情報検索における絞り込み語提示による検索者支援の試み

酒井 浩之　　大竹 清敬　　増山 繁
{sakai, otake, 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

近年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全文検索がある．全文検索は有効な検索方法であるが，問題点も指摘されている [1]．我々は全文検索の問題点のうち，次に述べる2点を解決する手法を提案する．さらに，手法を実装し，有効性を確認するために評価実験を行なったので報告する．

問題点 1 検索結果が大量に出力されたとき，検索者は検索結果を適切に絞り込む語を思いつけないことがある．

問題点 2 検索者の視点がシステム側には不明である．

問題点1に関する解決手法として，検索結果の分類 [2]，絞り込み語の提示 [3] といった手法が提案されている．

分類を行なう場合は，分類の視点と検索者の視点が一致する場合は非常に有効であるが，そうでない場合は逆に検索効率を下げる原因になりうる．本研究における視点とは，複数存在する属性において着目している属性を指す．たとえば，「鯨」は生物，食品といった複数の属性を持っているが，検索者はキーワード「鯨」で検索するとき「鯨」の属性のうちのいずれかに着目して検索をしている場合が多い．

また，絞り込み語の提示には，単純な方法としてシソーラスを利用する手法が考えられる．しかし，文献 [3] の手法は，データマイニング手法を適用し，効果的な絞り込み語を提示するものである．シソーラスを用いた場合は，そのシソーラスをどのように構成するかにもよるが，全文検索システムでは提示した語が検索対象文書に含まれていないと絞り込み語として機能しない．

そこで，我々は，入力したキーワードと関連性が高く，検索結果を絞り込める語群を提示するアプローチを採る．入力したキーワードと関連がある語として，検索結果文書集合における特徴語が挙げられる．そのような特徴語抽出に関する研究は数多くある（たとえば，文献 [8] 参照）．しかし，我々は全文検索システムを前提とした問題を解決することを目的としているので，より直接的で絞り込み語の導出が容易な pseudo-relevance feedback [7] に基づいた手法を提案する．一般的な pseudo-relevance feedback は検索結果上位の文書からキーワードを抽出し，検索質問を拡張する．それに対し，我々が提案する手法の概要は検索質問の拡張を自動で行なうのではなく，そこに人間を介して語を選択させるものである．この手法を用いた場合，提示する語は検索対象文書から抽出されることになる．そのため，検索者は提示された語を選び，その語と元のキーワードとで AND 検索することによって検索結果をさらに絞り込める．

問題点2の具体的な例としては，キーワード「鯨」で検索を行なったとき，従来のシステムでは生物として「鯨」と入力したつもりでも，「鯨食文化」といった食品としての「鯨」の文書も検索されてしまう．この問題の解決手法として関連フィードバック法 [4][5] が提案されている．しかし，この手法は検索者に文書の適合性の判定を要求するので，検索結果文書数が多い場合は検索者の負担が大きくなるという問題がある．そこで，我々は，あるキーワードに対して対象文書群中で可能な視点を網羅した絞り込み語群を検索者に示すことで解決を計る．たとえば，キーワード「鯨」に対しては，「捕鯨」，「鯨食」，「イメージ」というような異なる視点で「鯨」を表現している絞り込み語群を検索者に提示する．関連フィードバック法と比較すると，提示された語から検索者の視点に合致した語を選択するため検索者の負担が軽減される．

以下に我々の手法とその実験結果について述べる．

2 絞り込み語群の導出

本手法の絞り込み語群の導出アルゴリズムは，以下の2つの仮定に基づいている．

仮定 1 検索質問に対する該当文書の集合中に多数出現する語は，入力したキーワードとなんらかの関連がある．

仮定 2 絞り込みに有効な語は，該当文書集合中に分散している．

仮定2は，頻度が大きいだけでは絞り込み語として適当ではない場合があるので導入している．たとえばキーワード「パソコン」で検索した場合に，検索結果として数多くの該当文書が得られ，この中に「パソコンで堤灯を設計する」という文書があり，この文書には「堤灯」という語が数多く含まれているとする．このとき「堤灯」と「パソコン」との間に（強い）関連があるとはいえない．つまり，このような語は該当文書集合が大きいにもかかわらず，文書を絞り込みすぎてしまうといえる．よって，絞り込み語は該当文書集合中に分散している語を利用し，この「堤灯」のような絞り込み語として不適切な語を除外する．以下に手法を示す．

