

Supply and Demand Analysis in NDLTD Based on Patron Specialty Survey and Contents Statistics

Seonho Kim, Seungwon Yang, and Edward A. Fox
{shk, seungwon, fox@vt.edu}

Digital Library Research Laboratory, Virginia Tech
Blacksburg, VA 24061 USA

ABSTRACT

There have been many efforts to investigate the use of ETDs, based on web log analysis, yielding information such as transferred data type, data size, file name, connection time, browser type, user's IP address, error code, etc. It is more difficult, however, to assess how well the contents of collections match the demands in each scholarly area.

Since August 2005 we have been carrying out a user survey, as part of the registration process for our NDLTD search service providing clustered retrieval results, and 1,100 users' answers are available. Users are asked to enter their majors and detailed specialties, as well as the number of years they have worked in those areas. Through user modeling, building upon that data, we can measure the information demand for each user and area. To quantify the contents/supply, we classify all the ETDs in the NDLTD union catalog into each scholarly area, based on the information provided by the authors. Then we analyze the diversity and proportions of the ETDs, and compare them with the corresponding user demands.

1. INTRODUCTION

The Networked Digital Library of Theses and Dissertations (NDLTD) union catalog [1] contains over 242,000 electronic theses, dissertations (ETDs) and important documents provided by more than 325 member institutions, such as libraries and universities. There were many studies to characterize and classify these ETDs and their access rates. However, there were no prior efforts to measure the information demands in NDLTD. In this paper, we put our focus on figuring out how well the contents of NDLTD match the demands of patrons in each scholarly area. Additionally, we analyze the patron distribution according to major fields, and expertise years in their fields, as well as the distribution of NDLTD resources.

In Section 2 we describe the data set and preprocessing. In Section 3 we describe our approach to analyze data, difficulties we faced, and our method of measuring supply and demand in NDLTD. Section 4 gives details and charts about the results of our analysis. Section 5 presents conclusions and Section 6 describes future work.

2. DATA SET AND PREPROCESSING

We used two different types of data sets, from three different sources, according to the purpose of analysis. One is ETD data for measuring information supply of NDLTD, and the other is user

survey data and query log data, for measuring information demand of NDLTD patrons. Table 1 shows the types, sources, quantity of data, and brief descriptions.

Table 1. Data Sets

Type	Source	Number of Records	Description
Supply Analysis	ETD	242,688	Harvested from Online Computer Library Center (OCLC) using “OAI/ODL Harvester” [2]. Contains ETDs since Fall 2005 through part of the Spring 2006 graduation time.
Demand Analysis	User Survey	1,100	Online user survey conducted from August 2005 to April 2006 as part of our User Modeling Study [3]. Contains demographic information, major, research fields, and expertise years in the fields for each user.
	Query Log		Collected by User Tracking System [4] as part of User Modeling Study. Consists of queries and their frequencies for each user.

Figure 1 gives an example of ETD data, and Figure 2 shows an example of user data, which includes user survey data and query log data. ETD data is used for measuring the amount of resource supply in NDLTD, and user data is used for measuring the amount of information demand in NDLTD.

```
<dc oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
dc="http://purl.org/dc/elements/1.1/" xsi="http://www.w3.org/2001/XMLSchema-instance
schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/
OAI/2.0/oai_dc.xsd"><title>Composer-Centered Computer-Aided Soundtrack Composition
</title><creator>Vane, Roland Edwin </creator><subject>Computer Science</subject>
<subject>human Computer interaction</subject><subject>music composition</subject>
<subject>soundtracks</subject><subject>creativity</subject><description>For as long as
computers have been around, people have looked for ways to involve them in music....
</description><publisher>University of Waterloo</publisher><date>2006</date><type>
Electronic Thesis or Dissertation</type><format>application/pdf</format><identifier>
http://etd.uwaterloo.ca/etd/revane2006.pdf</identifier><language>en</language><rights>
Copyright: 2006
```

