ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 3, March 2013

Mobile Information Retrieval using Topic-Sensitive PageRank and Page Freshness

Suresh P[1], Jeril Kuriakose[2]

Department of Computer Science, Salem Sowdeswari College - Govt. Aided, Salem, India[1]

Department of Computer Science, The Kavery College of Engineering, Salem, India[2]

Abstract: Mobile users are on a rise in the past five years. Using of internet can be done anywhere and anytime because of mobile phone. PageRank Algorithm is used for ranking or sorting large number of hypertext documents. It can be used to tote up abundant sub graphs of the web with high efficiency. When the PageRank algorithm uses the large link arrangement of the web to retrieve the similarity of the required webpages, only a single vector comes as the outcome of the search query results. To overcome this we can use a group of PageRank vectors favoring a typical topic or subject matter. Using Topic sensitive PageRank a bit more accuracy can be obtained for the search-query. If the search is based on any catchword or keyword, the topic of the catchword or keyword is used by the topic sensitive PageRank to retrieve information accordingly. If the search is based on context, the PageRank scores of the topic can be figured out using the subject of the context. By grouping the above two based on linear combination, a more accurate ranking can be created, which will be better than a simple, basic PageRank vector. In our crawler we used Page Freshness and Age that retrieved the latest altered web pages compared to the old ones found in the remote date sources i.e., websites. This paper labels the techniques or methods for effectively retrieving information from web using mobile.

Keywords: Mobile computing, web search, information extraction, PageRank, page freshness, page age.

I. INTRODUCTION

More opportunity has risen because of the decline in mobile internet usage price. This term paper discuss on retrieving information through and for mobile devices. Information retrieval is comparable to retrieving information from mobile, whereas mobile information retrieval needs to encounter new frontiers as the mobile platform and technologies. They are not up to the mark of laptop platform, technologies and desktop platform, technologies. Retrieving information through mobile phone is done by a search interface. Progressing and measurement of retrieved information is done by a back-end server.

Access of landline has been low, usage of mobile phones are on a rise during the past five years [1]. People start using their mobile phones for searching, surfing, shopping and even to gather information. This is because of the smaller size, easy to use and carry, lower price and availability of network access everywhere. As web browsing using mobile are on a hike because of low internet rate, it's still a contest. The World Wide Web are used and designed for personal computer users and laptop users, not for mobile users. Carrying out the Information retrieval process through and for the mobile device is challenging, because methods and platforms used by personal computers and laptops can't be used by mobile devices for retrieving information. Fig 1 shows the rise and expected rise of mobile internet users verses the ordinary internet users.

Fig 2 shows the expected growth of people going to use mobile for searching the web. Fig 3 shows the growth of mobile users who use their mobile for reading news and searching information in web. According to [4] the total number

of mobile internet users will get past the PC (Personal Computer) based internet users. People will start preferring mobile for PC.

