IFLA WLIC 2015 CAPE TOWN

Submitted on: 19.06.2015

Developing ArXivSI to Help Scientists to Explore the Research Papers in ArXiv

Zhixiong Zhang
National Science Library, Chinese Academy of Sciences, Beijing, China.
E-mail address: zhangzhx@mail.las.ac.cn

Li Qian
National Science Library, Chinese Academy of Sciences, Beijing, China.
E-mail address: qianl@mail.las.ac.cn

Hongbo Shi
National Science Library, Chinese Academy of Sciences, Beijing, China.
E-mail address: shihb@mail.las.ac.cn

Abstract:

arXiv is an important e-print service in the fields of physics, mathematics, computer science etc. It becomes a driving force in scientific communication in those fields. But there are some weaknesses in search function and user interface provided by arXiv platform. The feature widely criticized is that arXiv search platform cannot help user to explore valuable research papers housed in arXiv in a convenient manner. ArXiv search platform just provides a simple search result set for user. It does not help user to "analyze" what is included in the result set and "reveal" some clues for user to better understand and use the search result.

In order to help scientists efficiently discovery information they wanted housed in arXiv, the authors developed a system named arXiv Search Interface (arXivSI)[1] with the help of CUL (Cornell University Library). arXivSI is an external search system to index the metadata of papers stored in arXiv. The most obvious difference between arXivSI and arXiv search platform is that arXivSI provides a more suitable search functions with faceted search features and arXivSI provides a navigable and exploratory interface with visualization analysis features. It can help scientists to explore the research papers stored in arXiv.

After arXivSI service is released publicly, it is widely used by scientists from more than 100 countries.

[1] http://arxivsi.las.ac.cn

Keywords: arXiv; arXiv Search Interface; arXivSI; Faceted Search; Knowledge Exploration

1. Introduction

arXiv is an e-print service hosted by the Cornell University Library, houses e-prints in the fields of physics, mathematics, computer science etc. Now it becomes a driving force in scientific communication in those fields. Every day, it draws in thousands of researchers to see the latest developments in their fields.

According to the survey of arXiv usage in China [1], the authors find some weaknesses in search function and user interface provided by arXiv platform. The feature widely criticized is that arXiv search platform cannot help user to explore valuable research papers housed in arXiv in a proper and convenient manner. In fact, arXiv search platform just provides a simple search result set for user, so it does not help user to "analyze" what is included in the result set and "reveal" some clues for user to better understand and use the search result.

The weaknesses of search function provided by arXiv platform include:

(1) No filtering aid is provided to filter the search result. ArXiv search platform just provides a flat search result to user, no filter aid to help the user to specify the search query and refine the search result.

(2) No sort aid is provided to sort the search result in different ways. That means user cannot sort the search result according to title, submission time, subjects etc.

(3) No navigation tool is provided to help user better understand and use the search result. ArXiv search platform just provides a flat search result for user. The users have to preview their search results page by page to find what they need and cannot get a specific subset result by authors, subjects, submission dates or other aspects.

(4) arXiv platform does not display more than 1000 hits in one query. While arXiv now houses more than one million papers, the arXiv platform does not display more than 1000 hits in one query, which means many valuable research papers are hidden by using arXiv search platform and user cannot fully explore the research papers in arXiv.

Till now, there are two other arXiv search interfaces have been developed to search arXiv papers, respectively INSPIRE [2] and NASA Astrophysics Data System (ADS) [3]. Some of the weaknesses have been improved in those search interfaces. But both of them don't provide a satisfied exploratory function to search and explore the arXiv search results.

In order to help scientists efficiently discover the knowledge they need housed in arXiv, with the help of CUL (Cornell University Library), the authors developed a system named arXiv Search Interface (arXivSI), which can help scientists to explore the arXiv search result in a more vivid way. This paper discusses the main ideas and the system implements of the arXivSI.

2. Main Ideas of ArXivSI

The main idea behind the arXivSI is trying to develop an exploratory system to turn the arXiv search experience from information retrieval to knowledge exploration. To be an exploratory system, arXivSI should analyze the arXiv search results, reveal useful patterns hidden in those results, visualize those patterns in a vivid way and provide an user-friended interface to help user explore the papers in arXiv easily and smoothly.

2.1 Overall Architecture

To develop the arXivSI, an overall architecture of the system is proposed. In this architecture, an automatic data harvester is built to harvest the metadata from the arXiv repository by using OAI-PMH [4] protocol. The harvested metadata is stored at one local repository and an automatic indexer is built to index the local metadata. By using Apache Solr [5] enterprise search platform, a suitable search platform with faceted search features is built. On top of the search platform, the authors develop the arXiv search Interface which provides exploratory functions with visualization analysis features. The data of arXivSI is powered by arXiv and synchronized with arXiv daily. As illustrated in Fig. 1.