Figure 1. An Example of ETD data

Each ETD record is in XML format and contains meta-information of the document, such as Dublin Core information, title, subjects, description, publishers, date, etc. Each user data record contains demographic information, research area, and years spent in these areas. It also contains user tracking information, such as query log data and interest in research topics. The “topics” are noun phrases appearing as cluster names of retrieval result sets processed by a document clustering system. In the user survey, users are asked to enter basic demographic information, one major, up to two specific research areas, and years of experience in each area. In the query log, illustrated in the “<query>” field in Figure 2, all queries the user have sent for search are recorded with their frequency. Our data shows each user sent on average 4.52 queries during the survey

period from August 2005 to April 2006. The “selected” fields in Figure 2 are topics, and names of clusters, that the user clicked to browse documents under the topics. The “proposed” fields represent all topics shown on the result screen. We use the SAX XML [5] parser to extract specific fields, such as “subject” and “date” fields from ETD data, and “userID”, “major”, “broadresearch”, “specific”, and “query” fields from user data.

```
<user> <userID>shk</userID> <email>shk@vt.edu</email> <name><first>Sh</first>
<last>King</last> </name><major>CS</major><broadresearch>Digital Library
<specific>User interface</specific><experience>8,2</experience></broadresearch><group
/><query><item freq="79">digital library</item> <item freq="33">computer science</item>
<item freq="25"> virginia tech</item> <item freq="9">artificial intelligence</item> <item
freq="5">digital library.</item> </query><selected><item freq="15">Digital Library</item>
<item freq="6">Electronic Theses and Dissertations</item> </selected><proposed><item
freq="80">Digital Library</item> <item freq="65">Data</item> </proposed></user>
```

Figure 2. An Example of User Data

3. CLASSIFICATION AND MEASUREMENT OF SUPPLY AND DEMAND

The goal of this study is to understand how well the ETDs in ND LTD match with the information demands of users in each scholarly field. Our approach is based on classifying both the ETDs and user data into the same scholarly classes with the same criteria to see their distributions. Then we compare these two distributions with each other. This section explains our classification approaches for both ETDs and user data.

3.1 Faculty Categories

We created our own classification categories based on faculty/college systems of universities. We considered five universities in Virginia, such as Virginia Tech, University of Virginia, George Mason University, Virginia Commonwealth University, and Virginia State University. 7 categories and 77 subcategories were identified as listed in Table 2.

Table 2. Seven Categories and Seventy Seven Subcategories

	7 categories	77 subcategories
1	Architecture and Design	ArchitectureConstruction, LandscapeArchitecture
2	Law	Law
3	Medicine, Nursing and Veterinary Medicine	Dentistry, Medicine, Nursing, Pharmacy, Veterinary
4	Arts and Science	Agriculture, AnimalPoultry, Anthropology, ApparelHousing, Archaeology, Art, Astronomy, Biochemistry, Biology, Botany, Chemistry, Communication, CropSoilEnvSciences, DairyScience, Ecology, EngineeringScience, English, Entomology, Family, Food, ForeignLanguageLiterature, Forestry, Geography, Geology, GovernmentInternationalAffair, History, Horticulture, HospitalityTourism, HumanDevelopment, HumanNutritionExercise, Informatics, Interdisciplinary, LibraryScience, Linguistics, Literature, Meteorology, Mathematics, Music Naval, Philosophy, Physics, Plant, Politics,

		Psychology, PublicAdministrationPolicy, PublicAffair, Sociology, Statistics, UrbanPlanning, Wildlife, Wood, Zoology
5	Engineering and Applied Science	Aerospace, BiologicalEnginerring, Chemical, ComputerScience, Electronics, Environment, Industrial, Materials, Mechanics, MiningMineral, Nuclear, OceanEngineering
6	Business and Commerce	AccountingFinance, Business, Economics, Management
7	Education	Education
8	Others (unclassifiable)	(Unclassifiable)

