Vol.2011-IFAT-104 No.1
2011/11/22

クロス集計による文献ファセット検索システムの提案

廣川 佐千男, 酒井 敏彦, 曾 駿, 殷 成久 †1

研究を行う上で、関連文献の検索は必須だが、重要文献の検索だけでなく研究動向の把握にはファセット検索が有効である。本発表では、研究者、所属、雑誌、国際会議、年、キーワードなどの論文メタデータを分析軸とするクロス集計により検索結果を表示する文献検索システムを提案する。

Cross-Tabulation Search Engine for Scientific Articles

Sachio Hirokawa, Toshihiko Sakai, Jun Zeng, Chengjiu Yin †1

Search of related literature is indispensable in any research fields. Facet search is effective not only to search for important literature but also to grasp the research trend. The present paper proposes a faceted search engine which displays search results as the cross tabulation. Users can chose a pair of metadata for analysis axes, such as researchers, affiliations, journals, conferences , publication years, and keywords.

1. はじめに

膨大な文書群から目的の文書を探し出す検索エンジンは、今の社会のあらゆる局面で必須の道具となっている。Web ページを対象とする商用検索エンジンでは、ユーザーの要求に合致する結果のランキングが重要となっている。そもそも検索結果に合致するデータの量が膨大すぎるので、ユーザーが全てを見ることが不可能なので必須となっている。実際、日常的にちょっとした調べごとをするときには、検索結果の上位数件しか見ることはない。ところが、本稿で対象とするような研究活動における調査では、網羅的な検索が必要となる。文献だけでなく特許情報も範囲とすれば、網羅性はより重要となる。

研究において、自身の研究に関する先行研究や関連研究の調査は必須である。引用数や雑誌のインパクトファクターで論文の重要度を評価する試みもある。しかし、最新の論文については、必ずしもそのような評価が定まっているわけではない。従って調査を始めるに当っては、できるだけ網羅的にデータを集めなければならい。Web ページの検索と比較すると検索結果の数は少数なので、検索結果のすべてを個別に読むことは不可能ではないが、多くの労力が必要となる。

検索結果が多く関連が低いものが多いときには、分類して全体像を把握し、一部分についての絞り込みが必要となる。一方、検索結果が少な過ぎる場合には、別の検索語で検索を直さなければならない。このような、検索や分析の軸となる関連語を発見し、検索拡張や検索絞り込みを繰り返す過程こそ、関連研究調査のプロセスといえる。このプロセスを繰り返すことで、納得できる文献リストや分類指標が徐々にユーザーの頭の中に出来上がる。

こうして得られる文献リストを調査レポートとしてまとめる時の困難の第一は、分類指標である。単一テーマの研究であれば、ランキングで十分だが、一般的に一つのテーマであっても、目的や手法、評価方法など様々な観点があり、調査レポートでは何らかの形で構造化されたものが求められる。

近年、注目を集めているのがファセットに基く検索である。Scopus[⋆1]や DBLP[⋆2] などの文献検索システムで使われているだけでなく、一般のショッピングポータルサイト[⋆3]でも広く利用されている。検索デザインに関する[6] では、大規模ウェブサイトの検索機能の 3 大要素として、横断性、高速性と並んでファセット性をあげている。

本稿では、文献情報に対するファセット検索を提案する。従来のファセット検索では検索条件として各ファセットの条件を設定したり、ファセットごとの件数を提示することで、検索の効率を上げている。本稿では、検索条件設定よりも、二つのファセットを選択することで、クロス集計の形で提示する検索結果を可視化するインターフェースを提案する。

また、本稿では、一回ごとの個別の検索ではなく、関連研究調査という一連の検索活動を分析し、想定される調査結果レポートのための中間データがどのようなものか考察した。その結果、ファセットの選択や結果表示を組合せた幾つかの典型的な手順があることが分っ