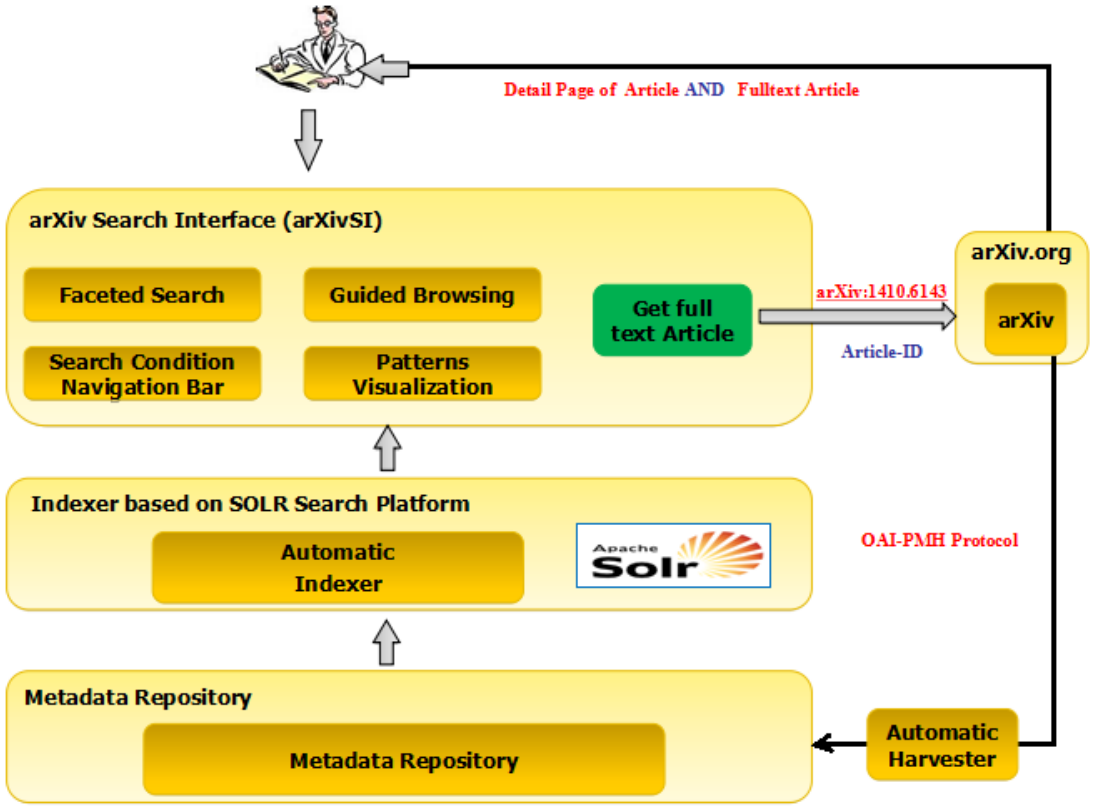

Fig. 1 Overall Architecture of arXivSI

Main processes of the arXivSI as follows:

(1) Metadata harvesting from arXiv based on OAI-PMH protocol. By using automatic harvester, arXivSI harvests metadata in arXiv format from arXiv OAI-PMH data-provider. Then arXivSI parses those metadata and stores the metadata in local metadata repository.

(2) Metadata indexing by Solr. An automatic metadata indexer is built based on the Solr software. It can index the harvested metadata efficiently and timely by the mechanism of Solr incremental indexing.

(3) Construct Search Interface. Based on Solr, arXivSI provides a suitable search functions with faceted search features. Furthermore, visualization analysis of search result is provided.

(4) Obtain full text article from arXiv. By using arXiv Article-ID in metadata, arXivSI can create a hyperlink for each article which can redirect user back to arXiv.org to get the detailed information about the article and download the article.

2.2 From Information Retrieval to Knowledge Exploration

In developing arXivSI, the authors are trying to turn the arXiv search experience from information retrieval to knowledge exploration. Instead of developing information retrieval services for user, which are the focus of current arXiv search platforms (including arXiv.org, INSPIRE & ADS), the authors propose to develop knowledge exploration services for user to explore research papers in arXiv. Fig. 2 illustrated the difference between information retrieval model and knowledge exploration model.

