社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－DBS－133 (12)
2004－ＦＩ－ 75 (12)
2004／5／14

特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案

戸田 浩之　長浜 光俊　片岡 良治
日本電信電話株式会社, NTT サイバーソリューション研究所
toda.hiroyuki@lab.ntt.co.jp

概要

コンピュータネットワークの発展により，アクセス可能な情報量は増大し，全文検索システムを代表とするナビゲーションシステムの必要性が高まっている．しかし，いわゆる全文検索システムでは，しばしば検索結果が膨大な量となり，その中からユーザが本当に望む情報を探すコストが少なくないことが指摘されている．本研究では，全文検索システムを用いた文書の検索において，検索結果から動的に「特徴的な固有表現」を抽出し，これを検索結果に対するインデクスとしてユーザに提示する，ラベル指向のナビゲーション手法の提案を行なう．これにより，検索結果の内容を概観可能とし，ユーザの検索効率を向上する．本提案手法について，毎日新聞 94，95 年の記事およびそれらの記事に対して IREX で作成されたトピックと正解を利用した評価を行い，従来手法と比較して非常に高い適合性を得たことを報告する．

A Label-Oriented Navigation Method using Informative Named Entities

Hiroyuki Toda, Mitsutoshi Nagahama, Ryoji Kataoka
NTT Cyber Solution Laboratories, NTT Corporation
toda.hiroyuki@lab.ntt.co.jp

abstract

Due to the growth of the Internet, the amount of information accessible has exploded. Retrieval systems that can efficiently locate the desired information are thus essential. Unfortunately, such systems often output too much useless information. Users are forced to manually prune the result list in order to get the documents desired. This is not efficient. The retrieval technique proposed herein uses automatic extraction of informative named entities. This method dinamically construct index for the retrieval result and it makes users easily prune the result list and rapidly get the desired document. We construct a prototype system based on this proposition and find that it yields much higher precision in terms of search results.

1. はじめに

コンピュータネットワークの発展により，アクセス可能な情報の量は増大し，効率的な情報検索手段の必要性が高まっている．

このようなユーザの要求を満たすため，一般的にランキング付きの全文検索を用いた検索システムが利用されている．検索方法，ランキング方法には様々な形態が利用されているが，ユーザは検索キーワードを入力し，ランキング付きの検索結果リストを取得，そのリスト中から所望の文書を選別するという手順を経る．

キーワードを用いたシステムの性質上，検索要求が曖昧となり，ユーザの求めるものと異なる文書が検索結果に含まれる事は不可避であるが，検索結果リストからの文書の選別は，ユーザに対して大きな

負担になっている.

我々は，従来システムの問題を，検索システムを利用するユーザの立場に基づき以下のように考えた.

- ある程度検索目的が明確な場合でも，適切な検索条件が作成できない.
- 検索目的があいまいな場合に，検索結果の概要が把握しにくい.

前者の例として「北朝鮮の核開発に対する国連の取り組みについて知りたい」という検索要求が挙げられる．一般的には「北朝鮮 核開発 国連」等の複数のキーワードで検索を行なう．もちろんこの検索条件で，ある程度の絞り込みは可能だが，文書の多義性等様々な要因により，不要な文書が検索結果に含まれる可能性がある．この場合，より効率的な絞り込みキーワードである「寧辺」や「国際原子力機関」「朝鮮半島エネルギー開発機構」等を利用することで，より多くの不要な文書を排除できるが，よほど分野について精通していない限り，このような検索キーワードを利用することは困難である.

また，後者の例としては，調査などの目的で，比較的緩い検索条件で検索を行なう場合が考えられる．例えば，「国連の活動内容について知りたい」という検索要求が挙げられる．この場合ユーザは，例えば「国連 活動」等の検索条件を入力し，検索結果を参照しながら，目的の情報を収集する．しかし，現状のリスト型の検索結果の提示では，個々の文書を参照する必要があり，非常にコストのかかる作業となる．この場合に，「国連安保理」や「国連平和維持軍」等の具体的な機関の名前や，「ゴラン高原」，「イラク」，「北朝鮮」等の具体的な地名などを含む検索結果があることが提示できれば，不要な文書へのアクセス頻度を低減させ，効率的な情報アクセスが可能となる.

