検索支援と分析のための多面的検索システム

関, 隆宏
九州大学大学評価情報室

和多, 太樹
九州大学大学院システム情報科学府

山田, 泰寛
九州大学ユーザーサイエンス機構

廣川, 佐千男
九州大学情報基盤センター

出版情報：電子情報通信学会 第18回データ工学ワークショップ論文集. E1-2, 2007
バージョン：

検索支援と分析のための多面的検索システム

関　隆宏† 和多　太樹†† 山田　泰寛††† 廣川佐千男††††

† 九州大学大学評価情報室　〒 812–8581 福岡市東区箱崎 6–10–1
†† 九州大学大学院システム情報科学府　〒 819–0395 福岡市西区大字元岡 744
††† 九州大学ユーザーサイエンス機構　〒 812–8581 福岡市東区箱崎 6–10–1
†††† 九州大学情報基盤センター　〒 812–8581 福岡市東区箱崎 6–10–1
E-mail: †tsekiuoc@mbox.nc.kyushu-u.ac.jp, ††t-wada@i.kyushu-u.ac.jp, †††yamada@usi.kyushu-u.ac.jp,
††††hirokawa@cc.kyushu-u.ac.jp

あらまし　検索における問題点として，適切な検索キーワードの選択が難しいため，一回の検索で所望の検索結果が得られるとは限らないことが挙げられる．ユーザは検索結果を見ることにより新たな検索キーワードを獲得し検索を繰り返すことになる．したがって効率よく検索を繰り返す機構が重要である．我々は半構造化文書群を対象として，その部分構造ごとに着目した多面的検索システムを開発し，文献情報等の実データに対して実装と実験を行ってきた．このシステムでは部分構造ごとの検索やクラスタリングさらにはクラスタの特徴語抽出を通じて，ある部分構造に対する検索結果を，ユーザが指定した 2 つの部分構造に関するマトリクスとして表示する．この利用により，検索支援と検索結果の分析が可能になる．

キーワード　多面的検索システム，検索支援，クラスタリング，特徴語抽出，半構造化文書

A Faceted Search Engine for Supporting Information Retrieval and Analysis

Takahiro SEKI†, Taiki WADA††, Yasuhiro YAMADA†††, and Sachio HIROKAWA††††

† Office for Information of University Evaluation, Kyushu University
Hakozaki 6–10–1, Higashi-ku, Fukuoka, 812–8581 Japan
†† Graduate School of Information Science and Electrical Engineering, Kyushu University
Motooka 744, Nishi-ku, Fukuoka, 819–0395 Japan
††† Kyushu University User Science Institute　Hakozaki 6–10–1, Higashi-ku, Fukuoka, 812–8581 Japan
†††† Computing and Communications Center, Kyushu University
Hakozaki 6–10–1, Higashi-ku, Fukuoka, 812–8581 Japan
E-mail: †tsekiuoc@mbox.nc.kyushu-u.ac.jp, ††t-wada@i.kyushu-u.ac.jp, †††yamada@usi.kyushu-u.ac.jp,
††††hirokawa@cc.kyushu-u.ac.jp

Abstract　The faceted search engine presented herein retrieves the semi-structured documents of the indicated search area, and a matrix of the distribution of the clustering from two aspects of the retrieval result is displayed. In addition, the characteristic words of each cluster are displayed in the matrix. The faceted search engine provides a visual and semantic bird's-eye view of the entire retrieval result. Viewing characteristic words makes the support of the narrowing and expansion possible. Furthermore, it is possible to immediately change the analysis criteria or the number of clusters and to use a zooming function. Thus, various retrieval conditions for a query can be attempted immediately and continuously.

Key words　Faceted Search Engine, Supporting Information Retrieval, Clustering, Semi-structured Document

1. はじめに

Web の発展によりさまざまな人がさまざまな情報や文書を電子的に公開するようになり，我々が扱える情報は増大し続けている．そのため，膨大な情報の中から必要な情報をさがし出す情報検索の重要性が高まっている．従来の意味での検索は求める情報を数件見つけることにあった．しかし，現在の検索においては，ある情報についての調査・分析を行うことも求められ，単なる検索ではなく，検索を支援し分析するための機構が必要になっている．これまで利用者の検索効率を上げるために，ランキング，クラスタリング，ディレクトリ表示などの検索結果表示上の工夫がされてきた．これらの多くは求める情報を含むもののリストを返すのみであるため，検索結果に現れる重要な事柄を発見し，検索結果全体が持つ傾向を理解することは難しい．また，検索の際に所望の結果を直ちに得るための検索条件を適切に与えることも一般的に難しい．そこで，絞り込みや拡張により検索条件を繰り返し変更しながら検索を行うが，所望の結果を容易にかつ正確に得られるとは限らない．次世代の検索にはこれらの問題を効率的に解決することが求められる．

HTML 文書や XML 文書に代表される，必ずしも定型的な形式ではなく，不規則な部分も含まれたある種の構造をもつ文書は半構造化文書と呼ばれる．半構造化文書の各部分構造には何らかの観点 (あるいは項目) が含まれることが多いので，半構造化文書はさまざまな観点をもつ文書とみなせる．半構造化文書の検索において，検索対象や検索結果を見る観点はさまざまあるため，検索結果を多面的に見ることが重要になる．その実現にあたっては，観点としての部分構造をうまく利用する必要がある．しかし，多くの検索エンジンで採用されている一面的な検索結果の表示は，この問題に迅速に対応できるといいがたい．XML 検索では，XQuery に代表される的確な検索結果が得られる代わりに複雑な検索要求を必要とする検索と，単純なキーワード検索による XML 文書の全文検索が一般的な検索手法である．現在これらの融合に関する提案が多くされており，例えば構造と内容の両方の観点からの検索方式 [6]，クラスタリングや分類技術あるいはランキング技術の応用 [3], [4] は重要なテーマである．しかし，これらの多くは入力と出力の一方に着目した研究であり，検索結果の全体像の把握や検索結果を多面的に見る点に関して十分であるとはいえない．