Step 1 検索者が入力したキーワード群で1回，検索を行なう．

Step 2 検索結果の上位 n 文書（本実験では $n = 20$）と，文書表題にキーワードが含まれている文書群を該当文書集合 (S) とする．

Step 3 該当文書集合をそれぞれ形態素解析し，複合名詞と，カタカナ，英字表記の語，地名，組織名を抽出し，それを語とする．

Step 4 抽出した語に対して関数 $W(w,s)$ を計算する．（絞り込み語を判別するため．）ここで，

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

$tf(w,s)$: 文書 s における語 w の頻度，
$df(w)$: 該当文書集合中で語 w を含む文書数，
$dt(w)$: 該当文書集合における語 w の頻度，

である．（仮定1から最初の2項が導かれ，仮定2に基づき，第3項を導入した．）

Step 5 $\max_{s \in S} W(w,s)$ を，語 w の指標とする．

Step 6 （カタカナ，英字表記語と複合名詞のみを対象とした文字種による重みづけを行なう．）

Step 6.1 該当文書集合中のカタカナ，英字表記語出現頻度と複合名詞出現頻度を求める．

Step 6.2 もし，カタカナ，英字表記語の方が出現頻度が高ければ，カタカナ，英字表記の語の指標に

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

を乗算する．

Step 6.3 もし，複合名詞の方が出現頻度が高ければ，複合名詞の語の指標に

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

を乗算する．

Step 7 計算結果の指標値が高い順に語を提示する．本実験では提示数は20とする．□

2.1 実験 1

上記の手法を実装して実験を行なった．検索対象として日経新聞93年1月から6月までの全記事約10万件を用いた．なお，本実験では全文検索エンジンとしてNamazu[1]，形態素解析器としてJUMAN[2] Version 3.5を使用した．キーワードとして「鯨」,「パソコン」をそれぞれ入力した時の結果を表1に示す．表1に示す結果から各キーワードに対して関連性が高い語を提示できたと判断される．

表 1: キーワードによって提示された絞り込み語

鯨	パソコン
クジラ	ウインドウズ
調査捕鯨	ＬＡＮ
商業捕鯨	マルチメディア
南氷洋	インテル
反捕鯨国	ＩＢＭ
大阪	ＣＤ
太地町	ネットワーク
メニュー	ＲＯＭ
ミンククジラ	ＮＥＣ
捕鯨国	ＤＲＡＭ
ＩＷＣ	ノート
…	…

3 様々な視点の絞り込み語の導出

検索者の視点と合致する絞り込み語が提示されない限り有効な支援とはならない．したがって，対象文書集

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

合中で可能な様々な視点を網羅した絞り込み語群を提示する必要がある．そのためには，提示された絞り込み語を全て使用しても検索されない文書（以下，これをロストファイルと定義する）が存在しなければよい．まずロストファイルの存在に関して調査を行なった．

絞り込み語提示数の変化によるロストファイル数と検索結果文書数の割合，すなわち

$$\frac{\text{ロストファイルの数}}{\text{検索結果文書数}} \times 100(\%)$$

を調べた．図1に調査結果を示す．なお，調査のための環境は実験1と同等である．結果から，指標値の高い上位20語の絞り込み語を用いた場合は，全検索結果文書の平均22%がロストファイルとなる．また，絞り込み語を100まで提示してもロストファイル数は0にならないことが分かった．

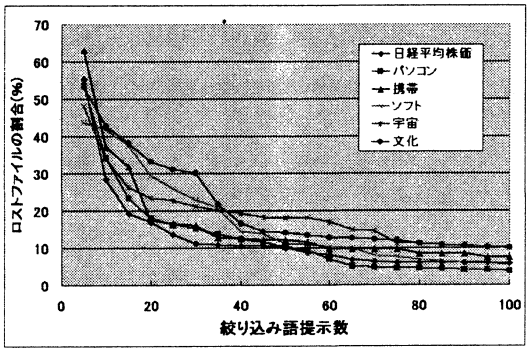

図 1: 絞り込み語提示数に伴うロストファイルの割合の変化

3.1 ロストファイルの削減手法

ロストファイルを減少させるためには絞り込み語提示数を増やすだけでは不十分である．そこでロストファイル数を効率よく削減し，様々な視点を網羅した絞り込み語群を導出するために以下の手法を導入する．

Step 1 一回，キーワードに対する絞り込み語群を提示する．

Step 2 キーワードと絞り込み語を個々に AND 検索することによって，絞り込み語を使用しても検索されない文書群 (ロストファイル群) を求める．

Step 3 ここで，ロストファイル群のうち，全文検索エンジンの出力した文書順位が上位 m まで（本実験では $m = 20$）を該当文書集合とする．

Step 4 新たな該当文書集合から絞り込み語群を生成し提示する．

Step 5 ロストファイル数が0になるか，絞り込み語群を生成できなくなるまでロストファイル群が減少したら処理を終了．そうでない場合は Step 2 へ戻る．□

3.2 実験 2

上記の手法を実装して実験を行なった．キーワードは「鯨」である．なお，実験環境は実験1と同一である．実験結果を表2に示す．

表 2: 様々な視点による絞り込み語の提示

最初の絞り込み語 (表1の「鯨」と同じ)	ロストファイル群から導出した絞り込み語
クジラ	米映画
調査捕鯨	トップ
商業捕鯨	東京
南氷洋	テーマ
反捕鯨国	イメージ
大阪	イラスト
太地町	モチーフ
メニュー	
ミンククジラ	
…	

