

単語のコミュニティ性に基づいた専門用語の抽出

Technical Term Extraction Method Using Community Feature of Terms

石橋 和樹 *1

Kazuki Ishibashi

南出 直樹 *1

Naoki Minamide

風間 一洋 *1

Kazuhiro Kazama

篠田 孝祐 *2

Kosuke Shinoda

*1和歌山大学

Wakayama University

*2電気通信大学

The University of Electro-Communications

This paper presents a new method to extract technical terms from bibliographic information such as a paper's title and abstract by a community feature of terms. We describe a measure of term technicality called ICF (Inverse Community Frequency), which is obtained from community structures in a co-authorship network, and a ranking method called TF-ICF. Furthermore, we evaluated FLR, MC-value, TF-IDF and TF-ICF using keywords and a co-authorship network extracted from bibliographic information of JSAI annual conferences. We show the effectiveness of TF-ICF and analyze its characteristics.

1. はじめに

研究開発の競争の激化により、新しく生み出される技術が年々増加するだけでなく、技術の陳腐化も早くなりつつある。例えば、学会における論文発表は、そのような世の中の技術動向を忠実に反映していることから、発表者だけでなく、論文の読者や、論文誌や会議を企画運営する学会自体からも、研究開発の状況や変化を容易に把握できる技術が望まれている。

例えば、CiNiiのような論文検索システムで探索的な検索を繰り返すことで網羅的に研究動向を調べることもできる。しかし、その場合は研究対象に対するクエリを適切に選択する必要があり、未知の技術や分野に対しては困難であることから、論文アーカイブから自動抽出した専門用語を、探索の手がかりとしてユーザに提示することは有用であると考えられる。

このような専門用語抽出のためには、例えば MC-value 法や FLR 法 [中川 03], TF-IDF [Jones 72] などの手法が用いられてきた。ただし、論文の題名と概要しか利用できない場合は、概要の長さの制約から専門用語の出現頻度が低くなると同時に、検索結果のように複数の論文に対して適用する場合は、相対的に一般的な論文用語の頻度が高くなり、これらの手法ではよい結果が得られなかった。

そこで本論文ではコミュニティ性という新しい専門用語の概念を導入し、学会の発表プログラムから得られる書誌情報の題名と概要から、ある研究分野を適切に表す専門用語を抽出する手法を提案する。

2. 単語のコミュニティ性

2.1 単語の専門性の判定

Kageura らは、専門用語の重要な性質としてターム性とユニット性を挙げている [Kageura 96]。ターム性とは、ある表現が対象分野固有の概念をどれだけ高い関連性を持って表現しているかという観念である。これは単語の出現頻度とその傾向に基づいて、統計的に計算される。ユニット性とは、ある言語単位（例えば、連語、複合語など）がコーパス中で安定して使用される度合いを表す。これは各言語単位の接続頻度に基づい

て、統計的に計算される。ただし、論文の題名と概要しか利用できない書誌情報アーカイブの場合には、概要でなるべく同じ専門用語を使わないようにしたり、逆に論文で多用されるような「提案」、「研究」などの一般的な単語の頻度が高くなることから、これらの手法では良い結果が得られなかった。

2.2 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、専門用語であれば、少数の関連する専門家のコミュニティで頻繁に使われるが、一般用語であれば多数のコミュニティで広く使われるなど、その単語の性質に応じて出現分布パターンに固有の特徴があると考えられる。そこで、このような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 と、それを用いた単語のスコアの計算法 TF-ICF を提案する。

2.3 ICF (Inverse Community Frequency)

ICF 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から求めた単語の専門性の指標である。

例えば TF-IDF は、単語-文書（論文）という 2 部グラフ構造（図 1a）から求められる。しかし、論文アーカイブの場合は、共著者情報を加えて単語-文書（論文）-著者という 3 部グラフ構造（図 1b）に拡張できる。さらに、論文は複数人の著者によって書かれることが多いことから、文書-著者の部分を変換した共著ネットワークは、頻繁に共同研究している著者達のコミュニティに分割できる。つまり、単語が決まれば、その単語を使用している著者のコミュニティの集合が決定される。

