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

2008－ＦＩ－ 89 (2)
2008－ＮＬ－183 (2)
2008／1／21

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「何を課題とした特許か(課題)」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、発明対象と課題といった複数の観点で自動的に特徴情報を抽出する手法について検討・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の特徴情報抽出は、人が手作業で行なっても難しい作業であること、多観点分類システムでもある程度の抽出が可能であることが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 extracting features of patent documents, we show promising result on Multi-viewpoint clustering of patent documents.

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]。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、発明の対象や課題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が有効である。しかし、こういった分類を人手で行なうには大変なコストがかかってしまうため、ここまで内容に踏み込んだ分析は十分には実施されていない。

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、特定の観点に基づいた分析を行ないたい場合には適さない。

富士通研究所では、特許文書について、発明の対象と課題といった複数の観点で分類を行う手法（多観点分類）について検討・実験を行ってきた[2]。特許流通促進事業では、専門家が人手によって特許分析を行なった結果を公開しており[3]、これまでは、この分析結果と多観点分類システムによって分類された結果から読み取られる傾向が一致するのを確認することで、多観点分類システムの実用性と妥当性を検証してきた。

本稿では、より精度の高い多観点分類システムの開発のために、人手による正解データの作成と、多観点分類システムによって抽出された結果の定量的な評価を行なった。

2 人手による正解データの作成

１１テーマ、計２０１２件の特許を実際に人が読んで、文書中で発明の対象や課題の観点で

重要だと思われる部分にタグ付けを行なった。

タグ付け作業の詳細について、発明の対象と課題についてそれぞれ説明する。

２．１　発明の対象

発明が何に関するものであるかという観点で特徴を表す部分にタグをつける。発明の対象は、主に、特許文書の中の「発明の名称」や「技術分野」、「請求項」などと言われる部分に書かれている。発明の対象である製品の一般的な名称を表す部分に <基本分類>タグを付与し、その製品の中で特に改良を加えている部分に <構成要素>タグを付与する。<基本分類> <構成要素>ともに、原則として１つの特許文書の中に１つだけタグ付けする。

例１：【発明の名称】<基本分類>車椅子</基本分類>用<構成要素>テーブル取付構造</構成要素>

<基本分類>と<構成要素>の関係は、なるべく同じカテゴリに属する上位・下位関係にする。すなわち、<基本分類>がモノであれば<構成要素>もモノであり、<基本分類> が作用であれば<構成要素>も作用であるとする。

このような方針に基づいて実際にタグ付けを行なったところ、以下のような問題があることがわかった。

ⅰ．特許文書によっては２つ以上タグ付けしたくなるようなこともある。

例えば、<基本分類> <構成要素>ともに、並列に書かれていることがある。

例５：【発明の属する技術分野】本発明は<基本分類>液晶ディスプレイ</基本分類>や<基本分類>プラズマディスプレイ</基本分類>等のデジタル駆動の表示デバイスを用いた表示装置における階調削減技術に関するものである。

ⅱ．<基本分類>と<構成要素>でモノと作用というのが食い違う場合がある。

例えば、プラズマディスプレイの特許文書では、次のように<基本分類>がモノでありながら、<構成要素>が抽象的コトである場合もある。

例６：【発明の名称】<基本分類>プラズマディスプレイ</基本分類>の<構成要素>階調表示方式</構成要素>

ⅲ．基本分類と構成要素のどちらかが見つからない場合がある

これらの問題に対して以下のように対処することにした。

Ⅰ．作業者が必要と認めた場合には、複数のタグを付けてもよい。

Ⅱ．他に相応しい部分が見つからない場合には、基本分類と構成要素でモノとコトで食い違うことを許す。

Ⅲ．基本的に<基本分類><構成要素>ともに整っていることが望ましいが、文書によっては<基本分類>だけ、または<構成要素>だけしか見当たらないことがある。また、<基本分類>か<構成要素>かの判断がつかないこともある。そのような場合、次のような方針で判断する。

①基本分類はあるが構成要素で適当なものが見当たらないときは、基本分類だけでよい。

②構成要素はあるが基本分類が見当たらないときは、その語句が明らかに構成要素であると判断できる場合（その語句が他の特許文書では構成要素になっている場合は、基本分類になるべきものが想定できてその構成要素と考えられる）は、構成要素のタグを付けて、基本分類のタグはなくてもよい。

③基本分類か構成要素か判断がつかない場合は基本分類のタグを付ける。

２．２　課題分類

発明が何を目的・課題とするものであるか、という観点において特徴を表す部分にタグをつける。「要約」や「発明が解決しようとする課題」「発明の効果」などと言われる部分に書かれている。

