ROC距離に基づく先読み検索手法の提案と性能評価

野村 賢[†] 河野浩之[†] 川原 稔[††]

データマイニング技術を検索式改善支援に応用したサーチエンジン「問答」を構築する過程において，多くの導出ルールを列挙しただけでは検索式との関係が把握しにくいといった問題が生じた．そこで，本稿では，適切なルール選択が行えるように，ROC グラフ上に，検索式と導出ルールの関係を可視化する．また，対象データを一括先読みすることで，導出ルールと対象データとの関係も可視化する．さらに，半構造データを用いた手法の性能評価およびサーチエンジンへの実装を行う．

Performance Evaluation of Materialized Query Processing Based on ROC Distance

Ken Nomura,[†] Hiroyuki Kawano[†] and Minoru Kawahara[††]

In a process to develop our search engine system, Mondou using data mining technology, it is difficult to show the relationship between a given query and associative keywords by the list indication. In this paper, we visualize this relationship based on ROC distance. Moreover, in order to visualize the relationship between a search result and associative keywords, we try to materialize the query process. we also evaluate the performance of our materialized query processing using semi-structured data.

1. はじめに

近年，大量の電子化データが蓄積されるにつれ，検索支援技術の重要性が高まっている．このような大量のデータから知識発見を行う研究は，データマイニング，もしくは，データベースからの知識発見（KDD: Knowledge Discovery in Database）と呼ばれ，データベース技術や統計処理，さらには論理学など，数多くの分野が強く関係している．しかし，いずれの分野においても，ノイズを含む膨大な実データから知識と呼びうる記述を導出するシステム構成技術を確立するまでに，解決すべき問題は数多く残されている．そこで，我々はデータベースシステムに対して，どのような知識発見アルゴリズムを実装し，また，設計制約をどのように緩和するべきかを明確にするための研究を行っている．また，半構造データにおける情報検索においては，一般に検索領域に対する領域知識に加えて，検索システムに習熟することが必要である．よって，様々な情報検索システム利用を容易にするために，検索支援技術などの研究を行っている[9)~11)]．

† 京都大学大学院情報学研究科
Department of Systems Science, Kyoto University
†† 京都大学大型計算機センター
Data Processing Center, Kyoto University

我々は，データマイニング手法の 1 つである相関ルール（association rule）導出アルゴリズムを拡張して，Web データや文献情報などの半構造データに対する検索式生成支援システム「問答（Mondou）」の構築を行っている[9),12)]．「問答」は，導出されたルールである関連キーワードを検索ユーザに提示することで，検索にかかわる知識を与えて検索支援を行うシステムである．現在，「問答」では，Web の次世代フォーマットである XML にも用いられるタグ情報などのメタ情報を活用したデータマイニングアルゴリズム研究のため，種々のデータを XML の上位互換である SGML フォーマットに変換して評価を行っている．

これまでの「問答」では，**図 1** のように，導出された関連キーワードをリスト形式で検索ユーザに提示しており，検索ユーザはここから関連キーワードの選択をすることになるが，このようなリストによる関連キーワードの提示方法では，提示される関連キーワードとクエリ（問合せ），および，関連キーワードと検索結果との関係を直感的に理解することが難しい．また，膨大な文書データから実用性の高いルールを効率良く導出するためのアルゴリズムの研究の結果[11)]，多くの有効な関連キーワードが得られるようになり，そのすべてをリスト形式で提示したのでは表示画面が煩雑になってしまう．そこで，本稿では，この問題の解

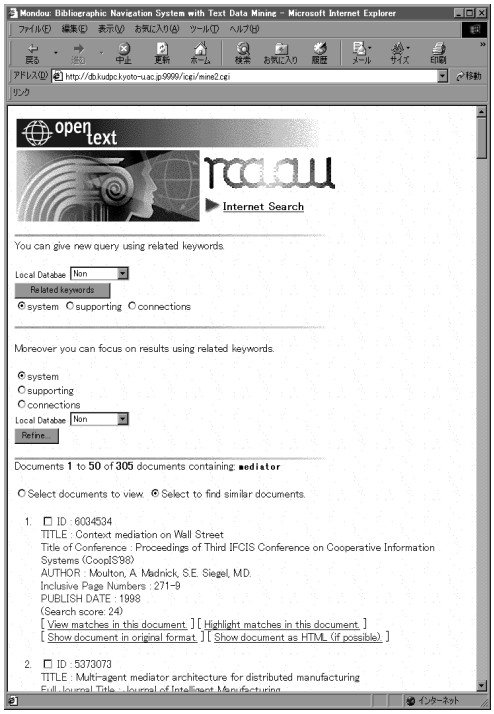

図 1 従来の「問答」による検索結果
Fig. 1 Result window by the old Mondou.

決手法として，情報可視化技術を応用した可視化インタフェースによる関連キーワード提示手法を提案する．

以下，2 章では，従来システムの問題点とその解決手法について述べる．3 章では，INSPEC データを用いて 2 章で提案した手法の性能評価を行う．4 章では，プロトタイプ検索支援システムの構成および実装について述べる．5 章では，結論と今後の課題を述べる．

2. 検索支援システムの問題点と可視化手法

「問答」では，相関ルール導出アルゴリズム[1),9),12)]を情報検索に応用している．データベースにおいて，キーワード集合 X と Y をともに含むタプルに成り立つ関係を相関ルール $X \Rightarrow Y$ として扱い，その条件部 X をユーザが入力するクエリ，結論部 Y に含まれるキーワードを関連キーワードとして提示し，それらを利用させることで，ユーザに対する検索支援を行っている．X と Y は共通のキーワードを含まず $(X \cap Y = \emptyset)$，データベース全体に対してともに含まれる割合 $p(X \cup Y)$ を表す支持度（support），および，X を含むタプルが Y も含む割合 $p(Y \mid X)$ を表す確信度（confidence）が，ヒューリスティックに定められたしきい値（最小支持度，最小確信度）を超えたものが相関ルールとして採用される．