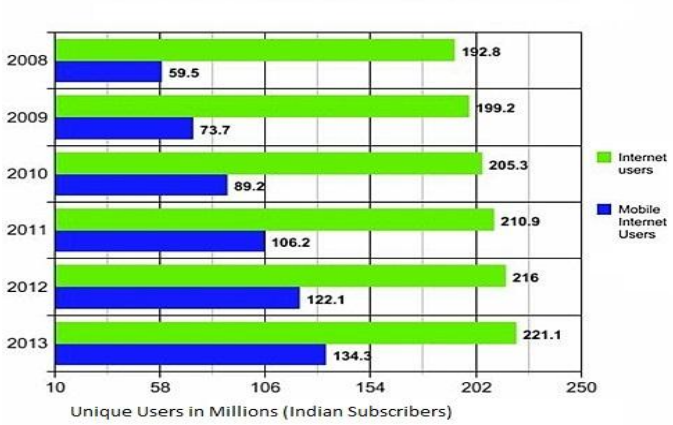

Fig. 1 Rise of Mobile Internet Users

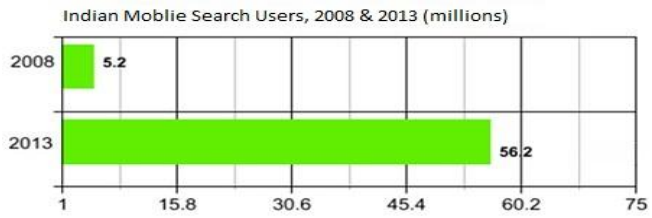

Fig. 2 Mobile Search Users

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 3, March 2013

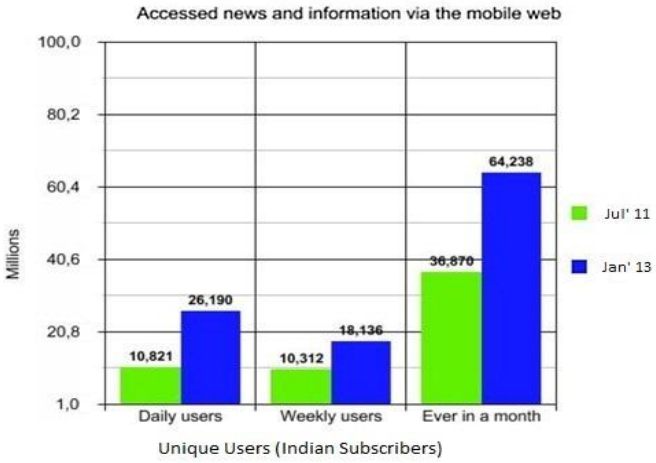

Fig. 3 People using Mobile Web

Some complications like network coverage in some areas, tiny screen size when compared to laptops and desktop and crouched processor speed leads to a new or different content change and limited platform base for application development in mobile devices.

II. APPROACHES FOR MOBILE INFORMATION RETRIEVAL

The popular action that a mobile user does is to search for information or content from some website. This gives rise to concretize the search query based on context, users profile, previous usage patterns and behavioral patterns of the user. Search engine retrieve information based on two chief activities called context awareness and content adaptation.

Reducing the information based on the user's mobile platform can be done using context awareness. This is because of the different mobile platforms like apple, windows, android etc., used by mobile users. Fig 4 shows the hierarchical representation of the manner in which content adaptation and content delivery techniques are carried out. Content adaptation center of attraction is to change the obtained results of the users query into the mobile device configuration.

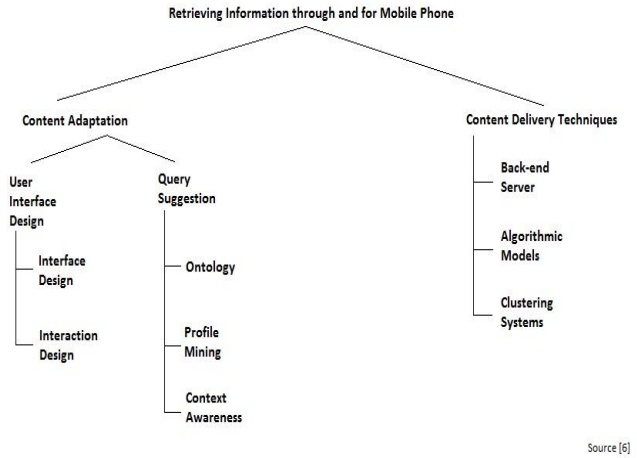

Fig. 4 Hierarchical Representation of Mobile Information Retrieval

Content adaptation takes care of user interface design and query suggestion. The interface design is responsible for the screen size and resolution, input type etc., whereas the interaction design is responsible for type of input given like total number of keywords given. Query suggestion is carried out using ontologies, profile mining and context awareness. On other hand content delivery techniques point out to original information retrieval procedures that can be used for mobile platform.

a) Content Adaptation

Content adaptation deals with the set of problems that takes care of the specific features in the mobile field like screen size and resolution, processor used, bandwidth etc.,

Content adaptation approach has two sub domains like User Interface Design and Query Suggestion.

b) Query Suggestion

Query suggestion deals with identifying what the mobile user require. It will be effectual for the system to confine the user's input before forwarding it to the search engine. The query suggestion can come from the database, ontology, previous user pattern or past history and from other users.

