Information Retrieval Features of Text Retrieval Engines: A Case Study of Lucene

Dr. Bijan Kumar Roy
Assistant Professor
Dept. of LIS,
The University of Burdwan,
West Bengal, India
bijankumarroy@yahoo.co.in

Dr. Parthasarathi Mukhopadhyay
Associate Professor,
Dept. of LIS
University of Kalyani,
West Bengal, India
psmukhopadhyay@gmail.com

Dr. Subal Chandra Biswas
Professor
Dept. of LIS,
The University of Burdwan,
West Bengal, India
scbiswas_56@yahoo.co.in

Dr. Rajesh Das
Assistant Professor, Dept. of LIS,
The University of Burdwan,
West Bengal, India
rajeshdas99@gmail.com

Abstract

The paper provides an overview of open access repositories movement in India and highlights on the features and functionalities of some popular open source text retrieval engines viz. Lucene, Solr and Zebra used by the most popular digital library software namely DSpace, NewGenLib and Koha. Paper describes the search features supported by Lucene search engine used by DSpace (selected for BURA software framework) repository software in details and shows how it functions. Also describes advanced search facilities supported by the model with different search syntax.

Keywords: Information retrieval, Open access, Digital repository, Open source software, Search engine.

1. Introduction

For a long time, '*Information Retrieval*' (IR) has been a classical topic for information systems both traditional and digital library environment. Due to the development of the World Wide Web (WWW) and Internet, the concept '*Information Retrieval*' has been shifted to '*Web Information Retrieval*' (WIR). The open access to knowledge movement has changed the process of information generation, dissemination, retrieval and preservation and users' expectations towards these open access knowledge resources has changed. But existing traditional keyword based IR systems are not capable of handling

non-textual OA knowledge objects and unable to fulfill the information needs of the users. Even these systems do not support several advanced level retrieval features such as contents indexing, searching of index, ranking of retrieved result etc. As a result, traditional IR system has lost its importance to the academic users and is being replaced by digital information retrieval system. The development and use of different sophisticated open source search engines or text retrieval engines by open access content providers has given birth to these new online information retrieval systems viz. Open repository system, open journal system, open harvesting system etc.

This paper gives an overview of the core topics underlying full-featured modern open source and open standard based search engines supported by some open source software (OSS) and examines a number of search features and finally settled on Lucene. The paper not only outlines the features of selected search engines with particular emphasis on recent developments but also describes browsing and searching facilities available in BURA (Burdwan University Research Archive) software framework developed by DSpace (http://www.dspace.org/).

2. Growth of open access repositories

Many research institutions and universities across the world are developing open access repositories (OARs) in order to provide global access to the public funded research outputs. Till date, there are 3344 (as of June 07, 2017) repositories throughout the world (http://www.opendoar.org/) and Fig. 1 shows the growth of OARs though out the World during last 10 years.

Fig. 1: Growth of OARs (World-wide)

(Source: OpenDOAR)

Almost all the countries are now maintaining OARs (Fig. 2 & Fig. 3) and our country, India is not the exception. As a developing country, India having 80 repositories ranks 10th position in the World after Brazil (Fig. 3) and 2nd position in Asia after Japan (Fig. 4).