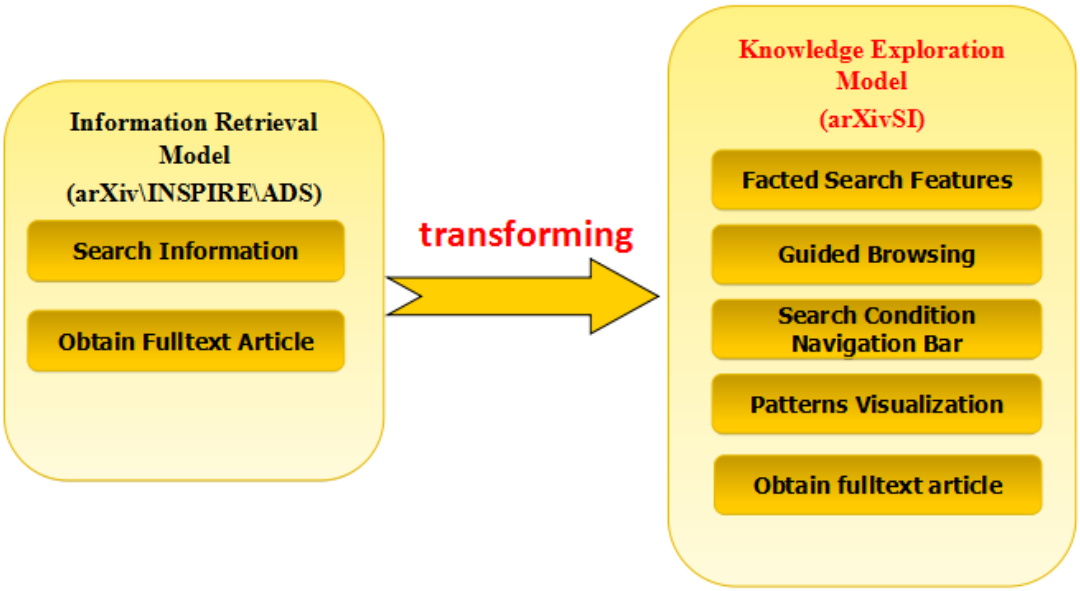

Fig.2 from Information Retrieval to Knowledge Exploration

Instead of just searching and getting full text article from arXiv, the authors try to realize a knowledge exploration system for user to explore papers in arXiv, with some new added features below.

(1) Faceted search features. arXivSI will provide advanced features for searching and filtering the search result with the faceted search abilities. It shows the hits counts by category in addition to search result.

(2) Guided browsing experience. arXivSI will provide guided browsing functions to help user to explore the papers by specific domain, subject, date or author, which means scientists can browse the new submitted papers or any paper they are interested easily and conveniently.

(3) Search conditions navigation. arXivSI will provide a search conditions navigation bar to trace user's search actions. The user can add or remove any search conditions easily to constrict or expand the search result.

(4) Patterns visualization. arXivSI will provide visualization tools to display the patterns of search result, which can help user to get a bird view of the information related to his search terms and promote knowledge discovery from search results.

3. Implementation of ArXivSI

To realize the proposed system, the authors implement the system with a focus on the follow main functions.

3.1 Metadata Harvesting

The authors develop an automatic metadata harvester based on OAI-PMH protocol that can harvest metadata timely from arxiv.org. ArXiv is a registered OAI-PMH data-provider and provides metadata for all submissions which is updated each night shortly after new submissions are announced [6].

From data-provider of arXiv, metadata for each paper is available in three formats: "oai_dc" format, "arXiv" format and "arXivRaw" format. The "oai_dc" format is a simple Dublin Core metadata, which does not include many useful fields, such as affiliation, ACM-Class etc. So the authors abandon this format. The "arXiv" format is a specific metadata format for arXiv submission which includes author names separated out, category and license information. It includes all the metadata except version history of the paper. Compared with "arXiv" format, the "arXivRaw" format includes version information of the submission. But the author names in "arXivRaw" format have not been preprocessed by arXiv, so if one chooses this format, he must separate author names out by himself. After detailed compare, the "arXiv" format is chosen to harvest metadata from arXiv data-provider.

Generally, metadata of arXiv is updated at about 8 pm (EST) each night and arXivSI will launch the task of harvesting incremented metadata from arXIv data-provider at about 10 pm (EST).

3.2 Metadata Indexing

The authors develop an automatic metadata indexer based on Solr software to index the new harvested metadata timely and efficiently.

There are two kinds of metadata in new harvested metadata. One is the metadata for new added paper, and the other is the metadata for updated paper. For new added paper, the indexer just adds the metadata to the Solr index for indexing. As to the metadata for updated paper, the indexer is needed to delete the old version of the metadata in the Solr index by arXiv Identifier and put the new updated metadata into the index. In this way, the Solr index can be kept up with the latest situation of arXiv.

Several subject classification systems are used in arXiv to organize submitted papers by subject or by concept, such as ACM Computing Classification System [7] and Mathematics Subject Classification [8]. In harvested metadata, the notation is used to represent the subjects or classes. For example, "I.2.7" in ACM Computing Classification System represents the subject term of "Natural Language Processing" and "14H10" in Mathematics Subject Classification represents the subject term of "Families, moduli". In order to support search from the subject by the term, the indexer not only index the notation of the subject, but also the corresponding subject term which is represented by the notation.

