4D-2

電子ジャーナルの多面的検索分析システム

関　隆宏†　　安元　裕司‡　　和多　太樹‡　　伊藤　希*　　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‡
筑波大学大学院生命環境科学研究科*　　九州大学情報基盤センター**

１．はじめに

文書群の増加に伴い，検索だけでなく検索結果の分析が重要になっている．一般の検索システムにおける単純なリスティングやランキングは検索結果を一面的に表示しているにすぎない．分析のためには，個々の検索結果だけからは分からない全体の概観が必要となる．複数項目からなる構造化文書群の検索においては，さらに検索結果を項目ごとに多面的に見ることが求められる．半構造化文書の検索例である XML 検索では，構造と内容の両方の観点からの検索方式，あるいはランキング技術との融合について研究されている[1,2]が検索結果の概観を与えるものではない．

クラスタリングにより関連性のある検索結果を画面上にまとめて表示する検索システム KartOO があるが，位置関係に意味はない．本稿では，検索結果を２つの観点でクラスタリングし，その分布を２次元マトリクス表示するシステムを提案する．さらにこのシステムでは各クラスタの特徴語を自動抽出するため，検索結果の全体像を視覚的にも意味的にも概観できる．これらの特徴語は検索の絞り込みにも利用できる．

筆者らは，本システムを大学の教員データに対して実装し[3]，病院の評判情報における品詞の分析に利用している[4]．本稿では，電子ジャーナルである日本動物学会学会誌 Zoological Science の概要ページ（全部で 638 件）にある複数項目を対象とする多面的検索分析システムとそれを用いた定性的評価実験について述べる．

２．多面的検索分析システム

本システムは，検索結果の文書群を２つの項目に着目しそれぞれの観点からクラスタリングを行い，結果を２次元マトリクス状に表示する．項目ごとのクラスタリングを実装するため，内部的には項目ごとに個別のインデックスを持つ．さらに，クラスタリングの際に各クラスタを特徴付けるキーワード群を抽出し，縦軸と横軸にこれらのキーワードを表示することにより，検索結果の意味的な構造認識を可能にし，複数の観点からの分析を実現する．ユーザは各クラスタの特徴語を見て所望のセルを選択し，その結果を見る．さらに，そのセルに属する数が多い場合，ズーミングによりさらなる絞り込みを行う．

３．多面的検索分析システムの実装

キーワードからの文書検索には国立情報学研究所で開発された汎用連想計算エンジン（GETA）を用いた．まず，検索対象となる Zoological Science の論文アブストラクト情報からタイトル，概要，著者，参考文献，巻の５項目を抽出し，各項目について個別のインデックスを作成した（図１）．

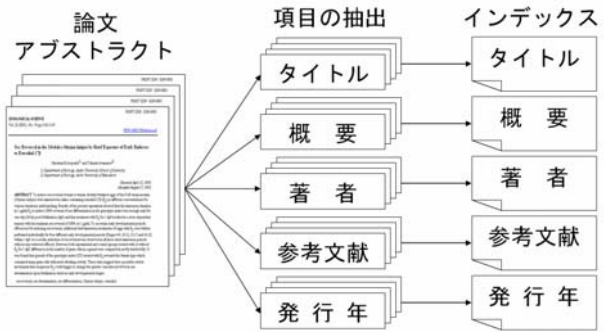

図１　インデックス作成

ユーザは検索対象項目，クラスタリングのために着目する２つの項目，クラスタ数を指定して，検索キーワードを入力する．検索結果はその指定に応じたマトリクスとして表示される．

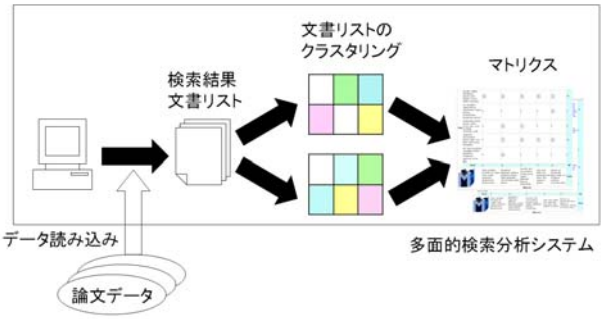

図２　システムの構成

マトリクス生成の際，内部では以下の処理を行っている(図２)．まず，検索要求からユーザが選択した項目で検索し，検索結果である文書リストを得る．次に，この文書リストを，ユーザが選択

Multiple-viewed Search and Analysis Engine for Electrical Journal
†Office for Information of University Evaluation, Kyushu University
‡Graduate School of Information Science and Electrical Engineering, Kyushu University
*Graduate School of Life and Environmental Science, University of Tsukuba
**Computing and Communications Center, Kyushu University