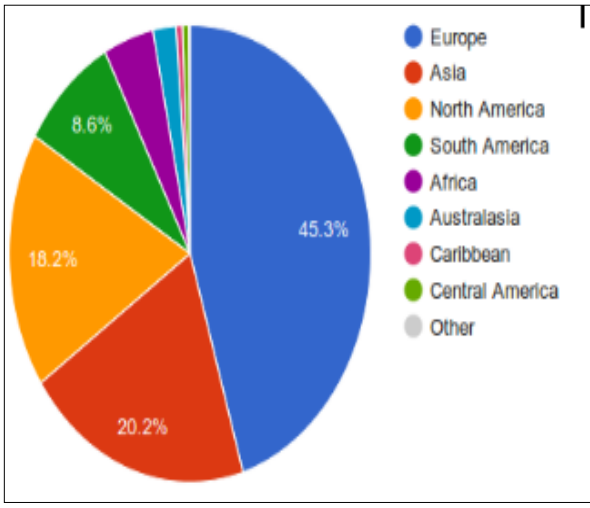

Fig. 2: Repositories by Continent-wise

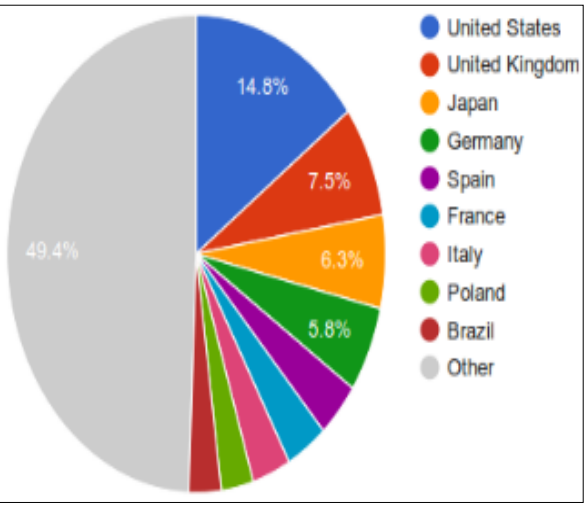

Fig. 3: Repositories by Country

(Source: OpenDOAR, 2017)

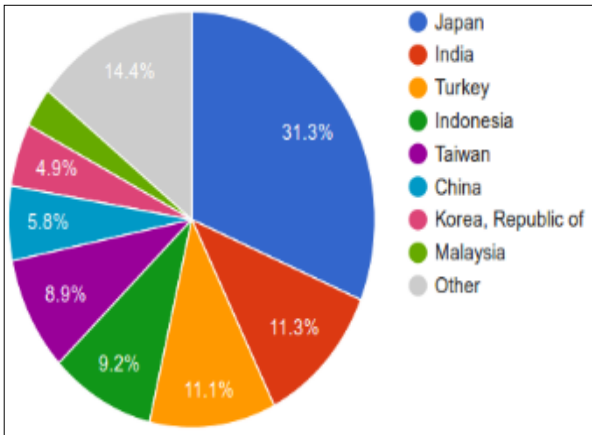

Fig. 4: Repositories by Country (Asia)
(Source: OpenDOAR, 2017)

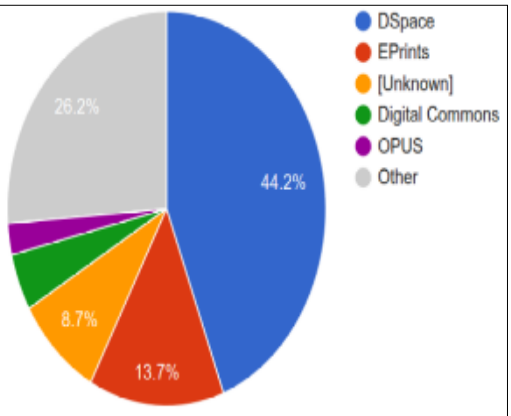

Fig. 5: Usage of Software
Source: OpenDOAR, 2017)

3. Review of Literatures

The development of OARs in India has been described by many researchers (Roy, 2007; Roy, Biswas & Mukhopadhyay, 2011; Roy, Biswas & Mukhopadhyay, 2017; Das & Chatterjee, 2015; Sengupta, 2012). Roy, Biswas & Mukhopadhyay (2012) gave an overview of current state of OARs in Asian countries with special reference to SAARC countries. In another research paper (Roy, Biswas & Mukhopadhyay, 2013), authors described OA and OARs movement in details and proposed university-specific model IDR using DSpace along with policy documentations.

Selection of software (both open source and commercial) for designing repository system is a vital issue (Roy, 2014). It is found that DSpace is the most popular software (Roy, 2015) and has the most installation from open source domain (Sourceforge, 2017). It is also used by majority of OARs registered in OpenDOAR (Fig. 5) and ROAR databases (http://roar.eprints.org/). A number of studies (Doctor, 2007; Doctor & Ramachandran, 2008; Jayakanth et al., 2008; Sutradhar, 2006; Anuradha, 2005; Shewale, 2012; Roy, Biswas, Mukhopadhyay & Das, 2017) have reported the implementation of the repository system using DSpace software in India. Another researcher study (Jayakanth, Minj & Dastidar, 2012) reported that many academic and research centers have made it mandatory to set up IDRs using OSS. Cherukodan, Santhosh Kumar & Humayoon Kabir (2013) described the design and development of a digital library at Cochin University of

Science and Technology (CUSAT) using DSpace. Another study (Vijaykumar, Murty & Khan, 2006) proposed a prototype model for Indian universities to preserve electronic theses and dissertations (ETDs). Krishnamurthy & Kemparaju (2011) reported the use of IDR in Indian universities and research institutes. In another paper, Krihnamurthy (2005) shared his practical experiences of using DSpace software.