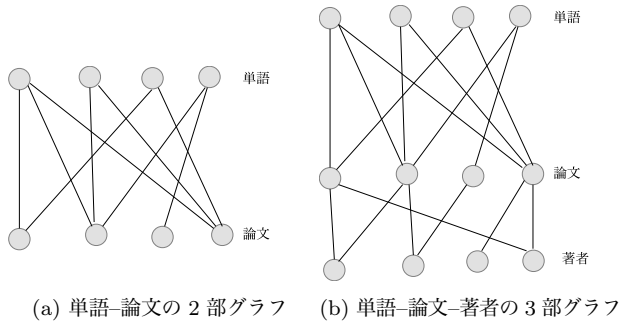
例えば、一般的に著者の論文生産性には大きな差があることから、一人の著者だけが用いる専門用語の頻度は、グラフ構造の論文のレベルでは大きく異なることになるが、著者のレベルでは同一となる。さらに、同じ専門分野の著者が共同研究するチームとして多数の論文を書く場合にも、論文のレベルでは差が出て、共著ネットワークでは、一つのコミュニティに集約されることになる。つまり、専門用語であれば少数のコミュニティに出現し、一般用語であれば多数のコミュニティに出現するので、ICF では、コミュニティ集合における単語の出現確率を定量化することで単語の専門性を判定する。

ICF は単語-文書（論文）-著者という 3 部グラフ構造を前提としていることから、既存のすべての分野に適用できるわけ

連絡先: 石橋 和樹 (s151002@center.wakayama-u.ac.jp)

和歌山大学システム工学部情報通信システム学科

〒 640-8510 和歌山県和歌山市栄谷 930

図 1: 単語と論文の n 部グラフ構造

ではないが、同様な3部グラフ構造を作成できる Twitter などのソーシャルメディアにも適用可能であると考えられる。

2.4 ICF の計算

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。 K は総単語数である。

1. データセットに含まれる全論文に対して、同一の論文を執筆した著者同士にエッジを張ることで、共著関係ネットワークを構築する。
2. 共著ネットワークを Clauset らの CNM 法 [Clauset 04] を用いて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取得する。
3. データセット中の単語 w_i に関して、単語 w_i が出現する論文の筆頭著者の集合を取得する。
4. 単語 w_i が出現する論文の筆頭著者の集合から、単語 w_i が出現するコミュニティ集合を求めて、そのコミュニティ数を $c(w_i)$ とする。
5. 単語 w_i のコミュニティ集合における出現率 $r(w_i)$ ($0 \leq r(w_i) \leq 1$) を求める。

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

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。

6. 単語 w_i の ICF の値 $ICF(w_i)$ を計算する。

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

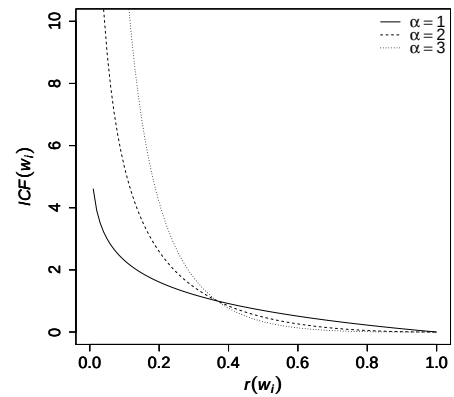
ここで、 α は定数である。

すなわち、図 1b 上で考えると、単語レベルから論文レベル、論文レベルから著者レベルに写像した上で、共著関係ネットワークにおける該当コミュニティを求めていることになり、この過程において個人やグループのアクティビティの差が除去され、著者のレベルからみた専門性をより忠実に反映させることが可能となる。

ここで $r(w_i)$ の逆数の対数を α 乗する理由は、 $r(w_i)$ の値の大小に対する $ICF(w_i)$ の効果を調節するためである。 $r(w_i)$ 、 $ICF(w_i)$ と α の関係を図 2 に示す。これから、 α の値を増加すると、 $ICF(w_i)$ の変化が大きくなることがわかる。

2.5 TF-ICF

TF-IDF は情報探索やテキストマイニングなどの分野で利用され、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 [Jones 72]。単語 w_i の TF-IDF の値

