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† 中渡瀬秀一† 片岡良治†

インターネットをはじめとするコンピュータネットワークの普及にともない，アクセス可能な情報量は増大し，全文検索を代表とする検索システムの必要性はますます高まっている．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，ユーザはこの中から実際に欲しい情報を探さなければならない．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，検索結果とともに，検索結果中の主要な話題を，検索結果に対するインデクスとして提示するラベル指向ナビゲーション手法を提案する．インデクスは話題を示すラベルの集合として構成される．具体的には，文書を検索システムに登録する段階で，固有表現抽出技術を用いてラベルの候補となるタームを抽出する．その後，検索が行われる段階で，検索結果集合中で主要な話題と考えられるタームをラベルとして選択し，提示する．これを利用することで，ユーザは検索結果の内容を概観できるとともに，効率的な絞り込み検索ができる．本稿では，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，検索結果の質と使いやすさの点から提案手法の優位性を示した．

A Label-based Navigation Method Using Informatively Named Entities

Hiroyuki Toda,† Hidekazu Nakawatase† and Ryoji Kataoka†

Due to the growth of the Internet, the amount of information accessible to the public has exploded. Retrieval systems that can efficiently locate the desired information are thus essential. Unfortunately, ordinary retrieval 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 automatically extracts informatively named entities and uses them to dynamically index the retrieval result. This allows the user to easily prune the result list and get the documents desired. Since it offers dynamic indexing, our method supports searching of daily-updated contents like news articles. We implement a prototype system based on this proposition and find it yields much higher performance than the existing method.

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，全文検索を用いた検索システムが利用されている．様々な検索方法，ランキング方法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，ユーザの望まない文書が検索結果に含まれることは不可避であり，ユーザは検索結果リストから本当に欲しい文書を選別しなければならない．

従来のシステムの問題点は，以下のように考えられる．

- **問題点** *1*：適切な検索条件が作成できない．
- **問題点** *2*：検索結果の概要が把握しにくい．

問題点 1 は，Belkin [1] の指摘にもあるように，情報検索システムの古くからの問題である．分かりやすい検索要求の例として「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」があげられる．この場合「北朝鮮 核開発 国連」等の複数のキーワードで検索することが考えられる．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，キーワードが一般的な語であるほど，不要な文書が検索結果に含まれる可能性が高い．より具体的なキーワードである「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，多くの不要な文書を排除できるが，よほど分野に精通していない限り，このような具

† 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
NTT Cyber Solutions Laboratories, NTT Corporation

体的な検索キーワードを，ユーザ自ら検索条件として利用することは困難である．

また，問題点 2 の例としては，調査等の目的で，比較的緩い条件で検索をする場合が考えられる．たとえば，「国連の活動内容について知りたい」という検索要求があげられる．この場合ユーザは，「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」等の具体的な地名等を含む検索結果の存在を提示できれば，検索結果の概要を把握することが容易になり，さらにそれらの語で絞り込み検索を行うことで，ユーザは所望の情報に容易にアクセスできる．

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，検索結果とともに，検索結果中の主要な話題を，検索結果に対するインデクスとして提示するラベル指向ナビゲーション手法を提案する．インデクスは話題を示すラベルの集合として構成される．具体的には，文書を検索システムに登録する段階で，固有表現抽出技術を用いてラベルの候補となるタームを抽出する．その後，検索が行われる段階で，検索結果集合中で主要な話題と考えられるタームをラベルとして選択，提示する．これを利用することで，ユーザは検索結果の内容を概観できるとともに，効率的な絞り込み検索ができる．

本稿では，文書から抽出したキーワードや意味のある表現をタームと呼び，そのタームのうち検索結果の概観および絞り込みに有効であると考えられ，検索結果とともにユーザに提示されるものをラベルと呼ぶ．また，1 つの検索結果に対するラベルのリストをインデクスと呼ぶ．

以下，2 章では関連研究について示し，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 章で評価に利用したシステム，リソースについて述べ，5，6，7 章でそれぞれ評価およびそれに基づく考察を行い，8 章でまとめる．

2. 関 連 研 究

ユーザの検索を支援する技術として，様々な技術が提案されている．

従来提案されている手法の 1 つとして適合性フィードバックがある[2]．この手法を用いたシステムのユーザは，検索結果の上位 10〜20 件程度について，自身の検索要求に適合しているか否かの判断を行い，その判断をシステムに入力する．システムはその情報に基づき，ユーザが適合するとした文書から重要なキーワードや表現を選択，それらに重み付けを行った検索式を作成し，再検索を行う．これによって，ユーザが適合すると考える文書が優先的に提示されるという手法である．

この手法は，ユーザが直接的に検索式を変更しなくても，ユーザの意図に合う検索式を作成できるという利点があるが，ユーザは少なくとも数件の文書を参照し，適合性の判定を行わなければならないので，ユーザの負荷が大きくなるという問題点がある[3),4)]．また，適合文書から不要なタームを選択してしまう等の問題点も指摘されている[4)]．

以上より，この手法は検索精度を向上させる検索式の作成を支援するという点で検索を支援する技術であるが，本稿でもう 1 つ指摘している「検索結果の概要が把握しにくい」という問題については考慮されていない．

一方，検索結果の概観性を改善し，絞り込み検索を効率化することで検索を支援する手法として，検索結果を組織化，分類して提示する手法があげられる．その中でもカテゴリ構造等の先験的な知識を必要としない手法は，以下に示す文書指向とラベル指向の 2 つのアプローチに大別できる．

文書指向のアプローチとは，いわゆるクラスタリングを用いた手法である．この手法では，検索結果中の個々の文書を文書ベクトル[5)] 等で表現し，文書間の類似度をもとにクラスタリングする．その後，クラスタリングされた個々の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とっている．

Cutting らは，検索結果等の大量文書を効率的に参照する手段として，Scatter/Gather[6),7)] を提案している．本手法では，文書ベクトルの類似度をもとに，高速にクラスタリングできる Fractionation 法[6)] 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う．この繰返しによって，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[8)] は，文書ベクトル間の類似度をもとに，凝集法によって検索結果をクラスタリングする手法について検討している．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[9] は，ラベル選択の基準として，TF-IDF[10] 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，Ohta ら[11] は，TF-IDF に語の出現する文書のランキングおよび個々のテキスト中でのタームの出現位置を加味した指標を提案している．また，そのほかにも TF-IDF を基準として採り入れている研究は多い[12]．

しかし経験上，TF-IDF 法に基づくターム抽出では，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といわれている[13]．