なお，これまで我々は，相関ルール導出アルゴリズムに対して，重み付け[9)] や SGML のメタ情報の利用[12)] など様々な改良を行ってきたが，本稿では，改良されたアルゴリズムによる影響を避けるため，一般的なアルゴリズム[1)] によりルール導出を行っている．

2.1 検索支援における問題点

従来の「問答」では図 1 の検索画面のように，検索結果の一部を表示するとともに，相関ルールとして導出された関連キーワードをリスト形式で提示する[9),12)]．検索は，クエリの入力から始まり，満足な検索結果が得られなかった場合，検索ユーザは関連キーワードを用い検索式を改善し再検索を行う形で進む．

選択した関連キーワードが目的の検索にふさわしいかは，検索結果を見て判断することになるが，満足な結果が得られない場合，以前の検索式に戻して別の関連キーワードと組み合わせて再検索を行うなど，バックトラックすることが少なくない．そして，再検索のたびに，検索ユーザは検索結果の提示を待たなければならず，特に長時間の検索においては，問合せ済み検索式を把握することも難しくなる．以下に，これらの問題点をまとめる．

問題点 1

クエリと導出された関連キーワードとの関係が明確に提示されていないため，検索ユーザは勘に頼ったキーワード選択を必要とする．

問題点 2

複数の関連キーワードが提示されるため，検索ユーザは検索作業を反復しなければならないことが多い．このとき，関連キーワードと検索結果との関係が提示されていないため，関連キーワードの取捨選択には試行錯誤を要する．また，検索ユーザは再検索のたびに問合せ結果を待たなければならない．

問題点 1 は，検索ユーザに関連キーワード選択の指針となるパラメータを提示することで解決を試みる．ただし，複数のパラメータをリスト表示に加えて提示するのでは，各関係を直感的に理解するのは難しい．

問題点 2 を解決するには，効率良く検索式の改善が行えるインタフェースが要求される．そこで，情報可視化技術を用いた解決を試みることにする．

2.2 可視化技術を用いたインタフェース構成

情報可視化（Information Visualization）はここ 10 年以内に精力的に研究されつつある，比較的歴史の浅い分野である[2)]．可視化手法は，計算コストや通信コストの大きさから実装が難しかったが，近年の計算機

とネットワークの性能の向上により様々な試みが行われつつある．なお，検索支援における可視化の必要性は，既存の検索システムの問題点として 90 年代初頭にも指摘されている[5]．

そこで，2.1 節で示した 2 つの問題点の解決方法として以下の 3 手法を提案する．

可視化 1

問題点 1 で指摘したクエリと関連キーワードの関係を可視化により提示する．可視化には ROC（Receiver Operating Characteristic）グラフを用いる[3),6),11)]．ROC グラフは，分類子のパフォーマンスをクラス分布やコスト分布から分離して視覚化することにより，クラス分布やコスト分布を正確に把握することが困難な場合でも，分類子のパフォーマンスを比較することを可能にする手法である．

ROC 解析では，ある事象が 2 つの事象クラス「正の事象クラス：P（positive）」と「負の事象クラス：N（negative）」に分類できると仮定する．そして，事象は分類子により「正：y（yes）」と「負：n（no）」に分類されるとする．このとき，正の事象 P が正 y と正しく分類される比率 TP（True Positive rate），および，負の事象 N が誤って正 y と分類される比率 FP（False Positive rate）は，

$$TP = p(y \mid P) \simeq \frac{\text{正と分類された正の事象の個数}}{\text{すべての正の事象の個数}}$$

$$FP = p(y \mid N) \simeq \frac{\text{正と分類された負の事象の個数}}{\text{すべての負の事象の個数}}$$

のように表すことができる．

情報検索においては，検索ユーザにより投入された初期クエリによる検索結果を，効果的なキーワードを用いて絞り込む必要性があるため，初期クエリの被覆する領域を正の事象クラス B と定義し，逆に検索領域を拡大する方向の領域を負の事象クラス $\bar{B}$ と定義した．よって，初期クエリから導出された関連キーワードが被覆する文献の集合を R とすると，初期クエリにより被覆される文書集合 B を用い，TP および FP は次式のように表せる．

$$TP = \frac{\mid B \cap R \mid}{\mid B \mid}, \qquad FP = \frac{\mid \bar{B} \cap R \mid}{\mid \bar{B} \mid}$$

この，FP 値，TP 値をそれぞれ X 軸の値，Y 軸の値とし，グラフ上にプロットする（ROC グラフ）ことで，分類子（ルール）による分類結果を視覚的に表すことができる．いくつかの文書集合についてプロットを繰り返していくと，TP と FP の定義上の関係から，ROC グラフ上 (0,1) の点に向かって凸状のカーブが描かれる．ROC グラフ上では，カーブが左上端に近づくほど，すなわち，TP が大きく FP が小さくなるほど，分類子による分類が正確であることを意味する．つまり，ROC グラフ上では，最高のパフォーマンスを示す点 $(0,1)$ に最も近い位置にプロットされる事象が最も高いパフォーマンスを示すことから，パフォーマンスを示すパラメータとして，最低のパフォーマンスを示す点 $(1,0)$ からの距離

$$L = \sqrt{(TP)^2 + (1 - FP)^2}$$

を用いることができ，これを ROC 距離と定義する[4),11)]．なお，ここで $(0,1)$ からの距離を用いず $(1,0)$ からの距離を用いたのは，$(0,1)$ に向かって事象が凸状にプロットされる傾向があるためである．

