ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Design and Implementation of a Novel Semantic Indexing Technique for Hidden Web Pages

Seema Rani, Anil Kumar, JyotiYadav

M.Tech Student, Dept of CE, YMCA University of Science & Technology, Faridabad, India

ABSTRACT: Hidden Web is a collection of web pages which cannot be accessed by a generic crawler by simply following the link structure. The high quality content of Hidden Web is hidden behind the search forms. The Hidden Web Crawler fills these search forms automatically and retrieves hidden information. The web pages fetched by Hidden Web Crawler need to be indexed for fast searching process. Searching process is directly based on the Indexing technique. There are many indexing techniques which index the contents retrieved by a general purpose web crawler. But these indexing techniques are not good for hidden web. So, the hidden web contents still need to be indexed efficiently. In this paper, an efficient indexing technique is designed and implemented. The ultimate goal of this indexing technique is to reduce query processing time and give more specific result pages to the user's query.

KEYWORDS:hiddenweb,indexing,first level indexer, second level indexer

I. 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. The information stored in hidden web is very important**.** Most of the users rely on traditional search engines to search the information on the web. Surface web are accessed by thosetraditional search engines which rely upon general purpose crawler. An indexer is a part of search engines which indexes the information. Mostly General purpose search engines use keywords for indexing. Search engines must be enabled with a special crawler for crawling hidden web, as hidden web sources store their content in searchable databases that only produce results dynamically in response to a direct request.

Hence Traditional crawlers retrieve content from the publicly indexableweb 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 crawl 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.

II. LITERATURE REVIEW

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:

- Seema Rani and Sonali Gupta[7] proposed an indexing techniquebased on keywords and attributes value pair. First crawler downloads the documents then extracts the keywords and attribute_value pairs from these downloaded documents.Then indexer indexes the documents based on these keywords and attribute value pair.

- Changshang Zhou, Wei Ding, Na Yang [9] proposed a double indexing mechanism for search engines based on campus Net. The Campus Net Search Engine (CNSE) is based on full-text search engine, but it is not general full-text search engine as it is basically a private net. The CNSE consists of crawl machine, Chinese automatic segmentation, index and search machine.

- RituShandilyaet 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 the documents containing these attributes and values. This technique provides more relevant result than keyword based indexing.

III. 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.
- When a user raises a query to a general search engine, a result page containing various links is displayed to the user. For getting the relevant information, user has to click on given links one by one. Sometimes user has to follow the procedure of form filling by clicking on each link. It is very time consuming and frustrating task from user's point of view.

IV. PROPOSED WORK

A Novel Semantic Indexing Technique for Hidden Web Pages isproposed.It is a very efficient indexing technique for indexing hidden web crawler's retrieved information. This technique uses attribute value pair for indexing the web pages. 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 attribute value pair indexing, indexing is performed on the basis of attributes and their corresponding values. It consumes less time to find the common document indexed by index files, as combination of attribute is also considered in this technique. In this indexing, separate indexing files are created for each attribute and for related combination of attributes. Hence it is an efficient way of indexing which has minimum query processing time.The architecture of Proposed Integrated Indexer is as shown in fig. 1.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