3.2 Classification

Classification of ETD was done by examining “subject” fields (see Figure 1). Likewise, classification of user data was done by examining the “major”, “broadresearch”, and “specific” fields (see Figure 2). We will call these fields, which are used for classification, “key fields”. For classification of both ETD and user data, we built a common matching table that consisted of identification string patterns for each of the 77 subcategories. We checked the key fields for inclusion of any pattern in the matching table. If any pattern in the table appears in the key fields, we classify the ETD or user data into the corresponding category. We created the matching table based on common knowledge, dictionary, and faculty/college information from the universities listed in Section 3.1. Table 3 shows a portion of the matching table.

Table 3. Matching Table Excerpt

77 categories	Identification Patterns
Education	/bildung/, /pedagog/, /fakul/, /educa/, /teaching/, ...
Geology	/geolog/, /geoscience/, ...
LibraryScience	/librari/, /library/, /informatik/, ...
...	...

3.3 Challenges

Our approach, classifying ETDs and user data based on searching for identification strings in key fields and classifying into the corresponding category, has several difficulties, as listed below:

- a. There are various expressions used in describing the same research subjects in the key fields. This is not only because there exist hundreds of research subjects, but also because people describe the same subject in slightly different ways.
- b. Various languages, such as English, Spanish, German, Portuguese, etc., are used for ETDs.
- c. There are many interdisciplinary subjects which can be classified into multiple categories. For example, should “Music Education” be classified into the “Music” class or the “Education” class? We classified these subjects to the class with the higher priority. In the case of this example, “Music Education” was classified into the “Music” category, because there are many types of education, e.g., “Math Education” and “Computer Education”, so the priority of “Education” is less than that of “Music”.
- d. Many ETDs have no entry in the key fields.
- e. Key fields in some ETDs and user data contains wrong information, such as “Ph.D”, “Georgia”, “gordd7@camu.edu”, etc., or typos.

- f. Some ETDs' key fields contain too much highly detailed information, that we couldn't understand, such as "muon", "cytochrome", "pulsars", etc., or unknown abbreviations, such as "MOCVD" and "OFDM".

ETDs and user data corresponding to cases 'd', 'e', and 'f' above, as well as some for 'a' and 'b', were placed into the "Unclassifiable" class.

3.4 Measuring Supply and Demand

To measure the resource supply of NDLTD, we counted the number of ETDs in each scholarly category after classifying ETDs as described in previous sections. However, measuring user demand in each scholarly category is controversial. Does the number of users in a category represent well enough the user demand of the category? The numbers of users just tells us how users are distributed in NDLTD. Thus, we attempted to use actual indicators of information demand, based on numbers of queries, as represented by equation (1) below.

$$Demand\ of\ a\ Category = \sum_{user \in category} number\ of\ queries \quad (1)$$

That is, the amount of information demand in a category is the sum of the number of queries sent by all users in the category.

4. ANALYSIS OF SUMMARY STATISTICS

4.1 ETD Distribution

We classified 242,688 ETDs into 77 subcategories as listed in Table 2. Because our classification method, described in Section 3.2, had too many ETDs in the "Unclassifiable" category, we followed the "correct and run over" approach, and kept updating the matching table until no ETDs with frequency greater than 9 were left. The final unclassifiable ETDs amounted to 40.17%. About 90% of the unclassifiable ETDs have a unique "subject", which makes it hard to handle them one by one.