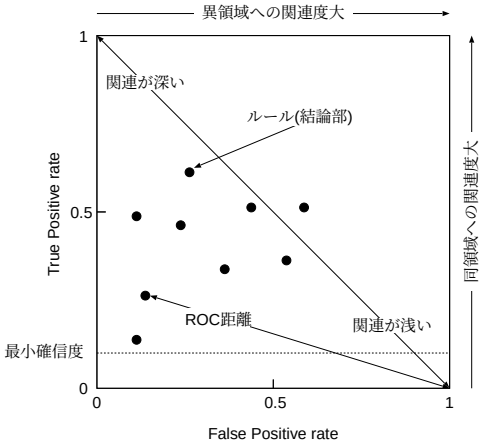

図 2 ROC グラフ上での意味
Fig. 2 Semantic property on the ROC graph.

通常，ROC 解析は分類子のパフォーマンスを測る手法であるが，分類子によって分類された事象，すなわち，導出された個々の関連キーワードに着目すると，分類子のパフォーマンスを高くするキーワードは，翻って，パフォーマンスの高いキーワードであると考えられる．したがって，パフォーマンスの高いキーワードを優先してクエリの改良に用いれば，より効果的な絞り込み検索が行えると考えられるので，導出される関連キーワードをそれぞれの (FP, TP) で ROC グラフ上に配置すれば，キーワードのパフォーマンスを視覚的に示すことができ，検索ユーザは ROC グラフ上の位置を指針として関連キーワードからの選択を行えると考えられる．

ROC グラフ上に配置された関連キーワードの，その位置における意味は**図 2** のようになると考えられる．個々の導出ルールを ROC グラフ上に配置した場合，TP の意味は，検索キーワードが検索空間に占め

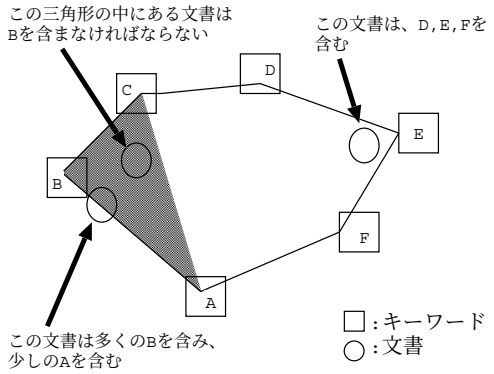

図 3 VIBE システムの概念図
Fig. 3 VIBE system outline.

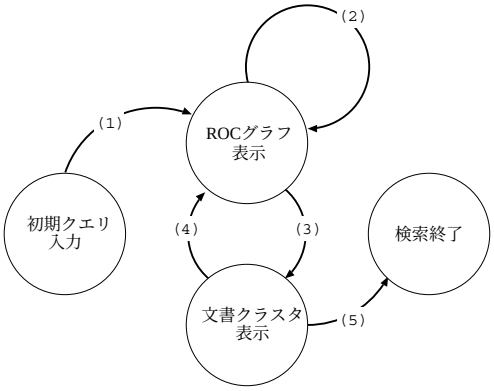

図 4 検索遷移図
Fig. 4 Search distribution diagram.

る領域をその導出ルールが占める割合であるから，相関ルール導出アルゴリズムにおける確信度を示すものとなる．TP が確信度に相当することから，TP 値が高い導出キーワードは，検索キーワードと慣用句や熟語の関係であるなど，検索キーワードと切り放されて出現する確率が低くなる．逆に，FP は，検索キーワードが検索空間を占有しない領域に対する導出ルールの占有率であるため，検索キーワードが関係しない領域に対する共起性を示す確信度と考えることができる．したがって，FP 値が高い導出ルールは，検索キーワードと関連が低いか，検索空間全体に対して関係するストップワードなどのキーワードであると考えられる．

可視化 2

問題点 2 で指摘した関連キーワードと検索結果の関係を可視化により提示する．まず，検索ユーザには，**可視化 1** を基にして任意の n 個の関連キーワードを選択してもらう．そして，それら n 個の関連キーワードを「含む（AND）/含まない（NOT）」の 2 値で表現したときの組合せ数 2^n の検索式が個々に被覆する文書集合を文書クラスタと呼び，検索式とクラスタを相互参照することで関係を確かめられるようにする．

可視化には VIBE（Visual Information Browsing Environment）システム[5]（**図 3**）を参考にして，関連キーワードを多角形の各頂点に配置し，文書クラスタをボタンとして配置するなど，簡単なシンボルとして配置し，検索ユーザはシンボルを選択することで，文書クラスタを選択できるようにする[8]．文書クラスタが選択されると，それに対応する検索結果を瞬時に画面内に提示する．クエリと選択された関連キーワードの AND/NOT の組合せによる検索式と，その検索結果であるクラスタを相互参照できるように，クラスタにはその検索式に対する検索結果がデータベースに存在すれば，サンプルを抽出して格納しておく必要がある．

先行一括問合せ

可視化 2 では，検索ユーザによる文書クラスタの選択に対し，瞬時に情報提示を行うことで，情報検索における反復作業の緩和を行う試みである．したがって，いかなる文書クラスタの選択に対しても即応することが重要であり，この実現のために，**可視化 1** から **可視化 2** に遷移する際に，上述の 2^n 個の文書クラスタに含まれる文献情報を可視化インタフェースの提示に先行して，あらかじめ一括して取得する．

