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

2004－NL－161 (10)
2004／5／14

特許文書の多観点分類について

田中一成
富士通研究所

特許分析では、内容を人手で判断して、「何についての特許か(発明対象)」や「何を目的とした特許か(目的)」といった観点で分類を行い、クロス集計してグラフを作成することなどが有効である。本稿では、特許文書を対象として、発明対象と目的といった複数の観点で自動的に分類を行う手法について検討・実験を行ったので報告する。実験の結果、特許文書の多観点分類を行うためには、各特許文書から観点ごとに内容を表す特徴情報を抽出した後に、抽出された特徴情報のレベルを合わせたり、複数の観点間で整合性が取れるように修正する処理が有効であることが分かった。

Multi-viewpoint clustering of patent documents

Kazunari Tanaka
Fujitsu Laboratories Ltd.

When we analyze patent trend, we often use various kinds of graphs, tables and figures. In this process, Multi-viewpoint clustering is very useful method to analyze patent documents. In particular, we focus on a clustering method based on viewpoints of "the subject of an invention" and "the purpose of an invention". Through experiments on clustering patent documents, we show promising result on Multi-viewpoint clustering of patent documents by amending extracted features. Finally, we summarize open problems to practical application of Multi-viewpoint clustering.

1 はじめに

特許調査、特に動向調査では、グラフや表、流れ図などの図を作成することで業界の動向を分析するという作業を行う。このため、特許調査を支援するツールの研究も行われており、特許の書誌情報やキーワードを利用したグラフや図を自動的に作成できるようになっている[1]。しかし、実際の特許調査では、元々特許に付けられている書誌情報を利用するばかりではなく、内容を人手で判断して、発明対象や目的といった観点で分類を行い、クロス集計してグラフを作成するなどしており、このような分類を行うために、大変な時間と労力を要している。

一般に文書を自動的に分類するための技術としては、文書クラスタリングがある。文書クラスタリングでは、共有される特徴素(文書を特徴付ける要素、キーワードなど)によって複数の文書がまとめられるので、クラスタごとに分類の観点が異なってしまい人にとって解釈の難しいクラスタができてしまう。

このような文書クラスタリングの問題点に対し、Web ページについては、「意見」「説明」「紹介」といった観点で分類する研究も行われている[2]。

本稿では、特許文書を対象として、発明対象と目的といった複数の観点で分類を行う手法について検討・実験を行ったので報告する。

2 アプローチと課題

複数の観点における分類(多観点分類)を行うためには、特定の観点において特許の内容を

表す特徴情報を選び出して利用することが重要であると考え、特徴情報を抽出し、抽出された特徴情報で特許文書をまとめ上げることで分類[1]を行うことを試みた。

今回は、特許調査においてよく活用される発明対象の観点と目的の観点において分類を行う。また、発明対象の観点については、全体を表す基本分類と、その構成要素を表す構成要素分類の２階層のレベルに分けて分類する。

２．１　基本アプローチ

以下に処理の手順を示す。

①特徴情報を抽出する

特許文書を係り受け解析し、観点ごとに定義した抽出ルールを適応することで、各観点での内容を表す特徴情報を抽出する。また、特許文書の構造を利用して、抽出ルールを適応する文書中の範囲を限定することでごみを減らし、抽出精度を高める。例えば、特許文書の「産業上の利用分野」の段落から抽出された係り受け組に対し「関する」の係り元を発明対象の特徴情報として抽出するというルールを適応して、「制御装置に－関する」という係り受け組から「制御装置」を抽出する。特徴情報は１件から複数抽出する。

また、予備実験により、発明対象の特徴情報から「の」で名詞に係る係り受け組では、係り先は係り元の構成要素になっている場合(「エレベータのかご」など)が多いことが分かったので、係り元を発明対象の観点での基本分類を表す特徴情報、係り先を構成要素分類を表す特徴情報として抽出する。

②意味的に近い特徴情報を適当な抽象度で統一化する

特徴情報を構成する文字が一定の割合以上同じものをまとめることにより、「乗心地」と「乗り心地」のように表記ゆれを統一化する。また、例えば、自動車に関する特許の構成要素分類において「フレーム構造」と「フレーム」をまとめるといったように、各観点において特徴情報を抽象化してまとめ上げることで分類を行う。