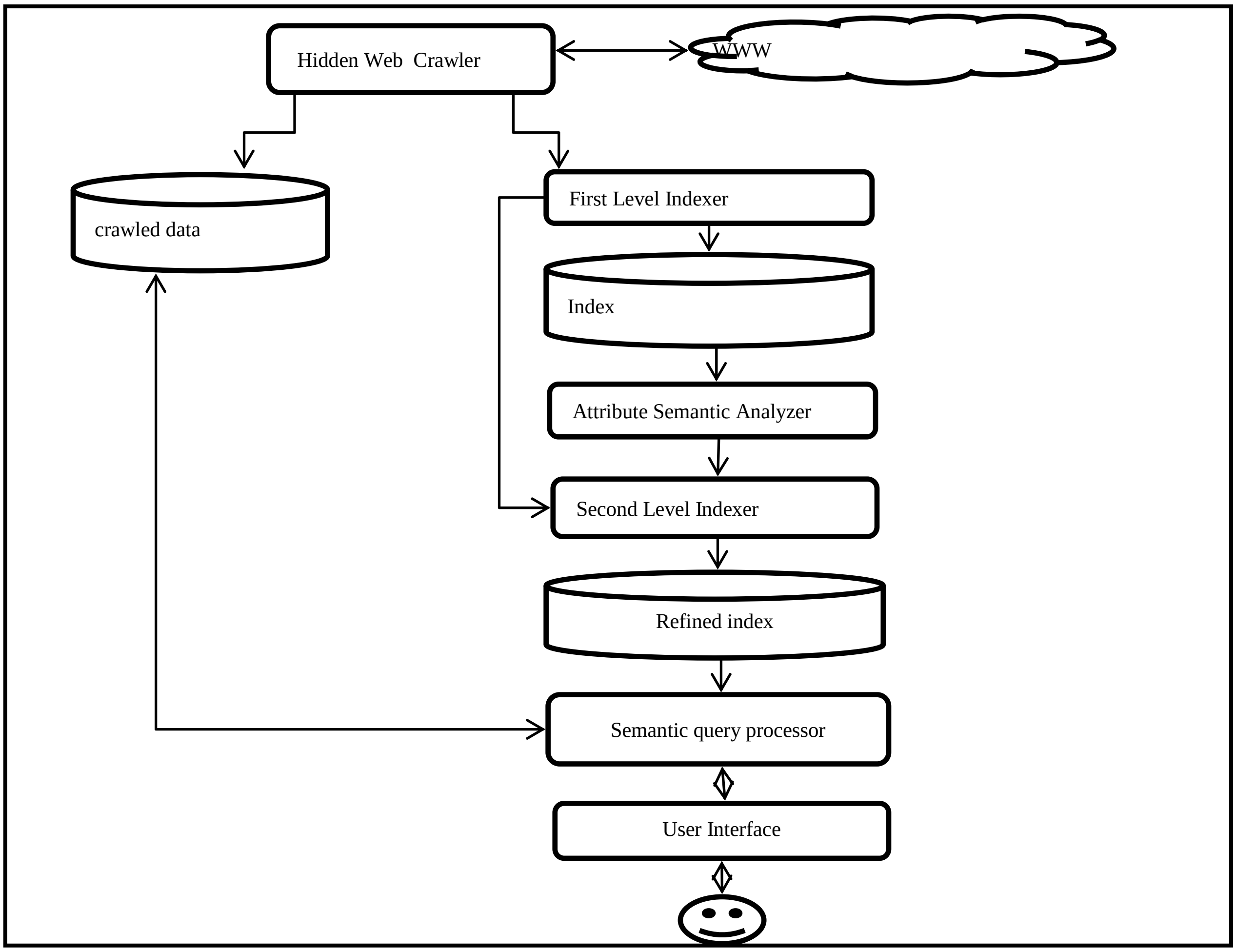


Fig. 1 Architecture of Proposed Indexing Technique

V. ALGORITHM FOR PROPOSED INDEXING TECHNIQUE VI.

Step1. Hidden web crawlercrawls the hidden web pages and also creates anattribute value pair file for all attributes and their corresponding values along with retrieved documents number while filling the search forms.
Step2. First Level Indexer indexes the all web pages according to attributes and their corresponding values.
Step3. AttributeSemantic Analyzer analyzes attributes & find relationship between them using user'sview of querying.
Step4. The Second Level Indexer indexes the web pages according to related attributes as given by Attribute Semantic Analyzer and their corresponding values.
Step5. Finally a Refined Index is created.
Step6. Stop**.**

VII. HIERARCHY OF INDEXING IN PROPOSED INDEXING TECHNIQUE

Hidden web belongs to different domains. Each domain deals with a particular topic. Each domain search form has various different attribute. So there is a need of creating different indexing files for each domain. At first level domain is indexed. Then further attribute based indexing is implemented on every domain simultaneously and index is created. The hierarchy of indexing has been shown in Fig. 2. If a query, mentioning the values 2 and 1 for attributes 1 and N

DOI: 10.15680/ijircce.2015.0305147

respectively in domain 1, is raised then document D2 is retrieved by query processor for such query using proposed indexing.