3.3 Exploratory Search Interface

On top the Solr index, an exploratory search Interface is developed for user to search papers in arXiv. The most obvious difference between arXivSI and arXiv search platform is that ArXivSI provides more suitable search functions for user to explore the research papers in arXiv. The main features implemented for exploratory search Interface of ArXivSI include the following functions.

(1) Faceted Search Features

At the arXivSI web sites, a quick search interface and an advanced search interface are implemented.

Comparing with the search functions of arXiv.org, arXivSI has filtering aid to filter the search result, has sort aid to sort the search result and it can display all hits in one query. With the help of faceted search features, arXivSI can display faceted result with the hits counts by category in addition to search result, which can help user to "analyze" what is included in the search result and "reveal" some clues for user to better understand and use the search result.

(2) Guided Browsing

Based on the faceted index, guided browsing is developed for user to select and explore the papers by specific domain, subject, submission time or author.

Guided browsing operates on the basis of constraints. Since arXivSi display all the search results to the user, not just the top 1000 hits in one query. Guided browsing presents the facets (specific domain, subject, submission time or author) in search and browse result for the user to choose. It will force constraints in the facets by extension papers are only presented for search and selection if the papers meet previous criteria as selected by the user, which means the user could browse any paper they are interested easily and conveniently in a large search result.

(3) Search Conditions Navigation Bar

A search condition navigation bar is implemented in arXivSI to trace user's search actions. The user can add or remove any search condition easily to constrict or expand the results. Fig. 3 and Fig.4 show the difference of user interface between arXiv and arXivSI platform. When one user searches "nano" in arXiv.org, the system cannot show all the hits because the

query resulted in more than 1000 hits. But when the user search “nano” from arXivSI, he can get all the hits and can navigate to any sub result set as he needs. For examples, Fig. 4 shows the user can get the exactly the 12 papers in Optics Research Domain that have keywords “nano” and submitted in 2015.

