deleted is assumed to have similarity -1/k with elements of the new document, where k is a fixed constant that defaults to 2.The efficiency of our technique since in general N'1 << N1 and N'1<< N2. The efficiency is confirmed by the execution times of the experimental results.

VI. EXPERIMENTAL RESULTS

The results of the experiments have performed to prove the effectiveness of our approach, what are the values of the parameters α,β,ɣ can be used in the definition of the element similarity formula.To assign effective values to them, and we performed some experiments to observe the variation of change detection accuracy depending on the values of these parameters. Once these parameters have been chosen in the best possible way, according to the tests performed, we test the real effectiveness of the approach by also performing other experiments to measure the execution time and the detection accuracy of the algorithm when applied to various categories of Web documents. The ZEDGRAPH Tool which used to draw the parameter result value of web page such accuracy and time. The results are presented in where the y-axis represents the alpha values of the page that has been changed, which was computed by differentiating the similarity.

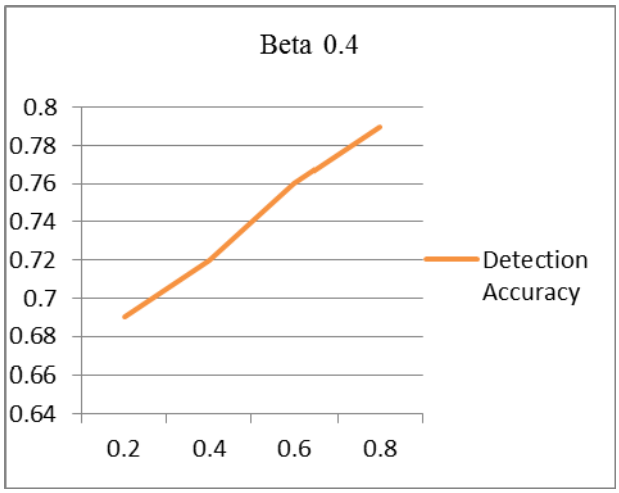
X- Diff:

Alpha Values	Detection Accuracy
0.2	0.71
0.4	0.76
0.6	0.79
0.8	0.81



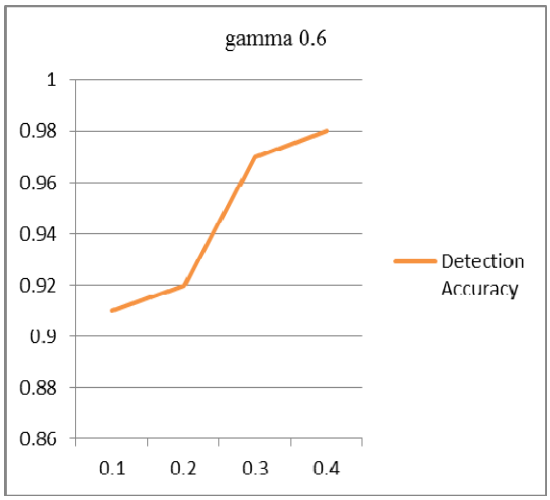
CWM Technique:

Alpha Values	Detection Accuracy
0.2	0.69
0.4	0.72
0.6	0.76
0.8	0.79



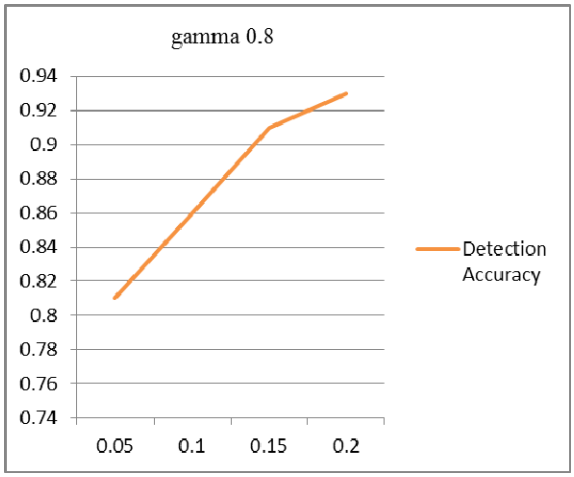
X-Diff:

Alpha Values	Detection Accuracy
0.1	0.91
0.2	0.92
0.3	0.97
0.4	0.98



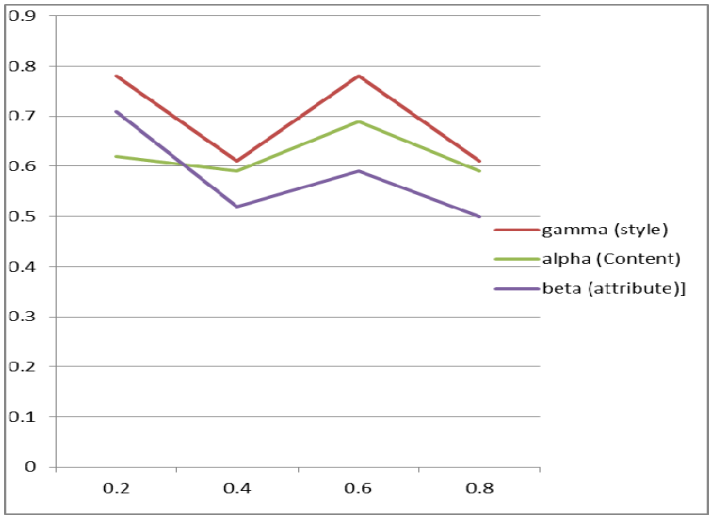
CWM Technique:

Alpha Values	Detection Accuracy
0.05	0.81
0.1	0.86
0.15	0.91
0.2	0.93



Comparison of parameters Table

Values	Time	Detection Accuracy
gamma (style)	0.8	0.87
alpha (Content)	0.6	0.78
beta (attribute)]	0.4	0.52



Coefficients.The graph reveals a linear trend, which indicates an exponential decay and corresponds to a Poisson distribution that resembles the output reported.

Tuning parameters: The element similarity function is essentially a weighted average of the three functions intersect, attdist, typedist, weighted by the coefficients α,β,ɣ. We performed experiments on a set of 10 Web pages, relating to auction bid and stock quotes.In particular, for each web page in the set we selected five modified versions and executed three different web update queries on it. For each triple of values α,β,ɣ we report the average accuracy obtained in the experiments, the accuracy level is an average of the results of the experiments, the successful experiments are weighted 1 and failing experiments are weighted 0.However, in this test set we want to be notified about the variation in bid prices and stock quotation, practically we are more interested in changes in the contents of the items than in structural changes.So,in this case the function that gives best results when a high weight is assigned to it is the intersect function that is weighted by the coefficient. In this case the information we are interested in is relative to structural (attribute) changes so we observe that attdist and typedist should have a greater value. Thus, the users are provided to selection of weights, depending on the type of the web pages.

Web page change result:

We present some statistics execution time of web updates and change detection. The dataset are countered as http://eBay.co.Uk. We performed experiments on different kind of web pages and we show the experiments on 10 test pages as reported and we made 100 tests on each of them. These pages are particularly interesting since they contain a high number of items rapidly changing. The average time interval which elapsed between the check request and the end of query evaluation, and the detection accuracy as the number of successful changes detected. The execution time does not include the download time of the document since tests are performed on local copies of the pages, pre downloaded from the Web site, so in this case our system is working like a personal web update server. In this figure we show the list of some items being monitored at the same time. We refer to this list as watch-list and it contains the information relative to the web pages that contain the item being monitored, the start date for monitoring and the last check time. When the time interval expires a statistic is reported, if relevant changes are detected then we refer these as major changes.

Time and accuracy of web page:

URL	PORTION MONITORING	TIME	ACCURACY
http://eBay.co.UK	DVD	3	90
http://eBay.co.UK	LAPTOPS	2	97
http://eBay.co.UK	GSM MOBILE	5	94
http://eBay.co.UK	AERONAUTICS	1	100
http://eBay.co.UK	COMMEMORATIVE	1	97
http://eBay.co.UK	SPORT	3	99
http://eBay.co.UK	FILM	4	99
http://eBay.co.UK	ROCK HARD	2	95
http://eBay.co.UK	CDs	2	99
http://eBay.co.UK	SCIENCE&NATURE	1	92

The experiments focused on the performance of the approach in terms of the number of node similarity computations and the time consumed to completely produce and store the similarity coefficients.