†1 九州大学
Kyushu University

⋆1 http://www.scopus.com
⋆2 http://dblp.l3s.de
⋆3 http://www.amazon.com

た。本稿では、これらの手順を検索シナリオとよび、具体例を通じクロス集計が検索シナリオでどのように使われるかを示す。

2. 関連研究

[3] では、検索結果をまとめる方法として、クラスタリングとカテゴリ階層に代表されるファセットの比較を行っている。よく設計されたカテゴリ階層はクラスタリングより有効であると述べている。[5] では、Wikipedia に含まれるリンク構造とカテゴリを活用し、検索結果に対して動的に分類階層を提示する方式を提案し、有効性を示している。[1] では、アメリカ労働統計局の 67,000 件の文書を対象に、単純検索、カテゴリ階層検索、複数ファセット制約検索の３通りのインターフェースについて、単純検索、複合検索、発見的検索という困難度が違う検索を多数のユーザーにやってもらい、複雑な検索にファセットの有効性を定量的に検証している。[8] では、11 人の利用者の 4 週間のビデオ検索行動を分析し、キーワードを組合せた複雑な論理検索よりも表示されたファセットに対するクリックの方が 5 倍よく使われると分析している。一方、[4] では、OPAC データについて階層的ファセット表示をしたとき、入力キーワード、カテゴリ分類の単語、個別検索結果の３つ領域についてのユーザー視点滞在時間を観測し、ファセット部分は 1/4 未満しか注目されていないことが示されている。このことから、検索時にはファセットより個別検索結果の方を重視しているという報告もある。一般のテキストデータには、ファセットが附属しているわけではない。対象とする文書データに自明なファセットが内在しない場合に、どのようにファセットを付与するかという研究がある。[2] では、Wikipedia や WordNet などの外部資源からファセットを表すフレーズを求め、それを用いて一般のテキストデータファセットを付与する方式を提案している。

これらの研究では、複数ファセットの利用は検索条件指定に使われていたり、検索結果の階層的に利用されていても、本稿のようなクロス集計表のような表示には使われていない。

特許文書についてのクロス集計は、特許マップとして知られているが、本稿のように柔軟に観点を切り換えるものではない。検索結果をクロス集計で可視化するシステムとして[7] がある。そこでは縦軸、横軸は観点ごとのクラスタを表し、その解釈はクラスタの特徴語を使っている。クラスタリング法と特徴語抽出法の選択に多様性があり、何が最適かは容易には分らない。また、軸の特徴語は事前に決っていないので、解釈が困難な場合もある。本稿も論文概要から特徴語を抽出するところでは同様の問題がある。しかし、研究会名、年、著者登録キーワードなど、解釈が固定の属性もあることで、得られるクロス集計表の解釈に揺れが少いといえる。

本稿では、クロス集計表を生成するための二つのファセットはユーザーが決めなければならない。[9] では、Yahoo の画像検索[⋆1]で利用されている手法の紹介している。そこでは、検索条件を満す画像群について上位 k 個のファセットを求めるランキングを示している。この方式による上位 2 個のファセットでクロス集計することも考えられるが、今後の課題である。

3. システム概要

3.1 論文メタデータと索引 DB

本研究で構築した検索エンジンは、電子情報通信学会研究会[⋆2] で検索できた 2004〜2011 年の期間の 42921 件 (2011 年 8 月 26 日現在) の論文概要を対象とする。一件の論文には、図 1 のように、研究会記号、年度、番号、タイトル、著者、論文概要、ならびに著者によるキーワードが登録されている。これを元データとして、我々は検索エンジン GETA[⋆3] のインデックスを作った。Mecab を使った形態素解析により、題名と概要から名詞を抽出した。概要中の単語の他に、研究会名などのメタデータも索引として登録した。研究会は「g:xxx」、出版年は「y:xxx」の形の単語としてインデックス化した。図 1 の論文は研究会、年、番号が「nlp2004-1」なので、g:nlp,y:2004 という単語をインデックスとして登録した。著者名、著者キーワードは、それぞれ「n:xxx」「k:xxx」という形で登録した (図 2)。