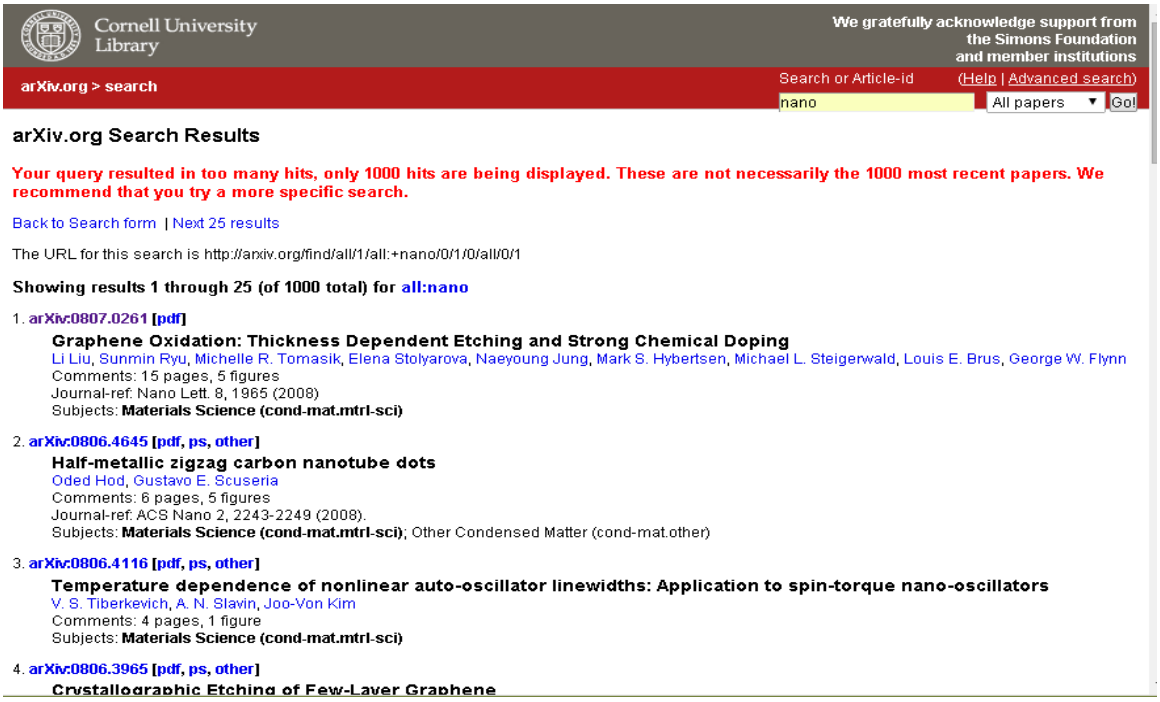

Fig. 3 the Search platform of arXiv.org

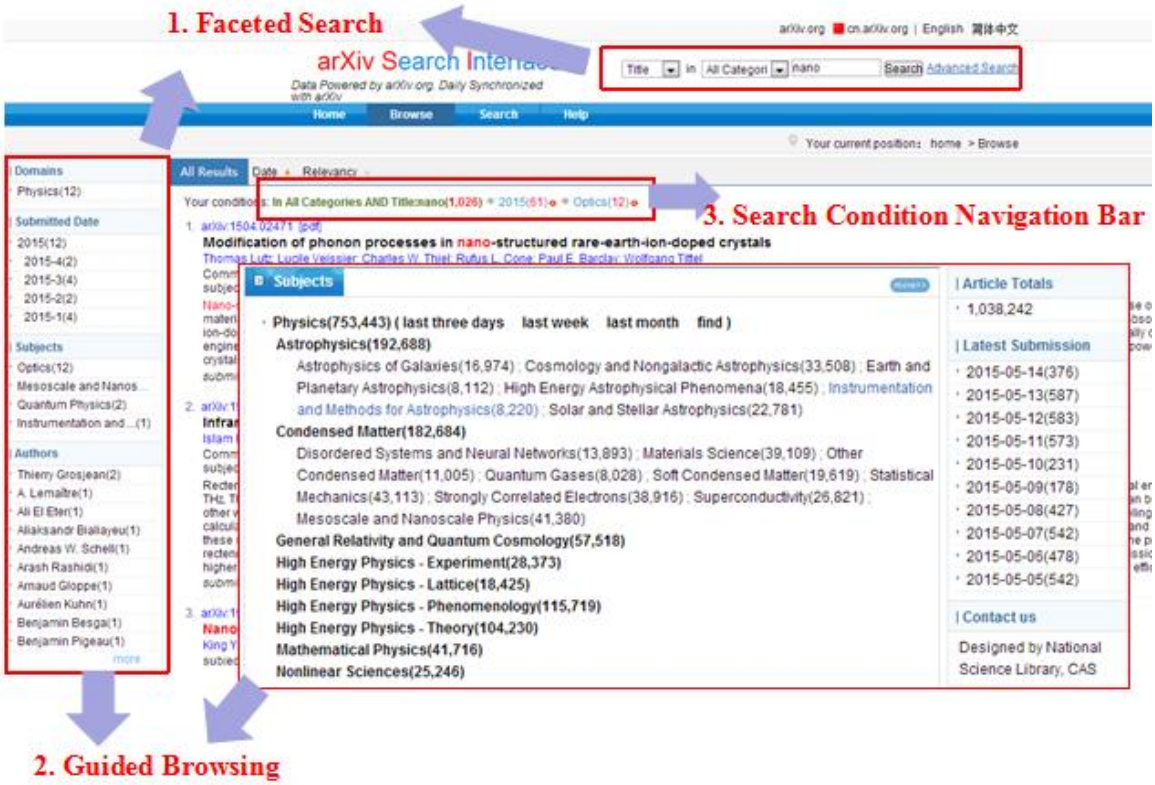

Fig.4 the Search interface of arXivSI

3.4 Patterns Visualization

Information visualization [9], as an effective information representation model, reveals the direct and indirect relationships between knowledge. In order to reveal useful patterns hidden in the search results, visualization functions have been implemented to visualize the patterns of search result. With the help of patterns visualization, the user of arXivSI can get a bird view of the information related to his search terms and explore the papers in arXiv easily and smoothly.

Some open source softwares, such as D3.js [10], sigma.js [11] are used in developing patterns visualization of arXivSI. The main functions implemented in patterns visualization of arXivSI include the followings.

(1)Visualization of Submission Count by Domain

From arXivSI, the user can get a vivid viewer of submission count of the papers submitted to the arXiv with the help of visualization tools. The Fig. 5 illustrates the paper submission count by domain in the past 10 years. From this graph, the user can see that the submission in each domain has an almost linear growth and the submission in physics increase more steadily compared with other domains.

