3C2-01

検索語の履歴を利用して再検索の支援を行うシステム

Refining keyword search based on the history of entered keywords

大井雅文　　岡夏樹
Oi Masafumi　　Oka Natsuki

京都工芸繊維大学大学院 工芸科学研究科
Graduate School of Science and Technology, Kyoto Institute of Technology

When we use a search engine of the Web, search results could become huge or too few, depending on search terms. For this reason, it is often difficult to reach information that we need. We propose a method of displaying useful related terms for the following search making use of a list of related terms, which is based on search terms used in the past. We have implemented the system with Namazu, and we have evaluated it with the test collection of NTCIR.

1. はじめに

近年，インターネット人口の増大と共に，あらゆる情報を探索する手段として検索エンジンが用いられるようになった。2004 年に行われた Yahoo!JAPAN の調査 [1] でも “82%のインターネットユーザは検索エンジンを介して商品やサービスを調べている” という結果が出ており，検索エンジンはさらにその役割を高めていくと考えられる。

現在主流となっているロボット型検索エンジンの課題は，与える検索式によっては求める情報が表示されない，または関係の無い情報が膨大に表示されてしまうということが挙げられる。ユーザが探したい情報に対する知識が少ない場合や漠然とした情報を探している場合には，適切な検索式を与えることは特に困難である。

本研究はそれまでに検索エンジンで使用された語句を利用し，入力語に対する関連語を提示，選択させることによってユーザの意思を明確にさせることを目的とする。

2. 現在の検索支援技術

2.1 適合フィードバック

検索システムを利用するユーザが目的とするページにたどり着くためには，システムに自らの要求を適切に伝える必要があり，そのための手法として適合フィードバックが提案されている。

適合フィードバックとは出力された検索結果に対してユーザが評価を行い，その評価を基にシステムが再検索を行うというものであり，サポートベクターマシンを用いる手法 [2] など様々な手法が提案されている。しかし従来の手法では，評価のために検索結果に表示されたページの内容を理解しなければならなかったり，多くの文章を評価する必要があったりとユーザへの負荷が大きかった。

そこで本研究では入力語句に対する関連語を提示し，ユーザの要求をページの評価ではなく提示語句の評価で汲み取ることを目指す。システムが適切な関連語を提示することにより，ユーザは自らの要求に合った語を選択し再検索を行う。なお，提示する関連語は今までに入力された検索式の履歴を利用して生成する。

連絡先: 大井雅文，京都工芸繊維大学大学院パターン情報処理研究室，E-mail:ooi@vox.dj.kit.ac.jp

2.2 検索履歴の利用

検索履歴の利用によって関連語を提案する利点としては，

- 検索したい分野に関する知識の少ない人が，その分野に関してキーワード設定能力の高い人との知識の共有が出来ることで，よりよい検索語を得ることが可能となる
- 関連語として提示される語句は人間が入力するものを用いるので，形態素解析時に時々起こるような分割ミスによる不自然な単語が出現することはなく，普段人が用いるような自然で有用な語句の提示が可能となる

といったことが挙げられる。ユーザーの検索履歴を用いた検索システムの提案は，これまでにもいくつかあるが [3][4]，関連語の学習のために，かなり多くのユーザーに検索してもらう必要があったり，2 語以上の単語で AND 検索を行った場合にしか学習が行われなかったりするもので，大規模なデータベースにおいて用いる場合はともかく，利用者数のそれほど多くない中小規模のデータベースで使用する場合は，ほとんど履歴が蓄積されず，提示する関連語がわずかになってしまうことが考えられる。

そこで今回の実装では，ユーザーが用いた検索語で表示された Web ページ全てとその検索語を関連付けることにより，少ない検索回数で多くの関連語を提示することを可能にした。

3. システムの概要

作成したシステムは学習フェーズと出力フェーズに分けられる。なお，基となる検索エンジンには Namazu[5] を使用した。

3.1 学習フェーズ

入力は一般の検索エンジンで用いられている方式と同様で，テキストボックスに検索語を打ち込むものである。

ユーザーが検索を実行すると，画面には，Namazu によって検索された URL 群が出力され，その際表示された各 URL と入力された検索語の組を生成し，ファイルに書き出す (図 1)。

