単語共起関係と共著関係を利用した論文探索システムの提案

風間 一洋[†, a]　石橋 和樹[††, b]　篠田 孝祐[††, c]　栗原 聡[††, d]

† 和歌山大学システム工学部　†† 電気通信大学大学院情報システム学研究科

a) *kazama@ingrid.org* b) *s151002@center.wakayama-u.ac.jp*

c) *kuri@is.uec.ac.jp* d) *kosuke.shinoda@riken.jp*

概要　本稿では，単語共起関係と共著関係を利用することで，論文の網羅的な探索を支援する方法を提案する．まず，単語共起関係から求めた検索結果の論文群に関連する関連語と，共著関係から求めた関連著者を提示することで，単語–論文–著者という 3 部グラフ構造における論文の探索を容易にする．次に，共著者ネットワークのコミュニティ構造に基づいて単語を順位付けすることで，研究分野を指定するのに適した専門的な関連語を選択できるようにする．さらに，共著関係に基づいて論文をグループ化すると共に，グループの代表的な著者と，その共著者を提示することで，検索結果の概要の把握を容易にする．実際に，2003~2014 年の 12 年間の人工知能学会全国大会の講演データを収集し，その論文アーカイブを探索するシステムを試作して，提案手法の有効性を検討した．

キーワード　論文検索, 単語共起, 共著関係, 3 部グラフ, TF-ICF

1　はじめに

インターネットの普及と共に技術情報やソフトウェアの公開や共有が進んだことにより，研究開発の進行と，生まれた技術の陳腐化がますます早くなっている．学会の論文投稿は実世界の研究開発の要求を反映していることから，その研究動向を的確に把握することは，研究開発の方向性を決めたり，有用性を探るために重要であると考えられる．例えば，論文の研究動向は，CiNii や ACM Digital Library のような論文検索システムで，調べたい研究内容を表すさまざまなキーワードで検索すれば調べることができる．しかし，研究動向を調べるためには，種々のキーワードを用いて試行錯誤的に検索を繰り返す必要があり，対象とする研究に対する網羅的な専門知識を持たないユーザの場合は，そのような探索的な論文検索をおこなうのは困難である．

そこで本稿では，単語共起関係と共著関係を利用して，論文アーカイブを容易に探索できるシステムを提案する．まず，単語共起関係から求めた検索結果の論文群に関連する関連語と，共著関係から求めた関連著者を提示することで，単語–論文–著者という 3 部グラフ構造における論文の探索を容易にする手法について述べる．次に，共著者ネットワークのコミュニティ構造に基づいて単語を順位付けすることで，専門的な関連語を選択できるようにする手法について述べる．さらに，共著関係に基づいて論文をグループ化すると共に，グループの代表的な著者と，その共著者を提示することで，検索結果の傾向の把握を容易にする手法について述べる．

2　単語共起関係と共著関係を用いる情報探索

2.1　単語共起関係と情報探索

論文検索システムでは，指定されたキーワードを含む論文を検索する機能を提供している．これは論文に含まれる単語と論文の間に，単語–論文という 2 部グラフ構造が存在し，単語が決まれば，その単語が出現する論文も決まることを示している．

つまり，情報探索の視点を変えるためには，単語を切り替えればよい．この際に連続的に視点を変更する一つの方法は，関連語を用いることである．関連語を求めるさまざまな方法が提案されているが，本稿では論文における単語共起関係から求めることとする．これは，ある単語が出現する論文集合を求め，さらにそれらの論文集合に出現する単語集合を求めることに相当する．

2.2　共著関係と情報探索

論文検索システムでは，ある著者によって執筆された論文を検索する機能も提供している．これは，著者と論文の間にも，著者–論文という 2 部グラフ構造が存在し，著者が決まれば，その著者によって書かれた論文も決まることを表している．

なお，論文は同じ学科や同じ研究プロジェクトに属する複数の著者によって共同で書かれることが多く，同一グループの著者は類似した研究テーマで執筆する傾向がある．つまり，情報探索の視点を変えるためには，著者または著者グループを切り替えればよい．

2.3　3 部グラフ構造の利用

すなわち，論文データには図 1 のような単語–論文–著者という 3 部グラフ構造が存在する．これは，この 3 部グラフ構造のエッジを辿ることで，単語と著者の両方の

