Cover Page

DOI: http://ijmer.in.doi./2022/11.05.66

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :7.816(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:5(4), May: 2022
Online Copy of Article Publication Available (2022 Issues): www.ijmer.in
Digital Certificate of Publication: http://ijmer.in/pdf/e-Certificate%20of%20Publication-IJMER.pdf
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd May 2022
Publication Date:10th June 2022
Publisher: Sucharitha Publication, India

VISIBILITY AND ACCESSIBILITY OF DIGITAL LIBRARY: SEARCH ENGINE OPTIMIZATION AND FEEDS

Dr. Bhoop Singh
Assistant Professor
Bhartiya Skill Development University
Jaipur, Rajasthan, India

ABSTRACT

With increasing the information generation, the information overload is increasing and there is challenge to provide proper information. Digital libraries are playing vital role in this scenario. In this paper the authors discussed the process of visibility and accessibility of digital library or institutional repository. The work process, newsfeed and optimization of search engine are introduced.

Keyword: Search Engine, Indexing, Web Crawler, PageRank, Newsfeed, Optimization.

INTRODUCTION

With the information explosion and use of ICT in the library field, the use of blogs very helpful to providing information to the users (Lata & Somvir, 2014). For most users of the Internet a search engine is the first entry point to find information. However, although the first search engine ("Archie") was launched in 1990, search engines are not part of the basic infrastructure of the World Wide Web. The search engines that most people use are commercial ventures supported by advertising revenue.

LITERATURE REVIEW

The presence of FOSS in the digital library (DL) software category, that is, DSpace, Eprint and GSDL, and in ILMS category, Koha and Newgenlib, was examined by Hanumappa et al. (2014). The migration or adoption of FOSS among Indian libraries has drawn considerable interest from institutions recently, as the benefits to the library become better known. But there are barriers to the implementation of FOSS. Rathee & Kaushik (2019) discussed the step and planning of the digital library. Rafiq et al. (2018) studied the barriers to digitalization in the central university libraries of Pakistan. The mixed methods of quantitative and qualitative research were used for their study. Rathee et al. (2020) carried out a study and reported through a paper entitled "Develop a multiple-criteria decision analysis (MCDA) cause and effect factor model for the implementation of the Greenstone Digital Library (GSDL) software". Somvir & Kaushik (2021) the relationship of Library 2.0 and Web 2.0, potential uses of social softwares to interact with users and information dissemination.

SEARCH ENGINE PROCESS

There are **two basic processes** that the search engines have in common:
Web crawling to collect and update content. A **Web crawler** (or Web spider) is a computer program that browses the World Wide Web in a methodical, automated manner:

1. it follows every link in a page;
2. it stores the resources that these links point to; and
3. It follows all the links in those resources.

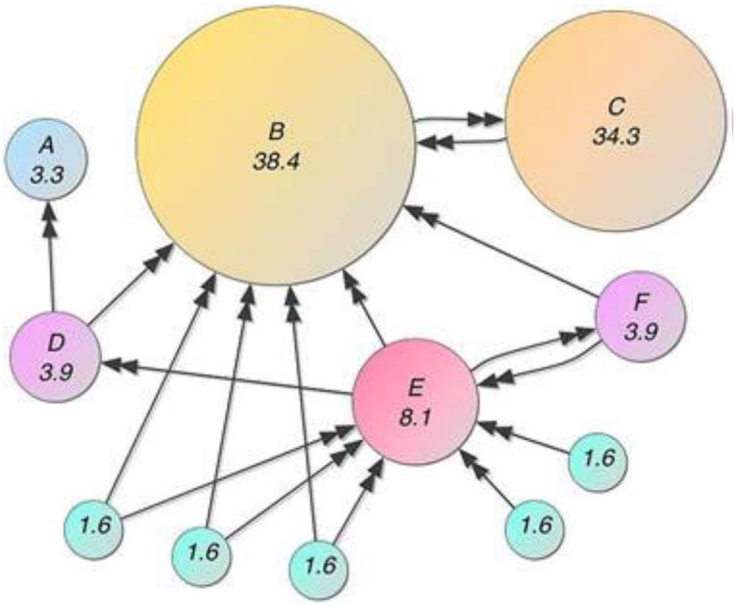
There may be a **limit to the number of links that the crawler follows**, and the site administrator may indicate that certain resources should not be crawled or the links in those resources should not be followed.

