土木学会図書館目録・書誌情報検索システムのログ分析について

森岡　倫子*　　野末　道子**　　嶋田　真智恵***

寺尾　洋子　　上田　修一****

本研究では、Web OPAC 並びに書誌情報システムである土木学会図書館システムの利用者検索ログを対象として、利用の多いデータベース、演算子の利用状況等を調査し全体的な傾向を把握した。さらに個々の利用者についての分析を行うために、IP アドレスを手がかりとして認識された個人の一連の検索フローをセッションとしてまとめた。このセッション別に、利用者の検索が主題検索、既知事項検索であるのかを分類し、検索の失敗と考えられる 0 ヒットやヒット過多を出した場合の理由とその後の対応、一覧表示画面から個別書誌事項の詳細表示を行った後の行動パターンについて分析を行い、ユーザの検索特徴の分析を試みた。

キーワード：書誌データベース，Web　OPAC ，情報検索，エンドユーザ、検索行動

１．はじめに

土木図書館がインターネット上で公開している蔵書検索システムは，2002 年から新たなシステムとなり，土木学会に設置されたファイルサーバー上にて運用されている。

本調査では、この検索システム上に残されているシステムログを利用して、利用の多いデータベース、演算子の利用状況等を調査し全体的な傾向を調査した。さらに、このシステムログを加工・編集することにより、個人毎の検索の一連の流れ（以下、セッションと言う）を作成したものを用いて、主に利用者の検索失敗から、検索行動の特徴を抽出することを試みた。

２．　調査対象と手法

（１）土木図書館蔵書検索システムの概要

土木図書館において，現在供用されている蔵書検索システムは，2002（平成 14）年５月の土木図書館のリニューアルにともない運用が開始された。検索対象となるデータの構成を表１に，検索システムの構成を表２に示す。なお，2003 年 12 月から土木学会８支部の年次講演概要集（約６万件）の書誌情報と土木関連雑誌（32 種 20 年分６万件）の検索についても，試験的に運用している。

表１　土木図書館の検索システムが対象とするデータの構成

目録（蔵書）検索システム	a.和図書　約 30,000 冊，b.洋図書　約 4,000 冊 c.和雑誌　約 600 種，d.洋雑誌　約 200 種
書　誌（学会論文）検索システム	a. 学会誌（1915 年創刊号～）約１万１千件 b. 学会論文集（1944 年創刊号～）約２万３千件 c. 年次学術講演会論文集（1937 年創刊号～）約７万 8 千件 d. 各種委員会論文集（約 70 種・創刊号～）約５万３千件

表２　土木図書館の検索システム構成

ハード	a. 検索用サーバ：１台（PC サーバ,PIII,1GB 72GBHDD） b. 管理用 PC：３台（データ登録・更新用,Celeron1GB 40GBHDD） c. 閲覧用 PC：４台（館内検索のみ） d. インターネット回線：B フレッツ（100MBps）
ソフト	a.web 用検索エンジン：NAMAZU b.登録・更新プログラム：Microsoft 社，ACCESS（VB，ASP） c.WEB サーバ：apache d.OS：turbo linux

*国立音楽大学附属図書館　QYB00077@nifty.ne.jp
**財）鉄道総合技術研究所　　輸送情報技術研究部（設備システム）　michiko@rtri.or.jp
***国立国会図書館
****慶應義塾大学文学部図書館情報学科

（２）調査の方法と期間

①　検索システム利用件数の推移

2002 年５月～2004 年２月までの 22 ヶ月間を対象としたアクセス件数を把握した。

②　ユーザの全体的な利用傾向の把握

ログの全数概要と、各検索コマンド中の利用データベース別検索演算子の利用状況、各検索式に対して得られたヒット数を調査した。

③　ユーザの検索セッションの分析

個別のユーザの傾向を分析するために、検索エンジンである NAMAZU のログデータならびに Web サーバのログデータの 2 種類を加工、編集して、ユーザの検索セッションを作成した。NAMAZU のログデータではどのキーワードでどの書誌データベースを検索し、何件のヒットがあったかがわかる。一方 Web サーバのログデータには、どのキーワードで、選んだ書誌データベース、出力件数指定、一覧表示や詳細結果表示についての情報がある。