4. Overview of Selected Search Engines

Search engines play a central role in helping users find information in digital libraries or on the Web. But the problem is that not a single search engine is capable of indexing all the content on the on-line environment**.** The following are the details of some selected search engines viz. Lucene, Solr and Zebra used by most popular digital library systems -

4.1 Lucene

Lucene was developed by Doug Cutting during 1997-98. It is Java-based open source toolkit for text indexing and searching. It is one of the projects of Apache Jakarta and is licensed under the Apache Software License (http://www.codemass.com/presentations/ 2007/luceneoverview.pdf). DSpace uses the Jakarta search engine, Lucene. Lucene search engine has very powerful search features that encompass many search approaches of the end-user. Lucene also facilitates Boolean search, range searches, term boosting and proximity searches (Prasad & Patel, 2005). Apart from the above, Lucene uses fuzzy logic which is based on the Levenstien's alogorithm that can replace and match terms by similarity (http://www. merriampark.com/ld.htm).

4.2 Solr

Solr is an open source enterprise search server. Solr written in java is the popular, blazing fast open source enterprise search platform from the Apache Lucene project. Solr is a standalone enterprise full text search engine with high performance search server with a web-service like API. Its major features include full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling.

4.3 Zebra

Zebra (http://indexdata.dk/zebra/) is a high-performance, general-purpose structured text indexing and retrieval engine. It reads records in a variety of input formats (eg. email, XML, MARC) and provides access to them through a powerful combination of Boolean search expressions and relevance-ranked free-text queries.

5. Browsing and Searching in BURA

This section describes with different screen snapshots the browsing and searching facilities supported by the model BURA (Burdwan University Research Archive) (Fig. 6). It is based on open standards and open source software (OSS) that organizes and preserves intellectual resources of an organization and provides global access to it. The browsing panel, as designed in BURA software framework, contains all the Communities, Sub-communities and networked resources placed under it.

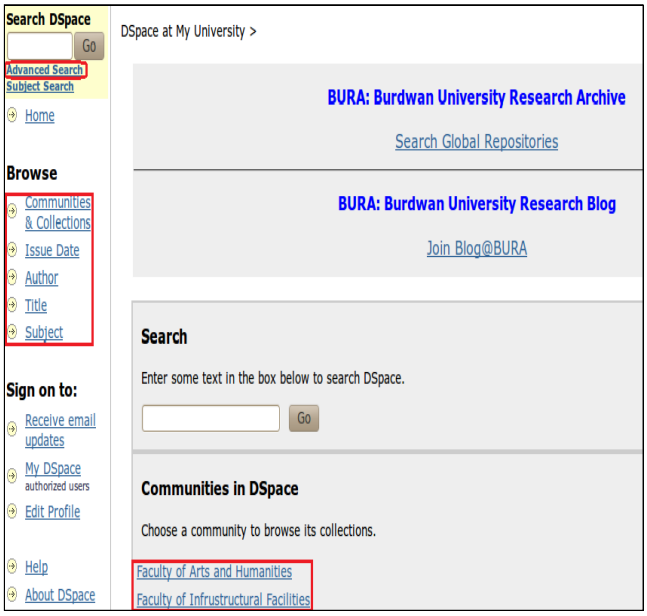

Fig. 6: Browse by Community

5.1 Search facilities in BURA

BURA supports sophisticated searching with the help of search operators (Boolean, positional and relational operators) both within the local repository and across the

repositories of multiple institutions. BURA offers by default the following search features: (i) *Search all DSpace* (Fig. 7) (ii) *Bounded search within a specified Community's Collection* (iii) *Simple search* and (iv) *Advanced search*. In the same fashion, documents can be browsed by '*Author'*, '*Title'*, '*Date'*, and '*Subject'*. Even documents can be browsed by '*Department*' and by '*Handle*' assigned to the document by the system.

