# NewNet- Crawling Deep Web

**Pradeep Rai**
Asst. Prof., CSE Department,
Kanpur Institute of Technology,
Kanpur-208001(India)

**Shubha Singh**
Asst. Prof., MCA Department,
Kanpur Institute of Technology,
Kanpur -208001(india)

**Abhishek Singh Yadav**
Research Associate Government projects
Bangalore - 560002

**Abstract**
Deep web are the WebPages which are not directly located by the search engines. Extracting the desired output from the searched query is not appropriate because our internet data is not structured. The desired information may store in the form of static pages or may be in the form of database, etc. So crawling those data with the same techniques is not responsive. In this paper we describe a new techniques called as NewNet by using Storewel so that this hidden web will no longer remains invisible. The Storewel will work as interface for the web crawler. Any crawler has to access the Storewel and Storewel will guide the crawler for the desired output. It simply gives you the output as desired by the Web developer or rather Web content holder. The web developer has also the capability to show the content for his desired query. This NewNet can also be used for securing the web content.

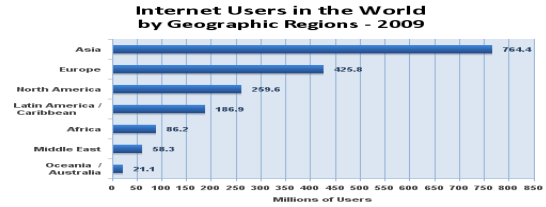
***Keywords:***
*Deep Web, NewNet, Storewel, Hidden web, Search Techniques, Web Security, XML*

## 1. Introduction:

In a current era searching the web is just as searching the sea for the "desired fish to eat". If that fish is on the surface and your net is capable enough to capture it then you will get your fish. If that fish is deep enough you will not get it and also if the structure of your net is not appropriate then also you will not get your fish if it is on the surface as well. Extracting the desired output from the searched query is not appropriate because our internet data is not structured. The desired information may store in the form of static pages or may be in the form of database, etc. So crawling those data with the same techniques is not responsive.

Google, the largest search database on the planet, currently has around eight billion web pages indexed. That's a lot of information. But it's nothing compared to what else is out there. Google can only index the visible web, or searchable web. But the invisible web, or deep web, is estimated to be **500 times** bigger than the searchable web. The invisible web comprises databases and results of specialty search engines that the popular search engines simply are not able to index.

The current statistics of users around the world is as given in the image below:-




This is the amount of peoples are using and searching the internet. And most of the time they are searching for the desired results. But they are getting nothing accurate as per their requirements. Also when we talk about the web content shown on web page the opinion differs what to show and whom to show.The web content should differ for different users.

We are proposing a new system for storing and retrieving the internet data. Our system NewNet is a new concept in storing and retrieving the internet data as it is totally based on the desire of the web master for "what to show, where to show & when to show "

## 2. NewNet:

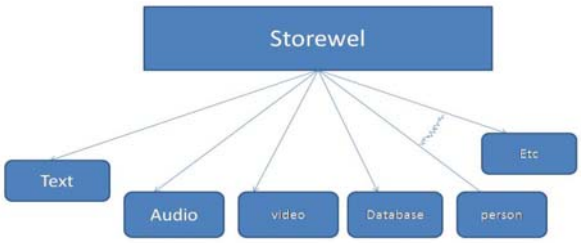
NewNet is altogether a new concept for web content access and storage. It majorly uses the storewel as a tool for all it's working. NewNet stores data about web content in it's new catalogue called Storewel. Storewel is a storage system or directory which has the capability same as index of any book. This complete facility is based on the desire of webmaster. As we know a index of an book requires only 1-2% of space for storing its contents , same as a storewel will increase the web content only upto 3%. But this excess storage will change every invisible site into a visible one. And the accessed information in the internet increase by almost 400 times. Every site has its storewel which is nothing but the index of data of website.

Manuscript received May 5, 2010
Manuscript revised May 20, 2010

The capability or excessive capability of storewel is to show what the webmaster or designer or the owner of the website wants to show. It will simply give a entry in the storewel as well as the link or links against it to show what he wants to show to the internet user.For example protocols---

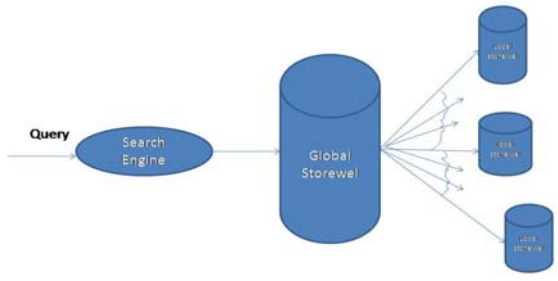
www.w3.org /Protocol/Activity.html#intro

The storewel stores data is in a alphabetical order. It can also stores data according to type of the data itself such as(called racks ) as :- i)Text ii) Video iii) Database iv) Image v) Audio vi) person vII) place etc
Each type of racks has its type of data. And composing these racks will form the storewel.