VII. CONCLUSION AND FUTURE WORK

A web change detection facility has become essential due to the fast rate of change of the information on the web.In this paper we have proposed a new technique which allows the efficient detection of Web page differences, in a quantitative way. Our technique, rather than being based on computing an edit script that produces the updated version of the whole document, focuses on the detection of changes in a specified portion of an web page.Using this technique has been possible to define a language that permits to express complex queries on web page changes.The CMW system has been developed to personal web update monitoring service. It is composed of a query editor that permits to specify web update queries in a fully visual and interactive way, and a query engine, that manages web query execution.Web page change detection system which will detect the changes in a web page. The proposed algorithm extracts change between different versions of web pages. This algorithm detects the changes, based on the change in the tag value. Future work will be devoted to the design and implementation of a Web based prototype of the system.

REFERENCES

[1] Hirschberg,"Algorithms for the longest common subsequence problem," Journal of the ACM, vol. 24, no. 4, pp. 664–675, 1997.
[2] WebCQProduct,http://www.cc.gatech.edu/projects/disl/ WebCQ.2006.
[3] L. Liu, C. Pu, and W. Tang,"WebCQ—Detecting and Delivering Information Changes on the Web,"Proc. Ninth Int'l Conf Information and Knowledge Management, pp. 512-519, 2000.
[4] CopernicTechnologies,CopernicTrackerProduct,2006
[5] ChangeDetect, http://changedetect.com/.
[6] M.AignesbergerWebSiteWatcherProduct,http://www.aignes.com, 2006.
[7] Wysigot, http://www.wysigot.com.
[8] D.Yadav,A.K. Sharma, J.P. Gupta"Change Detection In Web page" in a proceeding of 10th international conference on information technology.
[9] Heydon and .Najork.Mercator: "A scalable, extensible Web crawler",Worldwide web2 (4):219-229,1999.
[10] G. Srishti, Rinkle R,"An efficient for web page change detection" , IJCA, VOL. 48, NO.10, June.
[11] S.Chawathe,A.Rajaraman,H.Garcia-Molina,J.Widom,,Change detection in hierarchically structured information, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, June 1996, pp. 493–504.
[12] H.W.Kuhn,The Hungarian method for the assignment problem, Naval Research Logistics Quarterly 2 (1955) 83–97.
[13] L. Liu, C. Pu, W. Tang, J. Biggs, D. Buttler, W. Han, P. Benninghoff, Fenghua,CQ:a personalized update monitoring toolkit, in: Proceedings of theACMSIGMOD International Conference on Management of Data, 1998.
[14] L.Liu,C.Pu,W.Tang1999,Continual queries for internet-scale event-driven information delivery, in: IEEE Transaction on Knowledge and Data Engineering, 1999 (Special issue on web technology).
[15] L.Liu,C.Pu,W.Tang,WebCQ—Detectinganddelivering information changes on the web, in: Proceedings of CIKM_00, Washington, DC USA, 2000.

[16] D.Fetterly, M. Manasse, M. Najork ,A large-scale study of the evolution of web pages, Journal of Software – Practice and Experience 34 (2) (2004) 213–237.
[17] S.Flesca, E. Masciari2003,Efficient and effective web change detection, Data and Knowledge Engineering 46 (2) (2003) 203–224.
[18] Monica Peshave, Kamyar Dezhgosha],"How Search Engines Work and a Web Crawler WS Application". Department of Computer Science, University of Illinois, Springfield USA
[19] Junghoo Cho, Hector Garcia-Molina, and Lawrence**, "**Efficient crawling through URL ordering Page", In Proceedings of the 7th World-Wide Web Conference, page(s):161-171.
[20] E.Co.man, Jr., Z. Liu, and R.Weber,Optimal robot scheduling for web search engines", Proceedings of the 11th international conference on World Wide Web WWW '02 Honolulu, Hawaii, USA.ACM Press. Page(s): 136 – 147.

R.Kousalya M.C.A M.Phil (Ph.D),Assistant Professor, Head Department of Computer Application,
Dr.N.G.P Arts and Science College Coimbatore, India.
E-mail id: kousalyacbe@gmail.com

J.Rubana Priyanga M.phil Research Sholar, Dr.N.G.P Arts and Science College Coimbatore, India.
E-mail id: rubanapriyangacbe@gmail.com.