2.3 提案手法を用いた検索の遷移

2.2 節で示した手法を用いたシステムの検索過程の状態遷移図は**図 4** のようになる．ブラウザ上に表示される画面は，初期クエリ入力画面，ROC グラフ表示画面（図 12 参照），文書クラスタ表示画面（図 13 参照）の 3 つのノードからなり，それらのノードを以下のように遷移する．

初期クエリ入力画面

検索は，ユーザによる初期クエリ入力から始まる．検索式には論理和，論理積などを用いることができる．

ROC グラフ表示画面

可視化 1 による関連キーワードの提示を行う．検索ユーザは，ROC グラフ上のキーワードの配置から適当な関連キーワードを選択して，検索式の再構成を行う．これにより改善された検索式を用いて，さらに関連キーワードを導出する（遷移 (2)）．あるいは，検索結果を確認する（遷移 (3)）ことができる．

文書クラスタ表示画面

可視化 2 による関連キーワードと検索結果の提示を行う．ROC グラフ表示画面で選択した関連キーワー

ドのすべての組合せによる問合せに対し，該当文書の存在するすべての文書クラスタが提示され，ユーザは論理式を意識せずに検索結果の確認を行うことができる．その結果，さらに関連キーワードを導出する必要があれば ROC グラフへ戻る（遷移 (4)）．満足のいく結果が得られれば検索を終了する（遷移 (5)）．

遷移 (1)

クエリが入力されると，関連キーワードを導出し ROC グラフ表示画面に移行する．

遷移 (2)

関連キーワードを用いて再構成したクエリを用いて，再度 ROC グラフ表示画面で関連キーワードの導出を試みることができる．

遷移 (3)

ROC グラフで選択された関連キーワードの組合せによるすべての文書クラスタが先行して一括問合せされる．この問合せの数は，文書クラスタの数に等しい．

遷移 (4)

提示された文書クラスタから，必要な文書クラスタを選択し，さらに絞り込みを行うために，再び関連キーワードの導出を行うことができる．

遷移 (5)

ユーザが目的の情報を得ると検索は終了する．

この検索遷移のうち，**遷移 (3)** における先行一括問合せでは，キーワードの単純な組合せを考えると 2^n 通りのクエリがデータベースに対して発行されることになるが，データベースの規模を考えると，数十にも及ぶクエリを発行することは検索コストの増大をまねき現実的ではない．しかし，先行一括問合せで用いられる n 個の関連キーワードは，**ROC グラフ表示画面**において，検索ユーザ自身が任意に選ぶものであり，極端に多くの関連キーワードを選択することは考えにくい．また，選択されたすべての関連キーワードの AND/NOT による組合せによる検索式に対して，必ずしもすべてのクラスタに文書が含まれているとは限らない．したがって，キーワードの組合せに対して，存在するクラスタ数の傾向によってはクエリの発行数を減らすことができ，検索コストを低く抑えられる可能性が考えられる．

そこで，本稿では，**遷移 (3)** が特定の有限時間内に終了し，先行一括問合せに必要となる検索コストの見積りが可能であることを，大規模なデータベースを用いた評価実験により示し，提案する手法が検索支援システムに実装が可能な手法であることを示す．

3. 先行一括問合せ方式の性能評価

提案するシステムで検索コストが最も必要と考えられるのは，遷移 (3) で検索結果表示に先行して行う一括問合せである．しかし，関連キーワードの組合せによってできる文書クラスタの中には，該当文書が存在しない空の文書クラスタが存在する．そこで，これらの文書クラスタの数を調べることで，一括問合せの検索コストを評価する．

検索コストの見積りは，検索コストが最大となるようなクエリを用いて評価を行えば，検索コストがどの程度に抑えられるかを測ることができ可能となるが，本稿では，ROC 距離に基づいて関連キーワードを選択しクエリを構成する．ROC 距離の定義から，ROC 距離は TP が高く FP が低いものほど長くなる．TP が高いということは，クエリの被覆する検索空間に占める割合が大きいということを意味しており，その検索空間を 2^n のオーダで分割してクラスタを生成する可能性が高いと考えられる．また，クエリの被覆する検索空間がデータベース全体に占める割合がその規模と比較して小さいことを考慮すると，FP が高いということは，データベース全体に占める割合が大きいことを意味し，クエリの被覆する検索空間全体を覆う可能性が高くなり，クラスタが生成されにくくなると考えられる．したがって，ROC 距離に基づいてパフォーマンスの高い関連キーワードを選択すると，これらの語の組合せによる問合せは，最もコストが大きくなる問合せであると考えられる．なお，**可視化 1** を用いて関連キーワードのパフォーマンスを示すことにより，検索領域に対する知識を持たない検索ユーザは，検索を成立させるために，ROC グラフのパフォーマンスが高い位置，すなわち，ROC 距離が長くなる位置に配置されるキーワードを優先的に選択する可能性が高いと考えられる．

クエリに対する検索結果が 0 件であるか否か，すなわち，クラスタが存在するか否かは，インデックス検索により知ることができる．通常，インデックスは主記憶に置くことができるため，実際のデータベースアクセスに比較すると，そのコストは無視できる程度に収まる．また，クラスタの存在しないキーワードの組合せが判明すれば，それ以上のインデックス検索を行う必要もなくなる．したがって，先読み一括検索においては，クラスタ数が検索コストに大きく影響を与えると考えることができる．

3.1 サンプル抽出方法

本実験では，文献データベースに蓄積された IN-

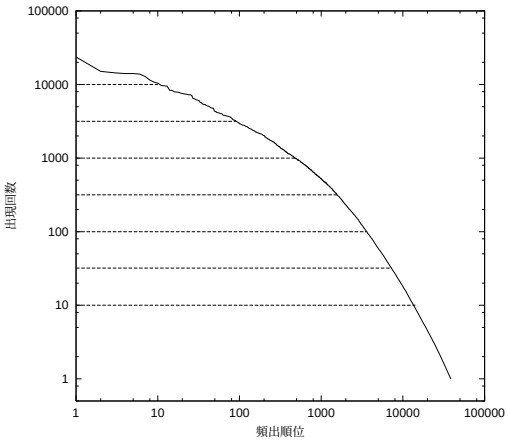

図 5 キーワード出現頻度
Fig. 5 Relationship between frequency order and frequency number.