上記問題を解決するため，本研究では，全文検索システムを用いた文書の検索において，検索結果から動的に「特徴的な固有表現」を抽出し，これを検索結果に対するインデクスとしてユーザに提示する，ラベル指向のナビゲーション手法の提案を行なう．これにより，検索結果の内容を概観可能とし，ユーザの検索効率を向上する.

以下，2章では関連研究に付いて示し，3章で本研究のアプローチおよび提案手法について述べる．4章で評価および考察を行い，5章でまとめる.

2. 関連研究

検索結果の概観性の改善，絞り込み検索の効率化を行なう手法に関する研究は多く行なわれている．主なアプローチとして，文書指向とラベル指向の2つのアプローチがある.

文書指向のアプローチとして，クラスタリングを用いた手法が幅広く研究されている．検索結果の文書から文書ベクトルなどの特徴表現を抽出，その特徴表現間の類似度を用いて，検索結果をクラスタリングする．その後，クラスタリングされた文書群から代表的なタームやセンテンス等をラベルとして取得し，各クラスタとともに提示する．以下の2つの研究はこのアプローチを取っている.

Cutting らは，効率的に検索結果等の大量文書を参照するために，Scatter/gather[1][2] を提案している．本手法では，ベクトルの余弦尺度を元に，高速にクラスタリングを行なうことのできる Fractionation 法 [1] を用いて，文書集合をクラスタリングして提示，ユーザが選択したクラスタの文書を対象に，再クラスタリングを行なう．この繰り返しによって，所望の文書に到達することを支援するという手法である.

また，Leuski[3] は，文書のベクトル間類似度を元に，凝集法によって検索結果のクラスタリングを行う手法について検討を行なっている．プロトタイプシステムを用いた評価では，検索結果のリストを提示するだけのシステムと比較し，ユーザが所望の文書に到達するまでに閲覧する文書数が低減しているとの報告を行なっている.

一方，ラベル指向のアプローチは，検索結果内のタームの出現状況から特徴的なタームをラベルとして抽出，検索結果とともに提示する手法である.

Sakai ら [4] は，特徴的なターム選択の基準として，tf-idf[5] に「絞り込み語に有効な語は検索結果中に分散している」との仮定に基づく値を考慮した指標を提案している．また，成田ら [6] は，tf-idf に語の出現する文書のランキングを加味した指標を特徴的なターム選択の基準として提案している.

しかし経験上，tf-idf 法によるターム抽出では，頻度の寄与が大きく，一般的すぎる不要語が排除できない [7] と言われている．上記で提案されている tf-idf の拡張手段についても，その頻度依存性の問題への抜本的な解決手段とはなっていない.

ラベル指向アプローチの関連手法として，久光ら

[7] の研究が挙げられる．これは，着目したタームの特徴を，共起するタームの分布で表現，文書コレクション中での一般的なタームの分布との類似度を元にタームの特徴度をはかる分野代表性 (representativeness) を用いて，検索を支援するタームの抽出について示している．この手法は，従来の tf-idf での問題点である頻度依存性の問題を解消している．ただしこの手法は，筆者が応用例として STOP 語リストの生成を挙げているように，検索結果のような動的な文書集合に対するものではなく，固定的な文書集合からの重要語抽出手法であり，提示するターム数が限定されるようなタスクを前提とした手法ではない．

また，技術の詳細は明らかにされていないが，上記の tf-idf を用いた技術に近いものとして，Vivisimo[1] や，mooter[2] が，Web 上の検索エンジンとして実用化されている．

3. 提案手法

3.1 アプローチ

文書指向のアプローチは，排他的なクラスタリングを行なう手法であり，クラスタリングの精度に依存する手法である．一般にクラスタの個数や，クラスタ生成のための類似度の閾値を調整し，クラスタリングの制御を行なうが，必ずしも一意に決定できる値とは限らず，ユーザの意図にあったクラスタリングを行なうことは困難である．また，クラスタ決定後に，クラスタを説明するラベルを抽出する点においても，クラスタの精度に大きく依存する等様々な問題がある．

一方のラベル指向のアプローチでは，文書中から特徴的なタームを抽出し、提示することで文書集合の一部へのアクセスを可能とする手法であり，非排他的なクラスタリングを行なう手法と考えることもできる．

検索を支援するという目的を考えた場合，提示するラベルを明確に制御でき，ユーザによって文書の着目点が違うような場合にも対応することが必要となることから，本手法ではラベル指向のアプローチを採用する．以下，本稿では，検索結果から抽出した個々の特徴的なタームをラベルと呼び，それらをまとめたものをインデクスと呼ぶこととする．

