絞り込み語提示による一検索支援手法の提案

酒井 浩之　　大竹 清敬*　　増山 繁
sakai@smlab.tutkie.tut.ac.jp, kohtake@slt.atr.co.jp, masuyama@smlab.tutkie.tut.ac.jp
豊橋技術科学大学　知識情報工学系

1　はじめに

最近の計算機の急激な性能向上とインターネットの普及により，膨大な情報が計算機上でアクセス可能になりつつある．そこで，必要な情報を効率良く得るための情報検索技術が，きわめて重要になってきている．現在の情報検索システムで用いられている主要な検索方法として，検索者のキーワード入力による検索がある．しかし，求める情報を的確に検索して利用することは容易ではない．その理由について，我々は，検索者にとって検索要求を数少ないキーワード[1]で素早く具体的に表現することが困難であるためであると考えた．

文献 [2] は，検索者が入力する検索式は検索者の検索対象分野に対する知識と経験によって異なるとし，専門家と一般の人とで検索式のキーワード数，および，検索精度を比較している．その結果，専門家が作成した検索式のほうが多くのキーワードを含み，かつ，精度の良い検索結果が得られたと報告されている．この結果は，検索対象分野に対する検索者の知識が不足している場合，検索者の知識から導くことのできる数少ないキーワードで，検索要求を具体的に表現することが困難であることを示している．もし，検索者が入力したキーワードが検索要求を具体的に表現するのに不十分であれば，その検索結果は次の2つのいずれかになる可能性が高い．

- 検索者の検索要求に適合しない文書が検索される．
- 検索者の検索要求に適合した文書の一部しか検索されない．

この問題を解決するには，入力した検索式に関連のある語を何らかの方法で導き出し，検索式を修正することが有効である [3]．検索要求に適合しない文書が検索される場合は，検索式に語を追加して AND 検索を行う．検索要求に適合した文書の一部しか検索されない場合は，検索式に語を追加して OR 検索を行う．文献 [4] は，データマイニングによって検索要求キーワードから導出されるルールを用いて，関連のある語を抽出する手法を提案している．文献 [5] は，tf・idf 法と文書構造から得られる情報を用いて語の重み付けを行ない，キーワードとして抽出する方法を提案している．文献 [6] は，検索式の拡張は全体興味を構成する要素を AND 条件で結びつけることであるとし，さらに，全体興味は部分興味を OR 条件でつないだものとして，抽出したキーワードを使った検索式拡張を提案している．文献 [7] は，分野別のデータベースを用意し，その中からファジィ発想推論を用いてキーワードを抽出し，OR 条件で検索式に拡張する手法を提案している．

我々は，文献 [8] で，絞り込みに適した語を検索結果の上位文書群から自動的に抽出する手法を提案した．上位文書群から抽出するので，既存の文書順位付けの機能を有する検索システムにただちに適用可能であり，また，どのような大規模検索システムにも適用が可能である．　文献 [8] では，検索者が選択した語を用いて，以下のように検索式を拡張している．

$$Q \cap W_1 \cap W_2 \cap W_3 \cap \ldots \cap W_n$$

ここで，Q は検索者が入力した検索式，$W_1, \ldots, W_n$ は検索者が選択した語である．拡張された検索式は，検索者が入力した検索式による検索結果の中で，選択した語を全て含む文書を検索する．しかし，この拡張では，検索者が複数の語を選択した場合は，検索結果を絞り込み過ぎてしまう場合が多い．提示された語の中から検索要求と最も関連のある語を1つ選択できる場合はよいが，検索対象分野に対する知識が不足している検索者にとって，そのような判断を行なうことは困難である．そこで，提示された語を用いて検索する場合，提示された語の中から検索要求と関連のある語を複数選択し，選択した語を複数含む文書を上位に順位付ける検索結果が得られる方が，検索者にとって検索しやすいと我々は考えた．我々は，この考えに沿って手法を改良した．本稿では改良された我々の手法について述べ，さらに，評価実験を行なったので結果を報告する．

*現在，ATR 音声言語通信研究所

[1] インターネット上の検索サイトのひとつ Excite (http://www.excite.com) に入力されるキーワードの数は，平均 2.35 個であるという報告がある [1].