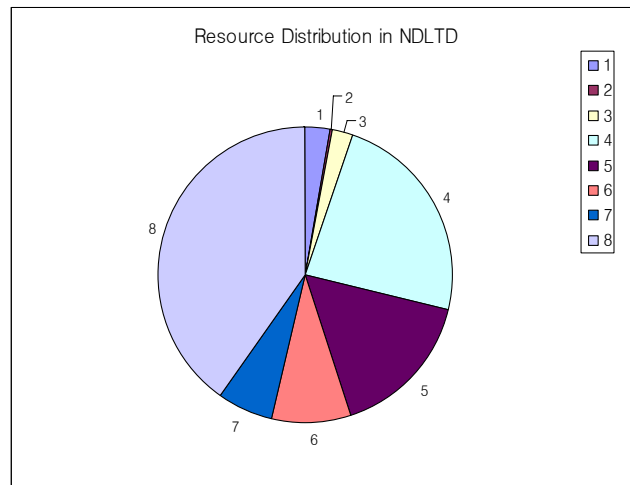


Figure 3. Distribution of ETDs in 7 categories

After classification into 77 subcategories, each subcategory was merged into one of the higher level 7 categories. Figure 3 shows the result. The numbers in this chart, and all other charts in this paper, correspond to the 7 categories listed in Table 2. We see that the "Arts and Science (4)"

and “Engineering and Applied Science (5)” categories, if one ignores “Unclassifiable”, are the most dominant areas among the 7 categories.

4.2 User Distribution

1,100 users are classified into 77 subcategories based on their “major”, “broadresearch” and “specific” research fields. After then, 77 subcategories are merged into higher 7 categories as we for the ETD classification discussed in Section 4.1. Figure 4 presents the result. Again, we had a “too many unclassifiable users” problem and reduced the number as we did in ETD classification.

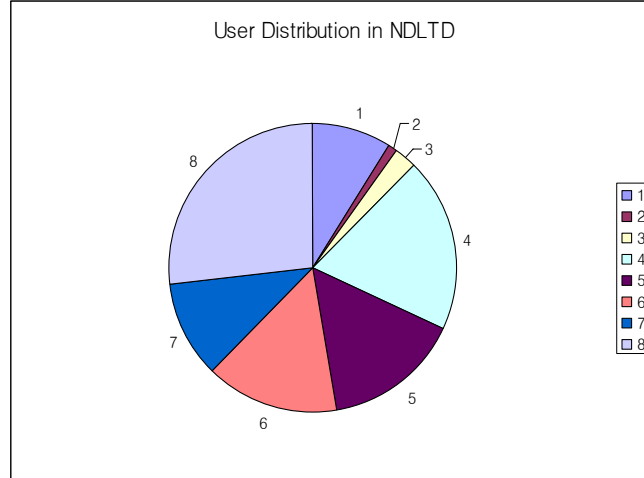


Figure 4. Distribution of NDLTD Users

From this chart, we can see that users are distributed similarly over all 7 categories except for the “Law (2)” and “Medicine, Nursing and Veterinary Medicine (3)” categories.

4.3 Query Distribution

Figuring out the query distribution in each scholarly category is done by adding up all frequencies of queries of all users in each category, as in equation (1). We regard these values as “demand” in NDLTD. Figure 5 gives the overall distribution of queries in NDLTD.

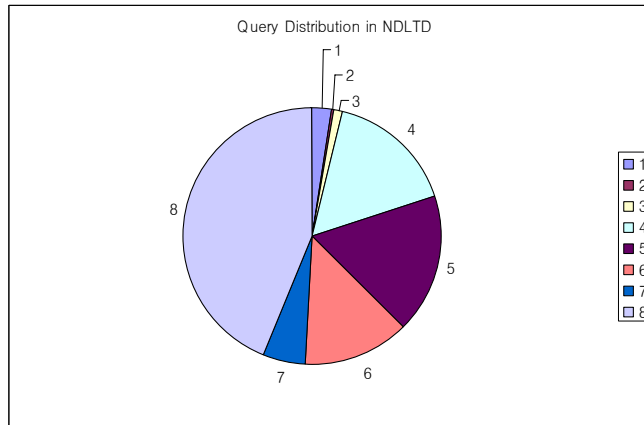


Figure 5. Distribution of Queries in NDLTD

We can see users from “Arts and Science (4)”, “Engineering and Applied Science (5)”, and “Business and Commerce (6)” have requested similar amounts of information. From Figures 4 and 5, we see that although the numbers of users in “Architecture and Design (1)” and “Education (6)” fields are similar to other categories, they didn’t use NDLTD actively.