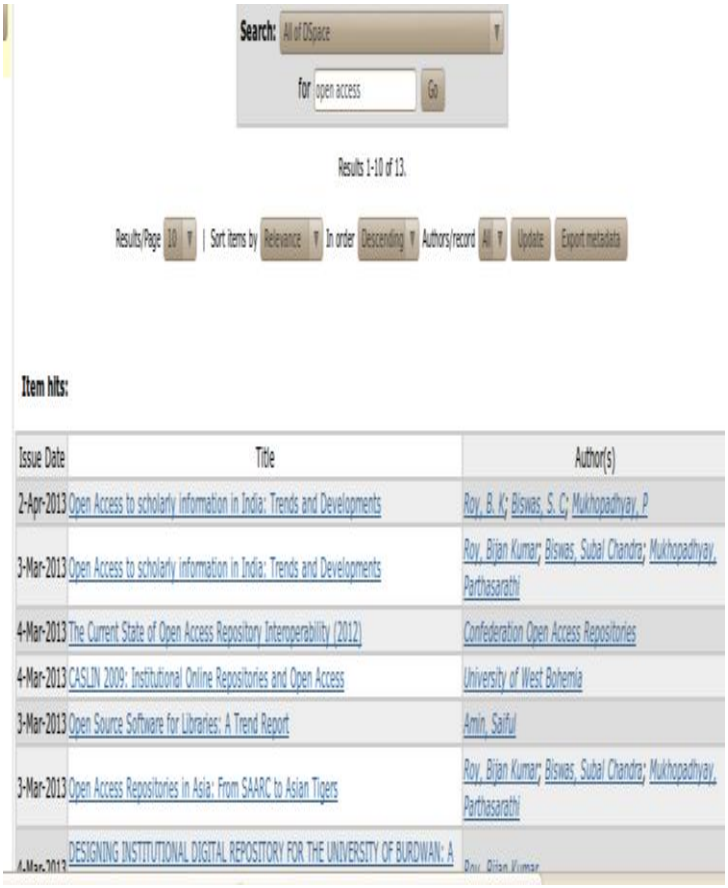

Fig. 7: Search results in all DSpace

A. Advanced Search

To navigate to the advanced search page, user can click on '*Advanced Search'* link (Fig. 6) at the top left corner of the BURA user interface. This interface (Fig. 8) allows user to specify the search fields and user can combine these searches with the Boolean operators '*AND'*, '*OR'* or '*NOT'*. This window is supported by drop down menu list from where user can pick up required value and can restrict search to a Community by clicking on the

arrow to the right of the top box. The window (Fig. 9) display the results against a search query matched (e.g. *author:bijan* or author:*biswas or author:mukhopadhyay*).