ラベル指向アプローチの関連研究として，Hisamitsu ら[13] の研究があげられる．この手法では，着目した語の特徴を，共起する語の分布で表現し，文書コレクション中での一般的な語の分布との類似度をもとに語の特徴度をはかる representativeness を用いて，検索に有効な語を抽出している．これにより，TF-IDF 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 語リストの生成をあげているように，検索結果のような動的な文書集合に対する手法ではなく，固定的な文書集合からの重要語抽出手法である．検索結果を分類するラベルを生成する場合のように，抽出する語の数が限定される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☆や，mooter☆☆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に基づく手法は，一般に排他的なクラスタリングを行う手法である．クラスタの個数や，クラスタ生成のための類似度を閾値とし，これを調整することでクラスタリングの制御を行うが，それらの値は必ずしも一意に決定できる値ではなく，ユーザにとって分かりやすいクラスタリングを行うことは困難である．また，クラスタ決定後に，個々のクラスタを説明するラベルを抽出する場合にも，クラスタリングの精度が大きく影響する．

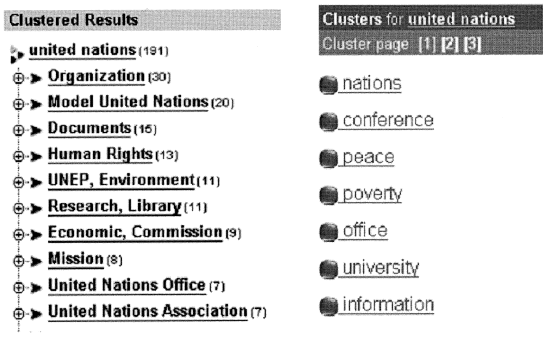

図 1 既存システムのラベル例
Fig. 1 Example of labels presented by ordinaly systems.

一方のラベル指向のアプローチに基づく手法では，検索結果の文書中から重要なタームをラベルとして複数個抽出，これらを，検索結果の概観や絞り込み検索を容易にするための情報として検索結果とともに提示する．この手法における，ラベルと文書のつながりを考えると，この手法は非排他的なクラスタリングを行う手法と考えることもできる．

検索結果を組織化，分類して提示することで検索を支援するという目的を考えると，ラベルは直接人目に触れるため，その選択はできる限り明確に制御できることが重要である．また，ユーザによって文書の着目点が違うような場合，複数の観点から文書に到達することが必要となると考えられる．

以上より，本稿ではラベル指向のアプローチを採用した手法を提案する．

3.2 ラベル指向アプローチの問題

2 章にも示すように，ラベル指向のアプローチにも問題が存在する．そこで，本節では，利用可能な Vivisimo，mooter を利用したうえで既存技術の問題点を示す．

それぞれのシステムは日本語には対応していないため，英語の検索条件を用いて利用した．

例として Vivisimo および mooter で “united nations” というキーワードで検索した場合に出力されるラベルを**図 1** に示す．Vivisimo の例を見ると，“Organization” や “Documents” 等の一般的なタームと，“Model United Nations” や “UNEP” 等の具体的な対象を示すタームがラベル中で混在していることが分かる．

このうち，一般的なタームである “Organization” および “Documents” について，それぞれ関連する検

☆ http://www.vivisimo.com/
☆☆ http://www.mooter.com/moot

索結果中の文書を参照し，ラベルとして存在している理由について考えた．

まず，“Organization” について見ると，多くの場合，様々な機関の名称中に存在し，それらの機関の名称の末尾から “Organization” というタームだけが抽出されていることが分かる．これにより，検索結果中での出現頻度が高くなりラベルとして提示されていることが考えられる．Mooter で出力される “Nations” や “Conference” 等も同様な例であると考えられる．

また，“Documents” というラベルに関連する文書には何らかの文書情報もしくは文書情報へのリンクが存在することが想定されたが，このラベルに関連付けられたサイトは国連に関係する機関のトップページが多く，想定した情報はあまり存在しない．このようなタームは単に文書に出現する頻度が高いためにラベルとして提示されているものであると考えられる．Mooter の例にある “information” や “office” 等も同様の例であると考えられる．

以上の例より，我々は以下のような点を，既存技術の問題点であると考えた．

- **問題** *1*：ラベルの候補となるタームの抽出品質が悪い場合がある
  ターム抽出時の区切り間違い等で，重要なタームが抽出されていないことがしばしばある．
- **問題** *2*：絞り込みに効果的でないラベルの選択/提示が行われている場合がある
  文書中で高い頻度で出現するが，重要でないタームがラベルとして提示されている場合がある．これは 2 章で示した TF-IDF の問題にも関連する．
- **問題** *3*：ラベルの単なる羅列では検索結果の概観性が低い場合がある
  様々な意味のラベルが提示される場合には，単純に羅列されているだけでは，検索結果の全体像がつかみにくく，絞り込み語を探す場合にも探しにくい．

3.3 提案手法

本節では前節でまとめた既存技術の問題点に対する我々の提案を示す．

上記の**問題** *1* に関して，我々は，固有表現抽出技術[14),15)] を利用して抽出できる固有表現をタームとして利用することを提案する．固有表現抽出技術では，文書中に含まれる「人名」,「組織名」,「地名」等の固有名詞を高い精度で抽出できる☆．これらの固有表現はもともと新聞記事等の文書中で「頻繁に重要になり，情報としての単位がはっきりしている」表現と定義されており[16)]，新聞記事等を対象にした場合，検索においても重要な表現であると考えた☆☆．

問題 *2* に関しては，ラベルとして有効な特性を検討したうえで新たなラベル選択基準を 3.4 節で提案する．このラベル選択基準に基づきラベルとなるべきタームを選択し，検索結果の絞り込みに有効なラベルの提示を可能とする．

問題 *3* に対しては，インデクスを提示する際に同種のラベルを分類して提示することで，概観性を向上させ，インデクスをより使いやすくできると考えた．そこで，インデクスを提示する際に，固有表現抽出技術によってタームを抽出するときに取得できるカテゴリ情報を利用して，同じカテゴリのラベルをまとめて提示する方法を提案する．これにより，検索結果の概観性を向上させるとともに，検索結果の絞り込み候補を容易に選択可能とする．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，カテゴリに対する優先度をカテゴリ優先度基準として 3.5 節で定義し，カテゴリの提示順序の指標とする．

3.4 ラベル選択基準

従来手法の取り組みにもあるように，絞り込みに効果的なラベルとして，検索結果集合内の重要タームを選択するために，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．