検索結果の全体像を把握するため，クラスタリングの利用が考えられる．関連性のある検索結果のページ群を近くにまとめて 2 次元表示する検索システム KartOO(注1) はこの一例であるが，各クラスタやその位置関係の解釈は必ずしも容易でない．多面的に検索結果を見る機構としては，OLAP やバイクラスタリングがある．OLAP とはデータベースを多次元的に解析し，視覚化するシステムやツールを指す．これは数値データあるいは分類が容易なデータに対して強力な機能を有するが，これらのデータを作る過程に困難が存在する．バイクラスタリングは，2 つの観点で見た個々のデータ間の関連度を示す数値からなる行列に対し，各観点でクラスタリングを行い，しかるべき計算によって得られる各クラスタの代表値からなる行列が表示される [1]．しかし，あらかじめ関連度計算結果を用意する必要があることや各クラスタの代表値の定義方法に課題がある．また，複数の観点から絞り込みを行うために，複数の独立したカテゴリ体系を設定するファセット分類が近年注目されており，例えば検索結果のナビゲーションに利用されている [2], [5].

筆者らはこれまで，同種の項目で同じ部分構造を持つ半構造化文書のある部分構造に対する検索結果について，ユーザが縦軸横軸として指定した 2 つの部分構造を観点とするクラスタリングを行い，その分布状況をマトリクス表示する多面的検索システム (以下，本システムという) を提案してきた．本システムの考え方に近いものに先述の OLAP やバイクラスタリングがある．これらは数値データを対象とするのに対し，本システムは文書データなどの非数値データを動的に扱える点が大きく異なる．本システムは検索結果全体を視覚的にも意味的にも概観できるため，検索結果の分析が行える．そして，通常の検索システムでは検索結果から発見した新たなキーワードを使って絞り込みを行うのに対し，本システムでは行と列の特徴語群を見るだけで注目するセルを決めて絞り込みを行うのみならず，絞り込みあるいは拡張のための検索キーワードの候補を得ることで検索の支援も可能にしている．また，ある検索要求に対する検索結果が得られた後の処理について，通常の検索システムではキーボード操作により別の検索キーワードを入力するしかない．一方，本システムでは検索キーワード変更以外に，観点や分割数の変更，あるいはズーミングといった条件の変更をマウス操作のみで行い，すぐに再検索できる．つまり，同じ検索キーワードであってもさまざまな検索条件の可能性を即座に試すことができる．

すでに，この多面的検索システムを大学の教員データ [13]，病院の評判情報における品詞の分析 [14]，電子情報通信学会の講演データ [11] に対して実装している．さらに，日本動物学会学会誌 Zoological Science や病院のカルテについても実装し，専門家による定性的評価実験を行っている [8]〜[10]．本稿では，多面的検索システムが持つさまざまな機能の検索支援や検索結果の分析における利用について実例を交えて説明する．

2. クラスタリングと特徴語抽出

多面的検索システムのマトリクス表示においてクラスタリングと特徴語抽出は重要な技術である．本システムでは既に知られている技術を採用しているので，それを中心に述べる．

クラスタリング手法は，単一リンク法や完全リンク法などの階層的手法と，k-means 法などの非階層的手法に大別される．さらに前者は，分枝型と凝集型に分けられる [7]．本システムでは凝集型の階層的手法を採用している．例えば，n 個の文書から k 個のクラスタを生成する場合，まず 1 個のみの文書からなるクラスタ n 個を用意する．次にクラスタが k 個になるまで，しかるべきクラスタ間の類似度計算を用いて最も類似する 2 個のクラスタを 1 個のクラスタに結合する操作を繰り返す．一連のクラスタリングの生成状況はデンドログラムと呼ばれる二分