The fundamental for query processing is to provide indexing which makes query suggestion easier. Designing of the ontology must be done cautiously, because it is necessary to make the domain knowledge explicit. Preprocessing of the user query can be done easily with the help of better ontology structure.

c) Mobile User Interface

The mobile user interface must be developed based on the search engine or the information retrieval application like meta-search engine etc. The whole retrieved information of the users query might not fit the small mobile screen, so categorization like tree-form, result window etc., can be done to minimize the small screen problem.

d) Content Delivery Techniques

The content delivery technique mainly focuses on delivering the best content to the users. This can be made possible by the approaches like using a back end server, having good algorithmic models for processing the results and better clustering systems for better results.

The main task of back end server is to redirect the request from the mobile user to its server and act as intercede between the mobile and the web search engine. The back end server then processes the requests and sends it to the web agent or the web search engine. This is done because the algorithmic operations cannot be done in the mobile domain due to its size and processing speed. Fig 5 explains the processes that come to pass during retrieving information from and through mobile devices.

Large numbers of algorithms are used in information retrieval. Most of the algorithms cannot be applied for mobile devices. Few algorithms like Real Time Query Expansion system [9] can be used by mobile users. It delivers appropriate suggestions to the mobile users if they are using the custom build user interface.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 3, March 2013

The step by step procedure of retrieving information through mobile using a back-end server is as follows;

1. When the mobile user wants to retrieve information, he first opens his application that has a search interface. The request typed by the user is directly routed to a back-end server through internet.
2. The users request reaches the back-end server. The back-end server is responsible for processing the request.
3. The back-end server searches the web (using crawlers, spiders, etc.,) for the relevant information.
4. The relevant information is retrieved (by crawlers, spiders, etc.,) from the web the web servers.
5. The retrieved information reaches the back-end server through internet.
6. The retrieved information is compressed or processed such that it fits the mobile users screen resolution, processing speed etc., the compressed information is then send to the mobile device through internet.
7. The user receives the retrieved information that supports his mobile device.

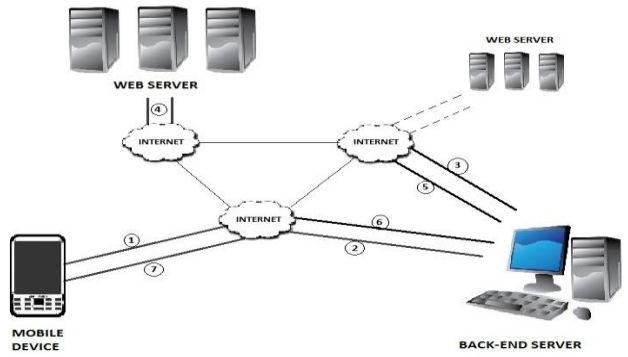

Fig. 5 Step by Step Method of Retrieving Information

Clustering the systems is done to obtain better results. A system called Mobile Findex [11] provide an efficient mobile user interface device. This system presents the result set or cluster to the mobile users by clustering content and user interface.

e) Search Interface

For every mobile device there need to be a search device to perform information retrieval. Mobile's user interface is connected to the back-end server by web services.

In CARSA Retrieval system [11], it submits its query to the Meta Searcher. Meta Searcher users User Profile, Intelligent Bookmark and Sense folder classification [5] to process the mobile users query. Ontology facilitated Interoperation [8] is carried out by the Meta Searcher to strengthen user mobile experience.

f) Grouping of the documents

Grouping of the information retrieved from the documents can improvise the performance of retrieval operations for the user. The grouping of the documents can improve the search speed by 50%.