TF-IDF による基準は，ターム i の重要度を I_i とした場合，以下の式で表される[10)]．

$$I_i^{TF-IDF} = TF_{R,i} \times \log\left(\frac{|D|}{DF_{D,i}}\right)$$

ここで，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，ターム i を含む文書数を示す．また，R は，検索結果の文書集合を示す．また，$TF_{R,i}$ は，検索結果 R 中でターム i の出現する頻度を表す．予備実験の結果，$TF_{R,i}$ を，検索結果 R 中でターム i の出現する文書の数 $DF_{R,i}$ で置き換えても同様な結果が得られることが確認でき

☆ 固有表現抽出技術では，上記で示した固有名詞の他，「金額表現」や「割合表現」等の数値表現を抽出することが可能であるが，今回は固有名詞のみを利用する．

☆☆ 現段階の固有表現抽出技術で対応可能な文書は，新聞記事等限られたテキスト情報に限定されている．しかし，本技術については，新聞記事を対象にした固有表現のカテゴリを従来の 8 種類から 200 種類のカテゴリに拡張する取り組み[17)] や，生物情報処理におけるたんぱく質や DNA の名前等を対象とする取り組み等，より多くのカテゴリやテキスト情報を対象に，「人が興味を持つ表現」を抽出する技術として研究が進められており，今後，適用範囲が広がる技術であると考えられる[18)]．また，それらを利用することで本稿の提案手法の適用範囲も広がると考えられる．

たため，本稿の評価では $DF_{R,i}$ を用いた基準を利用した．また，$DF_{R,i}$ を RDF（Retrieved Document Frequency）と呼び，これを適応した基準を RDF-IDF と示す．

TF-IDF は，もともと文書検索における索引語の重み付けのために提案された基準であり，注目している文書（今回の場合は検索結果文書群）中での出現頻度が多いという局所的重みと，文書集合全体では希少であるという大域的重みから構成される．

以下では，2 章で指摘した TF-IDF の頻度依存性の問題および，重要ラベルの要件について考え，局所的重み，および大域的重みのそれぞれについて新たな基準を提案する．また，5 章において，以下で示すそれぞれの重みをもとにしたラベル選択基準を評価し，特性について考察を行う．

3.4.1 局所的重み

TF-IDF における局所的重みである TF は，頻度依存性の原因となっている重みである．RDF-IDF における RDF も同様の傾向がある．我々はこれに代わる重みとして以下の 2 つの重みの利用を提案する．

1 つは対数化頻度を用いる手法である．TF の過度の影響を抑えるための直接的な対処法として，索引付けにおいても利用されている[5]．以下の式に基づき算出する．

- Logarithmic Retrieved Document Frequency（Logarithmic RDF）

$$LF_i^{LRDF} = \log\left(1 + DF_{R,i}\right)$$

また，効率的に検索支援を行うことができるラベルとして，検索結果中での頻度が過度に多くもなく少なくもないタームが重要であると考えた．そこで，2 つ目の指標として，以下の式に示すように検索結果数に対して 30〜40%程度の出現頻度で値が最大となるような重みを提案する．

- Original Local Factor（Original LF）

$$LF_i^{ORG} = DF_{R,i} \times \log\left(\frac{|R|}{DF_{R,i}}\right)$$

上記に示す 2 つの局所重みの関数形状を**図 2** に示す．上記でそれぞれ提案する重みでは，従来の TF に基づく基準である RDF と比較し，高頻度のタームに対する重みを低減させている．また，後者は検索結果に対して一定以上の割合を超えると重みが減少するようになっている．

3.4.2 大域的重み

一方，大域的重みとして利用されている IDF は，文書集合中での頻度が希なタームに対して重み付けする値であり，文書集合内での頻度の逆数を対数化した値として表現される．

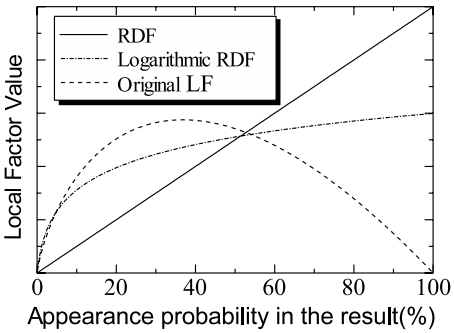

図 2 関数の形状
Fig. 2 The shape of functions.

ここで，索引付けと重要ラベル選択の違いについて考える．索引付けでは，文書集合全体と，1 つの文書に注目し，注目した文書中の語について重み付けを行う．一方，重要ラベル選択では，文書集合全体と，検索条件によって絞り込まれた部分文書集合に注目し，その部分文書集合から重要語を抽出することが求められている．

ここから，我々はラベルを抽出する対象がユーザの検索条件によって与えられている点に注目し，検索条件と関連性の高いタームが重要なラベルではないかと考えた．

そこで，検索条件と関連性が高いタームを評価する基準として，タームの文書集合 D 中での出現割合と，検索条件によって得られた部分文書集合 R 中での出現割合の比を用いた以下の基準を利用することを提案する．

- Original Global Factor（Original GF）

$$GF_i^{ORG} = \frac{DF_{R,i}/|R|}{DF_{D,i}/|D|}$$

この重みでは，文書集合 D 中と，絞り込まれた部分文書集合 R 中での出現割合を比較し，R 中での出現割合が増加したタームを，検索条件と関連性のあるタームとして評価する．

3.5 カテゴリ優先度基準

カテゴリ優先度基準とは，効率的な絞り込み機能と概観性を持ったカテゴリ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，各カテゴリに対するラベルの数は，ユーザが容易に全体を把握できる数ということで，ブラウザにおいてスクロールせずに表示できる 10 個程度を想定している．

表 1 カテゴリ優先度基準の算出法
Table 1 Equations of category ranking criteria.