(注1)：http://www.kartoo.com/

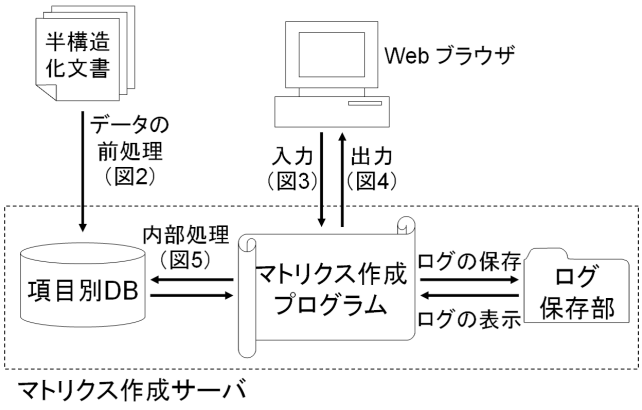

図 1　多面的検索システムの構成

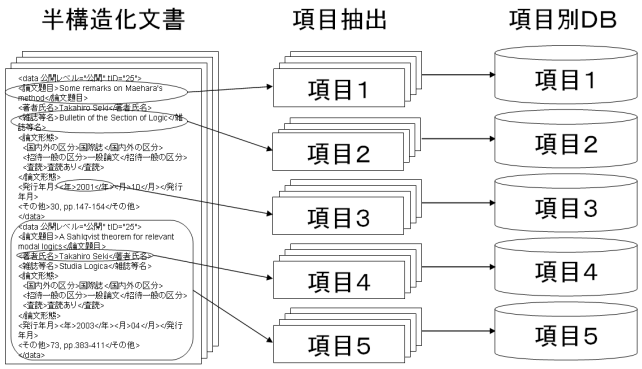

図 2　データの前処理

図 3　多面的検索システムの入力画面例

キーワード	mimo, 領域, 多重, チャネル, 分割, ds, 周波数領域等化, 空間分割多重, チャネル推定, チャネル容量	2	21	5	28
	通信, マルチ, 無線, 移動, 衛星, 光通信, 移動通信, 無線通信, 衛星通信, 可視光通信	3	90	66	159
	登録なし, システム, 制御, 通信, 情報, モデル, 符号, 画像, データ, 検索	11	259	1312	1582
	小計	16	370	1383	1769
		frequency, multiple, division, orthogonal, multiplexing, carrier, state, 次世代移動通信システム, duplex, サブチャネルブロック	提案, 方式, 特性, 通信, シミュレーション, 計算, 伝送, 受信, 送信, 計算機シミュレーション	登録なし, 提案, 本稿, 結果, 手法, システム, 研究, 利用, 実験, 情報	小計
		抄録			

図 4　多面的検索システムの出力画面例

木で描くことができる．デンドログラムを見ることにより，クラスタリングにおけるクラスタ間の結合状況が分かる．

クラスタリングは何らかの視点で似たものをまとめることであるから，生成された各クラスタに対する特徴づけが与えられる．そこで，クラスタの特徴語抽出を行う．階層的手法に基づくクラスタリングでは，生成されるクラスタ数 (以下，分割数という) が少ない場合に，結合の弱い複数のクラスタが無理矢理 1 つのクラスタに結合されることもある．したがって，1 語のみを特徴語として示すよりも，複数の語を特徴語として示す方が適当である．クラスタの特徴語を抽出する方法として自然言語処理分野で用いられる統計的指標を採用する．最も単純なものは頻度であるが，頻度が高い語でもそれが必ずしも適切な特徴語であるとは限らない．そこで，単語どうしの共起頻度を用いて定義されるイエーツの補正公式，カイ 2 乗値，コサイン，ダイス係数，自己相互情報量，対数尤度比，補完類似度 [12] などを利用する．

3.　多面的検索システム

多面的検索システムが対象とする文書群は，ある特定の文書群について同種の項目で同じ部分構造を持つ半構造化文書群である．例えば，ある大学の教員データの HTML 文書群あるいは XML 文書群は，教員名，所属，職位，研究テーマ等の各項目が同じ構造で記載されているから本システムで扱われる文書群である．一方，分野や内容を特定できない HTML 文書群や PDF 文書群のように，項目の解析が困難な文書群は本システムでは扱わない．したがって，本システムは同種の分野を扱った文書で，複数の項目があり，各項目について項目とデータを 1 対 1 に対応させることができる文書，つまり半構造化文書に変換できる文書が扱われる．本システムは，このような文書群を対象に，指定された検索領域に対してキーワード検索を行い，その結果を 2 つの項目に基づいてクラスタリングして，その分布状況をマトリクス表示するシステムである．本システムは図 1 のように構成されている．以下では既に半構造化文書群があり，検索領域や検索結果表示のための項目が決まっていると仮定する．

データの前処理は，管理者端末において以下の手順で行われる．個々の文書を 1 個のファイルとするファイル集合を用意し，項目別にデータを切り出す．次に，これらから単語の出現回数を記した頻度ファイルを生成し，項目別インデックスを作成する (図 2)．もし，ある項目が欠落していたら，データがないことを例えば「登録なし」のように明記しておく．単語の出現回数を求める際に，日本語文書を対象とする場合は茶筌[注2]等の形態素解析ソフトを利用して抽出された名詞等を単語として採用する．また，英語文書を対象とする場合はストップワードを除去し，ステミング処理あるいは TreeTagger[注3]等を利用して原形に直す処理を施したものを単語として採用する．

本システムのインタフェイスは，ユーザが検索領域とマトリクス表示するための 2 つの項目および分割数を選択し，検索キーワードを入力すると (図 3)，検索結果としてのマトリクスが表示される形式をとる (図 4)．したがって，ユーザが検索領域やマトリクス表示の縦軸項目・横軸項目を自由に切り替えることに

(注2)：http://chasen.aist-nara.ac.jp/hiki/ChaSen/
(注3)：http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html

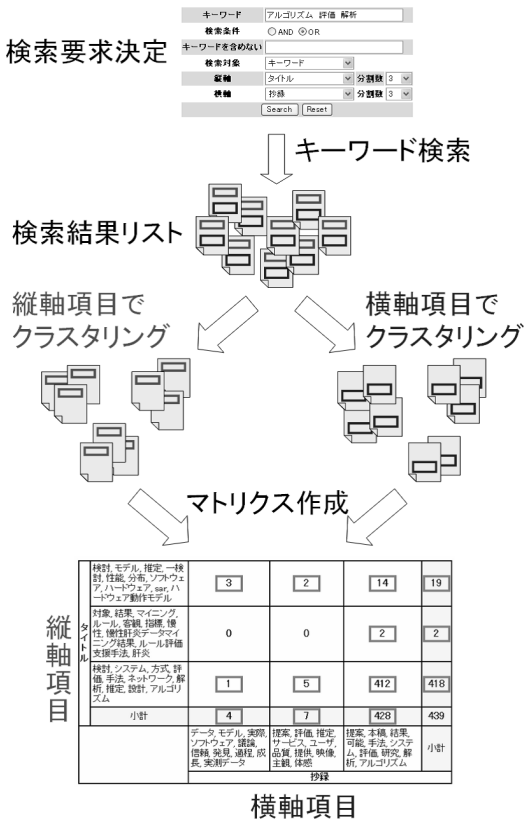

図 5　マトリクス生成の際の内部処理

より，多面的な検索を実現している．例えば，検索キーワードを固定した場合，検索領域，マトリクス表示の縦軸項目と横軸項目および各項目の分割数についてそれぞれ 5 つの選択肢を持っているとき，検索結果の表示について $5 \times 5^2 \times 5^2 = 3125$(通り) の可能性がある．

