3D-7

多面的解析システムにおけるデンドログラムの実装

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† 　九州大学大学院システム情報科学府‡
九州大学ユーザーサイエンス機構* 　九州大学情報基盤センター**

1．はじめに

情報の電子化や Web の発展は文書群の爆発的な増加と利用可能性の拡大を引き起こし，文書検索の重要性が高まっている．しかし，従来の検索システムの多くで採用される検索要求に合致する文書のリストを返す方法は，利用者にとって検索効率が必ずしもよくない．一つの解決法として，検索を行うと同時にその検索結果を分析して提示することが考えられる．

これまで筆者らは，同種の項目が同じ部分構造をも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 つの項目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解析システム（以下，本システム）を提案してきた [1,3,4]．本システムの考え方に近いものに OLAP やバイクラスタリングがあるが，これらが扱うのは数値データである点に大きな違いがある．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という特徴がある．しかし，これまでの実装においては，クラスタ間の結合の強弱が分からないため，検索結果全体の分類構造を解釈できない問題があった．

本システムは階層的手法によるクラスタリングを採用しているので，クラスタリングの生成状況はデンドログラムで表示できる．既に筆者らは電子情報通信学会で 2003 年以降に開催されたすべての研究会における講演のデータに対して本システムを実装している．本稿では，各軸に関するクラスタリングの生成状況をデンドログラム表示する機能を加え，それを用いた検索結果の大局的構造解析に関する定性的評価について述べる．

2．クラスタリングとデンドログラム

本システムで採用されているクラスタリング方法は，凝集型の階層的手法[2]である．与えられた n 個の文書から k 個のクラスタを生成する手順は次の通りである．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，しかるべきクラスタ間の類似度計算を用い，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．この操作をクラスタの個数（以下，分割数）が k になるまで繰り返す．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．

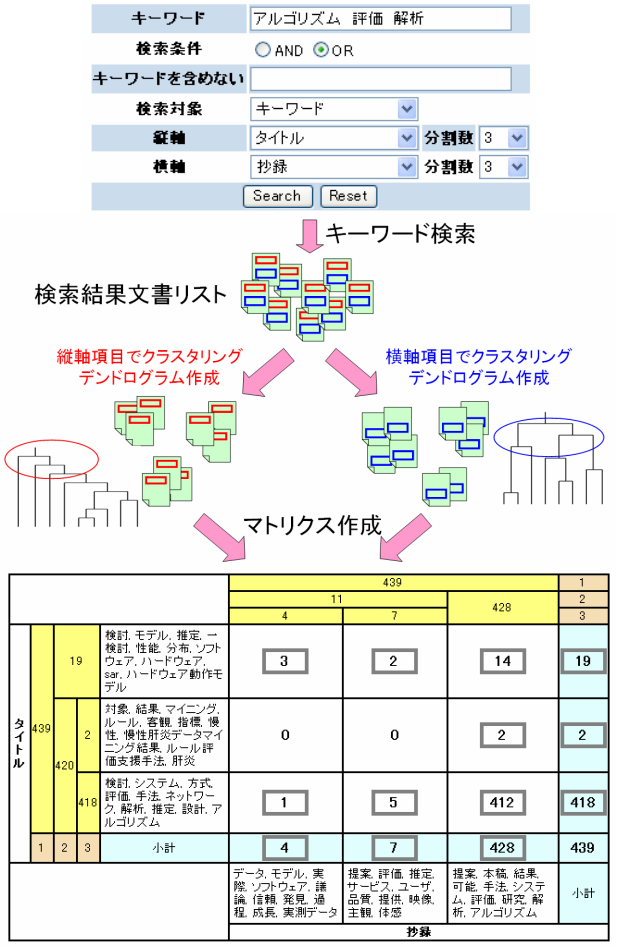

図 1　多面的解析システムの構成

3．多面的解析システム

多面的解析システムは，対象文書の指定された検索領域に対してキーワード検索を行い，その結果をユーザが選択した 2 つの観点でクラスタリングし，その分布状況を 2 次元マトリクスとして表示するシステムである（図 1）．具体的には，ある検索結果が縦軸，横軸のそれぞれ i，j クラスタに属しているとき，その検索結果はマトリクスの (i，j)セルの要素となる．マトリクスの表示について，各セルにはそのセルに含まれる検索結果ある

An Implementation of dendrogram in a faceted search engine
†Office for Information of University Evaluation, Kyushu University
‡Graduate School of Information Science and Electrical Engineering, Kyushu University
*Kyushu University User Science Institute
**Computing and Communications Center, Kyushu University