4.4 Supply-Demand Comparison

A Supply-Demand comparison tells how well the supply of ETDs in NDLTD matches with user demands in each category. Figures 6 and 7 are results of comparisons in the 77 subcategories. These figures show that supply for “Business” and “Economics” is inadequate, relative to other fields. It is similar for several engineering areas, such as “Computer Science” and “Electronics”. Figure 8 summarizes Figures 6 and 7 by merging subcategories into the higher level 7 categories.

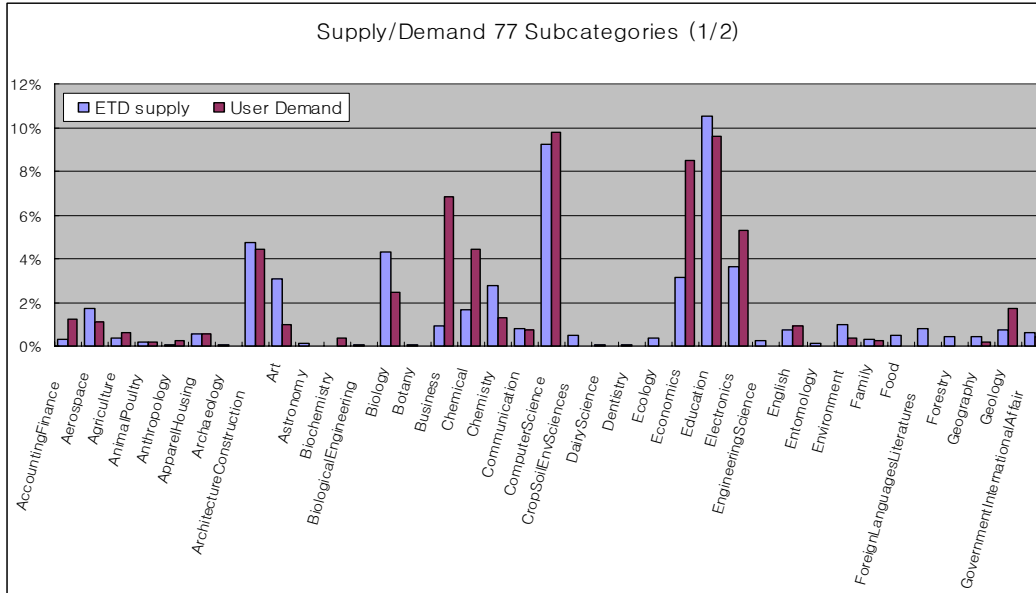


Figure 6. Supply-Demand Comparison in 77 subcategories (part 1 of 2)

