データベースシステム　122－53
(2000. 7. 28)

文献検索支援における可視化手法の提案とその評価

検索支援可視化インタフェースの実装

野村 賢
京都大学大学院情報学研究科
京都市左京区吉田本町
075-753-7420
nomura@ais.sys.i.kyoto-u.ac.jp

河野 浩之
京都大学大学院情報学研究科
京都市左京区吉田本町
075-753-5493
kawano@i.kyoto-u.ac.jp

川原 稔
京都大学大型計算機センター
京都市左京区吉田本町
075-753-7429
kawahara@kudpc.kyoto-u.ac.jp

あらまし　大量に蓄積されたSGMLなどの半構造データの効率的な検索支援を行うデータマイニング技術を応用した「問答」を構築してきた．その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づらいといった問題が生じた．そこで，本稿では，ROCグラフ上に，検索式と導出ルールの関係を可視化する手法を提案する．さらに，導出されたキーワードから適切なキーワード選択による検索式改善を行えるように，導出キーワードと対象データとの関係をクラスタリング技術を用いて可視化する．そして，INSPECデータベースを用いて，可視化インタフェースをもつ，文献検索支援システムの性能評価を行う．

キーワード　ROCグラフ，情報検索，データマイニング，ビジュアライゼーション，文書クラスタリング，相関ルール

Proposal and Evaluation of Information Visualization on Bibliographic Navigation System

Ken Nomura
Department of Systems Science
Kyoto University
+81 75 753 7420
nomura@ais.sys.i.kyoto-u.ac.jp

Hiroyuki Kawano
Department of Systems Science
Kyoto University
+81 75 753 5493
kawano@i.kyoto-u.ac.jp

Minoru Kawahara
Data Processing Center
Kyoto University
+81 75 753 7429
kawahara@kudpc.kyoto-u.ac.jp

Abstract　In order to retrieve a huge volume of semi-structured data like SGML efficiently, we develop Java applets for our information navigation system, **Mondou**. The applet visualizes ROC graph and document clusters. The ROC graph shows the relationship between a query and associative keywords derived by mining algorithm. On the other hand, the interface of document clustering is based on the extension of the VIBE system, and it presents the combination of retrieved keywords efficiently. Moreover, we evaluate the performance of our navigation system using INSPEC database.

key words　ROC graph, information search, data mining, visualization, document clustering, association rule

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には，データマイニング (data mining)，もしくはデータベースからの知識発見 (KDD: Knowledge Discovery in Database) に関する研究をはじめ，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，このような性質を持つデータの処理を効率良く行うためのシステム設計及びスキーマ設計に関する技術が必要となる．つまり，知識発見の源泉となるデータベースシステムに対して，どのようなアルゴリズムを実装し，また，設計制約をどの様に緩和するべきかを明確にするための研究が要求される．

このようにしてデータマイニング技術が活発に研究されるなか，電子化データのなかでも特にテキストデータに対しデータマイニング技術を適用するテキストマイニングという分野が形成されつつある．例えば，図書・文献に関わるデータベースを用いた情報検索では，一般に検索領域に対する領域知識に加えて，検索システムに慣れること（背景知識）が必要であるため，スムーズに検索を行うために熟練した図書館司書の支援に頼ることも多い．そして，このような困難さを解消あるいは緩和するために，様々な情報検索システム構築に関わる研究が数多く行われてきている．また，Web データといった膨大な半構造テキストデータからの情報検索では，AltaVista(www.altavista.com)，google(www.google.com)[1]，MetaCrawler(www.metacrawler.com) など多くの研究・開発がある．

我々の開発している Mondou(www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/) は相関ルールを用い検索キーワードを提示する検索支援システムである [3]．しかし，この従来の研究では検索式と導出ルールである関連キーワードの関係を直感的に把握することが難しいという問題がある．そこで，本稿では，この問題を解決するための可視化インタフェースの提案及び構築を行う．