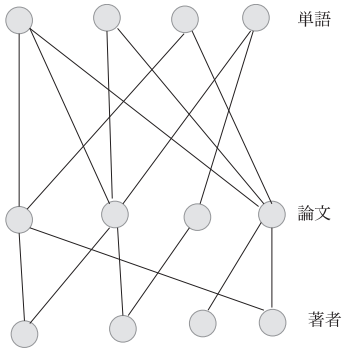

図 1 単語–論文–著者の 3 部グラフ構造

視点から論文を探索できることを示している．

このような 3 部グラフ構造を使った効率的な情報探索を支援するためには，現在注目している研究トピックに関連している単語または著者の推薦と，単語または著者の視点からの効率的・柔軟な論文の絞り込みが必要になると考えられる．

そこで，研究トピックに関連している単語を推薦するために，この 3 部グラフ構造を利用して，ある単語が出現する論文を著者の共著ネットワーク上のコミュニティ構造に射影する定量化指標を用いることで，検索結果の論文群に関連し，その研究分野を指定するのに適した関連語を提示する．

また，著者の視点から効率的・柔軟に論文を絞り込むために，共著関係に基づいて論文をグループ化すると共に，グループの代表的な著者と，そのより代表的な共著者を提示する．

3 関連研究

Paul らは，Web ページから特徴的な単語を抽出し，その単語共起の度合いに基づいて分散したサーバ上の情報を効率的に探索することができる広域情報探索インフラストラクチャIngrid を提案した [1]．風間らは，Web ページから人名とそれらの間の共起関係を自動抽出し，検索結果から作成した人間関係ネットワーク構造を手がかりに情報を探索する方法を提案した [2]．本稿で提案する手法は，単語共起関係と共著関係（人間関係）の両方を用いて情報探索をおこなうという点において，これらの手法を発展させたものであると言えるが，さらに関連語の抽出や検索結果のグループ化にも適用している点に新規性がある．

高野らは汎用連想検索エンジン GETA を用いた対話的な連想検索を提案した [3]．これでは，文書群同士・単語群同士・文書群–単語群間の類似性関連性に基づく連想計算に基づくものであり，文書から関連単語，関連単語から文書という処理の合成で実現される．本稿で提案する手法では，単語からの視点では同様な関連性に基づく探索を支援するが，著者からの視点ではそのような意味的類似性ではなく，研究者のコミュニティという別の種類の支援を試みる点が異なる．

4 論文データベースの作成

4.1 講演データの収集

人工知能学会 [1] は，毎年全国大会を開催しており，発表プログラムと論文の PDF を参加者に CD-ROM で配布すると共に，Web でも講演プログラムを公開している．これには，すべての講演の時間，演題番号，題目，著者に加えて概要も掲載されているので，wget コマンド [2] を用いて，2003∼2014 年の 12 年間の講演データを収集し，試作システムのデータとして用いた．なお，1999 年度（第 13 回）以降の発表プログラムが公開されているが，2001 年度までは概要がないこと，2002 年度は概要は掲載されていてもファイル構成が大きく違うことから，2003 年度以降を対象とした．

4.2 講演情報の抽出

収集した講演データから，Python で記述したプログラムを用いて，時刻，演題番号，題目，著者と概要などの書誌情報を抽出した．

まず，講演情報から，正規表現を用いて演題番号，題目，著者名，所属，時間，概要を抽出する．さらに次の手順で，題名と概要からキーワードとその出現頻度を抽出した．

1. 題目と概要のテキストを MeCab[4] を用いて日本語形態素解析する．
2. 形態素の大文字・小文字，全角・半角，カタカナ語の末尾の長音記号の有無などの違いを正規化する．
3. 伊藤らの手法 [5] と同様に，抽出された形態素の IPA 品詞体系 [6] における品詞情報に基づいて，名詞や接頭詞，接尾辞を連結して，複合名詞を抽出する．接頭詞は名詞接続のみ，接尾は助動詞語幹と副詞可能を対象とする．名詞は，ナイ形容詞語幹，形容動詞語幹，接続詞的，代名詞，動詞非自立的，特殊，非自立，副詞可能を除外した．
4. MeCab の品詞自動推定機能で名詞と誤判定されることが多い単独の記号類は，ストップワードリストを用いて除去する．

論文探索システムからアクセスすることを考えて，抽出された情報は MongoDB[3] に格納した．このデータベースには，他にも単語の出現数や出現文書数や，後述するような単語の ICF の値や，複合名詞の部分一致関係も格納する．

