
Pergamon

Library & Information Science Research
24 (2002) 265–291

Library & Information Science Research

Exploring behavior of E-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan

Hao-Ren Ke[a,*], Rolf Kwakkelaar[b], Yu-Min Tai[c], Li-Chun Chen[c]

[a]*Digital Library Section, Library of National Chiao-Tung University, 1001 Ta-Hsueh Road, 300 Hsinchu, Taiwan*
E-mail address: claven@lib.nctu.edu.tw (H.-R. Ke).
[b]*Elsevier Science (Asia Pacific), 1 Temasek Avenue, 17-01 Millenia Tower, Singapore 039192*
[c]*Institute of Computer and Information Science, National Chiao-Tung University, 1001 Ta-Hsueh Road, 300 Hsinchu, Taiwan*

Abstract

In the era of digital libraries, Web-based electronic databases have become important resources for education and research, providing functionality and ease of use superior to print products. Analysis of usage of such online systems can provide valuable information on user behavior, and on usage of electronic information in general. Furthermore, the findings can be used to improve effectiveness of these electronic systems and identify areas for improvement, ranging from user interface and functionality to documentation and product training. This article analyzes usage of the Taiwan-based ScienceDirect OnSite E-journal system, one of the largest and most heavily used full-text Science, Technology, and Medicine (STM) databases worldwide. © 2002 Elsevier Science Inc. All rights reserved.

Fostering information literacy is a primary goal of library science. Information literacy is the ability to access, evaluate, and use information from a variety of sources for problem solving, research, decision making, and continued professional development (Bruce, 1995; Doyle, 1992). Because information retrieval (IR) systems are indispensable in assisting these processes, the research into the assessment and improvement of IR systems to satisfy users' needs is an important part of library science.

* Corresponding author.

PII: S0740-8188(02)00126-3

Several criteria have been proposed for assessing IR systems, including information coverage, query functionality, precision and recall, response time, and user aids. Because the goal of an IR system is to assist users in locating and accessing information, user-centered assessment is imperative, and understanding user behavior is of primary importance. Only when user behavior is understood can developers of IR systems effectively enhance system functionality and increase user satisfaction.

The Internet and World Wide Web have introduced new and powerful ways for finding and sharing information. In the scientific arena, the proliferation of Web-based abstracting and indexing (A&I) databases and electronic journals (E-journals)[1] is revolutionizing the way researchers conduct research and communicate their research results. Furthermore, Web-based electronic resources facilitate the exploration of user behavior far beyond what is possible in a print environment, as Web services can be configured to record (log) all relevant user transactions. By analyzing these transaction logs, it is possible to obtain a detailed picture of what users are doing and how they use a service.

This article analyzes transaction logs of the Taiwan ScienceDirect OnSite (SDOS) E-journal system, interprets the results, and states the findings. The Taiwan SDOS system is one of the largest and most heavily used full-text Science, Technology, and Medicine (STM) databases worldwide. It hosts the bibliographic information and full-text articles of more than 1,300 journals published by Elsevier Science. There are an estimated 625,000 users for the Taiwan SDOS system. This system is hosted locally by Academia Sinica, a leading academic institution of Taiwan, and can be accessed by SDOS subscribers, including all major Taiwanese universities and research institutes.

Transaction log analysis has been used to obtain insight into user behavior for decades (Peters, 1993). Most of the studies have focused on online public access catalogs and traditional IR systems. With the widespread use of the Web, there has been some research related to the analysis of Web log files of E-journals (Borghuis et al., 1996; Eason, Yu, & Harker, 2000; Jones, Cunningham, McNab, & Boddie, 2000; Kaplan & Nelson, 2000; Spink, Wilson, Ellis, & Ford, 1998; Zhang, 1999). In Zhang's (1999) analysis results of an E-journal, *Review of Information Science*, he included an analysis of users and their distribution, information or services needed by users, and information on usage patterns. Zhang studied levels of use as revealed in the log files for a single E-journal, and his analysis was relatively basic. In comparison, this article will look at the usage of a full-text system containing over 1,300 journals, and accessed by a very large user population.

Jones et al. (2000) conducted a log analysis of the Computer Science Technical Reports (CSTR) collection of the New Zealand Digital Library. The log files that they considered were collected in a 61-week period from April 1996 to July 1997. They examined the user demographics, use of operators and options in queries, pattern of query construction and refinement, distribution of query terms, and common mistakes in searching. That study and

[1] In this article, the terms *Web-based A&I databases, Web-based electronic resources, databases*, and *electronic resources* are used interchangeably.

the present one are similar in that both investigate similar aspects of query behavior. However, there are some notable differences:

- CSTR contains about 46,000 computing-related technical reports, whereas SDOS holds well over 1,000,000 multidisciplinary articles, including computer science, engineering/energy/technology, and mathematics.
- CSTR is a publicly available collection, whereas SDOS is a commercial service and access is limited to subscribing institutions only.
- Jones et al. (2000) focused primarily on the query pattern of CSTR. In addition to query patterns, this study also examines browsing and full-text downloading patterns.
- Jones et al. (2000) manually examined queries to gain some qualitative understanding of query behavior. This study does not conduct a manual examination.

Transaction logs have been used for analyzing user behavior of Internet search services as well (Catledge & Pitkow, 1995; Choo, Detlor, & Turnbull, 1999; Hawk & Wang, 1999; Hert & Marchionini, 1998; Jansen, Spink, & Saracevic, 2000; Pu, 2000). Jansen, Spink, and Saracevic (2000) analyzed logs of a major Internet search service, Excite, to explore the characteristics of Web users in performing queries. Pu (2000) explored the searching behavior of network users in Taiwan via the analysis of query-term logs provided by two public Internet search services, Dreamer and Global Area Information Servers (GAIS).

1. Data collection and preprocessing

The SDOS transaction log files used for this study cover the period of January 1 to September 18, 2000, and hold about 1.3 gigabytes (GB). These log files are generated by the SDOS Web server component and adhere to the common log format specification defined by the National Center for Super Computing Applications (NCSA). The most relevant fields of this format include Internet Protocol (IP) address, date and time, requested URL, and bytes returned (see Table 1 for sample records).

Because the SDOS log files contain records that are irrelevant for this study, and the log records are more detailed than necessary, in our analysis, the following four steps have been taken in preprocessing the transaction logs:

- Data cleaning: This step removes the records unnecessary for the analysis, such as records related to loading embedded gif images and records for failed transactions (HTTP error codes).
- Data integration: The SDOS system was upgraded in April 2000, and as a result, the URL format changed, affecting the log files as well. This step unifies the pre- and post-upgrade log-file data.
- Data classification: SDOS usage is classified into 14 categories based on request types shown in Table 2.

Table 1
Sample of SDOS transaction logs (only relevant fields are displayed)

IP Address	Date and time	Request	Bytes
X.X.X.X	[01/Jan/2000:00:00:49 +0800]	"GET /cgi-bin/sciserv.pl?collection = journals&letter = t HTTP/1.1"	15,839
X.X.X.X	[01/Jan/2000:00:01:15 +0800]	"GET /cgi-bin/sciserv.pl?collection = journals&letter = u HTTP/1.1"	9,414
X.X.X.X	[01/Jan/2000:00:01:17 +0800]	"GET /images/rf.gif HTTP/1.1"	—
X.X.X.X	[01/Jan/2000:00:01:28 +0800]	"GET /cgi-bin/sciserv.pl?collection = journals&letter = v HTTP/1.1"	9,681
X.X.X.X	[01/Jan/2000:00:01:35 +0800]	"GET /cgi-bin/search.pl?collection = journals&search_field = xml& GetSearchResults = Search&fields = Any HTTP/1.1"	88,881
X.X.X.X	[01/Jan/2000:00:02:19 +0800]	"GET /cgi-bin/sciserv.pl?collection = journals&journal = 09214526& issue = vb208-209i1-4&article = 297_xsoremac&form = pdf&file = file.pdf HTTP/1.0"	129,011
X.X.X.X	[01/Jan/2000:00:02:23 +0800]	"GET /cgi-bin/sciserv.pl?collection = journals&journal = 01429418& issue = v18i0003&article = 181_tpocfc&form = pdf &file = file.pdf HTTP/1.0"	948,091

- Data transformation: Timestamp fields are normalized, and depending on the request category, the following information is extracted from the records:
  - Searches (SR): extract query fields and query operators;
  - Full-text article views (PS, PF): extract article identifier and journal ISSN;
  - Journal, issue, or article browsing (JL, IL, and AL): extract the ISSN, volume number, and issue number of the corresponding journals; and
  - Other requests: the URLs of the requested pages.