SPEC データベースのうち 1998 年に配布された 331,504 件の文献情報を評価対象とした．テストデータに対して，タイトル部分で使用されているキーワードを調べると，頻出順位と出現回数の関係は**図 5** のようになっており，使用されているキーワードに大きな偏りが見られる．なお，キーワードとして意味のない “and” や “the” などの無意味語は，SMART システム[7] に用いられているストップワードリストを用いて除去してある．

出現頻度を考慮したサンプルを得るために，出現回数 10 回未満のキーワードを除外したキーワードについて，出現回数に応じて図 5 の対数縦軸上でほぼ均等に分かれるように，キーワードを**表 1** のようなカテゴリにクラス分けし，それぞれのカテゴリからサンプルを選んだ．表 1 の各カテゴリは，ちょうど図 5 において上から下に向かって順に区切られた部分のそれぞれに相当する．これらのカテゴリのうち，カテゴリ 1 ～ 4 に属すキーワードは，約 30 万件のデータベースに対して 1,000 件に 1 回以上は出現し出現頻度が比較的高いキーワードであり，そのようなキーワードは初期クエリで指定される機会が多いと考えられ，また，カテゴリに含まれるキーワード数も考慮すると，抽出するサンプル数をある程度確保する必要がある．そこで，各カテゴリからは，キーワード数 20 とそのカテゴリに含まれるキーワード数の約 1%の多い方の数のキーワードをサンプル抽出するように，表 1 の “サンプル数” 欄に示した数のキーワードを無作為抽出した．

3.2 検索コストが最大となるクエリの生成

こうして選び出した 197 語のサンプルを初期クエリとして，それぞれについて関連キーワードを導出した．

表 1 検索キーワードのカテゴリ
Table 1 Categories of query.

カテゴリ	出現回数			キーワード数	サンプル数
1	10,000	～		10	10
2	3,163	～	9,999	81	20
3	1,000	～	3,162	392	20
4	317	～	999	1,086	20
5	100	～	316	2,073	21
6	32	～	99	3,722	37
7	10	～	31	6,894	69
8		～	9	72,231	—

さらに，導出された関連キーワードから選択したキーワードを組み合わせてできるすべての検索式について問合せを行った．この各々の検索式により得られる文献集合が文書クラスタとなる．なお，キーワードの選択は，導出される関連キーワードのうち，ROC 距離に基づくパフォーマンスの高い上位 10 語から無作為にキーワードを選択し，その文書クラスタに文献が含まれるか否かを調査した．そして，Web 上の検索エンジンでユーザがクエリとして入力する索引語数の平均は 2.35 であるという調査結果[13] を参考に，ルール提示による索引語数の増加を考慮して，索引語数 1 から 7 までの範囲について調査を行った☆．

3.3 カテゴリごとの平均導出文書クラスタ数の評価

索引語である関連キーワード 3，5，7 個を組み合わせた場合について，それぞれ**図 6**，**図 7**，**図 8** に示す．これらの結果より，3～7 個のどのキーワード数においても，文献が実際に含まれるクラスタ数は出現順位の高いキーワードほど多い傾向があることが分かる．すなわち，文献情報を含む文書クラスタはカテゴリ 1 に属する語の関連キーワードの組に最も多く，カテゴリ順に少なくなり，カテゴリ 7 で最も少ない．なお，関連キーワード数 1～6 についても同様な傾向を確認している．

図 9 は，それぞれのカテゴリについて関連キーワード数 7 までの範囲における平均クラスタ数である．図 9 より，たとえば，関連キーワード数 7 の場合，平均文書クラスタ数が最大となるカテゴリ 1 では 41.6，最小となるカテゴリ 7 では 7.7 であることが分かる．つまり，関連キーワードの組合せの数である 128 通りの文書クラスタに対して，文献情報の含まれる文書クラスタの数は実際には少ないことが分かる．

3.4 平均導出文書クラスタ数の評価

図 9 から，組み合わせるキーワード数によらず，出

☆ インターネット上の検索エンジン Excite（http://www.excite.com）に入力される平均索引語数も，約 3 語程度であると報告されている．

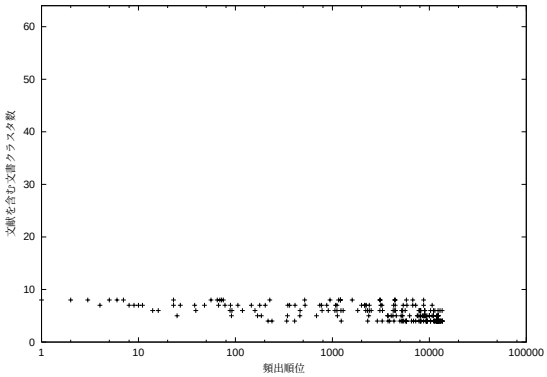

図 6 キーワード 3 語による文書クラスタ数
Fig. 6 Document cluster numbers from 3 associative keywords.