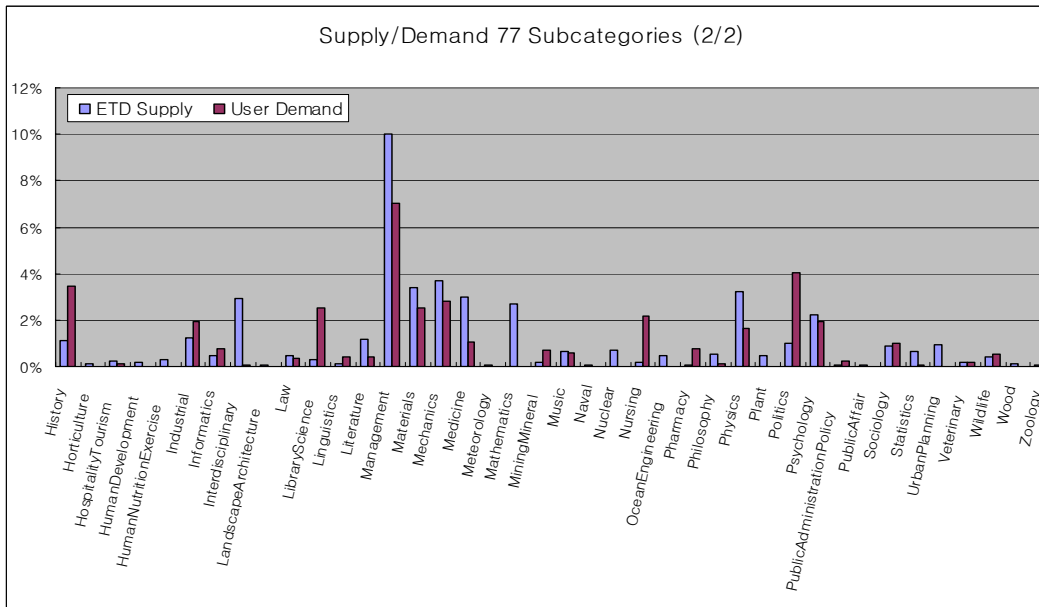


Figure 7. Supply-Demand Comparison in 77 subcategories (part 2 of 2)

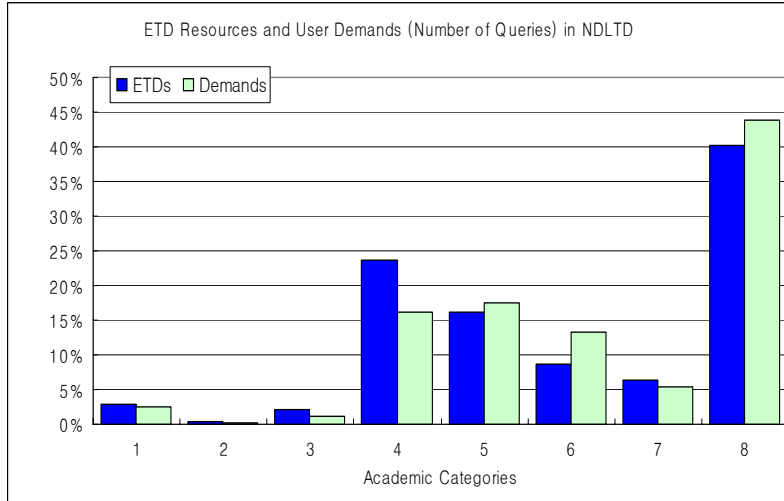


Figure 8. Supply-Demand Comparison in 7 categories

From these charts, we see that NDLTD is supplying enough ETDs in “Architecture and Design (1)”, “Medicine, Nursing and Veterinary Medicine (3)”, “Arts and Science (4)” and “Education (7)” fields, but not for “Engineering and Applied Science (5)” and “Business and Commerce (6)”.

4.5 Date Stamp Distribution

Each ETD record has a date stamp (“date” field, see Figure 1), which contains the document’s year of generation. Analyzing this information provides an overview of the age of NDLTD data.