(nlp2004-1) 高速計算を目指した ID 離散時間モデル
永嶋 宏和, 伝田 達明, 早川 吉弘, 中島 康治
高速計算を目指した ID 離散時間モデル ID(Inverse function Delayed) モデルは、出力関数に N 字型非線形逆関数を用いることで、負性抵抗領域を導入できる。このことにより、従来のニューラルネットワークで問題となっている、ローカルミニマム問題を回避できるという特徴をもつ。しかし、... ...
キーワード 負性抵抗領域 / 逆関数 / 遅延 / 離散時間 / 組み合わせ最適化問題

図 1 論文概要、論文メタデータの例

⋆1 http://images.search.yahoo.com
⋆2 http://www.ieice.org/ken/search/
⋆3 http://geta.ex.nii.jp/

Vol.2011-IFAT-104 No.1
2011/11/22

```
@nlp2004-1
1 y:2004
1 n:伝田達明
1 n:永嶋宏和
1 k:時間
1 k:組み合わせ
1 k:領域
..
1 k:遅延
1 k:化
1 k:関数
1 g:nlp
2 ため
1 サイズ
1 組み合わせ
...
@nlp2004-2
1 y:2004
1 n:吉田等明
...
```

図 2 文書インデックスの例

論文メタデータをこのように一般の単語と同じインデックスすることで、研究会や年を限定した検索や、検索結果からの特徴語として、メタデータが抽出できる (図 3)。

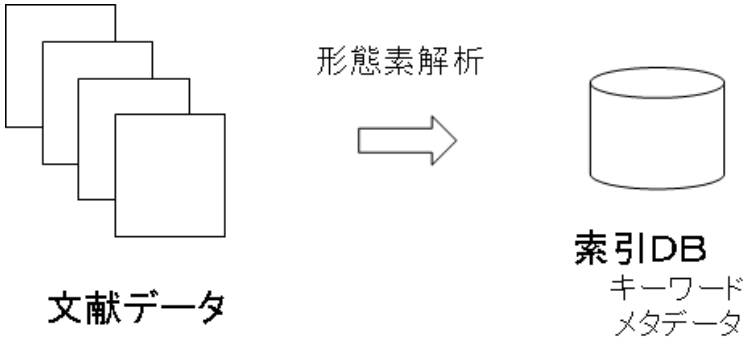

図 3 索引 DB

3.2 クロス集計表、反復的検索、階層的表示

図 4 と図 5 にそれぞれ提案システムの構成とインターフェースを示す。ユーザーは検索語の他に、分析の観点を規定する二つのファセットを選択する。システム内部では、検索結果リストの文献について、指定されたファセットの上位特徴語を求める。例えば指定されたファセットが研究会と研究者でそれぞれ分類数をいずれのファセットでも 5 個とするようなパラメーターの設定は図 5 の上部で行う。検索結果の文献の各ファセットの特徴語はその下の部分に表示される。研究会と研究者の上位単語がそれぞれ $g_1, \cdots g_5$、$n_1, \cdots, n_5$ だとすると、次にシステムは、$g_i \& n_j$ という検索を行い、検索結果の中で g_i と n_j の両方を満す文献の個数を求め、クロス集計表として、その下に表示する。一行目ならびに一列目には、選択した二つのファセットの特徴語が表示される。単語の後にある二つの数値は、検索結果中でその単語を含む文書数と全文書群における文書数である。クロス集計中の単語をクリックすると、その単語を検索語とする新な検索を行う。単語の後に二つ並んだ数値の一個目をクリックすると現在の検索条件にその単語を追加した絞り込み検索を行う。こうして、検索拡張と絞り込みをキーワード入力することく実現している。セル中の数値は、縦軸と横軸にある単語を含む文書数（検索結果中の）を示す。このセルをクリックすると、表 5 の左下の部分に該当する論文のタイトル一覧が表示される。さらにその中の各タイトルをクリックすると、右側にその論文の詳細が表示される。なお、特徴語の順位については、単純な出現頻度と SMART 法による順位を選択できる。また、年情報については、ユーザーの指定数とは無関係に全期間について提示している。

新な検索拡張や絞り込み、検索結果の俯瞰、着目する部分の詳細分析、具体的個別文献情報の確認、そして、分析観点の変更などの操作すべてについて、一番最初にキーワードを入力してしまえば、後は、表示画面で着目する部分をクリックするだけで、検索と分析を続けることができる。

