専門用語の専門度の指標に関する一考察

内山清子 †　鈴木崇史 †　相澤彰子 †

国立情報学研究所 †

E-mail:{kiyoko,t_suzuki,aizawa}@nii.ac.jp

本研究は、専門用語の専門度を示す客観的な指標を作成するために必要な基礎的情報として著者キーワードが出現する年度数、研究領域、文書数について分析を行った。著者キーワードは論文を特徴づける語であることから、専門用語の候補として出現頻度の傾向を分析することにより、専門性の度合いを推測する上で重要な手掛かりとなる。分析の結果、出現傾向は 4 つのグループに分類でき、専門度の指標を作成する上で有効であることを確認した。

1　はじめに

専門用語は、新聞に頻出するような一般用語と比較すると実際に目にする機会が少ないために、意味を理解することが難しい。さらに、特定分野における専門用語が初心者レベルの常識・基礎的な用語か、上級者レベルの専門・応用的内容であるかの段階が曖昧である。そのため、新しい研究分野の概要を把握できる論文や、抑えておくべき論文を検索する時には、他の人が勧めている論文やよく引用されている論文を探すことが多い。しかし、主観に基づく推薦では人によって観点が異なることや、引用が多い論文は専門的な論文が含まれていることがある。また、専門文書では専門用語が多く含まれているために、用語を理解せずに読み進めるのが困難である。

この問題を解決するために、専門用語の難しさの度合い（専門度）を客観的に判定し、読者に適した論文の提示が必要となる。分野の初心者に対しては特定分野の少ない背景知識で理解できる分野必須の用語（低専門度用語）を説明する、あるいは多く含む論文を提示し、上級者に対しては多くの背景知識を必要とする用語（高専門度用語）を多く含む論文を提示するレコメンデーションシステムへの応用のために、専門度の指標が重要となってくる。但し、専門度が高い用語を理解できる上級者であれば、論文の選択の際にあまり支障が出てこないと考えられる。

そこで本研究では、適切な論文提示を求めているのは初級者に多いという前提から、初級レベルの専門度の判定（一般用語との区別を含めて）に関連すると考えられる専門用語の出現頻度の分析を行う。具体的には、著者自身が著書論文に対して付与するキーワード（著者キーワード）について、その出現傾向（年度数、研究領域、文書数）の分析に基づき専門用語の指標との関連性について考察を行う。

2　関連研究

2.1　文書の難易度

専門用語の専門度の指標が重要な一つの理由として、専門用語の難しさは文書の読みやすさ（難易度）にも関連するからである。従来の文書の難易度を判定する研究 [2, 5] では、一般の文書（教科書など）を対象に文字種の頻度や一文の長さなどを基準にしている。しかし一般の文書と違い、専門文書（論文や技術文書等）では、専門用語の専門性が文書を読み進めるために重要なポイントとなる。

2.2　重要語抽出

専門度に関する研究は、先行研究で行われてきた重要語抽出 [4] の「重要度」に関連している。従来の重要度とは、対象分野の研究を行っている研究者あるいは書き手（論文の著者）にとっての重要さであるので、重要語抽出では分野や文書に特徴的な語や頻度が高い語、多くの背景知識を必要とする用語（高専門度用語）が抽出される。この手法を利用することで重要語＝高専門度用語の抽出が可能になる。本研究では、これに加えて比較的長く継続的に特定分野に共通して低頻度で出現する用語の抽出が必要となるため、時系列の情報が必要となってくる。

2.3　研究動向分析

時系列の情報については、タイトルに含まれる手がかり語（を利用した、を用いた等）から分野

と手法を抽出し、出現分野や年度に着目して、研究における手法がどの分野でいつごろ論文に出現しているかの傾向について可視化した研究[3]がある。この情報に加えて、本研究では出現頻度が継続的／瞬間的であるのか、分野を情報処理学会の研究会領域に絞って分野全体の傾向を概観するのが新たな視点である。

2.4 専門度推定

技術者がプレゼンテーションの時に専門外の人に対して、専門用語を使用せずに平易な用語に置き換えて、効果的に技術成果を伝えるために用語の専門度推定を行った研究もある[6, 7]。この研究では「専門用語ではなく意味の近い平易な用語を用いる」ことを前提としているため、専門用語内での専門度については専門外の人から見て比較的専門的な用語であるか、かなり専門的な用語かの2段階に分けている。本研究では、専門用語でも更に細分化された特定分野における専門度についての推定を行う点が特徴となる。