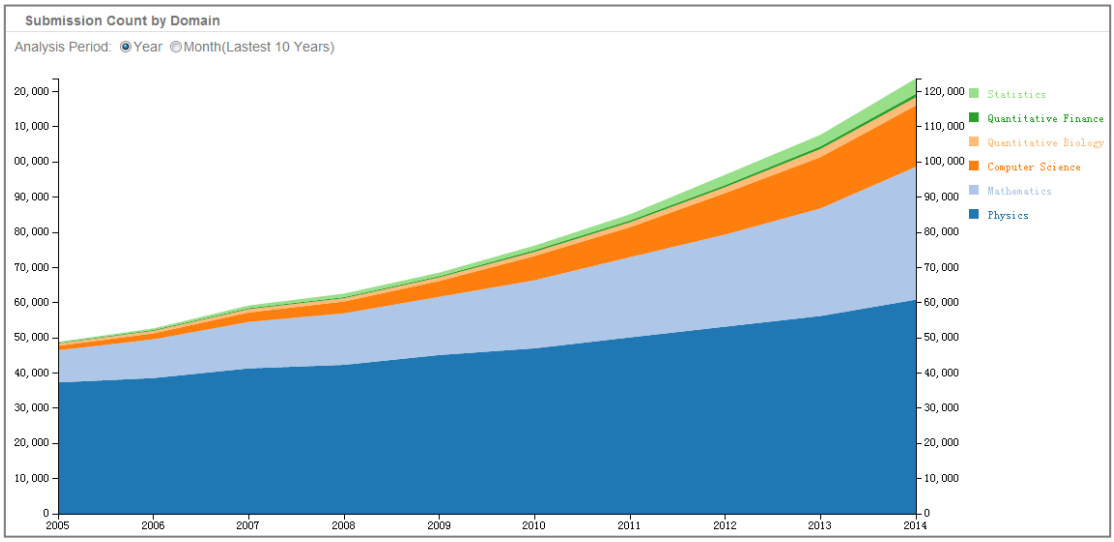

Fig. 5 Visualization of submission count by domain, 2005-2014

(2) Visualization of Submission Rate by Domain

In addition to visualization the number of submissions per year for each major subject area, the visualization of the fractions of total submissions for each subject area is also implemented. Fig. 6 is a graph from arXivSI which shows the visualization of submission rate of each major subject area in past 10 years. From this graph, we can see that the mathematics (gray) and the computer science (yellow) have grown to be the major areas of arXiv besides the physics in past 10 years and computer science (yellow) is the fastest-growing domain of arXiv in the past few years.

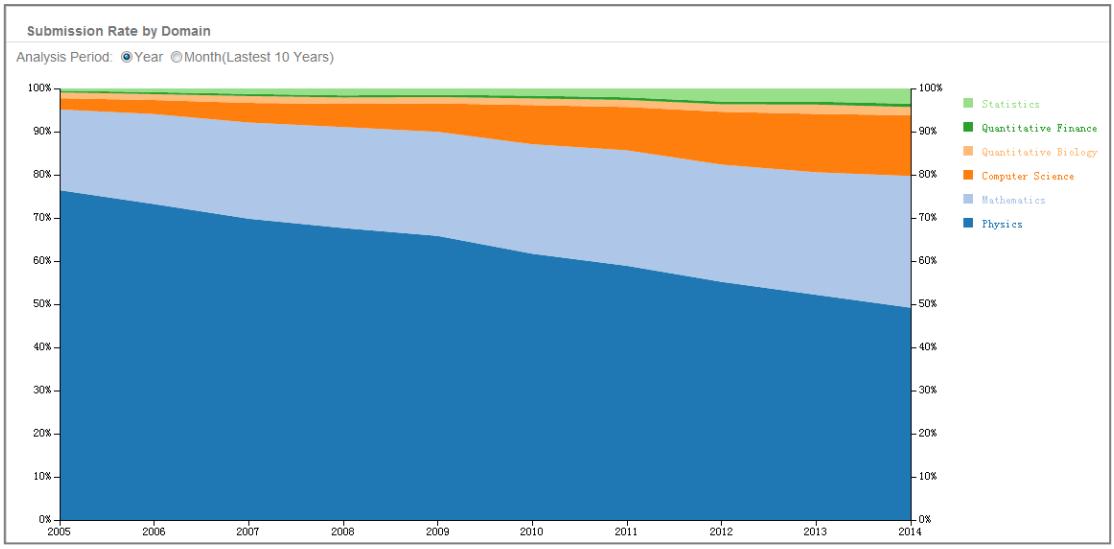

Fig. 6 Visualization of submission rate by domain, 2005-2014

(3) Visualization of Subject Distribution Patterns for One Domain

In arXivSI, visualization of subject distribution patterns for one special domain is implemented. By using this function, the user can analyze the subject distribution patterns of special domain.

Fig. 7 is a graph from arXivSI which shows the subject distribution patterns of arXiv research papers in “Physics” domain. The nodes represent the subjects which the arXiv papers in “Physics” domain belong to and the size of node represents the paper quantity of each subject. The edges represent the relationship of two subjects and the thickness of line represents the quantity of papers which belong to the both subjects. From this graph, everyone can see the arXiv research papers in “Physics” domain mainly distribute in subjects such as High Energy Physics-Phenomenology, High Energy Physics-Theory, Astrophysics, Quantum Physics etc. Moreover, one can find that there are strong relationships between subjects in General Relativity and Quantum Cosmology, High Energy Physics-Phenomenology, Quantum Physics, High Energy Physics-Theory and Mathematical Physics.