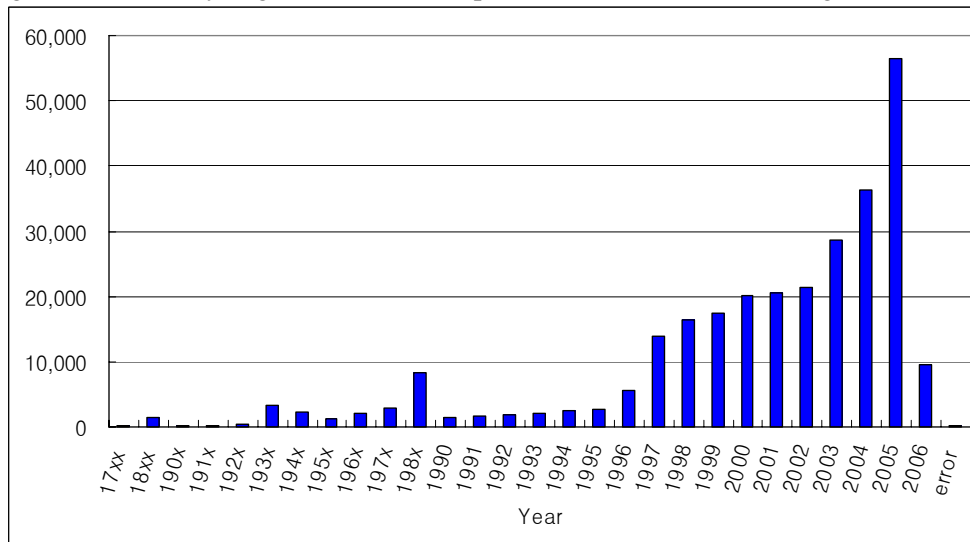


Figure 9. Date Stamps of ETDs

This chart shows that most of the ETDs in NDLTD are generated after 1996 when the NDLTD project started, and have increased in number steadily. Also, it shows some ancient document are digitized and have been entered recently. As we examined these old-date-stamped records, most of these records are documents describing old pictures or materials, and the date-stamps are the generation years of the old pictures or materials, not the document describing them.

4.6 User Expertise Distribution

We also analyzed the distribution of users' expertise years in their research areas and the amount of their demand, to investigate how many experienced users are using NDLTD. This is done by counting user data, which contains the answers to the user survey that asked their expertise years in their interest fields. Figure 10 shows the result. Because most users answered roughly, with simple numbers of years like 5, 10, 15, 20, etc., the values of expertise years tend to converge on those numbers.

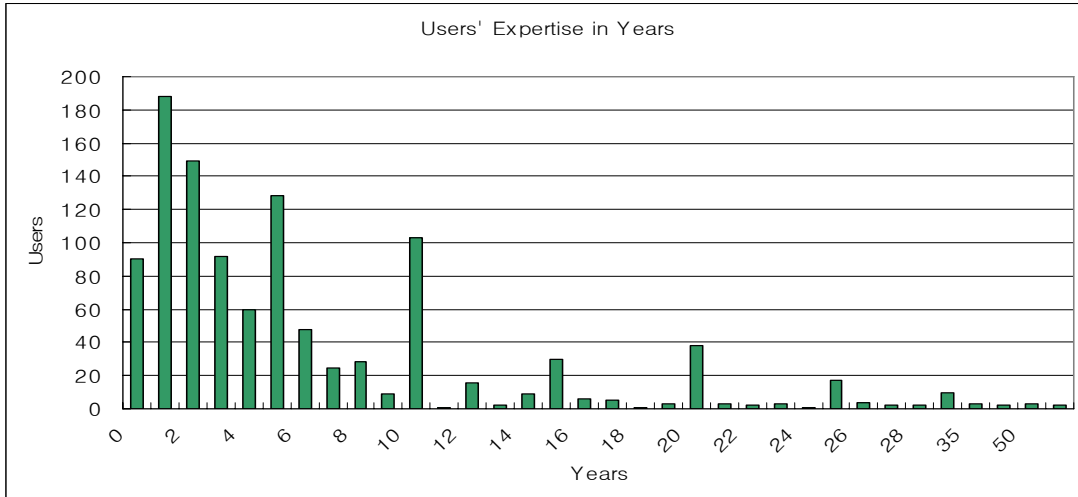


Figure 10. Users' Expertise in Years

This result tells us that most NDLTD users have experience of between 1 to 10 years, which means that most of NDLTD users are graduate students or novice researchers. However, about 15% of all users have more than 10 years of experience in their fields. Figure 11 presents user experiences and their information demands in NDLTD. This tells how NDLTD is used actively for different levels of users. We can see most active users were graduate students with one year experience in their field.