マトリクス生成の際，システム内部ではユーザが選択した検索領域に対する検索を行い，検索結果のリストを得る．次に，このリストをユーザが選択した 2 つの項目でそれぞれ選択した分割数にクラスタリングし，あわせて特徴語を抽出する．ある検索結果が縦軸項目，横軸項目のそれぞれ i, j クラスタに属しているならば，それはマトリクスの (i, j) セルの要素となる．そして，各セルに含まれる検索結果リストあるいは検索結果の数を記したマトリクスを表示する (図 5)．マトリクスにおける特徴語の提示は各クラスタの意味を与えている．これにより，検索結果全体を意味的に俯瞰し，ユーザは各クラスタの特徴語を見て所望のセル，すなわち縦軸の所望のクラスタと横軸の所望のクラスタが交わった位置にあるセルを選択して検索結果を閲覧する．

さらに，再検索の方法として，検索キーワードの変更，検索領域の変更，マトリクス表示の縦軸項目あるいは横軸項目の変更ならびに分割数の変更，セルあるいはクラスタ単位でのズーミングといったさまざまな可能性がある．ここで，ズーミングは当該セル全体に対して，ズーミング直前のマトリクス表示で用いた 2 つの項目・分割数でクラスタリングを行い，その分布状況をマトリクス表示する再帰的なマトリクス生成機能を指す (図 6)．特に，通常の検索システムでも行われる絞り込みのた

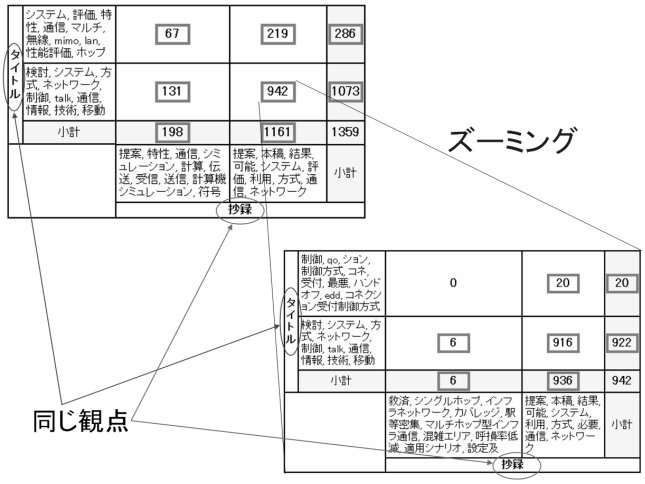

図 6　ズーミング

めの検索キーワードの変更，分割数の増加やズーミングを利用したセル当たりの件数の減少 (以下，細粒化という) は，所望のセルの該当件数が多い場合の再検索方法として有効である．また，これらは即時にかつ連続的に行うことができるため，ある時点から条件を変えて再検索する場面や複数の検索結果を比較する場面が通常の検索システムよりも多くなることが予想される．これはログの保存と再利用の重要性を示唆している．そこで，本システムではマトリクスを表示するたびにその画面を HTML ファイルにしてログ用のフォルダに保存する．

なお，本システムは Perl で記述した CGI プログラムを用いて実装されている．クラスタリングや特徴語抽出は 2 節で述べた方法を用い，インデックスの作成および文書検索は汎用連想計算エンジン (GETA)(注4)，クラスタリングのためのライブラリとして CPAN (注5) の perl モジュール Algorithm::Cluster を採用した．

4. 多面的検索システムによる検索支援

検索結果が数件しかない検索要求に対しては通常の検索システムで十分対応できる．しかし，通常の検索システムにおける検索の初期段階では，しばしば用いられる語を検索キーワードとする簡単な検索式が検索要求になるため，大量かつ多義的に解釈しうる検索結果が出力されることが多い．このような場合に，各クラスタの特徴語を提示することによって多義性の一つの解釈を与える多面的検索システムは有効である．例えば，論文データを対象に実装した本システムに対し，主流の研究テーマに関する語をキーワードとして検索する場合，そのテーマについて細分化された主題と研究者・研究グループの関連を見ることも可能であるが，これは通常の検索システムでは行えない．本システムのマトリクス表示は多義的な検索結果からの絞り込みを容易にしているが，さらなる絞り込みを行う必要がしばしば生じる．通常の検索システムでは極めて難しいさらなる絞り込みを，本システムではさまざまな機能を用いて容易に行える．

(注4)：http://geta.ex.nii.ac.jp/
(注5)：http://www.cpan.org/

表 1 「検索 or システム」に対する検索結果 (1)

	A	B	C	小計
a	0	0	4	4
b	0	0	4	4
c	8	4	1181	1193
小計	8	4	1176	1201

a	ベース，帯域，割り当て，内容，ＬＡＮ，映像伝送， サーベイランス，画像内容， マルチカメラサーベイランスシステム，帯域割
b	配分，移動通信，高度無線信号処理技術， 高度移動通信システム，伝送容量，移動衛星通信システム， 面的，リソース配分，マルチビーム移動体衛星通信システム， マルチリンク
c	検討，システム，評価，提案，通信，情報，開発，支援， 構築，検索
A	multiple，dynamic，tdma，carrier，parameter，controlled， 次世代移動通信システム，dpc，ci，interferometry
B	単体，ソフトエラー，キャッシュメモリ， コンピュータシステム，ソフトエラー発生回数， ソフトエラー数，ソフトエラー数見積，悲観的，宇宙線， メモリ階層
C	(登録なし)，提案，本稿，結果，可能，システム，評価， 研究，報告，利用

