2008－NL－188 (8)
2008／11／26

形態素出現パタンに基づく文書集合類似性評価

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

類似したテーマを類似した論述構造で扱う文書を集めた文書集合同士では、共通 した形態素利用パタンが出現すると予想される。一方で、形態素出現傾向の相違は、文書集合の間で扱われるテーマあるいは論述構造の相違を示すものと考えられる。本研究では、異なる学会の研究抄録を集めた文書集合について、形態素出現傾向が集合間の類似性や相違を評価する指標として有効であることを確認するとともに、どの形態素が文書集合間の類似性ないしは相違性に寄与しているかを評価する方法について発表する。

An evaluation of document set similarity based on morpheme occurrence patterns

Teruo KOYAMA* Koichi TAKEUCHI**

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

We can assume that two different document sets may show similar morpheme occurrence patterns, if the sets both discuss about similar topics with similar discussion manners. In this paper, the authors show the occurrence patterns of morphemes really indicates the similarity of the sets. The authors also show the difference of the patterns in both sets indicate the difference of topics or discussion manner between the sets. The authors also show how to find key morphemes that indicate the similarity or difference of the sets.

1. はじめに

学会抄録文書など、特定のテーマを扱う文書集合が存在する場合、文書集合の中では、共通するいくつかのテーマについて、類似した論述が行われていると期待することができる。ここで、テーマや論述を構成する基本単位は形態素であることから、二つの文書集合が、類似したテーマを類似の論述形式で扱っている場合、それぞれの集合における形態素出現パタンにも類似性が存在すると考えることができるであろう。また、類似性あるいは相違に対して各形態素がどのように寄与するかを評価することができれば、文書集合の間の類似性ないし相違がどのような性格のものであるかを明らかにできると期待される。

文書中に出現する形態素から文書ないしは文書集合の近接性を評価し、活用する試みとしては、自由テキストを検索要求として用いる文献検索[1],[2]や、検索対象文書をあらかじめ特定の観点から分類しておく文書クラスタリングの検索問題への応用[3]などが試みられている。しかし、計算された類似性ないしは相違がどのような要因による、どのような性格のものであるかについては十分に検討されてきたとは言えない。

本研究では、様々な学会における研究発表抄録データに基づき、類似のテーマを類似した論述形式で扱うと期待できる学会文書の間で、形態素出現パタンが類似していることを確認するとともに、類似性の認められる文書集合の間で、共通性を示す形態素ならびに相違性を示す形態素がどのようなものであるかを推定し、文書間の類似／相違の性格がどのようなものであるかを明かにすることを試みる。

2. 形態素出現パタンの比較

形態素出現傾向を評価する上で実際に用いるデータとしては、NTCIR-I[4]に収録された日本語学会発表データを用いる。このデータは合計 59 学会から集められた抄録を収録しているが、学会によってはごくわずかなデータしか存在しないものもある。今回は抄録が 1,000 以上収録されている 25 学会だけを検討の対象とした。これらの学会のうちで、どの学会が形態素出現パタンが類似しているかを判定した上で、類似する学会の間でどのような形態素が類似性に関わっており、また逆にどのような形態素が相違を構成しているか、また、それらの類似／相違は、対象に関するものなのか、それとも操作や論述に関するものであるのかを推定することを試みる。

最初に考察の対象とする形態素を限定する。これは、実際に予備的な解析を行った結果、多くの機能語や抽象名詞(「こと」、「もの」など)が文書集合間の類似を示唆するという結果が出てきたことによる。これらは、取り扱われる内容の類似性を示唆する効果はほとんど期待できず、また、論述の形式に関わることもないが、様々な文書集合で類似した出現パタンを持つ可能性があると考えられるものである。これらの形態素は、実際の文書集合で扱われているテーマや論述形態の類似性や相違を考える上であまり参考にならないと判断した。

実際に文書集合のテーマや論述の面から類似性や相違を考えるとき、記述の内容や形式をある程度反映する形態素に限定して検討を行う必要がある。記述内容に関わる可能性の大きい形態素を選び出す方法は、基本的には用語として有力な形態素を選出する基準に準ずると考えてよい。用語性を判定する基準としてはいくつかのものが考えられるが、ここでは簡単に Tf-IDf 基準を用いることとした。