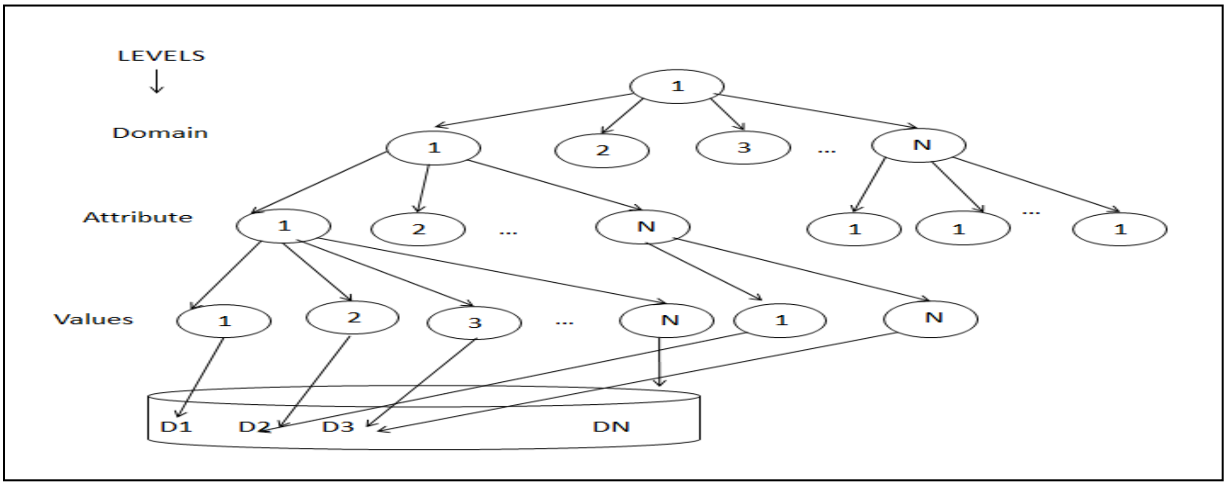


Fig. 2 Hierarchy of Indexing in Proposed Indexing Technique

VIII. EXPERIMENTAL EVALUATION

We have implemented this Technique for airline domain.Its query form has various attributes like source, destination, departure date, time etc. In below example some attributes has been taken to show the working of this technique.

User Interface: Fig. 3is designed user interface for searching flights. User has to query on this interface for his flights.

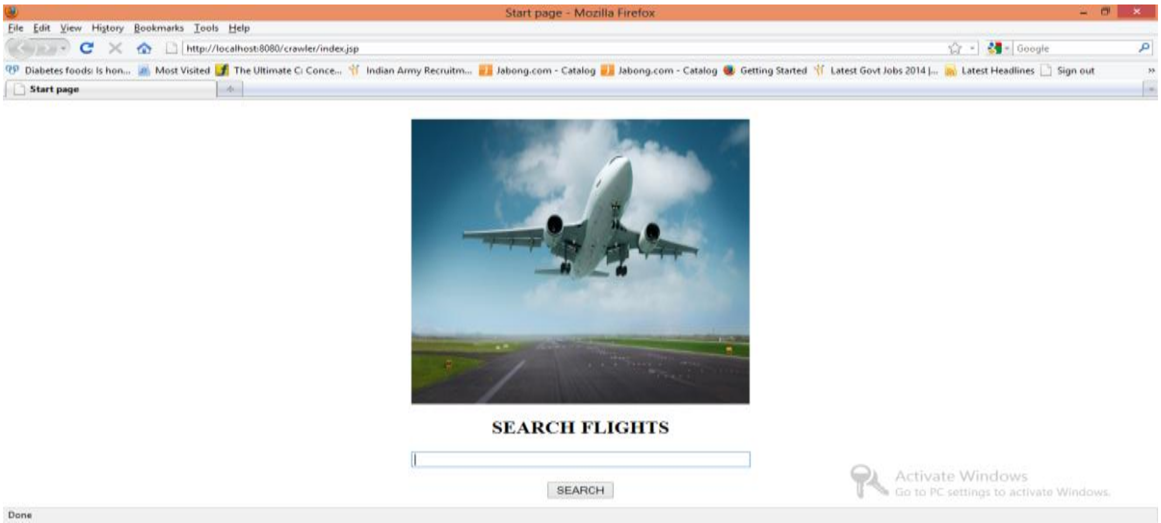
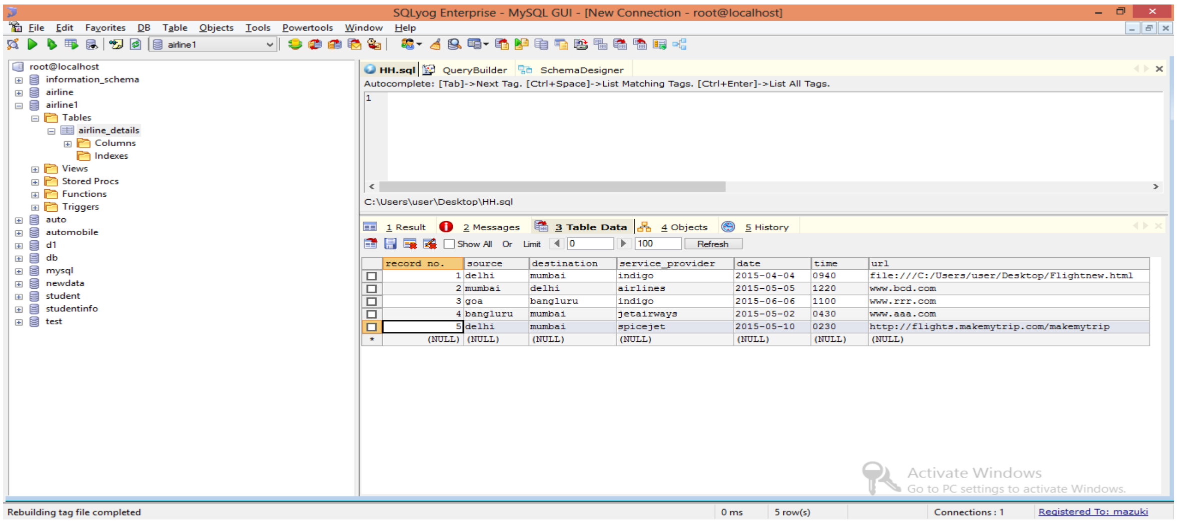