本節では，本システムのさまざまな機能について検索支援の観点から説明する．

本稿では，電子情報通信学会で 2003 年以降に開催された全ての研究会における講演のデータに対して実装したシステム(注6)を中心に説明する．講演データは全部で 10,903 件あり，各講演の情報として研究会名，講演年月，開催地，タイトル，講演者，抄録，キーワード等の項目が含まれている．

4.1 特徴語を利用した検索キーワード発見

通常の検索システムでは，検索結果として示されたリストから，ユーザが実際に個々のデータを見て獲得したキーワードを用いて絞り込みや拡張を行う．しかし，個々のデータを繰り返し閲覧することはユーザにとって必ずしも効率的ではない．本システムでは検索結果であるマトリクスに各クラスタの特徴語を提示することにより，個々のデータを見ることなく検索キーワードとして付加すべき語や除くべき語の候補を発見できる．

例えば，検索システムをテーマに調査する場合,「検索」または「システム」をタイトルに含む講演を検索するのは一つの考え方である．そこで，タイトルを検索領域に「検索 or システム」を検索キーワードとし，縦軸をタイトル，横軸を抄録として 3 × 3 表示した結果を表 1 に示す．表 1 を見ると，b 行と c 行の特徴語として「通信」を含む語がいくつか出現しているが，通信の文脈から検索していないユーザにとっては除くべき語の候補である．さらに，a 行の特徴語にある「帯域」も除くべき語の候補になるだろう．そこで，検索キーワードを「(検索 or システム) not (通信 and 帯域)」とし，他の条件は前と同一

(注6)：http://matu.cc.kyushu-u.ac.jp/works/matrix/ieice/ にて公開している

表 2 「(検索 or システム) not (通信 and 帯域)」に対する検索結果

	A	B	C	小計
a	0	0	3	3
b	0	0	6	6
c	16	4	1071	1091
小計	16	4	1080	1100

a	必要，経験，診療，バイ，機器システム， 大災害時高抗堪性診療，避難，必要条件，タルサ， 中越大震災避難所診療経験
b	マルチメディア，解答，軌跡，操作式解答，dril， マルチメディア試験システム，可搬型機器，領域解答， 機能検証，軌跡解答
c	検討，システム，評価，提案，情報，開発，利用，支援， 構築，検索
A	dynamic，tdma，carrier，parameter，controlled，dpc， パイロット信号，セルラシステム，ci，interferometry
B	単体，ソフトエラー，キャッシュメモリ， コンピュータシステム，ソフトエラー発生回数， ソフトエラー数，ソフトエラー数見積，悲観的，宇宙線， メモリ階層
C	(登録なし)，提案，本稿，結果，可能，システム，研究， 報告，利用，情報

表 3 「データベース or 検索」に対する 2 × 2 表示

	A	B	小計
a	0	8	8
b	2	310	312
小計	2	318	320

a	目的，発見，サーチ，エンジン，サーチエンジン， 同位，精製，半自動生成，類似度データ，自動索引付
b	システム，方式，手法，ネットワーク，提案，情報， 画像，利用，検索，データベース
A	ディジタル，１つ，簡便，撮影位置，ディジタルカメラ， 日常化，未整理，撮影機器，関連語抽出処理，姿勢情報
B	(登録なし)，提案，本稿，結果，手法，システム， 利用，情報，検索，データベース

にしたときの結果を表 2 に示す．特徴語を見ると通信関連の語を連想させるものは少なくなっていることが確認できる．

4.2 分割数変更による絞り込み支援

本システムで検索結果を絞り込むためには and 検索や not 検索による検索条件の厳格化だけでなく，分割数の増加やズーミングを用いて細粒化を繰り返すことも可能である．これらは通常の検索で行われる他のクラスタを排除する検索条件を考え出す必要がない点に特徴がある．分割数の増加は深い階層の状況を見るために用いられるばかりでなく，選択肢を広げるための操作でもある．例えば，2001 年から 2006 年までを 2 年ごとに見るように 3 分割されていたものを 6 分割にする分割数増加は，2 年単位で見ていたものを 1 年単位で見るように変更するイメージである．

全項目を検索領域に「データベース or 検索」を検索キーワードとし，縦軸をタイトル，横軸を抄録としたときの 2 × 2 表示した結果を表 3 に，3 × 3 表示した結果を表 4 に示す．例えば，

表 4 「データベース or 検索」に対する 3 × 3 表示

	A	B	C	小計
a	0	0	8	8
b	0	0	5	5
c	2	2	303	307
小計	2	2	316	320

a	目的，発見，サーチ，エンジン，サーチエンジン， 同位，精製，半自動生成，類似度データ，自動索引付
b	表示，インターフェース，アーカイブ，粒子，降雪， 検索結果，風景，風景画像検索， 降雪粒子画像アーカイブシステム，表示インターフェース
c	システム，方式，手法，ネットワーク，提案，情報， 画像，利用，検索，データベース
A	ディジタル，1つ，簡便，撮影位置，ディジタルカメラ， 日常化，未整理，撮影機器，関連語抽出処理，姿勢情報
B	genetic，言語的，遺伝的プログラミング，クラスタ内， 数値データ，検索ルール，ローン決定問題， 階層的ルール生成，commonality，カテゴリ化
C	(登録なし)，提案，本稿，結果，手法，システム， 利用，情報，検索，データベース