3 著者キーワードの出現傾向の分析

専門度の指標に必要な情報を分析するために、専門用語の出現傾向について調査を行った。対象とする専門用語の選定については、従来から重要語抽出手法が用いられているが、今回は著者キーワードを利用することにした。理由として、著者キーワードはその論文に特徴的な用語であり、他の研究との差異を強調するために付与するものであるため、一般的な用語が含まれにくく、専門用語の候補として適切であると判断した。将来的には重要語抽出手法も合わせて比較することで最適な選定方法を検討する。

3.1 著者キーワードの抽出

まず、文書から著者キーワードを抽出し集計を行った。データは、情報処理学会刊行誌掲載論文本文データの中から研究報告（35の研究会、1993年から2005年までの13年間、22720文書）のテキストを対象として専門用語の出現傾向について調査を行った。著者キーワードのうち、括弧の挿入や文で表現しているものについては、助詞、動詞、記号を削除するなどのフィルタリングを行った。著者キーワードは論文に特徴的な用語を設定するため、一般的な用語が含まれないと予想していたが、「大学」「自動化」などが多く含まれていた。これは計算機センタや学校教育という文脈で「大学」での開発や、何かを「自動化」することに意義があるとしてキーワードに設定していた。こうした一般用語についてのフィルタリングは今回は行わなかった。著者キーワードは15447の異なり語があり、その著者キーワードが何回利用されているか（使用頻度毎に該当する著者キーワード数）を数え、その両対数をとって図1に示した。著者キーワードとして使われた頻度が一回しかない単語が約10000個以上で全体の65 %であった。著者キーワード自体は半数以上が重ならない一方で最も多く共通して使われる著者キーワードは「計算機網」で1288回使用されていた。つまり共通で使われるものと著者独自に設定するキーワード群が二極化していた。

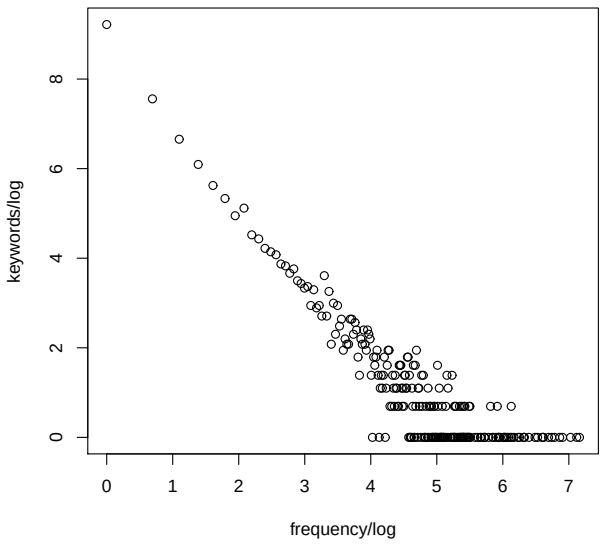

図1 著者キーワード使用頻度

3.2 著者キーワードの出現頻度

次に15447個の著者キーワードについて、そのキーワードが本文中で使われている出現数と文書数を計算した。その際部分文字列の重複カウントを避けるため、形態素解析MeCab[1]の辞書にあらかじめ著者キーワードを登録した上で、論文の出版年度と情報処理学会研究会の3つの領域であるコンピュータサイエンス(S)、情報環境領域(E)、フロンティア領域(F）毎に出現頻度を数えた。出現年度の集計では、継続的あるいは瞬間的に出現

[1] http://mecab.sourceforge.net/

するのかどうか、また研究領域の違いによる出現パターンについて分析を行った。この分析は、専門度の指標を考える際に、論文中に出現する専門用語の出現パターンの傾向を概観し、その出現パターンと専門度との相関関係を調べるための基礎データとなる。代表的なキーワードについて、出現する領域毎に出現年度（何年出現しているか）と年間平均出現文書数（出現文書数/出現年度）を研究会領域のラベル (S、F、E) を付与し図 2 に示す。