[1] http://www.vivisimo.com/
[2] http://www.mooter.com/moot

3.2 課題と提案手法

従来研究にも示す通り，ラベル指向のアプローチにも課題が存在する．そこで，実際に利用可能な Vivisimo，mooter を利用した上で課題を明らかにし，それに対する個々の提案手法について示す．

- **課題** *1*：ラベルの候補となるタームの抽出品質の悪さ
  従来の単語や，名詞句を利用したシステムでは，タームの区切り間違い等により必ずしも重要なタームがラベルの候補として抽出されていない．
- **課題** *2*：絞りこみに効果的でないラベルの選択/提示
  検索条件の絞り込みに効果的でないタームが，ラベルとして提示される．2 章で示した tf-idf の精度に関する問題とも関連する．
- **課題** *3*：ラベルの単なる羅列による検索結果の概観性の低さ
  ラベルが単純に羅列されているだけでは，絞り込み語を探す場合にも探しにくく，検索結果の全体像も掴みにくい．

上記課題 1 に関して，我々は固有表現抽出技術の利用を行なう事で対応する．本技術を用いる事により，テキスト中で検索語の候補になりうるタームの抽出を高精度で行なうことが可能となり，同時に，抽出したそれぞれのタームに対して「組織」「人物」等のクラス付けが行える．

また，2 点目の課題に関しては，ラベルとして有効な特性の検討を行なった上で新たな「ラベル選択基準」を定義し，その基準によるスコアに基づきラベルとなるべきタームの選択を行なう．詳細は 3.2.1 に示す．

3 点目の課題に関しては，インデクスを提示する際に，固有表現抽出によって得られたクラス毎にラベルをまとめて提示することを提案する．これにより，インデクス中に類似した種類のラベルを並べることが可能となり，検索結果の概観性の向上を行なうことができる．また，効率的な絞り込みを可能としつつ，概観性を確保するため，クラスに対する優先度を「クラス優先度基準」として定義し，クラスの提示順序決定の指標とする．詳細は 3.2.2 に示す．

以下では，上記で示した「ラベル選択基準」および「クラス優先度基準」について示す．

3.2.1 ラベル選択基準

従来手法の取り組みにもある様に，絞り込みに効果的なラベルとは，検索結果における重要タームであると言うことができ，タームの出現頻度や，tf-idfを用いた基準が利用されている．

ターム i の重要度を I_i とした場合，一般的な基準は以下の式で表される．

- 出現頻度を用いる方法
  最も基本的な手法として，検索結果中でのタームの出現頻度を用いる方法である．

$$I_i^{freq} = df_{R,i}$$

  ここで，$df_{R,i}$ は，検索結果 R 中での，ターム i が出現する文書の頻度を示す．

- tf-idf を用いる方法
  文書集合全体では希少であり，かつ検索結果中で頻度のあるタームが重要であるという考え方に基づく手法である．

$$I_i^{tf-idf} = df_{R,i} \times \log(\frac{|D|}{df_{D,i}})$$

  上式では，文書集合からの重要語の算出ということから，タームの頻度ではなく，タームの出現する文書の数を tf 値として用いているが，通常と同様に文書集合中でのタームの出現頻度を利用する場合もある．また，D は検索システムに登録されている文書の集合を表し，$df_{D,i}$ は，文書集合 D 中での，ターム i を含む文書数を示す．

一方，我々は，検索条件こそが「ユーザの検索要求を示すもの」であり，効果的なナビゲーションには，この「検索条件と関連性の高いターム」をラベルとして提示することが重要であると考えている．

この「検索条件と関連性の高いターム」は，一見，「検索結果と関連性の高いターム」と同義であると考えることもできる．しかし，この言い替えによる「検索結果と関連性が高い」という条件は「検索条件と関連性が高い」という条件に対して，必要条件ではあるが，十分な条件を満たしていない．

つまり，検索条件は，文書集合全体から検索結果を切り出すための情報であり，「検索条件と関連性が高い」と言う場合には，単に検索結果と関連性が高いタームを評価するだけではなく，文書集合を限定することによって文書集合との関連性が高まるタームを評価しなければならない．