本研究では、便宜的に同一日付の中での同一 IP アドレスからのアクセスを１セッションとして、まず機械的に取り出すことを試みた。しかし、実際にはこの機械的な判定では不充分なため、入力コマンド間の時間のあいているセッションについては、検索内容を見て、別セッションという判断を人手で行った。更に、アクセス元のサイト環境によっては、動的に IP アドレスを割り当てるために、同一セッションであっても別の IP アドレスが混在したログとして記録されている場合もあり、これについてはキーワードの傾向を見て人手により同一であるかどうかの推定を行い、複数の IP アドレスをまとめたデータを再度時間順に並べ替えを行って、同一セッションとしてまとめた。最後に個々の検索コマンドに対する検索結果の件数を NAMAZU のログデータから抽出した。

この編集プロセスを経て認識されたセッションごとに、利用者の検索が「主題検索」、「人名検索」、「人名以外の既知事項検索」のいずれかで分類した。ここで主題検索は**「主題キーワードで検索した検索式を含むもの」**、人名検索は**「検索の最初から最後まで人名による検索のみを行っているもの」**、既知事項検索は**「タイトルの一部と思われる記述や出典雑誌名による検索を行っているもの」**とした。

個々の検索ログに対しては、検索の失敗と考えられる 0 ヒットやヒット過多(100 件以上)を出した場合の理由とその後の検索行動、一覧表示画面から個別書誌事項の詳細表示を行った後の行動パターンについて分析を行った。（付表参照）さらに、この分析を進める上で、この検索ログを手がかりとして判断できる利用者のタイプとしては、利用者の検索スキルのレベル（熟練度）、利用者の忍耐強さ（忍耐度）、があると考えられたことから、それぞれセッション毎に３段階で評価を与えた。

今回分析の対象としたのは、２回以上のアクセスのあった 184 セッション、2435 のログである。また、対象を一般利用者からのアクセスに限定するために、図書館内のサーバからのアクセスについては除外した。

１）Namazu のログデータ

例）ファイル名　　キーワード　検索ヒット数　利用者 IP アドレス　日時

ronbun.Aug19.slog:西林　　13　　210.180.96.5　　Tue Aug 19 14:16:10 2003

ronbun.Aug19.slog:促進　　63　　210.180.96.5　　Tue Aug 19 14:17:09 2003

２）Web サーバのログデータ

例）利用者 IP アドレス　日時　実行コマンド　HTTP 形式

219.0.228.12 - - [19/Aug/2003:01:09:31 +0900] "GET /cgi-bin/namazu.cgi?reference=off&idxname=watosyo&idxname=wazassi&idxname=gakkai&idxname=ronbun&idxname=iinkai&max=20&sort=field%3Asubject%3Aascending&query=%B1%FE%CE%CF+and+%C9%E5%BF%A9&idxname=watosyo2&idxname=watosyo3&idxname=wazassi2&idxname=wazassi3&idxname=gakkai2&idxname=gakkai3&idxname=ronbun2&idxname=ronbin3&idxname=iinkai2&idxname=iinkai3&idxname=iinkai4&idxname=iinkai5&idxname=iinkai6&idxname=iinkai7 HTTP/1.1" 200 31392

⇒検索コマンド（実行コマンド中には、検索対象としたデータベースファイル名、文字化けしている検索キーワード、検索結果の出力件数、出力順位などが入っている。）

219.0.228.12 - - [19/Aug/2003:01:10:10 +0900] "GET /jsce/syosi/gakkai3/ID97130.html HTTP/1.1" 200 2090

⇒詳細閲覧コマンド

３．調査並びに解析結果

（１）検索システム利用件数の推移

表３は，2002（平成 14）年５月からの検索システムのホームページアクセス件数を記録したものである。2003 年 10 月と 2004 年１月の利用が多いがこれは，全国大会が終わった月であること, 2004 年１２月から試験運用で支部の研究報告論文の検索可能となったことなどが影響しているものと思われる。

