A personalised discovery service using Google custom search engine

Arup Dhara

Librarian, Ramsaday College, Amta, Howrah, West Bengal – 711401, Email: arupdhara1979@gmail.com

This paper provides an overview of Google custom search engine (CSE) and discusses some of the possibilities and methodologies regarding how these technologies and products might be helpful in developing college library discovery systems without deploying costly commercial tools or without venturing into the technicalities of open source tools. It is based on the experiences of designing a small-scale personalised discovery service for the Ramsaday College Library in Howrah, West Bengal and its integration with the library OPAC as one stop access to an array of library resources.

Introduction

Search engines - particularly Google - is usually the starting point for anyone who is looking for information on the Internet. Apart from general search, the search engine also offers advance search features which enable retrieving required information precisely on a given topic (e.g. allintitle: "environmental pollution" filetype:pdf to retrieve only those documents where the phrase environmental pollution appears in the titles of the documents available in pdf format). However, most of the end users do not apply special keywords (as provided by search tools and as shown in previous example) and a generic query fetches huge number of results, many of that are undesirable. Sometimes search engines produces unrelated results like a search on 'computer driver' may retrieve 'screw driver' and Google is no exception to this cross-disciplinary semantic drift. In order to minimize inconvenience of researchers, new age discovery tools such as Summon, Knimbus etc from commercial domain and Blacklight, VuFind etc from open source domain have emerged. These tools limit the patron's search only to the resources which are available within the library and subscribed by the library.[1] However, costs of such tools are enormous for a small college library. Open source discovery tools are available at no cost (like VuFind) but the deployment of such tools require high degree of technical knowledge. An alternative to costly commercial discovery services/tools or technically complex open source discovery software (for which local expertise may not be available) may be custom search engines. There are some open ended custom search engines like Rollyo, Yahoo search builder, Google Custom Search Engine (CSE) etc.

This study has applied Google CSE for three reasons – i) simplicity of the application and option for generating Widget (a combination of HTML and Java script); ii) no restrictions in inclusion of number of websites; and iii) extended support for special keywords as available in google Generic[2,9].

A brief review of literature and overview of services established this study into its proper perspectives. Gogoi et al opined that federated search is essential for rapid growth of scholarly literature and discussed the concept of federated search, different federated search engines, features of the search engines and search strategies of some of the federated search engines[2]. Jain and Saraf commented that Internet search engines are tools to retrieve the information from web. Google search engine is the most popular among all types of users, especially academic users[3]. Craven opined that federated search saves valuable time by bridging the gap between 'searching' and 'acquiring' relevant information in a timely fashion[4]. Enge and Biundo stated that Google CSE enables web site owners to define their own search engines and that people have the right to

determine which sites/URLs (including flexible URL patterns) are searched, and to define a set of rules that guide the ranking of results[5]. Kumar, Sanaman and Rai discussed about federated search engines and their usefulness for libraries in searching e-resources from one search box. They also discussed various federated search providers along with their advantages and drawbacks[6]. Perdananugraha argued that utilization of Google custom search in digital library has given another option in knowledge finder/tools to explore the information among hundreds, thousands, even millions or more collections that the digital library has[7]. Google Custom Search Engine is an online platform that enables one to create a search engine for a website, blog, or a collection of websites[8]. One can configure the engine to search both pages and images and can customize the appearance of the search results. This tool enables web searchers to refine and categorise scholarly information. Google launched this service on May 10, 2006[8]. Mukhopadhyay and Mondal reported that the unique feature of Google CSE to generate Widget may lead to possiblity of its integration in library search systems like OPAC. The authors demonstrated the process of integrating Widget generated by Google CSE on OA journals on LIS (https://cse.google.com/cse/publicurl?cx=-015294161869215487837: zpt9quza5by) in Koha OPAC through the application of widget[9].

All the above literatures have described its usefulness to explore the effective information from a large collection but none of them had described the technicalities involved in the design of a customised search engine using Google CSE. This case study demonstrates how Google CSE has been implemented in Ramsaday College Library, Howrah, West Bengal and also discussed its subsequent integration with library OPAC. The broad range objective of this study is to explore the process of implementing a cost-effective customized discovery service for college library users and its integration with library OPAC.