Start page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/crawler/index.jsp
Google
Diabetes foods: Is hon... Most Visited The Ultimate C: Conce... Indian Army Recruitm... Jabong.com - Catalog Jabong.com - Catalog Getting Started Latest Govt Jobs 2014 |... Latest Headlines Sign out
Start page
SEARCH FLIGHTS
SEARCH
Activate Windows
Go to PC settings to activate Windows.
Done

Fig. 3 User Interface

Crawled Repository: Fig.4shows crawled repository which needs to be indexed.



SQLyog Enterprise - MySQL GUI - [New Connection - root@localhost]

C:\Users\user\Desktop\HH.sql

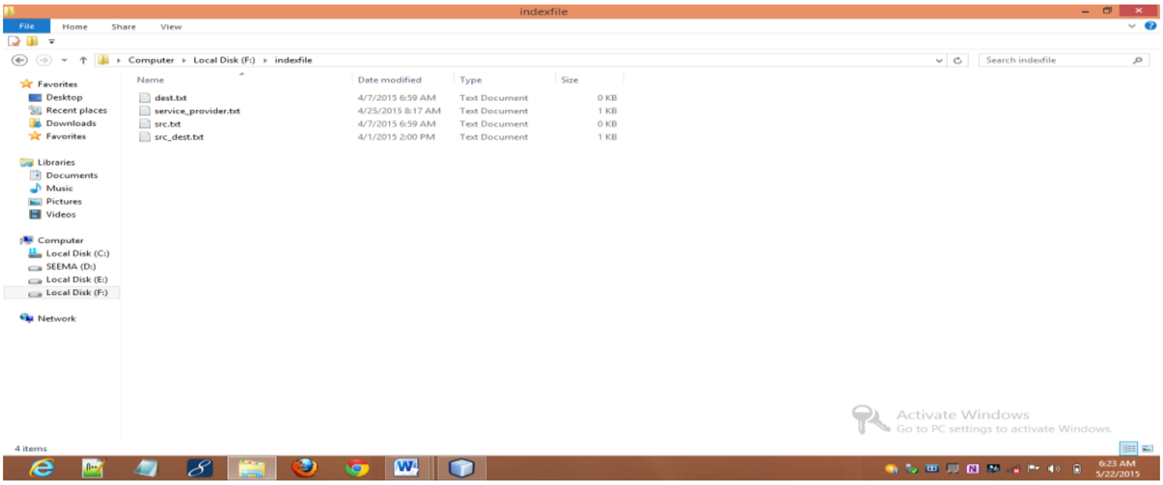
record no.	source	destination	service_provider	date	time	url
1	delhi	mumbai	indigo	2015-04-04	0940	file:///C:/Users/user/Desktop/Flightnew.html
2	mumbai	delhi	airlines	2015-05-05	1220	www.bcd.com
3	goa	bangluru	indigo	2015-06-06	1100	www.rrr.com
4	bangluru	mumbai	jetairways	2015-05-02	0430	www.aaa.com
5	delhi	mumbai	spicejet	2015-05-10	0230	http://flights.makemytrip.com/makemytrip
(NULL)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)

Rebuilding tag file completed

Fig. 4 Crawled Repository

Index created by First Level Indexer:First level index is created by first level indexer. It is created with the help of Hidden Web Crawler’s form filling process. In this index, documents are index using the particular value of particular attribute.

Refined Index: Fig. 5shows the created refined index with the help of second level indexer. It takes input from First Level Indexer and Semantic analyzer. Its output is integrated attributes indexing files.