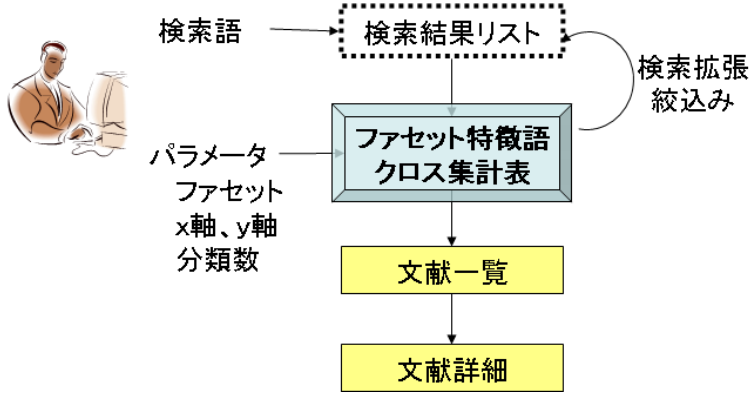

図 4 システム構成

本章に続く、4、5、6 章では、提案システムの各パーツの特徴を述べる。

Vol.2011-IFAT-104 No.1
2011/11/22

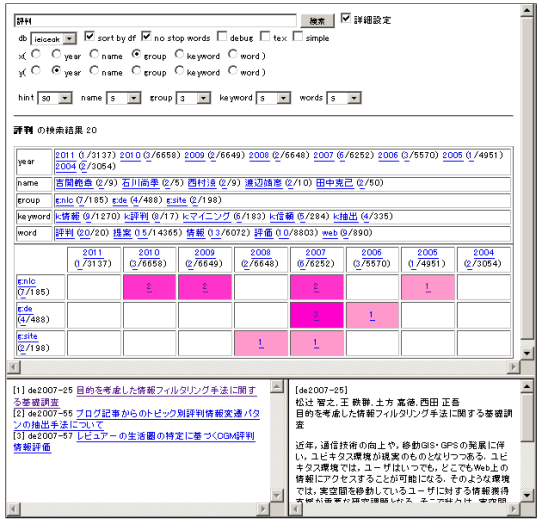

図 5 インターフェース

4. ファセット選択とファセット特徴語抽出

提案システムでは、検索語の他にクロス集計表で利用する縦軸と横軸の観点を指定する必要がある。今回対象とした論文データでは、発表年 (year)、著者名 (name)、研究会 (group)、著者登録キーワード (keyword)、タイトルと概要中の単語 (word) を選択することができる (図 6)。

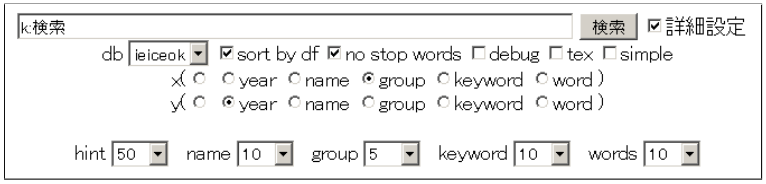

図 6 キーワード、パラメーター入力

それぞれの観点 (ファセット) での重要語を抽出し上位を表示することで、検索結果の解釈する手掛りを得ることができる (図 7)。ここに表示される単語をクリックすることで新な検索が可能となる。単語の横に表示された数値をクリックすると、現在の検索条件にその単語を追加した絞り込み検索ができる。

year	2011 (10/3137) 2010 (52/6658) 2009 (59/6649) 2008 (55/6648) 2007 (58/6252) 2006 (41/5570) 2005 (34/4951) 2004 (31/3054)
name	田中克己 (13/50) 小山聡 (9/18) 山井成良 (8/26) 木下和彦 (8/47) 村上孝三 (8/96) 山名早人 (7/21) 大月一弘 (6/13) 中村裕一 (6/46) 酒井善則 (6/34) 太島裕明 (5/14)
group	g:de (65/488) g:pmu (56/1793) g:ns (24/1632) g:in (22/1655) g:ie (21/689)
keyword	k:検索 (340/340) k:情報 (96/1270) k:画像 (55/1007) k:web (41/406) k:類似 (35/108) k:システム (34/1430) k:映像 (31/336) k:データ (30/554) k:- (26/2813) k:p (21/386)
word	検索 (303/793) 提案 (243/14365) 情報 (173/6072) 手法 (172/8647) 本稿 (164/9327) 結果 (153/10250) システム (146/8341) 利用 (128/7021) 実験 (124/6538) 可能 (113/8821)