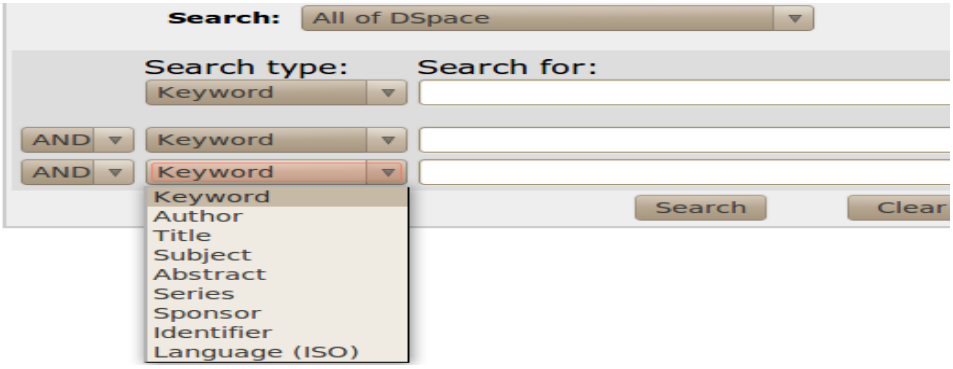

Fig. 8: Advanced Search Interface

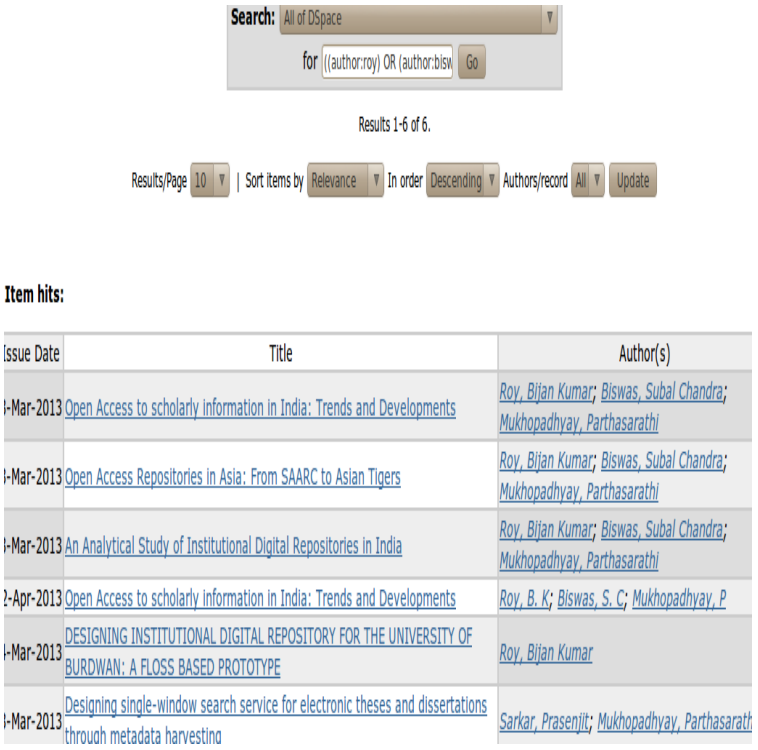

Fig. 9: Display of Results (Advanced Search)

A.1 Exact Term/Phrase Search

The search term can be a '*word*' or a '*phrase*'. One can use a search word, e.g. "*open access*" (Fig. 10) or a phrase "*open access repository*". For phrase search, the phrase should be enclosed with double quotes.

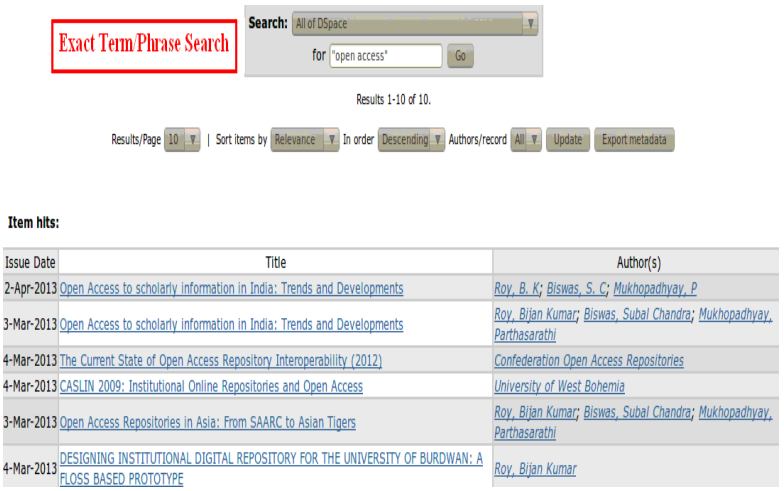

Fig. 10: Exact Term/Phrase Search

A.2 Fielded Search

It enables searching of specific field provided in the query (Fig. 11). One can search for a term in a particular field or any field by typing the field name followed by a colon ":" and then the term looking for e.g.: *author:bijan*

Fig. 11: Fielded Search by Author

A.3 Fuzzy Search

To do a fuzzy search, use the tilde symbol, "~", at the end of a single-word term (Fig. 12). To search for a term similar in spelling to "subal" use the fuzzy search: *subol~*. This search will find terms like subal. For example: *author:subol~* can match subal

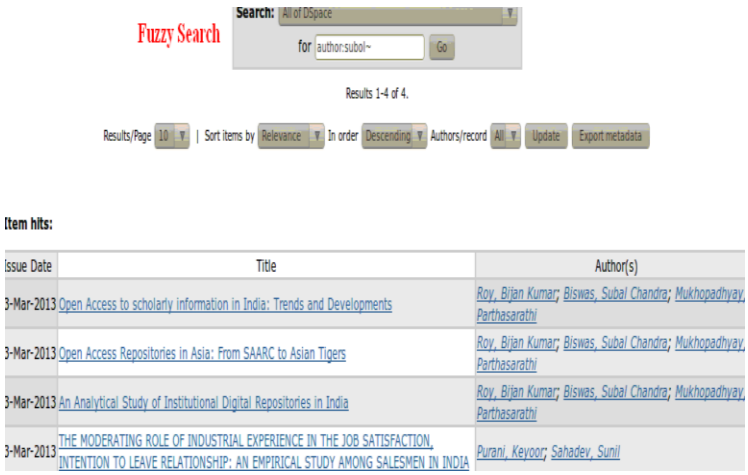

Fig. 12: Fuzzy Search

A.4 Proximity Search

Proximity search is used in a query to retrieve documents that have two words or phrases in proximity (Fig. 13). For example, *"library science"~3.* The system will retrieve records where the words 'library' and 'science' are within the three words distance.