indexfile

Computer ▸ Local Disk (F:) ▸ indexfile

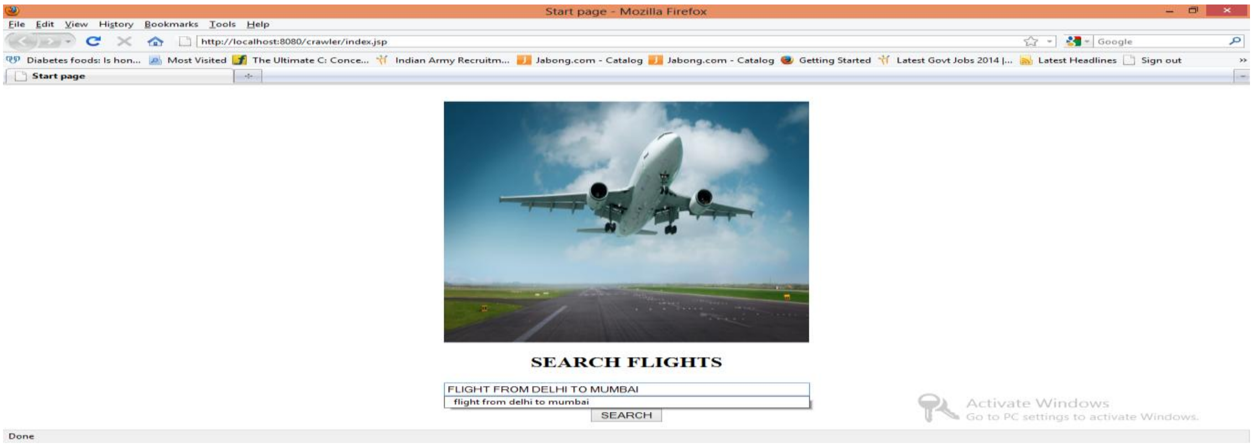
Name	Date modified	Type	Size
dest.txt	4/7/2015 6:59 AM	Text Document	0 KB
service_provider.txt	4/25/2015 8:17 AM	Text Document	1 KB
src.txt	4/7/2015 6:59 AM	Text Document	0 KB
src_dest.txt	4/1/2015 2:00 PM	Text Document	1 KB

4 items

Fig. 5 Indexing Files

User Query: Following are the snap-shots of user’s queries and their corrresponding results given by refined index.

Query type 1: When user raises query for flights from a particular source to a particular destination then this kind of query is raised. Fig. 6 shows such type of query for source delhi and destination mumbai.


Start page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/crawler/index.jsp
Google
Diabetes foods: Is hon... Most Visited The Ultimate C: Conce... Indian Army Recruitm... Jabong.com - Catalog Jabong.com - Catalog Getting Started Latest Govt Jobs 2014 |... Latest Headlines Sign out
Start page
Activate Windows
Go to PC settings to activate Windows.
Done

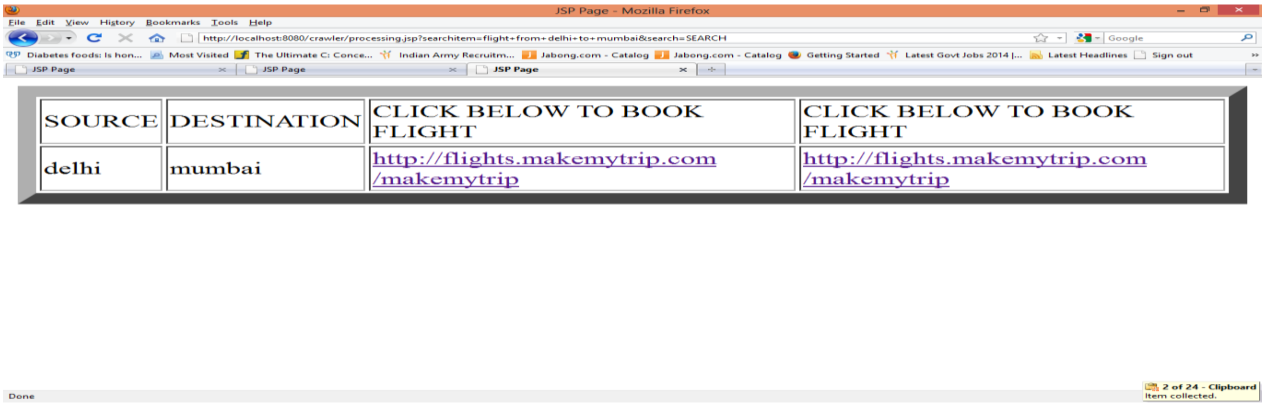
SEARCH FLIGHTS

FLIGHT FROM DELHI TO MUMBAI
flight from delhi to mumbai

SEARCH

Fig. 6 Query Type 1

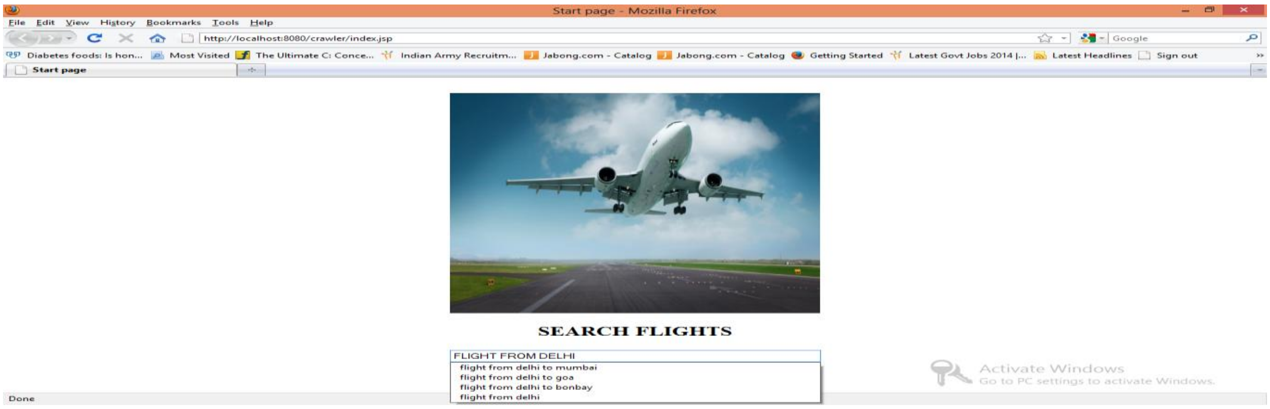
Result for query 1: Fig. 7 shows the result page for query 2. The refined index gives allurls for source Delhi and destination Mumbai.


