社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (13)
2008－ＮＬ－183 (13)
2008／1／22

用語クラスタリングに基づく部分研究領域推定と用語分類

小山 照夫* 竹内 孔一**

*国立情報学研究所 **岡山大学大学院自然科学研究科

研究抄録テキストコーパスから抽出された用語候補を体系的に整理する一つの有力な方法として、部分研究領域に関連付けた用語の分類を考えることができる。本研究では、動詞概念との共起に基づく用語クラスタリングによって、特定研究分野のいくつかの部分研究領域が同定できることを示すとともに、同定された部分領域との関連により、テキストコーパスから抽出された用語候補の分類が可能であることを示す。

Identification of Research Sub-Domain and Term Classification Based on Term Clustering

Teruo KOYAMA* Koichi TAKEUCHI**

*National Institute of Informatics
***Graduate School of Natural Science and Technology, Okayama University

Term classification associate to research sub-domain is an important approach for systematized classification of term candidates extracted from text corpora. The authors have developed a method which identify some of the important research sub-domains in research abstract corpora. The authors also proved that relatively frequent term candidates extracted from the corpus can be related to identified sub-domains.

１．はじめに

特定研究分野のテキストコーパスから抽出された用語を活用する上で、用語を様々な視点から分類する、用語の体系化は重要な課題である。用語を体系化する上での有力な方式の一つとして、コーパス全体が取り扱う研究分野にどのような部分領域が存在しているかを明らかにし、部分領域に関連付けて用語を分類することが考えられる。研究分野は一般に複数の部分領域を含んでおり、用語をこれらと関連付けることにより、様々な文書利用に役立てることができると期待できる。

筆者らは、先行研究[1]の中で、工学分野という比較的広い範囲の研究抄録からなるコーパスについて、動詞概念との共起に基づく用語クラスタリングによって、部分研究領域の同定が可能であることを示してきた。

本研究では、同様の手法を適用することにより、情報処理という相対的に狭い範囲の研究抄録からなる文書コーパスについても、当該分野の主要な部分領域のいくつかを同定できることを示すとともに、コーパスから抽出された複合語用語候補のうちで、比較的頻度の高いものを、同定された部分領域と関連付けて分類することが可能であることを示す。

特定領域コーパスにどのような部分領域が存在しているかを判定する方法として、各文書に出現する語の生起パタンに基づいて文書をクラスタリングする方法などが検討されている[2,3]。

部分領域推定の手法として、文書クラスタリングは有力な方法ではあるが、一面で、結果とし

て得られた集合がどのような部分領域に関連しているかを判定するのはかならずしも容易ではない。文書クラスタリングで求められた文書集合が関連する部分領域を推定するためには、例えば得られたクラスタに特徴的に出現する用語を求め、得られた用語との関連で部分領域の決定を行うなどが必要となる。これに対して用語クラスタリングを用いる場合、関連する部分領域はより直接的に判定が可能である。また、用語クラスタリングでは、代表的な部分領域を求めるために必要とされる用語の数は比較的少数で充分である。このことは、相対的に出現頻度の大きい用語を用いた推定が可能であることを示しており、統計的に安定したクラスタリング手法の適用が可能となることを意味する。以下では2節において用語クラスタリングの手法とその情報処理分野コーパスに対する適用結果を示す。3節ではあらかじめ抽出された複合語用語候補のうち、比較的頻度の高いものについて、得られた用語クラスタと関連づけて分類する手法と分類結果について述べる。最後に4節では結果に関する考察を述べる。

2．動詞概念との共起に基づく用語クラスタリング

自然言語では、動詞ないしは動詞的概念を中心に記述が行われる。このことから動詞概念は自然言語記述の内容を規定する重要な役割を担っていると考えられる。さらに、名詞概念を表す用語と一群の動詞概念との共起を調べることにより、特定の論述内容に関連する名詞的要素を同定できると期待できる。