2 本検索支援手法

2.1 検索の流れ

本検索支援における検索作業の流れは，文献 [8] と同様である．最初に，検索者は検索式を入力する．システムは入力された検索式で検索を行ない，検索結果の上位文書群から語を抽出し，提示する．検索者は提示された語群から検索要求に適した語を自由に複数選択する．システムは選択された語を用いて検索式を拡張し，拡張された検索式で再検索を行なう．その再検索結果における上位文書群から語を抽出し，提示する．検索者は新たに提示された語の中から，再び語を複数選択し，検索要求に適した語の選択数を増やしていく．

2.2 語の抽出手法

検索結果の上位文書群からの語の抽出手法は文献 [8] の手法を用いる．具体的には上位文書群 S の文書 s に含まれる複合名詞と，カタカナ，英字表記の語，地名，組織名を語とし，その語に以下の式で重み $W(w,s)$ をつける．

$$W(w,s) = tf(w,s) \times \log(|S|/df(w)) \times \log(dt(w)/tf(w,s)) \times \log(|S| - n)$$

$tf(w,s)$: 文書 s における語 w の頻度，
$df(w)$: 上位文書群 S で語 w を含む文書数，
$dt(w)$: 上位文書群 S における語 w の頻度，
n: 文書 s の順位，

さらに，抽出元の上位文書群における複合名詞と，カタカナ，英字表記の語の頻度を比較して，頻度が多い方の語の重みが大きくなるように重み付けをする．例えば，抽出元の文書群において複合名詞の頻度の方が大きければ，以下の式の計算値を複合名詞の語の重みに乗算する．

$$\frac{\text{複合名詞出現頻度}}{\text{カタカナ，英字表記語出現頻度}}$$

こうして算出された重みが高い語から順番に，検索者に提示する．

2.3 選択された語による検索式拡張

入力された検索式は，検索者が選択した語を用いて以下のように拡張される．

$$Q \cap (W_1 \cup W_2 \cup W_3 \cup \ldots \cup W_n)$$

ここで，Q は検索者が入力した検索式，$W_1, \ldots, W_n$ は検索者が選択した語である．拡張された検索式は，検索者が最初に入力した検索式による検索結果の中で，選択した語を 1 つでも含む文書を検索できる．次節で述べる文書順位付け処理と併用することで，拡張された検索式による再検索の結果は，選択した語を複数含んだ文書が上位に順位付けされる．検索者が検索要求と関連がある語を複数選ぶことが可能な場合は，再検索の結果は検索要求と関連のある語を複数含んだ文書が上位に順位付けされる結果となる．

2.4 文書順位付け処理

本システムにおける文書順位付けは，検索式の単語ベクトルと文書との単語ベクトルの内積によって類似度を算出することで行なった．各単語ベクトルは，語の重みを要素とするベクトルとして表現する．検索式の単語ベクトルは，検索式を構成する語 (検索者が選択した語も含む) の重みを 1，その他の語の重みを 0 として，ベクトルを作成する．文書の単語ベクトルは，その文書中に含まれる語の重みを tf·idf 法によって算出した値とし，計算した語の重みを要素とするベクトルを作成する．すなわち，ある検索結果文書群 S における文書 s に含まれる語 w の重み $W(w,s)$ は，以下の式によって求める．

$$W(w,s) = tf(w,s) \times \log(|S|/df(w))$$

ここで，$tf(w,s)$ は，文書 s における語 w の頻度，$df(w)$ は，検索結果文書群 S において，語 w を含む文書数である．こうして作成した検索式の単語ベクトルと文書の単語ベクトルとの内積を求め，値が大きい文書順に表示する．

2.5 選択できる語が提示されなかった場合の対策

もし，提示された語群から検索要求に関連した語を検索者が選べない場合は，以下の方法で新たに語を提示する．

提示された語の中に選択できる語が存在しないならば，抽出元の文書群には検索要求に適合する文書が存在しない可能性がある．そのため，抽出元の文書群を変更する必要がある．その際，提示された語の中で，検索者が選択しなかった語を含まない文書群を新たな抽出元とする．システムは，検索者が選択しなかった語を検索要求と関連がない語と判断する．検索要求と関連がない語を多く含む文書から語を抽出しても，検索要求と関連がない語が抽出される可能性がある．そこで，検索者が選択しなかった語を含まない文書群から新たに語を抽出することで，検索要求と関連がない語が提示されることを軽減する．以下の検索式で，提示された語群の中で，選択しなかった語を含まない文書群を検索する．

$$Q \; NOT \; (T_1 \cup T_2 \cup T_3 \cup \ldots \cup T_m)$$

ただし，Q は検索者が入力した検索式，$T_1, \ldots, T_m$ は提示された語群の中で検索者が選択しなかった語である．この検索式で検索された上位文書群より語を抽出して提示する．

