表題解析による科学技術論文の詳細分類

今井 俊

北陸先端科学技術大学院大学 情報科学研究科

1999 年 2 月 15 日

キーワード: 論文の分類, 表題解析, 機能語, 専門用語集.

様々なメディアから収集できる科学技術論文の数は増しているが、その数が増えるほどユーザは求めている論文だけを検索することが困難となる。現在、論文の検索にはキーワード検索システムが多く利用されている。しかし、キーワード検索システムには、検索もれや絞り込みが不十分などの問題がある。これに対し分野毎に整理された論文集を利用することは、その分野に精通していない人でも比較的容易に論文を特定することができる。しかし分類作業にはかなりの労力が必要なため、自動的に論文を分類する技術が必要とされる。

テキストの自動分類では、テキストの本文中に現れる単語集合とその頻度に基づいて、そのテキストがどの分野に属するかを決定する方法が標準的に用いられている。しかし、実際に人間が論文を分類する場合には、表題などのサマリー情報を利用することが多い。特に、ある分野の専門家は、その分野の論文に関しては、論文の表題を見るだけで、その論文がおおよそ何についての論文なのかを推定でき、適切な分類細目を決定できることが多い。たとえば、人工知能の専門家は、「プロダクション・システムによる線画の解釈」という表題を見ると、『この論文は「線画の解釈」に関する論文で、「プロダクションシステム」を手法として用いているのだろう』という推測ができるのが普通である。このような推測が可能な理由としては以下の 2 つが考えられる。

(1) 専門家は専門用語に関する知識を十分に持っている

(2) 論文表題は論文の最も短い要約となっており、論文の内容と密接に関連した専門用語が表題に含まれることが多い

もし、この仮説が正しいとすれば、ある分野の専門用語集を用意することによって、論文表題からその論文の分類細目（カテゴリ）を機械的に決定できる可能性がある。本研究で

は、このような考え方にたって、論文表題を専門用語集を用いて解析することにより、その論文の分類細目を決定する方法を提案する。

本研究で作成した表題解析を利用した自動分類システムは、標準化とコード割当の二段構成をとる。まず標準化では論文表題に対して、以下の処理を繰り返し適用することによっていくつかの部分要素に分割、整形する。

1. **文字列処理による不要部分の削除：**ここでは、先頭や末尾のシステム名などの固有名詞を削除する。
2. **文字列処理による分割：**表題が「ＸとそのＹ」という形ならば、「Ｘ」と「そのＹ」に分割する。
3. **文字列処理による木構造の変形：**表題が「ＸのＹへの応用」という構造であれば、「Ｘを応用したＹ」に変形する。
4. **単語列処理による不要部分の削除：**末尾の「〜の (方法 | 実現 |…)」といった、分類には寄与しない単語を削除する。
5. **単語列処理による分割：**表題が「Ｘ 付属語相当句 Ｙ」という構造であれば、「Ｙ」「付属語相当句」「Ｘ」に分割する。

標準化によって得られる論文表題の部分要素のほとんどは、複合名詞句、付属語相当句、となる。コード割当では、専門用語集を用いて複合名詞句に含まれる専門用語を見つけ、分類コードを決定する。専門用語集は、その分野で使われる専門用語とその分類コードを定義したものである。本システムでは、分類コードとして、岩波情報科学辞典の用語の木のコードを用いる。複合名詞句に含まれる専門用語を見つけるには、以下の方法を用いる。

1. 複合名詞句に対して、末尾から最も長く一致する専門用語を求める
2. 複合名詞句の末尾が「システム」「機構」などの省略可能語の場合は、これを省略したものに対しても、1. を適用する。
3. 複合名詞句に「から、の、へ」などの格助詞が含まれる場合は、格助詞以降を削除したものに対しても、1. を適用する
4. 複合名詞句に「的、型、の」が含まれる場合、これを削除したものに対しても、1. を適用する
5. 複合名詞句に「**サ変名詞**」や「**名詞**＋化」が含まれる場合、これを削除したものに対しても、1. を適用する

こうして、得られた専門用語から、それに対応する分類コードを求め、これを割り当てるべきコードとする。

人工知能学会誌に掲載された 369 論文を分類する実験を行なった。369 論文のうち 292 論文は人手で作成した分類コードと一致した。その精度は、79%となる。残りの 77 論文

のうち 36 論文は、正しい分類コードの決定のために、何らかの推論が必要なものである。その他の論文のほとんどは、辞書の不備、形態素解析の誤りといった、基本的には分類の誤りではないものであった。