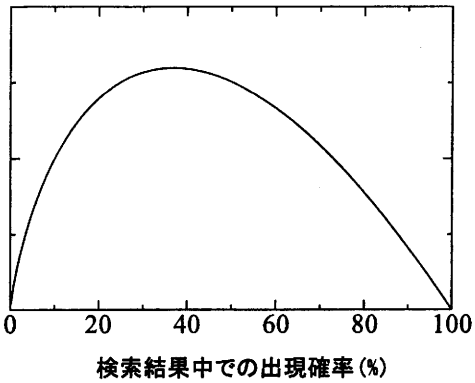

図 1　関数の形状

この考えを元に，我々は次に示す基準を提案する．

$$I_i^{proposition} = df_{R,i} \times \log(\frac{|R|}{df_{R,i}}) \times \frac{df_{R,i}/|R|}{df_{D,i}/|D|}$$

上式の前の 2 つの項では，「検索結果と関連性が高い」という指標を表現するため，検索結果中でタームの頻度する文書の頻度を元にした値を採用している．ここで適用した式は，図 1 に示すように，検索結果数に対して 30〜40%程度の出現確率で値が最大となるような関数である．これは，検索結果における頻度が少な過ぎず，多過ぎないタームが絞り込みに有益であるという直感的な感覚と一致する．

また，後の項では，「文書集合を限定することによって文書集合との関連性が高まるタームの評価」という観点から，文書集合全体および検索結果集合中のタームの出現確率の比を基準として用いた．

3.2.2 クラス優先度基準

クラス優先度基準とは，効率的な絞り込み機能と概観性を持ったクラスを優先的に提示するための基準である．

この基準の評価は，どのラベルをいくつ選択するかによって変わるが，ラベルの選択については，上記で示した基準を用いることとする．また，各クラスが保持するラベルの数に付いては，ユーザが容易に全体を把握できる数ということで，今回は 10〜20 個程度の数を想定している．

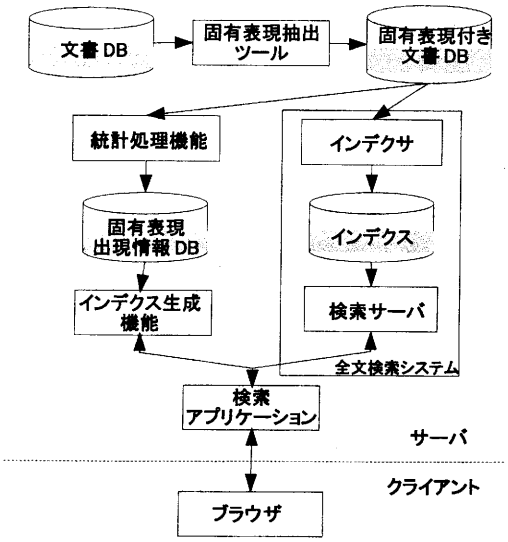

図 2　システム概要

基準としては，さまざまなものが考えられるが，仲川ら [8] は，複数のカテゴリを切替えて検索結果をカテゴライジングする手法の提案において，以下の基準を個々のカテゴリ (本稿で言うところのクラス) の有効性基準として挙げている．

- **分類の明確さ**：各ラベルを通してアクセス出来る文書集合間の差が明確であること
- **分類の均一さ**：各ラベルを通じてアクセスできる文書数のばらつきが少ないこと

上記の基準は，全ての文書を排他的に分類する場合について示されているが，我々の手法は，非排他的なアプローチであり，かつ全ての文書がいずれかの分類に属するという保証が無いためそのまま利用することはできない．そこで，上記の「分類の明確さ」の基準を「各ラベルを通してアクセス出来る文書集合間の重複が少ないこと」と読みかえ，また，上記に加えて，以下の基準を加える．

- **分類の網羅性**：インデクスを通して，アクセスできる文書が多いこと

これら個々の基準を用いた場合および組み合わせた場合について評価を行ない，有効なクラスを優先的に提示する手法に付いて検討を行なう．

これらの基準は表 1 に示す式にて定義できる．ここで，C_j をクラス j に含まれるラベルの集合，p_j をクラス j の優先度の値，D_j を検索結果中でクラス C_j のいずれかのラベルに関連付けられている文書の集合，$D_{j,i}$ を検索結果中でクラス j のラベル i に関連付けられる文書の集合を示す．

「分類の明確さ」については，ラベルに関連付けられる文書ののべ数と，異なり数の比率を元に，「分類の均一さ」については，平均エントロピーにより算出している．また，「分類の網羅性」に関しては，検索結果数とラベルに関連付けられている文書の異なり数の比率を用いている．

