情 報 学 基 礎 31－5
（1993．9．14）

各個人の使用するキーワードの違いを考慮した情報検索システム

鈴木 亮一
NTT基礎研究所

本稿では、FTPサーバからファイルを検索する時のユーザの行動記録を分析した。この分析結果から、実際多くのユーザが日常的に失敗を繰り返していることがわかる。それらの失敗の大半に対してはキーワードの綴間違いの指摘など表面的な処理で検索を有効に補助できることを示した。また、より深い意味的な問題に関しては、今後の研究課題を明らかにした。

An Information Retrieval System that incorporates meaning differences in each user's keywords.

Ryoichi SUZUKI
NTT Basic Reseach Laboratory

This paper shows a first hand analysis of user's log, when they attempt to retrieve files (or software packages) from a FTP server. This analysis reveals how users retrieve files and why they fail to do so. Significant amount of user's attempts fail daily, though one of the main cause is syntactic and is relatively easy to avoid. Yet another cause, semantic cause, is identified and we propose new research directions.

1 はじめに

さまざま情報があふれいてる現在では、そのあふれている情報をいかに選択するかが大きな課題となっている。そのために、従来、数多くの情報検索システムが研究あるいは利用されている[1)]。

多くのシステムでは、集めた情報の持つ領域固有の特徴あるいは検索目的の類型に適応して検索のアルゴリズムや検索に用いるインデクスを高度化するアプローチが採られている。また過去の検索事例を利用するようなアプローチもある。

本稿ではこれらと相補的な立場に立ち、検索システムに対峙しているユーザに焦点をあて、検索行動の特徴を明らかにすることを試みる。

実際の検索場面では、ユーザは試行錯誤を繰り返しながら必要な情報を得ている。これらのユーザの行動観察を通じて検索場面における障害要因を明らかにすることが重要であると考える。

ユーザ側のメンタルモデルの差異や認知的要因による行動の特徴を知ることは従来システムのアプローチにとっても有益である。

本研究では、より良い情報検索の方法を得るために、ユーザが情報検索時にどのように行動し、なぜ適切な情報が検索できないかをユーザの行動記録(ログ)を材料に考察する。さらに、その観察結果を基に、新しい検索補助手段を提案する。

2 ユーザの行動記録

本研究では、当研究所において利用されているftpサーバのログを解析した。

このftpサーバは、筆者らの研究所の全体を結んでいるネットワーク上のマシンで、いわゆるフリーソフトウェアや、さまざまなドキュメントなどを保存しておき、研究所内のユーザがUNIXのftpコマンドを使うことによって誰でも自由に取り出せる、あるいは登録できるものである。

このftpサーバとそのログを研究対象に選んだ理由は以下の通りである。

- 検索ファイル数が膨大である。
- 非常に多くのユーザが利用している。
- ユーザのほとんどがこのシステムを何度も利用している。
- ディレクトリ構成が複雑である。
- 検索コマンドが使える。

このサーバに保存されているファイル数は75000個弱である。また、実際に使用しているユーザの数は、600人程度である。

また、ユーザはほとんどが当社の研究者である。ログから300人以上のユーザが4カ月以内に10回以上アクセスしており、過半数のユーザがこのftpサーバに対して初心者ではないと仮定できる。

次に、このftpサーバの特徴を以下に挙げる。

- ill structuredなディレクトリ構成

  保存されているファイルはいわゆる“整理された”状態にない。そのため、ディレクトリ・ツリーをたどることによってファイルを探すことが著しく困難なシステムである。これは、ある程度の目安はあるものの、ソフトウェアがディレクトリ・ツリーのどこに保存されるかが一定していないためである。さらにその理由は、

  - ユーザが自由にディレクトリを作り、ファイルを保存することが可能(このため、同じファイルがいくつか登録されることがある)。

– ファイル、ディレクトリの名前付け方に、統一的なルールがない。

– USENET news も一部アーカイブされている。

普通の news spool のイメージで、一部のニュース・グループをアーカイブしている。これは、USENET news の local な記事番号がファイル名となっている。そのため、この中から必要なファイルを得るのが他のファイルに比べて困難である。

• ファイル検索コマンドを用意している。

ある程度容易にファイルのありかが分かるように、このサーバではファイル検索用に、2つのコマンドを用意している。UNIX のコマンドの grep と find である。grep は“path/ ファイル名”が、全ファイルについて記録されているファイルに対して正規表現によって検索を行なう。find はトップ・ディレクトリからファイルを find を使って検索する。ユーザは cd によってディレクトリを探検していくか、これらのコマンドを使って目的のファイルの場所を検索することもできる。grep、find を使った実際の検索例を付録 B に示す。