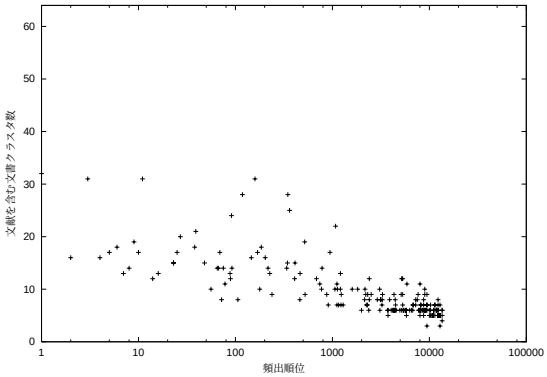

図 7 キーワード 5 語による文書クラスタ数
Fig. 7 Document cluster numbers from 5 associative keywords.

現頻度の最も高いカテゴリ 1 で最も多くの文献を含むクラスタが得られることが分かる．そして，カテゴリ 7 まで出現頻度の高いカテゴリ順に，クラスタ数も減っていくことが分かる．

表 2 および**図 10** は，カテゴリごとの結果を相加平均したものである．相加平均で評価するのは，初期クエリとして出現頻度が高い上位カテゴリに含まれるキーワードの選択比率を高く見積もり，また，出現頻度の高さからクラスタ数が多く得られる傾向のあるキーワードに比重を掛けることで，検索コストの見積りとして最悪の場合を見るためである．これらの結果を見ると，より多くの関連キーワードを組み合わせるほど空の文書クラスタ数の全体に占める比率は高くなることが分かる．つまり，実際に検索を行う必要のある検索式は，その組合せの数である $O(2^n)$ に対して，$O(n^2)$ あるいは $O(n \log n)$ 程度で済むことになる．

さらに，関連キーワードは**可視化 1** において検索ユーザにより選択されることから，選択される関連

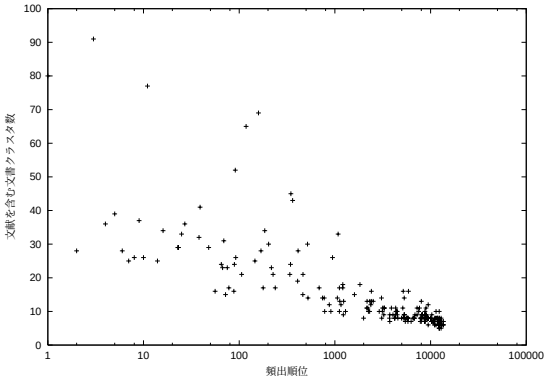

図 8 キーワード 7 語による文書クラスタ数
Fig. 8 Document cluster numbers from 7 associative keywords.

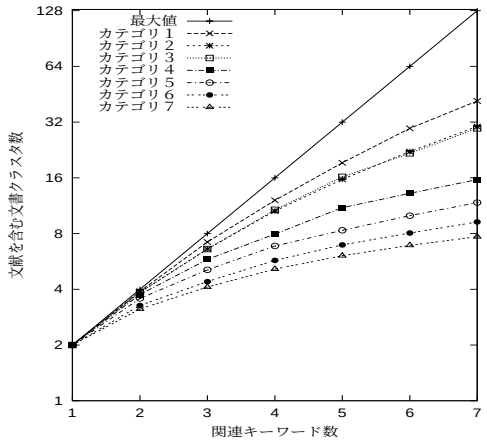

図 9 カテゴリごとの平均クラスタ数
Fig. 9 Average number of clusters containing some documents for each category.

キーワードは数語程度に抑えられると考えられ，この範囲内では文書クラスタの増加は線形 $O(n)$ の範囲内である☆．したがって，先行問合せによる検索コスト増は限られており，そのような検索システムは有効であると考えられる．

なお，検索領域に知識のあるユーザが本システムを用いた場合，ROC 距離に基づいたパフォーマンスのみに頼らず，パフォーマンスが低いキーワードであっても検索効率の良いキーワードを選択する可能性があるため，3 章冒頭で述べた理由から，導出される文書クラスタ数は $O(n)$ よりさらに少なくなると考えられる．

☆ $1 \leq n \leq 40$ の範囲に対して行った予備実験では，クラスタ数の増加は頭打ちとなっていた．

表 2 文献を含む平均文書クラスタ数
Table 2 Average number of document clusters containing some documents.

選択キーワード数	組合せ数	文献を含むクラスタ数
1	2	2.00
2	4	3.45
3	8	5.11
4	16	7.20
5	32	9.66
6	64	12.30
7	128	15.47

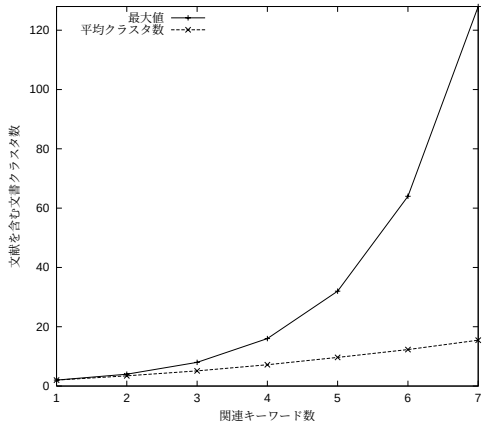

図 10 文献を含む平均文書クラスタ数
Fig. 10 Average number of document clusters containing some documents.