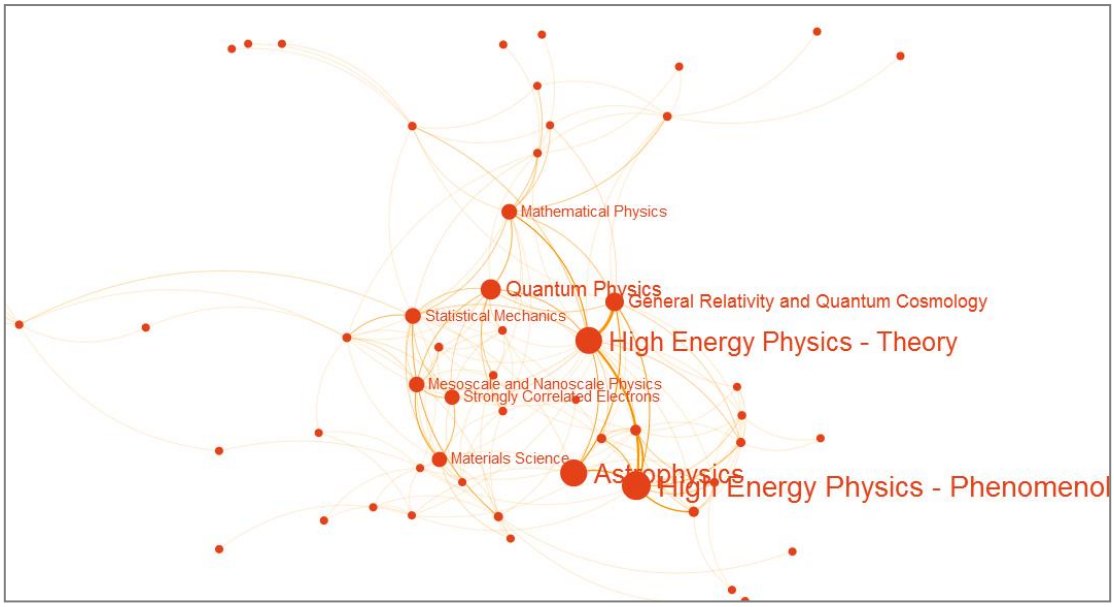

Fig. 7 Subject distribution patterns of arXiv papers in "Physics" domain

(4) Visualization of Subject Distribution Patterns for Search Result

In arXivSI, patterns visualization for the search results also has been implemented. It can help user to get a clear viewer of the subject distribution patterns in the search results.

Take the search term "nano" as an example. Fig. 8 illustrates the visualization of subject distribution of 1,026 arXiv papers that included in the search results. From the graph, the user can easily find that papers including "nano" in their abstracts are mainly categorized to following subjects: "Mesoscale and Nanoscale", "Materials Science", "Optics" etc.

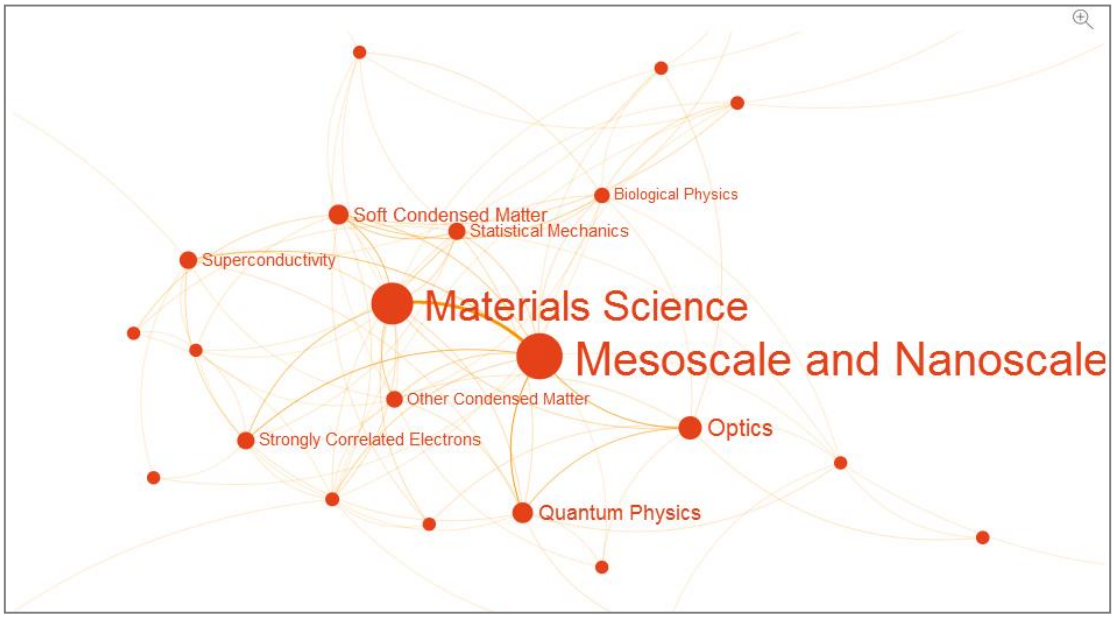

Fig. 8 Subject distribution patterns for search result of search term "nano"