[1] ここでは、説明上「分類」という言葉を用いるが、分類ルールや機械学習による分類とは異なる。

２．２　基本アプローチの課題

２．１節の方法で予備実験を行った結果、以下のような課題があることがわかった。

a. 特許間で特徴情報のレベルが合わない

発明対象の観点では、基本分類と構成要素分類の２階層で抽出を行ったが、例えば、自動車のエンジンに関する特許であっても、特許の書き方で「自動車に関する」と書かれる場合と「エンジンに関する」と書かれる場合があり、ある特許で基本分類の特徴情報として抽出されているものが別の特許では構成要素分類の特徴情報として抽出される。

b. 観点にそぐわない特徴情報が抽出される

少ないルールで特徴情報抽出の再現率を上げるために、様々な特許に使える汎用的な抽出ルールを用いるので、例えば、「関する」の係り元を基本分類の特徴情報として抽出するというルールを適応した場合に、「組立性に関する」という記述から、基本分類の特徴情報として目的を表す「組立性」のようなものが抽出されてしまう。

c. 複合語から基本分類と構成要素分類とを切り分けることができない

例えば、「エレベータの呼出装置」からは、「エレベータ」を基本分類、「呼出装置」を構成要素分類の特徴情報として抽出できるが、「エレベータ呼出装置」という複合語で表現されている場合には基本分類と構成要素分類の切り分けができない。

d. 抽出された特徴情報にごみが多い

係り受け解析の間違えや抽出ルールが甘すぎることにより誤ったものが抽出される。例えば、「並び」「ものに」など。

多観点分類を実現するにはこれらの問題を解決する必要がある。そこで、特徴情報の抽出結果に修正を施すことによってこれらの問題の解決を試みた。

２．３　改良アプローチ

２．２節で述べたような問題に対して、特徴情報抽出後に以下のような修正処理を加えることで解決を試みた。

①「産業上の利用分野」の記述を利用してレベルを統一(問題 a の解決のため)

特許の「産業上の利用分野」の段落では、発明の抽象的な技術分野から特に改良を施した構成要素までが１文で書かれる場合が多いため、これを利用して発明対象の観点の基本分類と構成要素分類の特徴情報のレベル合わせを行う。

発明対象の特徴情報は、２．１節の①で説明した方法で、基本分類と構成要素分類に分けて抽出するが、この両方が抽出できない場合には、全て基本分類として抽出する。

例えば、発明対象の特徴情報として「制御装置」のみが抽出された場合には基本分類として扱われる。そこで、抽出後の修正処理において、例えば、「産業上の利用分野」に「ハイブリッド車両の制御装置」という記述があった場合には、抽出時の基本分類の特徴情報「制御装置」の係り元である「ハイブリッド車両」が他の多くの特許で基本分類として抽出されているかを判断基準にして、基準を満たすときには、この特許についても、「ハイブリッド車両」を基本分類の特徴情報とし、これまで基本分類の特徴情報であった「制御装置」は構成要素分類として扱うことにする。

この例では、最も単純な例で説明したが、「ハイブリッド車両」と「制御装置」が「の」で係っていない場合や、「制御装置」の係り元の更に係り元が「ハイブリッド車両」であった場合にも同じように処理する。

②複数観点での抽出結果を利用して修正(問題 a、b の解決のため)

予備実験の結果、効果のありそうであった「用途」と「発明対象の基本分類」、「発明対象の基本分類」と「目的」、「発明対象の基本分類」と「構成要素分類」のそれぞれの組み合わせでの抽出結果を比較して抽出結果を修正する。

ここで、「用途」の観点と言っているのは、例えば、「車両用充電制御装置」の「車両用」の部分である。特許文書では、「車両の充電制御装置」と同じ意味でこのような書き方をされる場合も多い。そこで、発明対象の観点で特徴情報を抽出するとき、特徴情報が「･･･用～」となっていた場合、「･･･用」という文字列を用途の観点の特徴情報として抽出する。用途として抽出された特徴情報が、他の特許では発明対象の基本分類を表す特徴情報として抽出されている場合には、用途として抽出された特徴情報を発明対象の基本分類を表す特徴情報として観点を変更する。

抽出時に「の」で係る係り受け組と同じように処理する方法もあるが、「走行用電動機」や「歩行用ロボット」といったように基本分類と構成要素分類の関係にない場合も多いので抽出後に修正することにした。

発明対象の基本分類と目的の観点での抽出結果を比較して、同じ表記の特徴情報が抽出されている場合には抽出されている件数が多く確信度の高い観点にのみ特徴情報として残して、他の観点からは削除する。確信度は、抽出ルールの精度に応じて設定しておく。発明対象の基本分類と構成要素分類についても同じように比較して修正を行うが、この場合は単純に削除するのではなく、基本分類から構成要素分類、構成要素分類から基本分類といったように確信度の高いほうへ特徴情報の観点を変更することで特徴情報のレベルが合うように修正される。

③複合語からなる特徴情報を分割(問題 c の解決のため)

発明対象の基本分類の特徴情報には「エレベータ呼出装置」のように基本分類を表す特徴情報と構成要素分類を表す特徴情報がつながって複合語となっている場合がある。そこで、基本情報の特徴情報として抽出された中で部分一致するものを探すことにより分割を試みる。この例の場合「エレベータ」と部分一致するので、「エレベータ」の部分を基本分類とし、それより後ろの部分の「呼出装置」を構成要素分類の特徴情報として扱う。

