特許マップ作成のための文書処理技術

東京工業大学精密工学研究所客員教授
株式会社日立製作所　中央研究所
岩山　真

PROFILE
1992年(株)日立製作所入社。文書検索、自然言語処理等の研究に従事。また、NTCIRにおいて特許検索用テストコレクションの作成に携わる。2009年度より特許版産業日本語委員会委員。

makoto.iwayama.nw@hitachi.com　TEL 042-323-1111

1 はじめに

大量の特許情報を分析するためのツールとして特許マップがある。本 Year Book でも何度か取り上げられているし、実務でも広く活用されている。しかし、特許マップの作成には多大なコストがかかるため、その自動化もしくは半自動化が望まれている。

特許マップを入力情報という観点で分類すると、書誌情報（出願人、発明者、出願日、分類コードなど）から作成する場合と、明細書のテキストから作成する場合とに分けられる。前者は、自動化になじみやすく、自動作成したマップも解釈しやすい。それに対し、後者は、自動作成してもなかなか思い通りのマップが得られず、フラストレーションがたまるケースも多いと聞く。これはひとえに、明細書の文書処理の難しさのせいである。

本稿では、明細書から特許マップを自動作成する際に使われる文書処理技術を紹介する。技術の特徴や長所、短所を知ることで、自動作成ソフトウェアを効果的に使うヒントを提供できたら幸いである。

2 紹介技術の位置付け

特許マップの自動作成に限らず、データマイニングやテキストマイニングといったデータ分析法では、（1）まず、対象データを収集し、（2）それらを計算機で扱える形に変換し、（3）最後に、プログラムによりデータの集計、分類、データからの規則・パターン発見などを行う。ここで注意したいのは、前段のステップほど精度が要求されるという点である。対象データが不適切であれば、いくら分析しても意味がないし、変換時に情報が落ちてしまうと次の分析もうまくいかない。（1）のデータ収集は、検索に等しいため本稿では対象外とする。また、（3）の分析手法も既に多くの書籍があるために割愛する。本稿では残る（2）を扱う。

書誌情報から特許マップを作成する場合、（2）はほとんど必要ない。書誌情報は、既に表形式のデータで表現されており、各々の書誌情報の定義も明確なため（名前や日付や分類コードなど）、そのまま計算機で扱えるからである。

一方、明細書のテキストの場合は、発明内容を損なわずに計算機が扱える形式（具体的には表形式と考えてよい）に変換せねばならない。現状の文書処理技術では、完璧な変換は難しいため、ユーザの分析意図との齟齬も生まれやすくなる。

本稿では、比較的良く使われている二つの変換法を紹介する。一番目は、明細書の重要単語を抽出する方法であり、二番目は、発明の課題、発明で使われる技術、その効果など、発明の様々な観点を明細書から直接抽出する方法である。これらはいずれも、明細書から言葉を拾うため、出願人による表現の違いが生じてしまう。表現の違いを吸収するためには、同義語や上位・下位語の辞書が必要となる。本稿では、これらの言語資源を自動収集する技術についても紹介する。

3 重要単語を抽出する技術

文書を計算機で扱う際は、文書をそこで使われている単語の集合で表現することが多い。これは、bag-of-words (BOW) による表現と呼ばれている。BOW による表現は単語の出現順を全く考慮しないが、文書自動分類等では BOW 表現でもそれなりの精度を得ることができる。また、各文書を強く特徴付ける単語ほど重みを大きくするといったように、単語に重みを付与することも多い。

文書を重み付きの単語集合で表現すると、図１のような特許マップを自動生成することができる。

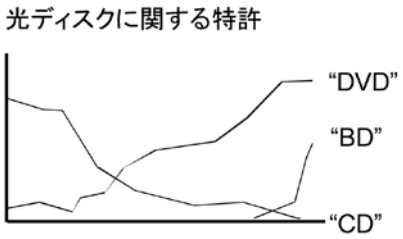

図１　単語頻度の時系列推移

これは、光ディスクに関する特許で使われている重要単語の頻度推移を調べた仮想的なマップである。明細書から抽出した単語を使うことにより、ブルーレイディスクのような技術の萌芽（分類コードがまだ定義されていないような技術）を捉えることができる。また、図２は、単語分布の類似度により、似ている文書をグルーピングして表示する文書クラスタリングの例である。