III. RETRIEVING OR EXTRACTING THE INFORMATION FROM THE WEB PAGES

There are many papers discussing how to extract or retrieve information from web pages. Document Object Modal based Content extraction from html pages or documents [13] focus on removing medley and arranging the content into a more readable format like changing the font size and font type or removing the html and data factors like pictures, advertisement etc., so that the content can be made available to mobile screen size.

Another paper called Extracting Structured Data from Web Pages [14] used an algorithm for extracting the information based on the set of words that have similar event arrangement in the given input, and it uses that to extract the information and develop a template.

Preliminary Discussion

In our paper we used back-end-server as computer and tested using that with a mobile phone. Both mobile phone and computer were tested individually to find the difference.

a) Cosine similarity

Cosine similarity is widely used to find the similarity between two documents. Cosine similarity is a similarity measure; we can convert the similarity measure to a distance measure. Thus by doing it we can use it in any distance based classifier like nearest neighbor classifier. The cosine similarity is used to select the topic and is also used by crawler to search for links that does not exceed the minimum similarity score.
The cosine similarity between a topic and page is as follows [16]

$$\mathbf{sim(q,p)} = \frac{\sum_{k \in q \cap p} f_{kq} \quad f_{kp}}{\sqrt{\sum_{k \in p} f_{kp}^2 \quad \sum_{k \in q} f_{kq}^2}}$$

where
q is the topic given,
p is the obtained page,
f_{kp} is the frequency of term k in p.

b) PageRank algorithm

The concept of PageRank algorithm is as follows; consider a webpage or page 'u' that has a link to another webpage or page 'v'. From this we can come to a conclusion that page 'v' has some importance from page 'u'. If many pages link to some particular webpage or page like Wikipedia, it's probably understood that it has more importance.
The PageRank is propagated as follows [17]

$$\forall_v \, Rank^{(i+1)}(v) = \sum_{u \in B_v} Rank^{(i)}(u)/N_u$$

where
N_u is the out link of u,
Rank represents the importance of the page,

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 3, March 2013

B_v is the set of pages that has link to v.

The PageRank of any particular page can also be computed as [18]

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

where
A is a webpage,
PR(A) is PageRank of a Website,
C(A) is the number of like going out of the webpage A,
d is a damping factor between 0 and 1,
$T_1 \ldots T_n$ are the links that point to the webpage A (i.e., citations).

The pseudo code of a PageRank crawler is as follows [19]

```
PageRank (topic, starting_urls, frequency)
{
        foreach link (starting_urls)
        {
                enqueue(frontier, link);
        }
        while (visited < MAX_PAGES)
        {
                if (multiplies(visited, frequency))
                {
                        recompute_scores_PR;
                }
                link := dequeue_top_link(frontier);
                doc := fetch(link);
                score_sim := sim(topic, doc);
                enqueue(buffered_pages, doc, score_sim);
                if (#buffered_pages >= MAX_BUFFER)
                {
                        dequeue_bottom_links(buffered_p
                ages);
                }
                merge(frontier, extract_links(doc),
        score_PR);
                if (#frontier > MAX_BUFFER)
                {
                        dequeue_bottom_links(frontier);
                }
        }
}
```

where
sim(topic, doc) acknowledge the cosine similarity among the documents and topic,
MAX_BUFFER can contain any specific number of pages.

c) Page Freshness and Age

From the pages retrieved through PageRank algorithm, we check for freshness of each page. This makes sure that each page is up-to-date and has a chock-full of latest information. A web page can be said a fresh page only when the page is found in both local and remote sources. Only information retrieved from a fresh page is sent to the mobile device. The freshness and age calculated by Cho and Gracia [20] is given below

The freshness of a local element is as follows

$$F(e_i; t) = \begin{cases} 1 & if\ e_i\ is\ up-to-date\ at\ time\ t \\ 0 & otherwise \end{cases}$$

In the above equation e_i is the element of the database at time t.