Preprocessing reduces the size of the transaction log file from 1.3GB to 700MB. The resulting log file contains approximately 5,000,000 transaction records, each of them representing one user "click."

Table 2
Classification of SDOS transaction logs

Category	Description	Category	Description
AR	Article abstract	AL	Journal TOC (article list)
CR	Copyright notice	ES	Expanded search screen
HM	SDOS homepage	HP	Help file
IL	Journal issue list	JL	Journal list
PF	PDF view	PS	Postscript view
QQ	Misc/unknown	SS	Simple search screen
SR	Submit search	TF	TIFF image view

2. Overall analysis

This section focuses on an overall inspection of the SDOS transaction logs. The findings reported in this section concern the distribution of access over registered IP addresses, access over time of day and day of week, and types of access.

For a proper understanding of terminology, it is important to realize that essentially the SDOS system is IP restricted, and users can only access the system from known, registered (SDOS subscriber) IP addresses.[2]

2.1. IP address analysis

For every SDOS access, the IP address is recorded in the transaction log. Although IP addresses are a means of identifying users, and there must be some kind of relationship between IP addresses and individual SDOS users, many factors indicate that it is not safe to assume a one-to-one relationship:

- Users may access SDOS from shared PCs and, as a result, IP addresses from shared PCs do not uniquely identify individual users. This is especially true for PCs located in the library.
- SDOS may be accessed through proxy or caching servers, in which case the proxy or caching server IP address is logged instead of the end user's IP address. The same applies when an institution uses so-called "address translation" through a firewall or other device. This can cause all accesses coming from an institution or department to be behind one single IP address.

Analysis of log-file data shows that SDOS was accessed from 30,008 different IP addresses (see also section 4.1), but that 46.7%, close to half of the full-text views, were requested from the 100 most active IP addresses. In other words, a small fraction of hosts generates a large part of the total number of article views. The busiest IP address made 3.6% of article downloads, while even the IP address at the bottom of this top 100 made 0.1% of the full-text accesses. These findings strongly suggest that most users are hidden behind a proxy or caching server address, or firewall or other device that applies address translation, and consequently, there is not a one-to-one relationship between individual users and IP addresses.

2.2. Repeat full-text access

One metric for evaluating how an electronic resource fulfills the information need of users is users' repeat visits. For lack of a mechanism to identify individual users, only repeat visits per IP addresses can be analyzed in this study.

[2] SDOS has provided authentication by username/password since June 2000. However, the logs analyzed in this article contained only a few accesses using this authentication mechanism; therefore, these accesses were excluded from this analysis.

If the IP addresses are ranked according to the number of full-text articles downloaded (see Table 9), the bottom 13.8% of hosts download only one article, and do not come back. Consequently, 86.2% of hosts download at least two articles.

Jones et al. (2000) reported 72.8% visited CSTR only once in the period captured in the log file. Although their findings are based on visits and individual users, while this study looks at full-text downloads per IP address, it appears that one-time usage of the Taiwan SDOS system is considerably lower.

2.3. Usage over time

Figure 1 depicts the average number of SDOS accesses per hour, per time of day. SDOS usage is highest during working hours (10:00 AM–12:00 PM and 1:00 PM–5:00 PM) and lowest during the early morning hours.

Figure 2 shows the number of page views per weekday. Usage is evenly distributed over the week, with the exception of the lower than expected usage on Mondays. Usage on Saturdays is half of that on weekdays, while usage on Sundays is one third of that on weekdays.

SDOS usage and server workload are strongly related. Understanding the usage trends within a day and within a week, system administrators can determine the best schedule for system maintenance. Similarly, users can benefit from better performance by accessing SDOS during low-usage hours.

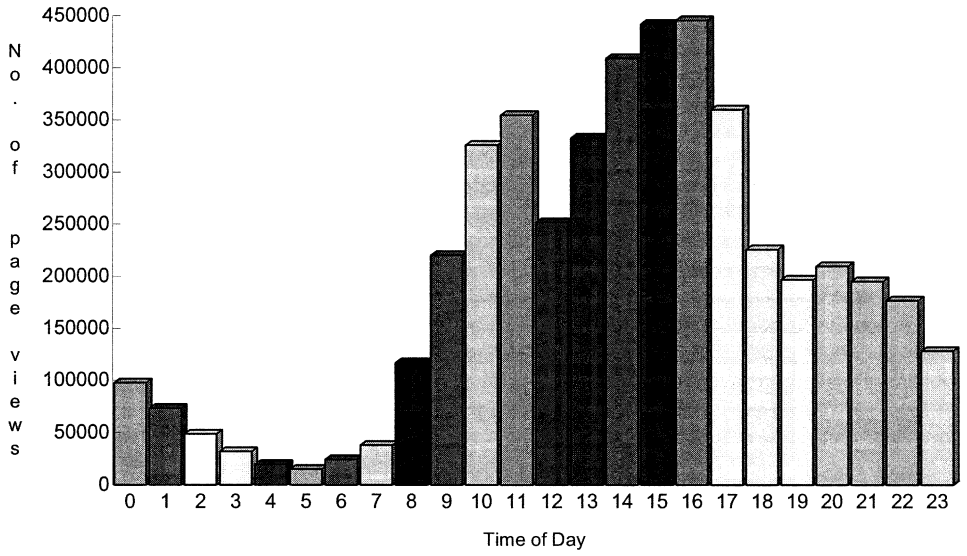

Fig. 1. Server loading per time of day.

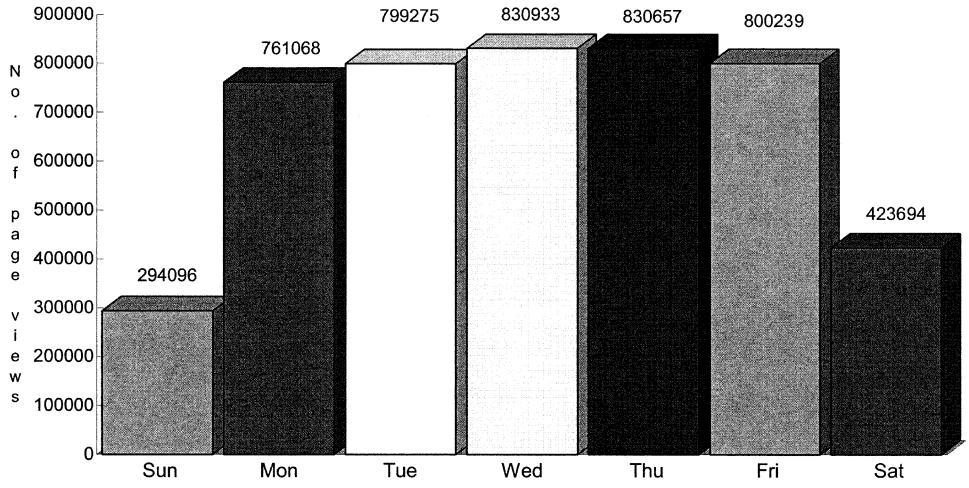

Fig. 2. Server loading within one week.

2.4. Analysis of log categories

As mentioned previously, SDOS usage can be divided into 14 categories. The distribution of the 14 categories of requests is shown in Figure 3.