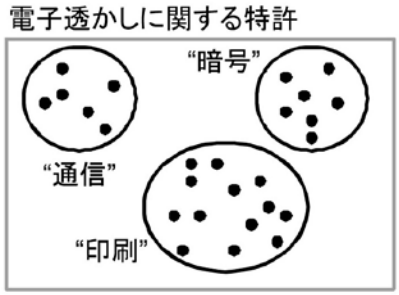

図２　文書のクラスタリング

結果の表示には様々な工夫がなされることが多い。ここでも、既存の分類コードの有無によらず、内容の似ている特許をまとめることができる。

文書を重み付き単語の集合に変換するには、（１）文書を単語に分割し、（２）不要語を除き、（３）最後に単語に重みを付ける。以下、それぞれについて説明する。

（１）形態素解析

文書を単語に分割するには、形態素解析と呼ばれる処理を行う。ここで難しいのは、数多ある分割の可能性の中から一番尤もらしいものを選ぶことである。

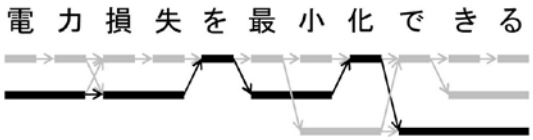

図３　形態素解析の様子

図３では、分割の可能性を、経路のパスで表現している。コスト最小化法と呼ばれる手法では、各パスの確率を推定し、確率が最大のパスを選ぶ。パスの確率は、各単語の生起確率（単語の現れやすさ）と、単語（もしくは品詞）間の推移確率（二つの単語（もしくは品詞）のつながりやすさ）から構成的に計算する。生起確率と推移確率はコーパス（大量の文書集合）から推定する。

よって、明細書の形態素解析の精度を上げるには、単語辞書を充実させることに加え、上記の確率値を明細書自身から推定することも必要となる。

（２）不要語除去

単語に分割した後は、明らかに不要な単語を除去する。例えば、助詞などの付属語を除去することが多い。また、「発明」や「請求項」など、どの明細書にも現れる単語を除去することも多い。

（３）単語重み付け

単語の重み付けは、概念検索でも良く使われている TF-IDF 法と呼ばれる手法で行うことが多い。つまり、その文書内で頻繁に現れる単語ほどその文書を強く特徴付けていると捉えて重みを高くする（TF）。一方、どのような文書にも現れる語の重みを低くする（逆を言うと、他の文書に現れにくい語の重みを高くする）（IDF）。

明細書で注意せねばならないのは TF の扱いである。特に請求項においては、解釈の曖昧性を避けるために、「これ」「それ」といった代名詞を用いたり指示先を省略したりせず、「前記分析装置」というように指示したい語を重複して書く傾向がある。そのため、例の場合は「分析装置」という句が必要以上に現れ、結果として TF が大きくなり、構成単語の重みも大きくなってしまう。請求項では TF を使わない方が良いという研究結果もある [6]。

請求項の構造を重みに反映させる方法もある [10]。「X において、Y を特徴とする Z」というパターンでは、Y に現れる単語の重みをより大きくすることで発明の特徴を際立たせる。

4 観点を抽出する技術

明細書全体を単語の集合で表現すると、発明の様々な観点（課題、従来技術、本技術、効果など）が入り混じってしまい、集合全体が漠然としたものになりやすい。そのため、前述の図２のようなクラスタリングを行っても、システムがどのような観点で分けたのかが判りにくくなる。また、自動的に集まったグループそれぞれが何を意味しているのかも読み取り難い。

そこで、文書全体から単語を抽出するのではなく、【発明の効果】などの特定の段落から単語を抽出することで、観点別に単語集合を作ることも試みられている。ただし、【発明の効果】に正しく「効果」が書かれるとも限らない。そこで、本節では、発明の「課題」「技術」「効果」などの観点を明細書のテキストからピンポイントで抽出する技術を紹介する。

例えば、

従来の電源トランスにおいては、…せねばならず、電力損失の最小値が大きくなるという課題があった。

それに対し、…冷却速度の制御により、電力損失を最小化できる。

という明細書内の文章から以下の観点を抽出する。

分野	技術	効果	
		属性	値
電源トランス	冷却速度の制御	電力損失	最小化

単純な単語の集合表現に比べ、より明確に発明内容が表現できていることがわかる。

このような観点表現を用いると、図４のような特許マップを作成することができる。

	被処理物質の前処理	流体の処理操作	処理条件の制御
環境負荷低減	文献1 文献7		
高効率	文献10	文献13	文献3 文献4
高品質	文献5 文献11 文献18	文献2 文献21	
高機能			文献6

図４　多観点マップ

これは、発明の技術とその効果という二つの観点で特許集合を整理したものである。F タームを用いることで同様のマップは作成できるが、ここでは、F タームが整備されていない分野や、F タームではまだ定義されていない技術についてもマップ化することができる。

発明の観点を自動的に抽出するには、（１）手掛かりパターンを用いる方法と、（２）機械学習による方法、の二つがある。以下ではそれぞれについて説明する。

（１）手掛かりパターンを用いる方法

例えば、発明の効果は、「X が Y できる。」「X の Y が可能になる。」といった文章に現れやすい。そこで、このような現れ方のパターン（手掛かりパターンと呼ぶ）を幾つか用意しておき、それらと明細書とを照合して発明の効果、さらにはその属性 X と値 Y を抽出する。例えば、「電力損失が最小化できる。」という文は「X が Y できる。」という手掛かりパターンと照合し、X に相当する「電力損失」が効果の＜属性＞に、Y に相当する「最小化」が効果の＜値＞になる。