The freshness of the database is

$$F(S;\ t) = \frac{1}{N} \sum_{i=1}^{N} F(e_i;\ t)$$

where S is the database at time t,
N is the number of elements retained up-to-date.

The age of a database is needed to find out how obsolete the database is. The age of a local element is as follows

$$A(e_i; t) = \begin{cases} 0 & if\ e_i\ is\ up-to-date\ at\ time\ t \\ t - t_m(e_i) & otherwise \end{cases}$$

where

t_m is the time of first modification of the local element e_i.

The age of the database is

$$A(S;\ t) = \frac{1}{N} \sum_{i=1}^{N} F(e_i;\ t)$$

where S is the database at time t,
N is the number of elements retained up-to-date.

IV. RESULT

Various tests have been conducted using mobile and computer (back-end-server) for retrieving information from 10 websites. For computer we used WIRE (Web Information Retrieval Environment) Crawler [21] because it is coded using C/C++, scalable, easily configurable and it also encompass several tools for analyzing and generating reports. The documentation and source code can be downloaded from http://www.cwr.cl/projects/WIRE/. The computer used is 2GHz processor with 2GB RAM. The network bandwidth for each device is 512 kbps. For mobile we developed a small Android application to retrieve and analyze the data obtained. The android application was successfully installed and tested in Samsung Ace plus S 7500. The reason for choosing Samsung Ace plus S 7500 was because it had 1GHz processor with 512 MB RAM and screen size 320 x 480 pixels. While crawling, in-order to achieve politeness we allow 10 seconds between accesses of each websites. The maximum exploration

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 3, March 2013

depth is 15 for each website. These are common for both computer and mobile.

Device Used	Time Taken (seconds)
Computer	201
Mobile	395

For testing purpose we used 10 Indian college websites to crawl and find out information regarding their academic departments. We used PageRank algorithm to retrieve top 20 Indian college websites frequently visited from a list of 100 college websites. Later we still filtered it to 10 colleges based on their freshness and age. The time taken to do the above is as follows,

Device Used	Time Taken (seconds)
Computer	86
Mobile	174

TABLE 1 Computer vs. Mobile Crawl Timings

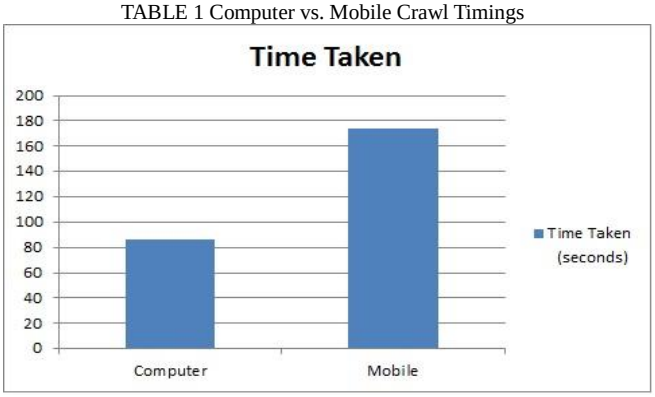

Fig 6 Computer vs. Mobile Crawl Timings

We used WIRE crawler and our android application to crawl from the above obtained list of colleges. The time taken for the crawler and android application is as follows,

Device Used	Time Taken (seconds)
Computer	115
Mobile	221

TABLE 2 Computer vs. Mobile Crawl Timings

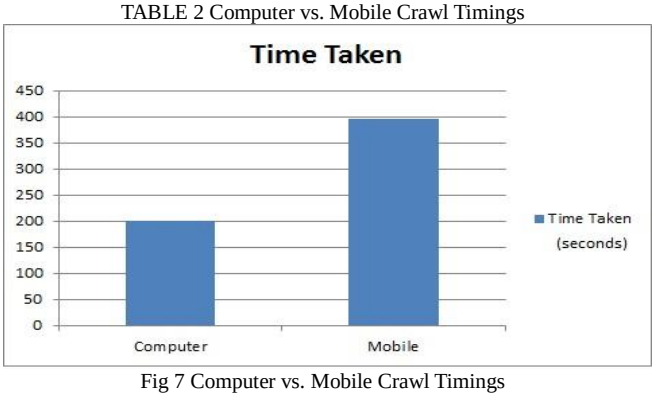

Fig 7 Computer vs. Mobile Crawl Timings

Hence the total time required for the complete process is found out by adding up the above two time taken,

