ヒューマンインタフェース 41－2
（1992. 3. 2）

サーチャのノウハウに見る検索インタフェース

下山 栄子　富士 秀　松井 くにお
（株）富士通研究所

概要

パソコン通信に代表されるネットワークシステムの普及で、大量の情報の交換が可能になってきている。これらの大量なデータにアクセスする情報検索システムは、従来からのキーワード検索が主流であり、現在においても実用化されているシステムはすべてキーワード検索の域を脱していない。検索インタフェースもユーザによる一方的なキーワード指定が主なもので、一般的なユーザにとっては満足する検索結果が得られることは少ない。

本報告では、サーチャの検索におけるノウハウを分析し、検索の再現率・適合率の向上のためには、キーワードの展開やユーザに対する適切なキーワードの提示等のユーザインタフェース機能が重要であることを示した。

Analysis of the Expert Knowledge of Database Searchers

Eiko SHIMOYAMA　Masaru FUJI　Kunio MATSUI
FUJITSU LABORATORIES LTD.

abstract

Recent development in computer network systems has resulted in the accumulation of huge quantities of electronic data. Currently, the most practical way of retrieving information from such data is by means of keywords, and most available systems are based on keyword techniques. However, these keyword-based systems require a certain amount of expert knowledge, and this makes it difficult for non-expert users to obtain the information they require.

This paper describes experiments in which we have tried to extract expert knowledge from the logs of database searchers. We have concluded from these experiments, that improved human interfaces may enable non-expert users to act like experts in information retrieval.

1　はじめに

パソコン通信に代表されるネットワークシステムの普及で、大量の情報の交換が可能になってきている。これらの情報は、送出側である程度の整理・分類を行なって提供してはいるが、受け手側ではそういった整理や分類が十分でなく、情報の失方位問題にぶつかっている。こうした受動的な情報にはある種のフィルタリングが必要であり、そのような研究は近年盛んに行なわれてきている [1]。

一方、能動的に大量なデータにアクセスする情報検索システムは、従来からのキーワード検索が主流であり、現在においても実用化されているシステムはすべてキーワード検索の域を脱していない。これは、情報検索システムの持つ宿命とも言うべき検索要求と検索対象のレベルの一致に原因がある。検索要求に対してユーザとの対話を行ないながらいくら意味解析したとしても、それらとマッチングする膨大な検索対象が意味レベルで表現されていなければならないのである。データが膨大であるから、コンピュータによって自動的かつ正確な解析が必要となるが、現在の意味解析技術では限界がある [2]。

そうは言っても電子化された文書は増加の一途をたどり、個人やオフィスにおける情報検索の必要性は高まり、使いやすいユーザインタフェイスを求める声は高い [3]。また、大規模文書検索においても検索精度を高めることや、システムの情報提示をもとに検索要求を明確化したい等のニーズは強まっている。しかしながらこれらのニーズをキーワードレベルでさえも十分に反映できない原因は、現在までに蓄積したデータベースが過去のシステムによって構築されたものであり、データの一様性を保つためにはこういった構築手法を踏襲していかなければならないことにある。

だが、電子化辞書の大規模化や、形態素解析の高速化によりキーワードの自動抽出が可能となり、過去の使いにくいシステムを一様性のためだけに使い続ける必要はなくなってきた。また、ネットワーク環境の変化により情報の収集や提供が容易になり、サーバ・クライアント方式等の分散処理環境の充実で、データの共有化をはかりながら手軽に検索できる環境は整いつつある。

こういった環境のもとでの情報検索は、今までの一方通行のデータの流れから、キーワードレベルにおける情報提示に基づく対話処理が可能となる。これは、キーワード間の関連性をたぐることにより検索要求を明確化して検索対象を探索することができる（図1）。このため、検索の素人である一般的なユーザでも計算機の支援を受けながら検索のプロであるサーチャに近いレベルの検索を行なうことができることができる。