表 1　クラス優先度基準の算出法

分類の明確さ	$p_j^1 = \|D_j\| / \sum_{i \in C_j} (\|D_{j,i}\|)$
分類の均一さ	$p_j^2 = \sum_{i \in C_j} (-\frac{\|D_{j,i}\|}{\|R\|} \times \log(\frac{\|D_{j,i}\|}{\|R\|}))$
分類の網羅性	$p_j^3 = \|D_j\| / \|R\|$

4. 評価

4.1 プロトタイプシステム

以上で説明した手法に基づいてプロトタイプシステムの実装を行った。全文検索に LISTA[9] を，また，固有表現抽出を行なうにあたり，磯崎 [10] の手法によるツールを利用した．システムは Web サーバ上に構築し，ブラウザを介してアクセスする形態となっている．

図 2 にシステムの概要を示す。前処理として，固有表現付きの文書を全文検索システムのインデクサで処理すると同時に，統計処理機能でターム (固有表現) の出現情報を生成する．実際の検索の際には，Web サーバ中に構築されたアプリケーションが検索サーバおよびインデクス生成機能にアクセスし，検索結果の提示に必要な「検索結果」とラベルから構成される「インデクス」を取得する．これらのプロトタイプシステムは，Java[3] 言語を用いて実装され，Web アプリケーションサーバ Tomcat[4] 上で動作している．

また，図 3 にユーザインタフェースを示す．ユーザインタフェースでは，ユーザが入力したキーワード

[3] http://java.sun.com/
[4] http://jakarta.apache.org/tomcat/

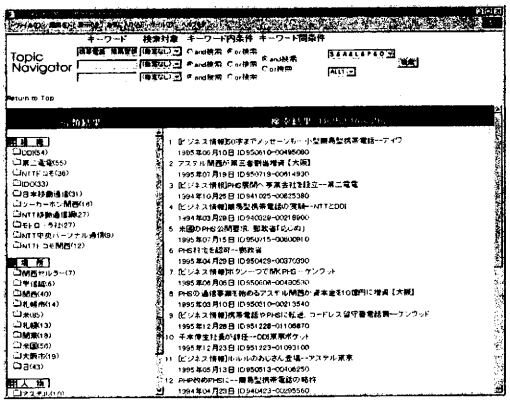

図 3　ユーザインタフェース

を元に，検索結果を表示した状態を示しており，右側に通常の検索結果リストを提示，左側に複数のラベルから構成されるインデクスを提示している．ユーザは，通常の検索結果から所望の文書を選択することに加えて，インデクスを参照することにより，検索結果の概観を行なったり，インデクス中に目的のラベルが存在する場合には，それを選択することで容易に検索結果の絞り込みを行うことが出来る．

4.2　評価手法

本報告では，3.2.1 に示した「ラベル選択基準」について評価を行なう．本研究の目的は，ユーザに効率的なナビゲーションを提供する技術の確立である．それに基づきラベル選択にあたっては，ユーザがそのラベルを選択することによって，効率的に目的の文書に到達することを目指している．

そこで「ラベル選択基準」の評価では，提示したラベルを用いた絞り込み検索の適合率を基準として採用する．評価のフローは以下の通り．

1. 初期検索条件により検索実行．検索結果リストと，ラベル群を取得．
2. 初期検索結果の上位に含まれる正解を記録．
3. 提示されたラベル群のうち条件に適合するラベルを用いて，絞り込み検索実行
4. 得られた検索結果について適合度を算出．
5. 条件に適合するラベルが無くなるまで，3，4 を繰り返す．

評価を行なうにあたっては，毎日新聞 94 年および 95 年の記事と，IREX の IR タスクにて規定されている 30 の検索トピックおよびその正解を利用した．初期検索条件には，検索トピックの DESCRIPTION 部分を形態素解析し，出力された形態素から品詞情報および STOP 語リストで不要語を除去した後 “or” で結合したものを用いた．

適合率の算出にあたっては，IREX[11] にて規定された正解集合 (A および B) を利用した．ただし，フローにもあるように初期検索条件で容易に取得できる正解文書を特定し，正解から除去して適合率の計算を行なっている．