実際にタグ付けを行なうと以下のような問題があることがわかった。

ⅳ．特許によって課題の抽象度が異なったり、１件の特許の中に抽象度の異なる課題が書かれている。

ⅴ．その特許で主に解決しようとしている課題と、改良によって副次的に解決される課題がある

そこで、課題を<ゴール><直接目的><効果>の３種類に分けることにした。<直接目的> は、その発明が主として目指している改良点で、発

明により直接変化がある具体的な目的を指す。<ゴール>は、複数の直接目的を達成することによって最終的に実現したいと考えているより抽象的な目標を指す。<効果>は、発明の第一の目的ではないが、副次的に得られる効果を指す。

例１３：【課題】【発明の効果】本発明の階調表示方法によれば、従来例よりも著しく<直接目的>動画疑似輪郭の見え方を低減し</直接目的>、<ゴール>動画像の画質を向上させる</ゴール>ことができる。

この発明が主として改良しようとしているのは「動画疑似輪郭の見え方を低減」することである。そして、この改良によって最終的には「動画像の画質を向上させる」ことを実現したいと考えている。そこで、前者を<直接目的>、後者を<ゴール>とする。<直接目的> は１つの特許文書に少なくとも１つは存在する。<ゴール>は１つの特許文書に存在しないこともある。存在する場合は基本的に１つだけである。<効果>は１つの特許文書に存在しないこともあるし、２つ以上存在することもある。

また、以下のような問題もあった。

vi．特許文書の中では同じような表現が、「課題」「発明が解決しようとする課題」「発明の効果」の項に何度も反復して書かれていることが多い。

このような場合、より明確に書かれている方にタグ付けを行い、同程度の表現の場合は先に書かれている方にタグ付けすることにした。

２．３　タグ付け結果の評価指標

タグ付け作業を複数人で別々に行なってみたところ、作業者によってタグを付ける場所や個数が思った以上に異なることがわかった。これは、背景知識の差などによって特許文書の特徴を捉えることが難しい場合があることを表している。そこで、特徴情報の抽出の難しさを定量化するために、作業者の評価を行なうことにした。ある作業者を評価するときには他の残りの作業者を正解として作業者の評価を行なった。評価値としては、以下のものを用いた。

２．３．１　一致率

タグ付けを行なった結果、発明の対象は単一の単語や複合語で表される場合が多いのに対し、課題については、「テーブルを容易かつ確実に取り付ける」のように複数の文節にまたがることが多かった。また、直接目的だけを見ても１件の特許から「高輝度」「高発光効率」「安定な放電」のように複数の課題にタグ付けされたり、「操舵に違和感が発生しない」「操舵上の違和感を防止ないし低減」「好ましい操舵感覚」のように同じような内容であるが、複数の場所にタグ付けされる場合が多く見られた。

そこで、発明の対象と課題を分けて一致率を以下のように定めた。

・発明の対象

評価される作業者と、他の作業者の誰か一人とがタグ付けを行なった個数と文字列が一致した特許の件数の割合（ただし、文字列は後方一致を一致とみなす）。

・課題

タグ付けを行なった箇所が１つでも部分一致した特許の件数の割合。

課題の一致率は、<ゴール><直接目的><効果>を混ぜてしまうと、その作業者が最も重要と考えた課題が一致しているかどうかがわからなくなるため、直接目的のみを用いて評価した。

２．３．２　抽出率

人手でタグ付けを行なった場合でも、全ての特許にタグ付けできるとは限らない。特に、構成要素では、人が読んでも判断が難しい例が多く見られた。そこで、それぞれの観点について１つ以上のタグを付けられた件数の割合を抽出率として評価した。

２．３．３　適合率と再現率

２．２節で述べたように、課題については、<ゴール><直接目的><効果>と分けてタグ付けを行なった。しかし、これらの厳密な区別は大変難しく、また、これらは抽象度や重要度の差はあれ、どれもその特許の課題を表している。そこで、課題を表すような表現がうまく取れているかを見るために、この３種類を混ぜて評価を行なってみた。

特許１件ごとに文字列の部分一致によって似ている課題をまとめた上で、その特許から抽出できた課題の数（評価される作業者がタグ付けした課題）と正解の数（その特許で他の作業者がタグ付けした課題）と抽出したものと正解が一致した数（部分一致を一致とみなした）をカウントし、テーマごとに適合率と再現率を計算した。