3 行動記録の分析

ログには、ユーザの login した時間と、そのユーザの cd、get、find、findp、grep のコマンドが記録されている。今回解析に使ったログは、1993 年 4 月 20 日から同年 8 月 12 日までのものである (ログの例を付録 A に示す)。

ftp サーバに個人のアカウントをもち、そのアカウントで ftp を利用した場合もログに記録されるが、アカウントを持っている場合は他の検索手段を用いていることが多いので、解析の対象とはしなかった。

表 1 ログの解析

<table>
<tr><td colspan="2">全セッション数
10830
(3000/ 月、100/ 日)</td></tr>
<tr><td rowspan="3">ファイルを持っていったセッション数 (成功)
7198 (66.4%)</td><td>検索コマンドを利用せず、ファイルを持っていったセッション数
3938</td></tr>
<tr><td>検索コマンドを利用して見つけたファイルを持っていったセッション数
2488</td></tr>
<tr><td>その他
772</td></tr>
<tr><td rowspan="3">ファイルを持っていかなかったセッション数 (失敗)
4542 (33.6%)</td><td>検索コマンドを利用したセッション数
2015</td></tr>
<tr><td>検索コマンドを利用しないセッション数
1891</td></tr>
<tr><td>ただ単に login しただけのセッション数
636</td></tr>
</table>

表 1 において、一つのセッションとは、ユーザが login してから logout するまでの一連の行動をさす。その中で、一つ以上のファイルをもっていったものを成功したセッション、全くファイルを持っていかなかったものを失敗したセッションとした (付録 A は成功した 1 セッションである)。

実際にはユーザはセッションを失敗しても、短時間のうちに再度 login し、前回 login した時に検索したファイルを取っていくこともある。また、成功しているセッションの中には grep 等で検索したもの以外のファイルを取って

いったものもある。しかしこれらは全体からすると小数であり、妥当性に影響はないものと考えられる。

表1によれば、全セッションの約$\frac{1}{3}$が検索に失敗していることが分かる。これは、ファイル名のような単純なものでさえ、検索に大きい障害があることを示している。

4 障害の解析

つぎに、ログから失敗したセッションを100例任意にとり出して調べ、失敗した理由を考察した。検索失敗の要因を分類した結果を示す(なお、かっこ内は各々の原因による失敗の割合(%)になっている)。また、いくつかの例においては、同ユーザの前後のセッションに渡って行動を調べ、より正確にユーザの行動を解析した。

4.1 サーバに保存されていないファイルの検索(30%)

4.2 さまざまな綴ミス(35%)

1. タイプ・ミス(10%)

   キーボード上の隣のキーをタイプするなど、単純なミス。

2. 複数の単語の連結(10%)

   複数の単語でファイル名が構成されている場合、その名前の連結方法は、一意に定まっていない。ユーザは連結方法のすべてを試みるわけではないので、検索できないことがおこる。

   - all-files.txt (hyphen でつなぐ)
   - FontPatchin'_2.0.1.cpt.hqx (under bar でつなぐ)
   - SwitchBack.hqx (何も入れない)
   - その他

3. 綴を不正確な記憶(自分の綴ミスに気がつかない)。(5%)

   日本人であるためか、日本語的な発音しか覚えていず、うまく検索できない例があった。たとえば、eudora というファイルを探すのに、eudra、eudla、eudola などで探している。

4. “to” と “2” の混乱(3%)

   フォーマット変換プログラムは XX to YY と呼ばれる。これがファイル名になると、rasttopbm のように単に単語を連結したもの、troff2ps のように to が 2 になるもの、dvidvi のように to が省略されたものなどがある。これらは、各ソフトウェアによってまちまちであるため、目当てファイルを見つけられないことがある。

5. acronym を見つけられない(3%)

   例えば、metamail が mm になっているなど、ファイル名に短縮形が使われている場合も定型がなく、検索時の障害となっている。

6. ファイル名内のバージョン番号(2%)

   多くのファイル名には、そのファイルのバージョン番号が入っている。しかし、そのバージョン番号の入れ方に定型がない(あるいは従わない)ため、極めてファイル名が探しにくくなっている。例えばユーザは以下のような行動を取る。

   - gcc-2.x.x.tar.z を探すため、gcc2 と gcc-2、gcc.2 などで検索している。
   - バージョン番号だけで探す。例えば、18.59 や 3.14 など。

   ということが行なわれている。

7. 綴が数種類になる可能性 (2%)

Micro Emacs というソフトウェアがある。これを日本語化した kemacs というものもある。このソフトウェアは、さまざまなファイル名になって、幾つか置いてある。例えば、uemacs、kemacs、memacs、microemacs などである。そのため、どれが最新版かなどが非常に分かり難くなっている。

4.3 中途半端な知識による思い込み (除く綴ミス) (2%)