JSP Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/crawler/processing.jsp?searchitem=flight+from+delhi+to+mumbai&search=SEARCH
Google
Diabetes foods: Is hon... Most Visited The Ultimate C: Conce... Indian Army Recruitm... Jabong.com - Catalog Jabong.com - Catalog Getting Started Latest Govt Jobs 2014 |... Latest Headlines Sign out
JSP Page
JSP Page
JSP Page
Done
2 of 24 - Clipboard
Item collected.

SOURCE	DESTINATION	CLICK BELOW TO BOOK FLIGHT	CLICK BELOW TO BOOK FLIGHT
delhi	mumbai	http://flights.makemytrip.com/makemytrip	http://flights.makemytrip.com/makemytrip

Fig. 7Result for Query 1

Query type 2:When user query for flights from a particular source and no destination specified then this kind of query is raised. Fig. 8 shows such type of query for source delhi.


Start page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/crawler/index.jsp
Google
Diabetes foods: Is hon... Most Visited The Ultimate C: Conce... Indian Army Recruitm... Jabong.com - Catalog Jabong.com - Catalog Getting Started Latest Govt Jobs 2014 |... Latest Headlines Sign out
Start page
Activate Windows
Go to PC settings to activate Windows.
Done

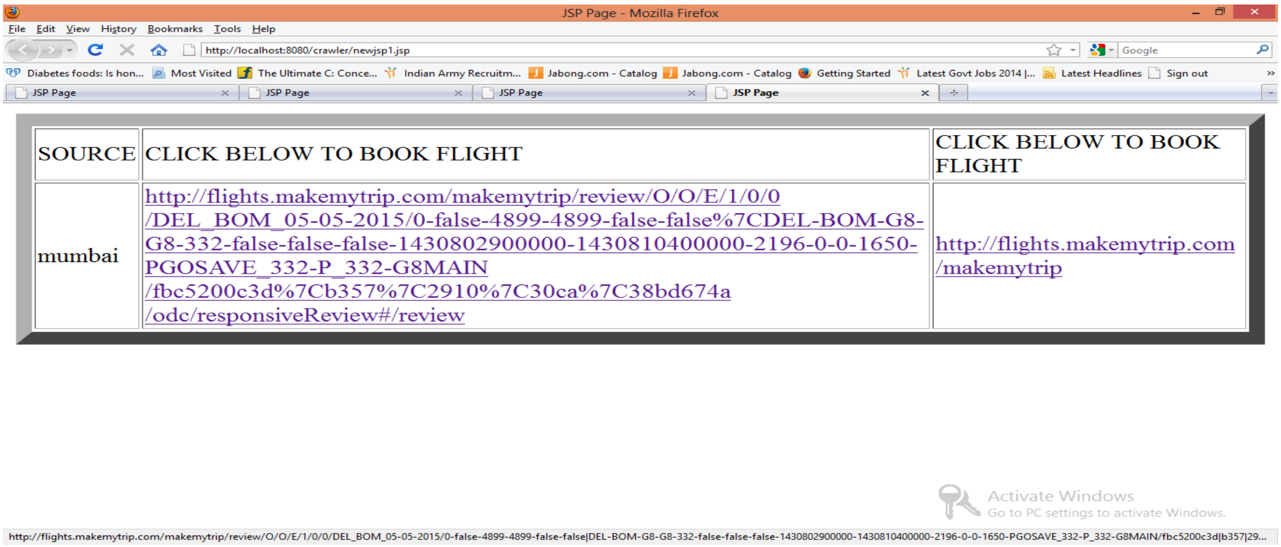
SEARCH FLIGHTS

FLIGHT FROM DELHI
flight from delhi to mumbai
flight from delhi to goa
flight from delhi to bonbay
flight from delhi

Fig. 8 Query Type 2

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

Result for query 2: Fig. 9 shows the result page for query 2. The refined index gives allurls for source Mumbai.