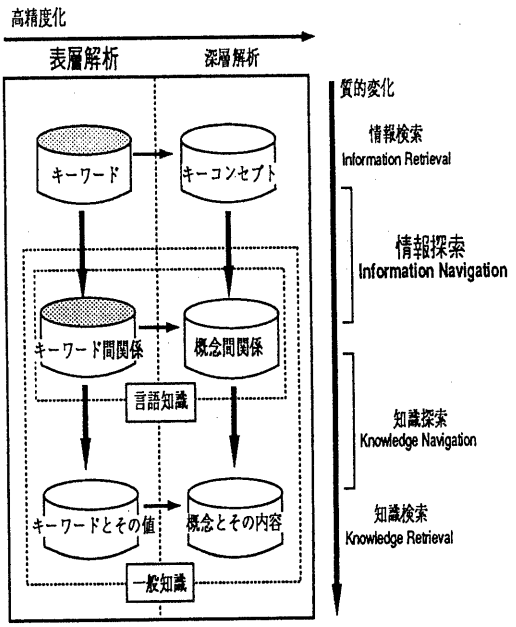


図1 情報・知識の収集／提供

このような背景を元に本論文では、検索の精度として基準となる再現率・適合率を向上させるためのヒューマンインタフェースの観点から、サーチャの検索におけるノウハウを分析した。

2　実験環境

実験は、一般ユーザがあるデータベースを検索した際の検索履歴と、この同じデータベースをプロのサーチャが検索した際の検索履歴を比較することによっておこなった。

2.1　検索対象データベース

検索対象となるデータベースは、社内のシステムエンジニア（ＳＥ）が利用するためのＳＥ事例を集めたものである。社内ＳＥは業務上必要な事例などをこのデータベースから得ることができる。

事例は社内で用意されホストの大型計算機上で集中管理される。利用する際はこのホストに接続されたパソコンやワークステーションなどの端末機から検索をおこなう。検索の結果、元データが必要な場合は社内メールなどで取り寄せることができる。

2.2 利用者

ユーザ 「ユーザ」とは自ら端末に向かってデータベースを操作するＳＥのことである。

サーチャ 「サーチャ」とは、専門でこのデータベースを検索する社内の人である。サーチャは、各地にいるＳＥから電話やファックスによって検索の要求を受け、この検索要求から適当な検索式を考え出してデータベースを操作する。

2.3 件数

およそ１ヶ月の間に蓄積されたログデータを評価した。検索件数は次の通りである。

サーチャ	4,855 件 (8,412 検索式)
ユーザ	21,974 件 (32,387 検索式)

2.4 書式

各検索履歴レコードの書式は以下のようになっている。

コード	利用者名	日付	時刻	検索件数	コマンド式

右端の「コマンド式」の文法は次のようになっている。

コマンド	項目名	関係演算子	文字式	論理演算子	...

コマンド SEA, OR, AND, NOT から一つ選択する。最初の検索は必ず SEA で始まり、AND, OR, NOT は一つ手前の検索式にかかる

項目名 KW （キーワード）、RC （登録コード） WD （作成日）などがある

関係演算子 EQ （項目の値が文字式と等しい）、GT （項目の値が文字式より大きい）

文字式 検索対象の項目値を指定する

論理演算子 AND, OR, NOT がある

```
1087   SEA   KW 流通 @
  27   AND   KW SIS
  15   AND   KW 事例

 210   SEA   SEC 2930

  12   SEA   KW EQ 役場

   1   SEA   RC QJ8002

1075   SEA   KW EQ 流通
  15   AND   KW EQ SIS AND 事例

 186   SEA   KW EQ DSLINK@
  13   AND   KW EQ HANDBOOK OR
             ハンドブック
  13   NOT   ST 0

   2   SEA   RC QJ8002 OR QJ8003

  12   SEA   KW EQ FTOPS
```

図２ 検索履歴の例（検索件数とコマンド式）

3 検索履歴の分析

3.1 検索コマンドの推移

図３はサーチャとユーザの検索コマンドの推移（上位 10 パタン）である。SEA、AND、OR、NOT はそれぞれ検索コマンドを表す。ただし、中カッコ {} 内の AND、OR、NOT は検索式の中の論理演算子である。検索コマンドの数は、言い換えると検索式数である。

