Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 9; April-June, 2015 pp. 99-102

http://www.krishisanskriti.org/acsit.html

Novel Indexing Technique for Hidden Web Pages using Semantic Analysis

Seema Rani[1] and Sonali Gupta[2]

[1]M.tech. Student, Computer Science Department, YMCA University of Science & & Technology Faridabad, India
[2]Department of Computer Engineering, YMCA University of Science & Technology Faridabad, India
E-mail: [1]seemanain44@gmail.com

Abstract—*Hidden Web is a collection of web pages which cannot be accessed by a crawler by simply following the link structure. The high quality content of hidden web is hidden behind the search forms.A hidden web crawler fills these search forms automatically and retrieves the hidden content.The webpages fetched by hidden web crawler need to be indexed for fastsearching process.Indexing reduces the query processing time up to large extent.Many indexing techniques exist in literature to index the content retrieved by a crawler from surface web still there is need to index the hidden web content efficiently. In this paper, we are proposing an indexing technique which indexes the hidden web pages more efficiently using semantic analysis.*

Keywords*:-Hiddenweb, indexing, integrated index*

1. INTRODUCTION

Web is a huge hypertext information resource and increases dramatically. A number of recent studies have noted that a tremendous amount of content on the Web is dynamic. Hidden Web contains very large amount of information which is nearly 500 times larger than surface web.The Hidden Web is generally defined as the content on the Web not accessible through a search through a normal search engines. This content is sometimes also referred to as the deep web. Hidden Web consists of much valuable information hidden behind their query interfaces. Most of the users rely on traditional search engines to search the information on the web. Surface web are accessed by thosetraditional search engines that encompasses normal crawler which work as basic keyword matching scheme whereas to access hidden web, search engines must be enabled with a special crawler. As Hidden Web sources store their content in searchable databases that only produce results dynamically in response to a direct request. So there is no keyword matching scheme and no url following for accessing hidden web data. [Wikipedia 2014]

Hence Traditional crawlers retrieve content from the publicly indexable Web ignoring the tremendous amount of high quality content hidden behind search forms. Research scholars have explored various techniques for crawling of hidden web pages[6].Hidden web crawlers crawls hidden web.The information retrieved by hidden web crawlers is very useful.So,it is very important to index the information retrieved by hidden web crawlers. Indexing technique directly affects the searching time of a query in databases. Hence to reduce the searching time of a query an efficient indexing technique is required.

2. LITERATURE SURVEY

Many researchers are trying to develop novel ideas to index hidden web pages in order to improve searching techniques for Hidden web. A brief overview at few of them is given in the following subsections:-

- The traditional proposed techniques[3] are based on keywords. First the crawler downloads the documents and then extracts the keywords from these downloaded documents. Then indexer indexes the documents based on these keywords.
- Ritu Shandilya et al[3] proposed indexing technique which is based on (attribute, value) pair. To retrieve the doc hidden behind the form, it assigns the values to the attributes of the form and then submits the form .This indexing technique uses these attributes and their values to index that doc. This technique provides more relevant result than keyword based indexing.

3. PROBLEM IDENTIFICATION

By doing a critical analysis of various papers some problems are identified as-

- Search engine does not index hidden web data properly.
- Results given by search engine to the user are not relevant and specific.
- User has to follow the procedure of form filling by clicking on each link. Sometimes user gets to know at last that this is not the information which he wants. It is very frustrating from user's point of view.

4. PROPOSED WORK

A Novel indexing technique is proposed .It is a very efficient indexing technique for hidden web. This technique uses the combination of two type of indexing i.e. keyword based indexing and attribute value pair indexing. It helps in reducing the query processing time up to great extent .Query processing time is a very critical aspect of search engine from the user's view .In keyword based indexing, indexing is implemented on the basis of terms only .This is used for indexing the surface web. In this terms and their corresponding documents are stored in index file .So, it takes a lot of time to search the relevant information in the large index file.