## 3. Querying the desired Web page:

When the user enters it query in the search engine. The search engine's spider will first go the **global storewel**. It is the index of the all the storewels exits in the web. It stores the global data in alphabetical order as well as in the form of racks – same as local storewel. Now this global storewel has the link to all the local storewels and these links will pass the required data to crawler.The Search engine's crawler has to crawl only the local storewel according to specific need of user's query. These storewel will index the complete data. Also the constraint over the data access can also be placed as per the requirement. The designer can design which page he wants to show with how much privilege, and also how much content he wants to show.




Query processing through global Storewels

Nobody can access the web content directly it can only be via the storewel. The access to the web content will become so secured that that access is based on codes written in different languages. Access to the web content of any website can be manipulates as per the directives of site owner.
This facility is only given to the sites having confidential data. The access is totally guided as per the codes written by the designer of the storewel. The storewel for a given site can only be formed by the third party if created automatically or it may be created by the web designer if created manually.

## 4. Storewel formation:-

Storewel can be formed using XML. The reason behind using XML is as follows:-

1. XML simplifies data sharing.
2. XML Simplifies Data Transport:-With XML, data can easily be exchanged between incompatible systems. One of the most time-consuming challenges for developers is to exchange data between incompatible systems over the Internet. Exchanging data as XML greatly reduces this complexity, since the data can be read by different incompatible applications.
3. XML Simplifies Platform Changes:- Upgrading to new systems (hardware or software platforms), is always very time consuming. Large amounts of data must be converted and incompatible data is often lost. XML data is stored in text format. This makes it easier to expand or upgrade to new operating systems, new applications, or new browsers, without losing data.
4. XML Makes Your Data More Available**:-** Since XML is independent of hardware, software and application, XML can make your data more available and useful.
5. XML is Used to Create New Internet Languages:-A lot of new Internet languages are created with XML.

Storewel so formed will be stored on the webspace allocated for the website for the general type of Websites. But in case of highly secured websites these storewels can be relocated at another web spaces. So these storewel will become the primary interface for all type of web contents.

In the formation of the storewel there are two type of strategies will be used:-

1. Manually by the storewel editor
2. Automated through website of the storewel itself.

Manually formation of the storewel will be done by the editor specifically used in storewel formation . In this type of storewel formation the web designer will open each page of the websites in the editor and do the necessary action as desired by him. Such as:-

1. Whom to show the complete web content and pass the parameter for that.
2. For general viewing what should be shown.
3. Decide the topics in which this particular page should be shown.
4. Also the alphabetical order of the index within storewel.

In the automatic formation of the storewel , the web contents are passed one by one to the automatic storewel formation tool and this tool will create the index based on the alphabetical order or on the basis of racks. After the automatic storewel formation the web designer will decide which should be the part of the storewel and which is not. Also the constraints on the web contents can also be applied.

## 5. Web Content Security by Storewel:

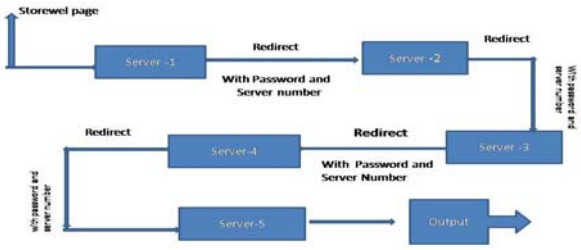
We can add several security features as per requirements, such as website security. We can provide a multiserver security for creating and maintaining the storewel. The basics motive of this NewNet technology is crawl the deep web. And change the static data into a dynamic database content so that it can respond according to the search engines.Storewel uses the techniques on the principal of indexing. We have two major techniques for storing and displaying data in our storewel:-

i) We have proper index in alphabhatical order for searching the entire website.
ii) Also we have categories such as person , audio , video etc for advance search

For making the storewel we can use any of the database of XML data storing techniques. As it can work on any platform. Nobody can directly access the web pages. One has to first access the storewel, and the storewel will give you the optimized webpage that will be directly access to any user. For complete user specific webpage we can pass the parameter to URL for the desired o/p. for example:-

www.storewel.com?p=”xxxxxxxxxx”

All the parameter passed will give u the desired o/p, on the basis of the XML or SGML code and that code is stored in storewel itself.We can use storewel for exposing the web pages into the internet. We can use multiserver security for accessing the web contents. We introduce a five server security system(5-3S)for web page security.




In the above security system storewel page will redirect the control to another server. As directed by web designer while creating the page through storewel editor. And this password and server number is only known to the web owner. And if unauthorized user wants to access the real page he/she should know the real server number and password combination.

## 6. Redirecting through storewel:

RSS redirection can be used for redirecting the URL. There are two ways to redirect your RSS feed:-

1. The http 301 redirect
2. The XML redirect:- You can also, or instead, post a special XML message to the address of your old RSS feed, advising RSS readers to redirect to a new location. Here's an example:

**Table : the XML redirect**

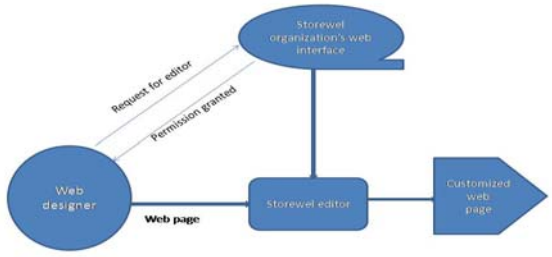
| address of old RSS URI | content of old RSS URI |
|---|---|
| http://radio.weblogs.com/0100887/rss.xml | <redirect><br><newLocation><br>http://weblog.infoworld.com/udell/rss.xml<br></newLocation><br></redirect> |

The latter redirection can be used to redirect the URL in the storewel.