コマンドの推移のパタンには次の特徴が見られる。

- サーチャはユーザと比べ、検索式の中で頻繁に論理演算子 OR を用いる。
- サーチャはユーザと比べ、少ない検索式で検索を終了するパタンが多い。

1928	SEA
479	SEA { OR }
207	SEA { AND }
188	SEA { OR OR
170	SEA AND
90	SEA { OR OR OR }
84	SEA AND { OR }
76	SEA { OR } AND
74	SEA { OR } AND { OR }
51	SEA AND AND

サーチャ

12043	SEA
3132	SEA AND
3024	SEA { AND }
1004	SEA AND AND
391	SEA { AND AND }
356	SEA AND AND AND
217	SEA { AND } AND
170	SEA { OR }
134	SEA AND { AND }
126	SEA AND AND AND AND

ユーザ

図3 検索コマンドの推移

3.2 検索項目の利用状況

ユーザが検索項目としてほとんどキーワード(KW)しか使わないのに対し、サーチャは公開レベル(ST)や登録コード(RC)などの項目も併用している(図4)。これはサーチャが一般ユーザに公開されていない情報を利用したり、情報の検索ではなく摘出を行なうなど、ユーザとはシステムの利用の仕方が異なる場合があることを示している。

5599	KW
1691	RC
615	ST
199	SF
130	DN

サーチャ

29632	KW
1089	RC
114	DN
56	OS
51	TI

ユーザ

図4 検索項目の利用回数

3.3 検索式数

3.3.1 1つの検索式で終了する検索

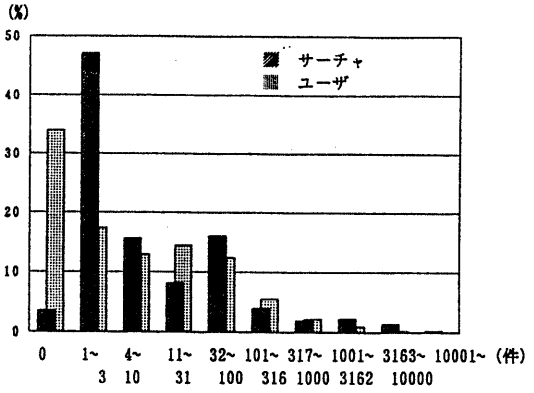
図3にも見られるようにサーチャ、ユーザとも1つの検索式で検索を終了する検索が目立って多く、その割合は次のとおり6、7割を占める。

サーチャ	62.3 % (4,855 件中 3,023 件)
ユーザ	72.2 % (21,974 件中 15,877 件)

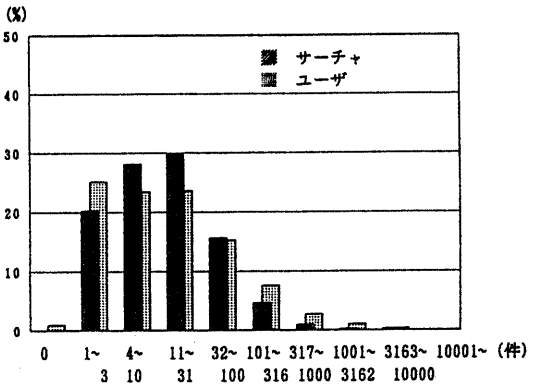
ここで検索式数と最終検索結果件数との関係を見ると(図5)、1つの検索式で検索を終了している検索には次の特徴がある。

- サーチャは、1～3件の検索結果を得て検索を終了するケースが多い。
- ユーザは、検索結果を全く得られずに検索を終了するケースが多い。

1つの検索式で終了する検索は、サーチャの場合は主として情報の摘出を行なったものと見るのが妥当なようである。このことは、1つの検索式で終了する検索の検索項目として、登録コード(RC)による検索が42 %を占めていることからも裏付けられる。一方、ユーザの場合はキーワードを用いた検索が75 %を占め、キーワード検索に失敗して検索結果件数が0件になってしまい、次の検索式につながらなかったケースと言える。



(a) 1つの検索式で終了する検索



(b) 複数の検索式からなる検索

図5 検索式数と検索結果件数の関係

3.3.2 複数の検索式からなる検索

図3からわかるようにサーチャは検索式のなかでOR展開をおこないながらANDで検索結果件数をしぼってゆく。一方、ユーザはOR展開をあまりせず、検索式の中でAND展開をおこなう傾向がある。そのため途中での絞り込みの失敗が多く、最終検索結果件数はサーチャと大差なくても、検索式数が増えてしまう。