ユーザが自分の思い込みで、正しくない知識に気がつかないことがある。

calen.sh というファイルに対して、calendar で検索したが検索できず、という例があった。この場合の特徴は、同じ正しくないキーワードで何度も検索することで、同じユーザの前後のセッションの様子を調べることで解析できる。

4.4 多くのエントリにマッチングしそうな単語の使用 (5%)

例えば、

- color、image など、多くのものにマッチしそうな単語を使う。
- appli(cation)、game など、大分類で検索しようとする。

などが挙げられる。これでは、全ファイル名のリストと事実上変わりがなくなってしまう。

4.5 抽象的なキーワードによる検索 (5%)

ファイル名のみが情報であるにも関わらず、必要とするファイルを包括する概念を表すキーワードを利用している。以下に、例を挙げる。

- "time" で探しているのが ntp(network time protocol というソフトウェア) であった。
- "encrypt" や "security" でファイルのセキュリティ関係のソフトウェアを探している
- "idealiner" や "outl" などで、アウトライン・プロセッサを探している (実際には、その後の様子から Mac のアプリケーションを探しているらしいことが分かった)。

4.6 システムの詳細な知識の欠如 (2%)

システムに慣れるに従って知識が増える。その知識を使えば、以前、よく検索できなかったものが検索できるように、不必要なものを選ばないように検索ができるようになる。例えば、いま、最新版の jgawk(日本語化された gnu awk) を手に入れるには、保存されている USENET news の記事の検索法という知識が要求される。jgawk*.tar.Z というファイルがあるがこれは古いファイルである。そのため、目的の新しい jgawk を入手できなかったユーザがいた。

4.7 適切なキーワードを見つけられない

例えば、Macintosh のソフトウェアの 9 割は binhqx というフォーマットで蓄積され、hqx というサフィックスをもっており、これはよく知られていることである。ところが、このような性質を利用している人が少ない。Macintosh 用の telnet を探す場合、
*"grep .*telnet.*hqx\$"* とするのは有効である。最後の hqx は事実上「自分は Macintosh のファイルを探している」と言っているのに等しいからである。事実上セマンティックに探すファイルを限定できるにも関わらず、この種のキーワードを使っている人は皆無であった。

4.8 各個人の持つキーワードの意味の違い

各ユーザによって、使っている意味が違っているため、検索が困難であると思われる場合があった。

“alpha” というキーワードを使って検索している例が数例あった。それら、1例を除いて、Macintosh 上のテキスト・エディタ、alpha を検索しているものであった。残りの1例は、DEC のワークステーション “Alpha” のソフトウェアについて検索していた(その例の人を O さんと呼ぶ)。O さんは目的のファイルがないので一旦セッションを終了し、Alpha 関係のソフトウェアをサーバに置いた。そのとき、“alpha” というディレクトリを作成したため、alpha というキーワードにマッチするエントリが増えてしまい、その後の検索が複雑になったのが観察できた。

これは、ユーザがキーワードに付与する意味が、個人毎に違うための問題である。また、この例は情報を追加するユーザへのアドバイスや補助手段についても工夫が必要であることを示唆している。

また、ファイル名に含まれるであろうキーワードよりむしろ、概念に近いキーワードで検索を試みている (4.5 参照) ユーザがいることも確認できた。このようなユーザは、目的のファイルが見つかることが稀である。

さらに、その場合に使われる単語は抽象度が高いため、ユーザの専門によって意味が異なる事もある。

5 解析に基づくユーザ・モデル

また、ログの解析から、ユーザの振舞いは、個人によってある傾向があり、以下のようなタイプに分けられる。

- 思いつくさまざまなキーワードを次々と試みるタイプ

  この人達に対しては、部分文字列などを提案するのが良いアドバイスであろう。

- 同じキーワードではあるが、部分文字列などを工夫するタイプ。

  この人達には、部分文字列のうまい使い方、短縮形についてのアドバイスが適当であろう。

- なるべく一般的なキーワード (言葉) に頼ろうとするタイプ

  この人達には、具体的なキーワードを考えるように促す。あるいは、蓄えられているデータが自己組織化して、一般的なキーワードにもマッチするような仕組みも必要である [2]。