表1. タグ付け結果の評価

テーマ名	タグ付け特許件数	タグ付け作業者数	基本分類の平均一致率	基本分類の平均抽出率	構成要素の平均一致率	構成要素の平均抽出率	直接目的の平均一致率	直接目的の平均抽出率
車いす	123	2	94.3	99.2	76.2	99.2	78.4	99.6
プラズマディスプレイの駆動技術	265	3	84.3	97.6	79.0	87.5	74.9	98.7
自律歩行技術	215	3	86.6	96.0	82.7	79.8	79.6	98.0
ハイブリッド自動車の制御技術	200	4	93.7	92.8	92.1	97.5	93.0	99.6
固体高分子型燃料電池	151	4	94.7	99.3	67.5	59.4	92.6	99.3
電子透かし技術	187	4	69.0	97.7	51.4	33.6	86.6	98.8
非接触型 IC カード	171	3	76.1	97.3	58.5	60.4	78.8	97.1
チャイルドシート	133	4	84.4	99.4	65.8	48.7	91.4	96.8
ゴルフクラブ	220	4	94.5	99.3	66.4	54.4	90.2	96.4
CRM・知的財産管理システム	143	4	88.5	99.0	20.4	14.3	87.3	99.1
バリアフリー住宅	204	4	62.0	81.0	63.5	58.1	91.0	98.5

2．4　評価結果

一致率と抽出率について作業者一人ひとりを評価し、最後に作業者全員の平均を算出した結果を表１に示す。

基本分類と直接目的に比べて、構成要素の一致率と抽出率が低い。特にCRM・知的財産管理システムや電子透かし技術のようなソフトウェア関連特許では、構成要素といっても実際に形のあるものではないため、判別が困難な場合が多かったようである。

また、ゴルフクラブについても構成要素の一致率と抽出率が低い。作業者ごとにタグ付けした結果を見てみると、「ゴルフクラブ」を基本分類として「ヘッド」を詳細分類にした例と、「ゴルフクラブヘッド」を基本分類にした例、構成要素として「シャフト」「ドライバー」にタグを付けた例など、作業者の考え方や知識の違いと思われるものが多かった。

表２に適合率と再現率の平均を示す。適合率は、７０～８５％程度で、人が文書を読んで判断しているだけにごみはそれほど多くないようであるが、人が読んでも適合率がこの程度であるのは、技術文書としての難しさであると思われる。再現率は、５５～７５％程度であり、データや人によってばらつきがあり、人手を割いても漏れなく抽出することは難しいようである。

3　多観点分類システムの評価

多観点分類システムでは、以下のような手法で特徴情報を抽出する。

3．1　特徴情報の抽出

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

表2. 人手でタグ付けした課題の適合率と再現率

テーマ名	平均タグ付け個数	平均適合率	平均再現率
車いす	407.0	71.0	71.2
プラズマディスプレイの駆動技術	679.3	78.4	56.5
自律歩行技術	569.7	78.4	57.6
ハイブリッド自動車の制御技術	359.8	88.2	80.1
固体高分子型燃料電池	391.8	87.9	79.2
電子透かし技術	419.8	84.1	70.6
非接触型 IC カード	504.7	79.5	57.9
チャイルドシート	407.3	86.4	76.0
ゴルフクラブ	794.5	84.2	69.4
CRM・知的財産管理システム	437.0	85.8	71.2
バリアフリー住宅	636.8	85.9	75.6

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、係り先を構成要素を表す特徴情報として抽出する。

３．２　特徴情報の修正

３．１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
b. 観点にそぐわない特徴情報が抽出される
c. 複合語から基本分類と構成要素とを切り分けることができない
d. 抽出された特徴情報にごみが多い

これらの問題に対して、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する手法を提案している[2]。

①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)

「用途」と「基本分類」、「基本分類」と「課題」、「基本分類」と「構成要素」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。

例えば、基本分類と課題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。

③複合語からなる特徴情報を分割(問題 c の解決のため)

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他の特許の基本分類と部分一致するものを探すことにより分割を行なう。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」を構成要素の特徴情報とする。

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。

３．３　多観点分類システムによる抽出の評価

２．３節で述べたタグ付け結果の評価手法と同じ指標で多観点分類システムによる抽出結果を評価した。ただし、多観点分類システムを評価する場合には、作業者全員を正解として用いているため、作業者を評価するときよりも一人分正解データが多いことになる。一致率と抽出率を表３に示す。