まず，2 章では，検索システムにおける一般的な問題点，及び，“Mondou” において解決すべき点について述べる．続く 3 章では，その解決に用いた情報可視化技術について述べる．4 章では，我々の開発した可視化インタフェースを用いた検索過程について述べ，5 章では，検索支援システムの構成を示す．また，6 章では ROC グラフによる可視化について，7 章では文書クラスタの可視化について，それぞれの手法の提案，及び，Java アプレットを用いた実装について述べる．

そして，8 章では，従来のシステムと比較を行うとともに，INSPEC データベースを用いての検索支援可視化インタフェースの評価を行う．

9 章は，むすびとし，結論と今後の課題を述べる．

2. 検索支援システム Mondou

WWW(World Wide Web) は，多様なデータ構造を用いた自由度の高い情報発信を可能とするシステムであり，インターネット上での普及が目覚ましい [2]．それにともなって，インターネット・ユーザも増加し，電子化された文書やデータを WWW を代表とするシステム上で検索する機会も増えてきた．こうして WWW データが広く利用される一方で，データの新規作成，更新，消滅が頻繁に発生することもあり，用語統制されていないデータが，あふれつつある．

このような大量の Web ページの中から，目的とする情報を含む文書クラスタを抽出するために，我々の開発した Mondou ではデータマイニング手法の一つである相関ルール (association rule) の解析により，文書クラスタをもとめ検索支援を行っている [6]．このシステムでは，導出されたルールである関連キーワードを，検索ユーザへの知識としてフィードバックした．1996 年当時，情報の可視化を行うインタフェースを Java アプレットで構築し，図 1 の X 軸には，Web ページのスコアを，Y 軸には Web ページへのアクセスに要する時間を表した．図中の 1 つの正方形が 1 つのコンテンツを表し．文書サイズが面積に相当する．また，

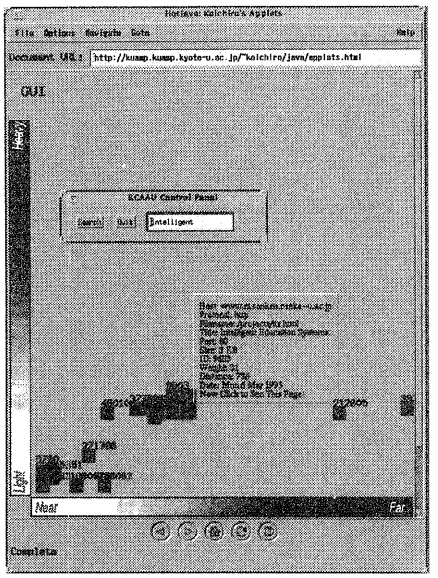

図 1　従来の Mondou による検索結果の可視化アプレット

該当する URL 等のデータを TIPS 表示した．

しかし，現在までに提案した検索インタフェースでは，導出した関連キーワードをリスト形式で列挙するにとどめるため，検索式と関連キーワードの関係は明確ではない．このため，検索ユーザは，適当に関連キーワードから検索キーワードを選択し，試行錯誤によって検索式を作り上げなければならない．また，検索結果もリスト形式で表示されているため，導出された関連キーワードと文献情報の関係を直感的に理解することも困難である．

3. 情報可視化

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 [1]．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．こうして提案された様々な情報可視化手法によって理解の難しい構造や特徴を視覚的効果により直感的に理解できる形で示すことが可能である．つまり，可視化の成功とは隠れた構造・特徴を，ユーザに理解させることであり，失敗とは逆に特徴を隠してしまうことである．

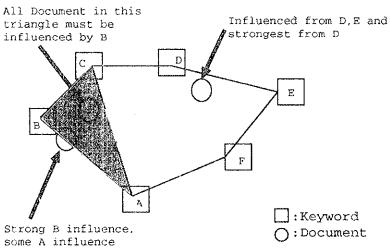

図 2　VIBE システムの概念図

3.1 focus versus contex 問題

情報可視化において，現在参照中のコンテンツのデータ構造全体における位置づけ (context) を見たいという要求と，現在参照中のコンテンツの詳細 (focus) を見たいという相反する 2 つの要求がある．この問題を “focus versus context” 問題と呼ぶ．[1]