3.3　著者キーワードの出現傾向

著者キーワードの出現傾向をまとめると大きく分けて以下の 4 通りになる。

(1) 継続的に全研究領域に出現（高頻度〜低頻度）

(2) 特定の研究領域に高頻度で出現

(3) 瞬間的に高頻度で出現

(4) 断続的に複数領域に低頻度で出現

(1) には、情報処理の基礎的な用語に該当するものとして「シミュレーション」、「インターネット」、「アルゴリズム」があり、9 年から 12 年間、3 つの領域全てに継続的に出現している。これはある時期に集中的に出現する用語（例：インターネットは 2000〜2002 年度頃がピーク）も含まれるが、ピークが収まってもコンスタントに出現するものは流行りではないと判断できる。(2) は、1 つの領域に 10 年以上数多く出現している分野特有の専門用語で「並列処理」、「機械翻訳」、「共有メモリ」などが含まれる。(3) は特定分野中で、専門性の高い語が 1 つの領域に特化して瞬間的（1 年から 3 年間）に数多く出現しているもので、「マルチメディア通信」、「無線」、「モバイル」などが該当する。これは他の年度には出現しないことから一種の流行りの用語であると判断できる。但し、新しい用語の場合は今の流行りなのか、今後も定着していく用語であるのかの判断が難しい。(4) は、複数の分野で長い間断続的に出現するが、比較すると出現する頻度が少ない「クラスタリング」、「遺伝的アルゴリズム」、「ウェーブレット変換」などである。分野共通あるいは分野に特化して用いられる手法に関する情報は論文を読む上で重要となるため、手法については、頻度以外の情報（以下で説明する文脈情報など）を用いる必要がある。

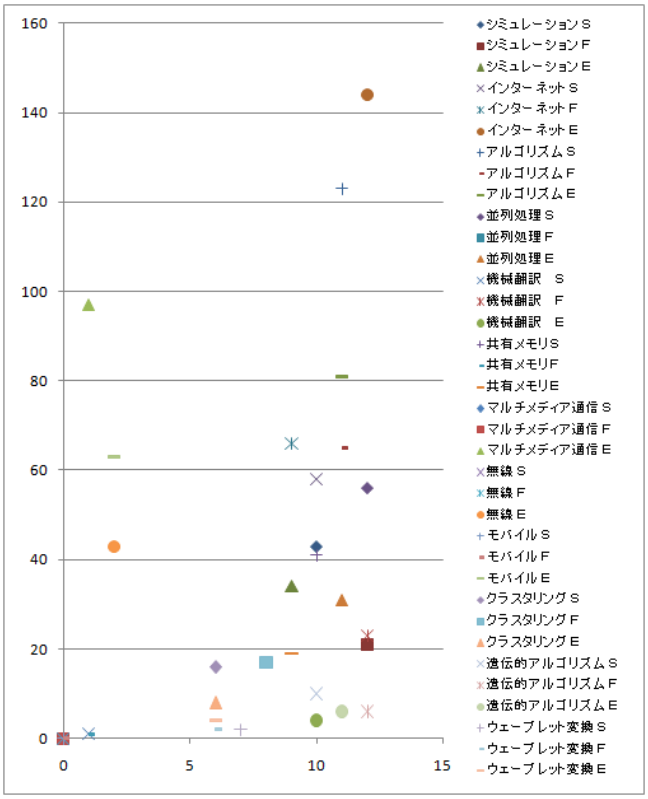

図 2　著者キーワードの出現傾向

このような結果に基づいて、(1) から (4) にかけて専門度が高くなると予測した場合、キーワードの出現傾向（出現頻度）は専門度の指標に有効な情報となる。重要な点として (2) の特定分野における基礎的な用語は、低専門度用語に判定できるが、これらの用語と共起する用語が各論文や分野の特徴語であると考えられる。たとえば、「機械翻訳」は分野における基礎的な用語であるが、共起する用語として「翻訳規則」「対訳コーパス」を抽出すれば、より専門性が高い用語の候補になると考えられる。一方で、著者による個人差として著者キーワードが論文の特徴を全く反映していないものを付与している場合もある。この場合はタイトルに関してはどの論文も特徴を反映しているため、タイトルに含まれるキーワードを利用することも検討課題である。今後は更に上記の出現頻度パターンを統計的手法により、客観的に専門度の指標として利用していく。