## 7. Commercial usage of the Storewel:

Storewel technique can be further used for the commercial purposes. We have to cultivate a fully web based system for generating storewels for the different web sites. The

storewel editor should be available as freeware. And also for highly secured government official web sites there should be a centre for managing and storing these storewels.




Use of Storewel Editor

For the management of the global storewel there should be fully functional organization is required. Also searching the web will also be changed and now the web crawler should access the storewels.So for granting the permission to search engines this organization is also required.

## 8. Conclusions and Future works:

In our paper we have presented a new and highly efficient way of indexing the web content.And the use of Storewel will convert each inaccessible invisible web content into a visible one. Altogether the internet storage and accessibility will change by using the Storewel. Now the shown web content will be at the web designer disposal. This theme will change the current scenario. Also a new security feature is added to web content that will convert the confidential web data to highly secured data.

In future a lot of work has to done for implementing the NewNet. A lot more new algorithms should be discovered so that access to these storewel should become efficient and lesser time consuming. The storewel itself can change its present representation into a newer one.

### Acknowledgement

The authors would like to express their cordial thanks to Mr. Anshul Pandey (Research Engineer Intel, USA) and Mr. Chandresh Verma (Project Manager, Satyam) for their valuable advice.

### References

[1] CRAWLING DEEP WEB CONTENT THROUGH QUERY FORMS Jun Liu, Zhaohui Wu, Lu Jiang, Qinghua Zheng, Xiao Liu

[2] Answering Web Questions Using Structured Data – Dream or Reality?Panel Discussion -Fernando Pereira, Anand Rajaraman, Sunita Sarawagi, William TunstallPedoe, Gerhard Weikum, Alon Halevy (moderator)

[3] Thanaa M. Ghanem and Walid G. Aref. Databases deepen the web. *IEEE Computer*, 73(1):116–117, 2004.

[4] BrightPlanet.com. The deep web: Surfacing hidden value. Accessible at http://brightplanet.com, July, 2000.

[5] Steve Lawrence and C. Lee Giles. Accessibility of information on the web. *Nature*, 400(6740):107–109, 1999.

[6] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener. A large-scale study of the evolution of webpages. In *Proceedings of the 12th International World Wide Web Conference*, pages 669–678, 2004.

[7] Ed O'Neill, Brian Lavoie, and Rick Bennett. Web characterization. Accessible at "http://wcp.oclc.org".

[8] GNU. wget. Accessible at "http://www.gnu.org/software/wget/wget.html".

[9] Kevin Chen-Chuan Chang, Bin He, and Zhen Zhang. Toward large scale integration: Building a metaquerier over databases on the web. In *CIDR 2005 Conference*, 2005.

[10] Hai He, Weiyi Meng, Clement Yu, and Zonghuan Wu. Wise-integrator: An automatic integrator of web searchinterfaces for e-commerce. In *Proceedings of the 29th VLDB Conference*, 2003.

[11] L. Barbosa and J. Freire. Siphoning hidden-web data through keyword-based interfaces. In SBBD, 2004.

[12] M. J. Cafarella, A. Halevy, Y. Zhang, D. Z. Wang,and E. Wu. Uncovering the Relational Web. In WebDB, 2008.

[13] M. J. Cafarella, A. Halevy, Y. Zhang, D. Z. Wang,and E. Wu. WebTables: Exploring the Power of Tables on the Web. In VLDB, 2008.

[14] Cazoodle Apartment Search. http://apartments.cazoodle.com/.

[15] Every Classified. http://www.everyclassified.com/

**Abhishek Singh Yadav** received his M. Tech degree in Computer Science and Engineering from Jawaharlal Nehru University, New Delhi in the year 2007 and B .Tech in Computer Science and Engineering degree from RML University, Faizabad in 2002. Currently, he is associated with research organization under Govt. of India. His area of interest includes Network Security, Data Mining and warehousing, Mobile Ad-hoc Network.

**Pradeep Rai** received his bachelor degree in computer Science & Engineering from KNIT, Sultanpur in the year 2002 and M.Tech in computer Science in the year 2008. Currently he is working as Asst. Prof. in CSE Department at KIT, Kanpur. His area of interest includes VPN, wi-fi networks, network Security.

**Shubha Singh** received her Master degree in Computer Applications from Agra university in year 2002 and M.Tech in computer science in year 2007. She has worked as associate in govt project at IIT, Kanpur. Presently she is working as Asst. Prof. in Compute Application Deptt. At KIT, Kanpur. She has more than 8 years teaching experience.Her areas of interest includes DBMS,Networks and Operating Systems.