Analysis shows that full-text (PDF) viewing is the most frequent type of SDOS usage, amounting to 31.7% of total access. Because SDOS's unique feature is its large full-text database, it is not a surprise that full-text article viewing is the most active usage category. Full-text articles are also available in postscript format (PS type of access), but the analysis

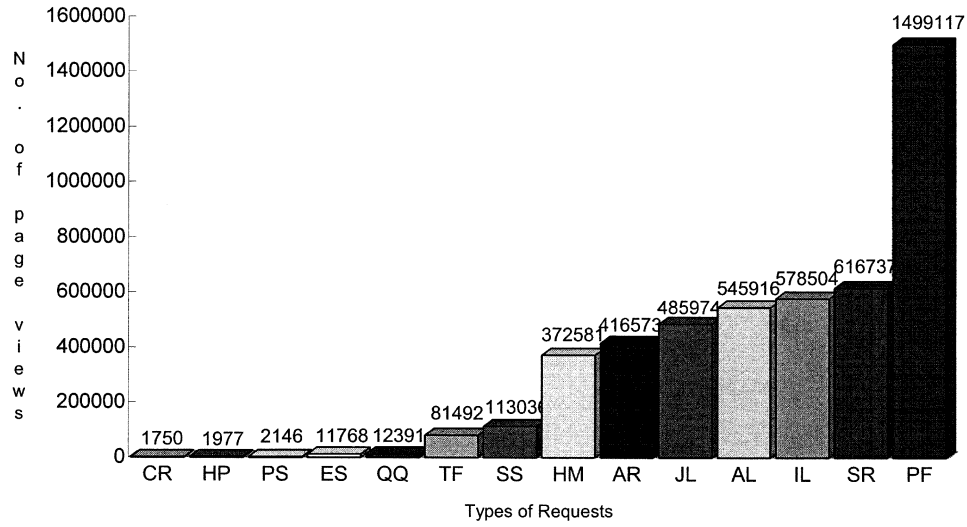

Fig. 3. Distribution of the 14 request categories (see Table 2 for a definition of the 14 request categories).

clearly shows the preference users have for viewing articles in PDF format. Because of the preference that the large majority of users have for viewing articles in PDF format, recent versions of the SDOS software no longer support the postscript option.

Another finding is that the number of article abstract views (AR) is much smaller than the number of full-text views. In other words, many users did not read the article abstracts before accessing the full text.

Full-text articles are available in PDF format only; therefore, by nature, accessing a PDF article, downloading and opening the file, causes a few seconds delay. Article abstracts are available in pure HTML format, and are available without delay.

Many possible explanations exist as to why users still prefer to access the PDF full-text article directly. These include the following:

- Because of the high speed of the Taiwan academic networks, access to the PDF files is apparently fast enough not to be an issue, and therefore users can afford the time to judge the usefulness of articles by going straight to the full text;
- Users may have read the abstract of an article in an A&I database and deemed that the article is useful; and
- Users may have accessed the article because it had been cited elsewhere, leading them to believe it was relevant to them.

Like most E-journal systems, SDOS supports two means of navigation, browsing, and searching. Before reporting how users browse, search, and access contents, the steps involved in getting to article level using these two access methods are described.

Browsing:

- Step 1: Choose "Category List of Journals" or "Alphabetical List of Journals" from the SDOS entry page. A list of categories or an alphabetic index appears. Clicking through will result in a journal list being displayed. These actions are categorized as JL (Journal List) access types.
- Step 2: Choose a specific journal from the journal-list page after which the relevant issue-list page will appear. An IL (Issue List) event is recorded.
- Step 3: After selecting a specific issue, a table-of-contents page appears, listing the articles in that particular issue. An AL (Article List) log is recorded in the log file.
- Step 4: After having selected an article, the abstract is available (AR type of access), or the full text can be viewed (PF type of access).

Searching:

- Step 1: Submit a query strategy from the Quick Search textbox in the SDOS entry page. An SR (Search) event is recorded.
- Step 2: Pick an article from the search result list to read its bibliographic page (AR type of access) or download the full text (PF type of access).

or

- Step 1: Choose Simple Search (SS) or Expanded Search (ES) from the SDOS entry page and a Simple Search or Expanded Search event is recorded.
- Step 2: Submit a query in the Simple Search or Expanded Search page and a Search (SR) event is recorded.
- Step 3: Pick an article from the search result list to read its bibliographic page (AR type of access) or download the full text (PF type of access).

Because browsing and searching serve the same purpose of navigating to (full-text) articles, there must be a strong relationship between browsing and searching and article access. Similarly, one can expect a relationship between journal and issue and article browsing. However, because users are likely to frequently use the browser's back button for navigation purposes, it is hard to interpret browsing and searching behavior or find relationships between the different access types using only log-file data, as usage of the back button is not logged and is therefore "invisible."

The numbers shown in Figure 3 indicate that roughly 31.7% of all recorded page accesses relate to full-text accesses (PF, PS), 34.0% of accesses relate to browsing (JL, IL, AL), 13.0% relate to searching (SR), and 8.8% of accesses are abstract page views (AR).

Users that start navigation using the SDOS browsing feature have two options available: 76% of users chose "Alphabetical List of Journals" and 24.0% chose "Category List of Journals."

On only a small fraction of SDOS accesses (1,977) was the SDOS online help consulted (HP type of access). In view of this, librarians and developers of E-journal systems should take action to increase use of online help, and improve its quality and accessibility. This would ensure that users are aware of the specific features of E-journal systems, the notion of query refinement, and all issues beneficial to correctly and efficiently discover the needed information. Proactive and context-sensitive mechanisms, such as *Today's Tip*, could be exploited to automatically guide users.

In addition, users rarely read the copyright disclaimer (CR). On average, only one 1 of every 22 IPs accessed the copyright announcement documents. This finding, combined with the assumption that usage terms and conditions may not always be strictly observed, reinforces the notion that libraries have to stress the significance of fair and legal use of electronic resources.

3. Analysis of query behavior

This section concentrates on SDOS searching behavior, particularly queries per IP address, query length, query modes, query operators, query refinement, and term occurrence. *Query* and *query terms* are defined as follows:

Query: One or more query terms, and possibly query operators. The end of a query is identified by the submission of a query in the Quick Search textbox in the SDOS entry page or in the Simple Search or Expanded Search page (an SR event is recorded).

Query term: A query term is any unbroken string of characters. Logical operators (AND, OR, NOT) are excluded.

3.1. Queries per IP address

As shown in Table 3, approximately half of the 30,008 valid IP addresses did not make any queries. This might be common in E-journal systems because users that know the articles' exact bibliographic information probably use the more-natural *browsing* mechanism, to locate articles. Of all users, 42.1% made 1 to 20 queries.

On the other hand, about 10.8% of users made more than 20 queries. Why did they submit so many queries? When they submitted queries, did they have a specific article in mind or did they just want to find articles pertaining to topics that they were interested in? The two questions are posed because from the authors' experiences in serving patrons, a few patrons make use of E-journal systems from an A&I database point of view. A&I databases and E-journal systems play different roles in assisting users to find necessary information. An A&I database aims to collect as much as possible the research information of a specific discipline; by comparison, an E-journal system encompasses only the research information published by a single or a few publishers. If E-journal systems are the main entry point of a user in search of information, much related information will be missed. Unfortunately, because E-journal systems provide a direct path to full texts, many patrons abandon A&I databases when searching for information. In view of this, it may be better for librarians to

Table 3
Number of queries per IP address

No. queries	No. IPs	%
0	14,130	47.09
1	2,524	8.41
2	1,804	6.01
3	1,328	4.43
4	1,067	3.56
5	858	2.86
10	2,740	9.13
20	2,316	7.72
30	1,048	3.49
40	673	2.24
50	370	1.23
60	287	0.96
70	161	0.54
80	149	0.50
90	71	0.24
100	56	0.19
> 100	426	1.42

clarify the different roles that A&I databases and E-journal systems play when conducting user training. In addition, this issue also reflects the significance of linking A&I databases and E-journal systems.