Methodology

A case study approach was used, with detailed analysis and failure checking of results. The performance of the system was assessed in terms of coverage, precision, quality and accessibility of results. A study of the documents retrieved was carried out and the reasons for the results which are not being retrieved from the queries were analyzed. A

report on the system development and customization has also been prepared for future reference.

Google CSE@Ramsaday College

Google Custom Search Engine is an online platform that allows to create a search engine for a website, blog, or a collection of websites. Anticipating the needs of the library users, this prototype has been developed as a cost-effective single window search platform to retrieve subscribed resources as well as open access resources available on web from a single-window search interface.

Once the search engine is ready, it can be accessed in two ways – via direct URL like https://cse.google.com/cse/publicurl?cx=01364923458 7080269501:zmze6gjpot or one can embed a search box on the webpage by pasting the code in the source code of the site's HTML where search box is to appear.

Ramsaday College is one of the oldest undergraduate colleges in Howrah district. It offers 23 subjects in Honours and general degree courses in arts, science and commerce streams. Approximately 5500 students have registered in different courses and most of them are using library on regular basis. Therefore, resource identification is one of the most important function during design of a discovery service. The Ramasday College Library has a bibliographic database in Koha with around 24000 records and it subscribes to NLIST package. Apart from these resources, users need to consult different abstracting and indexing databases and open access resources available in their domain of interests. But these resources are distributed in library catalogue database, subscribed resources and Open Access resources each with different user interface and specific sets of retrieval techniques. As a result, end users are suffering in retrieval of resources as they need to move from one search interface to another for comprehensive resource discovery. This situation calls for an immediate action by library in providing a single-window resource discovery system for end users. In response to this emerging demands, library explored the possibilities of implementing discovery services for the end users to help in retrieval. But commercial discovery services were too expensive for the already-strained library budget. On the other hand, library does not have technical manpower to implement freely available open source discovery tools like VuFind or Blacklight in view of the

complex installation process and convoluted configuration steps. Therefore, the college library attempted to use Google Custom Search Engine as a viable alternative. The resource base of the CSE developed by the college library includes an array of subscribed, free and Open Access resources. The web addresses concerning different publishers frequently sought by faculty, staff and student members are given in Table 1.

Using the interface of CSE, a basic discovery service was created and named as *Search Databases@Ramsaday College Library* (see Fig. 1) by filling-up minimum required fields like 'Site to search' option (selected URLs were added to activate the CSE e.g. http://nlist.inflibnet.ac.in/*). 'Name of the Search Engine' (as mentioned above) and description of the CSE (given as Searche My Databases). However, other options such as 'Search Engine Keywords', 'Edition', 'Language' etc. were left to their default settings. While adding site's URL, it is possible to configure the URL patterns that can enable CSE to include total or parts(s) of the website. For example, adding '/*' at the end of web address (URL) will enable CSE to include all the pages and subfolders derived from that particular web address. Use of the wildcards (* mark) in URL patterns allows access particular page or portion of the website. In the creation of the current CSE, '/*' at the end of each web address was added to cover whole website for indexing by the Google CSE developed by the college.

The Google CSE as developed by the college by following the above steps may be fine tuned in terms of search efficiency and appearenace. An overview of control panel is given to povide an insight about customization (Fig. 2). The major features that can be controlled from the control panel are as follows:

Setup: This feature provides basic details and preferences for the newly developed CSE. In the Basics tab of the Setup section, one can change the name of the search engine, specify sites to be included and excluded, enable image search, select the Custom Search edition (free CSE or paid Google Site Search) that one wants to use and set the language of search engine interface.

Look and feel: One can customize the look and feel of search engine on the look and feel page of the 'Control Panel'. On the 'Layout' tab, one can specify the layout of search results page; on the 'Themes' tab, one can select a design theme for the search engine; on the 'Customize' tab, one can specify exactly how one wants to set the CSE elements such as the search box, search button, refinements, and results to appear; on the 'Thumbnails' tab, one can enable or disable thumbnail images in search results.

Search features: By enabling these features, one can make search more effective, such as promotions, refinements, auto-Complete, synonyms and many such things.