222	SEA	KW EQ SX/A
222	AND	KW EQ JEFCONV
222	AND	KW EQ JEFCNV
12	AND	KW EQ FCAT
910	SEA	KW EQ FMR
598	AND	KW EQ MS-DOS
598	AND	KW EQ ハードウェア編
598	AND	KW EQ 3版

図6 複数の検索式からなる検索例

3.4 キーワード展開

3.4.1 ORを含む検索式

ORを含む検索式をユーザと比較すると、サーチャが論理演算子ORにワイルドカード(@)を組み合わせて広く展開しているのに対し、ユーザは（本来ワイルドカード展開で拾えるキーワード群の一部を）思いつく組合せでランダムにOR展開している。これでは検索洩れのでる可能性が高いし、入力しなければならない文字数も増えて無駄である。

OR展開の目的は、キーワードの同義語、反対語、対訳語、英語に対するカタカナ語の吸収などである。

42	AND	KW HANDBOOK OR ハンドブック
16	AND	KW HANDBOOK OR MANUAL
16	AND	KW 移行 OR CONVER@
15	AND	KW HANDBOOK OR MANUAL OR 手引
13	AND	KW 互換 OR 非互換
11	SEA	KW UNIX OR UNIX@
11	SEA	KW AIM@ OR AIM
11	AND	KW K シリーズ OR CSP@ OR CSP
9	SEA	KW RDB@2 OR RDB2@
9	SEA	KW K シリーズ OR CSP@ OR CSP

サーチャ

7	SEA	KW HHT OR ハンディターミナル OR ハンドヘルドターミナル
6	SEA	KW ISDN OR KW INS
4	SEA	KW SIMPLE OR KW SIMPLIA
4	SEA	KW FM OR KW FMR
4	SEA	KW FAX OR OCR
3	SEA	KW DMF OR KW DFM
2	SEA	KW XSP OR KW VTAM OR KW NCP OR KW RDB2
2	SEA	KW SFAMILY OR SUN
2	SEA	KW SCL@ OR F9082@

ユーザ

図7 OR を含む検索式

3.4.2 AND を含む検索式

サーチャ、ユーザともに AND 展開は OR 展開よりも少ない。ユーザは初期の段階から AND を使い、その結果検索を失敗する傾向がある。キーワードの例では、ソフトウェア名をレベルで限定したり、複合語の構成要素同士で AND をとるケースがある。

14	AND	KW SE AND HANDBOOK
13	SEA	KW AIM AND V20
9	SEA	KW SUPER AND CAPSEL
8	SEA	KW 販売 AND 管理
7	SEA	KW AIM AND V12
5	AND	KW 販売 AND 管理
4	SEA	KW FMG AND HANDBOOK
4	SEA	KW CAPSEL5 AND 会計
4	SEA	KW A シリーズ AND SA/SE
4	AND	KW 設計 AND 手引

サーチャ

22	SEA	KW AIM AND KW V20
14	SEA	KW AIM AND KW V12
13	SEA	KW AIM AND V20
11	SEA	KW VTAM-G AND KW V30
11	SEA	KW SX/A AND KW ハンドブック
11	SEA	KW AIM AND V12
10	SEA	KW RDB2 AND KW FSP
7	SEA	KW VTAM-G AND V30
6	SEA	KW YPS/APG AND MSP
6	SEA	KW SUPER AND CAPSEL

ユーザ

図8 AND を含む検索式

3.4.3 ワイルドカードの利用

サーチャはユーザよりも積極的にワイルドカード展開をおこなう。前方一致のほうが後方一致よりもはるかに件数が多い。ワイルドカード利用の目的は、日本語のキーワードでは複合名詞の吸収が多く、英文字では製品名などの吸収を目的としたものが多い。

131	MSP@
99	FMR@
98	SA/SE@
83	AIM@
66	FSP@
60	CSP@
57	K@
52	FMG@
51	R@
47	説明@

(a) 前方一致

33	@MSP
19	@FSP
7	@ハンドブック
6	@HANDBOOK
4	@XSP
2	@STANDBY
2	@SDEM
2	@FORM
2	@薬品
2	@スタンバイ