まず、すべての文書を一つの集合と考え、各形態素について Tf-IDf 値を求め、値の大きい順に適当数の形態素を選出する。ただし、形態素のうちでひらがなのみからなるもの、および漢字一文字の形態

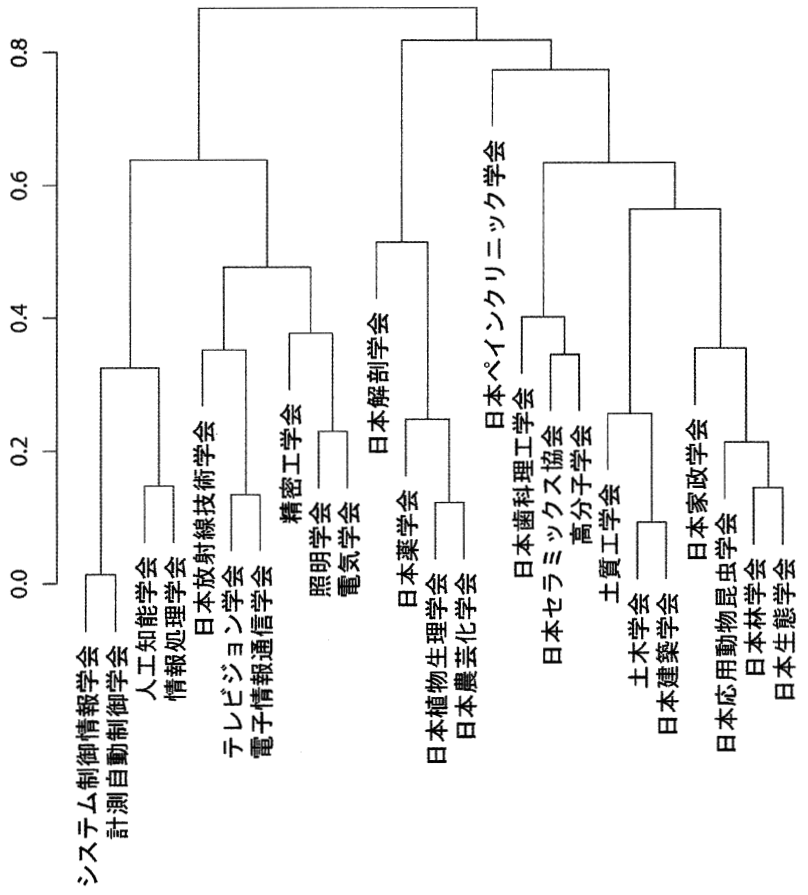

図 1. Cos 距離に基づく学会文書集合のクラスタリング結果

素を省略している。これらの形態素は、内容記述の効果が低いと判断したことによる。特定の形態素を除外した上で、Tf-IDf 値の大きい方から 500 形態素を選出し、これらの出現パタンを検討する。

選出された形態素について、各学会について出現総頻度を求め、これを当該学会に属する文書数で除し、正規化する。結果として各学会に対して 500 次元のベクトルが得られる。形態素出現パタンからみた二つの学会の間の類似度ないしは距離は、相当するベクトルの間の相関ないしは距離として評価することが可能である。

類似度ないし距離を考えるにあたって、どのような尺度を用いるのが適当かを判断するため、いくつかの尺度を想定した上で、それぞれの尺度を選択した場合、各学会の相互的関係がどのようなものになるかを調べる。この目的のため、25 の学会に対応する 25 の 500 次元ベクトルのクラスタリングを試みる。いくつかの距離尺度と Pearson の相関係数について結果を検討した結果、Cos 距離を用いた場合と相関係数を用いた場合、ほぼ同等のクラスタリング結果が得られ、かつ、これらの結果が最も直観的に妥当と判断できることが分かった。以下では Cos 距離を距離指標として採用する。Cos 距離を類似指標としたクラスタリング結果を図 1.に示す。

クラスタリング結果を見ると、類似度が特に大きい学会の組として、システム制御情報学会(シ)－計測自動制御学会(計)、人工知能学会(知)－情報処理学会(情)、テレビジョン学会(テ)－電子情報通信学会(信)、日本植物生理学会(植)－日本農芸化学会(農)、日本林学会(林)－日本生態学会(態)の 5 つが考えられる。これらに加えて、図からは必ずしも明らかではないが、集合間の距離が大きいものとして、人工知能学会(知)－日本解剖学会(解)の組を別途検討する。以下ではこれらの学会の組み合わせについて、どのような形態素が類似性あるいは相違に関係しているかを調べる。