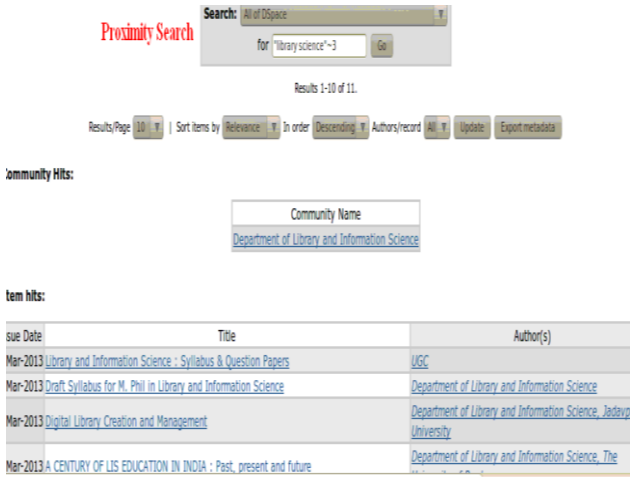

Fig. 13: Proximity Search

A.5 Range Search

Range Queries allow one to match documents whose field(s) values are between the lower and upper bound specified by the Range Query. If the search query is- *author:[rao TO rath]*. Then the system retrieves documents authored by names that fall between 'rao' and 'rath' (Fig. 14).

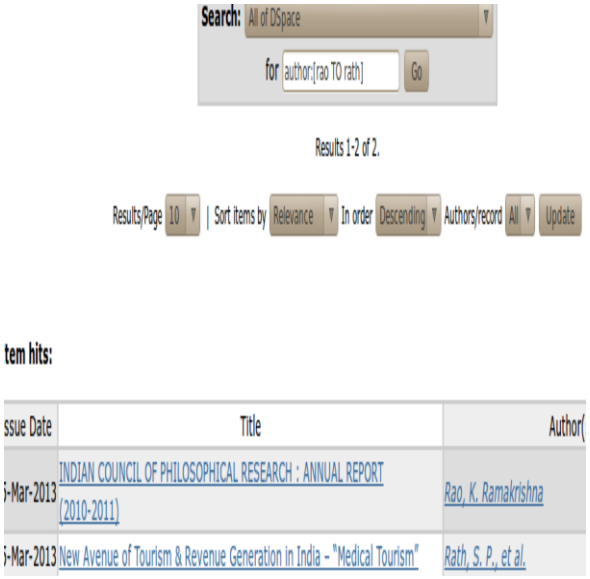

Fig. 14: Range Search by Author

A.6 Boolean Search

Boolean '*AND*', '*OR*', '*NOT*' are used for Boolean combinations. Boolean operators must be CAPITALIZED (Fig. 15).

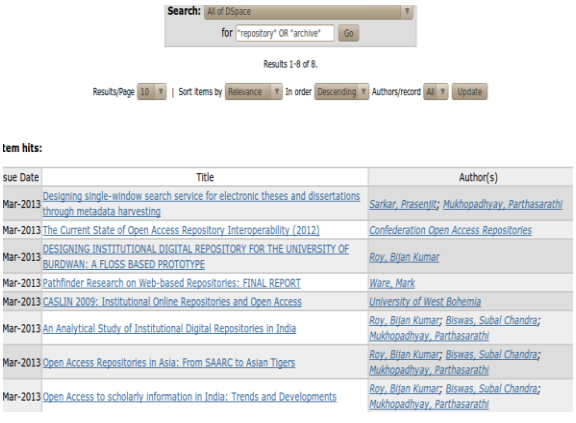

Fig. 15: Boolean Search by 'OR'

6. Multilingualism

The Indic-script based user interface is essential for any online information retrieval system in India (Roy, Biswas & Mukhopadhyay, 2016). Fig. 16 allows users browsing and searching resources in Bengali. In addition, it supports advanced searching with Boolean operators.