全体的に、一致率と抽出率は、かなり作業者を評価したときと近い値になっている。

電子透かし技術の構成要素の一致率がかなり低くなっている。電子透かしの構成要素

表3. 多観点分類システムの評価

テーマ名	基本分類一致率	基本分類抽出率	構成要素一致率	構成要素抽出率	課題一致率	課題抽出率
車いす	95.1	100.0	84.9	71.6	82.5	98.6
プラズマディスプレイの駆動技術	76.2	99.4	66.2	54.0	94.3	96.0
自律歩行技術	72.1	100.0	38.8	55.9	74.6	92.1
ハイブリッド自動車の制御技術	64.0	99.8	87.4	75.6	87.9	98.9
固体高分子型燃料電池	83.4	98.2	48.5	44.2	89.3	99.3
電子透かし技術	84.3	99.4	12.2	33.1	77.4	96.9
非接触型 IC カード	52.9	99.4	38.0	44.2	70.7	95.2
チャイルドシート	70.7	100.0	51.8	63.2	85.9	97.7
ゴルフクラブ	94.5	99.0	59.7	39.2	93.5	95.9
CRM・知的財産管理システム	71.3	99.6	55.0	33.3	73.8	90.8
バリアフリー住宅	85.1	98.6	59.5	53.7	86.4	97.0

について作業者ごとの一致率と抽出率を見てみると、抽出率の最も高い作業者が 98.4%で一致率が 13.6%、抽出率の最も低い作業者が 2.7%で一致率が 100%と作業者ごとのばらつきが大きく、元々難しいテーマであったことが伺える。CRM・知的財産管理システムでは、作業者ごとのばらつきはそれほど大きくないが、抽出率、一致率ともに低い。ソフトウェア関連特許では構成要素といっても具体的な形を持ったものではないので、人が読んでも抽出は難しいようである。

4　考察

一致率を見ると多観点分類システムでの抽出でもかなり人と近い品質で特徴情報の抽出が可能であると考えられる。これまで、全体傾向で多観点分類システムの妥当性が高いと評価してきたが、一件一件からの抽出結果としてもこれまでの評価を裏付ける結果が得られた。

ソフトウェア関連特許の構成要素としては、特定の処理部で行なう処理に特徴があるなどは考えられるが、よほど内容を理解していないと特徴を捉えることは難しいと思われる。処理するデータの種類など他の観点によって分析を行う方が、実用性が高いかもしれない。自律歩行技術や非接触型 IC カードの構成要素についても多観点分類システムでは余りいい結果は得られなかった。構成要素については、その特許が特に改良を加えた点を判別しなければならず、人手でタグ付けした結果を見ても他の観点に比べて全体的に一致率、抽出率ともに低いので、難しい問題だと考えられる。

課題として多観点分類システムで抽出された特徴情報は、課題として意味のあるものが大部分であるが、中には「位置簡単にする」や「兼用」のように課題として見るには情報が少な過ぎる特徴情報も見られた。課題は単一の単語や複合語では意味がわからないことも多く、抽出の仕方について検討が必要である。ただし、抽出された特徴情報をまとめて分類として扱うためには、長い特徴情報を抽出すればよいものではなく、必要最小限の情報でまとめやすい特徴情報の抽出が必要である。

今回の課題の評価では、部分一致を一致とみなして正解の判定を行なったが、多観点分類システムでは、係り受け解析の結果を元に特徴情報の抽出を行なっているため、元文書に直接タグ付けされているものと比較すると、副詞句を省略したり、活用形が変わっているなどのために、単純には一致しない場合も多い。そのため、正解データとの一致を見るときにも、タグ付けされた部分をいったん係り受け解析して多観点分類システムと同じような形に合わせるなどの工夫が必要であると思われる。また、多観点分類システムでは、同じような課題であっても、抽出ルールに当てはまる複数の表現を抽出するようになっているのに対し、人手によるタグ付けでは、はっきり書かれている部分を代表してタグ付けするとしていた。このため、タグ付け結果と多観点分類システムで課題の個数がかなり違うテーマもあった。同じような意味でも全ての課題にタグを付けるように正解データを見直すとか、多観点分類システムから抽出されたものをある程度意味的にまとめてから評価するなどが必要だと思われる。

また、課題については、人手では「ゴール」「直接目的」「効果」という３種類に分けてタグ付けを行なったが、多観点分類システムでは、１種類として扱っている。詳細な特許分析を行なうためにはこれらを判別する方法についても考えていく必要がある。

5　まとめ

特許文書の多観点分類のための特徴情報の抽出について、人が文書を読んで抽出した結果の評価を行なった。複数の人で抽出した結果を比較することで特徴情報の抽出の難しさが示された。また、人手による抽出結果を使って多観点分類システムで抽出した結果を評価することで、人に近い品質で特徴情報の抽出が可能であることが示された。

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，雑誌 Fujitsu, Vol56 No.4, 2005
http://img.jp.fujitsu.com/downloads/jp/jmag/vol56-4/paper17.pdf

[2]田中一成，特許文書の多観点分類について，情報処理学会　第１６１回　自然言語処理研究会，2004/5/13～14　p.p.69－74

[3] 特許流通支援チャート，
http://www.ryutu.inpit.go.jp/chart/tokumapf.htm