ユーザが表 3 を見たときに，所望の結果が Ab セルでも Ba セルでもない場合，ユーザは Bb セルをより細かく見ることになる．表 4 は表 3 における分割数を増加することで得られるが，特徴語の比較を通じて，表 3 の Bb セルは表 4 において Bb，Bc，Cb，Cc セルの 4 つに分割されたことが分かる．この分割の状況を特徴語から見ることで，風景，降雪，粒子の画像検索に関する講演が Cb セルに含まれるというように絞り込みが行える．ここで，縦軸も横軸も分割数を 1 ずつ増加したときに該当件数が最も多いセルが必ずしも分割されるとは限らないことと，特徴語が示されていない場合に分割状況が必ずしも容易に判別できないことを注意しておく．なお，分割数の増加に関する分析については 5.2 節と 5.3 節でも述べる．

4.3 ズーミングによる絞り込み支援

本システムのマトリクス表示における各セルで示されるリストは，通常の検索システムにおいて 2 つの項目とキーワードについて and 検索した結果のリストを連想させる．しかし，通常の検索システムでこれをより詳細に見るための条件を考えることは難しい．本システムにおけるズーミングは，ユーザが興味をもったセルあるいはクラスタを選んでクリックするだけでその詳細を見ることができ，この問題に対する一つの解決法を与えている．例えば，ユーザが興味をもったクラスタとそうでないクラスタが明確に分離されている場合に，興味をもったクラスタについてのみズーミングを行うことで直ちに詳細を見ることができる．

表 3 の Bb セルの 310 件をズーミングすると，表 5 が得られる．このとき，頑健なクラスタが依然として残っているため，表 3 と表 5 には同じような特徴語が示されている．Aa セルの件数は 0 (以下，このようなセルを空セルという) なので，該当件数の少ない Ab セルと Ba セルはそれぞれ A 列と a 行の特徴語を見ることでどのようなセルであるかが分かる．Ab セルは

表 5 表 3 の Bb セルのズーミング結果

	A	B	小計
a	0	16	16
b	2	292	294
小計	2	308	310

a	認識，相関，改良，文字，オンライン，文字認識， 歩行者，顔認識，一改良，相関法
b	システム，方式，手法，ネットワーク，情報，画像， 利用，web，検索，データベース
A	genetic，言語的，遺伝的プログラミング，クラスタ内， 数値データ，検索ルール，ローン決定問題， 階層的ルール生成，commonality，カテゴリ化
B	(登録なし)，提案，本稿，結果，手法，システム， 利用，情報，検索，データベース

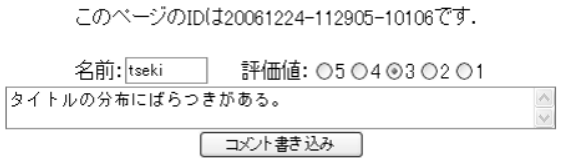

図 7 コメントの記入画面

遺伝的プログラミングに関する講演，Ba セルは認識に関する講演についてのセルであるといえ，そのような講演をさがしている場合は絞り込みが行えたことになる．別の見方をすると，縦軸項目 (タイトル) からは，認識に関連する講演が分離されたと見ることができ，横軸項目 (抄録) からは，遺伝的プログラミングに関する講演が分離されたと考えられる．

4.4 ログ保存機能の活用

さまざまな検索条件を試すことができる本システムの特徴は検索や分析の可能性を広げる一方で，後で過去の検索結果を参照したり，過去の検索結果からその検索を起点にした別の検索を行いたいと思う可能性がある．そこで，検索を行った時間，検索条件をログとして保存し，後から閲覧したり，保存されているログから条件を変えて再び検索したりできる機能を追加している．例えば (1) 最初の検索，(2) ズーミング，(3) 分割数変更，(4) ズーミング，(5) 条件変更，のような手順で操作をしていて，後になって (4) の時点から別の条件に変更したい場合，今までは (1) からやり直さなければならなかったので，手順を記録または記憶しない限り (4) までたどり着くことはできなかったが，このログ保存機能を使うとログを見てすぐに (4) の段階からやり直すことが可能であり，格段に効率がよくなる．

ところが，システムの利用が増えると保存されるログの数が多くなり，所望のログを見つけ出す作業に時間がかかる問題があった [11]．今回この問題を解決するためにログ検索機能を追加した．また，検索の際に意図した内容や検索結果を見た印象を記入したり，後でもう一度見たいことを表す目印をつけたりすることを目的に，検索結果に対するコメントを図 7 のような形で記入し，それをログとして残す機能も追加した．ログを閲覧する際の画面例を図 8 に示す．上部がログの検索部分である．下部がログの一覧になっており，閲覧したい検索要求について

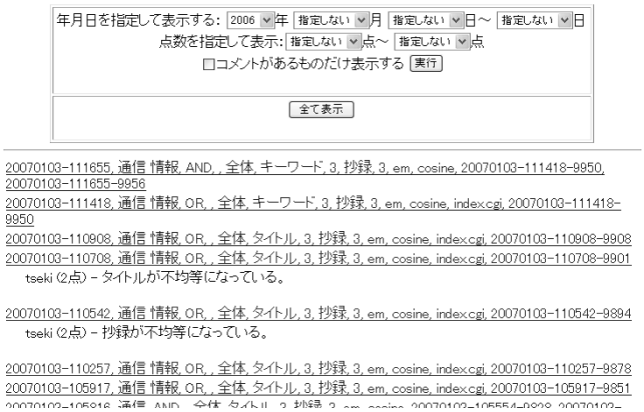

図 8　ログの表示画面

表 6 「検索 or システム」に対する検索結果 (2)