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	$p_j^2 = \sum_{i \in C_j} \left(-\frac{\|D_{j,i}\|}{\sum_{i \in C_j} (\|D_{j,i}\|)} \times \log \left(\frac{\|D_{j,i}\|}{\sum_{i \in C_j} (\|D_{j,i}\|)} \right) \right)$
分類の網羅性	$p_j^3 = \|D_j\| / \|R\|$

カテゴリの優先度に関する基準としては，様々なものが考えられるが，Takata ら[19] は，複数のカテゴリ（1 階層の分類情報）を切り替えて検索結果をカテゴライズする手法の提案において，以下の基準を個々のカテゴリの有効性基準としてあげている．

- **分類の明確さ：**各ラベルを通してアクセスでき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

これらの基準は，文書集合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，かつすべての文書がいずれかのラベルに関連付けられるという保証がないためそのまま利用することはできない．そこで我々は，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できる文書集合間の重複が少ないこと」と読みかえ利用することとした．また，すべての文書がラベルに関連付けられるとは限らないという点から，以下の基準も重要となる．

- **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

これらの基準は**表 1** に示す式によって定義できる．ここで，C_j をカテゴリ j に含まれるラベルの集合，p_j をカテゴリ j の優先度の値，D_j を検索結果中でカテゴリ 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，$D_{j,i}$ を検索結果中でカテゴリ 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」は，インデクス中のラベルに関連付けられる文書ののべ数と，異なり数の比率をもとに，「分類の均一さ」は，平均エントロピにより算出している．また，「分類の網羅性」には，検索結果数とインデクス中の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．

4. 評価リソース

4.1 評価用プロトタイプシステム

評価実施のため，以上で説明した手法に基づき，全文検索システム LISTA[20] および Isozaki らの手法[14] による固有表現抽出ツールを利用し評価用のプロトタイプシステムを構築した．

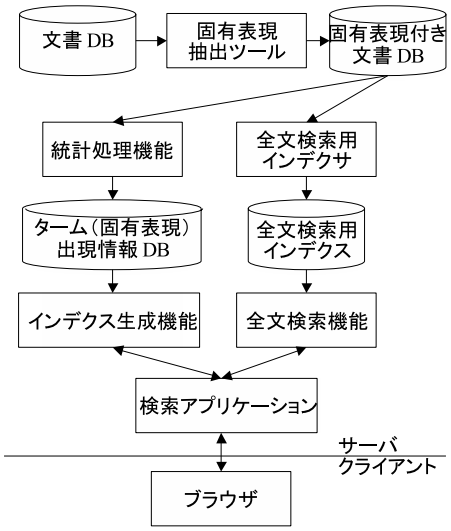

図 3 システム概要
Fig. 3 System overview.

LISTA のランキングアルゴリズムには，検索システム freeWAIS-sf で採用されている手法が利用されている[21]．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．

図 3 にシステムの概要を示す．前処理として，検索システムに登録する文書から固有表現を抽出した後，全文検索用インデクスおよびターム（固有表現）の出現情報を生成する．検索時には，Web サーバ中に構築されたアプリケーションが全文検索機能およびインデクス生成機能にアクセスし，検索結果とインデクスを取得し，ユーザに提示する．

また，**図 4** にユーザインタフェースを示す．図では，ユーザが入力した検索条件をもとに検索結果を表示した状態を示しており，右側に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，従来システムのように検索結果リスト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，検索結果を概観し

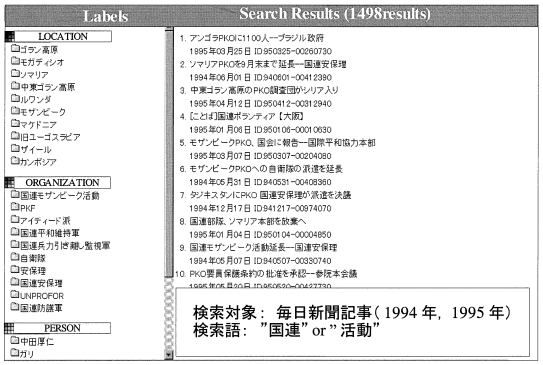

図 4 ユーザインタフェース
Fig. 4 User interface.

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を絞り込んだりすることができる．絞り込み検索は，既存の検索条件と選択されたラベルのタームを “and” で結合し，再検索することで行われる．

4.2 評価用コレクション

評価には，IREX（Information Retrieval and Extraction Exercise）[22] で利用された 1994 年および 1995 年の毎日新聞記事を利用した．文書数は約 20 万件である．

評価には，IREX で規定された 30 の検索トピックおよび正解文書の情報を利用した．各トピックについて平均約 100 件の正解文書が存在している（最小 29 件，最大 300 件）．

評価用の検索条件の作成には，各トピックについて，規定されている DESCRIPTION☆を利用した．これを形態素解析し，ストップワードを除去した後に “or” で連結し検索条件とした．

また，我々は上記の正解文書から「有効タームリスト」を作成した．作成手順としては，5 人の被験者に IREX で規定されたトピックおよび正解文書を提示，その正解文書の中から，「個々のトピックの検索において検索条件として有効なキーワード」を選択してもらった．このうち 3 人以上の被験者があげたものを「有効タームリスト」に追加した．これをそれぞれのトピックについて作成した．

5. ラベル選択基準の評価手法

5.1 評 価 法

本節では，3 章で示したラベル選択基準の評価法について示す．ラベル選択基準の評価としては，ラベルの質およびラベルを利用した絞り込み検索の結果の質に着目し，「ユーザが選択するであろうラベルを選択した場合に得られる検索結果の適合率」で評価を行った．手順は以下のとおりである．

(1) システムに検索条件を入力し，検索結果とインデクスを取得する．
(2) インデクス中の各ラベルを「有効タームリスト」と比較し，一致するラベルを「ユーザが選択するであろうと考えられるラベル」として取得する．
(3) 上記で取得した各々のラベルを用いて絞り込み検索を行い検索結果を取得する．
(4) 上記で得たそれぞれの検索結果のうち，規定ランキング以上の文書について「判定対象」とし，その適合性判定を行う．

(4) の規定ランキングには，ユーザが文書を参照する件数ということで，5 および 10 で評価を行った．これは，ユーザが検索結果の 1 ページ目の半分および 1 ページ目をすべて見ることを仮定した値である．またラベルの提示数は，1 画面において提示できるラベルの数ということで 20 とした．本章の評価では，ラベルのカテゴリは考慮せず，評価対象の基準で優先度が高いと判定されたラベルを対象とした．

評価は，3.4 節で提案した 2 つの局所的重み（Local Factor）と，1 つの大域的重み（Global Factor）を **表 2** に示すように組み合わせた基準について行った．それぞれの基準によるターム i の重要度 I_i は，以下の式によって算出する．

$$I_i = (LocalFactor) \times (GlobalFactor)$$

検索結果のランキング上位 m 件（m = 30，50，100，200，300，500 で評価）の文書に含まれるタームのうち，これらの基準のスコアが高いものから規定数分（今回は上述のとおり 20 個）をラベルとした．

また，単純に出現頻度を利用する方法（FREQ）および TF-IDF 法に基づく方法（RDF-IDF）を比較対象とした．

得られた検索結果ごとに以下の式で適合率を計算し，

表 2 ラベル選択基準
Table 2 Label selection criteria.