図 2: $r(w_i)$, $ICF(w_i)$ と α の関係

$TF-IDF(w_i)$ は、ある文書に単語が出現する度合いを表す TF (Term Frequency) の値 $TF(w_i)$ と、単語が文書全体に出現する度合いの逆数である IDF (Inverse Document Frequency) の値 $IDF(w_i)$ の積で与えられる。

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

$$IDF(w_i) = \log \frac{D}{d(w_i)} \quad (4)$$

$$TF-IDF(w_i) = TF(w_i) \times IDF(w_i) \quad (5)$$

ここで、 $n(w_i)$ は単語 w_i の出現回数、 D は総ドキュメント数、 $d(w_i)$ は単語 w_i を含むドキュメント数である。

本論文では、IDF の代わりに ICF を用いる TF-ICF を提案する。単語 w_i の TF-ICF 値である $TF-ICF(w_i)$ は次のように定義する。

$$TF-ICF(w_i) = TF(w_i) \times ICF(w_i) \quad (6)$$

なお、TF-ICF では、広く使われる一般語や特定のコミュニティだけが使う固有名詞を低く評価することを目的としているが、この際の TF と ICF のバランスは定数 α で調整できる。

3. 評価

3.1 JSAI データセット

人工知能学会は、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 で公開している。この発表プログラムには、すべての発表の時刻、演題番号、題目、著者に加えて概要も掲載されていることから、2003 年から 2013 年までの 11 年間の HTML 形式の発表プログラムを収集し、書誌情報とキーワードを抽出してデータセットとして用いた。本論文では、これを JSAI データセットと呼ぶ。

以下にデータセット作成方法と、その詳細について述べる。

3.1.1 講演プログラムの収集

まず、wget コマンドを用いて 2003~2013 年の 11 年間の発表プログラムを収集した。ただし、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除外した。年によって格納されているディレクトリ名が若干異なるが、修正して利用した。

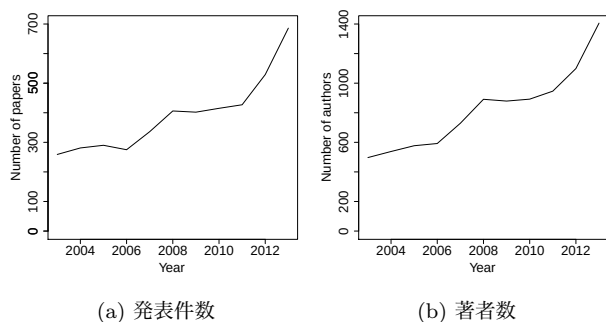


図 3: 人工知能学会の発表状況の推移

3.1.2 書誌情報の抽出

収集した発表プログラムから、Python で記述したプログラムを用いて、次の手順で時刻、演題番号、題目、著者と概要の書誌情報を抽出した。

1. 正規表現を用いて、演題番号、題目、著者、時間、概要を抽出する。
2. 著者名、組織名、題目と概要の表記、大文字・小文字、全角・半角などの違いを正規化する。

なお、HTML ファイルの文字コードや使用されているタグの使い方などの違いは、プログラムで対処した。

3.1.3 キーワードの抽出

題名と概要から、さらに以下の手順でキーワードを抽出した。

1. 題目と概要のテキストを MeCab[Kudo 04] を用いて日本語形態素解析する。
2. 抽出された形態素の品詞情報に基づいて、伊藤らの手法[伊藤 09]を参考に、名詞や接頭詞、接尾辞を連結して、複合名詞を抽出する。
3. 品詞が誤判定される記号類は、ストップワードリストを用いて除去する。

3.2 統計情報の分析

2003 年から 2013 年までの 11 年間で 4313 件の発表、6828 語のキーワードが抽出された。発表件数の推移を図 3a に示す。横軸が開催年、縦軸が発表件数である。さらに著者数の推移を図 3b に示す。横軸が開催年、縦軸が著者数である。ただし、著者数には発表者だけではなく、その共著者も含まれる。この結果から、人工知能学会全国大会は着実に発表件数と発表者が増加している活気のある会議であることがわかる。