Statistics and logs: Statistics about Search Engine usage can be integrated with Google analytics for more detailed reporting. This information can be gathered by day, week, month or year – even the entire lifetime of the CSE. The 'Statistics' page provides accurate data about the search usage dataset and the total number of searches completed.

Table 1—Resources added in the CSE

- NLIST: http://nlist.inflibnet.ac.in/
- Directory of Open Access Journals (DOAJ): https://doaj.org/
- Journals of Indian Academy of Sciences: http://www.ias.ac.in/
- JSTOR (up to abstract level): http://www.jstor.org/
- ACM Digital Library: http://dl.acm.org/
- Science Direct (up to abstract level): http://www.sciencedirect.com/
- Wiley Online Library (up to abstract level): http://onlinelibrary.wiley.com/
- IEEE: http://ieeexplore.ieee.org/Xplore/home.jsp/
- Project Gutenberg: https://www.gutenberg.org/
- DOAB: http://www.doabooks.org/
- Sodhganga: http://shodhganga.inflibnet.ac.in/
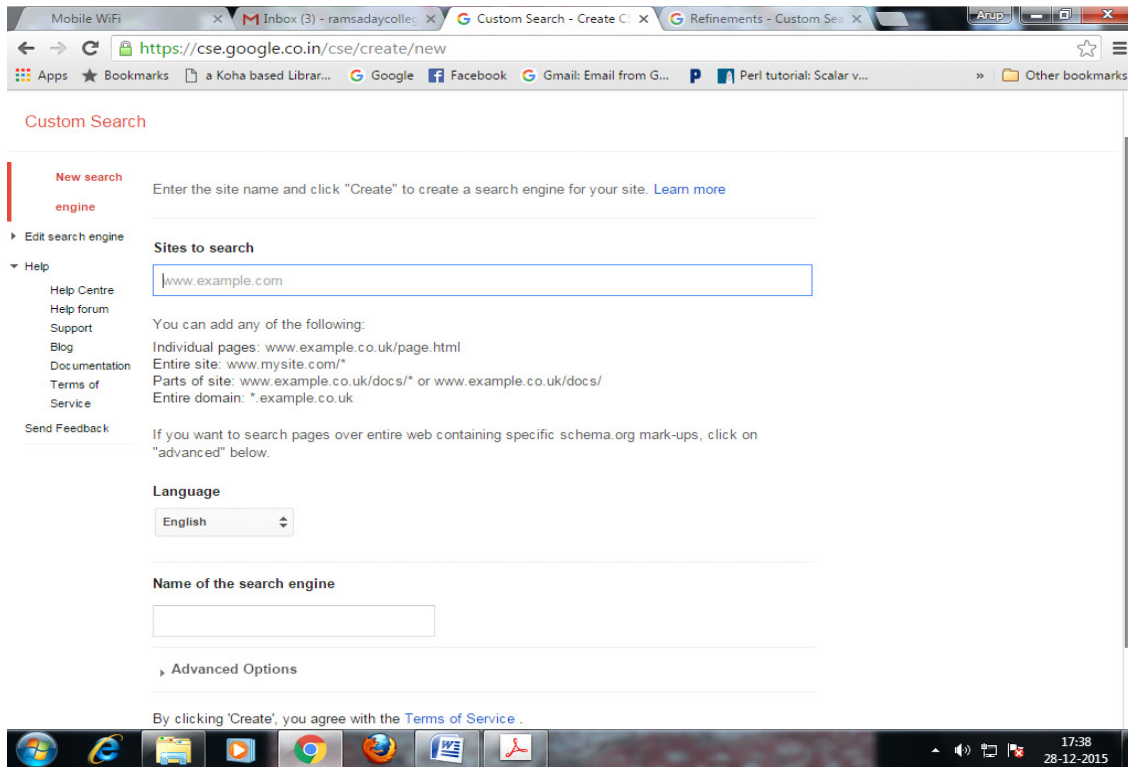- Nature: http://www.nature.com