多くの解決法が提案されているが，例えば，FishEye View[1] によるズーミングといった非線型な表示手法を活用するものがある．$H3^2$ - three-dimensional hyperbolic visualization [8] [9] では，ツリー構造をハイパボリック空間へマップすることで，表示の混雑を抑え，レイアウトスピードを改善し，巨大なグラフへのアクセスを可能としている．

3.2 VIBE システム

VIBE(Visual Information Browsing Environment) システム [7] (図 2) は文献と関連キーワードの関係を可視化するシステムである．多角形の各頂点に関連キーワードを配置し，その周囲に文献が配置される．文献の位置は各関連キーワードとの関連度により決定する．

関連キーワードと文献の関係がわかりにくい場合は，多角形の頂点を動かすことで文献の位置を変化させ，微妙な関係を理解することができる．

4. 文献検索支援可視化インタフェースの提案

従来のシステムでは導出される関連キーワードをリスト表示しており [4]，検索キーワードと導出ルールの関係を直感的に理解できる形で示していない．そこで，検索キーワードと求められたルールの関係を表す ROC(Receiver Operating Characteristic) グラフ [3] による可視化を行う．さらに，ルールである関連キーワー

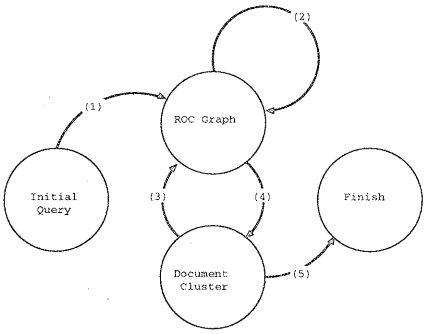

図 3　検索遷移図

ドと検索結果の関係の可視化を行う．

本稿で提案する検索支援インタフェースを用いた検索の過程を状態遷移図として図3に与える．ここで，ブラウザ上に表示される画面は初期検索式入力画面，ROCグラフ表示画面，文書クラスタ表示画面の3つからなる．

検索は初期検索式入力画面からはじまる．この画面で，検索ユーザが初期検索式を入力することによって，関連キーワードがROCグラフ表示される (遷移 1)．ROCグラフ表示画面では，ROCグラフ上に初期検索式から導出される関連キーワードが配置される．検索ユーザはここで，必要な関連キーワードを選択して検索式の改善を行い，再度，関連キーワードの導出を試みることができる (遷移 2)．また，関連キーワードと文献の関係を確かめるために，文書クラスタを生成することも可能である (遷移 3)．ここでは，関連キーワードの組み合わせによってできるクラスタを可視化し，検索ユーザが文書クラスタを選択することによって，その文書クラスタに含まれる文献のリストを表示する．また，得られた文書クラスタに対して絞り込みを行うため，さらに関連キーワードを導出しROCグラフの表示もできる (遷移 4)．そして，目的とする文書クラスタが確認できたなら，検索作業は終了する (遷移 5)．

5. 文献検索支援システムの構成

我々の開発する Mondou は全文検索データベース OpenText を拡張した検索支援システムである．Mondou のシステム構成図を，図4に与える．検索支援システムは，(a)Java アプレットを用いて構築した検索支援インタフェース，

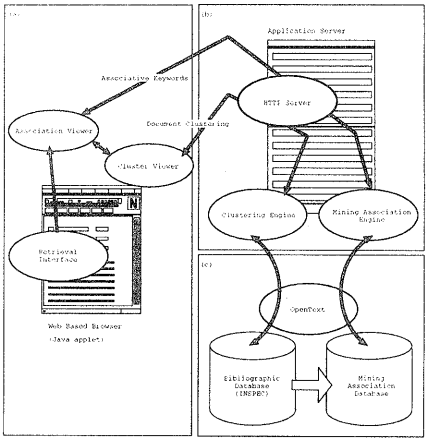

図 4　文献検索支援システムの構成

(b) アプリケーションサーバ，及び (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:300MHz SPARC64×8，メモリ:8GB，Solaris 8) 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:750MHz Pentium III × 2，メモリ:512MB，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．本実験では，文献データベースに INSPEC データベース[1] を用い，1998年に配布された 331,504 件の文献情報を評価対象とした．なお，INSPEC の文献二次情報データは，Web データとは異なってノイズの小さいなデータであり，基本的な性能を評価するのに適している．

