システムソフトウェアと
オペレーティング・システム 74－11
マルチメディア通信と分散処理 81－11
（1997．2．27）

WWW検索システムにおける不特定多数の操作履歴の活用

原田昌紀　東京大学大学院総合文化研究科広域システム科学系
清水　奨　日本電信電話株式会社ソフトウェア研究所

キーワードに基づくWWW検索システムでは適切なキーワードを含む検索式を与えることが重要である。熟練した利用者は様々な条件を指定して適切な検索式を与えることができるが、多くの利用者にとってこうした操作は困難である。
このため適合性フィードバックやデータマイニングによる利用者支援が行われている。しかしシステム側で多くの処理を行う必要から、サーバ負荷の増大要因となっていた。
本論文では、不特定多数の利用者の検索履歴に基づき、検索結果の関連語を抽出して提示する利用者支援手法について述べる。従来の手法より容易に実装可能で、自然な関連語を提示できるのが特徴である。本手法をWWW検索システムODIN上で実装し運用した結果、全体のおよそ2割の利用者が提示された関連語を使い、AND条件による絞り込みを実施するようになった。これは導入前の3倍の割合である。一方、サーバ負荷は導入前とほぼ同程度に抑えられた。

A simple way of guidance: making relevant keyword from anonymous user's behavior on WWW search.

Masanori Harada, Univ. of Tokyo
Susumu Shimizu, NTT Software Labs.

Providing user guidance is a key component of WWW search services. Some search services are providing guidance through relevance feedback or data mining. These systems, however, are getting increasingly complex as new functions are added. In this paper, we examine a simple approach based on anonymous monitoring of user search behavior. User behaviour information is collected and presented to users as a recommended relevant keyword list. This method was tested on an actual WWW search service. It was found that, as a result of the relevant keyword list, the percentage of queries using the AND operation increased from 6% to 18%.

1　はじめに

WWW上に存在する多種多様な情報に対する検索のニーズが高まっている。すでに従来の情報検索技術を応用したさまざまなシステムが稼働しており[1]、広くサービスを提供している。

このうち、主流となっているのはフリーキーワードに基づく検索システムである。この種のシステムの動作は、キーワードを利用者に入力してもらい、あらかじめ生成しておいたインデックスに基づいて検索を実行し、結果を利用者に返すというものである。

しかし、従来から指摘されているように[2]、情報検索の専門知識を持たない多くの利用者にとって、自分の探したい情報に関連する適切なキーワード(の組み合わせ)を入力するのは困難である。WWW上の検索システムでは一般の利用者が対象となるため、システムとのインタラクションを通じて検索式やキーワードを適切に導く利用者支援機能が重要である。

たとえば京都大学のrcaau - mondou - システム[3]ではデータマイニングの手法を使って関連語を文書集

Contact: M9-311A 3-9-11 Midoricho Musashino TOKYO Japan
Email: harada@graco.c.u-tokyo.ac.jp, shimizu@slab.ntt.co.jp

合の中から求め、利用者に提示している。またNTTのNTT Directory[4]や富士通のInfoNavigator[5]では適合性フィードバックの手法を使って利用者の検索式を拡張し、より適切な内容をもった文書を検索結果として返そうとしている。またリクルートのWebdew[6]ではおよそ7万語の同義語辞書を使ってOR検索を行い、検索範囲を広げる工夫をしている。

これらの手法は文書を語の集合として捉える必要があるため、日本語のように分かち書きされていない言語では辞書を用いて形態素解析を行うのが一般的である。しかし、複合語の扱いに問題が残るほか、検索キーワードとして重要な固有名詞や新語に弱いという問題点がある。たとえば「イントラネット」という語が「イン」と「トラ」と「ネット」に分割されるなどして、元の意味が失われてしまうことがある。

こうした問題の影響を抑えるためには、高度な自然言語処理技術を必要とし、システムはより複雑化する傾向にある。特にWWW文書のように内容が多岐に渡り、かつ短期間に生成／消失する情報を扱う場合、自動処理を実現するために多くの問題を克服しなければならない。

このため、自動で処理する部分を単純化して問題を小さくし、人間が行なっている処理を積極的に利用するアプローチを考える。このアプローチでは、人間にかかる負荷を最小限に抑える事が重要であり、人間が無意識に行なっている処理を活用することが理想となる。

本論文では、こうした視点から、これまで活用されてこなかった不特定多数のサービス利用者の検索履歴に着目し、ある利用者が入力したキーワードをそのまま再利用することで、他の利用者にとっても有用な語を抽出する方法について述べる。本手法は形態素解析をはじめとする複雑な処理が不要であり、システムに対する負荷も少ない。