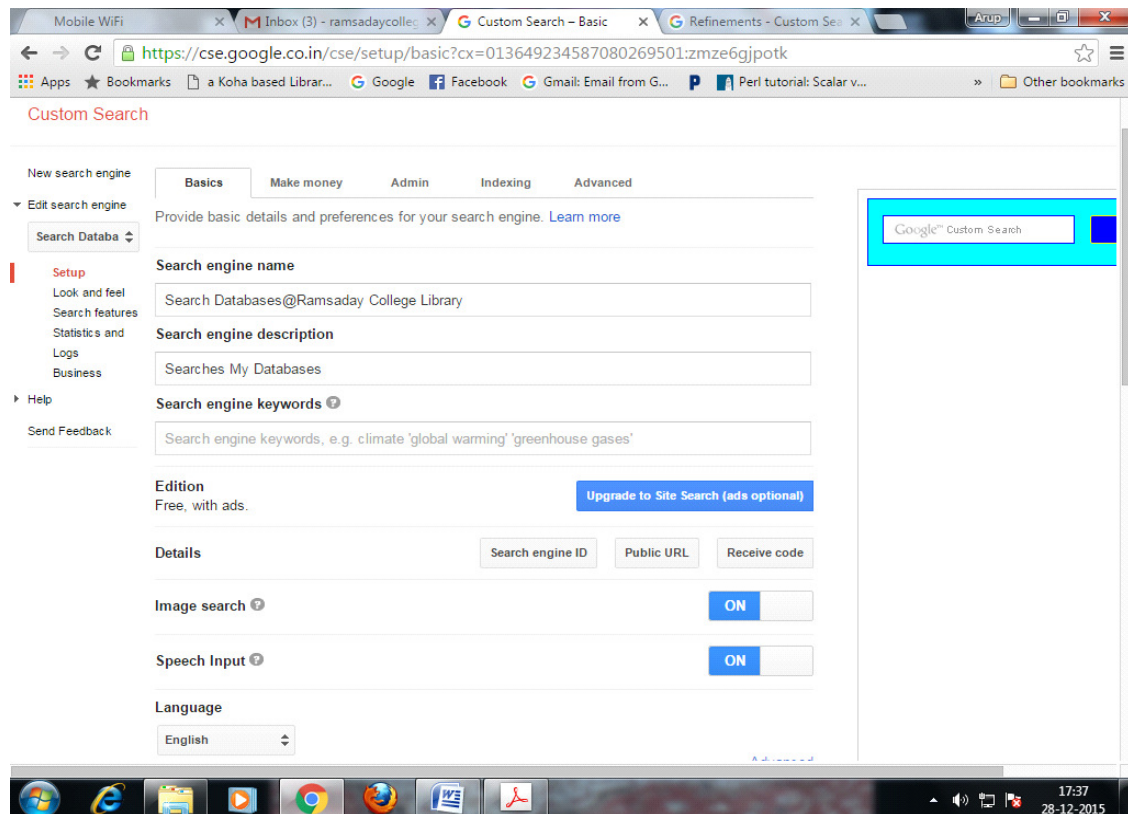Fig. 1—Google CSE interface

Fig. 2—Google CSE control panel

Customization of CSE

The most importat feature of customization of CSE is related with serach refinements. Refinements are labels that are applied to web sites in the search engine index to help users to find the information according to sources. In the Ramsaday College Library CSE, 'labels' such as NLIST, JSTOR, ScienceDirect, E-books, E-journals, Wiley Online etc (as listed in Table 1) were added (Fig. 3) for helping end users to filter results accoring to their favourite sources. The 'Manage labels' option from the Label tab of Setup section can also be used for this purpose. While adding labels, 'Search only the sites with this label' option was chosen so that searches can be restricted only in specific URL(s).

Unlike search engine labels, refinement labels are visible to users, and appear at the top of the search results page. When a user clicks a refinement label, site(s) that were annotated with that label are searched exclusively and results retrieved and displayed in the result page.

The next important section is the 'Look and feel' settings to tune the appearance and the layout of CSE. Here 'Full-Width' layout and 'default' theme were selected for layout of search result page and design of search engine respectively. On the Customize tab, the default values of font and colour of different CSE elements such as the search box, search button,

refinements, and results to customize their appearance were changed to suitable values to get enhanced visual effects. This section also provides 'Save and Get Code' option to generate Widget for further integration of the CSE in web services like College Website, library OPAC etc. Another important setting is auto complete option configuration. This option was enabled to help end users to see lists of useful queries as they start typing in the search box.