- 同じキーワードを繰り返し試すタイプ。

  他のキーワードを加えるように促す。また、ユーザインターフェースについてもアドバイスするべきである。

それぞれ検索に際し明らかに違う傾向を持つため、各個人毎のキーワードの意味の違いに加え、ユーザ・モデルの情報も個人情報に加味されるべきである。

6 情報検索システムの提案

4 節に述べた分類に従って対策を考察する。

4.1 の ftp サーバに無いものを探している場合は、一般に対策が難しい。本当にユーザが意図しているものを正確に把握しなければならないからである。

4.2 から 4.3 のような、スペル・ミス、短縮形など、syntactic な障害については、ある程度機械的にユーザに対してサポートが出来る。

例えば、grepへの引数が"gcc-2-2"のような場合、'-'の部分を'_'.'などにもマッチングがとれるようにすることは、そう難しいことではない。特に、このようなセパレータに使われている文字は限られているからである。また、単語の連結についても、幾つかの代表的なパターンを蓄えておき、入力されたキーワードに前処理を施してからgrepするなどすれば、比較的容易に補助できる。また、特にこれらについては、専門用語であることが多いため、まず、辞書を完備し(ftpサーバ上のデータは有限なので、これは可能)、キーワードと辞書のエントリの距離の近いものについては、ユーザに注意を促すといったことが、綴の間違いに対して有効である。このような補助手段を用いれば、実際に検索に失敗している場合の37%に対して補助できる。

残りの4.4、4.5、4.7で述べたような一般的な(汎用的な)キーワードで探そうとしたり、意味で検索することは、前述の機械的な手法ではうまくいかない。

各個人が持っている言葉の意味を考え、一般的なシソーラスとは別に、個人毎の意味を表す疑似シソーラス[3]を、システムを良く利用する人に限っても用意すべきであろう。この場合、情報提供者と情報検索者の疑似シソーラスが異なるためどちらかのシソーラスだけでは不十分で、どのようにこれら二つのシソーラスを情報検索時に適用していくかは今後の課題である。

また、ユーザ側だけでなくファイル名データベース自身も文献4)のような自己組織化を行なうことなども、今後の課題である。

4.6のような、システム対する知識の欠落に対しては、システムの利用頻度、ログから、ユーザのシステムに対する習熟度、5節で述べたようなユーザの傾向から、ユーザ・モデルを想定し、同じく5節で述べたような対策と合わせて、ユーザに対し細かい補助を考えていかなければならない。

7 まとめ

本研究では、実際の検索場面におけるユーザの行動の記録を解析し、検索時にどのようにつまずき、検索に失敗しているかを調べ、ユーザ・モデルについて考察した。

また、各ユーザがキーワードに対して持っている意味を考えた場合に、システムがどのような対策をとるべきかについて考察し、syntactic な障害と semantic な障害に分け、syntactic な障害は比較的簡単に補助を出来ることを示した。また、semantic な障害を取り除くに当たって、各個人のキーワードに対する意味の違いを考慮するとともに、各ユーザの行動のタイプにも注意を払うべきであることを述べた。

今後は、今回の分析で得られた知見をもとに実際にシステムを試作し、分析結果を検証、修正していくとともに、実用的な情報検索システムを構築する。特に semantic な障害を除去する方法を中心に実験、考察を進めていく。

参考文献

1) 菊池芳秀他: 全文検索の技術動向とシステム事例. 情報処理学会情報学基礎処理研究会資料,25-1,1992.

2) 岡田直之: 語の概念の表現と蓄積. 東京、電子情報通信学会,1991.

3)V. V. Ranghavan:A Machine Learning Approach to Automatic Pseudo-Thesaurus Construction . 情報処理学会情報学基礎処理研究会資料, 25-3, 1992.

4) Xia Lin et al A Self-Organazing Semantic Map for Information Retrieval. Proceeding of 14th SIGIR'91, pp262-269, 1991.

付録 A ログの一部

```
ftp@nttiba (XXXXX@nttiba.ntt.jp) Thu Jul  1 10:04:48 1993
        grep fax
        find fax
        findp fax
        get //pds/unix/newgnu/fax-3.2.1.tar.Z
        cd pds/unix/newgnu
        find fax
        findp fax
        get /pds/unix/newgnu/fax-3.2.1.tar.Z
        cd ../../ucs/ucs-unix
        get /pds/ucs/ucs-unix/fax2.0.tar.Z
        findp tex
        cd ftp
        cd TeX
        cd Mac
        cd jtex
```

付録 B grep の出力の例

```
ftp> quote grep truetype
214-grep truetype
 pds/ucs/apple/truetype.readme
 pds/ucs/apple/truetypedisk1.sit.hqx
 pds/ucs/apple/truetypedisk2.sit.hqx
 tape/pub/TrueType
 tape/pub/TrueType/rakfonts.cpt.hqx
(略)
 tape/pub/TrueType/windsordemi.cpt.hqx
ftp> quote find truetype
230-find truetype
 tape/pub/TrueType
ftp> quote findp truetype
230-find truetype
 pds/ucs/apple/truetype.readme
 pds/ucs/apple/truetypedisk1.sit.hqx
 pds/ucs/apple/truetypedisk2.sit.hqx
 tape/pub/TrueType
```