[1] http://www.ai-gakkai.or.jp
[2] https://www.gnu.org/software/wget/
[3] http://www.mongodb.org

5 論文探索システム

5.1 論文検索

本システムでは，キーワードまたは著者名を用いて，論文データベース中の論文を検索する機能を提供する．なお，現在利用しているデータは論文の全文ではなく題名・概要だけであることと，関連キーワード推薦とキーワード検索の挙動を一致させるために，全文検索ではなく，データベースに格納されたキーワードを用いて検索する．

ただし，再現率を向上させるために，次の二つの改良をおこなった．

5.1.1 キーワードの正規化と頻出形の表示

表記のゆれに対応するために，キーワードを正規形に変換してからデータベースに登録する．たとえば,「LDA」の正規形は「lda」,「オントロジー」の正規形は「オントロジ」である．ただし，このような正規形をそのまま表示しても，理解しにくかったり，違和感があることが多い．

そこで，まず正規形が同じキーワードの元の表記の出現頻度を求めて，一番頻繁に用いられている形式を表示形として，正規形と表示系の対応をデータベースに格納する．つまり，キーワードは入力された時点で正規化され，以後システム内部では正規形に統一して扱う．ただし，ユーザには，このデータベースを用いて正規形から表示系に変換してから表示する．

5.1.2 複合名詞の部分一致

本稿で用いている複合名詞抽出法では，例えば「マルチエージェントシミュレーション」のような長い複合語が抽出される傾向があるが，このままでは「エージェント」,「シミュレーション」では検索できないことになる．

そこで，入力されたキーワードに部分一致するキーワードでも検索可能にした．まず，以下の手順で各キーワード w_i に部分一致するキーワード集合 $W^{pa} = \{x|x = W_i^{pa}\}(i = 1, \ldots, N)$ を求めた．

1. 全単語集合 W から単語 w_i を取り出す．すべて処理し終えたら 5 に移動する．
2. w_i と前方一致する単語集合 W_i^{pr} を求める．
3. w_i と後方一致する単語集合 W_i^{po} を求める．
4. 1. に戻る．
5. $W^{pr} = \{x|x = W_i^{pr}\}$ と $W^{po} = \{x|x = W_i^{po}\}$ から，w_i と部分一致する単語集合 W_i^{pa} を求める．この部分一致は，文字単位ではなく，実際に使われる言語単位に基づく．

この対応関係を用いることで,「エージェント」というキーワードから,「マルチエージェント」,「エージェントシステム」,「マルチエージェントシステム」も検索できるようになる．

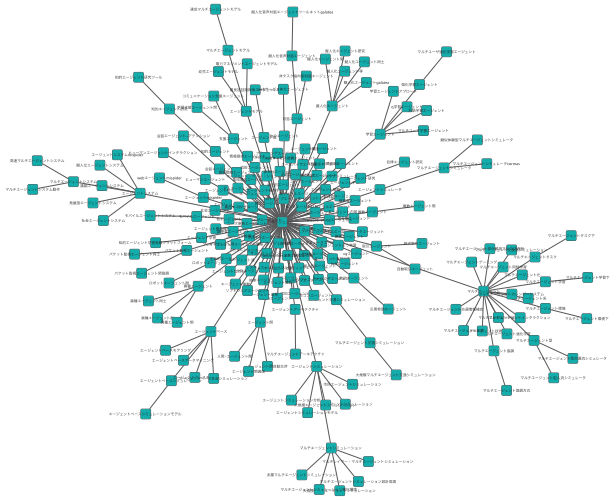

図 2 「エージェント」から派生する複合語のネットワーク構造

参考のために，上記手順で作成された「エージェント」から派生する複合語の木構造を図 2 に示す．中心部の単語が「エージェント」であり，それからさまざまな単語が派生していることがわかる．分岐状況の違いは，そのノードに対応する複合語のユニット性に関係すると思われる．なお，右にある少し大きな構造は「マルチエージェント」であり，これは「エージェント」とは少し違う文脈で使われていることを示していると思われる．

5.2 システム構成

ユーザが論文を検索・探索する場合には，Web ブラウザを使って Apache HTTP Server にアクセスすることで，mod_wsgi の daemon モード [4] で待機しているプログラムを呼び出すことができる．