Method ID	Local Factor	Global Factor
FREQ	RDF	—
RDF-IDF	RDF	IDF
LRDF-IDF	Logarithmic RDF	IDF
ORG-IDF	Original LF	IDF
RDF-ORG	RDF	Original GF
LRDF-ORG	Logarithmic RDF	Original GF
ORG-ORG	Original LF	Original GF

☆ 3 つまでの自立語から構成される検索要求の簡潔な表現である [22].

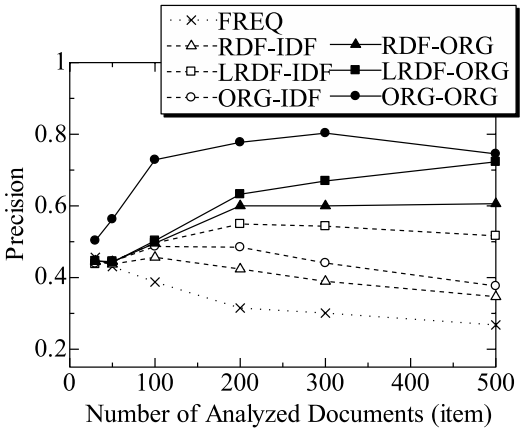

図 5 ラベル選択基準の評価（検索結果上位 5 件を評価）
Fig. 5 Evaluation results of label selection criteria (for each top 5 results).

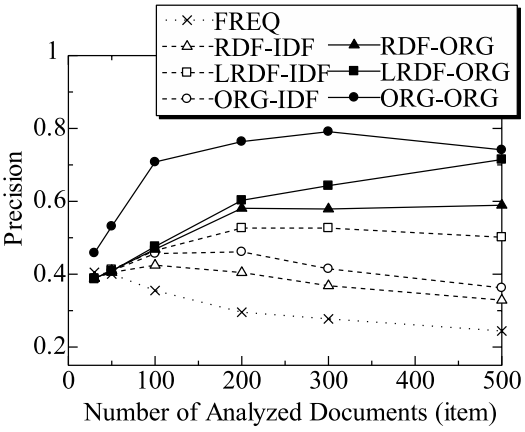

図 6 ラベル選択基準の評価（検索結果上位 10 件を評価）
Fig. 6 Evaluation results of label selection criteria (for each top 10 results).

平均した値を評価に利用した．

$$\text{適合率} = \frac{\text{「判定対象」中の正解文書数}}{\text{「判定対象」の文書数}}$$

5.2 評価結果

図 5 および**図 6** にラベル選択基準の評価結果を示す．それぞれのグラフは検索結果の上位 5 件および 10 件を適合性判定の対象とした場合の結果である．横軸はラベルを出力するために処理した検索結果の数を示し，縦軸に適合率を示している．

まず大域的重みに注目する．Original GF を用いた基準と IDF を用いた基準の適合率を比較すると，両方のグラフとも，ほとんどの場合に，提案手法である Original GF を用いた基準が IDF を用いた基準と比較し上回っている．これは，ラベル選択の基準として，3.4.2 項で示した検索条件と関連性の深いタームが重要であるとした考えが正しいことを証明している．

一方，局所的重みに注目すると，大域的重みの指標にかかわらず，従来手法である RDF を利用する手法と比較し，今回提案した Logarithmic RDF および Original LF を利用した場合の適合率が上回っている．これは 3.4.1 項で示した考えが正しいことを証明している．

また，Logarithmic RDF と Original LF について見ると，前者に比べて後者の場合は比較的文書処理量が少ない値で最大の適合率を記録し，その後処理する文書数が多くなるにつれて，適合率が低下する傾向が見られる．この傾向は大域的重みに Original GF を用いた場合に顕著に現れ，処理する検索結果の件数が 300 件以下の場合には Original LF が高い適合率を示しているが，500 件では Logarithmic RDF とほぼ同等の値を示している．

それぞれの適合率について見ると，大域的重みに Original GF を用い，局所的重みに Original LF もしくは Logarithmic RDF を用いた指標では絞り込み適合率が 60〜80%の値となっている．これはいい替えると，絞り込み検索で得た検索結果のうち半分〜4/5 は適合文書であるといえる．

以上の結果は，従来手法と比較し，提案手法によって選択したラベルが，適合文書と高い関連性を持ち，提案手法を利用することで，高い精度の検索結果を得ることができるということを示している．

一方，TF-IDF 法に基づいた指標（RDF-IDF）では提案手法を用いたすべての指標の評価値を下回り，単純に頻度を利用した場合と大きな差がない結果となっている．

6. インデクス提示法の評価

6.1 評価法

本章では，3 章に示したインデクス提示法の評価について示す．本評価では，システムの使いやすさについての評価を行うため，12 人の被験者による検索システム評価実験を実施した．被験者は日常インターネットの検索サイトでキーワード検索を行う人たちで 20 代から 40 代の男女である．

評価には，2 つのシステムを用意した．1 つは本稿での提案システムで，検索結果に対するインデクスでラベルを固有表現カテゴリに応じて分類するシステム（システム 1；**図 7** 参照），もう 1 つはインデクスでラベルを分類しないシステム（システム 2；**図 8** 参照）である．

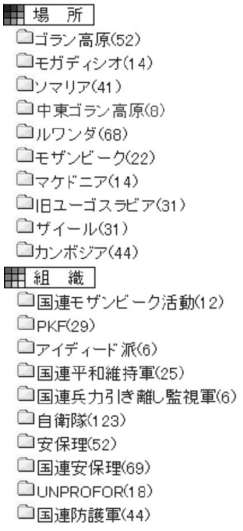

図 7 システム 1 のラベル出力例
Fig. 7 Example of labels presented by system1.

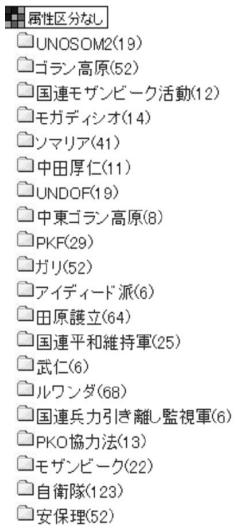

図 8 システム 2 のラベル出力例
Fig. 8 Example of labels presented by system2.