Now the Custom Search Engine, Search Databases@Ramsaday College Library is ready to use. The Search Engine can be accessed either through Public URL or from search box embedded in the website (the Receive code button in the setup page can generate the Widget code of the Google CSE). The Public URL of the CSE (https://cse.google.com:443/cse/publicurl?cx=013649 234587080269501:zmze6gjpotk) has been shared with users through library website. Appearances of search box and search result page have been depicted in Figs. 4 and Fig. 5 respectively.

Google CSE in Library OPAC

Apart from accessing Google CSE as developed by the college library in the Web as individual service, the library used the Widget facility of Google CSE to include the search box in library OPAC by following the methodology as provided by Mukhopadhyay and
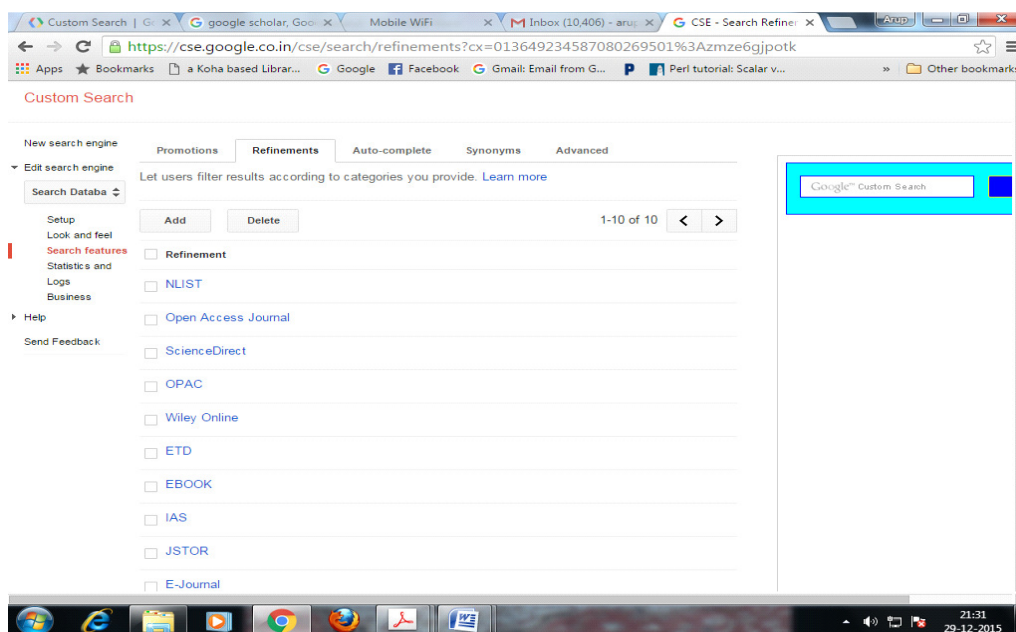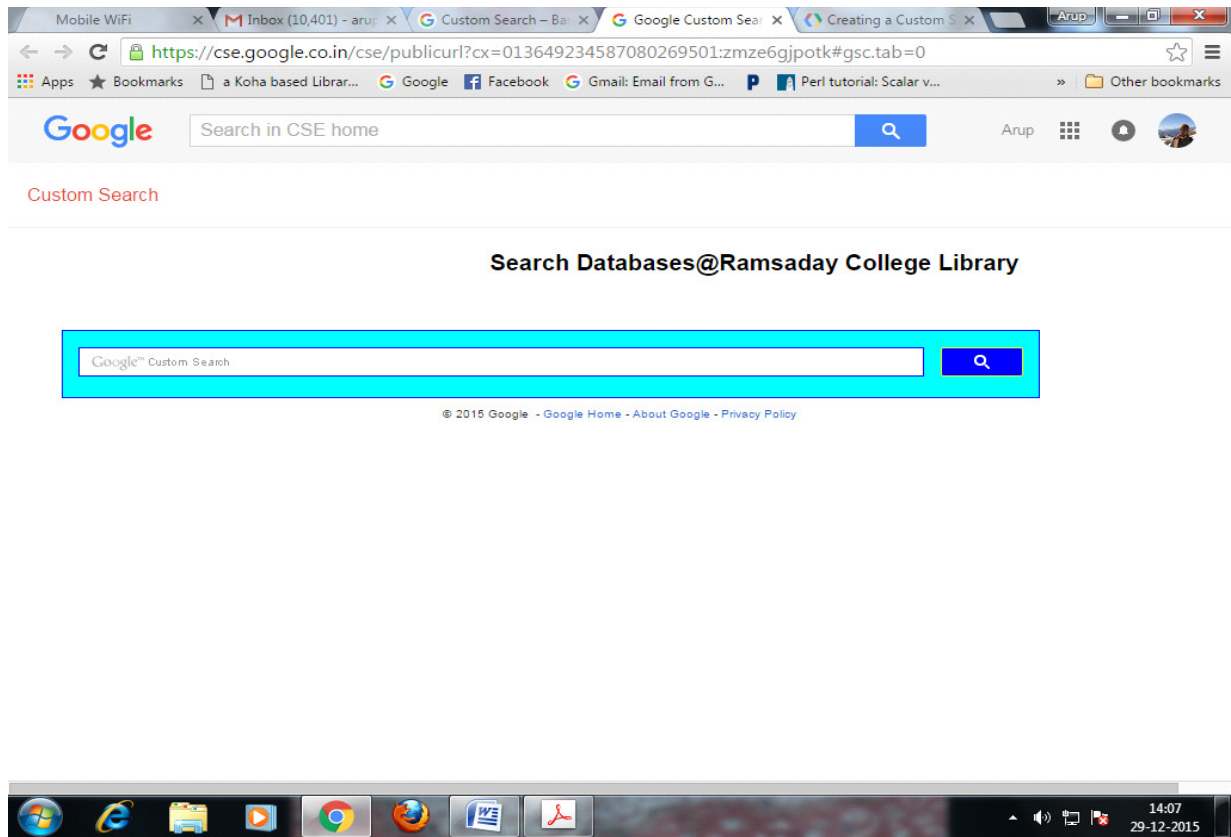