以下2章では従来の利用者支援手法について、適合性フィードバックとデータマイニングを代表として概観する。3章では利用者行動の観察結果から、本論文で提案するシンプルな支援手法について述べる。4章で稼働中のWWW検索システムODINにおける実装についてのべ、導入前と比べた利用者支援効果をAND検索の使用率の変化から検証する。5章で結びとする。

2 従来の利用者支援手法

本章では情報検索システムにおける従来の利用者支援手法について述べる。

2.1 適合性フィードバック

Saltonの適合性フィードバック(Relevance Feedback)[13]は、ある検索結果に対し、利用者が適合すると判断した文書を指定することにより、はじめの検索式を自動修正する手法である。適合すると指定された文書のキーワード群に基づき利用者が与えたキーワード群の重みが再計算され、再検索が行われる。この結果、利用者の情報要求に適すると期待される文書が集められる。

一方、日本語のWWW検索システムでは、ほとんどの利用者は単一のキーワードしか入力しない。第一筆者が運用しているODIN[7]では9割の利用者が単一のキーワードを用いており、千里眼[8]やHole-in-one[9]といった同種のサービスにおいても同様の傾向が確認されている[10]。

このため、日本語の検索システムでは適合性フィードバックを用いて検索式の自動拡張を行なうことが多い。NTT DirectoryやInfoNavigatorで使われているのは、利用者が適合すると判断した文書が含むキーワードリストを使って再検索を行い、それらの共起度に基づくスコアリングを行う手法である。

この方式には次のような問題点がある。

- 利用者に適合性判断という負担を強いる。
- 適合性判断の結果を処理する過程で、扱うデータが多くなるため、サーバ負荷が大きい。
- 適合すると判断した文書の組み合わせによっては、ノイズの多い結果が返る。
  ここでいうノイズとは、利用者の要求とはあまり関係のない文書である。利用者の適合性判断の結果から多くのキーワードが生成されるため、中には本来の要求とは直接関係のないキーワードが含まれる。こうしたキーワードに高いスコアが与えられる[1]と、再検索の結果あまり関係のない文書が多数含まれる事になる。

この方式で効果を出すには、システムの実装に応じた正しい適合性判断をする必要がある。しかし、う

[1]例えば、ある文書に特定のキーワードが多数含まれる場合

まくフィードバックして素早く目的の情報にたどり着くには、そのシステムにおけるフィードバックの評価やスコアリングの手法を理解する必要があり、不慣れな利用者には違和感が残ることも多い。

2.2 データマイニング

一方、データベースと AI の分野ではデータ集合の間に成立する関連を見つける研究が進められており、データマイニングや KDD(Knowledge Discovery in Databases) と呼ばれている [11][12]。

WWW 検索システムにおいては、WWW 文書の集合をデータ集合と考え、文書間のハイパーテキストリンクや文書から抽出されたキーワード群の共起度に基づいた計算から、キーワード集合を相関ルールとして導出する。利用者からの問い合わせの結果にこのキーワード集合をつけて返すことで、再検索を行う際のキーワード入力を支援する [3]。

前節の適合性フィードバックと異なり、キーワード集合が直接利用者の目に触れるため、キーワードの品質を高めることが重要である。しかし、実行時の計算コストが比較的大きいため、文書を表現する語の数を小さく抑える必要があり、出現頻度などの判断基準で語を選んでいる。また、日本語においては単語の切り出しが難しく、固有名詞や複合語、新語のキーワードを導出するのは困難である。こうした理由から現状では利用者の期待するようなキーワードを自動的に導出することは必ずしも容易ではない。

3 不特定多数の利用者の検索履歴に基づく手法

本章では、前章で述べた従来の利用者支援手法とは異なるアプローチで利用者支援を行う手法について述べる。

3.1 WWW 検索システムの利用者行動

既存のキーワードベース WWW 検索システムでは、キーワードの入力、検索結果の評価、情報閲覧という大きな流れの中で、利用者は常に検索とその見直しを行っている。利用者行動を観察して状態遷移を図示すると図 1 のようになる。