プログラム本体は Python で記述されており，データベースとして MongoDB に格納された講演データを検索しながら動作する．なお，データアクセス回数は一般的な検索システムよりもかなり多くなるので，比較的小規模なデータについては，サービス起動時にメモリ上に読み込むことで，システムを高速化している．

5.3 システム実行例

単語「SVM」で検索した例を，図 3 に示す．

通常の検索システムと異なる点は，キーワードまたは著者名で検索された論文を表示するだけではなく，各年の検索結果数を棒グラフとして表示すること，探索する情報を切り替えるための主な共著者や主な関連キーワードを提示すること，検索結果をグループ化して代表的な著者と一緒に表示することである．以降で，コミュニティ性に基づく関連語提示と共著関係に基づく論文のグループ化について詳しく説明する．

[4] https://code.google.com/p/modwsgi/

図 3 「SVM」で検索した例

6 コミュニティ性に基づく関連語提示

6.1 単語のコミュニティ性

コミュニティ性は，ある単語がどのような著者達のコミュニティで活用されているかについての性質である．例えば，一般的な単語は世の中で広く使われるのに対して，特殊な単語は使う人たちが限られるはずである．さらに，専門用語は関連する専門家のコミュニティの中だけで頻繁に使われると思われる．つまり，単語が人間のコミュニティでどう使われているかという性質に応じて，人間のコミュニティ構造における単語の出現分布パターンに固有の特徴が生まれると考えられる．

そこで本稿では，「専門用語とは，特定の専門家達の間で共有される言葉である」という前提に基づいて，著者のコミュニティ群における単語の分布から専門性を定量化する指標である ICF(Inverse Community Frequency) を用いる．さらに，検索結果に対する関連語の選択のために，コミュニティ性に基づいて単語の専門性を定量化する指標 ICF を用いたスコア TF-ICF を用いた [7].

6.2 ICF

単語 w_i の ICF の値 $ICF(w_i)(1 \leq i \leq K)$ は以下のように計算する．ここで K は総単語数である．

1. データセットに含まれる全論文の共著者同士にエッジを張って構築した共著関係ネットワークを Clauset らの CNM 法 [8] でコミュニティに分割し，著者と所属コミュニティの関係を求める．
2. 単語 w_i が出現する論文の筆頭著者の集合から，著者と所属コミュニティの関係を用いて単語 w_i が出現するコミュニティ数 $c(w_i)$ を求める．
3. 単語 w_i のコミュニティ集合における出現率 $r(w_i)(0 \leq r(w_i) \leq 1)$ を計算する．

$$r(w_i) = \frac{c(w_i)}{C} \tag{1}$$

ここで，C はクラスタリングによって得られた全コミュニティ数である．

4. 単語 w_i の ICF の値 $ICF(w_i)$ を計算する．

$$ICF(w_i) = (log(\frac{1}{r(w_i)}))^{\alpha} \tag{2}$$

ここで，α は定数である．

6.3 TF-ICF

さらに，単語 w_i の TF-ICF 値である $TF\text{-}ICF(w_i)$ を次のように計算する．

$$TF(w_i) = \frac{n(w_i)}{\Sigma_{k=1}^{K} n(w_k)} \tag{3}$$

$$TF\text{-}ICF(w_i) = TF(w_i) \times ICF(w_i) \tag{4}$$

ここで，$n(w_i)$ は単語 w_i の出現回数である．

情報探索やテキストマイニングなどの分野で利用される同様な指標として、文書中に出現した単語がどのくらい特徴的であるかを識別するための指標である TF-IDF が存在する [9]．しかし，本システムで検索結果に出現する単語とその出現頻度から TF-IDF を用いて関連語を求めた場合には，探索に有効な専門用語よりも，論文特有の「提案」,「手法」,「分析」などの単語のスコアが高くなりがちだった．これは，論文の本文ではなく題名と概要しか利用できないことから専門用語の TF が低く抑えられ，複数の論文に対して適用するために一般的な用語の TF が高くなってしまったからと考えられる．

このような場合に ICF を用いれば，研究者の論文生産性の違いの影響や共同研究のチーム規模の影響を受けにくくなる．さらに，TF-ICF における ICF の α を調節することで TF と ICF のバランスを変更し，ランキング結果の特性を調節することができる．なお，ICF では単語を論文空間に射影してから，さらに著者コミュニティに射影するが，一般的に論文数より著者コミュニティ数の方がかなり少ないために，IDF よりも小さくなりがちな ICF の影響を α を大きくして強めることができる．

