OPAC ログ分析による検索過程の類型化
土木学会図書館書誌データベース検索システムを対象として

野末道子(鉄道総合技術研究所), 森岡倫子(国立音楽大学附属図書館)
嶋田真智恵(国立国会図書館), 寺尾洋子, 上田修一(慶應義塾大学)

Ⅰ. 既往研究と本研究の目的

Web 上に自館の目録や書誌データベースを公開する図書館が増加する中、利用者は図書館の開館時間を気にせず、自分の端末を利用して図書館の所蔵確認ができるようになった。時間的、空間的制約を受けずに書誌情報にアクセスできる機会が持てるようになることを大きなメリットとしてとらえ、土木学会図書館においても、2002年より自館の目録を公開し始めた。Web 上の利用者のアクセスは順調に増加傾向をたどっているが、その大多数が土木学会図書館に足を運ぶことなく、数回の検索トライでシステムの前から立ち去っている。そこで満足の行く検索結果を得ているのかということが、疑問点として挙がってきた。

我々は、これまで土木学会図書館のログのみを対象として、そこから浮かび上がる利用者の検索行動を分析してきた。[1]これまでの調査において、我々は利用者の検索行動に現われた熟練度と忍耐度という二つの軸を仮定し、人手により判定付与した熟練度、忍耐度のレベルで分割した利用者の検索行動の差異を抽出することを試みてきた。熟練度は、後天的な検索スキルの取得によってもたらされるものである。また忍耐度とは本来、個人の性格的な影響を受けたものであるが、本研究では、検索行動自体が網羅的な検索を行うという目的に基づくものについても忍耐力のある検索として含んでいる。本研究では、これまでの手法の問題点から分類手法を見直し、この熟練度と忍耐度を軸とした利用者の検索行動の類型化を行うことを試みる。

II. OPAC ログ分析に関する既往調査研究

OPAC を対象としたログ分析に関する研究調査は数多く行なわれている。Borgman[2]はログ分析の目的として以下の5項目を挙げている。

1) システムインタフェースの改良
2) 情報検索システムの検索過程の明確化
3) オンライン目録の操作方法の評価
4) コンピュータ支援設計の一連のタスク評価
5) コンピュータを基本としたメッセージシステムの評価

我々の調査は、個々のユーザの行動を分析、類型化することで、3の操作方法の評価に分類される研究と位置付けている。

ユーザインタフェースの利用者の図書館情報学分野においてトランザクションログ分析の歴史について Peters [3]らは表1のようにまとめている。1995 年代までは、トランザクションログを対象とした研究の数も多いが、全体として、最近の発表論文数としては減少傾向にある。Web OPAC の検索ログ分析には、アクセス元、時間帯、曜日、検索フィールドなどの個々のユーザではなく検索システム全体としての特徴傾向をとらえる研究が主流となりつつある。[4]

表1　トランザクションログ研究の年代別トピック

第1期:1960 年代中期-1970 年代末	ユーザの行動よりもシステムの評価に重点。
第 2 期:－1980 年代中期	OPAC システムの使われ方、利用者の検索行動。
第 3 期:その後	様々に分化。大部分は一般利用者による情報検索システムの実際の利用に焦点。

トランザクションログを分析する上で、Peters[3]らはトランザクションログ研究における一般利用者の利用行動の評価項目を整理している。この項目を表2に示す。

この表2において、最も関心が高い項目は検索の失敗要因の分析である。Tenopir は、全米の大学・公共・専門図書館のレファレンス担当職員に対するアンケート調査を行い、オンライン検索の失敗の原因として次の項目を挙げた。

- 入力エラー:単数形と複数形
- インタフェース:CD-ROM、オンラインサービス、WWW など図書館が提供するシステム

*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)
**国立音楽大学附属図書館
***国立国会図書館
****慶應義塾大学文学部図書館・情報学科

の検索方式がそれぞれ違うことが多い
- ブール演算：最もよく利用されている検索方式だが、利用者にはあまり理解されていな
- タームエラー：一般的すぎる単語を入力して、多数のヒットを生じてしまう
- 概念エラー：利用者は通常、質問をきちんと概念化せず、思いついた言葉で検索してしまう。

表2　検索評価項目とログ研究例

検索評価項目
コマンド、反応時間、セッションの長さ（所要時間／コマンド数）
コマンドの連鎖と検索の移行
検索モード：メニュー/コマンド形式の比較
同じシステムの場所や組織を超えた比較
複数システムの比較
失敗：エラー、ゼロヒット、結果過多、機会を逃す
ヘルプの利用
特定の検索
照合の拡大
アクセスフィールド
印刷・ダウンロード
ユーザの根気（対応できるヒット数）
終了