3. 文書集合間の類似と相違を表す形態素

類似あるいは相違に対して各形態素がどの程度寄与しているかを評価する指標として、次のものを考える。今、二つの学会について、それぞれ全体として 500 次元のベクトルが得られており、これらのベクトル間での Cos 距離が求められている時、それぞれのベクトルから特定の一形態素に相当する要素を除外することにより、499 次元のベクトルを得る。これらのベクトルの間の距離は、相当する形態素の効果を無視した場合の距離になると考えられるから、この縮退したベクトルに対する Cos 距離を計算し、

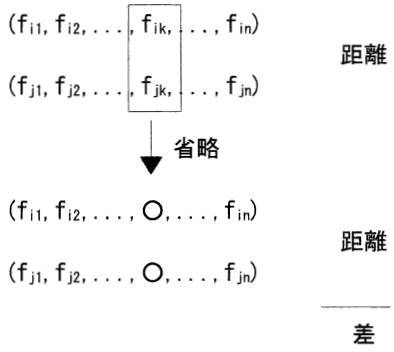

図 2. 形態素 k の寄与計算

元の距離との値の差を求めてやれば、指定された形態素が類似性に対して寄与する強さが得られると考えてよい(図2)。

比較の対象とする各学会の組合わせについて、選択した形態素の全てについてこの指標を計算し、値の大きさにしたがって形態素をソートした上で、最も類似度および相違に関連するものとして順序最上位 10 位と最下位 10 位の形態素を調べた。該当する形態素を表 1 に示す。

最上位の形態素は類似性に寄与するものと考えられるが、実際の類似性としては、当該形態素が比較対象とする学会以外の学会と比較して、共通して出現頻度が高い場合も、逆に頻度が低い場合も寄与が大きいと判定されると考えられる。そこで、他の学会と比較して頻度的に高いいものであるか、低いものであるかを、後に付与した符号で示している。一方、最下位の形態素は、比較対象とする二つの学会の間での相違に寄与するものと考えられ、一方の学会で比較的出現頻度が高いのに対して、他方では相対的に頻度が低いことを示唆している。後につけた文字は、どちらの学会に対してより頻度が大きいかを示すものである。

以下、得られた結果に関する考察を行う。

3.1. システム制御情報学会／計測自動制御学会

類似性に関しては、「制御」「モデル」が上位に来ているのは自然な結果と考えられる。「時間」「状態」も、制御に関連する可能性があると言えよう。その他は一般的な論述に関わるものと考えられる。一方相違では、「測定」「センサ」が計測関連の用語として相違に関わっており、また「運動」「変化」「温度」についても、「測定」と関連がある可能性が考えられる。一方、「アルゴリズム」「データ」「処理」は、情報処理との類似性を示唆している。

システム制御情報学会／計測自動制御学会				人工知能学会／情報処理学会			
類似要素		相違要素		類似要素		相違要素	
制御	+	測定	計	システム	+	知識	知
モデル	+	生産	シ	情報	+	学習	知
方法	+	アルゴリズム	シ	提案	+	処理	情
提案	+	運動	計	手法	+	並列	情
研究	+	変化	計	利用	+	推論	知
手法	+	データ	シ	設計	+	データ	情
場合	+	処理	シ	方法	+	方式	情
構成	+	実験	計	モデル	+	画像	情
時間	+	温度	計	構造	+	管理	情
状態	+	センサ	計	言語	+	オブジェクト	情
テレビジョン学会／電子情報通信学会				日本植物生理学会／日本農芸化学会			
類似要素		相違要素		類似要素		相違要素	
システム	+	画像	テ	活性	+	植物	植
方式	+	信号	テ	gt	+	生産	農
検討	−	映像	テ	lt	+	細胞	植
結果	−	通信	信	結果	+	酵素	農
特性	+	開発	テ	解析	+	検討	農
可能	+	提案	信	遺伝子	+	物質	農
処理	+	ディジタル	テ	合成	+	分解	農
実現	+	記録	テ	阻害	+	化合	農
実験	+	解析	信	存在	+	照射	植
情報	+	表示	テ	培養	+	タンパク質	植
日本林学会／日本生態学会				人工知能学会／日本解剖学会			
類似要素		相違要素		類似要素		相違要素	
調査	+	個体	態	構造	+	細胞	解
結果	−	植物	態	研究	+	システム	知
変化	+	作業	林	結果	−	知識	知
影響	+	処理	林	機能	+	神経	解
関係	+	被害	林	可能	+	学習	知
研究	−	試験	林	方法	+	問題	知
環境	+	発生	林	検討	−	設計	知
明らか	+	検討	林	組織	+	提案	知
速度	+	培養	林	関係	+	モデル	知
比較	+	濃度	林	解析	−	観察	解