In attribute value pair indexing, indexing is performed on the basis of attribute and their corresponding values .It consumes more time to find the common document indexed by index files of each attribute. Because in this indexing, an index file is created for every attribute and every time AND operation is performed on documents retrieved by these files to findcommon document. Proposed indexing technique is eliminating the problems of these two techniques by creating separate index files for attributes value pair and also for the related attributes value pair .Hence it is an efficient way of indexing which has less query processing time .The architecture of Proposed Integrated Indexer is as shown in fig.1.:-

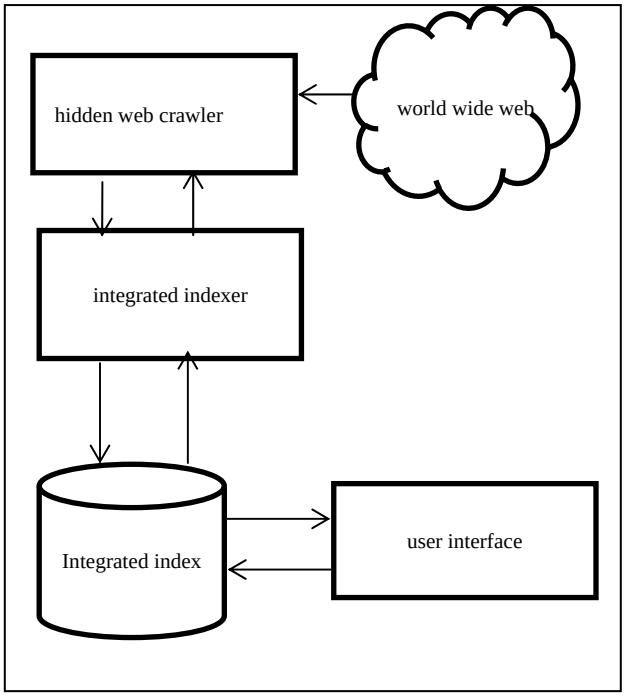

Fig. 1: Architecture of Proposed Integrated Indexer

An integrated index is created with the help of two separate indexer .i.e. attribute based indexer and keyword based indexer. Web pages retrieved by hidden web crawler are the input to attribute based indexer and keyword based indexer. The webpages are indexed on the basis of attributes and keywords also. These indexed files are given to integrator module which integrates these files and creates an integrated index. When a query is fired to the search engine, it first comes to integrated index to find the indexed document corresponding to the query. The query is first searched in attribute based indexing file, all the records matching to the user query in attribute based indexing file are shown to user and process of searching terminates. If user's query does not match with any attribute based indexing file then this query is searched in keyword based indexing file. And finally the results are shown to the user.