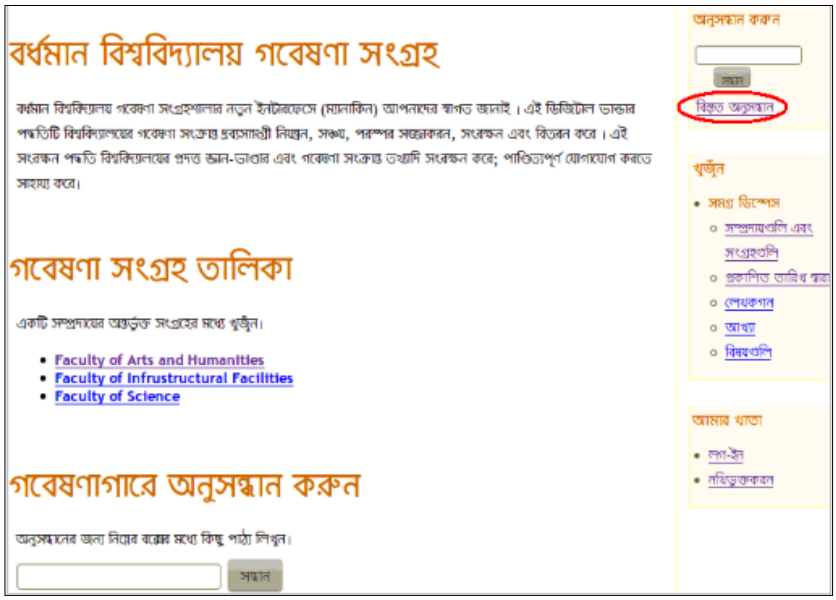

Fig. 16: Search Interface in Bengali

7. Subject Access System

Browsing and searching open knowledge resources through any subject access system is another advantage to the end users (Roy, Biswas & Mukhopadhyay, 2017). Here, users can search resources using DDC (Dewey Decimal Classification) both in English and in Bengali (Fig. 17).

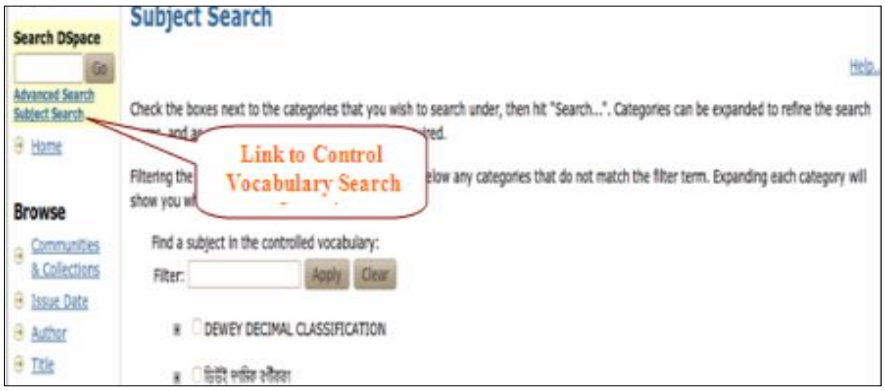

Fig. 17: Searching through DDC

6. Conclusion

Retrieval of non-textual information effectively and efficiently from heterogeneous knowledge sources has been a challenging task to the digital library developers. In this context, open source text retrieval engines could play an important role in handling non-informational objects and our professionals have achieved much success in retrieval of non-textual documents in a short time. The development and application of these licensed-free open standard based search engines in Web and digital library (DL) environment is now a genuine alternative to commercial and proprietary systems. And, the successful deployment of open source search engines claim that this open standard based search tools have matured to support many unmet uses.

References

Anuradha, K. T. (2005). Design and development of institutional repositories: A case study. *The International Information & Library Review*, 37(3), 169-178.

Cherukodan, S., Shanthos Kumar, G., & Humayoon Kabir, S. (2013). Using open source software for digital libraries: a case study of CUSAT. *Electronic Library, The*, 31(2), 217-225.

Das, D., & Chatterjee, P. (2015). Institutional Repository at Central Library IIT Kharagpur: An Overview. *International Journal of Engineering Development and Research*, 3(2), 321-328.

Doctor, G. (2007). Knowledge sharing: developing the digital repository of SIPS. *VINE: The journal of information and knowledge management systems*, 37(1), 64-73.