Indexing to facilitate fast searching (without an index all the documents that are collected would have to be scanned for each search). Indexing of text files means creating a list of words and pointers to all the documents where that word occurs.
Most search engines use additional intelligence, such as: **stemming** indexing only the stem of a verb, or the plural and singular form of a noun; and **proximity**, how close a word is to other words. All this semantic intelligence needs to be applied in different languages, making the whole process rather complex. Because of the dimensions of the web, the number of results for many queries is more than anybody can scroll through, and therefore users would like to find the most important results on top. For that purpose, search engines use a technique that is often referred to as **PageRank**. In a way the PageRank can best be compared to a popularity vote and the search line of a Web page is a vote. The term **PageRank** was first introduced by Google, and it is named after one of the founders of Google, Larry Page. Pageranks are calculated for individual web resources not for sites or domains.

Cover Page

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :7.816(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:5(4), May: 2022
Online Copy of Article Publication Available (2022 Issues): www.ijmer.in
Digital Certificate of Publication: http://ijmer.in/pdf/e-Certificate%20of%20Publication-IJMER.pdf
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd May 2022
Publication Date:10th June 2022
Publisher: Sucharitha Publication, India

DOI: http://ijmer.in.doi./2022/11.05.66



B is the largest, and it receives seven links. It is much larger than **E**, although **E** receives six links. The reason is that **B** receives more links from sites that have received links themselves. **The more votes a site receives the more weight is assigned to the votes that it casts**. For example, **C** receives only one link, so it gets one vote. But that vote is from **B**, the most popular site, and it is the only vote that **B** casts. So, **C** ends up as the second largest circle. Of course, the entire Web is much larger and the calculation will become more and more complex. The Web is also dynamic and the **PageRank is recalculated on a regular basis**.

FEATURING DOCUMENT COLLECTIONS IN SEARCH ENGINES

If we want the documents in our digital library to be indexed in search engines, we should bear in mind that **web crawlers** follow links in a systematic way, they **stop at a search box**. A **browsable list of latest additions with links to those documents** offers a partial solution. **Sitemaps** are another technique that can be used to submit the content from databases like repositories to search engines.

Sitemaps

Sitemaps are an easy way for webmasters to inform search engines about pages on their sites that are available for crawling. In its simplest form, a sitemap is an XML file that lists URLs for a site along with additional metadata about each URL (when it was last updated, how often it usually changes, and how important it is, relative to other URLs in the site) so that search engines can more intelligently crawl the site.

To create a sitemap XML file, we can use the database that is behind the repository. The file should be available in the directory where the documents are stored, or in a higher directory (or the root directory of site). If a file can put there, the webcrawler assumes this the owner of the site.

The Sitemap protocol (http://www.sitemaps.org/index.php) is a standard for preparing XML sitemaps, created and used by all of the most important search engines.

Using the Sitemap protocol does not guarantee that web pages are included in search engines, but provides hints for web crawlers to do a better job of crawling site. GSiteCrawler (http://gsitecrawler.com/) is a free software that generates Sitemap files by crawling a static website. So, if we have a browsable index of the documents in repository/digital library the software can use to create a Sitemap.

SEARCH ENGINE OPTIMIZATION

Creation of sitemaps is one technique that is used for search engine optimization. Search engine optimization in general is meant to make sure that web crawlers can find the items on site, and improve their ranking in the results from the search engine. As this

Cover Page

DOI: http://ijmer.in.doi./2022/11.05.66

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :7.816(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:5(4), May: 2022
Online Copy of Article Publication Available (2022 Issues): www.ijmer.in
Digital Certificate of Publication: http://ijmer.in/pdf/e-Certificate%20of%20Publication-IJMER.pdf
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd May 2022
Publication Date:10th June 2022
Publisher: Sucharitha Publication, India

is very important in commercial environments, search engine optimization has become quite an industry. There are two forms of search engine optimization:

The white hat method: This works on search engine accessibility that is 'transparent' and that search engines approve of. There are a number of recommended approaches for **White hat** search engine optimization of site. The most important ones are:

- Think carefully about which words we want the user to enter in order to find page. Check that these words do in fact lead them to page.
- Use clear titles for pages, because that is what the user clicks on in the results from the search engine. In technical terms, the <title> in the <head> of web page.
- Make sure each item can be reached from a static link. Dynamic links (links with a question mark '?' in the URL) within site will cause difficulties. In technical terms: dynamic links are links that do not go to a static HTML page but call an external program (like a database management program) that generates the HTML dynamically. Their URL usually contains a question mark. One should also be careful with links that are generated by Java scripts on a page.
- If we choose a Content Management System (CMS) to maintain website, an important requirement should be that search engines are able to crawl sites which use this CMS.