日本語の学術論文では、主要な動詞概念は「サ変名詞」として現れる。このことから、分野において重要であると期待できる名詞概念を、サ変名詞との共起関係から分類することにより、名詞概念の分類が可能となると考えられる。

今回の研究では、NTCIR-1に収録された学会発表データベースから選出された、情報処理学会の抄録約28,000を用いた。

コーパスに対してあらかじめJumanを用いて形態素解析を行った後、各形態素に対してTf-IDf値を求めておく。各形態素について、動詞概念を表すものとしてサ変名詞、名詞概念を表すものとして普通名詞および未知語を選び、それぞれをTf-IDf値によってソートする。

それぞれのグループに属する上位のもののうち、名詞系形態素については、明らかに用語性が低いと考えられるもの（例えば「方式」など）、また、動詞系形態素については、あまりに一般的過ぎると考えられるもの（例えば「利用」など）を除いた上で、名詞系形態素40および動詞系形態素15を選択し、共起傾向を調べることとした。

今、動詞系形態素 v に対する名詞系形態素 t の共起強度を

$$O_{tv}=(t_v/d_v)/(t_h/d_h)$$

とする。ここで、

t_v ： v と共起する t の出現頻度
t_h ：全文書に対する t の出現頻度
d_v ： v を含む文書数
d_h ：全文書数

である。t_x/d_x は文書集合 x についての、t の文書あたり出現期待値と考えられるから、この値は名詞性形態素 t の出現に関する尤度比を表していると考えることができる。

今、動詞系形態素15を考えるとき、名詞性形態素のそれぞれについて、15の共起強度が得られることになる。これらをまとめてベクトル $\vec{O}_t=\{O_{tv}\}$ と考えると、各名詞性形態素は15次元空間内に分布することとなる。このベクトル空間において、クラスタ分析を適用する。クラスタ分析にあたって、距離尺度としてはユークリッド距離、クラスタリング基準としてはコンパクト法を用いている。図1は解析結果を示したものである。ここで、視認に基づいて、全体を6つのクラスタに分かれるように閾値を決定すると、それぞれのクラスタに属する名詞系形態素は次の通りとなる。