Doctor, G., & Ramachandran, S. (2008). DSpace@IBSA: knowledge sharing in a management institute. *VINE*, 38(1), 42–52.

Jayakanth, F., et al. (2008). ePrints@IISc: India's first and fastest growing institutional repository. *OCLC Systems & Services*, 24(1), 59-70

Jayakanth, F., Minj, F., & Dastidar, P. G. (2012). Setting up an open access digital repository: A case study. *Annals of Library and Information Studies*, 59(1), 16-24.

Krishnamurthy, M. (2005). Digital library of mathematics using DSpace: a practical experience. *SRELS Journal of Information Management*, 42(3), 245-256.

Krishnamurthy, M., & Kemparaju, T. D. (2011). Institutional repositories in Indian universities and research institutes: A study. *Program: electronic library and information systems*, 45(2), 185-198.

Prasad, A.R.D., & Patel, D. (2005). Lucene Search Engine: An overview. In A.R.D.Prasad (Ed.), *International Workshop on Building Digital Libraries using DSpace* (March 7th – 11th, 2005, Bangalore) (Paper I). Bangalore : DRTC.

Roy, B. K. (2007). Indian Initiatives in the Development of Institutional Digital Repository. *Digital Media and library Information Services, Proceedings of 26th IASLIC Conference* (December 26-27, 2007, Jamia Millia Islamia University, New Delhi) (pp. 253-262). Kolkata: IASLIC.

Roy, B. K. (2014). *Designing Institutional Digital Repository for the University of Burdwan: A FLOSS Based Prototype*. A PhD Thesis, Department of Library and Information Science, The University of Burdwan, Burdwan.

Roy, B. K. (2015). *Institutional digital repository: from policy to practice*. LAP; Saarbrücken, Germany.

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2011). An Analytical Study of Institutional Digital Repositories in India. *Library Philosophy and Practice.* Paper - 692. Retrieved November 3, 2016, from http://digitalcommons.unl.edu/libphilprac/692

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2012). Open Access Repositories in Asia: From SAARC to Asian Tigers. *Library Philosophy and Practice.* Paper - 808. Retrieved December 12, 2016, from http://digitalcommons.unl.edu/libphilprac/808

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2013). Global visibility of Indian Open Access Institutional Digital Repositories. *International Research: Journal of Library & Information Science*, 3(1), 182-194.

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2016). Open access repositories for Indian universities: towards a multilingual framework. *IASLIC Bulletin*, 61(4), 150-161.

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2017). BURA: An Open Access Multilingual Information Retrieval and Representation System for Indian Higher Education and Research Institutions. *Library Philosophy and Practice* (e-journal). Paper- 1541. Retrieved November 3, 2017, from http://digitalcommons.unl.edu/libphilprac/1541

Roy, B. K., Biswas, S. C., Mukhopadhyay, P., & Das, R. (2017). Developing Open Access Institutional Digital Repository Using Open Source Software: A Step by Step Guide. *International Research: Journal of Library & Information Science*, 7(2), 244-257.

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2017). DDC in DSpace:

Integration of Multi-lingual Subject Access System in Institutional Digital Repositories. *International Journal of Knowledge Content Development and Technology* [Mss. Submitted].

Shewale, N. (2012). Building Digital Library using DSpace: Case Study of GIPE's Dhananjayarao Gadgil Digital Library. *DESIDOC Journal of Library & Information Technology*, 32(5), 417-420.

Sutradhar, B. (2006). Design and development of an institutional repository at the Indian Institute of Technology Kharagpur. *Program: Electronic Library and Information Systems*, 40(3), 244-255.

Sengupta, S. (2012). Status of E-Theses Repositories with Special Reference to India. *Library Philosophy and Practice.* Paper – 764. Retrieved May 25, 2016, from http://digitalcommons.unl.edu/libphilprac/764

Sourceforge. (2017). *Download statistics of software in the archive*. Retrieved July 12, 2017, from http://sourceforge.net/

Vijaykumar, J. K., Murty, T. A.V., & Khan, M. T. M. (2006). Experimenting with a Model Digital Library of ETDs for Indian Universities Using DSpace. *Library Philosophy and Practice,* 9(1), 1-17. Retrieved March 12, 2016, http://digitalcommons.unl.edu/libphilprac/105/