本研究では、Peters や Tenopir らがまとめている従来のログ分析手法を踏まえつつ、WebOPAC システムの検索ログのみを用いて、特徴的なユーザ群とその行動を分析することを試みてきた。

表3　土木図書館の検索システムの構成

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

表4　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

III.　土木学会図書館システムの概要

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。検索システムのハードウェア・ソフトウェア構成を表3に、データベースの構成を表4に示す。

次に検索システム利用件数の推移を知るために、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ログ件数を把握した。（表5）

IV.　ユーザの検索セッションの分析方法

ユーザの検索セッション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 種類を利用した。NAMAZU のログデータの例を図1に示す。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。

この NAMAZU と Web サーバのログをマージしたデータを作成し、2003・7・20~7・29 日までの図書館内からのアクセスを除く全ログに対する、セッションの認定を次の手順で行った。

1) 同一日付の中のログを IP アドレス順にソートする。
2) 検索コマンド間が 30 分以上あいている場合には、同じキーワードもしくは同様のキーワードが出現している場合に同一セッションとする。
3) 他の IP アドレスのログと混配 IP アドレスを混配すると流れが理解できるログはマージ。また、別のセッションでマージしたほうがよいとわかったアドレスはほかのログでも同様の処理を行う。
4) マージした結果を含め全ログを複数人の確認を通して一セッションとして確定。

このプロセスを経て、セッションとして確定した総数は 961 セッションとなった。また、この期間から2回以上の検索コマンドからなる 186 セッション、2431 のログについて、各コマンド単位の検索行動に対し、

1) 次にどのような検索行動をとったのか
2) 検索に結果ヒット過多、

表5 アクセスログ数の推移

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

0ヒットの検索結果である場合、その理由は何か

を人手により判定した。既往調査 [1]においては、この判定は一名の判断で行ったが、今回は複数人による確認を行うことで、判定内容の精度の向上を図った。

既往調査においては、各セッションに対し、人手による熟練度、忍耐度について3段階での判定を行った。さらに、利用者の検索が「主題検索」、「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」**、既知事項検索は**「人名、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。ここで熟練度、忍耐度の判定は主観的な評価であり、以下のような問題が挙げられた。

1) 判定結果が人により大きく異なる。
2) 判定結果が検索内容によって、同一判定者であっても判断が揺れる
3) 結果的に、判定者は個々に何らかの基準を持って判定を行っている。

この3の基準が、それぞれの判定レベルで分けられた利用者の特徴的な行動と重なって現われた。そこで、前回の熟練度、忍耐度により分けられたユーザの行動を判定者間で確認し、主観的判断によらずかつあらゆるセッションについて機械的に熟練度、忍耐度の判定ができる基準を検討した結果、熟練度については、「演算子の利用の有無」、忍耐度については、「10回以上もしくは10 分以上の検索セッションであるかどうか」を基準とした。また前回は三段階評価を行ったが、実際には「熟練度が低い」「忍耐度が低い」と判定できるデータは少なく、「そのほとんどが普通または不明」に分類されることから、今回は2段階での判定とした。前回の高－普通･不明－低のうち、普通･不明と低をまとめた判定結果と、今回の機械的な基準による判定結果との合致率を表6に示す。

表6 前回判定結果との合致率

	合致	非合致
忍耐度	76.09%	23.91%
熟練度	77.17%	22.83%

V. セッション分析の結果

1) セッションの全体傾向

961セッションによる、検索ログ数と熟練度の関係を表7に示す。検索ログ1回で検索を終了するセッションが全体の約10%を占め、10回未満で終了するログが63%である。また、ログ数6回以上から、検索演算子の使用の割合が逆転した。

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

図1 NAMAZU と Web サーバのログデータ

表7　検索ログ数と検索演算子使用（熟練度）

セッション中の検索ログ数	熟練度高 演算子を使用		熟練度普通・不明 演算子を使用せず		計	
1回	12	11.8%	90	88.2%	102	100%
2回	27	23.9%	86	76.1%	113	100%
3回	25	29.1%	61	70.9%	86	100%
4回	18	27.7%	47	72.3%	65	100%
5回	28	40.6%	41	59.4%	69	100%
6回以上	85	50.3%	84	49.7%	169	100%
10回以上	124	59.6%	84	40.4%	208	100%
20回以上	110	73.8%	39	26.2%	149	100%
合計	429	44.6%	532	55.4%	961	100%

次に、AND、OR 検索演算子と忍耐度の関係を表8に示す。約 75%の忍耐度無しの利用者が論理演算子のいずれも使用しないのに対し、約63%の忍耐度のある利用者が演算子を使っている。