4. Using of ArXivSI

On December 12, 2014, arXivSI service is released publicly. Till now, the using statistics shows arXivSI is widely used by scientists from more than 100 countries. Many scientists from world leading research institutes, such as Stanford, Oxford, CERN, MIT are using arXivSI to explore the research papers in arXiv.

After arXivSI released, there are some promotion activities in the Chinese Academy of Sciences (CAS). So the user from the CAS accounts for a very large proportion of the total access to arXivSI. Besides the user from the CAS, there are lots of users from some Chinese famous universities, such as Zhejiang University, University of Science and Technology of China, Tsinghua University. Since arXivSI is linked by arXiv.org, it is a great honor for us to see scientists from many world leading research institutes, such as Stanford, Oxford, CERN, MIT are using arXivSI.

Table 1 shows a general access count of arXivSI in 2015. Table 2 and Table 3 show top institutional users inside and outside China in 2015.

Table 1, General access count of arXivSI in 2015

Month	access count
2015-01	33552
2015-02	42408
2015-03	25642
2015-04	37895

Table 2, Top institutional users of arXivSI in China

No.	Institution Name	Total access
1	Chinese Academy of Sciences	45244
2	Zhejiang University	531
3	University of Science and Technology of China	245
4	Tsinghua University	205
5	Shenzhen Institutes of Advanced Technology	144
7	Beijing Normal University	133
8	The Hong Kong University of Science and Technology	129
9	Huazhong University of Science and Technology	78
10	Shanghai Jiao Tong University	74

Table 3, Top institutional users of arXivSI outside China

No.	Institution Name	Total access
1	Stanford University	229
2	University of Oxford	121
3	Nanyang Technological University	106
4	University of California	99
5	CERN	92
6	The University of Tokyo	87
7	Massachusetts Institute of Technology	86
8	Columbia University	70
9	Universität Hannover	69
10	University of Toronto	65

5. Conclusion

In this paper, the authors developed a search interface service named arXivSI to help scientists efficiently explore and discover the knowledge housed in arXiv.

arXivSI harvests metadata in arXiv format from arXiv OAI-PMH data provider and uses Solr to index the harvested metadata. On top of the system, besides displaying the ordinary search result, arXivSI offers faceted search features, guided browsing, search conditions navigation bar, and patterns visualization functions to build a knowledge exploration interface for user. arXivSI reveals useful patterns hidden in the search results, visualizes those patterns in a vivid way and provides a user-friended interface for user.

After the arXivSI service is released, it is widely accepted and used by Chinese and other international institution users. In further, arXivSI will provide more deep knowledge exploration services with Nature Language Processing, Machine Learning and other technologies.

Acknowledgments

This article is partially supported by the project “Developing Shared Services Platform for Scientific and Technological Knowledge Organization System” (Grant No. 2011BAH10B03), funded by the National Science & Technology Pillar Program of China.

References

[1] Zhang Zhixiong, Zhang Shanshan, Ku Liping, Li Lin. Survey and Analysis on Cognition and Using of arXiv for China Mainland Researchers. New technology of library and information services. 2014, 30(7/8): 1-8.
[2] HEP - INSPIRE-HEP. http://inspirehep.net [accessed May 14, 2015]
[3] The SAO/NASA Astrophysics Data System. http://adsabs.harvard.edu/ [accessed May 14, 2015]
[4] Open Archives Initiative (OAI). http://arxiv.org/help/oa/index [accessed May 14, 2015]
[5] Apache Solr. http://lucene.apache.org/solr/ [accessed May 14, 2015]
[6] arXiv, Open Archives Initiative (OAI), http://arxiv.org/help/oa/index [accessed May 14, 2015]
[7] The 1998 ACM Computing Classification System. http://www.acm.org/about/class/1998/ [accessed May 14, 2015]
[8] Mathematics Subject Classification. http://www.ams.org/mathscinet/msc/msc2010.html [accessed May 14, 2015]
[9] Information visualization - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Information_visualization [accessed May 14, 2015]
[10] D3.js - Data-Driven Documents. http://d3js.org/ [accessed May 14, 2015]
[11] Sigma.js. http://sigmajs.org/ [accessed May 14, 2015]