To improve the ranking of pages the best recommendation is to work together with peer sites that work in the same subject area and link to each other's pages wherever appropriate.

The black hat method: This tries to trick search engines in various ways. Search engines penalize black hat optimization and may remove sites that apply those techniques from their index.
The Black hat method may involve:

- cloaking, which is exposing different content to the web crawler than is exposed to human visitors (by using specific tags in the source, or by having the font colour the same as the background colour); or
- link farms, which is creating artificial links to items to improve the pagerank.

There are examples of reputable firms that have been punished that way (although it is has been repaired quickly after apologies from those firms). So best **avoid this method,** and if a consultant approaches us to improve performance in search engines, ask carefully how this will be done.

NEWSFEEDS AND AGGREGATORS

Feeds are another way to make the content of digital library known beyond own website.

A feed is a regularly updated summary of content – blog entries, headlines, publications, multimedia – in the form of metadata about the source and the contents. It includes links to the full versions of those contents at their original location.
Feeds are often referred to as newsfeeds, as they became popular when news media (newspapers, television networks) started providing them.

From a content producer perspective, a feed is a data format used to provide users with frequently updated content. It allows the publisher to 'syndicate' its content to many sources at once. Content producers can serve either all of their content or specific selections as feeds.

From consumer perspective feeds offer a way for web content to be selectively tracked, subscribed to, and perhaps customized into a personal service. Webmasters can also re-publish such feeds on their own websites.
Feeds are often called RSS feeds (acronym for Really Simple Syndication or RDF [Resource Description Framework] Site Summary).


What all feed standards have in common is that they are all in XML format, and sometimes feeds on a website are indicated by a button like this...


can also find a button like this...

Most recently the makers of popular Internet browsers (Internet Explorer, Mozilla have agreed to use this button.

RSS and Atom standards

Feeds are delivered using different versions of the RSS standard, such as: RSS 0.91, RSS 1.0 and RSS 2.0, etc. These versions are technically quite different, and they were developed by different groups of people. To end the confusion, it was decided to do further development on a new standard, Atom. Feeds may be delivered using this standard as well. This all sounds very confusing but as a feed

Cover Page

DOI: http://ijmer.in.doi./2022/11.05.66

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :7.816(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:5(4), May: 2022
Online Copy of Article Publication Available (2022 Issues): www.ijmer.in
Digital Certificate of Publication: http://ijmer.in/pdf/e-Certificate%20of%20Publication-IJMER.pdf
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd May 2022
Publication Date:10th June 2022
Publisher: Sucharitha Publication, India

provider there is no need to worry about all this: software that handles feeds handles all versions. Feeds allow users to create their own menu of dynamic content by subscribing to different feeds. This process is called aggregation. Consumers of any kind (individuals or services) can opt to get specific feeds displayed in a special program, a feed reader.

Subscribing to an RSS feed reader

First steps

Subscribing to an RSS feed is easy.

To get an RSS feed: go to the web site of interest with dynamic content (content that is frequently changed, added to or updated) such as a blog or a news site;

look for the RSS link on the page (it is usually a small icon with the letters "RSS", "Atom ", "XML" or a link on the page...

In some browsers the RSS icon can also appear next to the site address in browser address bar.

Using web browser as feed reader

If web browser used as feed reader: when click on the RSS link, will receive instructions to add this feed to bookmarks; once there, when click on the new link, will see a summary of the latest headlines pulled out from the feed subscribed.

Using a feed reader application

If we use a feed reader client, we can add a New Subscription and write down the **feed URL** (or copy and paste it from web browser address bar). Feed reader application will start pulling out the new content the next time it appears on the web.

Once subscribe, will notice that feed reader shows the post in **different formats**:

- as **headlines only** - for when want to go quickly over a large amount of information;
- as **summary of the posts** - for when not interested in the details or following links, but want to have a general idea of what the content is about;
- as **entire posts** - for when want to read the complete text and follow links;
- as **links to the original posts on the web** - for when want to visit the posts on their original formats as they are posted by the author on their website.

