

表題解析による科学技術論文の詳細分類

1 V-6

今井俊 佐藤理史

北陸先端科学技術大学院大学 情報科学研究科

1 はじめに

論文の自動分類では、論文の本文中に現れる単語集合とその頻度に基づいて、その論文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、人間が論文进行分类する場合には、このような方法を用いることはない。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らう』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のは、(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、という理由によると考えられる。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。

本稿では、このような考え方に立って、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について検討する。対象分野は情報科学であり、専門用語集として岩波情報科学辞典 [2] を用いる。

2 システムの概要

作成したシステムの概要を、図1に示す。本システムは、標準化とコード割当の2つのモジュールから構成される。

2.1 標準化

標準化では、動詞や機能語を手がかりに論文表題をいくつかの部分要素に分割し、整形する。具体的には、以下の標準化手法を繰り返し適用することによって標

Automatic classification of technical papers by using title analysis.

Shun IMAI and Satoshi SATO.

School of Information Science, Japan Advanced Institute of Science and Technology.

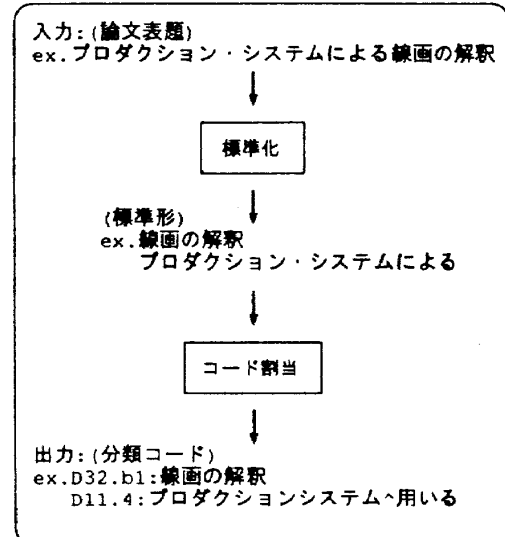


図 1: システム概要

準化を行なう。

1. 不要部分の削除（文字列処理）：先頭や末尾のシステム名などの固有名詞を削除する。
例：「MBT1: 実例に基づく訳語選択」→「実例に基づく訳語選択」
2. 分割（文字列処理）：表題が「XとそのY」という形ならば、「X + そのY」に分割する。
例：「事例ベース推論の対話型モデルとその機械調整支援への適用」→「事例ベース推論の対話型モデル + その機械調整支援への適用」
3. 不要部分の削除（単語列処理）：末尾の「～の（方法 | 実現 | …）」といった、分類には寄与しない部分を削除する。
例：「専門用語辞書の自動的ハイパーテキスト化の方法」→「専門用語辞書の自動的ハイパーテキスト化」
4. 分割（単語列処理）：「X 助詞（動詞 | 機能語）Y」という構造であれば、「Y + X 助詞（動詞 | 機能語）」に分割する。
例：「知識獲得のための知識表現」→「知識表現 + 知識獲得のための」、「知識工学を応用した機械設計 CAD」→「機械設計 CAD + 知識工学を応用した」

5. 木構造の変形: 「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。

例: 「知識工学の機械設計CADへの応用」→「知識工学を応用した機械設計CAD」

2.2 コード割当

標準化の結果として得られる論文表題の部分要素のほとんどは、(a) 複合名詞句、(b) 複合名詞句+付属語、(c) 複合名詞句+格助詞+動詞、のいずれかの形をとる。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。

例: 「分散型問題解決」(下線部は専門用語を表す。)

2. 複合名詞句の末尾が「システム」「機構」などの削除可能語の場合は、これを削除したものに対しても、1. を適用する。

例: 「医療知識ベースシステム」

3. 複合名詞句に「の」が含まれる場合は、「の」以降を削除したものに対しても、1. を適用する。

例: 「類推の定式化」

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。

3 予備実験

人工知能学会誌に掲載された100論文(1986年9月～1990年3月)を分類する実験を行なった。分類の実行例を図2に、実験結果を表1に示す。

100論文のうち、人間でも表題からは適切な分類コードを決定できないものが2論文あった。残りの98論文のうち80論文に対して、システムは正しい分類コードを割り当てることができた。その精度は、

$$\frac{\text{システムが正しく分類した論文 (80 論文)}}{\text{人間が分類できる論文 (98 論文)}} = 81(\%)$$

となる。

残りの18論文のうちの13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。

表 1: 予備実験の結果

	システムによる分類		
	正しい	誤り	分類できない
人間は分類できる	80	11	7
人間も分類できない	0	0	2

1. 類推の定式化とその実現法
→ D12.35:類推

2. 命題自己認識論理における決定手続き
→ A44.422|C12.631|C21.514|C52.226:手続き
A41.3|A51.227:命題論理における

3. 仮説生成に基づく分散型問題解決
→ D12.1:問題解決
D12.36|D14.344:仮説生成に基づく

4. 病歴管理システムを統合した医療知識ベースシステム
→ C33.d4|D11.3|D13.111:知識ベース

5. 多重世界機構による常識推論
→ D12.3:推論
D11.35:多重世界による

図 2: 分類の実行例

例えば、「知識ベースの保守方式」は、「知識ベース管理(D11.32|D14.21)」に分類すべきであるが、これには、「知識ベースの保守」→「知識ベースの管理」→「知識ベース管理」といった推論が必要になる。前者は、「保守」は「管理」に含まれるという知識が、後者には、「XのY」が「XY」に変形できるという文法的知識が必要である。このような推論は、いわゆる「言い換え」によって実現できると考えられる。

4 おわりに

本論文では、表題を解析することによって、科学技術論文を自動的に分類する方法を提案し、予備実験の結果を示した。今後、実験の規模を拡大して、本方法の有効性を検証するとともに、(1) 文献[1]に見られるような機能語の分類を取り入れる、(2) 言い換えを実現することによって本方法の精度を向上させる、などの改良に取り組む予定である。

参考文献

- [1] 松村敦, 池田和幸, 高須淳宏, 安達淳. 構造化インデックスを用いた情報検索システム. アドバンス・データベース・シンポジウム '97 論文集, pp.151-158, 1997.
- [2] 長尾真 他編. 岩波情報科学辞典, 岩波書店, 1990.