4. 可視化を用いた試作システムの構築

4.1 検索支援システム構成

我々の開発している「問答」は，全文検索システム OpenText を拡張した検索支援システムである．「問答」のシステム構成図を**図 11** に示す．検索支援システムは，(a) 検索支援インタフェース，(b) アプリケーションサーバ，(c) 文献情報データベースからなるクライアントサーバシステムである．データベースサーバとして富士通 GP7000/900（CPU：450 MHz SPARC64×8，メモリ：8 GB，OS：Solaris 7）上の OpenText を用い，アプリケーションサーバとして PC/AT 互換機（CPU：733 MHz Pentium III × 2，メモリ：512 MB，OS：FreeBSD 4.0）を用いた．アプリケーションサーバは，関連キーワードの導出部分と文書クラスタ導出部分から成っている．そして，検索支援インタフェースは Java アプレットを用いて構築してある．

4.2 ROC グラフ表示インタフェース

図 12 は初期クエリ “information” に対して，導出

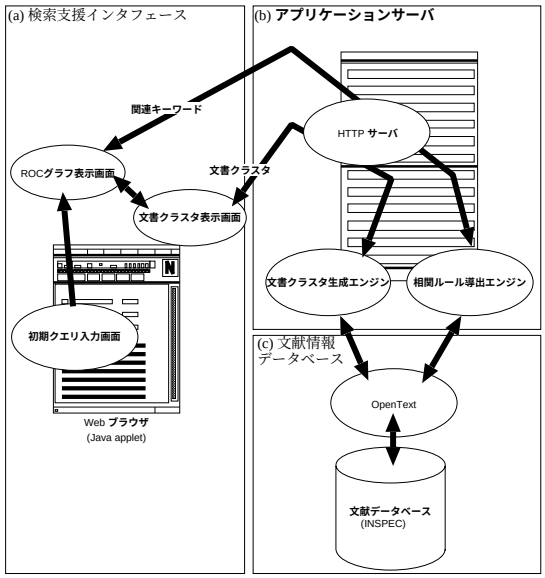

図 11 文献検索支援システムの構成
Fig. 11 Search engine system outline.

キーワード “systems”，“based”，“development” などが三次元表示された例である[14),15)]．なお，ROC グラフに，検索空間における出現頻度により重み付けされたサポート値[9)] の軸を加え，三次元表示へ拡張をしている．軸上で値が高いほど重要度の高いキーワードであり，同様の (FP,TP) を持つ導出キーワードであっても，サポート値が異なることにより表示位置がずれる．また，グラフの回転，拡大，縮小機能と併用することにより，サポート軸を持つ ROC 空間に数多くのキーワードが導出される場合でも判読可能となる．全体に対する位置の把握を容易にするためには，フィッシュアイ型の非線型ズーミングによる拡大を用いた．さらに，各関連キーワードの詳細なパラメータを見るためにマウスポインタによる TIPS の表示を行う．

4.3 文書クラスタ表示インタフェース

図 13 は，キーワード “information” から導出された関連キーワードのうち，“systems”，“based”，“development”，“data”，“application”，“design” の 6 個の選択キーワードに関する文書クラスタを表示した実行例である．素数多角形の各頂点に VIBE システムと同様に関連キーワードを配置した[14),15)]．また，文書クラスタを表すシンボルとして，文書クラスタに含まれる関連キーワード数を用い，その数字を多角形の周囲に配置した．この数字にマウスのポインタが触れると，その文書クラスタの文献リストが右側のウィンドウに表示される．現在選択されているクラスタの数字は四角で囲まれる．図 13 では，右上の 3 が選択され

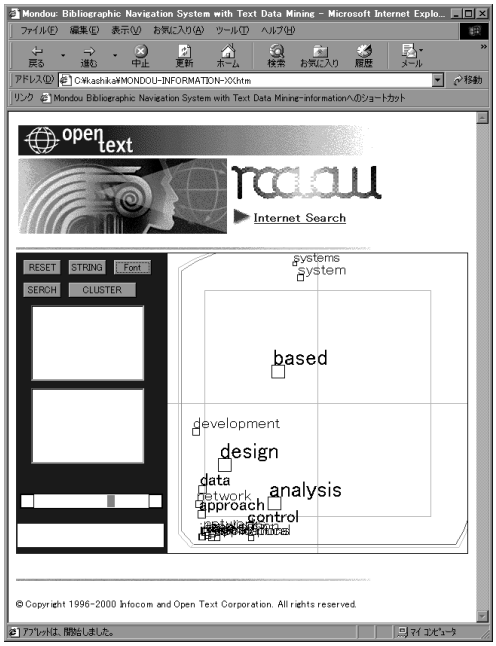

図 12 ROC グラフ表示画面
Fig. 12 ROC graph display screen.

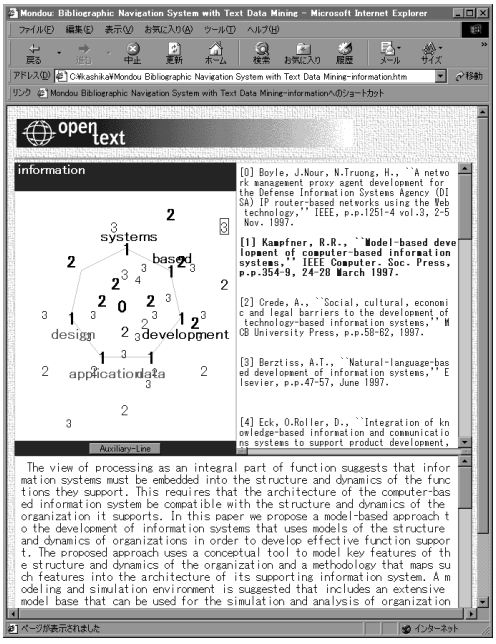

図 13 文書クラスタ表示画面
Fig. 13 Document cluster display screen.