クラスタ1：
文字、音声

クラスタ2：
事例、データベース、文書

クラスタ3：
構造、モデル、階層、空間、データ、効率、グラフ、アルゴリズム、関数、概念、知識、言語、論理

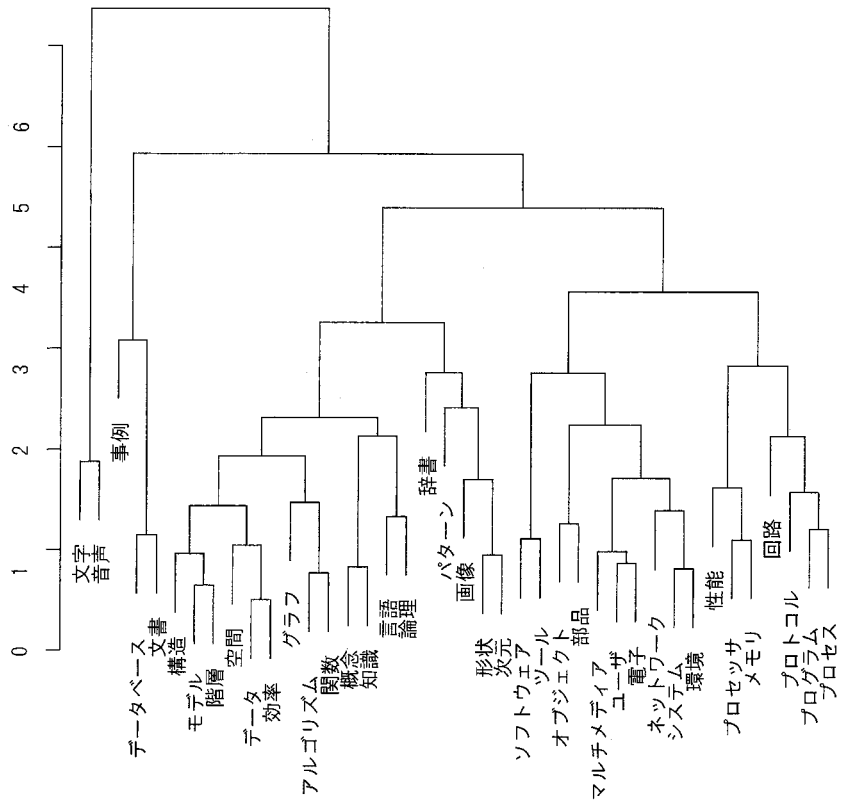

図１．クラスタ解析結果

クラスタ４：辞書、パターン、画像、形状、次元

クラスタ５：
ソフトウェア、ツール、オブジェクト、部品、ユーザ、電子、ネットワーク、システム、環境

クラスタ６：
性能、プロセッサ、メモリ、回路、プロトコル、プログラム、プロセス

これらの用語から推定される部分領域はそれぞれ、

1. 音声・パターン処理
2. データベース
3. 知識処理
4. メディア・パターン処理
5. ソフトウェア開発
6. ソフトウェア効率

に相当すると考えることができるであろう。これらはいずれも情報処理分野の部分領域として妥当なものである.

３．求められたクラスタに応じた用語候補分類

以上の結果を用いて、別途抽出された複合語用語候補の分類を試みる。用語候補は文献[4]で提案した方法により、あらかじめ抽出を行っている。実際に抽出を行った結果、今回は

46,021の候補が抽出されているが、統計的手法の信頼性を考慮して、コーパス内出現頻度40以上のもの882候補を対象とした。

分類の基本的考え方は、まず第一にすべての文書について、各クラスタに関連する用語の出現のしやすさを推定し、次いで、クラスタ関連語が出現しやすいと判断された文書に、相対的に高い頻度で出現する用語候補を、クラスタ関連用語と判定する。

まず、各クラスタについて、分類された形態素ベクトルの平均値を求める。これは、各クラスタを代表する、仮想的な用語を求めることに相当すると考えられるであろう。次にこの平均ベクトルを用いて、各文書に対する仮想用語の出現傾向を求める。

用語ベクトルの各要素が尤度比であることから、動詞性形態素が用語出現のしやすさに対して独立した効果を持つと仮定するなら、特定文書に対して用語の出現しやすさは、文書に出現する動詞要素に該当する尤度比の積になると考えることができる。

ここで尤度比が1より小さい要素について考えるならば、これは動詞が文書中に出現することにより、用語出現の尤度が低下することを意味する。しかし、実際には動詞の出現が用語の出現の確からしさを減少させる直接の因果関係があるとは考えにくい。むしろこの様な結果は偶然によるものと考える方が自然であろう。

これに対して尤度比が1より大きい場合、当該動詞は用語を用いた記述に関係していると考えられるところから、直接の因果関係によって尤度比を向上させる効果を持つと考えられる。以上より、あるクラスタ c を代表する用語に対するベクトルを

$$\vec{O}=\{O_{cv}\}$$

とする時、文書 d に当該クラスタに属する用語が出現する可能性に対する尤度比を

$$R_{cd}=\prod_{v=1}^{n} M_{vcd}$$

ただし

$$M_{vcd}=O_{cv} \quad : \quad d \text{ が } v \text{ を含み、かつ } O_{cv}>1 \text{ の時}$$
$$1 \quad : \text{それ以外}$$

によって計算する。この方式では、尤度比が1より小さい動詞概念の影響は無視することになる。

次にこの値を用いて各用語候補がそれぞれのクラスタに属する傾向を計算する。ある用語が特定のクラスタに属する傾向は、当該クラスタに属する用語が出現しやすい文書に現れる割合が大きいほど強いと考えることができる。そこで用語 t が文書 d に f_{td} 回出現する場合に、この用語がクラスタ c に帰属する度合いを、

$$S_{tc}=\frac{\sum_{k=1}^{n} R_{ck}\times f_{tk}}{\sum_{k=1}^{n} f_{tk}}$$

n ：全文書数

によって評価することとする。

この式を用いて、先に述べた、比較的頻度の高い用語候補について、各クラスタに属する指標を計算し、評価値順にソートした結果の上位に位置する候補を調べた。各クラスタに対する上位10候補は次の通りである。