図 7 特徴語抽出

5. クロス集計と階層的表示

表 8 の上部は、縦軸、横軸をユーザーが選択した観点に基き検索結果をクロス表として表示する。縦軸、横軸ともに件数の降順 (sort_by_df) か、または、単語の重要度の降順で特徴語を表示することができる。この例では、研究会を縦軸、年を横軸としている。年は降順に固定している。件数の多い各セルは濃い色で表示されている。セル中の数値は検索結果の文献のうち縦軸、横軸の単語を満す件数を表している。この数値をクリックすると、該当する論文のリストがが左下の領域に表示される。更に、左下の論文タイトルをクリックすると、右側にその論文の情報が提示される。このように、同一画面内で、クロス表として全体像を把握し、それぞれの部分の一覧を確認し、特に注目する論文の概要をクリックだけで見ることができる。

6. ファセットの選択と切替による検索シナリオ

関連研究の調査や分野のサーベーでは、一回の検索操作ではなく、連続する一連の検索作業に意味がある。毎回同じ目的で検索するのではなく、最終的なレポート作成に向けて必要となる部分的調査を組合せることで全体像が明かになる。一連の操作の中には、全体を把握するための上昇指向の分析と、部分構造を捉えるための特徴語や、各部分をさらに絞り込むための手掛り語の抽出という二つの方向の操作が必要となる。ファセット機能やクラスタリング機能や間連語提示機能がない検索システムでは、ユーザーは、ブックマークを残したり、別のタブやウィンドを開いたり、検索過程のメモを別途残さなければならない。そし

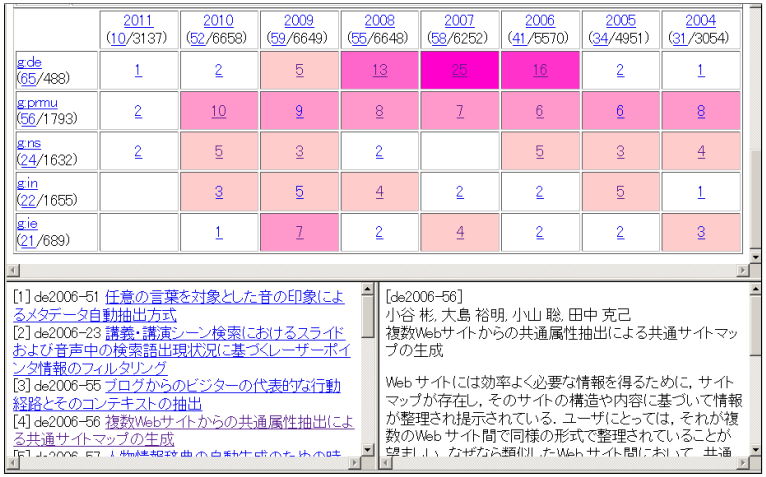

図 8 クロス表と階層的表示

て、ある段階でそれらのメモをもとに、最終的な調査レポートをまとめなければならない。論文情報の検索においても、一つあるは数編の論文を探すだけなら、このようなメモを残したりレポートまで書く必要はない。しかし、本システムのように、網羅的な調査では中間的な記録が重要である。