④確信度により特徴情報の足切り(問題 d の解決のため)

ここでは、各特許で確信度の低い特徴情報を削除することでごみを排除する。

⑤辞書登録を利用して特徴情報を削除(問題 d の解決のため)

「装置」や「方法」などのように特許文書の中には一般的によく出るが単独では意味を持たない単語があるため、これらを辞書に登録しておき削除する。

3　実験

２．３節で示した特徴情報抽出結果の修正処理の有効性を検証するために、修正処理がある場合とない場合の分類結果の比較を行った。

３．１　評価方法

今回はハイブリッド電気自動車関連の特許９１４件を対象にして実験を行った。表１に発明対象の基本分類での分類結果を示す。

表２に、２．３節で述べた修正が起こった回数と修正の正解率を示す。修正の正当性は、１件１件の特許の内容を正確に把握した上でどう分類されるべきかを明確にし、その最終的な分類に対して適当な修正がされたかを評価すべきであるが、９１４件もの特許の内容をすべて把握して人手で分類を行う作業は相当にコストがかかる作業であり、そこまでは行えていない。

しかし、個々の修正の内容については、人の知識によって妥当性を評価できるため、人の主観による評価から修正の正解率を算出した。

ただし、２．３節の④の修正については、どの特徴情報が残るべきなのかを評価すべきものなので、今回の実験対象からはずした。また、⑤は一般的に意味のないものだけを削除しているため、これも実験対象からはずした。

表１．発明対象の基本分類の分類結果

修正なし		妥当性
ハイブリッド車両	523件	○
動力出力装置	47件	
制御方法	46件	
エンジン	45件	
電動機制御装置	44件	
前後輪駆動車両	31件	○
内燃機関	30件	
電気自動車	26件	○
駆動装置	25件	
車輌	10件	○
パワートレーン	8件	
アイドルストップ車両	6件	○
自動変速機	6件	
モータ制御システム	6件	
電動駆動力アシスト車両	4件	○
回生協調ブレーキ制御装置	4件	
複合駆動システム	4件	
移動体	4件	
複数	4件	
バッテリー制御装置	3件	
トルク変動低減装置	3件	
回生制動トルク	3件	
パワープラント	3件	
発電機	3件	
回転電機	3件	
電池電圧検出装置	3件	
冷却装置	3件	
妥当な分類の件数	600件	

修正あり		妥当性
車両	453件	○
ハイブリッド車	86件	○
電気自動車	85件	○
エンジン	80件	
内燃機関	59件	
動力出力装置	42件	
電動機	34件	
車輌	17件	○
変速機	15件	
複数	14件	
パワートレーン	7件	
回転電機	5件	
回生制動トルク	3件	
妥当な分類の件数	641件	

表２．修正回数と修正の正解率

	修正回数	絶対精度	相対精度
技術分野を利用して抽象度合わせ(3章の①)	1008 回	46%	59%
用途を使って基本分類を修正(3章の②)	62 回	95%	98%
基本分類と目的分類を比較して特徴情報を修正(3章の②)	92 回	51%	99%
基本分類を使って構成要素分類を修正(3章の②)	660 回	71%	73%
部分文字列の一致による特徴情報の分割(3章の③)	38 回	11%	74%

表３．構成要素分類の分類結果

構成要素分類	
制御装置	516 件
制御方法	72 件
ハイブリッド自動車	27 件
駆動装置	23 件
発電機	20 件
発電装置	11 件
制御手段	11 件
異常検出装置	10 件
トルク変動抑制装置	9 件
モータ	9 件
・・・	

表４．目的分類の分類結果

目的分類	
燃費	71 件
エネルギ効率	54 件
ショック	51 件
燃料消費率	27 件
応答性	26 件
運転性	24 件
トルク変動抑制する	21 件
効率低下する	21 件
ドライバビリティ	18 件
駆動力確保する	16 件
・・・	

今回の評価は、２つの基準で正解率を算出した。

１つは、元々のデータがハイブリッド電気自動車に関する特許であることから、発明対象の基本分類は自動車全体を表すもので、構成要素分類は「エンジン」や「発電機」のように自動車の構成要素になると考えて、このようなレベルになるように修正されたものを正解とした。これを絶対精度として算出した。

もう１つは、基本分類が自動車全体を表すことにこだわらず、行われた修正を１つずつ人が見たときに妥当だと思えるかを判断して評価した。例えば、「遊星歯車装置」が発明対象の基本分類として抽出されていたときに、「変速機」を基本分類にして「遊星歯車装置」を構成要素分類にするように修正が起こった場合には、基本分類と構成要素分類の関係が正しくなるように修正されているので正解とした。逆に、「変速機」が発明対象の基本分類の特徴情報として抽出されていた時に、「遊星歯車装置」を基本分類にして「変速機」を構成要素分類にするような場合にはレベルが逆転してしまうために不正解とした。これを相対精度とした。