4　専門度の指標に用いる情報

著者キーワードの出現傾向の分析による情報に加えて、今後取り入れたい情報について以下で述べていく。

4.1　引用情報

引用情報とは、対象とする専門用語の引用文脈を利用した情報を指す。ここでいう引用文脈とは、同一文内に参照番号 [n](n は任意の数字) が付与されているものとする。たとえば、対象とする専門用語が「機械学習」であれば、「機械学習のうち教師あり学習の自己組織化マップを用いた研究 [n]」という文を引用文脈としてその中に含まれるキーワード「教師あり学習」「自己組織化マップ」を抽出する。論文中における引用文脈は、論文テーマに対する現状、問題点、改善点などをまとめた重要な用語が含まれているため、引用文脈に含まれる用語の専門度、重要度を他と差別化するような重みづけをすることが必要となる。

4.2　文脈情報

文脈情報では、専門用語が文中で表現される語のパターンや共起語に着目する [1)]。たとえば、対象とする専門用語 A が「A などの B」の場合、A は B の一種であり、B は基本的な用語で、低専門度語であると判定できる。また、対象とする専門用語が同一文内に二つ以上出現する場合、その文章を抽出すると、専門用語間の意味関係を抽出するための基礎データとなる。このデータから関係を示す表現パターン（手がかり語）として、「A 手法を用いた B」、「A の基盤技術である B」などを抽出し、専門用語間の関係を特定する。特に著者キーワードの出現傾向で概観した手法に関する用語は専門用語として重要であると考えられるので、手法を抽出するための表現パターンをうまく設定する必要がある。

4.3　語彙情報

語彙情報とは、専門用語を構成している形態素（以下、語構成要素と呼ぶ）の数や語構成要素が単独で文中で用いられる特徴に関する情報を指す。語構成要素の数が多くなるほどより詳細な専門用語を表している可能性が高くなり、専門度が増すと予測される。また語構成要素間の関係が、専門用語の意味を理解するための手掛かりとなることから、語構成要素の文中での役割（主語、目的語、修飾語、述語など）については後続する助詞の種類や出現傾向を調べた上で、専門用語内での役割、たとえば、「決定木学習」では「決定木」は「学習」の一手法であるといった役割を推定していく。語構成要素間の関係は専門度だけでなく、用語の上下下位関係を判定する上で重要な役割を果たす。

5　おわりに

本研究では、著者キーワードを対象として出現傾向を分析し、専門度の指標として利用できる情報と今後取り入れたい情報について整理を行った。出現年度、出現する研究領域における頻度は、専門度の指標、特に初級者レベルの低専門度用語の判定に有効であることがわかった。今後は全ての情報を統合して専門度推定を行い、専門度を客観的に評価していくことが課題である。

謝辞

情報処理学会刊行誌掲載論文本文データに関して、研究利用することを許諾していただいた、社団法人情報処理学会に感謝いたします。

参考文献

1) 相澤彰子, 大規模テキストコーパスを用いた語の類似度計算に関する考察, 情報処理学会論文誌, Vol.49. No.2, PP.1426-1436, 2008.

2) 小島健輔, 佐藤理史, 藤田篤, 文字 bigram モデルを用いた日本語テキストの難易度推定, 言語処理学会第 15 回年次大会論文集, pp.897-900, 2009.

3) 近藤友樹，難波英嗣，奥村学，新森昭宏，谷川英和，鈴木泰山, 論文データベースからの研究動向情報の抽出, 言語処理学会 第 13 回年次大会，pp. 470-473, 2007.

4) 中川裕志, 森辰則, 湯本紘彰, 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.

5) Satoshi Sato, Suguru Matsuyoshi and Yohsuke Kondoh, Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus, LREC-08, 2008.

6) 千田恭子、篠原靖志、奥村学, アンケートによる用語調査とWWW上の頻度分布を用いた用語の専門度推定, 言語処理学会第 10 回年次大会ワークショップ「固有表現と専門用語」発表論文集, pp.36-39、2004.

7) 千田恭子、篠原靖志、奥村学, 技術成果を効果的に伝える表題作成支援手法:開発と評価, 情報処理学会論文誌、Vol.46, No.11, pp.2728-2743, 2005.