本稿では、論文調査で必要となる中間的レポートには特定の様式がある点に着目した。一番単純なものは検索結果の論文のリストである。ただし、どのような順序のリストとするかは、重要度順、著者順、発表年順など何通りか異なるリストが考えられる。個々の論文だけでなく、どのような研究者や研究グループがいるかという事も、関連研究調査として必須である。自身の研究の位置付けを捉えるためには、他の論文全体を時間軸で並べ、それぞれの時点の特徴語で研究動向を見ることも重要である。個別の論文を正確に評価するためには、具体的に論文本文を読まなければならない。しかし、効率よく大局的に論文を比較するためには、著者が与えたキーワードや、タイトル、概要中の特徴語により比較するしかない。分析に使うファセットを切り替えることで、対象論文を分類しあるいは絞り込み、その結果、特徴語によるクラスタリング、研究者グループの抽出、時系列による研究動向分析などの検索の中間結果が得られる。このような分析手順は、従来の単純な検索エンジンにはない検索シナリオといえる。

次章では、検索シナリオの具体例を述べる。

7. 分 析 事 例

検索語から始め、まず、研究会、研究者、著者キーワード、年などの単独のファセットを見ることで重要ファセットを抽出し、次に、個々のファセットの内容を見ることで、論文や研究者の塊をみることができる。横軸を年とすることで、研究者や研究グループ、研究会、特定のキーワードに関る研究の趨勢を見ることができる。ある程度の予想が考えられる場合には、論文リストや各論文概要をみることで、その予想の確認ができる。

本章では、どの研究会が適当か探すシナリオ、研究者グループを探すシナリオ、研究動向を分析するシナリオの３つの目的に応じてどのようにファセットを選択するかを、具体例について述べる。分析事例として、集合知によるテキストマイニングとして評判情報の分析をする場合を考えてみる。

7.1 研究会選択シナリオ

新なテーマについての研究を始めるとき、どの研究会が適当か調査する必要がある。評判情報に関連する研究であれば、「評判情報」、「集合知」、「テキストマイニング」などが検索語として考えられる。まず、「評判情報」だと検索結果が 0 件なので、「評判 情報」で検索し、13 件の結果が求まる。分析目標が研究会 (group) なので縦軸、横軸ともにファセットを group として図 9 が得られる。研究会としては、nlc、de、ai の 3 つの研究会に多いことが分り、それぞれのセルをクリックすると、該当する論文リスト (表 1) が得られる。この表自体が関連研究の分析ともいえる。

7.2 研究者グループ分析シナリオ

次に、どのような研究者や研究者グループがこの分野の研究を行っているか見るため、ファセットを研究者 (name) とする。上位 10 人に広げたクロス表 (図 10) を見ると、5 つのグループが分る。色がついた長方形が共著論文を出している研究者グループになっていることが分る。これは、形式概念の理論 でいうところの概念になっている。ただし、同じグループなのに離れて表示されることもある。行や列の入れ換え (matrix re-ordering) などで直感的に分りやすい表示する改善は今後の課題である。

7.3 研究動向分析シナリオ

次に、研究動向を見るため、横軸を年 (year)、縦軸を著者キーワード (keyword) として図 11 が得られる。図 11 でも、横軸を年 (year)、縦軸を研究会 (group) とした図でも、いずれも 2007 年をピークとしてその後、減っていることが分る。

nlc(4/185)		
1	nlc2007-87	各属性のレビュー・評価値の関係を用いた評判情報の検索支援
2	nlc2007-88	構文片を用いた分野の同定を必要としない意見・評判情報抽出
3	nlc2005-28	日本語と英語の文タイプの自動付与とその特徴素
4	nlc2009-51	集合知を用いた物体認識及び評判情報の取得
de(4/488)		
1	de2007-25	目的を考慮した情報フィルタリング手法に関する基礎調査
2	de2007-55	ブログ記事からのトピック別評判情報変遷パタンの抽出手法について
3	de2007-57	レビュアーの生活圏の特定に基づく CGM 評判情報評価
4	de2006-1	単語の印象を考慮した言い換え処理に基づくクエリ展開
ai(2/364)		
1	ai2011-TBD1591	評判情報に基づく言語サービスの選択
2	ai2010-38	観光イベントについての「といえば検索」の提案