6. 関連キーワードの可視化

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．以下，ROC空間におけるデータの可視化について簡単に述べる．なお，詳しい議論は文献 [3][5] において行った．

[1] INSPEC データベースは，英国 IEE からの独立組織である INSPEC が，文献の収集・整理を行い全世界に配布している理工学系の代表的な文献二次情報であり，計算機・制御・情報工学・電子・電気工学・物理学の分野における文献データベースである．

6.1 ROC空間

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，その事象に対する分類子による分類を，“正:y(yes)”と“負:n(no)”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，及び，負の事象Nが誤って正yと分類される比率FPは，それぞれ，事象Pが分類yとなる事後確率$p(y \mid P)$，および，事象Nが分類yとなる事後確率$p(y \mid N)$を用い，次のように表すことができる．

$$TP = p(y \mid P) \simeq \frac{\text{正であると分類された正の事象}}{\text{すべての正の事象}}$$

$$FP = p(y \mid N) \simeq \frac{\text{正であると分類された負の事象}}{\text{すべての負の事象}}$$

いくつかの事象Iに対して，FP値をX軸の値，TP値をY軸としてプロットするとROCカーブ(図5)と呼ばれるグラフが描かれ，これを分類子のパフォーマンスを表すのに用いる．ROCグラフでは，グラフが左上端に近づくほど，すなわち，TP値がより高くなるほど，分類子により事象が正確に分類されたことになる．逆に，グラフが右側に近づくほど，すなわちFP値がより高くなるほど，分類子による分類にノイズが入ってくることになる．従って，TPがより高くFPがより低い点の方，つまり，左上端にROCグラフが近づくように描かれるほど，よりパフォーマンスが高いといえる．例えば，図5において，分類子Aは分類子Dより常に左上に存在しているので，分類子Aの方がよりパフォーマンスが高いことになる．

6.2 ROCグラフを用いたキーワードの可視化

導出された関連キーワードのTP(True Positive rate)およびFP(False Positive rate)により，それらのキーワードをROCグラフ上に配置[2]する．

なお，導出キーワードをROCグラフ上に配置する場合における問題点として，ROCグラフ上で近い位置に配置されるキーワード同士が重なり，判読しづらくなることがある．そのため，3次元表示など，各種ビジュアライゼーション手法を用いて，表示空間の有効利用を図る必要がある[5].

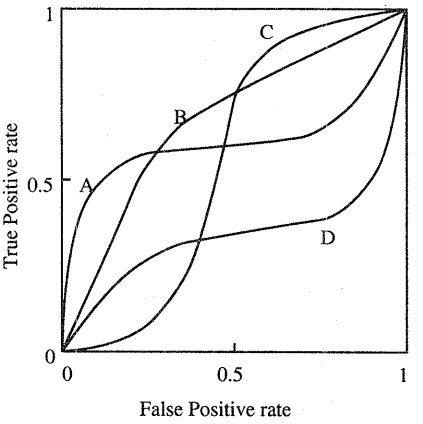

図5 ROCグラフ

そこで，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[6]の軸を加え，3次元表示への拡張を行う．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って，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．拡大には，全体に対する位置の把握を容易とするためにFishEyeタイプの非線型ズーミングを採用する．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．

図6は初期検索キーワード“mediator”に対し，導出キーワード“information”，“data”，“distributed”等が3次元表示された例となっている．

7. 文書クラスタの可視化

検索キーワード“mediator”によって得られる文書クラスタ全体は，導出される関連キーワード“information”，“data”，“distributed”の3つによって8つの文書クラスタに分割される(図7)．すなわち関連キーワードの数nに対