システム 1 では，固有表現のカテゴリとして「人名」,「組織名」,「地名」,「その他」を利用し，検索が行われるごとにカテゴリを 3.5 節で提案した 3 つの指標の積により重み付けし，重みの大きい方から順に提示した．システム 1 およびシステム 2 ともラベルの数は 20 個程度になるように設定した．

個々の被験者はそれぞれのシステムで 10 個のトピックについて検索を行う．1 つのトピックについてユーザはすべての正解文書を見つけるか，30 分経過するまで検索を行う．

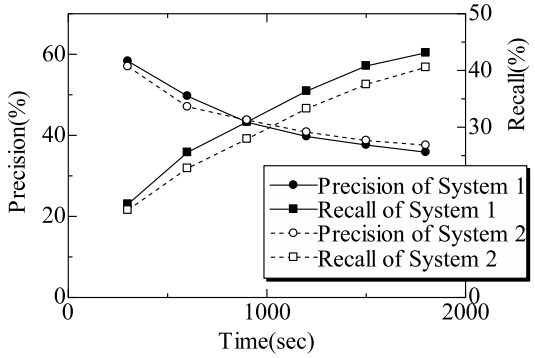

図 9 被験者による検索実験の適合率および再現率
Fig. 9 Recall and precision of search tests by subjects.

このユーザの行動をログとして記録し，時間ごとの適合率と再現率およびラベルの利用回数を評価した．ここで適合率，再現率は以下の式で算出する．

$$\text{適合率} = \frac{\text{ユーザが参照した正解文書数}}{\text{ユーザが参照した文書数}}$$

$$\text{再現率} = \frac{\text{ユーザが参照した正解文書数}}{\text{正解数}}$$

上式の「ユーザが参照した正解文書数」は，ユーザが参照した正解文書の異なり数であり，同じ正解文書が複数回参照された場合には，最初の 1 回のみを正解，それ以外を不正解の扱いとした．

また，被験者には，この検索実験終了後に，以下の項目についてのアンケートに回答してもらった．

- どちらのシステムが使いやすかったか．
- それぞれのシステムの使いやすかった/使いにくかった点について（自由文回答）．

6.2 評価結果

前節の手法に従って，被験者による検索実験を行い，システムを用いた実験のログと実験後のアンケート結果について比較分析した．前節での評価結果をもとに，ラベル選択基準には表 2 に示す ORG-ORG を用い，文書処理数は最大 300 件とした．最大としたのは，入力される検索条件によって検索結果が 300 件に到達しない場合があるからである．

図 9 に，ユーザの検索時間とそれにともなう適合率および再現率の変化を示す．グラフから適合率はシステム間にほぼ差がない．一方，再現率は今回の提案手法であるシステム 1 の値がシステム 2 を上回っている．

また，**図 10** に，検索時間に対するキーワードの平均入力回数の推移およびキーワードを 1 回入力するごとの平均ラベル使用回数の推移を示す．この図から，システム 2 に比べてシステム 1 を利用した場合の方

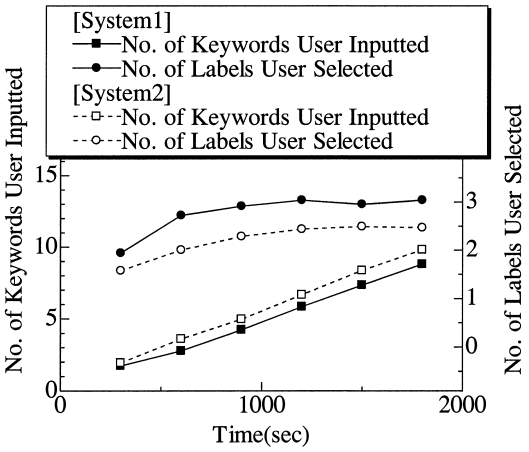

図 10 キーワードおよびラベルの利用頻度
Fig. 10 Number of keywords and labels.

が，キーワードを入力する回数が減少し，ラベルを利用する回数が増加していることが分かる．図 9，図 10 の結果から，システム 1 を利用するユーザはシステム 2 の場合と比べて検索にラベルを利用する割合が多く，より多くの検索結果を参照し，より多くの適合文書を得ることができたといえる．

次にアンケートの結果について示す．

「どちらのシステムが使いやすかったか」という問いに対しては，83%のユーザ（12 人中 10 人）が提案手法であるシステム 1 の方が使いやすいと答えた．

また，「それぞれのシステムの使いやすかった/使いにくかった点について」尋ねた自由記述文アンケートでは，システム 1 の使いやすかった点として，5 人のユーザが「ラベルがカテゴリで分類されている点」をあげており，残りの 7 人は「ラベルが提示されている点」そのものをあげている．一方，システム 1 の使いにくさとしては，システム上の共通の問題（「検索結果のランキングが文書適合順ではなく時間順の方がよかった」，「1 度選択したラベルが参照していないラベルと同じ色で表示され分かりにくい」）を除いては，「抽象的な内容のものは見付けにくかった」「ラベルを見ても分からない分野はどうしようもなかった」「2 つ以上のカテゴリにまたがった分類がなされることがあった（固有表現抽出のカテゴリ誤り）」「欲しいと思うラベルがないことがあった」との意見があげられた．

また，システム 2 の使いにくさとして 6 人のユーザが「ラベルがカテゴリで分類されていない点」をあげ，カテゴリについては言及していないもののそれ以外の 4 人のユーザが「ラベル自体の質の悪さ」を指摘していた（システム 1 とシステム 2 で提示されるラベルは同様）．残りの 2 人は，上記に示したシステム上の共通の問題をあげた．使いやすさとしては 11 人が「ラベルによって検索結果が分類されている点」をあげ，残る 1 人のユーザは「特に使いやすかったとは思えない」と答えた．

以上より，重複を除くと 12 人中 8 人が「システム 1 はラベルをカテゴリで分類してあり使いやすい」もしくは「システム 2 はラベルが分類されていないので使いにくい」という点について言及していた．

以上，2 つのアンケートの結果からラベルを固有表現のカテゴリごとに分類して提示することで，ユーザ自身，検索結果を概観しやすく，効率的な絞り込み検索ができたと感じていることが確認できた．

以上の結果より，インデクス中の同種のラベルを分類して提示することで，インデクスを使いやすくできるということが証明できた．

7. 従来手法との比較評価

7.1 評 価 法

提案手法の有効性評価のため，2 章で示した従来手法との比較を行った．今回比較対象としたのは，NTCIR-4 の Web Task D として実施された「トピック分類タスク」(“NTCIR-4 Web D”)[23] で最も高い精度を示した Ohta らの手法[11] である．