3 システムの実装

上記の手法を実装した．システムは Linux(Vine Linux2.0) 上で動作し，語の抽出部分を perl，インター

フェイス部分を JAVA で実装した．本検索支援手法は既存の検索システム上にただちに適用できる．我々は，検索エンジンとしてフリーの検索エンジンである Namazu[2] を用いて検索を行ない，その検索結果に対し，上記に示した方法で文書順位付けを行なっている．語の抽出元として，検索結果の上位 100 文書を採用した．また，形態素解析器として JUMAN[3] Version 3.5 を使用した．図 1〜図 2 にシステムの実行例を示す．

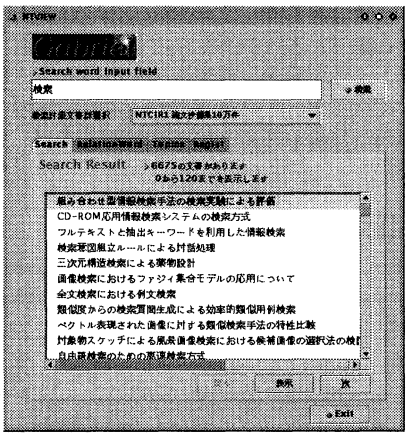

図 1: システムの実行例 1 検索結果表示

図 1 の実行画面では，検索式としてキーワード「検索」

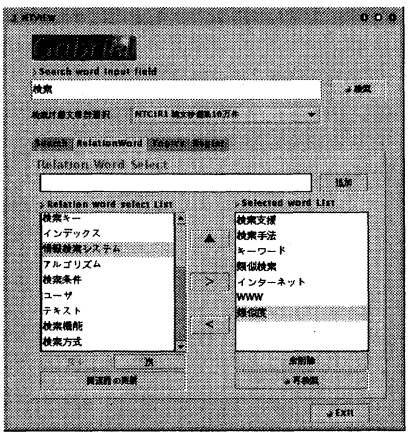

図 2: システムの実行例 2 語の提示

を入力したときの検索結果を表示している．図 2 の実行画面では，キーワード「検索」に対して抽出された語が左のリストに提示されている．検索者は左のリストから検索要求と関連した語を選択する．また，検索者自身が語を入力して追加することもできる．検索者が選択した語は右のリストに表示されている．語が複数選択できなければ「関連語の更新」ボタンを押して新たな語を提示する．語を複数選択できたなら「再検索」ボタンを押す．システムは検索式を拡張し，再度，検索を行なって検索結果を表示する．

4 評価実験

4.1 実験方法

我々は実装したシステムを使って評価実験を行なった．従来，本手法のような検索者のインタラクションを介した検索システムと，インタラクションなしのシステムでの性能の比較評価を行なう試みがなされてきた．そこでは，インタラクションなしのシステムにおける平均適合率が 46.8 %であるのに対し，インタラクションありのシステムにおける平均適合率は 42.5 %であるという結果が報告されている [9]．しかし，本稿で提案したような検索支援システムの真価は適合率，再現率といった古典的な評価尺度では評価できないと考える．なぜならば，検索者のインタラクションを介する検索システムでは，必ずしも最初の検索結果が適合率の高い結果である必要がなく，インタラクションを繰り返して，自分の求める文書を検索できることを目的としているからである．我々は複数の被験者に，複数の検索課題を与えて実際に検索を行なってもらい，アンケートを行なうことで，目的の文書を検索する場合の検索のしやすさを評価することを試みた．その際，本検索支援システムが検索の役に立っているならば，検索時間の減少が期待できると考え，語の提示ありの場合と語の提示なしの場合の検索時間を比較する．我々は，以下のような方法で評価実験を行なった．

- 複数の検索者に複数の検索課題を与えて検索を行なう．ここで，1 人の検索者は，与えた検索課題の半分は語の提示あり，残りの半分は語の提示なしで検索を行なう．
- 検索課題に対して適合していると検索者が判断したら，その文書を選択する．
- 選択した文書が一定数集まったら終了する．終了するまでの時間，および，実験終了後のアンケートで評価する．

本評価実験では，学生 4 人の被験者に 6 つの検索課題を与えて評価実験を行なってもらった．すなわち，1 人の被験者に対して 3 つの検索課題が語の提示ありで，残りの 3 つの検索課題が語の提示なしということになる．そして，1 つの検索課題に対して 7 文書〜10 文書の適合文書を選択してもらった．なお，検索対象と

[2] http://openlab.ring.gr.jp/namazu/
[3] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html

してNTCIRテストコレクション1[4] (論文抄録 約33万件) を採用した．被験者に与えた6つの検索課題は，NTCIRテストコレクション1の検索課題から無作為に選んだ．