5. DETAILED ARCHITECTURE OF PROPOSED INTEGRATED INDEXER:-

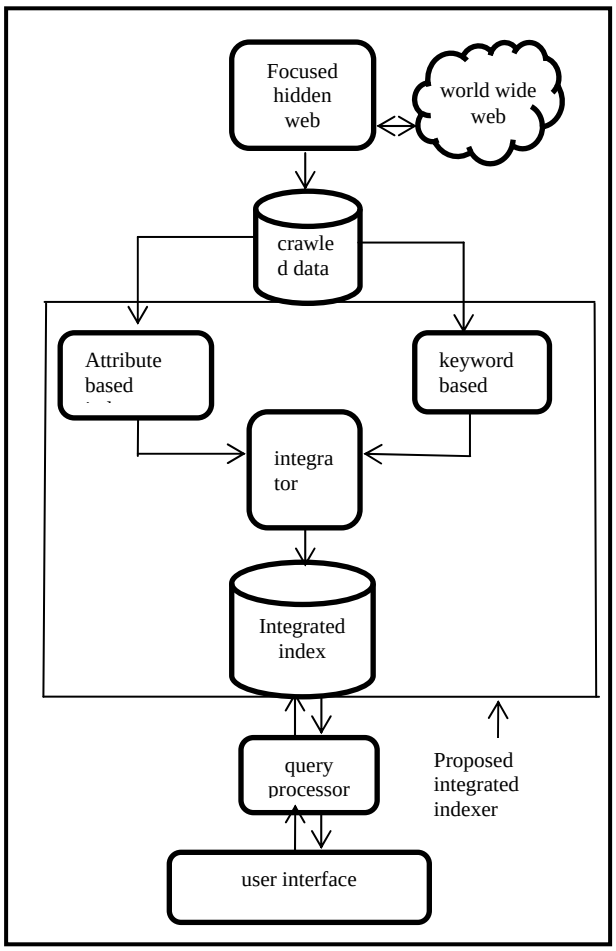

Fig. 2: Detailed Architecture of Proposed Integrated Indexer

5.1. Attribute Based Indexer

The Attribute Based Indexer is based on a indexing technique which index all the webpages according to the attributes and their corresponding possible values .This indexer first extracts

the attributes and its values from the webpages. Then possible combination of attributes are taken for creating indexes and also For more specific results and fast access we create index by combining two or more attributes.

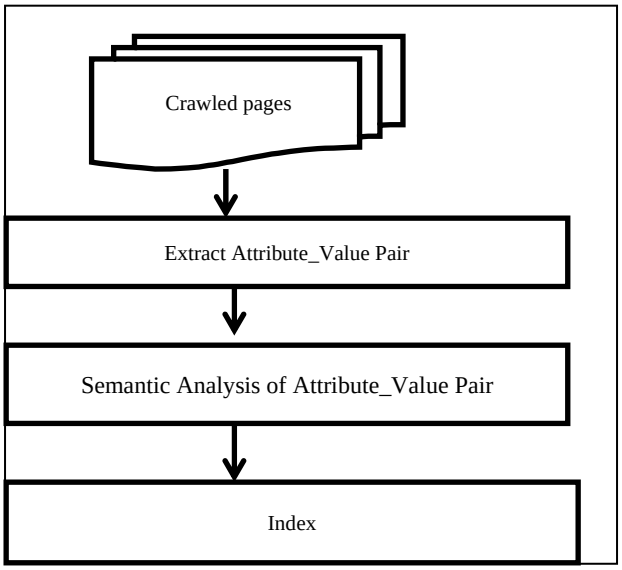

Fig. 3: Attribute Based Indexer

5.1.1.Working:-Attribute based indexer takes webpages as input and apply the following algorithm on the input webpages to create its attribute based index files.

Algorithm

Step1. Take input webpages from the crawled repository.

Step2. Extracts the headings from HTML tags.

Step3. Extracts the attribute and values from these extracted headings.

Step4. Now perform semantic analysis of attributes and values to find the correct value to a particular attribute and also find the relationship between attributes on the basis of query type to index more relevant webpages.

Step5. Now index the documents corresponding to attributes and their values.

Step.6. Stop**.**

6.2. Keyword Based Indexer

The Keyword Based Indexer takes webpages as input, then it tokenize the whole documents to get the keywords from the documents. It follows the algorithm given below to index the webpages. This index is searched only when user's query does not contain any attribute matching to the attribute based indexer's attribute list.It is good to give some results to user instead of showing nothing in response to his query. Keyword based indexing solves this problem.

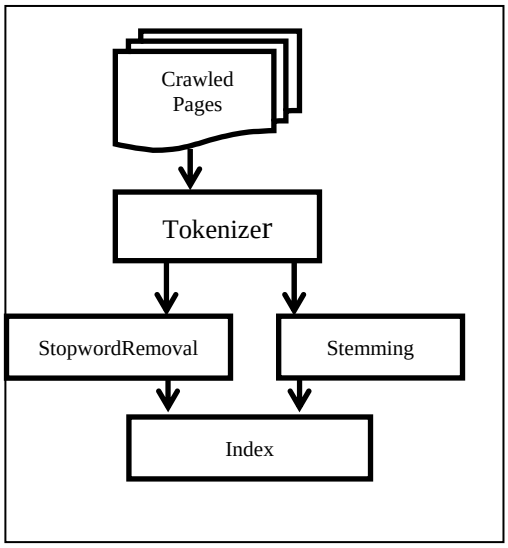

Fig. 4: Keyword Based Indexer

5.2.2. Working:-Keyword based indexer takes webpages as input and apply the following algorithm on the input webpages to create its keyword based index files.