このようにしてユーザーに繰り返し検索を実行させ，データがファイルに蓄積されれば関連語リストの作成を行う。ここで関連語リストは外部 CGI により

1. 一つの URL に同じ検索語が複数回関連付けられていれば，その回数をカウントし，重複を取り除く。
2. カウントされた数値を検索語と組み合わせる。

図 1: ファイルへの出力

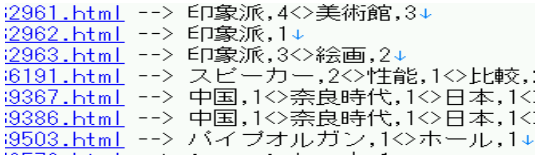

図 2: 関連語リスト

3. もし一つの URL に異なる検索語が関連付けられていれば，一行にまとめる。

以上の処理を行うことによって生成する（ 図 2)。なお検索回数をカウントしているのは，関連語提示の際に重みとして利用するためである。

3.2 出力フェーズ

まず，ユーザがある検索語にて検索を行った際，作成された関連語リストを基に表示された web ページ群に関連付けられている単語を読み込む。読み込まれた単語群は，その単語が含まれている web ページ数と単語の検索回数によってソートされ，表示される。ここで検索回数が少なすぎて不要と思われる単語は表示しない。ユーザーは提示された単語の中から、自分の求める情報に近い単語を選択し，それを基にシステムは再検索を行う。今回の実装では 20 ページを閾値として，それ以上のページが出力された場合ば関連語選択にて AND 検索，それ以下だと OR 検索を実行するようにした。

なお表示された関連語は全ていずれかの URL と組になっているため，関連語を用いて再検索を行った際，1 件以上の検索結果が出力されることは保証されている。

4. 評価実験

実験の評価は，国立情報学研究所の提供する Web 検索用テストコレクション [6]NTCIR － 3 WEB を用いることによって適合率と再現率の推移を計算することによって行った。

4.1 適合率と再現率

検索結果にはほとんどの場合，ユーザーの目的となるページ以外に不要なページも含まれている。適合率とは検索結果中の正解の割合，つまり

$$\text{適合率} = \frac{\text{表示された目的のページ数}}{\text{表示された全てのページ数}}$$

を示し，再現率とは検索結果として現れた正解がすべての正解に占める割合，つまり

$$\text{再現率} = \frac{\text{表示された目的のページ数}}{Web\text{ 上に存在する全ての目的ページ数}}$$

を示す。もちろん両方の値が高ければ優れたシステムであると言えるのだが，一般的に適合率と再現率はトレードオフの関係にあり，正解として出力するページ数を多くすれば再現率を向上させることができるが，適合率が低下してしまうことになる。

今回は，提示される関連語を用いて再検索を行った場合に，適合率や再現率がどのように変化するのかを測定する。

4.2 実験方法

実験ではテストコレクション中より，データベース用に約 20MB 分（966 ページ）の文書データ，検索履歴の作成用に 50 件の検索課題を使用した。また，別に 3 件の検索課題を関連語リスト作成後にテスト用に使う検索課題として用いた。

1. 検索履歴作成
   10 人の学生に分野の異なる 50 の検索課題を提示し，自分ならそれぞれの課題に対してどのようなキーワードで検索を行うか答えてもらい，その結果を入力することで検索履歴の作成を行う。また，比較のためにその半分である 25 の検索課題を提示したときの検索履歴も作成する。

2. 関連語リストの作成
   外部 CGI を実行して検索履歴を基にした関連語リストを作成する。

3. 検索実行
   関連語リストを作成した後，履歴の作成に使ったものとは別の検索課題を用いて検索を行う。最初の入力に用いる検索語は，10 人の学生にアンケートを取り，最も多かった語を一般的なものとして採用する。

4. 関連語の記録
   検索実行時に提示された関連語を記録する。複数個表示された場合は上位 10 個のみを対象とする。

5. 適合率と再現率の遷移を測定
   検索実行時に出力された結果の適合率と再現率を記録し，関連語の選択を行った際の適合率と再現率の変化を見る。今回は適合率や再現率の向上が大きかった単語に対して推移を示す。