クラスタ1：
誤認識、文字認識、認識結果、認識率、音声認識、パターン認識、自動認識、物体認識、文書画像、特徴ベクトル

クラスタ2：
検索機能、検索効率、情報検索システム、検索処理、情報検索、キーワード検索、検索キー、検索方法、検索条件、検索システム

クラスタ3：
グラフ表現、表現形式、内部表現、知識表現、意味表現、表現法、言語表現、表現力、表現方法、記述力

クラスタ4：
認識結果、誤認識、音声認識、文字認識、自動認識、認識率、物体認識、パターン認識、構造解析、文書画像

クラスタ5：

排他制御、性能評価、制御方式、評価結果、制御システム、通信性能、評価法、並行処理制御、制御構造、経路制御

クラスタ６：
成果物、プロジェクト管理、開発支援環境、一元管理、設計情報、設計支援、設計支援システム、支援ツール、ワークフロー管理システム、支援環境

この結果は、先に用語クラスタリングの結果から推定した部分領域にほぼ該当すると考えられる。ただし、クラスタ１とクラスタ４は、実際には明確な区別のない結果となった。これらはいずれもパターン認識／メディア処理という部分領域であることが推定できる。

もう少し順位の下位のものまで調べ、上位50位までの中で、推定された分野として不適切な用語と考えられるものがどの程度含まれているかを調べた結果は、

クラスタ１：５
クラスタ２：６
クラスタ３：３
クラスタ４：４
クラスタ５：４
クラスタ６：１１

となっている。クラスタ６に対して相対的に不適切なものが多くなっているが、この主な原因は、例えば「論理設計」に見られるように、ソフトウェア開発／設計に関わる用語候補が８候補入ってきていることによる。このことから、クラスタ５とクラスタ６は、実際にはあまり明確には分かれていないと考えられる。想定された部分領域がそれぞれ、ソフトウェア開発、ソフトウェア効率であったことを考えると、それほど意外な結果ではない。

全体的に見ると、上位50位までの内で想定された分野に対して不適切と考えられるものはおよそ10%程度であると考えられ、ほぼ妥当な結果が得られていると言える。

クラスタ１とクラスタ４は、上位50位までを見ても、ほぼ同一の部分分野であると考えることができる。ただし、上位にランクされる候補はある程度異なっており、候補として共通するもの36、異なっているもの14という結果になっている。異なっているものの内訳を見ると、クラスタ１でむしろ画像／物体認識関係の候補が上位に来ており、クラスタ４では自然言語解析に関する候補が入ってきている。これは当初用語クラスタリングの結果から推定した分野とは逆の結果となっている。

これら二つのクラスタ決定に対して効果の大きいサ変名詞を調べると、クラスタ１では「認識」が大きな割合を占めており、「検索」がわずかに影響しているという結果であるのに対して、クラスタ４では「認識」が最も大きな要素であることは同様だが、「解析」、「表現」もかなりのウェートを示している。候補のランク付け結果にもこの辺の事情が影響しているものと考えられる。

用語の分野期属性は必ずしも排他的なものではない。同じ用語が複数の部分領域に関連することも考えられる。例えば今回の結果では、「一元管理」がクラスタ６（ソフトウェア開発）の４位になっていると同時に、クラスタ２（データベース）の36位に位置している。分野に依存してややニュアンスは相違するものの、この結果は妥当と言えるであろう。一般には用語が一般的、あるいは概念粒度が大きい場合、複数の分野に帰属しやすくなると考えられる。