ただし，TF-ICF を用いるためには，図 1 で示したような単語–論文–著者という 3 部グラフ構造が存在しなければならない．このために，すべての種類のデータに適用できるわけではないが，例えば同様な 3 部グラフ構造を作成できるソーシャルメディアには適用可能である．

6.4 TF-ICF と TF-IDF の比較

TF-ICF と TF-IDF の違いを知るために，「ロボット」の検索結果に対して提示される関連語の違いを調べる．なお，$\alpha = 2$ とした．

TF-ICF を用いた場合の上位 20 件の関連語は「ロボット, インタラクション, 人間, 発話, 操作者, 移動ロボット, 物体, 人, 動作, 対話, コミュニケーションロボット, 対話ロボット, 自律ロボット, 知能ロボット, ヒューマノイドロボット, 獲得, 学習, 行動, ユーザ, 物体概念」, TF-IDF を用いた場合の関連語は「ロボット, 人間, 人, 実現, インタラクション, 学習, 実験, 行動, 獲得, 動作, 研究, 提案, ユーザ, 本稿, 発話, 対話, 物体, 構築, 手法, 影響」となり, TF-ICF の方が具体的な研究テーマを示すようなキーワードを提示できていると言える.

7 共著関係に基づく論文のグループ化

7.1 論文のグループ化と共著関係

論文検索システムでは, 検索結果は発表日や被引用件数に基づいて順位付けされることが多い. これは新規に投稿された論文や, 有用と考えられる論文を発見する場合には適している. ただし, 研究動向を知りたい場合には, 必ずしも使いやすいとはいえない.

このような場合に, 検索結果をグループ化する手法がさまざまな手法が提案されている. 風間らは, 検索結果の Web ページのディレクトリ階層に着目してグループ化し, さらにその索引としてふさわしい Web ページを表示する手法を提案した [10]. 馬場らは検索エンジン基盤 TSUBAKI を用いて, Web ページに含まれる複合名詞に着目してクラスタリングする手法を提案した [11].

本稿では, 著者の共著関係に注目する. 一般的に, 研究者は固有の研究テーマを持ち, 類似した研究テーマの他の研究者と共同で研究をおこなったり, 外部資金を獲得する. 研究者が複数の研究テーマを持つことも多いが, その場合は研究テーマに応じて異なる研究者と共同研究することになる. このような研究アクティビティによって論文をグループ化できれば, 単なる研究内容や手法の類似性ではなく, 研究方針や研究に用いるインフラなどの類似性で検索結果を概観できるはずである.

そこで, キーワード検索の検索結果から代表的な著者を発見して, それに基づいてグループ化して表示する手法を提案する.

7.2 グループ化手法

検索された論文のグループ化の手順は以下の通りである.

1. キーワードで検索し, 論文リスト P を取得する.
2. 論文リスト P の各著者（共著者も含む）の執筆論文数をカウントし, 得られた著者リスト A を論文数の降順にソートする.
3. 著者リスト A から著者 a_i を取り出し, 著者名と執筆論文数, そして共著リスト C_i から, より執筆論文数がより多い共著者を出力する. リストが空の場合は終了する.
4. 論文リスト P から論文 p_i を取り出す. リストが空の場合は終了する.
5. 論文 p_i の著者に a_i が含まれない場合は, 未処理リスト R に追加する.
6. 論文 p_i の著者 a_j が a_i と異なるならば, a_j の共著リスト C_j に a_i を追加する.
7. 論文 p_i の情報を出力する.
8. $P = R$ とする.
9. 4. に戻る.
10. 3. に戻る.


論文 (67件) :