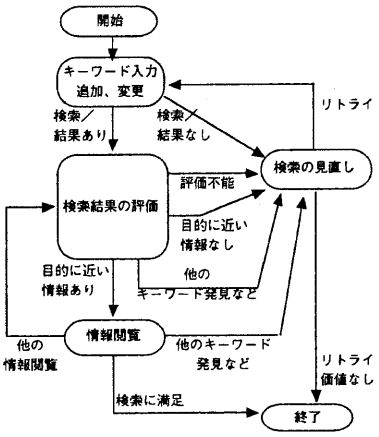

図 1: WWW 検索システムの利用者行動

検索は、利用者が検索目的を表わすのにふさわしいと思いついたあるキーワードから始まる。検索結果がなければ直ちに見直しが行われ、キーワードや検索条件を変更して再検索が行われる。利用者の過去の経験からリトライしても改善されないと判断されると、検索終了となる。

WWW 検索システムの多くでは、検索結果に情報内容の一部が含まれる。HTML のタイトルや最終更新時刻、アンカーに含まれるテキスト、あるいは文書の最初の一部分などである。この場合、利用者は情報の一部から全体の内容を推測し、自分の検索目的に合致するかどうかを逐一判断する。すなわち、検索結果の評価を行う。目的の情報に近いと判断できるものに対して、情報自身を閲覧する。

この評価の過程で、検索結果の数が多すぎて判断できなくなったり、目的の情報に近い物が見つからない場合は検索の見直しに移る。何度かセッションを繰り返して見直しをしても意味がないと判断した時点で検索終了である。

一方、WWW 検索システムのような大規模文書データベースにおいては、思いがけず大量の検索結果を得てしまうことは避けられない。検索システムは Salton の tf · idf 法 [13] などの方法に基づき検索結果をスコアリングして表示し、利用者はスコアの高い情報から評価していくことができるが、現実的に評価できるのは数十件程度である。したがって、利用者は結果を絞り込むための検索条件を検討し、再検索の必要に迫られる。

3.2 検索履歴からの関連語抽出

前節で述べたとおり、利用者はシステムとのインタラクションを通じてキーワードをいろいろ工夫して与えている。そのうえで、限られた情報から適切なURLを判断して閲覧している。

利用者AとWWW検索システムとの標準的なインタラクションは次のようになる。

1. キーワードTの入力
2. 検索の実行
3. 検索目的に合致する情報のURL U を選択
4. 情報の閲覧

このとき、利用者Aのインタラクションから、キーワードTとURL Uの相関を抽出できると考える。図1に示したとおり、情報の閲覧の前に利用者は自分の入力したキーワードと検索結果を比較し、適切かどうか判断を下しているからである。

さらに、不特定多数の利用者群についてキーワードのリストとURLのリストの相関を集めることにより、各URLに対して利用者群が与えたキーワードリストを集めることができる。

たとえば、検索エンジンに関するページ[1]にたどり着くのに、利用者Aは「検索」というキーワードを使い、利用者Bは「サーチエンジン」というキーワードを使い、利用者Cは「ロボット」というキーワードを使ったとする。この場合、[1]のURLと「検索」「サーチエンジン」「ロボット」というキーワードの間の相関は単にそのキーワードが含まれているという条件よりはるかに大きいと期待できる。また、同じURLを介しているので、これらのキーワード同士は互いに関連していると考えることにする[2]。

次に、利用者Dが「検索」というキーワードで検索をかけたとき、システムは先のURLが持っているキーワードリストから「サーチエンジン」「ロボット」を関連語として提示できる。

以上のように、利用者が入力したキーワードと、その利用者が選択したURLの間の相関を集めることで辞書を使ってデータを処理することなく関連語の候補を抽出できる。ある利用者がキーワードであると判断して入力した単語がそのまま関連語になるため、自然な語の提示が期待できる。しかも、システムは利用者の検索行動を記録して処理するだけでよく、比較的軽い処理ですむことが予測できる。ただし関連語を提示できるまでに、ある程度数のインタラクションを集める必要がある。

[2]後に、この前提が元で発生するノイズについて触れる

4 実装と評価

我々は本手法をWWW検索システムODIN[7]に適用した。本章ではその実装および運用状況について述べ、明らかになった問題点について言及する。

ODINの利用者はWWWブラウザを用いて、検索システムが置かれているサーバにアクセスし、キーワードを入力することで、検索結果をHTML形式で受けとることが出来る。検索結果には入力されたキーワードを含むWWWページのURLと要約が列挙されており、加えて本手法で導出された関連語が表示される。

利用者はそれらの関連語の一つを選択することで、はじめのキーワードと選択した関連語をAND条件で結合した検索式を作り、再検索を行うことができる。本手法のアルゴリズムの性質上、導出された関連語は検索結果に含まれていたWWWページのうち少なくとも一つには含まれているので、AND条件を満たすWWWページが存在することは保証されている。

