Web 情報検索サービスの教育利用についての考察 — 大学生の利用履歴の調査から

津田塾大学情報数理科学科
来住 伸子
187-8857 東京都小平市津田町 2-1-1
Tel:042-342-5160 Fax:042-342-5161 E-mail:kishi@tsuda.ac.jp

概要
Web 上の各種サービスの中でも、 goo などの Web 情報検索サービスは、高校教科「情報」をはじめ、教育全般で、非常に重要なサービスになると考えられる。この報告では、大学生の Web 情報検索サービスの利用履歴の調査を元に、入力誤りや入力形式の誤解や、情報検索サービスの利用目的の現状について報告する。続いて、 Web 情報検索サービスを教育利用するための課題や改善すべき点について考察する。

1 はじめに

Web 上の各種サービスの中でも、 goo や infoseek などの Web 情報検索サービスは、高校教科「情報」をはじめ、教育全般で、非常に重要なサービスになると考えられる。そこで、この報告では、大学生の Web 情報検索サービスの利用履歴の調査の一部を紹介し、それを元に、現状の Web 情報検索サービスを教育利用する際の問題点や、改善すべき点を考察する。

この報告では、まず、調査方法について説明する。次に、入力形式の特徴として、入力誤りや入力形式の誤解がどの程度あるかを報告する。続いて、誤入力を除いた後の 検索文字列を読むことによって、ユーザの利用目的を推定し、利用目的の分布について報告する。最後に、これらの報告から、 Web 情報検索サービスのユーザインタフェースを初心者向きに改善すべきか、利用方法をどのように教えるべきか、学生の Web 情報検索サービスの利用を制限すべきか、等の点について考察する。

Educational Use Of Web Search Engines – What Are College Students Searching?
N.Kishi, Tsuda College

2 調査方法

Web 情報検索サービスの利用履歴データは、二大学の Web Proxy Server の管理者の協力を得て、各大学で実際に 運用している Proxy Server の約 6ヶ月間のアクセス記録から収集した。この際、調査対象となるユーザには、この報告のために調査をすることを事前に知らせていない。なお、二大学のうち一方は、 Proxy Server の運用開始時から、 Web ページ上で「このキャッシュ・サーバの利用統計はプライバシに関わる情報を除いて公開することがあります. 」という表示をしていた。もう一方の大学は何も表示していない。また、ユーザは学部生に限らず、大学院生や教員や職員を含み、ユーザ数は二大学合わせて、 一万人を越えると推定している。

二大学とも利用している Web Proxy Server は Squid[2] というソフトウェアで、そのアクセス記録には次のような特徴がある。

1. 機密保護のため、標準の設定では、通常の形式 の CGI スクリプトの引数文字列の情報をアクセス記録に残さない。たとえば、暗号化していない Web サーバーに送信した、クレジットカード番号のような情報を残さない。

2. ユーザのプライバシー保護のため、標準の設定では、どの計算機からの要求であるかという情報を残さない。たとえば、a.b.c.d という IP アドレスの計算機からの要求があっても、a.b.0.0 からの要求があったという記録を残す。

特徴 1 のため、情報検索サービスに対する 検索文字列は、通常の設定のアクセス記録には多くの場合残らない。ところが、情報検索サービス goo は、通常と異なる形式でも 検索文字列を送信するので、その 検索文字列がアクセス記録に残る。一方、特徴 2 のため、複数個の文字列があっても、同じユーザが同じ文字列を複数回入力したか、複数ユーザが同じ文字列を 1 回ずつ入力したかは区別できない。

これらの特徴を踏まえて、goo に対する検索文字列を、Squid の アクセス記録から抽出したところ、約 194,909 個の文字列、重複入力を除くと 約 60,136 種の文字列が得られた。1 種の文字列の平均入力回数は、3.25 回で、1 回しか入力されなかった文字列は 30,539 種、最も頻繁に入力された文字列は 1607 回入力された。図 1は、この文字列種と入力回数の関係を示している。ヨコ軸が入力回数 (軸目盛は 2 を底とする対数)、タテ軸が文字列種数 (軸目盛は 10 を底とする対数) を示している。この図から、入力回数が 256 回以上の文字列は 10 種未満であることがわかる。なお、固有名詞を除いた後の、入力回数上位 10 種は次の通りである。

入力回数	文字列
1607	(空文字列か印字できない文字)
711	女子アナ
446	mp3
378	レイプ
303	midi
288	ナコルル
274	草野満代
256	フラクタル圧縮
231	広末涼子
219	爆笑問題