比較評価の実施にあたっては，文献 11) をもとに手法を再現し，今回利用した新聞記事データを対象としたシステムを構築した☆．

この手法では，検索結果の文書中に出現する名詞および未知語の連続からなる語およびそれを含む複合語をタームとして抽出し，タームの出現する文書のランキングならびに文書中での出現位置等に応じてスコア付けを行いラベルを選択している．複合語は 2 階層目のラベルとして利用される．今回比較評価に用いた各パラメータは，文献 11) で示されている “NTCIR-4 Web D” のフォーマルランで最も高い評価を得たものを採用した．

比較評価は，以下の 2 つの手法で行った．

- 容易にアクセスできる n 件の適合率での評価
  “NTCIR-4 Web D” で実施された評価手法に基づき，優先的に提示されるラベルを利用して取得できる文書 n 件を取得，その適合率を評価した．最上位のラベルが n 件以上の検索結果と関連する場合，上位 n 件の文書のみを取得し，評価し

☆ 形態素解析後の不要形態素の除去，複合語の作成方法等，参考文献だけで十分な情報が得られない内容については，同著者らの先行文献である文献 24) 等を参考にした．

た．また，n 件以下の場合は n 件に到達するまで，次のラベルに関連する文書を取得し，各トピックにおける適合率を平均し評価した．今回 $n = 5$，10，20 として評価した．また，ベースラインとなる全文検索システムの結果も同時に比較した．適合率は各トピックについて以下の式で求め，全トピックの平均で評価を行った．

$$\text{適合率} = \frac{\text{取得した } n \text{ 件中の正解文書数}}{n}$$

- ユーザが選択するであろうラベルを選択した場合に得られる検索結果の評価

  5 章の評価と同等の評価である．Ohta らの手法については，1 階層目のラベルについて「有効タームリスト」との一致を判定し，ラベルが選択されるか否かを判定した．2 階層目のラベルの選択には，“NTCIR-4 Web D” での評価に従い同一の親ラベル配下にあるラベルのうち最も正解の多いラベルを選択するという理想的な状態を仮定した．ラベルを選択した場合の絞り込み検索は，“NTCIR-4 Web D” の手法に従い，最初の検索結果のうち上位 200 件に含まれる文書のみを対象とした．また，ここでは比較手法との特性の差を評価するため，「有効タームリスト」と一致するラベルをすべて選択すると仮定した場合の検索結果集合を特定し，再現率の比較も行った．各トピックの再現率は以下の式で求め，全トピックの平均で評価を行った．

$$\text{再現率} = \frac{\text{検索結果集合中の正解文書集合}}{\text{正解文書集合}}$$

  再現率，適合率の評価は，それぞれの絞り込み検索結果の上位 p 件（$p = 5$，10，20）について行った．

7.2 評価結果

前節で示した手法により，Ohta らの手法との比較評価の結果を示す．まず「容易にアクセスできる n 件の適合率」の評価結果を**図 11** に示す．ベースラインとして，全文検索で得た結果も示している．$n = 5$ の場合には，提案手法は，従来手法と比較し約 10%，全文検索と比較すると 15%以上適合率が上昇している．しかし，件数が増加するにつれて差が小さくなり，$n = 20$ の場合には，どの手法もあまり有意な差がなくなっている．

この “NTCIR-4 Web D” に基づく評価は，優先的に選択されるラベルを，ラベルに関連する正解文書の多さで決定するという，理想的な状態を仮定している．このため，実際にユーザが選択するラベルかどうか分からない評価であるといえる．

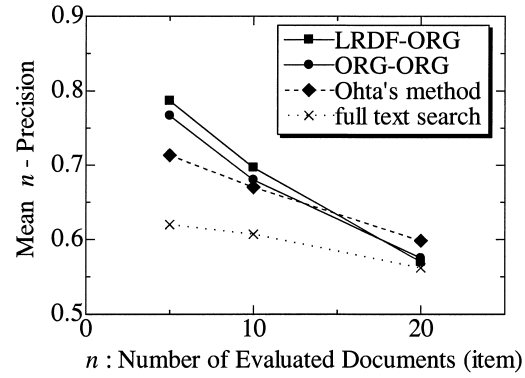

図 11 容易にアクセスできる n 件の平均適合率（$n = 5$，10，20）
Fig. 11 Mean n-precision ($n = 5$, 10, 20).

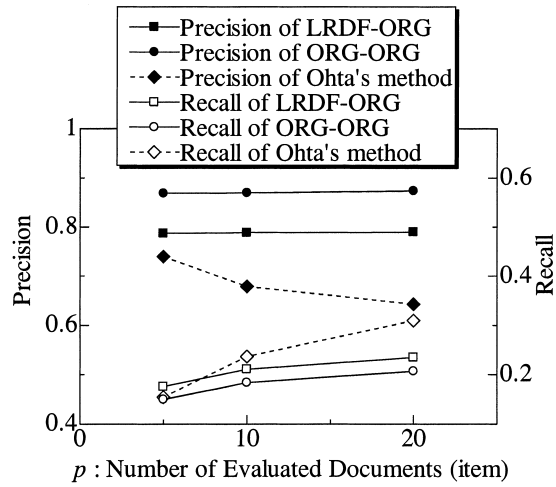

図 12 検索精度の評価（$p = 5$，10，20）
Fig. 12 Evaluation of search accuracy.

そこで，そのラベルが選択されるかどうかを人手によって作成した「有効キーワードリスト」と比較して評価する「ユーザが選択するであろうラベルを選択した場合に得られる検索結果の評価」の結果を**図 12** 示す．

提案方式によると，適合率は従来手法と比較してつねに高い値をとる．一方，再現率は $p = 5$，10 の場合は従来手法と同等の値を示しているが，以降の再現率の上昇は従来手法と比較して低い．これは，提案手法が，関連する文書が比較的少ないラベルを生成する傾向にあることに起因すると考えられる．

以上の結果より，提案手法は，従来手法と比較し，ユーザが選択しそうなラベルで絞り込んだ検索結果の上位に正解を提示できるという特性を持つといえる．一方で，正解文書集合が大きい場合には，十分に正解文書を網羅できない場合が考えられる．

この提案手法の特性と，インターネット上等での検索において，多くのユーザが検索結果の上位しか見ないという現実をあわせて考えると，提案手法は，従来手法より有益な検索結果をユーザに提示できる手法であるということができる．