5. 結果

今回テスト用に使用した検索課題は以下の 3 つである。

検索課題 1

ピラミッドに関連した古代エジプト美術を見るのに良い美術館はどこか？

－この課題に対して正解となるページはデータベース中に 14 件存在する
初期入力検索式は「ピラミッド」「美術館」「古代エジプト」

検索課題 2

京都の寺や神社について，歴史的背景，地域での存在など，一歩踏み込んだ情報を知りたい

－この課題に対して正解となるページはデータベース中に 8 件存在する
初期入力検索式は「京都」「寺」「神社」「歴史」

検索課題 3

情報処理や IT と言った分野の資格試験にはどのようなものがあるのか知りたい

－この課題に対して正解となるページはデータベース中に 7 件存在する
初期入力検索式は「情報処理」「ＩＴ」「資格」

なお 10 人に 50 の検索課題を与えて検索を行わせた結果，履歴に蓄積された語句は 179 個，半分の 25 の検索課題を用いた場合は 87 個であった。

5.1 適合率・再現率の変化

先程の入力語によって提示された語句を選択した際，適合率・再現率がどのように変化するかを測定した。提示語句が複数個ある場合は，選択によって最も適合率・再現率の向上が見られた語句を使用した。なお UniSci の調査 [7] により，７０%のユーザーは検索結果の上位２０件までしか見ないという結果が出ていることに基づき，適合率と再現率は上位２０件のページのみで計算している。

検索がうまくいかなかったときの再検索支援のため，実験にはそれぞれの検索課題に対して最も検索結果の適合率の低かった入力語を使用した。検索課題１では「美術館」を入力した場合，検索課題 2 では「京都」を入力した場合，検索課題 3 では「ＩＴ」を入力した場合である。また，履歴の作成の際には 50 の検索課題を用いた。

検索課題 1

1. 「美術館」を入力
   - 検索結果： 60 件
   - 適合率 ： 35.0%
   - 再現率 ： 50.0%
   - 提示された語句
     「印象派」「絵画」「梅」「画家」「東京」「宮崎駿」「夢」「名所」「統計」「作品」

   提示された語句を選択した時に，最も適合率・再現率の向上が見られた語句は「絵画」であった。
2. 提示された語句の「絵画」を選択
   - 検索結果： 14 件
   - 適合率 ： 35.7%
   - 再現率 ： 35.7%

検索課題 2

1. 「京都」を入力
   - 検索結果： 125 件
   - 適合率 ： 5.0%
   - 再現率 ： 16.7%
   - 提示された語句
     「安倍」「日本」「歴史」「中国」「夢」「インターネット」「平安時代」「奈良時代」「文化」「株」

   提示された語句を選択した時に，最も適合率・再現率の向上が見られた語句は「歴史」であった。
2. 提示された語句の「歴史」を選択
   - 検索結果： 47 件
   - 適合率 ： 10.0%
   - 再現率 ： 25.0%

検索課題 3

1. 「IT」を入力
   - 検索結果： 75 件
   - 適合率 ： 5.0%
   - 再現率 ： 14.3%
   - 提示された語句
     「株」「インターネット」「ネット」「入門」「石川」「オンライン」「著作権」「デジタル」「株式投資」「ネットワーク」

   提示された語句を選択した時に，最も適合率・再現率の向上が見られた語句は「ネットワーク」であった。
2. 提示された語句の「ネットワーク」を選択
   - 検索結果： 21 件
   - 適合率 ： 20.0%
   - 再現率 ： 57.1%