3.3 共著ネットワークの分析

データセットから抽出された総著者数は 4951 人、孤立ノードを除く連結成分数は 293 個であった。2003 年から 2013 年の共著ネットワークを、Cytoscape[Shannon 03] の Force Directed Layout を用いて可視化した結果の最大連結成分を図 4 に示す。ノードは著者で、共著した論文が存在する場合にエッジが張られている。さらに、クラスタリング係数は 0.744、平均直径は 14、平均次数は 5.294 であった。単独で執筆している著者は孤立ノードとなり可視化結果には表示されていないが、ICF の計算時には 1 人だけのコミュニティとして扱った。その結果、537 個のコミュニティに分割された。

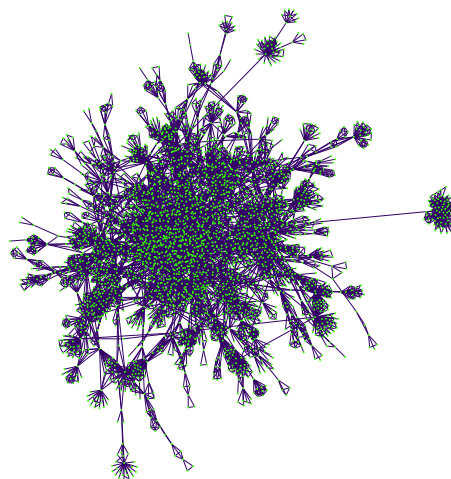


図 4: 共著ネットワーク

表 1: 手法毎のランキング上位の比較 (1)

順位	MC-value 法	FLR 法	TF-IDF
1	提案	情報	オノマトペ
2	手法	システム	ユーザ
3	ユーザ	提案	提案
4	利用	データ	情報
5	情報	分析	利用
6	分析	研究	手法
7	システム	ユーザ	分析
8	構築	手法	研究
9	研究	学習	ロボット
10	検討	表現	システム

3.4 ランキング結果の評価

JSAI データセットを用いて、各年に使用されたすべてのキーワードを以下の 4 種類の指標に基づいてランキングするプログラムを Python で作成し、その実行結果を評価した。

1. MC-value 法
2. FLR 法
3. TF-IDF
4. TF-ICF

α は 1.0, 2.0, 3.0 で計算し、CNM 法には SNAP(Stanford Network Analysis Platform) *1 の community プログラムを用いた。

3.4.1 ランキング上位の比較

2003 年から 2013 年までの 11 年分のデータのうち、2012 年の MC-value 法、FLR 法、TF-IDF のランキング結果の上位 10 件を表 1 に、TF-ICF の α の値を変更した場合のランキング結果の上位 10 件を表 2 に示す。MC-value 法や FLR 法では、「提案」、「分析」、「研究」のような論文特有の単語が上位を占めており、TF-IDF では「オノマトペ」、「ロボット」などの特定の専門分野を表す単語がいくつか現れている。これに対して、TF-ICF では論文特有の単語は少なく、「オントロジ」、「物語生成システム」などの具体的な研究内容を表す単語が上位に現れていると言える。

3.4.2 ランキング精度の評価

ランキング性能を、上位 20 件の精度である $P@20$ を用いて評価する。ただし、「ソーシャルメディア」や「オノマトペ」のように特定の分野で使われる専門的な用語を正解とし、「提案」や「研究」のような論文に多用されるが、特定の分野を示

*1 <http://snap.stanford.edu/snap/>

表 2: TF-ICF のランキング上位の比較 (2)

順位	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$
1	ユーザ	オノマトベ	オノマトベ
2	提案	ロボット	フォント
3	オノマトベ	ユーザ	オペレーション
4	情報	オントロジ	物語生成システム
5	手法	フォント	照度センサ
6	ロボット	分析	デザイン知
7	分析	オペレーション	照度
8	利用	物語生成システム	知的照明システム
9	システム	デザイン	ブランドネーム
10	構築	仕掛け	作家

表 3: 抽出精度 ($P@20$) の比較

手法	2011 年	2012 年	2013 年	平均
MC-value 法	0.20	0.25	0.15	0.20
FLR 法	0.30	0.35	0.35	0.33
TF-IDF	0.25	0.30	0.25	0.27
TF-ICF($\alpha = 1.0$)	0.25	0.30	0.30	0.28
TF-ICF($\alpha = 2.0$)	0.65	0.95	0.65	0.75
TF-ICF($\alpha = 3.0$)	0.90	1.00	0.95	0.95

さない単語は不正解とした. 2011 年から 2013 年の各手法の $P@20$ を表 3 に示す.