(a)　タイトル × 抄録

	A	B	C	小計
a	0	0	4	4
b	0	0	4	4
c	8	4	1181	1193
小計	8	4	1176	1201

(b)　研究会 × 開催年

	A	B	C	小計
a	3	4	4	11
b	4	20	4	28
c	182	526	454	1162
小計	189	550	462	1201

そのリンクをクリックすると，その条件のもとでの検索結果が表示され，そこから再検索できる仕組みになっている．なお，再検索の結果もログとして保存される．本稿の執筆で，今回追加した機能は以前のログ保存機能に比べ所望のログを発見するための効率化に大きく寄与した．

5.　多面的検索システムによる検索結果の分析

本節は多面的検索システムを利用した検索結果の分析方法を例を交えて説明する．

5.1　多面的分析

本システムは検索結果を各クラスタの特徴語を提示しながらマトリクス表示することにより，検索結果の全体像を 2 つの観点から把握し，縦軸項目や横軸項目の変更を通じてさまざまな切り口から検索結果を分析することができる．

表 6 はタイトルを検索領域に「検索 or システム」を検索キーワードとし，縦軸をタイトル，横軸を抄録として 3 × 3 表示した結果を (a) に，縦軸を研究会，横軸を開催年として 3 × 3 表示した結果を (b) に示したものである．ここで，(a) は表 1 の再掲である．表 6 から分かるように，縦軸項目・横軸項目を変えると，検索結果の総数は同じでもマトリクスにおける分布状況は異なる．

検索結果の分析は各クラスタの特徴語を手掛かりに行われる．まず，縦軸項目と横軸項目が関連性をもつ場合，2 つの項目に出現する特徴語の比較を通じて，特徴語 w が 2 つの項目に共通して現れるのでそのセルは w について強い関わりがありそうだとか，予想していなかった特徴語が示されているのであるセルは珍しいものを含んでいるといった分析ができる．一方，縦軸と横軸に共通に出現する特徴語が少ない場合はセルの特徴がつかみにくく，件数が多ければ多いほどその傾向は強まる．このような場合は分割数増加あるいはズーミングを用いてセルを詳細に分析する必要がある．表 6 (a) に則して説明すると，空セルである Aa，Ab，Ba，Bb の 4 つのセルと該当件数の多い Cc セル以外の 4 つのセルの解釈は容易である．これらのセルの該当件数は少ないので，特徴語から Ac セルは次世代移動通信システムや TDMA (Time Division Multiple Access)，Bc セルはソフトエラー，Ca セルはマルチカメラサーベイランスシステム，Cb セルは移動通信に関連するセルであると見当がつく．

次に，縦軸項目と横軸項目が直接に関連しそうにない項目を選択した場合について記す．表 6 (b) はこれに該当する．例えば，両軸ともに数値や地名等の固定属性を持つような場合には各属性からの分類が示される一方，そうでない場合にはさまざまな見方が可能になると予想される．これにより一方の項目から見たときの特徴的なセルや異質な特徴を示すセルを見つけたり，両方の項目の隠された関係を明らかにしたりできる．例えば，数十年の論文データに対し，縦軸に論文の内容がわかる項目，横軸に論文の発行年をとれば，研究のはやりすたりが見えてくるだろう．

表 7 「検索 or システム」に対する検索結果のデンドログラム表示

				A	B	C	D	小計
				1201				
				8	1193			
					4	1189		
						14	1175	
a	1201	8	4	0	0	0	4	4
b			4	0	0	0	4	4
c		1193	8	0	0	0	8	8
d			1185	8	4	14	1159	1185
小計				8	4	14	1175	1201

5.2　分割数変更とデンドログラム表示

本システムは階層的手法に基づくクラスタリングを採用しているので，分割数の増加はあるセルを分割することを意味する．1 次元で分割数を 1 増やせばどれか 1 つのクラスタが 2 つに分割されるのに対し，2 次元で分割数を 1 ずつ増やすと 1 つのセルが最大 4 つのセルに分割されるため，分割の追跡と解釈は必ずしも容易ではない．ところが，各軸のクラスタの結合状況はデンドログラム表示できるので，クラスタ間の結合の強弱の分析が行える．

表 7 はタイトルを検索領域に「検索 or システム」を検索キーワードとし，縦軸をタイトル，横軸を抄録として 4 × 4 表示した結果をデンドログラム表示したものである．つまり，表 6 (a) (したがって表 1) のマトリクス部分に関して 4 × 4 表示したものである．いま，n 個のクラスタ $X_1, \cdots, X_n$ が 1 個に結合されたクラスタを $\langle X_1, \cdots, X_n \rangle$ と書くことにすると，縦軸の 4 つのクラスタ a，b，c，d について，結合の弱い順に c と d，a と b，⟨a,b⟩ と ⟨c,d⟩ と書くことができる．横軸の 4 つのクラスタ A，B，C，D に関しては，C と D，B と ⟨C,D⟩，A と ⟨B,C,D⟩ となる．これを見ると，3 × 3 表示した場合には，c 行と d 行が 1 つになり，C 列と D 列が 1 つになることが分かり，表 6 (a) ではそのようになっている．言い換えると，表 6 (a)

の 3×3 表示から表 7 の 4×4 表示への分割数増加で，表 7 における c 行と C 列がそれぞれ，表 6 (a) における c 行と C 行から分割されたクラスタであることが分かる．

5.3 細粒化による分析

本システムで該当件数が多いセルは，and 検索や not 検索の利用による検索キーワードの厳格化，分割数の増加，そのセル自身やそのセルを含む行や列に関するズーミングといった，複数の手法に基づく細粒化の利用を通じて分析可能である．例えば，特徴語を見ると明らかに異なるクラスタに属するものが混在する場合，これらの操作を通じて混在していたものが別々のクラスタに分離できると考えられる．個々の細粒化手法については既に述べているので，ここではこれらを比較していく．