Fig. 3—Refinement Labels

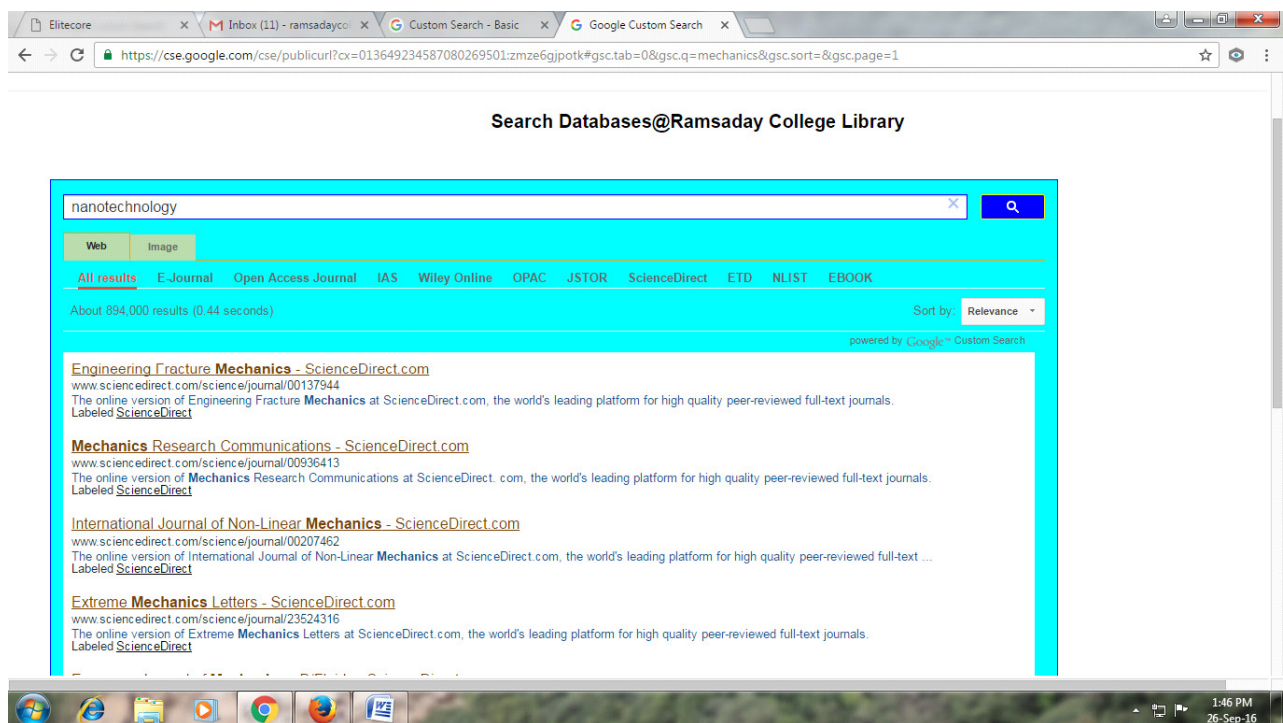Fig. 4—Interface of Search Databases@Ramsaday College Library