3.2. Length of query strategies

The length of a query is defined as the number of query terms contained in it. The distribution of the query length in SDOS is shown in Table 4. The queries with a length of 0 resulted from the submission of queries without entering any query terms. Approximately 85.2% of queries contained one, two, or three terms, although the average query length was 2.27 terms. This average length is close to the average lengths reported for CSTR (2.43; Jones et al., 2000) and Excite (2.21; Jansen et al., 2000). However, this average length is very different from the average lengths of query strategies composed of English query terms for Dreamer (1.10) and GAIS (1.22) reported in Pu (2000). This observation emphasizes the need to further investigate several issues. For example,

- Do users of Internet search services and E-journal systems exhibit similar search behavior in terms of query length (Jansen et al., 2000; Jones et al., 2000; and the current study report similar results)?
- Is there a structural difference in terms of query length between Asian and Western users of E-journal systems (Jones et al., 2000, and the current study report similar results)?
- Do Asian users of Internet search services and E-journal systems exhibit different search behavior in terms of length of English-language query strategies (Pu, 2000, and the current study report different results)?

Table 4
Distribution of query length

Query length	No. Queries	%
0	596	0.14
1	182,754	42.97
2	104,095	24.48
3	75,268	17.70
4	27,900	6.56
5	16,399	3.86
6	6,086	1.43
7	3,979	0.94
8	2,102	0.49
9	1,625	0.38
10	1,169	0.27
11–15	2,589	0.61
16–20	602	0.14
20+	122	0.03

Typically, queries containing a limited number of keywords result in relatively large result sets. Without further refinement of query strategies, it is not an easy task for users to locate relevant information. In view of this, the following two suggestions are proposed:

- Librarians should educate users about query refinement and the use of query operators for more effective searching; and
- Electronic resources should exploit techniques, such as relevance feedback, query expansion, and data mining, to learn the interests of users, and proactively lead users to relevant information.

3.3. Query modes

When searching in SDOS, many (bibliographic) key fields can be used. In many circumstances, users can increase the likelihood of locating relevant articles by limiting their queries to particular fields in an article. The fields available for SDOS searches include Author Name, Article Title, Abstract, Journal Title, Author Keywords, ISSN, Publisher Item Identifier (PII), Article Full Text, and Journal Category. To facilitate the formation of query strategies, SDOS offers users two query modes: Simple Search and Expanded Search. Simple searches allow users to match their queries against any of the previously described fields or limit their queries to a specific field. Expanded searches allow users to specify a maximum of two queries that match any field or a specific field. These two queries can be combined via Boolean operators—AND, OR, NOT. In addition, several query options are available in expanded search, which allow users to specify journal categories, article types, article language, date-range limitations, ranking options, and the maximum number of returned articles displayed on each page.

A total of 90.8% of the queries were of the Simple Search type, while only 9.2% of the queries were of the Expanded Search type. Because a simple search is the default search mode, this phenomenon indicates that most users do not switch to a more powerful query mode. However, a simple query mode is inadequate for advanced and professional searchers because it cannot filter out irrelevant information effectively. Many electronic resources therefore offer two query modes, one for novice and the other for advanced searchers. Some even offer a query mode for professional searchers.[3] SDOS usage suggests that users do not use advanced search functionality, despite the benefits.

3.4. Query fields

Table 5 shows the distribution of the key fields in user queries. ‘‘Any Field’’ is the default query field, matching any of the fields that can be searched, and is used in 84.4% of simple searches. On the other hand, about half (49.3%) of Expanded Search usage included fields

[3] As of SDOS 4.0, three query modes are available in SDOS: Simple Search, Expanded Search, and Expert Search.

Table 5
Distribution of query fields

	Simple search		Expanded search	
Fields	No. queries	%	No. queries	%
Any field	197,983	84.4	104,847	50.7
Author's name	16,455	7.0	28,779	13.9
Article title	13,021	5.6	34,245	16.6
Abstract	7,009	3	22,004	10.6
Journal title	NA	NA	7,956	3.9
Author keywords	NA	NA	2,567	1.2
ISSN	NA	NA	NA	NA
PII	NA	NA	192	0.1
Text	NA	NA	662	0.3
Category	NA	NA	5,447	2.7

NA = not applicable.

other than the default field. Article Title, Author's Name, and Abstract are the three query fields most frequently used in Expanded Search mode. The ISSN and PII fields are seldom used in either of the query modes, probably because the two fields are mostly searched by librarians; therefore, their relatively low use by end users seems reasonable.

3.5. Query operators

In addition to the Boolean AND, OR, and NOT operators, SDOS supports several query operators, including operators for wildcard search (*), proximity search (NEAR), adjacent search (ADJ), and fuzzy search (Soundex and Typo[4]). As shown in Figure 4, the Boolean "AND," the wildcard operator " * ", and Boolean "OR" were the three operators used most frequently. Because many Internet search services and electronic resources offer the three operators, users may be very familiar with these operators; therefore, their usage was high. As for other operators, perhaps users who had not consulted the SDOS online help document were unaware that these operators were available. Because query operators help users to search more effectively, librarians need to emphasize the usage of query operators when training users.

3.6. Query refinement

An ideal IR system only retrieves relevant documents. In reality, however, IR systems retrieve many nonrelevant documents, and it is nearly impossible for a user to retrieve all relevant information via a single query. Experienced searchers often refine queries as part of a search process.

[4] The Soundex operator matches keywords that "sound like" a term a user specifies. The Typo operator expands a term that a user specifies to keywords with a similar spelling.

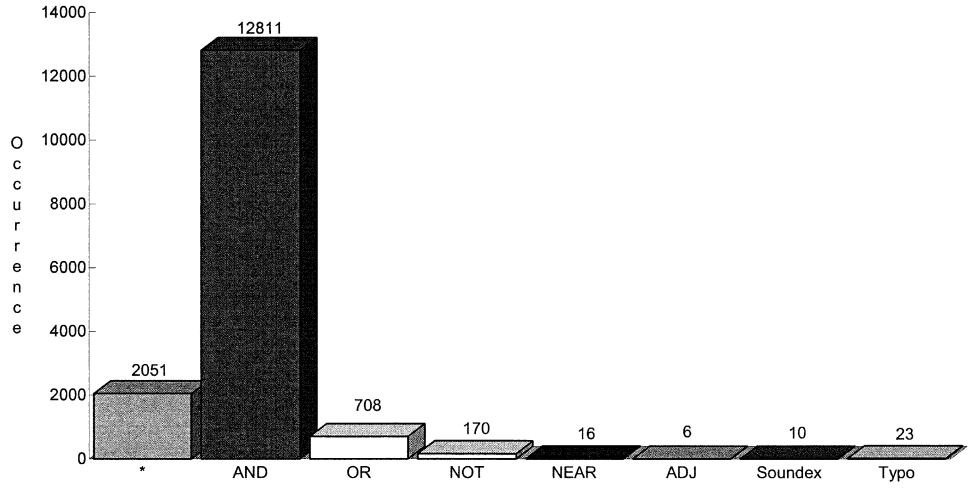

Fig. 4. Boolean and query operators used in queries.

The query strategies are classified into three types for analyzing the query refinement behavior: unique query, identical query, and modified query.[5] A unique query does not have any terms in common with its preceding query. An identical query is identical to its preceding query. A modified query has some (but not all) terms in common with its preceding query. As presented in Figure 5, the proportions of unique query, identical query, and modified query were 55%, 13%, and 32%, respectively. Identical queries were ignored because it was impossible to interpret them meaningfully in the context of query refinement.

Suppose a unique query indicates that a user changes his or her information need and a modified query represents that a user refines his or her query strategy based on the result of a preceding query. The percentage of modified queries in relation to unique and modified queries suggests a clue to query refinement. In this study, this percentage is 36.8% (32%/[32% + 55%]), which is similar to the 38% presented in Jansen et al. (2000) but significantly lower than the 66% presented in Jones et al. (2000). According to this analysis, query refinement in SDOS may not be a usual action.

A further study was conducted to examine how users refine their queries. The result is shown in Figure 6. Percentages in Figure 6 are based on the number of queries in relation to all modified queries. About half of the modified queries did not change the number of terms, compared with their preceding queries. In other words, users only substitute query terms without changing the number of terms. As indicated in Figure 6, users typically do not refine queries extremely with respect to the number of terms in their consecutive queries. Modification to queries is done in small increments or decrements; in addition, users both increase and decrease terms.