Algorithm

Step 1. Take input webpages from the crawled repository.

Step 2. Now create tokens from these webpages with the help of a tokenizer.

Step 3. Stopword Removal process is applied to eliminate the irrelevant terms like is,am,are,of etc.

Step 4. Now the Stemming of terms is done to change the derived forms of terms to its root terms.

Step 5. And finally these root words are indexed to the documents containing these terms.

Step 6. Stop.

5.3. Integrator :-The integrator module takes input from keyword based indexer and attribute based indexer. It combines the index files provided by these two indexers and create a common index for the whole system.The output of integrator is an integrated index.

6.4. Integrated Index:-This is the index repository. It is accessed by the query processor to find the relevant and specific indexed webpages to the user's query.The web pages matching with more attribute value pairs are more relevant and specific.

6. EXAMPLE

Take an example of airline domain.A user makes a query flight from Delhi to Mumbai.This query is processed by integrated indexer as below:-

STEP.1

A query is submitted by the user to the user interface as shown below.

STEP.2

This becomes input to query processor. It processes the query .The query is converted into tokens.These tokens are semantically matched with the index files as shown below. On the basis of tokens Query Processor decides to which index files query should be submitted in integrated index.

1. SRC_DEST INDEX FILE

Source	Destination	Book_flight_by_clicking_on_link
src1	dest1	url1,url2
src1	dest 2	url3,url4
src1	dest3	url5url,6
src2	dest4	url7,url8,url9

2. SRC INDEX FILE

Source	Book_flight_by_clicking_on_link
src1	Webpage1, Webpage2, Webpage3, Webpage4
src2	Webpage7, Webpage8, Webpage9

3. DEST INDEX FILE

destination	Book_flight_by_clicking_on_link
dest1	Webpage1, Webpage2
dest2	Webpage3, Webpage4
dest3	Webpage5, Webpage6
dest4	Webpage7, Webpage8, Webpage9

4. SERVICE PROVIDER INDEX FILE

service_provider	Book_flight_by_clicking_on_link
Sp1	Webpage1,webpage2,webpage4,webpage6,webpage7
Sp2	Webpage3,webpage5,webpage8
Sp3	webpage9,webpage10

STEP.3

In integrated index,query is matched with the index files.The selection of index file is done by query processor.This index returns the matched webpage to the query processor.

STEP.5

Finally semantic query processor returns the results to the user shown as below-

Source	Destination	Book_flight_by_clicking_on_link
Src1	Dest1	Webpage1
Src1	Dest1	Webpage2

7. CONCLUSION

A technique for indexing of hidden web pages is proposed. This is an efficient indexing technique for reducing query processing time up to a large extent. It avoids the terms matching process used in keyword based technique. It also avoids the intersecting process done in attribute based indexing. Because it already creates indexing files using intersection of related attributes. It gives good performance to user at the time of searching information.So, it is a good technique of indexing hidden information.This work can be further extendedfor increasing the memory utilization.

REFERENCES

[1] http://en.wikipedia.org/wiki/ Deep_Web

[2] http://en.wikipedia.org/wiki/Index_term

[3] http://en.wikipedia.org/wiki/Search_engine_indexing

[4] Ritu Shandilya, Sugam Sharma, and Shamimul Qamar, - A Domain Specific Indexing Technique for Hidden Web Documents published in CISME Vol.2 No.2 2012.

[5] Shobhit University, India,2 Iowa State University, USA,3 Salman Bin Abdul Aziz University, KSA .

[6] .Komal kumar Bhatia, A.K.Sharrma, Rosy Madaan:AKSHR[2010] Novel Framework for a Domain-specific Hidden web crawler.

[7] Usha Gupta-Fetching the hidden information of web through specific DomainsinIOSR Journal of Computer Engineering Volume 16, Issue 2, Ver. VII (Mar-Apr. 2014)

[8] Sudhakar Ranjan, Komal K. Bhatia-query interface integrator for domain specific hidden web in International Journal of Computer Engineering & Applications, Vol. IV, Issue I/III.