いは検索結果数が記されている．あわせて，各クラスタの意味を理解するための特徴語を提示する．さらに，本稿では分類構造を見るためのデンドログラムを新たに加えている．

これにより，検索結果全体を意味的に俯瞰し，各クラスタの特徴語を見て所望のセルを選択する．また，所望のセルはズーミングあるいは分割数変更を通じて絞り込みが行える．さらに，縦軸横軸の観点を自由に切り替えることで，多面的な分析を可能にする．

本システムの実装は Perl で記述した CGI プログラムを用いた．ベクトル空間モデルに基づくクラスタリングや特徴語抽出を行うため，その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[1]），クラスタリング計算およびデンドログラム生成は CPAN [2] の perl モジュール Algorithm::Cluster を利用している．

4．考　察

本システムは階層的手法によるクラスタリングを採用している．1 次元の場合，分割数を 1 増加することはあるクラスタが 2 分割されることを意味し，比較的容易に分割前後の解釈が行える．一方，2 次元の場合，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，その分割の解釈は必ずしも容易でない．しかし，デンドログラム表示と連動しながら分析を行うと，セルの分割状況が見えてくる．

本稿では電子情報通信学会の講演データを用いて実装した本システムを例に，デンドログラム表示の利用について述べる．キーワードとして「アルゴリズム」または「評価」または「解析」を含んでいる講演について，タイトルを縦軸，抄録を横軸として 3×3 表示した結果を表 1 に示す．

表 1 のデンドログラム表示から 2×2 表示したとき，b 行と c 行，A 列と B 列がそれぞれ一つのクラスタになることが分かる．逆に，分割数を増加し，2×2 表示から 3×3 表示にしたとき，次のことが分かる．縦軸に関して，420 件のクラスタが b 行の 2 件と c 行の 418 件に分割される．表 1 から b 行について慢性肝炎に関するタイトルの講演が分割されたと予想され，実際その通りであった．横軸に関して，11 件のクラスタが A 列の 4 件と B 列の 7 件に分割され，この分割の観点は表 1 で下線を付さなかった特徴語にあると予想される．実際のデータを見ると，439 件の全検索結果のうち A 列の「成長」，B 列の「体感」が抄録に含まれているのはそれぞれ A 列の 4 件，B 列の 7 件のみであった．

マトリクス表示において，数値や地名等とは異なる，固定属性をもたない観点を選択した場合，

[1] http://geta.ex.nii.ac.jp/　　[2] http://www.cpan.org/

表 1　「アルゴリズム or 評価 or 解析」の検索結果

				A	B	C
				439		
				11		428
				4	7	
a	439	19		3	2	14
b		420	2	0	0	2
c			418	1	5	412

a	検討，モデル，推定，一検討，性能，分布，ソフトウェア，ハードウェア，sar，ハードウェア動作モデル
b	対象，結果，マイニング，ルール，客観，指標，慢性，慢性肝炎データマイニング結果，ルール評価支援手法，肝炎
c	検討，システム，方式，評価，手法，ネットワーク，解析，推定，設計，アルゴリズム
A	データ，モデル，実際，ソフトウェア，議論，信頼，発見，過程，成長，実測データ
B	提案，評価，推定，サービス，ユーザ，品質，提供，映像，主観，体感
C	提案，本稿，結果，可能，手法，システム，評価，研究，解析，アルゴリズム

(注)　A と B で下線を付した特徴語は 2×2 表示における特徴語であることを表す．

クラスタリング結果が，1 つのクラスタのみ構成要素数が多く，他のクラスタの構成要素数が少ない状況がしばしば生じる．この状況はチェイニング効果[2]と呼ばれ，デンドログラム表示においても明確に示されていた．

5．まとめと今後の課題

本稿では，多面的解析システムにおける検索結果全体の分類構造を把握するため，新たにデンドログラムを表示する機能を追加した．そして，デンドログラム表示の観点から分析を行い，分割数の増加に伴うセルの分割状況が明確になることが分かった．また，デンドログラム表示の利用により多くの場合でチェイニング効果が現れることが確認された．今後クラスタリング方法についてさまざまな面から検討する必要がある．

参 考 文 献

[1] T.Seki, T.Wada, Y.Yamada, N.Yetow, S.Hirokawa. “Multiple Viewed Search Engine for e-Journal ― a Case Study on Zoological Science”, Proceedings of the HCI International 2007, 2007, to appear

[2] 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，pp.59－65，2003

[3] 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，pp.13－18，2006

[4] 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，pp.665－672，2005