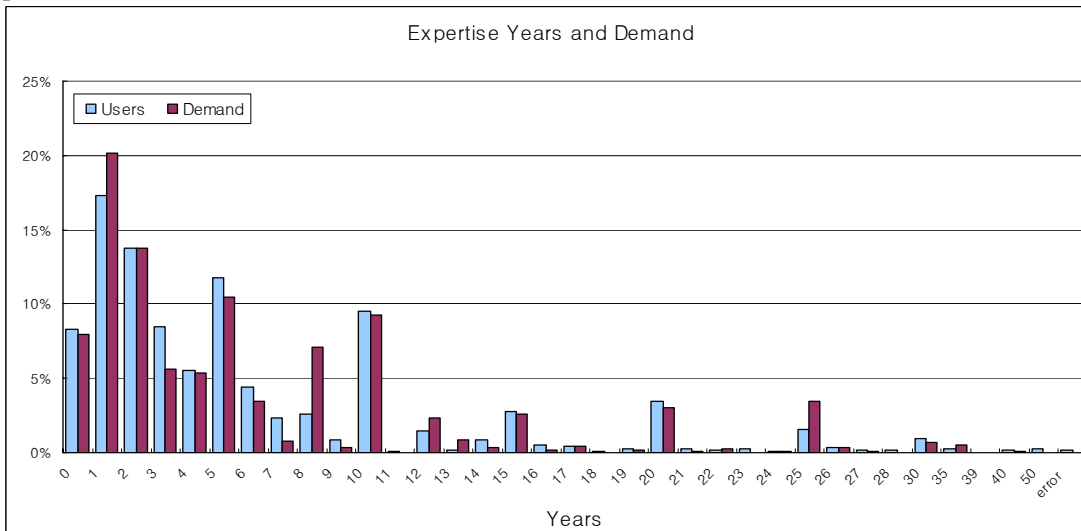


Figure 11. Users' Expertise and Demand

5. CONCLUSIONS

We analyzed ETDs and users in NDLTD to understand how well NDLTD supplies ETDs for users in each scholarly area. A total of 242,688 ETDs and 1,100 users were classified into 7 categories and 77 subcategories using the same classification criteria, based on matching identification strings from “subject” fields in ETDs and “major”, “broadresearch”, and “research” fields in user data. User demand in a category was measured by adding up query frequencies sent by all users in the category. From the comparison between supply and demand in NDLTD, we found the supply in the “Engineering and Applied Science” and “Business and Commerce” areas may be relatively insufficient. We also presented the date stamp distribution of ETDs to help in understanding of the age of the data in NDLTD, and users’ expertise distribution and their amount of demand – to reveal the expertise level of major users.

6. FUTURE WORK

Classification of ETDs and users was the most important process in this study, however, our classification method showed a problem that left too much data “unclassifiable”. In the future, we will add more user data and improve the precision of classification by using a popular text classification system, such as SVM^{Light} [6]. Using “oai_etdms” meta-data together with “oai_dc” also would improve the classification. Currently, we use only query log information to measure the amount of information demand, however, we can improve this method by using additional user tracking information. Finally, our future work will include use of more advanced data mining techniques and visualization.

REFERENCES

1. NDLTD, Networked Digital Library of Theses and Dissertations, available at <http://www.ndltd.org>, 2006
2. Hussein Suleman, “Introduction to the Open Archives Initiative Protocol for Metadata Harvesting”, in Proceedings of ACM/IEEE 2nd Joint Conference on Digital Libraries (JCDL 2002), p. 414, software available at <http://oai.dlib.vt.edu/odl/software/harvest/>
3. Seonho Kim, Uma Murthy, Kapil Ahuja, Sandi Vasile, Edward A. Fox, “Effectiveness of Implicit Rating Data on Characterizing Users in Complex Information Systems”, Springer-Verlag Lecture Notes in Computer Science, LNCS 3652, 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005), 2005, 186-194
4. Search Interface Embedded User Tracking System, available at <http://boris.dlib.vt.edu:8080/controller/index.jsp>, 2006
5. SAX XML parser, available at <http://www.saxproject.org/>, 2006
6. SVM^{light}, Support Vector Machine, available at <http://svmlight.joachims.org/>, 2006