4.1 システム構成

関連語導出機能を加えたWWW検索システムの構成を図2に示す。

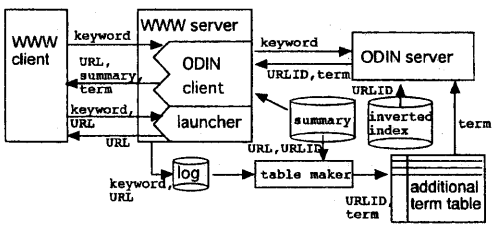

図2: WWW検索システムの構成

本手法を実装するには利用者が入力したキーワードとジャンプ先URLの組を記録する必要がある。しかし、WWWにおいてクライアント・サーバ間の通信に用いられるHTTPはステートレスなプロトコルであり、利用者の行動を追跡する手段がないため、ジャ

ンプ元のページのどのリンクを辿ったかを知ることはできない。そこで検索システムの結果出力部に手を加え、

```
<A HREF="url">タイトル</A>
```

というようにWWWページを指すリンクを生成する代わりに、

```
<A HREF="/cgi-bin/launcher?
URL=url&KEY=key">タイトル</A>
```

のように検索サーバ上のCGIプログラムを起動するリンクを生成するようにした。launcherは、ブラウザに別のURLへ移動するように指示するだけのプログラムであるが、CGIプログラムの起動がサーバのログに記録されるため、キーワードとジャンプ先、その時刻などを集計することが可能になる。現在はブラウザに余分な動作をさせることによる利用者の負担の増加を抑えるため、launcherプログラムをWWWサーバプロセスにモジュールとして組み込むことで高速化したり、JavaScript言語を用いることで、外見を通常のリンクと同じようにするなどの工夫を施している。

さて、十分な量のキーワードとURLの組(相関ルールリスト)がログに蓄積されてきたら、集計してURLとそのURLに関連づけされたキーワードリストの表を求める。これはオフラインでバッチ的に行なう処理である。ODINでは複数のキーワードをスペース句切りで並べることでOR条件やAND条件で検索することが可能だが、本実装では簡単のため、キーワード一語のみで検索が行なわれた場合に限って相関ルールリスト作成に利用することにした。

実際の検索エンジン内部ではURLは数値で表現されているので、その数値からキーワードリストを得るための表を作ることになる。以下ではそれぞれURLID、関連語表と呼ぶことにする。

検索時の関連語の導出は以下のようになる。検索サーバは与えられたキーワードを用いて、まず通常の文字列検索を行ない、スコアリングされたURLIDのリストを得る。次に得られたそれぞれのURLIDに関連づけられたキーワード群を関連語表から読みだし、関連語候補として数え上げていく。そして関連づけられているURLの数を基準に関連語候補をソートし、上位に来る語を関連語であると決定する。つまり、検索結果中で関連づけられているURLの数がキーワードとの関連の強さを表していると考えていることになる。

4.2 ODINでの利用経過

国内のWWWサーバから回収した約22万URL分のデータベースに対して97年1月8日から11日間の間に24万件の検索が実行され、19万6千組の相関ルール(ジャンプ)が記録された。このログから生成された関連語表では、1URLあたり平均して0.45語が関連づけられており、ジャンプ先として1回以上関連づけられたページは約5万URLとなった。この段階で関連語表のファイルサイズは約7MBと十分小さく、また検索時にランダムにアクセスされるものであることから主記憶上に置かれている。そのため、本機能による速度低下はほとんど見られない。

本手法により提示される関連語の例を表1に示す。キーワードの後ろの括弧内の数字は検索結果のURL数を、関連語の後ろの括弧内の数字は関連づけられているURL数である。

"天気" (637)	天気 (17), 天気図 (8), 天気予報 (8), プレゼント (5), 気象 (5), 風速 (4), 気象庁 (4), 気象予報士 (4), 気象衛星 (4), realaudio(4), 占い (4), ニュージーランド (3), 地震 (3), うらない (3), ペルー (3)
"経済企画庁" (242)	経済企画庁 (9), 郵政省 (6), 厚生省 (5), 大蔵省 (3), 野村証券 (3), 港湾 (3), 環境庁 (3), 弁護士 (3), シンクタンク (3), 経済白書 (3), 農林水産省 (3)
"namco" (92)	ソウルエッジ (11), namco(10), ナムコ (7), playstation(5), レイジレーサー (4), tekken(3), enix(3)
"源氏物語" (52)	源氏物語 (7), 樋口一葉 (3), 夏目漱石 (3)