[5] The three categories were also proposed in Jansen et al. (2000); however, their and this study's definition of unique query is somewhat different.

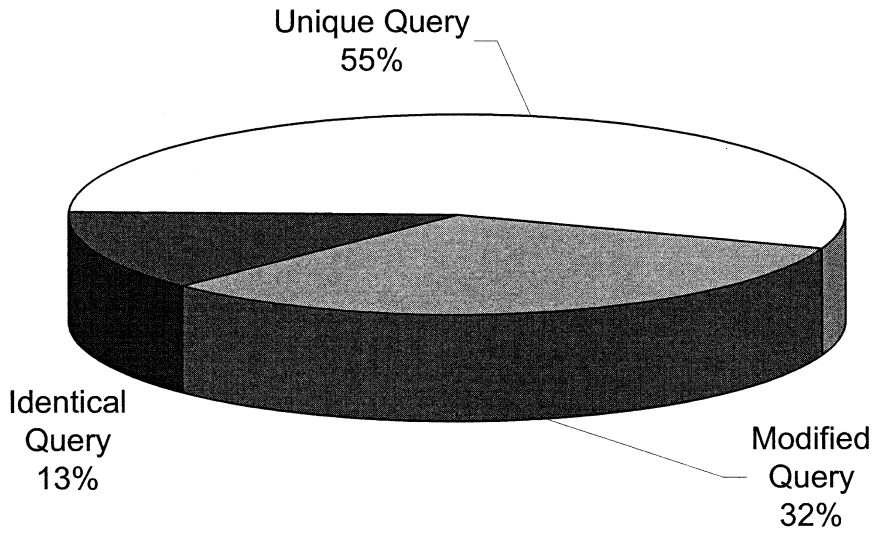

Fig. 5. Proportions of identical, unique, and modified queries.

3.7. Term occurrences

A total of 887,309 terms occurred in all queries. After eliminating duplicate terms, 81,014 unique terms remained. Table 6 shows the distribution of term frequency. About half of the unique terms (49.7%) were used once, but only 623 terms (0.8%) were used more than 200 times.

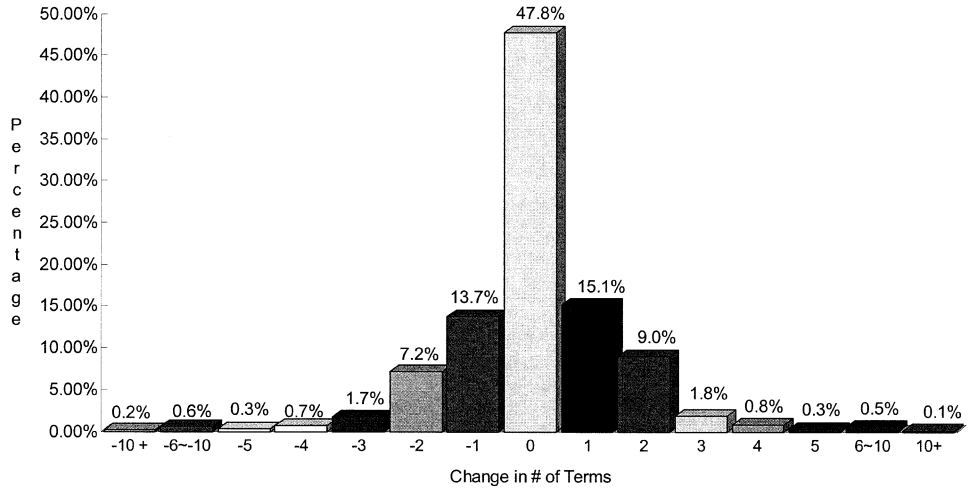

Fig. 6. Changes in number of terms in consecutive modified queries.

Table 6
Distribution of term frequency

Frequency	Query terms	%
1	40,237	49.67
2	13,117	16.19
3	6,199	7.65
4–5	6,267	7.74
6–10	5,944	7.34
11–20	3,790	4.68
21–50	2,947	3.64
51–100	1,184	1.46
101–200	706	0.87
200+	623	0.77

A complete rank-frequency table was built for all terms. From the complete rank-frequency table, the top 50 most commonly occurring terms were extracted and shown in Table 7. The

Table 7
The 50 most commonly occurring query terms

Rank	Term*	Occurrence	Rank	Term	Occurrence
1	***and***	62,203	26	chain	1,556
2	***of***	13,707	27	surface	1,496
3	***in***	4,655	28	design	1,418
4	***the***	3,658	29	supply	1,402
5	acid	3,422	30	heat	1,395
6	***for***	2,854	31	copper	1,366
7	review	2,760	32	thermal	1,362
8	cell	2,388	33	data	1,322
9	carbon	2,355	34	plasma	1,313
10	management	2,289	35	transfer	1,269
11	analysis	2,240	36	silicon	1,262
12	***a***	2,152	37	network	1,235
13	control	2,013	38	sensor	1,205
14	polymer	1,979	39	neural	1,198
15	protein	1,937	40	DNA	1,191
16	system	1,893	41	stress	1,186
17	fuzzy	1,831	42	phase	1,174
18	film	1,800	43	diamond	1,166
19	***&***	1,737	44	thin	1,146
20	model	1,725	45	plant	1,144
21	***on***	1,708	46	flow	1,135
22	***or***	1,699	47	information	1,114
23	water	1,653	48	membrane	1,104
24	oxide	1,616	49	process	1,095
25	metal	1,574	50	GaN	1,090

* Terms in italic and boldface are stop words.

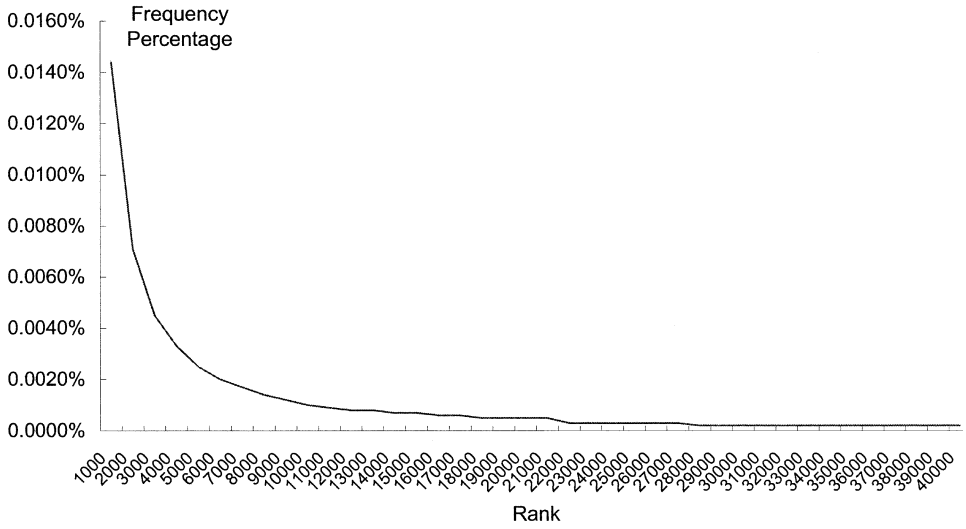

Fig. 7. Rank-frequency graph of all terms.

top 50 terms represented only 0.1% of all unique terms, yet they appeared 159,192 times, accounting for 17.9% of all terms in all queries.