また，絞り込み検索に利用するラベルの条件は以下の 2 条件とした．

- **評価条件** *1*：出力されたラベル全て
- **評価条件** *2*：絞り込みに有効なラベルのみ

前者では，出力されたラベルを選択した場合の平均適合率を得ることができる．

後者の「絞り込みに有効なラベル」は有効タームリスト[5] と一致させることで特定した．これによって，ユーザが選択するであろうラベルを利用した場合の平均適合率を得ることができる．

4.3　評価結果および考察

上記評価手法にて，今回の提案手法を，3.2.1 にて示した従来手法と比較した．今回の評価では，ラベルのクラスは考慮せず，それぞれの評価基準で優先的に提示されたラベル 10 件を対象とし，それを用いた絞り込み検索の結果から最大上位 10 件の文書を取得し適合率の評価を行っている．インデクス中のラベルを抽出するために解析した文書は，初期検索の結果得られた文書のランキング上位 500 件である．

結果を図 4 および 5 に示す．それぞれのグラフより，提案手法は tf-idf 法と比較して，評価条件 1 で約 15%，評価条件 2 で約 30%の適合率の向上を確認することができる．相対的には tf-idf の約 2 倍の適合率

[5] 有効タームリストは，被験者に，IREX で規定されたトピックおよびその正解文書の両方を提示，「それぞれのトピックの検索において検索条件として利用可能なターム」として選択してもらったタームによって構成される．

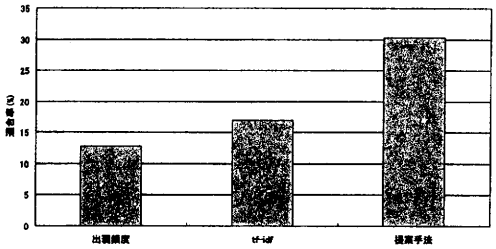

図 4　評価条件 1 の結果

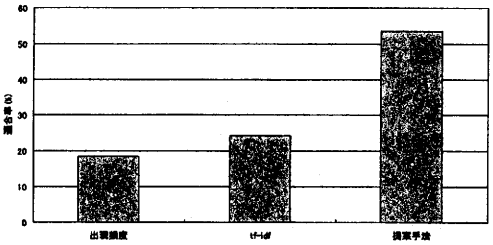

図 5　評価条件 2 の結果

を示している．特に，絞り込みに有効なラベルのみを利用した場合の適合率は 50%を越えており，絞り込みを行なった検索結果の 2 件に 1 件は正解文書であるということを示している．

一方，tf-idf 法の精度は，頻度だけの手法を多少上回っているが，5%程度の向上に留まり，あまり差は無かった．

また，実際のラベル抽出結果を図 6 および 7 に示す．タームの出現頻度によるラベルの生成が容易だったと思われる「携帯電話、簡易型携帯電話のサービス」(トピック 1) と困難だったと思われる「社名変更」(トピック 2) に関しての結果である．

両方のトピックを通して，tf-idf の基準を用いた場合には，頻度を元に抽出したものと同傾向のラベルを選択している事がわかる．事実この二つのトピックに関して言えば，40 件中 34 件が同じラベルであった．サンプル的な評価ではあるが，2 章で述べたように tf-idf を用いた手法が頻度による影響を受けやすいということを裏付ける結果となった．

また，個々のラベルについて見てみると，トピック 1 のようにラベル抽出が容易な場合は，多少ふさ

「携帯電話，簡易型携帯電話のサービス」に関して		
頻度	tf-idf	提案手法
郵政省	郵政省	DDI
NTT	NTT	第二電電
米	DDI	NTT ドコモ
日本	第二電電	IDO
米国	NTT ドコモ	日本移動通信
第二電電	IDO	ツーカーホン関西
DDI	日本移動通信	NTT 移動通信網
日	NTT 移動通信網	モトローラ社
東京	モトローラ社	アステル
関西	米	NTT 中央パーソナル通信
NTT ドコモ	関西	NTT ドコモ関西
IDO	日本電信電話	DDI 東京ポケット
日本移動通信	日本テレコム	アステル関西
阪神	ツーカーホン関西	アステル東京
モトローラ社	米国	NTT パーソナル
NTT 移動通信網	シャープ	NTT 関西パーソナル通信
日本電信電話	日本	NTT パーソナル通信
日本テレコム	NTT ドコモ関西	モ社
大阪市	日	東京デジタルホン
シャープ	DDI 東京ポケット	NTT 関西移動通信