表8　論理演算子の使用と忍耐度

	両方使用せず		ANDのみ使用		ORのみ使用		両方使用		計	
忍耐度無し	309	75.2%	101	24.6%			1	0.2%	411	100%
忍耐度有り	154	36.7%	249	59.3%	3	0.7%	14	3.3%	420	100%
	463	55.7%	350	42.1%	3	0.4%	15	1.8%	831	86%

表9　詳細分析対象の熟練度/忍耐度セッション数

	忍耐度無し	忍耐度有り	総計
熟練度高	36	63	99
熟練度普通・低	49	38	87
総計	85	101	186

2）　ユーザの検索行動詳細分析

前述のユーザの中から2回以上の検索コマンドからなる 186 セッションを無作為に抽出し、それぞれの検索過程について、0ヒット、ヒット過多、詳細閲覧後の行動を分析した。本研究では、表9の4カテゴリにおけるユーザの行動の特徴的な検索行動を明らかにすることを試みた。分析結果の一部を図2、図3、図4に示す。

これらの結果から以下の傾向が見られる。

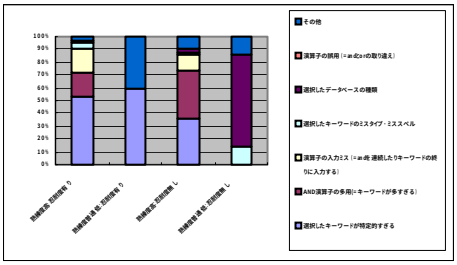

図2　4カテゴリ別ユーザの0ヒットの原因と理由
（キーワードを変更して入力した理由）

【熟練度高－忍耐度有り】・詳細閲覧するのは、目的の文献を見つけたということにとどまらず、その文献を元にキーワード変更を行う。0ヒットが現われた場合には、自分の入力したキーワードを修正し、簡単には検索方針を変えない利用者である。

【熟練度普通・低－忍耐度有り】・詳細の閲覧を丹念に繰り返し、一覧の出力も多くみるが、その後のキーワード変更はあまり行わない。ヒット過多になると、詳細の出力を数多く行うユーザ群である。

【熟練度高－忍耐度無し】・ヒット過多になった場合には、まず詳細や一覧継続表示は行わず、AND による絞込みを行うことが特徴できである。

【熟練度普通・低－忍耐度無し】・0ヒットの原因となるのは、選択したデータベースの種類であることが他のユーザ群に比べて多い。

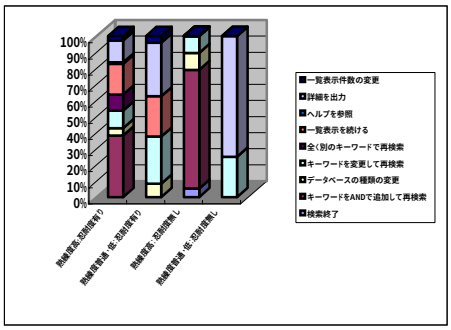

図3　ヒット過多後の行動
（キーワードが一般的すぎる場合）

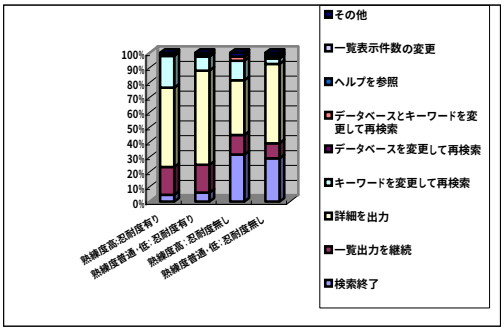

図4　詳細閲覧後の行動

5. まとめ

本研究では、ユーザの特徴的な行動を具体的な検索行動から類型化することを試みた。

【注･引用文献】

1) 森岡倫子；野末道子；嶋田真智恵；寺尾洋子；上田修一,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,日本図書館学会春季大会 2004 予稿集.
2) Borgman. C. L.; Hirsh, S.G.; Hiller, J. Rethinking Online Monitoring Methods for Information Retrieval Systems: From Search Product to Search Process. Journal of the American Society for Information Science. Vol. 47, No.7. p.568-583(1996)
3) Peters, Thomas A. The History and Development of Transaction Log Analysis. Library Hi Tech. 11(2): 41-66 (1993)
4) Cooper, Michael D. Usage Patterns of a Web-Based Library Catalog. Journal of the American Society for Information Science and Technology. 52(2):137-148(2001)
5) Wiberley, S.E. ; Daugherty, R.A. Users' Persistence in Scanning Lists of References. College and Research Libraries, vol. 49, no.2, p.149-56(1988)