Figures 7 and 8 depict the rank-frequency relation of query terms. From Figure 7, the 1,000th frequently appearing term accounted for 0.016% of all 887,309 terms in all queries,

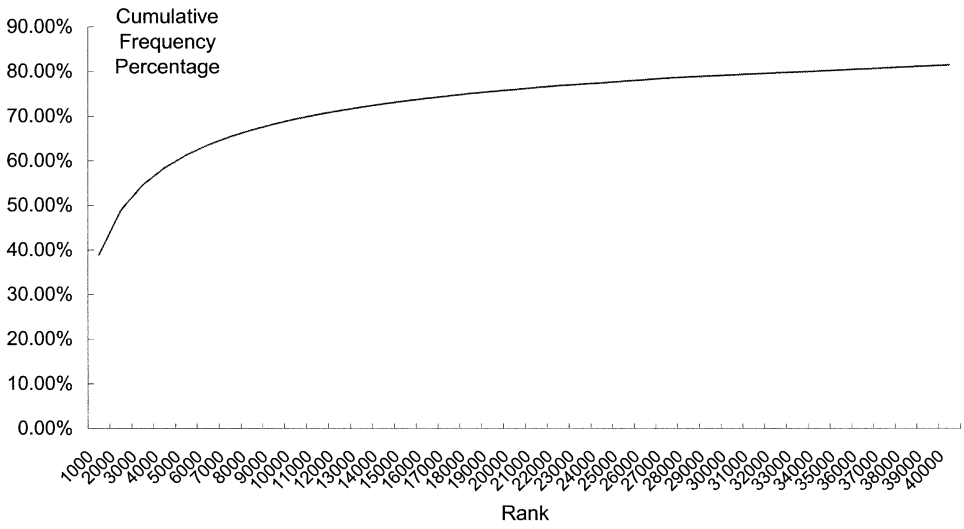

Fig. 8. Rank-frequency graph of all terms, showing cumulative frequency percentage.

the 3,000th accounted for 0.00494%, and the 5,000th accounted for 0.00272%. Figure 7 is highly skewed. At the end of the high-ranking terms, the curve descends abruptly, expressing a relatively large difference in the frequency of two adjacently ranking terms; at the end of low-ranking terms, the curve becomes smooth, representing terms with a frequency of one. Figure 8 is a cumulative total. The cumulative frequency of the top 1,000 terms accounted for 38%, the top 3,000 accounted for 59%, and the top 5,000 accounted for 61%. As indicated in Table 6 and Figures 7 and 8, a few terms were used repeatedly and many terms were used only once. This result is consistent with the studies presented in Jansen et al. (2000), Jones et al. (2000), and Pu (2000).

Nine of the top 50 terms shown in Table 7 are stop words (and, of, in, the, for, a, &, on, or). In general, the use of stop words does not affect the query results or impose a heavy processing load on the system; nevertheless, if users understand the use of stop words, they can focus on their information need when planning their query strategies. Furthermore, "or," "and," and "&" are commonly used in Internet search services and electronic resources as query operators ("or" for Boolean OR; "and" and "&" for Boolean AND). Mistakenly using these stop words in queries may lead to results diverging from the users' original intent. For instance, in previous versions of SDOS (including the version from which this study collected transaction logs), "or" and "and" are stop words but not query operators. If a user submits a query "digital or communication" (with the information need "digital OR communication" in mind), this query will be interpreted as "digital communication," whose meaning is "digital AND communication" and different from the original query. The current version of SDOS considers that "or" and "and" have the same meaning as query operators OR and AND, respectively. The use of "&" causes a more severe problem. The current version of SDOS does not treat "&" as a stop word or query operator; therefore, if a user submits a query "digital & communication" (with the information need "digital AND communication" in mind), this query will be interpreted as "digital AND & AND communication." The data show that, the query "digital AND communication" returned 372 results, but the query "digital AND & AND communication" returned no results.

4. Analysis of download behavior

4.1. Article downloads

Over the full period of analysis, of the approximately 1,000,000 articles available in SDOS, 1,503,418 requests for full-text views were recorded in the log files relating to 346,776 unique articles. As indicated in Table 8, 49.5% of the 346,776 articles were downloaded once, and 90.3% of articles were downloaded less than five times.

Table 9 lists the number of article downloads per IP address (host). Of the 30,008 valid hosts (IP addresses) that accessed the SDOS system during the period of the study, 22,282 hosts downloaded one or more articles. Among those, 38.4% hosts downloaded 1 to 5 articles, 14.4% downloaded 6 to 10 articles, 29.9% downloaded 11

Table 8
Frequency of downloads per article

Downloads per article	No. articles	%	Downloads per article	No. articles	%
100+	19	0.01	11	1,667	0.48
51–100	117	0.03	10	2,260	0.65
21–50	1,595	0.46	9	3,144	0.90
20	255	0.07	8	4,153	1.19
19	269	0.08	7	6,219	1.78
18	357	0.10	6	9,177	2.63
17	409	0.12	5	13,326	3.82
16	533	0.15	4	22,314	6.40
15	663	0.19	3	40,249	11.54
14	806	0.23	2	66,121	18.96
13	1,055	0.30	1	172,744	49.53
12	1,310	0.38	—	—	—

to 50 articles, 7.3% downloaded 61 to 100 articles, and 8.9% downloaded more than 100 articles.

4.2. Downloading subscribed versus nonsubscribed journal articles

Most Taiwanese SDOS participating institutions (SDOS customers) have subscriptions to the print versions of only a portion of the journals available in electronic format in SDOS. For the year 2000, as a temporary arrangement, Elsevier Science allowed SDOS subscribers to access the full SDOS collection, including full-text access to journal titles that were not subscribed to in print. It is interesting to see to what extent SDOS customers access titles not subscribed to in print. From January to July of 2000, 1,149,974 article downloads were requested. The majority of downloads, 735,810 (64.0%) were for articles in nonsubscribed journals.

Table 9
Number of article downloads per IP address

No. downloads	No. IPs	%	No. downloads	No. IPs	%
1	3,078	13.81	21–30	1,652	7.41
2	1,909	8.57	31–40	1,082	4.86
3	1,563	7.01	41–50	798	3.58
4	1,122	5.04	51–60	580	2.60
5	885	3.97	61–70	446	2.00
6	916	4.11	71–80	347	1.56
7	667	2.99	81–90	261	1.17
8	586	2.63	91–100	236	1.06
9	569	2.55	101–200	1,069	4.80
10	471	2.11	201–300	367	1.65
11–20	3,129	14.04	>300	549	2.46

Table 10
Distribution of nonsubscribed articles, access ratios over SDOS customer basis

Nonsubscribed ratio (see Eq. 1)	No. Customers	%
0%–37.5%	2	3.92
37.5%–50%	5	9.80
50%–62.5%	5	9.80
62.5%–75%	9	17.65
75%–87.5%	16	31.37
87.5%–100%	14	27.45

A further study was conducted to understand the tendency to use nonsubscribed journal articles. Equation 1 calculates the ratio of frequency for downloading nonsubscribed journal articles to that for downloading all journal articles for each customer. Among the 51 SDOS customers, the minimum ratio is 26.2% and the maximum is 99.2%. Table 10 shows the distribution of this ratio for the 51 SDOS customers. A considerable proportion of customers (76.5%) have a nonsubscribed access ratio greater than 62.5%; in other words, more than 62.5% of articles accessed by users belonging to these institutions are nonsubscribed.

$$Ratio = \frac{Download\,Frequency\,of\,Non\text{-}Subscribed\,Journal\,Articles}{Download\,Frequency\,of\,All\,Journal\,Articles} \quad \text{(Eq. 1)}$$

Another indicator would be to examine the number of journal titles that have had articles downloaded by customer users. After inaccuracies were excluded, this indicator was calculated for 34 customers using Equation 2, the result of which is shown in Table 11. Once more, a majority of journal titles, whose articles had been downloaded by almost all of the 34 customers, were not held in print.

$$Ratio = \frac{Non\text{-}Subscribed\,Journal\,Titles\,Downloaded}{All\,Journal\,Titles\,Downloaded} \quad \text{(Eq. 2)}$$

For each journal, the ratio of download frequency by nonsubscribing customers to that by all customers was calculated according to Equation 3. When computing this ratio for 822

Table 11
Distribution of nonsubscribed journals, access ratios (34 SDOS customers only)

Nonsubscribed Ratio (see Eq. 2)	No. Customers	%
0%–50%	1	2.94
50%–60%	0	0
60%–70%	3	8.82
70%–80%	4	11.76
80%–90%	15	44.12
90%–100%	11	32.35