Fig. 5—Appearance of search result pag

Mondal way back in 2011[9]. The steps for integrating search box in Koha OPAC are simple and straightforward and mainly a two-step process – generate the code for Widget (a combination of HTML and Javascript) and inclusion of the code in 'OpacMainUserBlock' in Koha system preference (Fig. 6).

This integration produces a search box in Koha OPAC from where users can search the resource bases that are included in the Google CSE. It provides a user-friendly search environment in where users can switch from OPAC search to discovery and vice-versa (Fig. 7). The advantages are manifold like – single-window search, one stop access, metadata search and full-text search in one interface (along with the special keywords provision of Google), simple to use and thereby requires no user training and most importantly an instant hit in new generation students who are acquainted with basic web searching in Google.

Limitations

Google CSE is an excellent free service for developing personalized search engines but at the same time it is not free from limitations. First of all, the CSE will not able to search the site if it is not indexed by Google. Secondly, in case of subscribed databases, search for whole site have to be specified. As a result, the non-subscribed journals in a journal portal can be indexed but the access to full-text for those journals won't be possible although the results will be included in retrieved set of results. As it is a free edition, another problem is the future availability of the Google CSE facility (as free tool). For example, Google Reader, a free RSS feed aggregator service served users from October, 2005 but discontinued by Google Inc in July 2013.

Conclusion

The present scenario of library discovery systems is characterized by two extremes - big tax – small task zone dominated by commercial discovery tools (like Summon, EBSCO, ExLibris, WorldCat Local etc) and No tax – Big task zone dominated by open source discovery tools (like VuFind, Blacklight etc). Academic libraries require discovery systems to manage vast array of diversified resources distributed from local level to global services. Naturally, users' demands for single window search service are increasing due to the growth of different types of databases in a library. College libraries in India are presently providing different search interfaces with different sets of retrieval techniques. As a result, users are suffering a lot as far as search service of the libraries are concerned. College libraries in the state of West Bengal are no exceptions and are striving hard to cope with this problem and looking forward to implement discovery tool as meta search service.

```
<hr>
<center>Search Databases@Ramsaday College Library</center>
<script>
  (function() {
    var cx = '013649234587080269501:zmze6gjpotk';
    var gcse = document.createElement('script');
    gcse.type = 'text/javascript';
    gcse.async = true;
    gcse.src = 'https://cse.google.com/cse.js?cx='+ cx;
    var s = document.getElementsByTagName('script')[0];
    s.parentNode.insertBefore(gcse, s);
  })();
</script>
<gcse:search></gcse:search>
<hr>
```

OpacMainUserBlock   Show the following HTML in its own column on the main page of the OPAC:

```
<hr>
<center>Search Databases@Ramsaday College
Library</center>
<script>
  (function() {
    var cx = '013649234587080269501:zmze6gjpotk';
    var gcse = document.createElement('script');
    gcse.type = 'text/javascript';
    gcse.async = true;
    gcse.src = 'https://cse.google.com/cse.js?cx=' +
cx;
    var s = document.getElementsByTagName('script')[0];
    s.parentNode.insertBefore(gcse, s);
  })();
</script>
```