した２つの項目でそれぞれ指定した数にクラスタリングし，あわせて特徴語を抽出する．なお，本システムでは GETA で実装されているクラスタリング計算方法を選択し，特徴語は各クラスタ中で出現頻度の高い 10 語を抽出する．ある文書のクラスタリング結果がそれぞれクラスタ i，j に属するとき，その文書はマトリクスの i 行 j 列セルの要素となる．これを検索された文書リストすべてに対して行い，セル内にそのセルに含まれる文書リストならびに文書数を記した２つのマトリクスを表示する．さらに，あるセルをクリックすると，そのセルに含まれる文書リストを基に同じ観点で再度マトリクスを生成するズーミングを行う．

４．定性的評価実験

クラスタリング結果の生物学的意味を調べるため，638 の全論文のタイトルと概要について分割数を変えながらマトリクス表示した．なお，クラスタリング計算方法として，確率的クラスタリングを用いた．図３は縦軸をタイトル，横軸を概要として３×３表示(a)と４×４表示(b)したものを小計付きで示したものである．また，３×３表示での特徴語を図４に示す．下線を付した語は行と列の両方に現れた語である．

1	100	27	128
72	66	113	251
5	75	179	259
78	241	319	638

(a)

1	45	55	27	128
57	10	21	34	122
15	15	20	79	129
5	43	32	179	259
78	113	128	319	638

(b)

図３　全論文の分割数を変えたときの変遷状況

１行	species, japan, genus, crustacea, mitochondrial, sequences, dna, amphibia, sp
２行	biology, symposium, endocrinology, acknowledgments, biochemistry, genetics, physiology, author, index
３行	expression, cells, medaka, oryzias, latipes, hormone, cell, japanese, ascidian
１列	特徴語なし（ほとんどが元データに概要なし）
２列	species, japan, sp, populations, population, genus, nov, genetic, females
３列	cells, expression, results, levels, brain, activity, mrna, cell, development

図４　３×３表示での特徴語

(a)の１列，３列ならびに１行，３行は分割数を変えても変化しないが，２列と２行は(b)においては２つのクラスタに分離している．(a)２列目の特徴語のうち，species は(b)の２列と３列の両方に現れ，japan, sp, genus, nov といった分類学を連想させる語は(b)では２列目に，populations, population, genetic, females といった個体群生態学を連想させる語は(b)では３列目に，それぞれ分割された．一方，共通語を持つタイトル群である１行目は分割数を増やしても８分割までは分割されない頑健なクラスタであり，タイトルのクラスタリングだけでは分類学と個体群生物学の識別は難しい．(a)２列目の特徴語のうち，タイトルの特徴語でもあった語は全て(b)では２列目に現れているのも興味深い．これは，タイトルと概要の両方を使うことで分割されたものの関係を保ちつつ分割結果を表示できていることを示している．

タイトルと概要の両方でクラスタリングして抽出された特徴語から，分類学や個体群生態学といったいわゆる自然史に該当するものと生理学，発生学ないし分子生物学に該当するものとがクラスタとして見えることが分かった．(a)でいえば１行１列と３行３列がそれにあたる．一方，これらのクラスタ以外にはユニークな研究が見られた．たとえば，(a)１行３列の 27 の論文を調べると，ホヤによるバナジウム集積といったきわめて珍しい現象に関する研究などが含まれていた．行と列の特徴語の比較により，共通する語を持つセルにはよく言えば主流の，悪く言えばありきたりな研究が見つかり，そこから外れたセルには一風変わった研究が見つかることが分かった．

共通する特徴語による分析は同一の項目について分割数を変えたマトリクスについても可能であり，実際，前述のような分類学と個体群生物学の分離といった解釈を助ける上で有効なツールとなることが経験された．分割数の違いはすなわち粒度という観点の違いであり，本手法の新たな応用のひとつといえる．

５．まとめと今後の課題

電子ジャーナル Zoological Science の概要ページの複数項目に関して，任意の観点について検索を行い，任意に選んだ２つの観点から結果をマトリクス表示する多面的検索分析システムについてシステムの構成と，定性的評価実験について述べた．

特徴語抽出法の改良や，クラスタリング手法の選択基準検討が今後の課題である．提案手法の検索効率についての定量的評価も今後の課題である．

参考文献

[1] L.Guo, F.Shao, C.Botev, J.Shanmugasundaram. “XRANK: Ranked Keyword Search over XML Documents”, SIGMOD2003, 2003

[2] C.Yu, H.Qi, H.V.Jagadish. “Integration of IR into an XML Database”, INEX Workshop 2002, 2002

[3] 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

[4] 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，pp.1-4，2005