この結果から、TF-ICF の精度が既存手法よりも高くなる傾向があり、 $\alpha = 3.0$ の場合に一番良いことがわかった。ただし、単純に α の値を増やせば性能が向上するわけではない。例えば、 $\alpha = 6.0$ の場合には $P@20$ の性能という点では大きく変化しないが、 $\alpha = 3.0$ の時にランキング上位に出現していた「クラウドソーシング」等のその年に注目された分野の専門用語が減少し、代わりに「交代取引ゲーム」などの特定の論文にしか出現しない単語が増加する傾向が見られた。これは、複数の論文をまとめて指定することができなくなるので、論文探索効率が低下し、ユーザの利便性を損なうと考えられる。

4. JSAI 全国大会論文探索システム

人工知能学会全国大会の発表プログラムの書誌情報から抽出された主要キーワードと主要著者を手がかりに、単語—文書（論文）—著者という3部グラフ構造を辿って論文を探索することができるシステムをPythonとMongoDBを用いて試作した。このシステムでは、著者名またはキーワードを入力すると、検索結果に含まれる複数の論文を著者グループごとに表示するのに加えて、論文の探索に有効な主要著者と主要キーワードも提示する。この主要キーワードのランキングにTF-ICFを用いた。

このシステムで実際に「ロボット」で検索した時の実行例を、図5に示す。TF-ICF の場合の上位 20 件の主要キーワードは「ロボット、移動ロボット、知能ロボット、インタラクション、自律ロボット、コミュニケーションロボット、対話ロボット、物体概念、操作者、発話、教示、ヒューマノイドロボット、物体、SIGVerse、商業施設、連携作業、知識状態、遠隔操作ロボットシステム、セマンティックロボットサービス、子ども達」、TF-IDF の場合は「ロボット、人間、実現、人、インタラクション、学習、実験、行動、ユーザ、動作、提案、獲得、発話、物体、対話、研究、手法、コミュニケーション、構築、移動ロボット」となる。すなわち、提案手法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。

5. おわりに

本論文では，専門用語コーパス，特に論文データからの専門用語の抽出法について検討した．コミュニティ性を利用した指



図 5: 論文探索システムの実行例

標である ICF と、それを用いたスコアリング手法である TF-ICF を提案すると共に、既存の手法である TF-IDF や FLR 法、MC-value 法と比較して評価した。その結果、今回用いたデータセットの場合には、TF-ICF は他の手法と比較して優れていることがわかった。

今後の課題は、まず論文探索に適切な専門用語を選ぶための α の最適値の評価方法を確立することである。次に、共著ネットワークの作成に Jaccard 係数や Simpson 係数などの関連性の強さを用いた場合の性能の変化を定量的に調べる必要がある。さらに、専門用語抽出やランキングにおいて単語のユニット性を考慮していないことから、「ロボット」、「移動ロボット」、「知能ロボット」同じ部分文字列を含む複合語が同時に提示されてしまうという問題があることから、複合語の扱いを改善する必要がある。

参考文献

- [Clauset 04] Clauset, A., Newman, M. E. J., and Moore, C.: Finding Community Structure in Very Large Networks, *Physical Review E*, Vol. 70, No. 6 (2004)
- [伊藤 09] 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)
- [Jones 72] Jones, K. S.: A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of Documentation*, Vol. 28, No. 1, pp. 11–21 (1972)
- [Kageura 96] Kageura, K. and Umino, B.: Methods of Automatic term recognition: A review, *Terminology*, Vol. 3, No. 2, pp. 259–289 (1996)
- [Kudo 04] Kudo, T., Yamamoto, K., and Matsumoto, Y.: 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)
- [中川 03] 中川 裕志, 湯本 紘彰, 森 辰則: 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, Vol. 10, No. 1, pp. 27–45 (2003)
- [Shannon 03] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T.: Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks, *Genome Research*, Vol. 13, pp. 2498–2504 (2003)