8. まとめ

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，検索結果とともに，検索結果中での主要な話題を，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ナビゲーション手法を提案した．また，それに基づくプロトタイプシステムを試作し，従来手法との比較評価を実施した．

ラベルの候補の抽出に固有表現抽出技術を利用すること，および，新たなラベルの選択基準を提案した．IREX の正解セットを用いた評価において，提案手法の基準を用いることで従来手法と比較し，ラベルおよび検索結果の質が高くなることを示した．

また，インデクスの提示法として，固有表現のカテゴリを用いて個々のラベルを分類し提示する方法および優先的に提示するカテゴリの評価基準を提案した．被験者を用いた検索システム評価実験を行い，提案手法が検索効率に寄与することを示した．また，被験者のアンケートから，8 割以上のユーザが，本提案による検索システムが，ラベルを分類しないシステムと比較して使いやすいと感じているとの結果を得た．

さらに，NTCIR-4 の「トピック分類タスク」[23] で最も高い精度を示した手法[11] と比較評価を行い，実際の検索エンジンにおいては提案手法の方がユーザに有益な検索結果を提示可能であるとの知見を得た．

参考文献

1) Belkin, N.J.: Anomalous states of knowledge as a basis for information, *Canadian Journal of Information*, Vol.5, pp.133–143 (1980).
2) Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern Information Retrieval*, Addison-Wesley (1999).
3) 杉本雅則：情報収集システムにおけるユーザモデリングと適応型インタラクション，人工知能学会学会誌，Vol.14, No.1, pp.25–32 (1999).
4) 徳永健伸：情報検索と言語処理，東京大学出版会 (1999).
5) 北 研二，津田和彦，獅子堀正幹：情報検索アルゴリズム，共立出版 (2002).
6) Cutting, D.R., Karger, D.R., Pedersen, J.O. and Tukey, J.W.: Scatter/Gather: A cluster-based approach to browsing large document collections, *SIGIR '92: Proc. 15th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, pp.318–329, ACM Press (1992).
7) Hearst, M.A., Karger, D.R. and Pederson, J.O.: Scatter/Gather as a tool for the navigation of retrieval results, *AAAI Fall Symposium on Knowledge Navigation*, pp.65–71 (1995).
8) Leuski, A.: Evaluating document clustering for interactive information retrieval, *CIKM '01: Proc. 10th international conference on Information and knowledge management*, New York, NY, USA, pp.33–40, ACM Press (2001).
9) Sakai, H., Ohtake, K. and Masuyama, S.: A Retrieval Support System by Suggesting Terms to a User, *ICCPOL2001: 19th International Conference on Computer Processing of Oriental Languages*, pp.77–80 (2001).
10) Salton, G. and Yang, C.G.: On the Specification of Term Values in Automatic Indexing, *Journal of Documentation*, Vol.29, pp.351–372 (1973).
11) Ohta, M., Narita, H. and Ohno, S.: Overlapping Clustering Method Using Local and Global Importance of Feature Terms at NTCIR-4 Web Task, *Working Notes of NTCIR-4*, Vol.Supl.1, pp.37–44 (2004).
12) Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y. and Ma, J.: Learning to cluster web search results, *SIGIR '04: Proc. 27th annual international conference on Research and development in information retrieval*, New York, NY, USA, pp.210–217, ACM Press (2004).
13) Hisamitsu, T., Niwa, Y. and Tsujii, J.: Measuring Representativeness of Terms, *IRAL 1999*, pp.83–90 (1999).
14) Isozaki, H. and Kazawa, H.: Efficient Support Vector Classifiers for Named Entity Recognition, *COLING*, pp.390–396 (2002).
15) Bikel, D.M., Schwartz, R. and Weischedel, R.M.: An Algorithm that Learns What's in a Name, *Machine Learning*, Vol.34, No.1–3, pp.211–231 (1999).
16) Grishman, R. and Sundheim, B.: Message Understanding Conference – 6: A Brief History, *COLING*, pp.466–471 (2002).
17) Sekine, S. and Nobata, C.: Definition, dictionaries and tagger for Extended Named Entity Hierarchy, *LREC 2004*, pp.1977–1980 (2004).
18) 関根 聡：固有表現から専門用語，NLP2004：言語処理学会第 10 回年次大会「固有表現と専門用語」ワークショップ (2004).
19) Takata, Y., Nakagawa, K. and Seki, H.: Flexible Category Structure for Supporting WWW

Retrieval, *2nd International Workshop on the World Wide Web and Conceptual Modeling*, pp.165–177 (2000).
20) Hayashi, Y., Tomita, J. and Kikui, G.: Searching text-rich XML documents, *ACM SIGIR 2000 Workshop on XML and Information Retrieval*, pp.27–35 (2000).
21) 富田準二，竹野　浩，菊井玄一郎，林　良彦，池田哲夫：グラフモデルの提案とテキスト検索システムへの適用による評価，情報処理学会論文誌：データベース，Vol.43, No.SIG2(TOD13), pp.94–107 (2002).
22) 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，pp.1–5 (1999).
23) Eguchi, K.: Overview of the Topical Classification Task at NTCIR-4 WEB, *Working Notes of NTCIR-4*, Vol.Supl.1, pp.ov-48–ov-55 (2004).
24) 成田宏和，太田　学，片山　薫，石川　博：Web 文書検索のための非排他的クラスタリング手法の提案，DEWS2003：第 14 回データ工学ワークショップ DEWS2003 (2003).

(平成 17 年 3 月 20 日受付)
(平成 17 年 5 月 9 日採録)

(担当編集委員　岸田 和明)

戸田　浩之（学生会員）

日本電信電話株式会社サイバーソリューション研究所所属．1999 年名古屋大学大学院工学研究科材料プロセス工学専攻博士課程前期課程修了後，日本電信電話株式会社に入社．2005 年 4 月より，筑波大学大学院博士課程在学中．情報検索，情報抽出関連の研究開発に従事．日本データベース学会会員．

中渡瀬秀一（正会員）

日本電信電話株式会社サイバーソリューション研究所所属．1992 年神戸大学大学院工学研究科修士課程修了後，日本電信電話株式会社に入社．以来，情報探索，自然言語理解等の研究に従事．

片岡　良治（正会員）

日本電信電話株式会社サイバーソリューション研究所所属．1987 年千葉大学大学院電子工学専攻修士課程修了後，日本電信電話株式会社に入社．以来，トランザクションの並行処理制御方式の研究，マルチメディア情報システムの研究，ポータルサービスシステムの研究開発に従事．電子情報通信学会会員．