Table 12
Journal distribution of nonsubscribed access ratio

Nonsubscribed Ratio (see Eq. 3)	No. Titles	%
0%–10%	4	0.5
10%–20%	12	1.5
20%–30%	31	3.8
30%–40%	56	6.8
40%–50%	82	10.0
50%–60%	114	13.9
60%–70%	109	13.3
70%–80%	130	15.8
80%–90%	136	16.5
90%–100%	148	18.0
Total	822	100.1*

* Subject to rounding.

journal titles, as indicated in Table 12, 77.4% of the journals had a ratio greater than 50%. In other words, three fourths of the journals examined are mostly accessed by institutions that did not subscribe to the journal in print.

$$Ratio = \frac{Articles\,Downloaded\,by\,Non\text{-}Subscribing\,Customers}{Articles\,Downloaded\,by\,All\,Customers} \qquad \text{(Eq. 3)}$$

Based on the previous analysis, two issues are worth further investigation: First, do the journals subscribed by a library meet the requirement of its users? To answer this question,

Table 13
Ranking of journal categories selected in Expanded Search mode

Rank	Journal category	Frequency	%
1	Engineering, energy, and technology	1,788	14.4
2	Materials science	1,482	11.9
3	Physics and astronomy	1,386	11.1
4	Computer science	1,351	10.9
5	Chemistry and chemical engineering	1,344	10.8
6	Social sciences	1,154	9.3
7	Environmental science and technology	983	7.9
8	Agricultural and biological sciences	803	6.5
9	Life science	710	5.7
10	Mathematics	607	4.9
11	Clinical medicine	579	4.7
12	Earth and planetary science	255	2.0
	Total	12,442	100.1*

* Subject to rounding.

Table 14
Ranking of journal categories selected in Browse mode

Rank	Journal category	Frequency	%
1	Chemistry and chemical engineering	13,767	21.0
2	Materials science	9,047	13.8
3	Life science	6,881	10.5
4	Engineering, energy, and technology	5,820	8.9
5	Environmental science and technology	5,752	8.8
6	Computer science	5,648	8.6
7	Social sciences	4,955	7.6
8	Physics and astronomy	3,452	5.3
9	Clinical medicine	2,802	4.3
10	Agricultural and biological sciences	2,739	4.2
11	Earth and planetary science	2,661	4.1
12	Mathematics	1,921	2.9
	Total	65,445	100

the article download frequency of individual SDOS journals can be computed for each customer. Second, should publishers appreciate the tendency toward downloading non-subscribed journal articles and review their pricing strategies? In view of this tendency,

Table 15
The 20 most commonly downloaded journals

Rank	Journal title	Download frequency
1	*Synthetic Metals**	28,424
2	*Thin Solid Films**	26,574
3	*Tetrahedron Letters**	23,658
4	*Sensors and Actuators A: Physical**	21,576
5	*Journal of Power Sources**	17,867
6	*Atmospheric Environment**	16,332
7	*Journal of Chromatography A**	15,549
8	*Polymer**	15,407
9	*Analytica Chimica Acta**	13,845
10	*Journal of Crystal Growth**	12,389
11	*European Journal of Operational Research*	11,695
12	*Surface and Coatings Technology*	11,198
13	*Fuzzy Sets and Systems**	11,171
14	*Materials Science and Engineering: A*	11,038
15	*Journal of Materials Processing Technology*	10,561
16	*Sensors and Actuators B: Chemical*	10,233
17	*Journal of Alloys and Compounds*	9,749
18	*Fuel and Energy Abstracts*	9,595
19	*Electrochimica Acta**	9,488
20	*Tetrahedron**	9,435

Journal titles marked with an asterisk also appear in Table 16.

Elsevier Science adjusted its pricing strategy in 2001 so that customers can access all the SDOS journals for a limited extra fee.

5. Miscellaneous analysis

5.1. Ranking of journal categories

A user can pick a specific journal category to search and browse in "Expanded Search" and "Browse." Tables 13 and 14 show the ranking of journal categories selected in Expanded Search and Browse, respectively. In Expanded Search, the top three categories are "Engineering, Energy, and Technology," "Materials Science," and "Physics and Astronomy"; in Browse, the top three are "Chemistry and Chemical Engineering," "Materials Science," and "Life Science." The only overlapping category is "Materials Science." This indicates that users majoring in different categories possess different usage manners. Overall, users prefer browsing a specific category to searching a specific category.

5.2. Ranking of the most popular journals

Tables 15 and 16 list the 20 most commonly downloaded and browsed journals, respectively.

Table 16
The 20 most commonly browsed journals

Rank	Journal titles	Browse frequency
1	*Tetrahedron Letters**	22,322
2	*Journal of Chromatography A**	11,530
3	*Atmospheric Environment**	10,975
4	*Polymer**	9,267
5	*Thin Solid Films**	9,065
6	*Synthetic Metals**	8,488
7	*FEBS Letters*	7,697
8	*Trends in Biochemical Sciences*	7,567
9	*Trends in Biotechnology*	6,892
10	*Tetrahedron**	6,878
11	*Journal of Power Sources**	6,874
12	*Biomaterials*	6,227
13	*Journal of Crystal Growth**	6,198
14	*International Journal of Heat and Mass*	6,106
15	*Analytica Chimica Acta**	6,053
16	*Electrochimica Acta**	5,988
17	*Journal of Membrane Science*	5,705
18	*Fuzzy Sets and Systems**	5,657
19	*Sensors and Actuators A: Physical**	5,524
20	*Trends in Cell Biology*	5,512

Journal titles marked with an asterisk also appear in Table 15.

6. Conclusion and future research

This article analyzed usage of the Taiwan-based SDOS E-journal system, one of the largest and most heavily used full-text STM databases worldwide.

To increase effectiveness and usage levels of the SDOS system in particular and E-journal systems in general, suggestions for librarians, developers of electronic resources, and administrators of electronic resources are as follows:

Librarians:

- Instruct users about the most relevant and useful features of electronic resources to compensate for the fact that users seldom read online help documentation;
- Inform users about usage terms and conditions and the importance of fair and legal use to compensate for the fact that users seldom read the copyright notice; and
- Educate users about search functionality and strategies to increase effectiveness of use of electronic resources.

Developers of electronic resources:

- Make use of proactive and context-sensitive mechanisms to automatically inform users about the distinctive system features and usage terms and conditions;
- Provide multiple query modes to satisfy the needs of novice and advanced users and professional searchers; and
- Make use of mechanisms such as relevance feedback and query expansion to enhance the interaction between users and systems, and then assist users in formulating more effective queries.

Administrators of electronic resources:

- Understand the server load patterns over time and take them into consideration when scheduling hardware or software maintenance tasks; and
- Achieve an optimal trade-off among cost, efficiency, and size of storage devices, by taking into consideration the frequency of use of (segments of) electronic collections.

Transaction logs provide a quantitative description of system usage and general user behavior over a specific period. However, if users are anonymous, and no identification is possible beyond the IP address being used (as is the case in this study), no user level information is available. By means of only the SDOS transaction logs, individual users, user sessions, or repeat use cannot be tracked, and therefore information-seeking needs and user satisfaction cannot be determined.