図 6　ラベルの抽出結果

「社名変更」に関して		
頻度	tf-idf	提案手法
日本	日本	ランサー
東京	自民党	辻本一義
米	東京	ミラージュ
自民党	米	太東興産
阪神	阪神	ニッサン
大阪	衆院	ベネッセコーポレーション
米国	社会党	日本サッカー協会
社会党	大阪	福武書店
衆院	日本サッカー協会	辻本
ワシントン	米国	改革連合
神戸市	東証	ホンダ
参院	ホンダ	アシックス
中国	トヨタ自動車	河野正
外	神戸市	大証
東証	ワシントン	新東京国際空港公団
村山富市	参院	KBS 京都
トヨタ自動車	名古屋	近畿放送
京都	大証	護憲民主連合
名古屋	外	グランディー
東京都	京都	新緑風会

図 7　ラベルの抽出結果

わしくないラベルが混在する程度であるが，トピック 2 のように頻度でのラベル取得が難しい場合，ほとんど有効なラベルを抽出できていない．

一方，提案手法を用いた場合，トピック 1 では，正式な表現と略語の両方が選択されていること以外，ほぼ目的にあったラベルが選択されている．また，トピック 2 でも半分のラベルは，検索目的と関連性の深いものが選択されている．

以上の定量的および定性的な評価より，提案したラベル選択基準は，非常に有効な基準であると言うことができる．

5. まとめ

本稿では，全文検索システムを用いた文書の検索において，検索結果から動的に「特徴的な固有表現」を抽出し，これを検索結果に対するインデクスとしてユーザに提示することで，検索結果の内容を概観可能とし，絞り込みを支援するラベル指向のナビゲーション手法の提案を行ない，それに基づくプロトタイプシステムの試作，評価を行なった．

ラベルの選択法として，検索条件との関連性を利用した新たな評価基準の利用を提案した．IREX の正解セットを用いた評価では，従来手法を大きく上回る絞り込み検索精度があることを示し，非常に有効な基準であると言うことができた．

また，クラスの優先度付けに付いては複数の評価基準を利用し，クラスに優先度付けを行なうことを提案した．

今後，クラスの優先度付けについての評価およびシステム全体としてのユーザビリティについての考察が必要である．

謝辞

本研究を進めるにあたり，固有表現表現抽出のツールを提供頂いた NTT コミュニケーション科学基礎研究所の磯崎様，多言語対応 XML 検索エンジン LISTA を提供頂いた NTT サイバーソリューション研究所の富田様に深く感謝します．

参考文献

[1] Cutting, D., Karger, D., Pedersen, J., and Tukey, J. W.: "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections" Proceedings of the 15th Annual International ACM/SIGIR Conference, 1992.

[2] Hearst, M., Karger, D. and Pedersen, J.: "Scatter/Gather as Tool for the Navigation of Retrieval Results" Working Note of the 1995 AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval, 1995.

[3] Leuski, A.: "Evaluating Document Clustering for Interactive Information Retrieval" Proceedings of CIKM 2001, 2001.

[4] Sakai, H., Ohtake, K. and Masuyama, S.: "A Retrieval Support System By Suggesting Terms to a User" Proceedings of ICCPOL 2001, 2001.

[5] Salton, G. and Yang, C. G.: "On the Specification of Term Values in Automatic Indexing" Journal of Documentation 29, 1973.

[6] 成田宏和，大田学，片山薫，石川博 "階層的クラスタリングを利用したメタサーチエンジンの提案" 情報処理学会研究報告 DBS-128-050, 2002.

[7] 久光徹，丹羽芳樹，辻井潤一: "タームの representative を測る" 情報処理学会研究報告 NL-133-16, 1999.

[8] 仲川こころ，高田喜朗，関浩之: "可変なカテゴリ構造を用いた文書検索支援方法" 情報処理学会論文誌 Vol.42, No.10, 2001.

[9] Hayashi, Y., Tomita, J. and Kikui, G.: "Searching text-rich XML documents" ACM SIGIR 2000 Workshop on XML and Information Retrieval, 2000.

[10] 磯崎秀樹，賀沢秀人: "固有表現抽出のための SVM の高速化" 情報処理学会論文誌 Vol.44, No.3, 2003.

[11] 関根聡，井佐原均: "IREX プロジェクト概要" IREX ワークショップ予稿集, 1999.