JSP Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Google
Diabetes foods: Is hon... Most Visited The Ultimate C: Conce... Indian Army Recruitm... Jabong.com - Catalog Jabong.com - Catalog Getting Started Latest Govt Jobs 2014 |... Latest Headlines Sign out
JSP Page
JSP Page
JSP Page
JSP Page
Activate Windows
Go to PC settings to activate Windows.

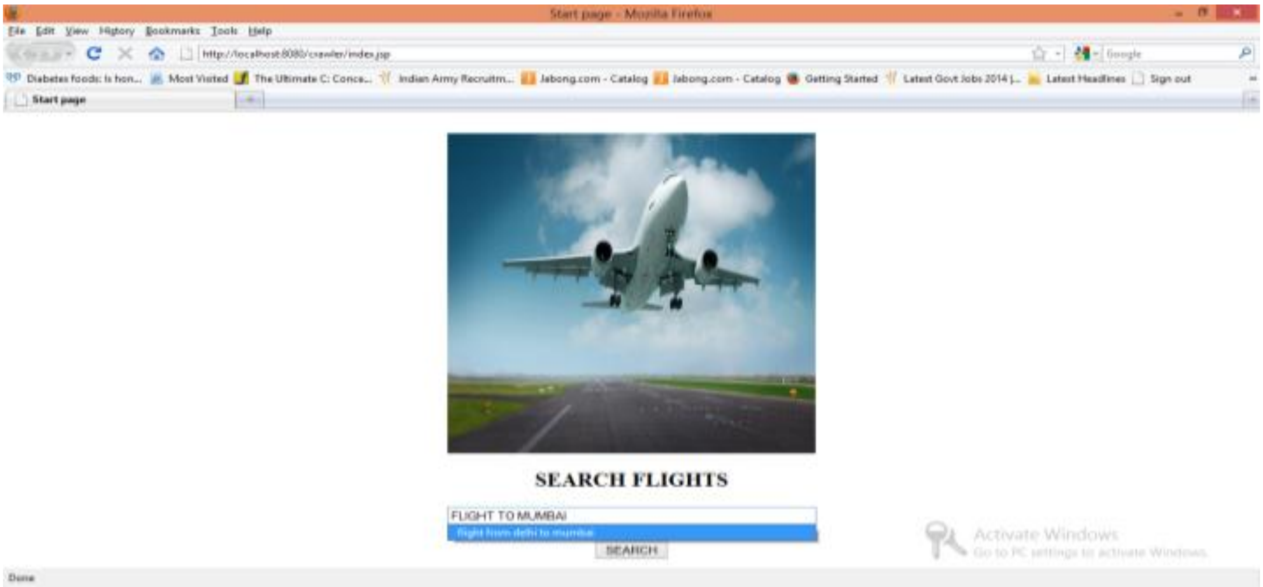
http://localhost:8080/crawler/newjsp1.jsp

SOURCE	CLICK BELOW TO BOOK FLIGHT	CLICK BELOW TO BOOK FLIGHT
mumbai	http://flights.makemytrip.com/makemytrip/review/O/O/E/1/0/0/DEL_BOM_05-05-2015/0-false-4899-4899-false-false%7CDEL-BOM-G8-G8-332-false-false-false-1430802900000-1430810400000-2196-0-0-1650-PGOSAVE_332-P_332-G8MAIN/fbc5200c3d%7Cb357%7C2910%7C30ca%7C38bd674a/odc/responsiveReview#/review	http://flights.makemytrip.com/makemytrip

http://flights.makemytrip.com/makemytrip/review/O/O/E/1/0/0/DEL_BOM_05-05-2015/0-false-4899-4899-false-false|DEL-BOM-G8-G8-332-false-false-false-1430802900000-1430810400000-2196-0-0-1650-PGOSAVE_332-P_332-G8MAIN/fbc5200c3d|b357|29...

Fig. 9 Result for Query 2

Query type 3: When user query for flights from a particular destination and no source specified then this kind of query is raised. Fig. 10 shows such type of query for destination Mumbai.


Start page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Google
Diabetes foods: Is hon... Most Visited The Ultimate C: Conce... Indian Army Recruitm... Jabong.com - Catalog Jabong.com - Catalog Getting Started Latest Govt Jobs 2014 |... Latest Headlines Sign out
Start page
SEARCH
Done
Activate Windows
Go to PC settings to activate Windows.