鳥海不二夫 (12件),

1. 松本慎平, 川口大貴, 鳥海不二夫: Twitter利用者の投稿活動に基づく定量化手法の災害情報支援システムに向けての展望 (2014)
2. 風間一洋, 鳥海不二夫, 榊剛史, 栗原聡, 篠田孝祐, 野田五十樹: Twitterのイベントの因果関係の分析 (2014)
3. 馬場正剛, 鳥海不二夫, 榊剛史, 篠田孝祐, 栗原聡, 風間一洋, 野田五十樹, 大橋弘忠: 災害情報の分類の妥当性の評価 (2014)
4. 池田圭佑, 岡田佳之, 榊剛史, 鳥海不二夫, 風間一洋, 野田五十樹, 篠田孝祐, 諏訪博彦, 栗原聡: マルチエージェント型情報拡散モデル(AIDM)の提案 (2014)
5. 岡田佳之, 榊剛史, 鳥海不二夫, 篠田孝祐, 風間一洋, 野田五十樹, 沼尾正行, 栗原聡: 拡張SIRモデルによるTwitterでのデマ拡散過程の解析 (2013)
6. 松澤有, セーヨーサンティ, 鳥海不二夫, 陳昱, 大橋弘忠: リツイート時系列の3パラメータ混合対数正規分布による分析 (2013)
7. 原久美子, 木野泰伸, 鳥海不二夫: 字・町名をキーとした災害時Twitter情報の抽出と地図への展開 (2013)
8. 松本慎平, 川口大貴, 鳥海不二夫: 情報量に基づく投稿活動定量化手法を用いた東日本大震災前後のTwitter利用者の特徴付け (2013)
9. 石原裕規, 諏訪博彦, 鳥海不二夫, 太田敏澄: 震災時の情報流通を支えるTwitterアカウントの発見 (2013)
10. 白井滉士, 榊剛史, 鳥海不二夫, 篠田孝祐, 風間一洋, 野田五十樹, 沼尾正行, 栗原聡: Twitterにおけるデマツイートの拡散モデルの構築とデマ拡散防止モデルの推定 (2012)
11. 風間一洋, 鳥海不二夫, 榊剛史, 篠田孝祐, 栗原聡, 野田五十樹: 東日本大震災時のTwitterデータを用いた単語間の関係の時系列変化の分析 (2012)
12. 白井翔平, 鳥海不二夫, 石井健一郎, 間瀬健二: 震災による情報伝播ネットワークの変化 (2012)

榊剛史 (9件), ⇒ 鳥海不二夫

1. 丸井淳己, 則のぞみ, 榊剛史, 森純一郎: 分散表現を用いたコミュニティにおける単語使用傾向の分析 (2014)
2. 丸井淳己, 榊剛史, 松尾豊: ソーシャルメディアと時系列 データを用いたイベント抽出及び自動ニュース生成に関する研究 (2013)
3. 榊剛史, 松尾豊: ソーシャルブックマークとしてのTwitterリスト機能の応用 (2010)

風間一洋 (7件), ⇒ 鳥海不二夫

1. 風間一洋, 今田美幸, 柏木啓一郎: Twitterの情報伝播ネットワークの分析 (2010)

松尾豊 (4件), ⇒ 榊剛史

1. 那須野薫, 松尾豊: Twitter における候補者の情報拡散に着目した国政選挙当選者予測 (2014)
2. 保住純, 飯塚修平, 大澤昇平, 中山浩太朗, 高須正和, 鵜飼絵理子, 須賀千鶴, 西山圭太, 松尾豊: Webマイニングを用いたコンテンツ消費トレンド予測システム (2013)

斉藤和巳 (3件),

1. 大原剛三, 斉藤和巳, 木村昌弘, 元田浩: Twitter上の情報拡散系列からの変化点検出 (2013)
2. 加藤翔子, 小出明弘, 伏見卓恭, 斉藤和巳: ジニ係数による多重有向グラフとしてのFavoritesネットワークの分析 (2012)
3. 小出明弘, 斉藤和巳, 長屋隆之, 伊藤健二: ネットワーク粗視化による情報拡散過程の可視化法 (2012)

図 4 「Twitter」の検索結果のグループ化例

グループ化の結果, 共同研究している研究者がグループ化されると共に, 一番論文数が多い代表的な著者が選択される. なお, グループ間に重複が存在する場合は, より論文数が多い著者のグループの論文として表示され, それより論文数が少ない著者には, それらに含まれない論文だけが表示される. この結果, 共同研究しているグループとそれらの間の関係が示されることになる.

7.3 グループ化の例

「Twitter」で検索した時のグループ化された検索結果を, 図 4 に示す. この結果では,「鳥海不二夫」が代表者である研究コミュニティが最大であり, そのコミュニティに属している「榊剛史」,「風間一洋」,「松尾豊」らは, 別のコミュニティとしても発表していることがわかる. また,「鳥海不二夫」のグループとは別に,「斉藤和巳」らのグループもあることがわかる.