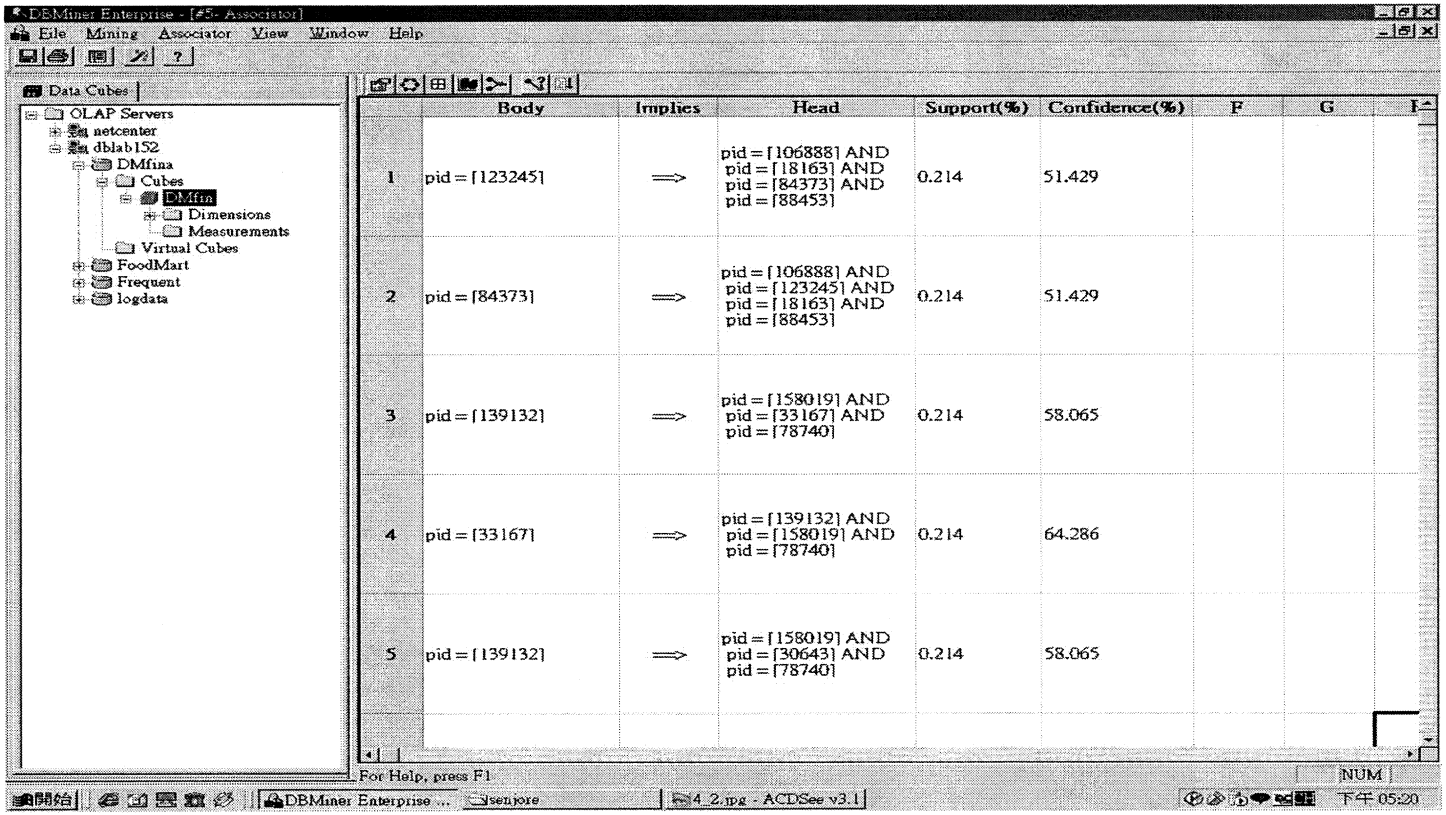

Fig. 9. An association-mining example using DBMiner 2.0.

Qualitative and quantitative user–behavior analysis, such as online surveys, real-time observation, real-time surveys, and feedback gathering, is complementary to the transaction log–based quantitative usage analysis. General usage patterns can be identified by means of transaction log analysis, and qualitative analysis tools can be applied to understand the causes of the phenomena observed. Several issues are worth further investigation through qualitative user–behavior analysis:

- The similarity and difference of users' behavior in using paper-based journals and electronic journals;
- The reasons why many SDOS users download full texts of articles without reading bibliographic pages and abstracts;
- The similarity and difference of motivation, information need, and behaviors of users of electronic resources such as SDOS and Internet search services; and
- The influence of the linking mechanisms between A&I databases and E-journal systems on information-seeking behavior of users.

This study uses statistical methods for log-file analysis. In the future, data-mining techniques will be used. One possible direction is discovering association rules among SDOS articles and exploiting the rules to make recommendations to users based on queries they execute or articles they access.

A data-mining package, DBMiner 2.0 (DBMiner, 2002), was used to see what can be learned from further log-file analysis. In this study, focusing on SDOS articles of which the full texts had been downloaded more than 30 times, 77 association rules were discovered, some of which are illustrated in Figure 9. For example, the first rule in Figure 9 shows that if a user has downloaded a full-text article with identifier (pid) 123245, then the possibility that the user has also downloaded full-text articles with pid 106888, 18163, 84373, and 88453 is 51.429% (Confidence); furthermore, the percentage of users that downloaded these five full-text articles is 0.214% (Support).

An E-journal system can make use of the discovered association rules to guide users when they browse or issue queries. For instance, when a user finds a useful article, the system can display the titles of the articles that have been downloaded by other users who also downloaded the article. Furthermore, data mining can be leveraged to investigate the association between query terms and articles selected from search result lists. The system can then use the discovered rules to recommend articles to users based on queries issued. In this manner, an E-journal system can proactively and intelligently assist users in finding the information to meet their information-seeking goals.

Acknowledgments

The authors thank Yi-Fen Chen, Ming-Chun Yang, Dai-Yi Wang, and Chung-Wei Huang for their help with several statistics reported in this article and the preliminary data-mining study.

References

Borghuis, M., Brinckman, H., Fischer, A., Hunter, K. van der Loo, E., ter Mors, R., Mostert, P., & Zijlstra, J. (1996). *The TULIP final report* (chapter 4 on user behavior). Retrieved March 2, 2002, from http://www.elsevier.nl/homepage/about/resproj/trchp4.htm

Bruce, C. (1995). Information literacy: A framework for higher education. *Australian Library Journal*, *44* (3), 158–170.

Choo, C. W., Detlor, B., & Turnbull, D. (1999). Information seeking on the Web — An integrated model of browsing and searching. In L. Woods (Ed.), *Proceedings of the 62nd ASIS Annual Meeting* (pp. 3–16). Medford, NJ: Information Today.

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the World Wide Web. *Computer Network and ISDN Systems*, *27*, 1065–1073.

DBMiner Technology Inc. Homepage. (2002). Retrieved June 19, 2002, from http://www.dbminer.com.

Doyle, C. (1992). *Outcome measures for information literacy within the national education goals of 1990*. Rockville, MD: Educational Resources Information Center (ERIC Document Reproduction Service No. ED 351 033).

Eason, K., Yu, L., & Harker, S. (2000). The use and usefulness of functions in electronic journals: The experiences of SuperJournal. *Program*, *34*, 1–28.

Hawk, W. B., & Wang, P. (1999). User's interaction with the World Wide Web: Problems and problem solving. In L. Woods (Ed.), *Proceedings of the 62nd ASIS Annual Meeting* (pp. 256–270). Medford, NJ: Information Today.

Hert, C. A., & Marchionini, G. (1998). Information seeking behavior on statistical Websites: Theoretical and design implications. In C. M. Preston (Ed.), *Proceedings of the 61st ASIS Annual Meeting* (pp. 303–314). Medford, NJ: Information Today.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing & Management*, *36*, 207–277.

Jones, S., Cunningham, S. J., McNab, R., & Boddie, S. (2000). A transaction log analysis of a digital library. *International Journal on Digital Libraries*, *3*, 152–169.

Kaplan, N. R., & Nelson, M. L. (2000). Determining the publication impact of a digital library. *Journal of the American Society for Information Science*, *51* (4), 324–339.

Peters, T. A. (1993). The history and development of transaction log analysis. *Library Hi Tech*, *11* (2), 41–58.

Pu, H. T. (2000). An exploratory analysis on search terms of network users in Taiwan [in Chinese]. *National Central Library Bulletin*, *89* (1), 23–37.

Spink, A., Wilson, T., Ellis, D., & Ford, N. (1998). Modeling user's successive searches in digital environments. *D-Lib Magazine*, *4* (4). Retrieved November 12, 2001, from http://www.dlib.org/dlib/april98/04spink.html.

Zhang, Z. (1999). Evaluating electronic journals and monitoring their usage by means of WWW server log analysis. *Vine*, *111*, 37–42.