また、図 2は、入力回数別の累計回数分布を示しており、ヨコ軸が入力回数 (軸目盛は 2 を底とする対数)、タテ軸が累計入力回数が全体に占める割合 (パーセント) を示している。この図から、入力回数が 1 回しかないものが、入力回数全体の 15 % を占め、入力回数上位のものは、累計入力回数全体の中ではそれほど高い割合を占めていないことが分かる。

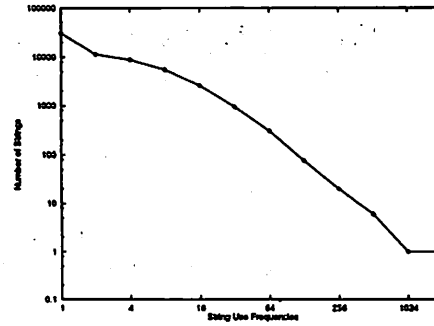

図 1: 入力回数別の文字列種の分布

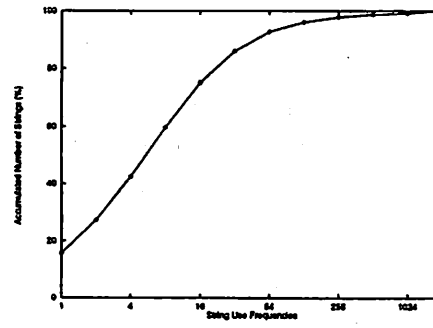

図 2: 入力回数別の入力回数累計分布

3 文字列形式の特徴

上記のようにして得られた 検索文字列には、意味のとれない 検索文字列がかなり含まれている。そこで、perl などのスクリプトを利用して、次のような加工を行なった。この加工の過程で、文字列形式から、ユーザ入力には、次のような特徴があることが観察できた。

3.1 誤入力

空白文字列をはじめ、全く意味のとれない文字列は約 1989 個、重複を除くと約 229 種あった。つまり、回数で見ると 1 % 程度、種類で

は、0.1 % 弱あった。これは、空文字列をはじめ、誤入力であっても、同じ文字列を繰り返し入力するユーザがいるために起きたと考える。また、このような誤入力の原因はつぎのようなものではないかと推定している。

- カットアンドペースト、置換などの編集操作の誤り
- 仮名漢字変換の誤変換

さらに、本来の文字列がほぼ推定できる誤入力は数多くみられ、誤入力の判定を厳密に行えば、誤入力は一割弱にのぼると思われる。これらの原因は、誤変換、表記の間違い、区切り文字の誤用、半角全角の混用であるが、固有名詞の表記が誤りかどうかの判定が難しいため、この意味での誤入力の回数を正確に数えることはできなかった。

3.2 複合検索

goo では、半角空白文字で文字列を区切って入力すると、4種の複合検索 — AND(すべての語を含む), OR(いずれかの語を含む), 近接(フレーズ), 人名 —が可能になる。この複合検索がどの程度使われているかを調べた。これは、検索文字列を 、どのような区切り文字が含まれるかによって、 次の4グループに分類することによって行った。

標準複合 サービス提供側の指示通りに、「スペース」(半角空白文字) で 区切った 検索文字列。

複合 A 半角空白文字でない文字のうち、カンマ、ピリオド、全角空白文字などの区切り文字を使用した 検索文字列。「スペース」と同じように複合検索が行われる。

複合 B スラッシュ、かっこなどを区切り文字とした 検索文字列。フレーズ検索と似た動作はするが、複合検索すべてに対応しない。

単独語 上記の3種類のどれにも該当しない 検索文字列。

表 1: 複合検索の使用回数

種類	種類数	(%)	入力回数	(%)
単独語	26089	43.5	102748	53.3
標準複合	27573	46.0	73823	38.3
複合 A	5832	9.7	15152	7.9
複合 B	413	0.7	1197	0.6
合計	59907		192920	

4グループに分けた結果を 表 3.2 に示す。

この表から、種類数でみると、複合検索は単独語より多く使われ、入力回数でみても 4割以上に使われていることがわかる。また、複合検索利用者の数分の一, 複合 A に該当する 検索文字列を利用した人たちは、複合検索の使い方を意識的には学習していないと考えられる。