なお，松尾らによるWebから人工知能学会の人間関係ネットワークを抽出法 [12] や風間らによる人間関係ネットワークを用いた情報探索法 [2] では，人間関係のネットワーク構造を可視化して，そのままユーザに提示している．ただし，これは既存の論文検索システムが検索結果を表示するために使うリストベースのユーザインタフェースとは親和性が悪い．またネットワーク構造を可視化した場合は，その構造が複雑になるほどユーザが理解しにくくなる．

これに対して，本手法は計算コストが比較的小さい簡単なアルゴリズムでありながら，クエリに応じて動的に生成した共著者ネットワークをコミュニティ分割すると共に，その代表的な著者を選出することと同等の効果があり，既存のリストベースのユーザインタフェースとの親和性も良いことから，検索結果の理解が容易になると考えられる．

8 おわりに

本稿では，単語共起関係と共著関係を利用して関連語と関連著者を提示することで，論文アーカイブを探索できるシステムを提案した．さらに，共著者ネットワークのコミュニティ構造に基づいて単語を順位付けすることで専門的な関連語を選択する手法と，共著関係に基づいて論文をグループ化して，グループの代表的な著者とその共著者を提示する手法について述べた．

本研究の課題は以下の通りである．

まず，提案した論文探索システムの効率を数値的に定量化する手法を明確化することである．限られた数の関連語と関連著者を選択することで探索する情報空間を変更するとしたら，各単語によって指定される論文の重複はなるべく少なく，検索結果の広い範囲をカバーできた方がよいはずであり，その定量化指標を用いてTF-ICFのαを決定するのが望ましい．

次に，関連語の選択のためにTF-ICFを提案したが，関連著者の選択は論文数の上位から選択しているだけである．そこで，共著関係をうまく利用することで，上記の情報探索指標をさらに向上される手法が必要となる．

さらに，図2に示したような単語のユニット性に基づく複合名詞同士の関係を考慮することで，関連語提示における複合名詞の扱いを向上させることが考えられる．

また，共著関係に基づく論文のグループ化における代表的な著者の選択において，同じ論文数の著者が存在した場合に不自然と思われる結果が表示されることがある．そこで，共著者リストにおける順序も考慮することを検討する必要がある．

参考文献

[1] Paul Francis, Takashi Kambayashi, Shin-ya Sato, and Susumu Shimizu. Ingrid: A self-configuring information navigation infrastructure. In *Proceedings of the 4th International World Wide Web Conference*, pp. 519–538, Boston, Massachusetts, USA, December 1995.

[2] 風間一洋, 佐藤進也, 福田健介, 村上健一郎, 川上浩司, 片井修. Web空間における人間関係を用いた情報探索の一手法. 情報処理学会論文誌：データベース, Vol. 46, No. SIG 13 (TOD27), pp. 26–39, 2005.

[3] 高野明彦, 西岡真吾, 丹羽芳樹. 連想に基づく情報アクセス技術：汎用連想計算エンジンGETAを用いて. 情報の科学と技術, Vol. 54, No. 12, pp. 634–639, 2004.

[4] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp. 230–237, 2004.

[5] 伊藤直之, 西川侑吾, 田村直之, 中川修, 新堀英二. 品詞結合規則と外部辞書データを用いた複合名詞の生成. 情報科学技術フォーラム講演論文集, 第8巻, pp. 311–312, 2009.

[6] 浅原正幸, 松本裕治. Ipadic Version 2.7.0 ユーザーズマニュアル. 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座, 11 2003.

[7] 石橋和樹, 南出直樹, 風間一洋, 篠田孝祐. 単語のコミュニティ性に基づいた専門用語の抽出. 人工知能学会第28回全国大会 2J4-OS-16a-5in, 2014.

[8] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, Vol. 70, No. 6, 2004.

[9] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Vol. 28, No. 1, pp. 11–21, 1972.

[10] 風間一洋, 原田昌紀, 佐藤進也. サーチエンジンの検索結果のマルチレベルグルーピングの評価. コンピュータソフトウェア, Vol. 17, No. 4, pp. 58–69, 2000.

[11] 馬場康夫, 新里圭司, 黒橋禎夫. 検索エンジン基盤TSUBAKIを用いた大規模ウェブ情報クラスタリングシステムの構築. 情報処理学会研究会報告 2008-FI-89, pp. 67–74, 2008.

[12] 松尾豊, 友部博教, 橋田浩一, 中島秀之, 石塚満. Web上の情報からの人間関係ネットワークの抽出. 人工知能学会論文誌, Vol. 20, No. 1, pp. 46–56, 2005.