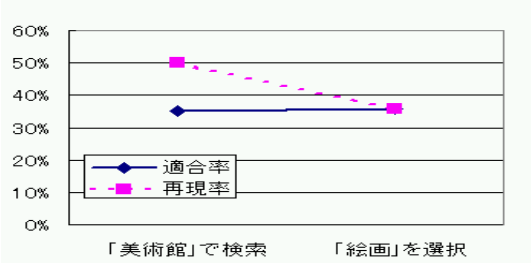

図 3: 検索課題 1 にて提示された関連語「絵画」を選択した時の適合率・再現率の推移

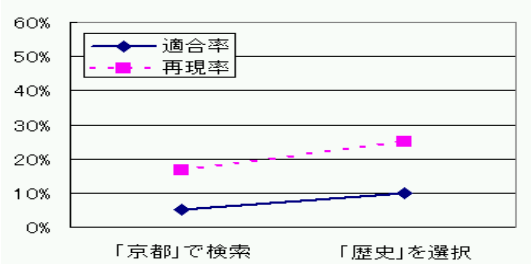

図 4: 検索課題 2 にて提示された関連語「歴史」を選択した時の適合率・再現率の推移

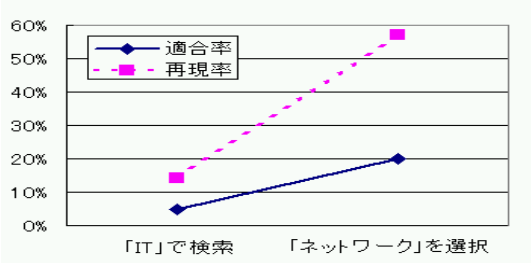

図 5: 検索課題 3 にて提示された関連語「ネットワーク」を選択した時の適合率・再現率の推移

5.2 考察

- 履歴作成の際の検索課題の数
  提示された関連語は履歴作成の際に用いた検索課題の数によって大きな違いが見られ，50 個の検索課題を用いた方が多くの関連語を得ることが出来た。今後履歴をさらに多くの検索課題を用いて作成すると，提示される語句の数も多くなるであろう。だが関連語の表示順位に気をつけなければ，不要な語句ばかり表示される可能性がある。

- 履歴作成の際の検索課題の取り方
  履歴の作成の際の検索課題を増やせば多くの関連語が得られた一方で，提示された語句が検索課題の数によらなかったものも存在した。これは，履歴作成の際の検索課題の取り方に原因があったと考えられ，もし特定の分野のデータベースを探索する際にこのシステムを利用すると，少ない検索履歴でも有用な語句の表示が可能であると推測できる。特定の分野で検索を行う場合，ある検索目的で使用される検索語が，他の多くの検索目的に対しても有効であることが多いためである。

- 適合率・再現率の遷移
  関連語選択の際の適合率・再現率については，最初から数値の高かった検索課題 1 を除いて向上[*1]が見られ，今回の実験で提示された関連語は有用であったことが確認された。しかし検索課題 3 の「ネットワーク」という語句のように，最も適合率・再現率の向上が大きい語句が下位に表示されることもあり，更なる表示順位の工夫が必要である。

6. おわりに

本研究では，検索履歴に基づく再検索支援の手法を提案した。今回の実験で提示された語句の中には，選択することによって有効な再検索が行えるものが存在し，検索履歴を利用することによる利点が見受けられた。

だが，それと同時に不要であると思われる語句も多く存在した。これは主に履歴に蓄積された語句が 179 個と非常に少なかったことや重み付けによる不要語句の除去がうまくいかなかったこと，また今回は一般の web ページを基にデータベースを作成したため，検索対象となるページは様々な分野のもので，履歴の作成の際に用いた課題にも共通する部分があまり無かったことなどに原因があったと考えられる。

今回は関連語選択による AND 検索のみでしか再検索を行わなかったが，今後は関連語を用いた OR 検索や NOT 検索，またその他の手法による再検索の手法も考えなければならない。さらに蓄積された語句の中から，ユーザの意図を汲み取るために有用な語句のみを選別すること，またその語句を用いてユーザに問い掛けるなどしてより対話性を高めること，従来のシステムとの比較を行い利点をさらに明確にすることなどを課題とする。

謝辞

本研究の一部は、平成１６年度科学研究費補助金基盤研究 (B)(2)「伝統産業技術の継承と発展のための新しいマルチメディアアーカイブ技法の研究」（研究代表者：黒川隆夫）の支援を受けた。

*1 なお適合率と再現率の両方が向上しているケースが存在するのは，今回上位２０件のみを対象に計算を行ったためである。

参考文献

[1] Yahoo! JAPAN: “第１６回インターネット利用者アンケート結果”, http://docs.yahoo.co.jp/info/research/wua/200410/, 2004 年.

[2] 村田博士、小野田崇、山田誠二：“適合フィードバックにおける非適合文書からの文書検索” , 第 18 回 人工知能学会全国大会、2004 年.

[3] 原田昌紀、清水奨：“WWW検索システムにおける不特定多数の操作履歴の活用”, 情処ＤＰＳ研究会 97-DPS-81, 1997 年.

[4] 栗本亜実:“WWW情報検索ナビゲーションシステムの設計と実装”, http://www.sfc.wide.ad.jp/thesis/2002/bachelor/ami/b-thesis.pdf, 2002 年.

[5] namazu http://www.namazu.org/.

[6] ＮＴＣＩＲ 情報検索システム評価用テストコレクション構築プロジェクト: http://research.nii.ac.jp/ntcir/index-ja.html.

[7] UniSci:“Search Engine Users Look Less For Sex, Entertainment” http://unisci.com/stories/20022/0403026.htm, 2002 年.