http://localhost:8080/crawler/index.jsp

SEARCH FLIGHTS

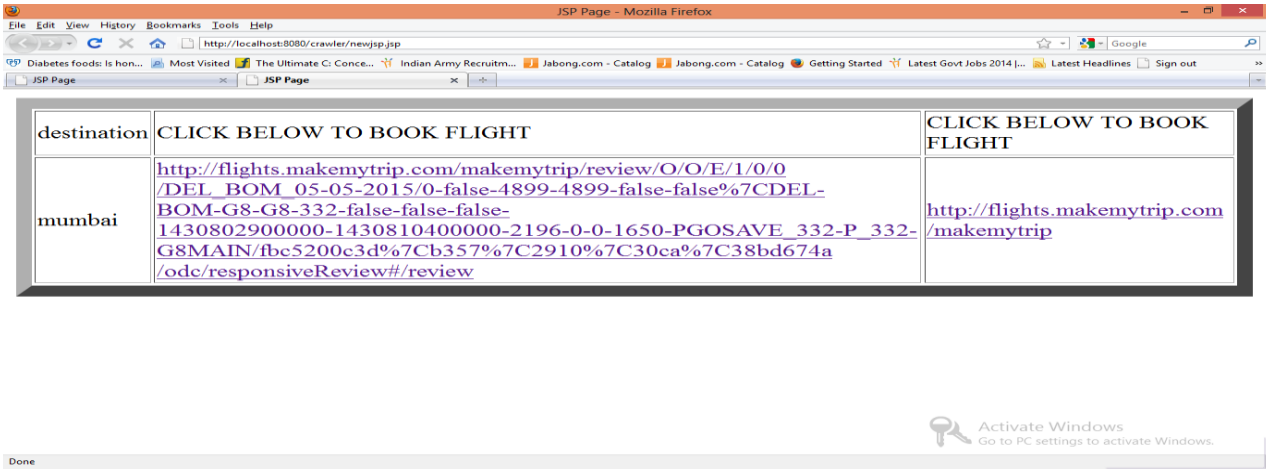
FLIGHT TO MUMBAI

flight from delhi to mumbai

Fig. 10 Query Type 3

Result for query 3: Fig. 11 shows the result page for query 3. The refined index gives allurls for destination Mumbai.

DOI: 10.15680/ijircce.2015.0305147


JSP Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/crawler/newjsp.jsp
Google
Diabetes foods: Is hon... Most Visited The Ultimate C: Conce... Indian Army Recruitm... Jabong.com - Catalog Jabong.com - Catalog Getting Started Latest Govt Jobs 2014 |... Latest Headlines Sign out
JSP Page
JSP Page
Activate Windows
Go to PC settings to activate Windows.
Done

destination	CLICK BELOW TO BOOK FLIGHT	CLICK BELOW TO BOOK FLIGHT
mumbai	http://flights.makemytrip.com/makemytrip/review/O/O/E/1/0/0/DEL_BOM_05-05-2015/0-false-4899-4899-false-false%7CDEL-BOM-G8-G8-332-false-false-false-1430802900000-1430810400000-2196-0-0-1650-PGOSAVE_332-P_332-G8MAIN/fbc5200c3d%7Cb357%7C2910%7C30ca%7C38bd674a/odc/responsiveReview#/review	http://flights.makemytrip.com/makemytrip

Fig. 11 Result for Query 3

IX. CONCLUSION

sA technique for indexing of hidden web pages has been designed and implemented. This is an efficient indexing technique for hidden web pages. Query processing time is a very vital aspect from user's point of view. Thistechnique reduces the query processing time up to a great extent, as it does not index web pages on the basis of keywords only. It also avoids the intersecting process for finding common documents which contains attributes and values matching to user's query. Hence it eradicates the problem of indexing faced in attribute value pair based indexing. It gives good performance to user at the time of searching information.So, it is a good technique of indexing for hidden information retrieved by hidden web crawler.This work can be further extendedfor efficient memory utilization.

REFERENCES

1. http://en.wikipedia.org/wiki/Index_term
2. http://en.wikipedia.org/wiki/Search_engine_indexing
3. **RituShandilya, Sugam Sharma, and ShamimulQamar**, *"A Domain Specific Indexing Technique for Hidden Web Documents"* published in CISME Vol.2 No.2 2012.
4. **Komalkumar Bhatia, A.K. Sharrma, Rosy Madaan**, *"Novel Framework for a Domain-specific Hidden web crawler"* in 2010.
5. **Usha Gupta**, *"Fetching the hidden information of web through specific Domains"* in IOSR Journal of Computer Engineering Volume 16, Issue 2, Ver. VII (Mar-Apr. 2014)
6. **SriramRaghavan, Hector Garcia, Molina**, *"Crawling the Hidden Web"*,CA 94305, USA.
7. **Seema Rani and Sonali Gupta,"***Novel Indexing Technique for Hidden Web Pages Using Semantic Analysis"in* Advances in Computer Science and Information Technology (ACSIT).
8. **A.K. Sharma, Komal Kumar Bhatia**, *"Merging Query Interfaces In Domain -Specific Hidden Web Databases"*in 2008.
9. **Changshang Zhou, Wei Ding, Na Yang**, *"Double Indexing Mechanism of Search Engine based on Campus Net"*, in Proceedings of sthe 2006 IEEE Asia-Pacific Conference on Services Computing (APSCC'06).

DOI: 10.15680/ijircce.2015.0305147