表 1 研究毎の「評判 情報」関連論文のリスト

8. まとめと今後の課題

本稿では、2004〜2011 年の信学会研究報告 42921 件を対象として、研究会記号、年度、番号、タイトル、著者、論文概要、ならびに著者によるキーワードを観点とするファセット検索システムを提案した。ユーザーが指定する二つのファセットに基き、検索結果をクロス集計表として可視化する。最初に検索語を入力する必要はあるが、それ以降の検索と分析では、ファセットの選択や切替、着目するセルに該当する文献一覧表示、検索拡張と絞り込みなどすべての操作が、表示画面で注目する部分のクリックだけで行うことができる。

本稿では事例を述べただけなので、今後、操作性の定量的評価が必要である。現在の実装では、一行目と一列目に表示する各ファセットの特徴語の順序は、出現頻度か SMART スコアで決めているが、例えば再配置することでクラスタとしての認識性向上が考えられる。着目部分のクリックについては、一覧表示と再検索の二通りの操作が考えられる。今は、表示場所に応じて一意に決めているので、柔軟性を持たせることが考えられる。

参 考 文 献

1) Capra, R., Marchionini, G., Oh, J.S., Stutzman, F., Zhang, Y., Effects of structure and interaction style on distinct search tasks, Proc. ACM ICDL, pp. 442-451, 2007
2) Dakka, W., Ipeirotis, P.G., Automatic extraction of useful facet hierarchies from text databases, Proc. ICDE, pp. 466-475,2008
3) Hearst, M.A., Clustering versus faceted categories for information exploration, CACM, Vol.49, No.4, pp. 59–61, 2006

図 9 「評判」関連研究 group*group

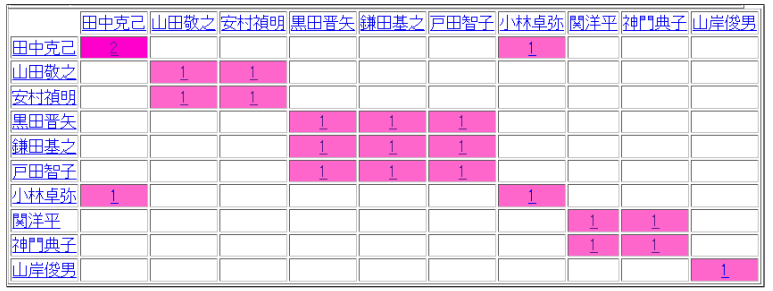

図 10 「評判」関連研究グループ

4) Kules, B.,Capra, R.,Banta, M.,Sierra, T., What do exploratory searchers look at in a faceted search interface?, Proc. ICDL, pp. 313-322, 2009
5) Li, C., Yan, N., Roy, S.B., Lisham, L., Das, G., Facetedpedia: Dynamic generation of query-dependent faceted interfaces for Wikipedia, Proc. WWW 2010, pp.

Vol.2011-IFAT-104 No.1
2011/11/22

	2011 (1/3137)	2010 (1/6658)	2009 (1/6649)	2007 (6/6252)	2006 (2/5570)	2005 (1/4951)	2004 (1/3054)
k:情報 (9/1270)		1	1	5	1	1	
k:評判 (6/17)			1	4	1		
k:マイニング (5/183)				2	2		1
k:分類 (3/149)				2	1		
k:抽出 (3/335)		1		1	1		

図 11 「評判」関連研究動向

651-660,2010

6) Peter Morville, Jeffery Callender, Search Patterns:Design for Discovery, O'Reily, 2010

7) T. Seki, T. Wada, Y. Yamada, N. Ytow, S. Hirokawa, Multiple Viewed Search Engine for an e-Journal - A Case Study on Zoological Science, Proceeding of 12th International Conference on Human-Computer Interaction, Part 4, pp.989-998, 2007

8) Wilson, M.L.,Schraefel, M., A longitudinal study of exploratory and keyword search, Proceedings of the ACM International Conference on Digital Libraries, Proc. ICDL, pp. 52-55, 2008

9) Roelof van Zwol, Borkur Sigurbjornsson, et. al., Faceted Exploration of Image Search Results, Proc. WWW 2010, pp.961–970, 2010