[2] 検索要求キーワードにより被覆される文献の集合をB，検索要求キーワードにより導出されるある1つのキーワードが被覆する文献の集合をRとし，N(x)を文献集合xの文件数を求める演算子と定義すると，TP及びFPは，$TP = \frac{N(B \cap R)}{N(B)}$, $FP = \frac{N(\bar{B} \cap R)}{N(\bar{B})}$で表すことができる．

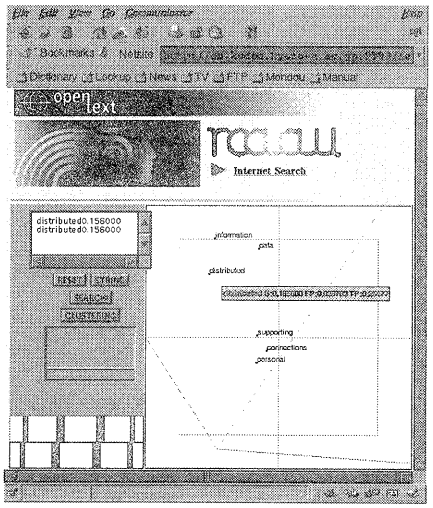

図 6　3 次元拡張した ROC グラフ

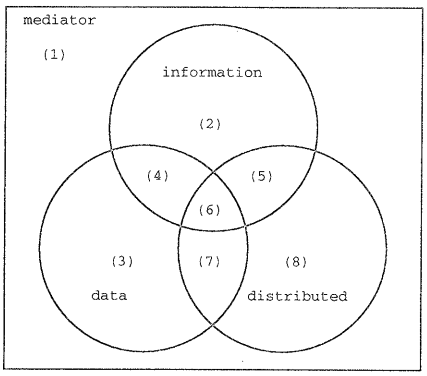

図 7　検索式の組み合わせ

し 2^n 個の文書クラスタが存在する．ここで，n 以上で最小の素数 p による p 角形により，この全ての文書クラスタの一つの表示領域内への配置を試みる．

まず，素数多角形の各頂点には VIBE システムと同様に関連キーワードを配置する．そして，文書クラスタを，p 角形の重心-頂点間ベクトルの合成により配置する．例えば検索キーワード “information”，“data”，“distributed”，“personal” の内 “information”，“data” を含み “distributed”，“personal” を含まないクラスタの配置位置は図 8 の丸印 (1) の位置となる．なお，素数多角形を使用することにより各文書クラスタは互いに重なることなく，表示領域内に効率的に配置できる．ただし，全ての関連キーワードを含む文書クラスタと全て含まないクラスタは同位置に配置されるため，どちらかを中心からずらす必要がある．

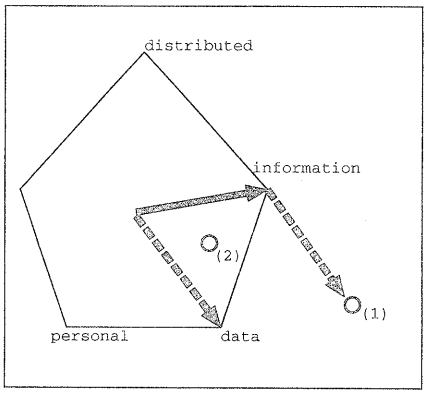

図 8　クラスタの表示位置の決定法

他の素数多角形を利用した文書クラスタの配置方法として，クラスタに含まれる関連キーワードの頂点を結んでできる多角形の重心に文書クラスタを配置する方法を考えることもできる．この場合，図 8 の丸印 (2) の位置に配置されることになる．しかし，この方法では一つ一つの文書クラスタ表示位置が直感的にわかりやすい反面，その配置は中心付近に偏るといった問題が生じる．多くの文書クラスタを配置した場合，個々の文書クラスタの判別が困難であることも多い．そこで，実装では比較的一様に配置できる前述のベクトル方式を用いた．

図 9 は，“mediator” による検索結果のうち，“distributed”，“informaion” ，“data” の 3 個の関連キーワードのみを含む文献情報クラスタが表示した例である．多角形の周囲に文書クラスタを表す数字が配置されている．数字は文書クラスタに含まれる関連キーワード数を表している．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 9 右上の 3). この状態で数字をクリックすると，選択した文書クラスタに対し，関連キーワードの導出が行われ，その ROC グラフが表示される (図 3 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．