3.3 自然言語入力

また、区切り文字ではなく、格助詞などを利用した自然言語で入力したものが1 %未満程度あった。主なパターンは次のようなものである。

- 本や映画のタイトル
  例　あの日に帰りたい
  　　いまを生きる
- 格助詞「の」や「と」を利用した複合検索風の表現
  例　岡山県のキャンプ場
  　　マルチメデイアと外国語教育
- 格助詞以外の修飾句を利用した複合検索風の表現
  例　よく当たる姓名判断
  　　核実験によるオゾン層の破壊
- 質問表現
  例　ワールドカップ日本代表について
  　　水質汚染とは？
- その他
  例　Ｗｎｎのファイルでありません
  　　これで終わりと思ったら大間違いだ

3.4 半角全角文字の混在

検索文字列中の半角文字と全角文字の使用割合を調べたところ、表 2のようになった。英単

語や略語の検索が3分の1程度を占めることが分かる。

表 2: 全角半角文字の使用状況

種類	種類数	(%)	入力回数	(%)
半角文字のみ	15013	25.1	43942	22.8
半角全角混在	8767	14.7	20557	10.7
全角文字のみ	36070	60.3	128294	66.5
合計	59850		192793	

また、全角英数字と半角英数字を合わせた、英数字について、半角と全角の分布を調べると表 3のような分布になった。この表から、英数字入力の 2-3 %、全体では、約 1 % が、全角と半角の混在した英数字を含むことが分かる。

表 3: 英数字の全角半角文字の使用状況

種類	種類数	(%)	入力回数	(%)
半角文字のみ	23805	89.5	64174	88.4
半角全角混在	621	2.3	2255	3.1
全角文字のみ	2179	8.2	6191	8.5
合計	26598		72620	

4 利用目的の特徴

検索文字列の集合を実際に読むことにより、検索文字列を利用した、ユーザの目的を推定することを試みた。全体で6万種の文字列のうち、入力回数15回以上の検索文字列1536種、全体の 2.5 % にあたる 検索文字列について調べることにした。検索文字列全体を調べなかった理由は、時間と労力に制約があったためである。また、入力回数の少ない検索文字列は、意味が簡単には分からない場合や、複数の目的に該当する場合が多く、分類が難しいと考える。

利用目的は、主観により、次の6グループに分類した。

- 研究教育関係
  大学の講義内容と直接関連がある語。例: mp3, フラクタル圧縮、 死刑廃止, 創価学会, 英語教育 + インターネット
- 計算機関係
  パソコン用語とゲーム用語。計算機科学の用語の場合、上記に分類したが、パソコン用語で趣味に近いものは、別分類にした。例: midi, mule, バッテリー, 画像,vesa
- 生活 -A
  日常生活の用語、地名、大学名、企業名などの語。例: アルバイト募集, トラベラーズチェック, 河合塾, フロムエー, 弁理士
- 生活 -B
  趣味、音楽、映画、テレビ、芸能関係の用語。例: 女子アナ, ナコルル, 草野満代, 広末涼子, 爆笑問題
- 生活 -C
  アダルト情報の検索が目的と思われる語。例: レイプ, lolita, アイコラ, 盗撮。
- その他
  意味が分からなかったり、上記の分類ができなかった語。例えば、人名は固有名詞と認識できても、芸能人、研究者、知人等の区別が難しい。

分類した結果を 表 4 に示す。

表 4: 利用目的の分類

種類	種類数	(%)	入力回数	(%)
研究	286	18.6	10197	20.2
計算機	207	13.5	5604	11.1
生活 A	183	11.9	5512	10.9
生活 B	349	22.7	13691	27.1
生活 C	81	5.3	4646	9.2
その他	430	28.0	10950	21.6
合計	1536		50600	

なお、検索文字列全体について生活 C グループ に該当する語を調べたところ、種類数 で 649 個 (1.1%) , 入力回数は 4801 回 (2.5 %) であった。

5 考察

5.1 利用方法の教育について

文字列入力方法の特徴は、Web 情報検索サービスの利用方法を教育するにあたって、次のような課題があることを示していると考える。

- 基本入力操作の事前学習
  Web 情報検索サービスを利用する前に、基本的な入力操作に習熟する必要がある。たとえば、全角文字と半角文字の区別、改行キーの使い分け (かな漢変換の確定とWeb サーバーへの送信の区別) などは、初心者は十分に学習していないことが多い。

- 複合検索の学習
  複合検索は 検索文字列の約半分を占め、非常によく使われているが、複合検索の正しい利用方法を知らないユーザもかなり多い。現状では、複合検索の操作方法が検索サービス毎に異なることが多く、学習がしにくい状況である。しかし、操作方法として、どの方法を教えるべきかという問題はあるが、複合検索でどんな検索ができるかを、少なくとも意味や機能の解説をすべきだと考える。