News aggregation services

News aggregation services or websites are sites which collect news. The feeds are either selected by the site owner or registered by the feed owners. Examples of news aggregation services are: AgriFeeds and Google News. Web browsers and e-mail programs can now often handle feeds as well. There are also websites that offer these options, and there are tools for webmasters to feature feeds (from their own site, or from others) as dynamic content on their site. There are specialized news aggregator services that may be connected with search engines. There are also specialized aggregator services.

Although feeds are commonly used for news items, they are also an excellent way of:

- keeping the users of website and repository up to date with the latest developments. Feeds are also particularly useful in low bandwidth environments. Users do not need to connect to a site unless they are alerted that there is something new.
- letting other services re-publish content

Providing newsfeeds for document collections

When providing newsfeeds for document collections we will need to:

- **select what feeds want to provide:** one general feed (latest from the repository) or specialized feeds on different subject areas;
- **decide how to present these feeds to users** (using the icons)
- go into the **technicalities of producing newsfeeds**. If a CMS used or a library system it may very well have that option so it is worth checking. If selecting a software solution for digital library it can be on list of requirements.

NEWSFEEDS FROM DOCUMENT COLLECTIONS

If digital library platform does not provide with the option to produce feeds, we can:

- as an **short-term solution**, maintain a **blog** that features recent acquisitions of digital library, probably as a manual process of cutting and pasting. There is specialized blogging software available, and there are public platforms where creator can start a blog. Blogs come with the option of providing an RSS or Atom feed.
- as a **long-term solution** creator may opt to develop and implement **feed creation as a feature of digital library platform**; this will require involving programmers with XML experience. Programmers with XML experience need to get familiar with


Cover Page

DOI: http://ijmer.in.doi./2022/11.05.66

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY EDUCATIONAL RESEARCH
ISSN:2277-7881; IMPACT FACTOR :7.816(2022); IC VALUE:5.16; ISI VALUE:2.286
Peer Reviewed and Refereed Journal: VOLUME:11, ISSUE:5(4), May: 2022
Online Copy of Article Publication Available (2022 Issues): www.ijmer.in
Digital Certificate of Publication: http://ijmer.in/pdf/e-Certificate%20of%20Publication-IJMER.pdf
Scopus Review ID: A2B96D3ACF3FEA2A
Article Received: 2nd May 2022
Publication Date:10th June 2022
Publisher: Sucharitha Publication, India

RSS/Atom specifications. They would be wise to make use of third-party programming libraries with routines such as Feedcreator (http://feedcreator.org) or Rome (https://rome.dev.java.net/).

SUMMARY

Search engines use webcrawlers to scan the web systematically and index the pages that they find to make them searchable. They use techniques like Google's PageRank to determine the relative importance of pages, and show the most important pages on top. There are several search engine optimisation techniques to make sites better accessible to web crawlers and provide search engines with useful terms to index. To feature electronic documents from digital libraries and repositories owners can submit Sitemap files to search engines. Black hat methods, which aim at providing the search engine with other information than that given to the human user, should be avoided. Newsfeeds enable users to create their own collection of dynamic content, and the information provider to syndicate the same content to several channels. They offer an important opportunity to feature the latest additions of a digital library.

REFERENCES

- Hanumappa, A., Dora, M. and Navik, V. (2014), Open-source software solutions in Indian libraries. Library Hi Tech, 32 (3), pp. 409-422.
- Rafiq, M., Ameen, K. and Jabeen, M. (2018), Barriers to digitization in university libraries of Pakistan: a developing country's perspective. Electronic Library, Vol. 36 No. 3, pp. 457-470.
- Rathee, S., Kumar, A., Kaushik, S., Kazimieras Zavadskas, E., Banaitis, A., & Garza-Reyes, J. A. (2020). An MCDA cause-effect factors model for the implementation of Greenstone Digital Library software. Management Decision, 58(11), 2543–2564.
- Lata, P. & Somvir (2014). Blog: A useful tool for dissemination. Indian Journal of Library Science and Information Technology, 4 (1), 37-40
- Rathee, S. & Kaushik, S. (2019). Steps and Planning in Setting up Digital Libraries and Repositories. Journal of Advancements in Library Sciences, 6(2): 73–76p.
- Somvir & Kaushik, S. (2021). Digital Libraries and Web 2.0. Indian Journal of Library Science and Information Technology, 6(2), 78-81