表1: 導出された関連語の例

最終的に利用者に提示する関連語はこの数値が4以上になるものとした。また、キーワード自身は関連語として意味がないので取り除いている。したがってこの例では結果として「源氏物語」の関連語は提示されないことになる。しかし、ある程度の検索結果数になる語に関しては、すでに述べたような

1. 他の利用者がキーワードであると判断した語が提示される

2. 複合語や新語の問題がない

という特長から十分実用的な語が提示されていると思われる。

このような関連語による絞り込み提案機能を公開したところ、全検索においてAND機能による絞り込みを利用した検索の占める割合は公開前(1/13 – 1/19)が約6%であったのに対し、公開後(1/21 – 1/27)は約18%となった(表2)。

	公開前	公開後
総検索要求 (a)	166,861	216033
AND 検索 (b)	10,055	39,025
AND 使用比 (b/a)	0.060	0.181

表2: 関連語絞り込みの利用状況

この利用率の変化から現実のサービス上での有効性が確認できたと考えられる。

4.3 課題

本システムでは、しばしば入力されたキーワードとはまったく無関係な語が関連語として出力されてしまうことがある。これは主に雑多な内容を持ったページが存在するためである。

特にファイルサイズの大きなページは多くの語を含んでおり、様々なキーワードにマッチするため、関連語の質を下げる傾向が見られた。「リンク集」と呼ばれる類のページにはそうしたページが多いようである[3]。現在は1ページあたりの関連語候補の数を制限するという対策をとっているが、検索結果がページ単位となっている限りは、避けるのが難しい問題である。本手法で求められる関連語は、検索結果の文書集合と関連する語であって、必ずしも入力キーワードとは関連しないともいえる。

また、検索結果数が少ない場合には関連語が抽出できないことが多い。今回は絞り込みに利用したので自然な結果であるが、関連語が一つ以上出ることの保証がないために他の応用が限られてしまうのが問題である。

5 終わりに

本論文では、不特定多数の利用履歴に基づくシンプルな関連語抽出を行うことで、サーバにほとんど負担をかけずに利用者支援を行えることを示した。従来あまり使われなかったAND演算子による絞り込みが多くの利用者に使われるようになったという点で効果があった。

本論文で述べた環境は、不特定多数の利用者による協調作業環境であると考えることができる。この場合、協調作業に参加する利用者が何も特別な作業をせずに知的作業の結果(キーワードとURLの相関)を共有できる点が大きな特徴の一つである。

一方、効果を上げるためにはある程度の利用者数を獲得する必要がある。また、キーワードとURLの相関を判断する手がかりが現在はタイトルやタグづけされたテキストからなる数行の要約だけであることから、この数行に重要な情報が現われない場合は有効なキーワードを抽出することができない。しかし、書誌情報の検索システムなどでは本手法が直接適用できるのではないかと考えられる。

今後は、この手法のクリティカルマス(有効に機能する利用者数の規模)がどの位であるか調べる。また、ノイズとなる語を減らす方法について考える。

[3] 入力キーワードと一見関係のない関連語でAND検索をしてみるとリンク集的なページが残ることを確認できる

参考文献

[1] 清水, 日本の *Search Engine* のリスト, http://www.ingrid.org/w3conf-bof/search.html

[2] P. Ingwersen, *Information Retrieval Interaction*, 邦訳:細野ほか、情報検索研究 -認知的アプローチ-, トッパン 1995

[3] 河野, 長谷川 *WWW* データ資源検索におけるデータマイニング手法, 情報処理学会研究報告 DBS No.45(108-5), pp.33–40(1996), http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/

[4] NTT Directory, http://navi.ntt.co.jp/

[5] InfoNavigator, http://infonavi.infoweb.or.jp/

[6] Webdew/あちゃら Navi, http://www.acara.com/

[7] ODIN – Open Documentary Information Navigator, http://kichijiro.c.u-tokyo.ac.jp/odin/

[8] 千里眼, http://www.info.waseda.ac.jp/search.html

[9] Hole-in-one, http://www.hole-in-one.com/

[10] w3-search メーリングリストアーカイブ, http://rodem.ingrid.org:8080/hypermail/w3-search/0047.html

[11] Fayyad, M.U., *Proceedings of the KDD-95*, AAAI Press,1995

[12] ACM, *Data Mining and Knowledge Discovery in Databases*, CACM Vol.39,No.11,1996

[13] Salton,G. and McGill,J.M. *Introduction to Modern Information Retrieval*,McGraw-Hill,1986