TABLE 3 Total Time Taken

.

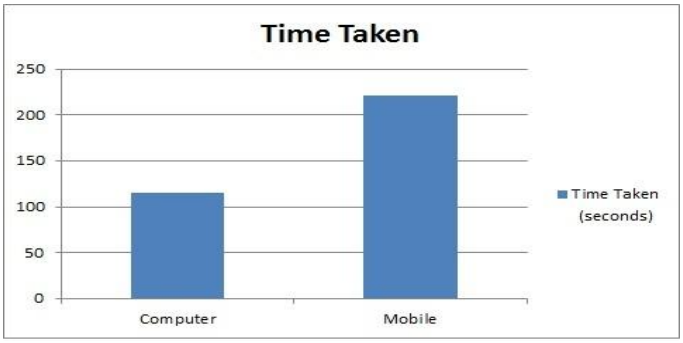

Fig 8 Total Time Taken

V. DISCUSSION

a) Other Mobile Intelligent Agent

Hewlett-Packard's SiteOnMobile [2] to create specific kind of task called "tasksets" which extract the relevant data from web site and make it available to the mobile users. This can be used by people who want to get specific set of information. Apple iPhone uses a personal assistant application called Siri [3] to send message, to make appointments, call a person, to reserve a ticket or to hire a cab and many more just by the voice instruction in natural language. There are still many intelligent agents going to mount in the market as the mobile usage is going on massive up rise.

b) Pros

Reduced search: Having information retrieved through and for mobile devices makes the search easy. Suppose we want buy a digital camera we need to do a lot of search like opening various websites, checking the price, selecting our desired camera with specific requirements like resolution, lens, pixel etc., this will take a pretty much time. But having all our information retrieved through mobile can make our work and search simpler.

Anywhere anytime use: We can use the mobile anywhere and anytime, no need to go and sit in front of a personal computer or a laptop. Consider a scenario where there is no laptop or a personal computer like museum or a movie hall and we want to book a holiday travel pack. We can do that manually but the thrill of the movie might go, so just tell the mobile agent to search for travel destination and hotels in the current place and continue watching the movie.

c) Cons

User privacy: In some cases the mobile users are required to give their passwords, bank account details etc., which affects the privacy of the mobile user.

User security: There can also arise some security issues like identity theft etc., some fraud mobile agents can also send our internet history, personal details etc., to unknown people or they can use it for other malicious purpose.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 3, March 2013

User acceptance: Because of the above problems it is a challenging task to make mobile users welcome the information retrieval through and for mobile.

Network and Mobile processor speed: Network speed and processor speed act as the barrier in retrieving information through and for mobile device. In some cases the person might have a better processor and might not have a better internetwork and vice versa.

VI. CONCLUSION

A small technique for retrieving information through mobile is discussed in this paper. We analyzed how pages are retrieved using Topic-Sensitive PageRank Algorithm. Combing it with Page Freshness and Age gave us a more efficient and well-run results. Our method combines Topic-Sensitive PageRank with Page Freshness and Age, which can be found useful in retrieving updated information using a crawler. Best results can be obtained from the mixture of Topic-Sensitive and Page Freshness. An experimental proof done by us also implies that it can retrieve relevant information. Future works can also be done using other search algorithms and combing it Page Freshness and Age. Comparative study can made between search algorithms combined with Page Freshness and Age.