(b) 後方一致

図9 ワイルドカードの利用例

4 システムに求められること

これまでの分析結果からユーザの検索方法には次の問題点がある。

ユーザはキーワードとして何が適当か知らない 1つ目の検索式で検索結果件数が0件になってしまうケースに象徴されるとおり、キーワードの選び方がまずい。

ユーザは検索式を作るのが下手 サーチャの検索が成功するのは、検索式の中で適当な展開を行なっているからである。サーチャの検索式には、始めに大きく広げて検索し（OR展開、ワイルドカード使用）、続いてきっちり限定する、というパタンがあるが、ユーザはANDばかりを多用し、ワイルドカードやORをうまく使っていない。

これらを踏まえて、一般ユーザにとって使いやすいシステムを作るには、検索インタフェースとして何が求められるだろうか。

4.1 提示

ユーザはデータベースの中身や検索に有効なキーワードもわからないまま検索を始めるが、それを知らなくてもうまく検索できる枠組が欲しい。

例えば、1つ目の検索式で検索結果件数が0件になってしまっては検索にならないので、最初のキーワードでつまづかないようにしたい。一度も失敗しないようにするのはほとんど不可能であるが、失敗を繰り返さないようなアドバイスをシステムがしてやることは可能なはずである。例えば、検索結果件数0件を表示すると同時に、検索に失敗したそのキーワードの構成語や同義語など、再検索時のキーワードの候補をユーザに提示することが考えられる。

また、次に入力すべきキーワードがわからないユーザを支援する手段としては、入力キーワードに対する関連キーワードの提示が有効と考えられる。関連キーワードデータはデータベース中の文書から関連語を抽出／正規化して作成できる[4]。

ユーザを有効な検索式の作成に導くには、以上のようなキーワードの提示のほか、コマンドシーケンスの提示も重要と考えられる。これらの提示用のデータはサーチャの検索履歴からもある程度抽出することができそうである。またどのようなデータをどのようなタイミングで提示するかも、サーチャのコマンドシーケンスおよび検索結果件数のシーケンスから判断できそうである。

4.2 展開

展開を行なわれければ有効な検索は望めない。現状では、サーチャは同じような展開を繰り返し行なわなければならず、ユーザは展開の必要性すら認識していないという問題点がある。展開をなるべく自動的におこなう

ことによって、このような問題を軽減することができる。キーワードの自動展開用のデータ、例えば同義語、英語／訳語、英語／カタカナ語、複合語／構成語などの変換テーブルの作成には、サーチャの検索履歴も利用できる。

4.3 デフォルトの導入

サーチャは現在のシステムの枠内でさまざまな工夫をしている。例えば、サーチャは検索式の書式のルールを利用して、論理演算子 OR の後の項目名 KW を省略しているが、ユーザでは繰り返しタイプしていることがほとんどである。サーチャのこのような工夫はシステムに採り入れ、一般ユーザにも共有させたい。この例なら、利用する項目名は KW のことがほとんどなのだから (図3,4)、書式のルールを変更し項目名として KW をデフォルトとすることが考えられる。

5 おわりに

検索履歴を分析することによってサーチャとユーザの検索のモデルをある程度作ることができた。しかしながら、サーチャのモデルだけをもとにシステムを作ってしまうと、確かにサーチャにとっては効率が良いがユーザにとっては使い勝手の悪いものができかねない。サーチャの検索方法を強要するのではなく、ユーザがうまく検索できないところを支援してゆくようなシステム（「安心して失敗できるシステム」）を目指したい。

6 参考文献

1. T.W.Malone, K.R.Grant, F.A.Turbak, S.A.Brobst, and M.D.Cohen, "Intelligent Information-Sharing Systems", Commun. ACM, Vol.30, No.5, May 1987.

2. 秋山「テキスト情報の知的検索における諸問題」情報処理学会データベース・システム研究会、1988.3.

3. 三輪「情報検索システムにおけるユーザインタフェースの条件」、情報処理学会情報メディア研究会、1991.9.

4. 富士「自然言語文書からの特徴キーワード抽出」、情報処理学会第43回全国大会論文集, 1991.