- 検索ユーザインタフェースの共通化
  上記の教育を容易にするには、検索サービス提供者間の協力により、検索ユーザインタフェースの共通化が行われることが望ましいと考える。たとえば、検索文字列に利用できる区切り文字や自然言語句が各サービスで共通になれば、ユーザの学習の負担はかなり減るはずである。また、全角半角文字の区別は、初心者には現実的には無理だと考えられるので、半角文字と全角文字を区別せずに扱うことも期待したい。

5.2 利用目的の制限について

利用目的の分類は、分類基準が主観であり、分類できない 検索文字列も多かったため、不十分な調査であるが、次のような点が観察できた。

- 広い意味での研究教育目的の文字列は 3 分の 1 程度であったと考える。

- アダルト情報の検索と考えられる利用は、入力回数上位では 数 % , 下位では 1 % 程度と考える。

- 残り、つまり半分以上は、生活や趣味に関する利用と推定される。

これらの点や、大学生は 18 才以上であることを考えるとフィルタリングは大学生には不要と考える。一方、生活 C グループの 検索文字列を検索サービスで実際に検索すると、アダルト情報が簡単に入手できることから、18 才未満のユーザに対して、フィルタリングを行なう必要性は高いと考える。

フィルタリングの方式 [1] は数種あるが、今回の利用目的の分布の調査から、各方式について、次のような考察ができる。

- Web 情報検索サービスで検索できるアダルト関係の情報の多くは、「利用は 18 才以上の方に限ります」と表示されていることが多く、ブラックリスト方式のフィルタリングは、ある程度有効であると考えられる。

  なお、有害情報全体の中には、フィルタリングの対象とならないような対策を行っていたり、頻繁に移動する Web ページがあり、そのような情報に対しては、ブラックリスト方式のフィルタリングは必ずしも有効でない。

- 一方、学生の Web 利用を研究教育利用に制限したい場合は、ホワイトリスト方式のフィルタリングが有効と考えられる。研究教育利用のためのホワイトリストは、上記のブラックリストより大きくなるが、運用が可能な程度の大きさであると考える。

  しかし、生活や趣味に関する利用を認めたホワイトリストは、非常に大きくなるので、運用は非常に難しいと考える。さらに、ホワイトリスト方式のフィルタリングによって、利用目的を研究教育に制限すると、学生の Web の利用回数が大幅に減る可能性が高い。

- ブラックリスト方式やホワイトリスト方式ではチェックできない、Web ページに対するフィルタリングの第 3 の方式として、コンテンツチェック方式、内容情報に基づいた自動判別が提案されている。将来、リスト方式の短所を補うことが期待できるが、実用化にはまだ多くの課題があると考えられる。たとえば、今回の調査対象語と

同じ程度の規模、3万語程度のデータベースでは、内容情報の自動判定は非常に難しい。また、固有名詞、流行語に対する対応などの研究がさらに必要である。

5.3 調査方法について

今回の調査方法は、偶然得られたデータを元に、短期間で行ったもので、少ない時間や労力で、比較的多くの情報が得られる有効な調査であると考える。

しかし不十分な点も多く、個別の問題点については、より本格的な調査が必要である。たとえば、Web 情報検索サービスの入力方式について詳細な検討を行うには、実験室環境での個々のユーザの観察を行う必要があると考える。一方、Web 利用目的の調査については、より多数のユーザを対象にした調査が必要である。

6 おわりに

この報告では、Web Proxy Server のアクセス記録を利用し、大学生の Web 情報検索サービス利用の実態について報告した。さらに、Web 情報検索サービスの利用方法の教育や、教育利用のための使用についての課題について考察した。

現状では、Web 自体が未成熟な技術であるため、Web 情報の利用教育は課題や問題点が多い。今後、高機能な検索サービスの開発, フィルタリング技術の改良、 XML の普及、などが実現すれば、教育しやすい環境になると思われるが、短期間での実現は難しいと思われる。

そのため、「現状の Web を利用するための、最低限の知識と技術とは何か」について、教育する側と技術やサービスを提供する側の合意が望まれる。今後、検索サービス以外の利用も含めた、様々な面からの Web ユーザの調査研究が必要である。

参考文献

[1] Nagasawa,「特集 フィルタリングソフトの現状と課題」Internet Watch, 1999年5月31日号, http://www.watch.impress.co.jp/internet/www/search/article/9905/3114.htm

[2] Duane Wessels., et al., "Squid Internet Object Cache" http://squid.nlanr.net/Squid/