4.2 実験結果

評価実験終了後にアンケートを行なった．アンケートでは語を提示する機能が検索に役に立ったかどうかを“1．おおいに役に立った”“2．役に立った”“3．あまり役に立たなかった”“4．役に立たなかった”の4段階で評価してもらった．その結果，被験者全員が“2．役に立った”を選んだ．検索時間の方は，被験者が文書を選んだときの経過時間の平均を，語提示ありの場合となしの場合で比較した．図3に，被験者が最初の1文書，2文書，3文書 ...と選んだときの経過時間の平均を示す．

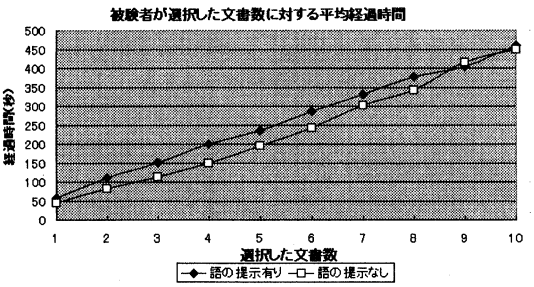

図 3: 実験結果

5 考察

図3の結果をみると，被験者が文書を選択できるまでの経過時間の平均は，提示有りの場合もなしの場合も差がなく，本支援手法は，検索時間の短縮には貢献できなかったことが分かった．今回の評価実験では，被験者に与えた検索課題として，NTCIRテストコレクション1の検索課題を採用している．そのため，各検索課題は明確に文書化されているので，被験者は検索課題の文書中の語を使って，最初から適合文書が検索できる検索式を構築できた．そのような検索式を思いついた被験者は，たとえ語の提示がない場合でも短時間で検索課題を終了していた．そのため，語の提示が検索時間短縮に有効に働かなかったと考える．

しかし，アンケート結果では，被験者が全員，本支援システムは検索の役に立ったと回答している．その理由として，最初に必ずしも適切でないキーワードで検索を行なっても，語を選択することで検索結果を絞り込むことができた，AND検索では検索結果が少なく，OR検索では多過ぎる場合に絞り込みが楽に行なえた，といった意見があった．以上の結果より，検索時間に差は現れなかったが，被験者は全員，役に立ったと回答しており，本検索支援手法は検索に有効であると考える．

[4] NTCIR(NII-NACSIS Test Collection for IR Systems)Project, http://www.rd.nacsis.ac.jp/ ntcadm/

6 むすび

本稿では，絞り込み語提示による一検索支援手法を提案し，提案した手法を実装した．そして，実装した検索システムを被験者によるアンケートで評価した．その結果，本検索支援は検索の役に立ったという回答を得た．我々は，本検索支援システムの有用性を定量的に評価しようと試み，検索時間を指標としたが，結果に差が現れなかった．そのため，本支援手法の有用性を定量的に評価できる評価実験方法を考察し，評価実験を行なうことが今後の課題である．

参考文献

[1] Jansen, M. B. J., Spink, A.,Bateman, J. & Saracevic：Real life information retireval: A study of user queries on the web, SIGIR Forum, 32(1), pp.5-17, 1998.

[2] 木谷強，高木徹，木原誠，関根道隆：フルテキストと抽出キーワードを利用した情報検索，情報処理学会報告，96-NL-115，pp.129-134，1996.

[3] 徳永健伸：情報検索と言語処理，東京大学出版会，1999.

[4] 河野浩之，長谷川利治：WWWデータ資源検索におけるデータマイニング手法，情報処理学会報告，96-DBS-108，pp.33-40，1996.

[5] 神林隆，清水奨，佐藤進也，Paul Francis：インターネット情報探索に適したキーワード抽出，情報処理学会報告，97-NL-118，pp.79-84，1997.

[6] 砂山渡，大澤幸生，谷内田正彦：ユーザの興味の構造を用いて関連検索キーを提示する検索支援インターフェイス，人工知能学会誌，Vol.15，No.6，pp.1117-1124，2000.

[7] 宮田裕次郎，古橋武，内川嘉樹：ファジィ発想推論を用いた情報検索システムの質問拡張，T.IEE Japan，Vol.119-C，pp.632-637，1999.

[8] 酒井浩之，大竹清敬，増山繁：絞り込み語の提示による検索支援の試み，言語処理学会第6回年次大会発表論文集，pp.443-446，2000.

[9] Micheline Beaulieu, Stephen Robertson, Edie Rasmussen： Evaluating Interactive Systems in TREC, JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 47(1), pp.85-94, 1996.