Retrieving information through and for mobile devices is an emerging field as the mobile users are on aggrandizement. As the security increases people will normally start retrieve information and through and for mobile. In this field there are plenty of adventitious channel for research.

REFERENCE

[1] http://arstechnica.com/tech-policy/2011/03/world-mobile-data-traffic-to-explode-by-factor-of-26-by-2015/
[2] http://www.hpl.hp.com/india/research/siteonmobile.html
[3] http://www.apple.com/iphone/features/siri.html
[4] http://www.mobile-ent.biz/news/read/mobile-web-use-to-pass-pc-based-internet-access-by-2016/019707
[5] Ernesto William De Luca and Andreas Nürnberger, "*Supporting Information Retrieval on Mobile Devices*," 2005.
[6] W. Aisha Banu, P. Sheik Abdul Khader, and Shriram Raghunathan, "*Mobile Information Retrieval: A Survey*," 2011.
[7] The Second Strategic Workshop on Information Retrieval in Lorne February 2012 about *Frontiers, Challenges, and Opportunities for Information Retrieval*.
[8] V. Anna Zhdanova, Ying Du, and Klaus Moessner, "*Mobile Experience Enhancement by Ontology-Enabled Interoperation in a Service Platform*."
[9] L. Eunshil, J. Kang, J.Choi, and J. Yang, "*Specific Web Content Adaptation to Mobile Devices*," 2006.
[10] H. Tomi and M.Käk, "*Mobile Findex: Facilitating Information Access in Mobile Web Search with Automatic Result Clustering*," 2007.
[11] K. Bade, E. W. De Luca, A. Nürnberger, and S. Stober, "*CARSA - An Architecture for the development of Context Adaptive Retrieval Systems*," 2005.
[12] E. W. De Luca and A. Nürnberger, "*Improving Ontology-Based Sense Folder Classification of Document Collections with Clustering Methods*," 2004.
[13] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, "*Dom-based content extraction of html documents*," 2003.
[14] A. Arasu and H. Garcia-Molina, "*Extracting structured data from web pages*," 2003.
[15] Ifeyinwa Okoye, Jalal Mahmud, Tessa Lau, and Julian Cerruti, "*Find This For Me: Mobile Information Retrieval on the Open Web*," 2010.
[16] J Cho, H Garcia-Molina, and L Page, "*Efficient crawling through URL ordering*," Computer Networks, 30(1 - 7):161 - 172, 1998.
[17] Taher H. Haveliwala, "*Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search*," 2002.
[18] S Brin and L Page, "*The anatomy of a large-scale hypertextual Web search engine*," Computer Networks, 30(1 - 7):107 - 117, 1998.
[19] Filippo Menczer, Gautam Pant, and Padmini Srinivasan, "*Topic-Driven Crawlers: Machine Learning Issues*," 2002.
[20] Junghoo Cho and Hector Garcia-Molina, "*Effective Page Refresh Policies for Web Crawlers*," ACM Transactions on Database Systems, 2003.
[21] Carlos Castillo and Ricardo Baeza-Yates, "*WIRE: an Open Source Web Information Retrieval Environment*

BIOGRAPHY

Suresh P is the Head , Department of Computer Science , Salem Sowdeswari College[Govt. Aided] , Salem. He received the M.Sc., Degree from Bharathidasan University, the M.Phil. Degree from Mononmaniam Sundaranar University, and the Ph.D. Degree from Vinayaka Missions University in 1993, 2003 and 2011 respectively, all in computer science. He is an Editorial Advisory Board Member of Elixir Journal. His research interest includes Data Mining and Natural Language Processing. He is a member of Computer Science Teachers Association, New York.

Jeril Kuriakose received the B.Tech. degree from Jeppiaar Engineering College, Chennai, in 2010 and M.Tech. degree from University of Mysore at Mysore, in 2012, all in information technology. He is an Assistant Professor in The Kavery College of Engineering, Salem. His research interests include data mining, networking and security.