表３　アクセス件数の推移

期　間	件　数	累　計	件/日
2002年5月-11月	70,000	70,000	334.9
12月18日-2月20日	27,877	97,877	449.6
2003年2月21日-3月22日	9,300	107,177	300.0
3月23日-７月20日	74,949	182,126	640.6
7月21日-8月31日	22,244	204,370	556.1
9月1日-9月26日	13,020	217,390	520.8
9月27日-10月31日	22,319	239,709	656.4
11月1日-12月2日	19,329	259,038	623.5
2004年12月3日-1月12日	19,160	278,198	491.3
1月13日-2月12日	24,162	302,360	833.2

（２）ユーザの全体的な利用傾向

今回抽出の対象としたデータベース別の解析対象ログの概要を表４に示す。

表４　データベース別分析対象ログの概要

データベース種別	対象期間	検索数	異なる検索語		異なる IP アドレス	
			数	割合	数	割合
和図書	2002/8/15-2003/8/20	120234	46174	38%	6213	5%
洋図書	2002/8/15-2003/8/20	38453	17557	46%	2301	6%
和雑誌	2002/12/2-2003/8/20	92345	40875	44%	4573	5%
洋雑誌	2002/8/15-2003/8/20	37872	17292	46%	2236	6%
学会誌	2002/8/16-2003/8/20	133853	51066	38%	6022	4%
学会論文集	2003/4/11-2003/8/20	112330	40875	36%	4364	4%
各種委員会論文集	2002/8/16-2003/8/20	162761	59434	37%	5987	4%

対象期間が異なるのは、データベースの公開時期がずれているためである。検索の際には、複数のデータベースを対象として検索をすることが可能であるため、上記の検索数の列はデータベース相互の重複を含む。

次に各データベースに対し、利用者がどの程度検索演算子を利用して検索しているかを表５に示す。ここで挙げている、and が連続２以上、３以上は検索ミスとなり０ヒットとなるが、検索エラーとして利用者には明示されない。いずれのデータベースにおいても２％近く当該エラーが見られる。論理和(or)ならびにトランケーション(*)については

ほとんど利用されていないことがわかった。

表５　各データベース別の演算子の使用状況

データベース	and			" or "	*	全件
	" and "	うち連続2以上	うち連続3以上			
和図書	21057	180	167	495	539	46174
	45.6%	0.9%	0.8%	1.1%	1.2%	
洋図書	7976	66	63	209	181	17557
	45.4%	0.8%	0.8%	1.2%	1.0%	
和雑誌	17504	146	138	390	395	37441
	46.8%	0.8%	0.8%	1.0%	1.1%	
洋雑誌	7839	71	67	208	174	17292
	45.3%	0.9%	0.9%	1.2%	1.0%	
学会	24930	218	205	567	479	51066
	48.8%	0.9%	0.8%	1.1%	0.9%	
論文	19868	161	152	445	312	40875
	48.6%	0.8%	0.8%	1.1%	0.8%	
委員会	30833	283	267	653	523	59434
	51.9%	0.9%	0.9%	1.1%	0.9%	

次に、それぞれのデータベースに対して検索を行った結果、どれだけの検索ヒット数が得られているかを図１に示す。このグラフからは、相当数の検索結果が０ヒットとなっていることがわかる。実際の利用場面では、横断的な検索を行う場合には、ヒット件数が０件として表示されるとは限らない。