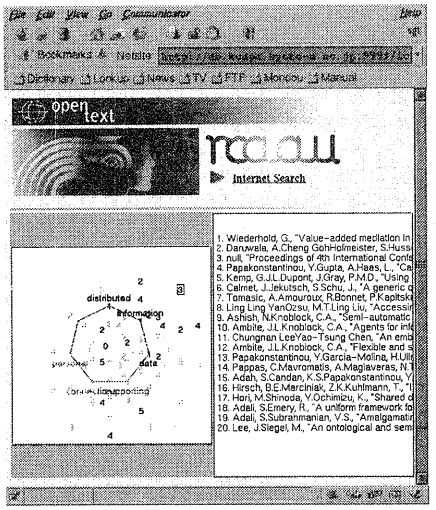

図 9　文書クラスタの表示

8. 性能評価

関連キーワードの組み合わせによってできる文書クラスタの中には実際には文献の含まれていない空の文書クラスタが存在する．従来の検索システムでは，検索結果としてこのような空の文書クラスタが導出され，そのため再検索しなければならないこともあった．本インタフェースでは，全ての文書クラスタを一括表示するので，検索ユーザは，このような再検索を強いられることがない．

従って，文献が含まれない文書クラスタ数が多いほど本インタフェースによる検索は効率が良くなると考えられる．そこで，実際に文献の含まれない文書クラスタ数を INSPEC データベースをテストデータに用い調べた．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，縦軸に出現回数，横軸を頻出順位として，図 10 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た．ただし，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[10] システムに用いられているストップワード辞書を用いて除去している．

適切なサンプルを得るために，出現回数に応じて図 10 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

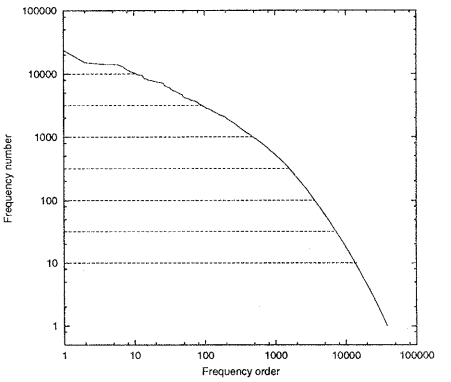

図 10　キーワード出現頻度

表 1　検索キーワードのカテゴリ

カテゴリ	出現回数			キーワード数	サンプル数
1	10000	～		10	10
2	3163	～	9999	81	20
3	1000	～	3162	392	20
4	317	～	999	1086	20
5	100	～	316	2073	21
6	32	～	99	3722	37
7	10	～	31	6894	69
8		～	9	72231	—

クラス分けして，それぞれのカテゴリからサンプルを取る．表 1 のカテゴリは，ちょうど図 10 において上から下に向かって順に区切られたそれぞれの部分に相当する．各カテゴリからは，カテゴリに含まれるキーワード数の約 1 %以上のキーワードをサンプリングするように，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．

こうして選びだした 197 語のサンプルを初期検索キーワードとして，それぞれについて関連キーワードを導出した．さらに，この関連キーワードの中でも ROC パフォーマンス [3] の高い語を無作為に組み合わせてできる文書クラスタのうち，実際に文献の含まれる文書クラスタの数を調べた．関連キーワード 7 つを組み合わせた場合について図 11 に示す．この場合の平均クラスタ数は 15.36 であり，127 通り全てのクラスタを生成したとしても，実際に検索にかかるコストは 10 分の 1 程度である．同様にして，組み合わせるキーワード数 2～6 の場合について，その平均を調べた結果を表 2 に示す．また，そのグラフを図 12 に示す．組み合わせた関連キーワード数が 2 語の場合には，ほとんどの文書クラスタに文献が存在するが，それ以上の関連キーワードを組み合わせた場合には，