４．考察

本研究では、動詞概念との共起に基づき、用語をクラスタ分類することによって、情報処理分野の研究抄録コーパスに対して妥当な部分研究領域を推定することができることを示した。筆者らは既に先行研究[1]において、工学というある程度範囲の広い分野の研究抄録コーパスを用いて部分領域推定が可能であることを示してきたが、今回の結果から、情報処理という相対的に狭い領域においても、同様の方法により、部分研究領域を同定することが可能であることが明かとなった。

また、部分領域を同定する上で利用したサ変名詞との共起関係を用いることにより、別途抽出された用語候補を、同定された部分領域に関連付けて分類することを明らかにした。

部分領域の同定と、部分領域に関連付けての用語候補の分類は、今回検討した範囲では妥当な結果が得られたと考えている。

これらの結果からは、統計的には動詞概念の存在が文書内で取り扱われている名詞的概念を規定しているという仮説に妥当性のあることを示していると言うことができる。

今回の研究で、研究抄録コーパスに出現する

用語を、部分研究分野に関連付けて整理できることが示されたが、一方でいくつかの課題も残されている。

その第一は一つの研究分野に存在する部分領域の粒度に関する問題である。情報処理分野の部分領域としては、今回用語クラスタリングで明かになったもの以外にも様々なものが考えられるが、今回の用語クラスタリングという方法によって、どこまで推定が可能であるかは今後の課題である。

クラスタリングの手法に基づいてより詳細な部分領域を同定するためには、分類すべき名詞系用語、また、基準となる動詞系概念ともに数を増やす必要があると考えられる。しかし、このことは相対的に頻度の低い要素まで取り扱う必要が出てくることを意味している。

クラスタリングなどの統計的手法がどこまで有効に適用できるかは、今後とも検討を必要とする課題である。

もう一つの課題として、今回の分類基準に見られるような、統計的基準に基づく方法で取り扱うことのできる用語候補の範囲がどのようなものであるかという問題がある。今回は、比較的出現頻度の高い用語候補に限定して分類を試みたが、各クラスタに対するスコア上位のものを見る限り、妥当な結果が得られていると考えられる。

しかし問題は、用語候補をある程度頻度の高いものに限定することにより、実際に分類対象とすることのできる候補数が著しく少なくなってしまうことである。今回、全体としては46,021の候補が抽出されているにもかかわらず、出現頻度を40以上という条件で絞り込むことにより、そのうちわずか882候補のみを対象とすることになった。

実際に抽出可能な複合語用語候補を見ると、コーパス内出現頻度が小さいものが大部分である。このことは統計的手法に基づく分類に一定の限界があることを予想させる。統計的手法を適用する限り、例えば出現頻度が10以下という候補に対する解析精度を期待することには問題があると考えざるを得ない。

この問題に対処する一つの考え方は、形態素レベルでの分類に基づいて複合語と部分領域の関係を推定することであろう。しかし、この問題については、領域性が高いと判定された形態素がどのようなものであり、それが複合語の中でどのような役割を果たしているかを検討する必要があろう。

これらの問題をさらに検討することにより、今後、より実用的な形で用語候補を分類・整理する方法論の確立を目指す予定である。

謝辞：
本研究の一部は科学研究費補助金19500135の援助の下に行われた。

参考文献：

[1]Koyama, T., Kageura, K., and Takeuchi, K., Term Extraction Using Verb Co-occurence, Proc. 3rd International Workshop on Computational Terminology, pp.79-82, 2004.

[2]Glover, E. Pennock D.M., Lawrence, S., and Krovetz, R. Inferring Hierarchical Descriptions, Proc. 11th International Conference on Information and Knowledge Management, pp.507-514, 2002.

[3]Chuang, S. and Chien, L., A Practical Web-based Approach to Generating Topic Hierarchy for Text Segments, Proc. 13th International Conference on Information and Knowledge Management, pp.127-136, 2004.

[4]小山照夫、影浦峡、竹内孔一、日本語専門分野テキストコーパスからの複合語用語の抽出、情報処理学会自然言語処理研究報告、2006-NL-176、pp.55-60、2006.