検索結果中に入っていては困る語があったり，逆に入っていなくては困る語を検索キーワードに加えなかった場合はまず検索キーワードの厳格化あるいは拡張を行うべきである．また，大きめの頑強なクラスタがあって，さらにそのクラスタの特徴語に上記のような語が含まれる場合には，and 検索や not 検索を利用して検索条件を洗練した方がよい．

単一のセルないしクラスタを詳細に見たい場合には分割数増加かズーミングの利用が考えられる．分割数増加は直前の全体的な状況を踏まえた細粒化であるのに対し，ズーミングはその場の状況のみに着目した排他的な細粒化であるという違いがある．例えば 100 件の検索結果のマトリクス表示中に 50 件のセルが存在し，そのセルを細粒化する場合を考える．分割数増加の場合はもともとあった 100 件のデータを基にクラスタリングを行った結果としての細粒化であるのに対し，ズーミングの場合は 50 件のデータを基にクラスタリングを行った結果としての細粒化である．

このように考えると，検索結果全体あるいは複数のセルやクラスタを詳細に見たい場合には分割数増加が適切な方法であるといえる．しかし，4.2 節で指摘したように所望のセルないしクラスタが分割されるという保証はなく，分割数増加を繰り返すと分割数の多い巨大なマトリクスが生成されることを覚悟する必要がある．また，該当件数の多いクラスタが頑健な場合は，分割数を増やしても該当件数が少ないセルあるいは空セルがいたずらに増えるばかりである．一つの目安としては，セルの半分近くが空セルになった時点で分割数の増加以外の方法を考えるべきである．

6. まとめと今後の課題

本稿では，教員データや文献データ，講演データ等の，ある特定の文書群について同種の項目で同じ部分構造を持つ半構造化文書群に対し，任意の観点について検索を行い，任意に選んだ 2 つの観点でクラスタリングを行った結果をマトリクス表示する多面的検索システムを開発し，それを用いた検索支援や分析の可能性について論じた．

現在，数量データのようにクラスタリングよりも分類 (classification) が適切なデータに対する扱いが十分に行えていない問題がある．また，各種クラスタリング方法に関する検討が課題として残っている．さらに，ある語が複数のクラスタの特徴語にはならないような特徴語抽出方法については十分な検討を要する．以上の問題を解決した上で検索の効率に関する定量的な評価実験を行うと同時に，本システムを用いた検索が通常の検索システムを用いたときと比較して，どの程度ユーザの満足度を向上させるかについて調査したい．今後，同種の項目であっても同じ部分構造を持つとは限らない半構造化文書に対する本システムの実装可能性についても検討する予定である．

謝 辞 有益なコメントをいただいた査読者の方々に感謝します．本研究は，平成 17 年度科学研究費特定領域 16016267 ならびに科学技術振興機構独創的シーズ展開事業平成 18 年度大学発ベンチャー創出推進による．

文 献

[1] Y. Cheng, G.M. Church. Biclustering of expression data, *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, 2000, pp.93–103

[2] V. Christophides, D. Plexousakis, M. Scholl, S. Tourtounis. On labeling schemes for the semantic web, *WWW2003*, 2003, pp.544–555

[3] L. Candillier, I. Tellier, F. Torre. Transforming XML trees for efficient classification and clustering, *INEX2005*, 2005, pp.469–480

[4] L. Guo, F. Shao, C. Botev, J. Shanmugasundaram. XRANK: Ranked Keyword Search over XML Documents, *SIGMOD2003*, 2003, pp.16–27

[5] Y. Tzitzikas, N. Spyratos, P. Constantopoulos, A. Analyti. Extended Faceted Taxonomies for Web Catalogs, *WISE2002*, 2002, pp.192–204

[6] C. Yu, H. Qi, H.V. Jagadish. Integration of IR into an XML Database, *INEX Workshop 2002*, 2002, pp.162–169

[7] 神嶌 敏弘. データマイニング分野のクラスタリング手法 (1) － クラスタリングを使ってみよう！ －, 人工知能学会誌 *vol.18, no.1*, 2003, pp.59–65

[8] 関 隆宏, 安元裕司, 和多太樹, 伊藤 希, 廣川佐千男. 電子ジャーナルの多面的検索分析システム, 情報処理学会 第 *68* 回全国大会 *3*, 2006, pp.43–44

[9] 関 隆宏, 安元裕司, 和多太樹, 中島直樹, 廣川佐千男. XML 文書としての電子カルテ検索システム, 情報処理学会 第 *68* 回全国大会 *3*, 2006, pp.51–52

[10] 関 隆宏, 安元裕司, 和多太樹, 伊藤 希, 廣川佐千男. 多次元マトリックス法による電子ジャーナル Zoological Science の検索と分析, ディジタル図書館 *No.30*, 2006, pp.72–80

[11] 関 隆宏, 和多太樹, 山田泰寛, 廣川佐千男. 学会講演データの多面的解析システム, 人工知能学会研究会資料 *SIG-FPAI-A601*, 2006, pp.13–18

[12] 中條清美, 内山将夫, 長谷川修治. 統計的指標を利用した時事英語資料の特徴語選定に関する研究, 英語コーパス研究第 *12* 号, 2005, pp. 19-35

[13] 廣川佐千男, 関 隆宏, 安元裕司, 山田泰寛. 教員データに対する多面的検索システム, 情報処理学会研究報告 *2005-DBS-137*, 2005, pp.665–672

[14] 安元裕司, 和多太樹, 関 隆宏, 廣川佐千男. 病院評判情報の多面的解析, 人工知能学会研究会資料 *SIG-KBS-A501*, 2005, pp.1–4