表 1. 二つの学会を比較した場合の類似性および相違を表す形態素

3.2. 人工知能学会／情報処理学会

類似性に関しては、「システム」「情報」「モデル」「構造」「言語」など、情報処理分野の基本的な形態素が現れている。また、「設計」が共通する操作として現れている。相違では、「知識」「学習」「推論」が人工知能学会に特有の形態素として現れているのが特徴的で、より一般的な情報処理関連の形態素が相違を構成していると考えられる。ただし、「データ」や「オブジェクト」などが相違に関わり、「システム」「モデル」が類似に関わることについて、どのような要素がこれらの区別を生じさせているかはこのデータだけでは不明である。

3.3. テレビジョン学会／電子情報通信学会

類似性を示すものとして、「システム」「処理」「情報」という、情報処理関連の形態素が見られる。「特性」は一般的な形態素ではあるが、これらの分野で共通するものとして興味深い。「方式」についても同様である。これらについては、学会の文脈から特定の対象を指す形態素として用いられている可能性がある。操作として「実現」「実験」が現れていること、「検討」がむしろ使われない傾向にあることが特徴的である。相違では、「画像」「映像」「表示」といった、画像処理関係の形態素がテレビジョン学会に特徴的なものとなっている。「ディジタル」「記録」がテレビジョン学会に特徴的に現れているのは、時期

的な要素も考えられるかもしれない。「信号」がテレビジョン学会、「通信」が電子情報通信学会といういのも、自然であろう。「開発」「提案」「解析」といった、一般的操作については、相違に関わってくることの背景はあまり明らかではない。

3.4. 日本植物生理学会／日本農芸化学会

類似要素として「活性」「遺伝子」「合成」「阻害」「培養」という、生物学／植物学関連の形態素が上位に来るのは自然である。「gt」「lt」についていえば、数値比較が頻繁に行われていることを示唆している。その他の一般的操作を示すものについては背景はあまり明らかではない。相違では、「植物」「照射」「タンパク質」が植物生理学会に、「生産」が農芸化学会に特徴的であるのは自然である。農芸化学会で「酵素」「物質」「分解」「化合」が見られるのは、化学というアプローチに関連する可能性がある。

3.5. 日本林学会／日本生態学会

類似要素として「環境」があるのは自然であると考えられる。「変化」「影響」「速度」が共通するのも理解できる。「関係」も、おそらくはこれらと関連するものと考えられるが、位置づけはあまり明確ではない。操作／論述として「研究」がむしろ少なく、「調査」「比較」が多いのも興味深い。相違では、「個体」「植物」が生態学会に偏っており、「被害」「発生」「濃度」が林学会に偏っている。「作業」「試験」「培養」という操作が林学会に現れているのも特徴的である。

3.6. 人工知能学会／日本解剖学会

これまでの組が全体として類似性の高いもの同士の組み合わせを見てきたのに対して、この組は全体としての相違が著しいものである。しかしながらここで共通する要素として、「構造」「機能」「組織」というものが出現していることが注目される。これらはそれぞれの分野における重要な概念構成要素となっていると考えられる。しかし、一方でそれぞれの分野でのこれらの形態素が表す対象を考えるなら、人工知能学会ではコンピュータシステムや応用対象となるシステムに関する概念であり、解剖学会では生体システムに関する概念であると考えられる。実際には、それぞれの示す対象は、同型異義とまでは言えないにしても、相当程度異なった対象であることが予想される。これらは「システム」という、幅広い分野を横断して出現しうる、いわばメタ概念に関連するものであり、分野の扱う対象がたとえ異なっていても、共通要素として出現する可能性のあるものと考えられる。その他の類似要素は一般的な対象、操作がほとんどであり、強いて挙げれば「解析」が用いられることが少ないことが注目される。相違では「細胞」「神経」が解剖学会、「システム」「知識」「学習」「モデル」が人工知能学会というのは理解しやすい。操作として解剖学会では「観察」人工知能学会では「設計」「提案」が多いのも妥当な結果といえる。これらの操作を「解析」との対照としてとらえることも可能であろう。