3.3 考察

実験結果を検討すると，最初の絞り込み語群は捕鯨に関するものが多い．そのロストファイル群から導出した絞り込み語は「米映画」,「イラスト」といった語であった．これらの視点に対応する文書は最初の絞り込み語群では網羅されていなかった．しかし，ロストファイル群から導出した絞り込み語群で網羅することができた．これで,「鯨」と入力した検索者の視点が「捕鯨」である場合にも「イラスト」に関することである場合にも対応できた．

4 評価実験

ここまで述べてきた2つの手法によって1節で述べた問題点は，おおむね解決できたと考える．しかし，これは我々の主観的な判断に過ぎないため，第三者に使用してもらった客観的な評価が必要である．そこで，本手法による検索支援の有効性を検証する評価実験を行なった．

20 才から 26 才までの学生 9 名に「企業の合併事例を 7 例検索せよ」という検索課題に取り組んでもらい検索支援の評価を行なった．絞り込み語群提示機能を使う被験者と使わない被験者に分けて，課題完了までの時間によって評価した．表 3 に評価結果を示す．

表 3: 評価結果

絞り込み語提示	無し	有り
被験者数 (人)	4	5
課題完了までの平均時間 (分)	11.3	6.4
課題完了までの平均検索回数	2.3	6.0

4.1 考察

評価結果より，平均検索回数が 2.3 から 6.0 に増えているにもかかわらず，課題完了までの平均時間が 11.25 分から 6.4 分まで短縮され，本手法による支援が有効であると考えられる．検索回数の増加は，絞り込み語の使用によるものであると判断される．検索回数が増加したにもかかわらず時間が短縮されたのは，絞り込み語が検索質問に対して妥当であり，支援システムが効果的に機能したからであると考える．

しかしながら，評価実験の課題は「検索者は検索要求についての知識をあまり持っておらず，しかしながら明確な目的を持って検索を行なう場合」を想定しており，我々の支援手法が効果的に機能するであろうことも予想できる．他にも様々な検索の状況が存在し，そのような検索の状況において，本支援システムが検索に与える影響について実験・評価する必要がある．また，1 節で述べたように様々な特徴語抽出の手法を用いた絞り込み語の導出も可能である．このような手法，たとえば Carpineto らの手法 [9] との比較・検討も必要となるだろう．

本研究で提案した手法は，全文検索を前提とした検索支援手法であるが，手法自体は全文検索手法からは独立している．そのため，検索結果文書を順序づけて出力する様々な全文検索支援システムに適用させることができる．またこの際，絞り込み語の導出の過程から提示される絞り込み語が，検索結果文書の順位づけによって異なることが考えられる．しかし，可能なかぎりの視点を網羅する枠組によって，絞り込み語が多少異なっても，検索者に与える影響は小さいと予想できる．

5 むすび

本研究では，検索支援のための絞り込み語群導出手法を提案し，その有効性を確かめた．しかし，様々な検索状況における実験・評価がさらに必要であり，WWW などへの応用を考慮すると絞り込み語導出の高速化も重要な課題である．今後，これらの課題に取り組んでいく予定である．

参考文献

[1] 馬場肇: 日本語全文検索システムの構築と活用，ソフトバンク株式会社 (1998).

[2] 有田英一，安井照昌，津高新一郎: 単語集合の自動構造化機能を持つ「情報散策」方式，情報処理学会研究報告 NL-108，pp.69–74(1995).

[3] 河野浩之: 問答：検索支援システム構築技術としてのデータマイニング，画像電子学会第 9 回メディア統合技術研究会 (1997).

[4] 徳永健伸: 情報検索と言語処理，東京大学出版会 (1999).

[5] 長尾真，黒橋禎夫，佐藤理史，池原悟，中野洋: 言語情報処理，岩波書店 (1998).

[7] D. Harman: Overview of the Sixth Text REtrieval Conference(TREC-6), Proceedings of the TREC-6, pp.1–24, (1997).

[8] 久光徹，丹羽芳樹，辻井潤一: タームの representativeness を測る，情報処理学会研究報告 NL-133, pp.115–122, (1999).

[9] C. Carpineto, R. De Mori and G. Romano: Informative term selection for automatic query expansion, Proceedings of the TREC-7, pp.363–370, (1998).