この結果から、０ヒットに関する原因を探ることが重要な課題である

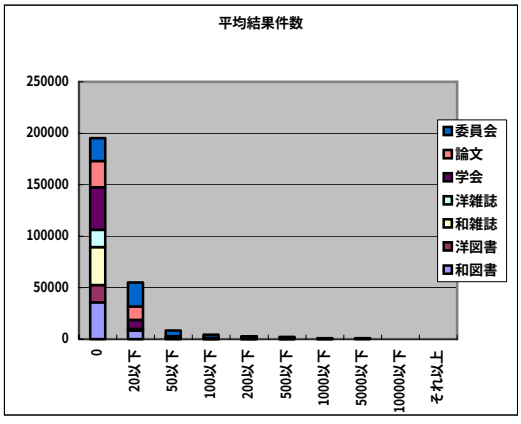

図１　各データベース別の検索結果件数

ことが明かになった。

（3）ユーザの検索セッション

各セッションの概要として、熟練度の3段階毎の検索ログ数（検索コマンド、詳細アクセス、一覧表示によるページ送りなど全てのアクセス記録を含む）、0ヒットの数、ヒット過多の数、詳細ページのアクセス数のヒストグラムを表6～7に示す。それぞれの傾向については､あまり大きな差異は見られない。

表6 検索ログ数（熟練度別）

	熟練者	非熟練	普通	全体
10	30	8	71	109
20	19	1	22	42
30	8	1	5	14
40	9	0	1	10
50	2	0	3	5
100	1	0	2	3
次の級	1	0	0	1
合計	70	10	104	184

表7 0ヒット数（熟練度別）

	熟練者	非熟練	普通	全体
0	26	3	71	100
1	13	2	11	26
2	8	1	4	13
3	7	0	7	14
4	7	2	0	9
5	3	1	2	6
6	1	1	4	6
7	1	0	0	1
8	2	0	1	3
9	0	0	0	0
10	1	0	0	1
次の級	1	0	4	5
	70	10	104	184

表8 ヒット過多（熟練度別）

	熟練者	非熟練	普通	全体
0	30	6	71	107
1	11	2	14	27
2	7	1	5	13
3	3	1	1	5
4	2	0	0	2
5	1	0	1	2
6	2	0	0	2
7	1	0	2	3
8	3	0	1	4
9	2	0	1	3
10	0	0	1	1
次の級	8	0	7	15
合計	70	10	104	184

表9 詳細閲覧数（熟練度別）

	熟練者	非熟練	普通	全体
0	14	5	25	44
1	11	2	18	31
2	10	1	17	28
3	9	0	10	19
4	1	0	6	7
5	0	1	8	9
6	2	0	7	9
7	1	0	0	1
8	5	0	2	7
9	2	0	4	6
10	3	0	1	4
次の級	12	1	6	19
合計	70	10	104	184

次に、熟練度と忍耐度の3段階で評価したセッション数を、表10に示す。

表10 忍耐度と熟練度による評価結果

熟練度	忍耐度高	忍耐度低	普通･不明	総計
熟練度高	45	1	24	70
熟練度低	3	4	3	10
普通･不明	30	3	71	104
総計	78	8	98	184

データとして忍耐度、熟練度の低いユーザとして得られたサンプルが少ないが、熟練度の高いユーザは忍耐度も高いという傾向は現わた。【熟練度高】としたユーザセッションには①OR 演算子や前方一致コマンドの使用 ②対象データベースの変更 ③一覧表示件数の変更、を行うことが特徴的な行動として抽出された。一方【忍耐度高】からは、①投入したキーワード数が5以上 ②セッションを10 分以上継続 ③20 件以上の一覧表示、が特徴的なものとなった。

4．まとめ

今後は、自動的な検索失敗理由の判別や認識されたユーザタイプを用いて、効果的な支援機能についても検討していきたい。

付表1 **0ヒットの理由と継続行動**

0ヒットの理由	継続行動
選択したデータベースの種類	キーワードを変更して入力
選択したキーワードが特定的すぎる	検索終了
選択したキーワードのミスタイプ･ミススペル	全く別のキーワードで入力
AND 演算子の多用(=キーワードが多すぎる)	データベースの種類を変更して再検索
演算子の誤用 (=and と or の取り違え)	ヘルプを参照
演算子の入力ミス (=and を連続したりキーワードの終りに入力する)	その他
その他	キーワードを変更して入力