全般として、形態素出現傾向から見た学会相互の類似点、相違点は妥当なものと考えられるが、操作、論述に関する形態素については必ずしも意味づけが明解でないものも多い。

4. 考察

様々な分野の文書集合に出現する形態素に基づき、集合間の距離および、問題とする形態素が距離に寄与する傾向を求めることにより、それぞれの集合がどのような形で類似ないし、相違しているかに対する手がかりを得ることができる。厳密に言えば、形態素同士の共起に関わる傾向などまで考慮に入れる必要があると考えられるが、とりあえずの近似としては十分に納得できる結果が得られたと考えられる。しかし、一方で分野ごとの類似や相違がどのような性格のものであるかについては、さらに検討を必要とするであろう。

今回、文書集合間の類似と相違に関わる形態素として取り出されたものを大きく分類するならば、1.「システム」や「知識」などのように、領域が取り扱う対象に関わるもの、2.「設計」や「処理」などのように、対象に対する操作と位置付けることができるもの、3.「提案」や「研究」などのように、論述の方向性を示すもの、などに分けられると考えられる。

対象にかかわるものは、分野の類似や相違を表すものとして有効なものが多いと考えてよいが、しかし中には「手法」「方法」「関係」など、一般的で、分野の特徴を代表するとは考えにくいものも含まれている。これらはむしろ、分野で採用されている議論の構成を表すものと考えることもできるであろう。また、人工知能学会と解剖学会との比較で見られたように、「システム」という広い概念が、全く異なる複数の分野で、異なる対象に対して適用されることもある。これらは、広い意味では類似性と言うこともできるが、一般に研究分野の類似性として意識されるものとは異なっている可能性がある。

操作について言えば、多くの操作は操作の対象と密接な関連を持つことが予想され、操作対象を想定することにより、分野の特徴を示していると言うこともできる。ただし、操作と対象の対応は様々な組み合わせが考えられることから、かなりの曖昧性が残ると考えられる。また、対象と同様に例えば「比較」

などのように、対象の特定にあまり役立たないものも存在する。

論述については、一般的な述語であることが多いことから、分野の比較を行うという観点からは必ずしも有用ではない可能性がある。林学会／生体学会で「研究」があまり用いられていないのは、例えば「調査」という操作が類似に関わるものとして出現していることと併せて考えるなら、たしかにこれらの学会の特徴と言うこともできようが、ここに見られる特徴が、例えば文書分類などの何らかの目的に対して有効であるかどうかはあまり明らかではない。

全般に、分野の類似や相違、特に類似を示すと考えられる形態素として、記述の内容を表すものの他に、記述の構成を表すと考えられるものが相当程度入ってきており、どのような視点から類似ないし相違の尺度が構成されているか、幾分分かりにくくなっている面があると考えられる。これらの点についてはさらに検討が必要であろう。

5. 結論

いくつかの異なる学会からの研究抄録を集めた文書集合において、形態素を適切に制限することにより、学会ごとの形態素出現頻度に基づいて、あらかじめ想定される学会間の類似度をほぼ再現する距離尺度を構成できることを明かにした。また、距離に対する各形態素の寄与を求めることにより、学会間の類似や相違がどのような要素に影響されているかを明かにする手法を提案した。

今後は、文書や文書集合の類似ないしは相違に関して、どのような観点から評価が行われる必要があるか、また、そのような観点を反映させるためにどのような形態素の組を設定すべきかに関して検討を進めていく必要がある。

謝辞:
本研究の一部は科学研究費補助金 19500136 の援助の下に行われた。

参考文献:

[1] 徳永健伸、情報検索と言語処理、東京大学出版会、1999.

[2] 佐々木稔、北研二、ランダムプロジェクションによるベクトル空間情報検索モデルの次元削減、自然言語処理、vol. 8、no. 1、pp. 79-84、2002.

[3] 川谷隆彦、多文書間の共通性分析に基づく文書クラスタリング(情報検索)、情報処理学会論文誌、vol. 47、no. 6、pp. 1903-1917、2006.

[4] KANDO, N., and NOZUE, T. eds., Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Proc. NTCIR Workshop I, 1999.