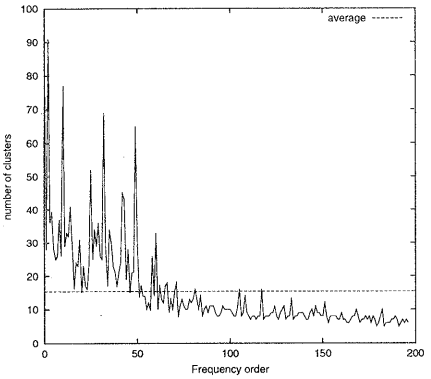

図 11　キーワード 7 語による文書クラスタ数

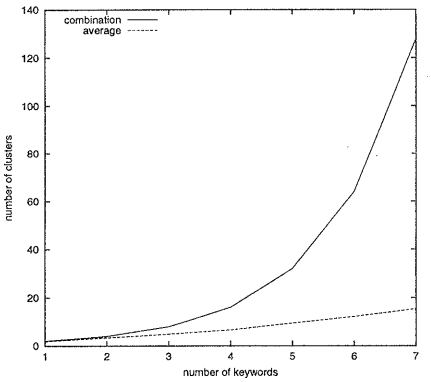

図 12　文献の含まれる文書クラスタ数

表 2　文献を含むクラスタ数

選択キーワード数	組み合わせ数	文献を含むクラスタ数
1	2	1.87
2	4	3.35
3	8	4.88
4	16	6.61
5	32	9.51
6	64	12.24
7	128	15.36

実際に文献の含まれる文書クラスタは，その組み合わせの数である 2^n よりかなり少ないことがわかる．例えば関連キーワード数7語の場合，従来の検索方法では，128 通りのなかから目的の文書クラスタを探していたのに対し，提案法では平均 15.36 通りの中から文書クラスタを探すこととなる．従って，明らかに検索効率は良くなるといえる．

9. むすび

情報可視化技術を用いた検索支援インタフェースの提案と構築を行った．まず，ROC グラフを用いることで検索キーワードと導出ルールの関係を視覚的に判断することを可能とし，検索ユーザが効果的なキーワードを容易に選択することができるようになった．また，文書クラスタの一括表示を行うことにより，何通りもの検索式を作成し再検索するという反復作業と，文献の含まれていない文書クラスタによる再検索を解消した．また，文献の含まれていない文書クラスタ数を調べ，評価した．従って，本インタフェースの使用により検索操作性は良くなると判断できる．今後，Web などの一般的なデータへの適用，複数クライアントによるアクセス負荷の解析や実験を行う必要がある．

謝辞　本稿の一部は，文部省科学研究費 (11130211, 12780278) の研究成果による．本インタフェースの構築をサポートした，京都大学大学院情報学研究科の安村賢英氏に感謝する．

参考文献

[1] C. Chaomei, "Information Visualisation and Virtual Environments," Springer-Verlag, 1999.

[2] M. Gray, "Growth of the World Wide Web," http://www.mit.edu:8001.

[3] 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, Vol.40 No.SIG3(TOD 1), 1999.

[4] 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.165-166, 1999.

[5] M. Kawahara and H. Kawano, "Performance evaluation and visualization of association rules using receiver operating characteristic graph," Proc. of DANTE'99, Japan, pp.334-342, 1999.

[6] 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," Proc. of ADBS'96, Japan, pp.27-34, 1996.

[7] R. R. Korfhage, "To see, or not to see- is that the query?," Proc. of 14th ACM/SIGIR, Chicago, IL USA, pp.134-141, 1991.

[8] T. Munzner, "Drawing Large Graphs with H3 Viewer and site Manager," Proc. of Graph Drawing 98, Montreal, Canada, Springer-Verlag, London, UK, 1998.
(http://graphics.stanford.edu/papers/h3draw/)

[9] T. Munzner, "Exploring large graphs in 3D hyperbolic space," IEEE Computer Graphics and Applications, pp.18-23, 1998.

[10] G. Salton and M. J. McGill, "Introduction to modern information retrieval," McGraw-Hill, New York, USA, 1983.