た文書クラスタであり，これは “systems”，“based”，“development” の 3 語を含むクラスタである．この状態で数字をクリックすると，該当する各文献のアブストラクトが画面下に表示される．さらに，□をクリックすると，選択した文書クラスタを生成するクエリから，関連キーワードの導出を行い，その ROC グラフを表示することができる（図 4 の遷移 (4)）．また，文献が実際には含まれないクラスタの数字は，目立たないように表示し，触れても反応しない実装としている．

5. むすび

相関ルール提示を行う文献検索支援システム構築において，導出される関連キーワードおよび検索結果をどのような手法を用いて提示するかは重要な問題である．本稿では，検索式生成の困難さを緩和するために情報可視化技術を用いたインタフェースの構築と先行一括問合せ処理の性能評価を行った．そして，検索結果を一括表示するうえでボトルネックとなる先行問合せ処理のコストが実用性の高い範囲に収まることが分かった．すなわち，最もパフォーマンスが高い関連キーワードを用いた場合でも，一括問合せのコストは $O(n)$ 程度という結果である．また，プロトタイプシステムを構築し，システムが動作することも確認した．

今後，さらに多くの選択キーワード数を用いた場合のクラスタ数の解析や，Web ページなどの他の半構造データへの適用，複数クライアントによるアクセス負荷の解析や実験評価を行う予定である．

謝辞 本稿の一部は，文部省科学研究費（12792015, 1380482, 11130211, 12780278）の研究成果による．また，本稿に対して貴重かつ有益なご指摘ならびにコメントをいただいた査読委員の方々に感謝する．

参考文献

1) Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, *Proc. 20th International Conference on Very Large Data Bases*, Santiago, Chile (1994).
2) Chaomei, C.: *Information visualisation and virtual environments*, Springer-Verlag (1999).
3) Kawahara, M. and Kawano, H.: Performance Evaluation and Visualization of Association Rules using Receiver Operating Characteristic Graph, *Proc. 1999 International Symposium on DANTE*, Kyoto, Japan, pp.334–342 (1999).
4) Kawahara, M. and Kawano, H.: The Other Thresholds in the Mining Association Algorithm, *Proc. 14th International Conference on Systems Engineering* (*ICSE2000*), Coventry, UK, Vol.1, pp.305–310 (2000).
5) Korfhage, R.: To see, or not to see—is that the query?, *Proc. 14th ACM/SIGIR*, Chicago, IL USA, pp.134–141 (1991).
6) Provost, F. and Fawcett, T.: Analysis and Vi-

sualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions, *Proc. 3rd Int'l Conference on Knowledge Discovery and Data Mining* (*KDD-97*), pp.43–48 (1997).
7) Salton, G. and McGill, M.J.: *Introduction to modern information retrieval*, McGraw-Hill, New York, USA (1983).
8) Spoerri, A.: Visual tools for information retrieval, *Proc. 1993 IEEE/CS* (*VL'93*), pp.160–168 (1993).
9) 河野浩之，長谷川利治：WWW 情報空間における文書データマイニングを用いた知的検索システム，アドバンストデータベースシンポジウム，東京，pp.27–34 (1996).
10) 川原 稔，河野浩之：相関ルール実体化を行う文献情報検索支援システムの性能評価，電子情報通信学会論文誌，Vol.J82-D-I, No.1, pp.165–173 (1999).
11) 川原 稔，河野浩之：文献情報検索支援システムの ROC 解析による相関ルール選択基準，情報処理学会論文誌データベース，Vol.40, No.SIG3 (TOD1), pp.105–113 (1999).
12) 川原 稔，河野浩之，長谷川利治：文献データベース情報検索に対するデータマイニング技術の適用，情報処理学会論文誌，Vol.39, No.4, pp.878–887 (1998).
13) 徳永健伸：情報検索と言語処理，東京大学出版会 (1999).
14) 野村 賢，河野浩之，川原 稔：文献検索支援における可視化手法の提案とその評価，データ工学研究会，夏の DB ワークショップ，岩手，pp.31–38 (2000).
15) 野村 賢，河野浩之，川原 稔：文献検索支援システムにおける可視化インターフェイスの構築と性能評価，第 14 回人工知能学会全国大会，東京，pp.406–407 (2000).

(平成 12 年 12 月 20 日受付)
(平成 13 年 3 月 31 日採録)

（担当編集委員 江口 浩二）

野村 賢

平成 11 年京都工芸繊維大学工芸学部電子情報工学科卒業．同年 4 月京都大学大学院情報学研究科システム科学専攻入学，現在に至る．情報可視化，データマイニングの研究に興味を持つ．人工知能学会会員．

河野 浩之（正会員）

昭和 60 年京都大学工学部数理工学科卒業．平成 2 年同大学大学院工学研究科数理工学専攻博士課程研究指導認定退学．同年 4 月同大学工学部数理工学教室助手．同時に応用システム科学教室助手を兼任．平成 9 年 10 月同大学院工学研究科応用システム科学専攻助教授．平成 10 年 4 月同大学院情報学研究科応用システム科学専攻助教授，現在に至る．平成 5 年カナダ・サイモンフレーザー大学客員研究員．平成 12 年カリフォルニア大学アーバイン校客員研究員．工学博士．情報伝送システム，データベースシステムの研究に興味を持つ．ACM，IEEE，AAAI，電子情報通信学会，人工知能学会各会員．

川原 稔（正会員）

昭和 63 年早稲田大学理工学部電気工学科卒業．平成 2 年京都大学大学院工学研究科応用システム科学専攻修士課程修了．同年 4 月同大学大型計算機センター助手．平成 7 年 4 月同大学院工学研究科応用システム科学専攻助手兼任．平成 10 年 4 月同大学院情報学研究科システム科学専攻助手兼任．データベースシステム，データマイニングの研究に興味を持つ．IEEE，人工知能学会各会員．