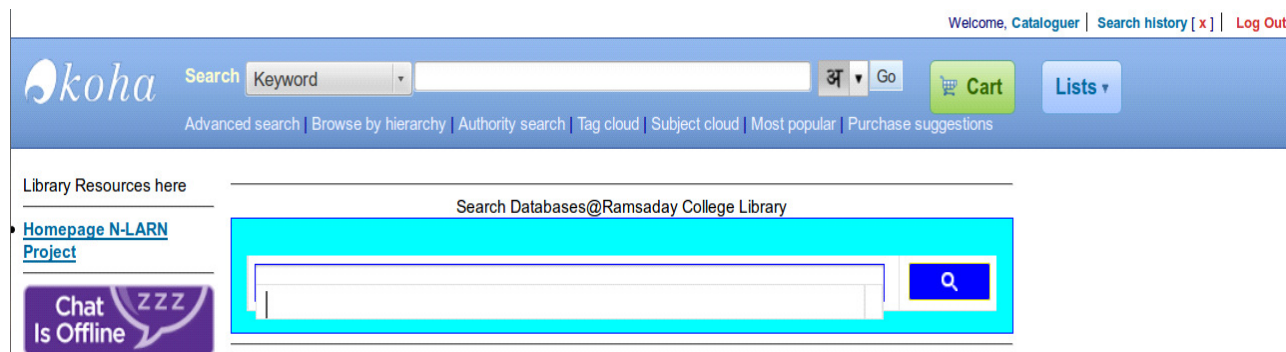Fig. 6—Widget and its integration in Koha

Fig. 7—Discovery interface in library OPAC

However, many college libraries find it difficult to subscribe or implement commercial discovery tools due to financial constraints. On the other hand, open source discovery tools like VuFind, Blacklight etc are available for no cost implementation but complex installation and configuration, command based harvesting of resources and many such issues are too technical to implement in a college library. Here, Google CSE fits-in well , as it is zero-cost and easy-to-implement solution. This economic and time-saving application is very useful for librarians who have to deal with overwhelming information resources. Library professionals need to put in efforts in using such new age tools ignoring some of its inherent limitations.

References

1. Gavali R, Discovery Service for Engineering and Technology Literature through Google Custom Search: A Case Study, DESIDOC Journal of Library & Information Technology, 35(6) (2015) 417-421.

2. Gogoi K, Barthakur J and Sarmah M, Federated Search: An Information Retrieval Strategy for Scholarly Literature. In Proceedings of the paper presented at International CALIBER-2008. Sikkim March 1-3, 2012, p. 224-232.

3. Jain V and Saraf S, Google search engine and its usefulness to library professionals, DESIDOC Bulletin of Information Technology, 26(5) (2006) 23-25.

4. Craven V, Federated Search Engines –the Unified Search Environment: The Western Health Library Experience. In Proceedings of the paper presented at 13th Exhibition and Conference, Sydney January 30 – February 1, 2007, p. 19-23.

5. Enge E and Biundo J, Google Custom Search Engines (Google CSEs). Available at: https://www.stonetemple.com/-google-custom-search-engines-google-cses/(Accessed on 24 December, 2015).

6. Kumar S, Sanaman G and Rai N, Federated search: New option for libraries in the digital era. In Proceedings of the paper presented at International CALIBER-2008. Ahmedabad February 28 – March 1, 2008, p. 267-285.

7. Perdananugraha G M, Utilization of Google's Custom Search in Knowledge Finding on Digital Library, International Journal of Engineering & Computer Science, 13(02) (2013), p. 24-30.

8. Google Developer Documentation. Available at: https://developers.google.com/custom-search/ (Accessed on 28 December, 2015).

9. Mukhopadhyay, Parthasarathi and Mondal, S. Custom search engine (CSE) in library OPAC: a methodology for integrating Google CSE in Koha OPAC, Information and Knowldge Dissemination : Present Status and Future Direction (IKD 2011), Allied Publishers Pvt. Ltd., 2012, pp. 139-148.