３．２　実験結果と考察

表１に修正処理を行った場合と行わなかった場合に分けて、発明対象の基本分類の分類結果を件数の多い順に並べた結果と、基本分類として妥当であるかを評価したものを示す。修正を行わなかった場合には、構成要素と思われる分類がかなり混ざっているが、修正を行うことにより、発明対象の基本分類として妥当ではない分類の数が減るとともに、基本分類として妥当な分類に含まれる件数が増えており、そのほとんどが件数の多い上位３分類に含まれるようになった。また、構成要素分類の特徴情報の変化を見たところ、誤って基本分類として抽出されてしまった特徴情報が構成要素分類に修正されることで、構成要素分類の特徴情報が得られなかった特許文書の件数が、２５９件から１４１件に減少しており、抽出漏れを減少させる意味でも効果が見られた。

これらのことから、特許文書の多観点分類において抽出結果の修正は有効であると考えられる。

表２からかなりの回数の修正が起こったことがわかる。修正の内容を見たところ、「制御装置」や「発電装置」「駆動装置」などが発明対象の基本分類から構成要素分類へ変更されて特徴情報のレベルが適切に修正され、また、初めは用途の観点で抽出されていた「車両」や「電気自動車」などが発明対象の基本分類の特徴情報へ変更されて観点が適切に修正されていた。

また、いくつかの修正で絶対精度と相対精度の差が大きいのがわかる。この多くは、例えば、「エンジンの制御装置」のように「の」によって係るが基本分類と構成要素分類という関係になるとは必ずしも言えないものであった。今回は、このようなものも相対的には正解として評価したため、絶対精度と相対精度に差が出たが、本来であればむしろ、「エンジン」と「制御装置」は、制御という処理を行うものと、その処理対象の関係にあり、新たな「処理対象」という観点の導入が必要であると考えられる。構成要素分類については、基本分類の特徴情報から「の」で名詞に係る係り受け組で全体と構成要素という関係にあることが多いという知見から、「の」を使って基本分類と構成要素の判断を行ったが、今回の実験から処理を行うものと処理対象という関係もかなり多いことが分かったためこの２種類の関係を判別するための方法について検討する必要がある。

表３，表４に構成要素分類と目的分類の観点での分類結果のうち、件数の多い１０分類を示す。構成要素分類としてはレベル合わせ処理の誤りにより「ハイブリッド自動車」という分類ができてしまっているが、それ以外は「制御装置」「駆動装置」「発電機」などの分類ができ、目的分類として「燃費」「ショック」「応答性」などの分類ができており、それぞれの観点に応じた分類が行えることが分かった。

しかし、「制御方法」と「制御手段」や「燃費」と「燃料消費率」のように意味的に近い分類がうまくまとまらない場合があるなどの問題も残っている。

4　今後の課題

今回の実験により、本手法によりある程度の多観点分類が可能であることが分かったが、多観点分類の実用化に向けては、２章で述べたアプローチの各処理について今後以下のような課題に取り組む必要がある。

①特徴情報抽出処理
- ●再現率の向上
- ●抽出精度の向上
- ●特徴情報抽出ルールの整備
- ●発明対象と目的以外の観点の導入

②特徴情報の修正処理
- ●目的分類については、現在はレベルをあわせる仕組みが入っていないので、目的分類についてもレベルをあわせる仕組みを検討

③特徴情報の統一化処理
- ●現在の統一化処理の評価及び統一条件の検討
- ●構成文字一致以外の統一化方法の検討
- ●ラベル付けの方法の検討

また、今回は行えなかったが、人手で特徴情報を抽出した結果を利用して分類結果の精度や再現率の評価も行う必要がある。

5　まとめ

特許情報の多観点分類を実現するための手法の説明を行い、特徴情報抽出結果の修正の効果について実験で示した。最後に、今後取り組むべき課題をまとめた。

実験の結果、特許文書の多観点分類を行うためには、各特許から観点ごとに内容を表す特徴情報を抽出した後に、抽出された特徴情報のレベルを合わせたり、複数の観点間で整合性が取れるように修正したりする処理が有効であることが分かった。これにより、特許文書の多観点分類がある程度可能になった。今後は、多観点分類技術の実用化に向けて４章で挙げた課題を一つ一つ解決していく必要がある。

参考文献

[1]渡部勇，テキストマイニング技術による公知例調査の支援，ロボット学会誌，Vol22 No.3，2004

[2]木村託巳，山田寛康，島津明，WWW 探索支援のための記述意図によるテキスト分類，言語処理学会　第９回年次大会　発表論文集，pp.505－508，2003