手掛かりパターンを使う際に注意すべき点は、漏れとノイズの対策である。全ての現れ方をパターン化することは容易ではない。漏れを無くすためにパターンをいた

ずらに増やすと、パターンによるノイズも増えてしまう。

例えば、前述の「XのYが可能になる。」というパターンは、「カートリッジの書き換えが可能になる。」という文には適切だが、「複数回の書き換えが可能になる。」という文には不適切である。一般に、精度が１００%に近いパターンはそう多くないし、このような高精度のパターンのみでは網羅性が担保できない。

そこで、パターンを適用する際に何らかのノイズ除去を施す必要がある。良く使われる方法は、複数のパターンの合わせ技で決める方法である。上記の「カートリッジの書き換えが可能になる。」は「XがYできる。」というパターンでも、「XのYが可能になる。」というパターンでも出現する。一方、「複数回の書き換えが可能になる。」の場合は、「複数回が書き換えできる。」とは言えない。

（2）機械学習による方法

手掛かりパターンを用いて観点を抽出する方法では、あらかじめ何らかの方法でパターンを準備し、かつ、パターンのノイズ対策を行う必要があった。

それに対し、正解例から自動的に観点の抽出器を学習する方法がある。正解例とは、既に人手で観点が抽出されている事例のことで、訓練データとも呼ばれる。訓練データは人手で作成せねばならないのだが、前述のパターンの作成や、ノイズ対策に比べれば、高度な知識やノウハウを必要としない。また、訓練データは一度作成してしまえば、手法が変わっても再利用できる。

本稿では、系列ラベリングという手法で観点抽出を行う方法について説明する。図５に系列ラベリングが行われる様子を示す。

図５　系列ラベリングの様子

ここでは、「最小」という単語にどのような観点を付与すべきかを、その前後２単語の情報（文脈情報）から決めている。各単語の情報としては、その表記、品詞、に始まり様々な情報（図中では特徴量と表記）を使うことができる。明細書特有の情報については後述する。系列ラベリングでは、図中の太線で囲われた部分の情報から「最小」の観点を決めることになる。

太線内がどのようなパターンの時どのような観点になりやすいかは、訓練データにより学習する。学習の方法としては、SVM(Support Vector Machine) やCRF(Conditional Random Field) といった手法を使うことができる。これらの学習手法については、各種書籍（例えば [8]）を参照してほしい。

機械学習を用いる利点は、人手によるチューニング無しに、様々なタイプの情報を考慮することができる点にある。最適な設定は機械学習アルゴリズムが訓練データから自動的に決定してくれるので、設計者は思いつく情報を比較的自由に取り込むことができる。

明細書に特化した情報としては、以下のものが利用されている [2][3][5]。

- どの見出しに現れたか：例えば、発明の効果は【発明の効果】という見出しに現れやすい。
- 見出し中の位置：従来技術に対し、本発明の特徴は各見出しでも後半に現れやすい。
- 手掛かり句や意味ラベルの出現有無：「できる」「可能になる」といった手掛かり語の出現有無は発明の効果という観点の強い要因になる。また、「速度」「温度」のように必ず属性になる語や、「大きい」「高い」のように必ず値になる語（これらは意味ラベルと呼ばれる）も重要である。手掛かり語や意味ラベルは人手で収集することが多い。

その他、明細書に特化した方法として、前後数単語の情報に加え、係り元や係り先の単語も図５の文脈に加えることも多い。請求項内では、単語間の修飾関係が遠くなりやすいためである。例えば、請求項では「X と、Y と、Z から構成される」といった並列構造が現れやすく、この場合、「X」と「構成」の物理的な距離が遠く

なってしまい、「X」の観点を決める際に、「構成」という有力な語の情報が使えなくなってしまう。そこで、係り受け関係を使い、「X」でもその係り先の「構成」を文脈に入れることで観点抽出の精度向上を試みる。

5 言語資源を収集する技術

これまでに説明した方法は、いずれも明細書内の単語をそのまま抽出していた。そのために、F タームなどが定義されていない分野でも、新技術や発明の観点などを抽出することができた。しかし、同じ概念でも出願人が異なれば異なる書き方をするため、これらを同一化しないと正確な集計や分析は行えない。明細書から生成した特許マップの大きな弱点は、概念の統制にある。例えば、図 1 のマップを説明した際、ブルーレイディスクのような新技術を捉えることができるという利点を強調したが、ブルーレイディスクという概念は「BD」と略号で書かれたり、「Blu-ray Disc」と英語表記されたり、固有名称ではなく「青紫色半導体レーザを使用する光ディスク」と書かれたりする。これらを同一化（データクレンジングに相当する処理）しないと正確なマップは作れない。

一般には、同義語辞書や上位・下位語辞書を使って、同じ概念に統一するのだが、これらの辞書の作成や保守にはコストがかかる。そこで、本節では、同義語、上位・下位語を明細書から自動的に収集する技術について説明する。

（1）上位・下位語の自動収集

上位・下位語の抽出には、観点抽出と同様に手掛かりパターンを用いることが多い。例えば、論文 [9] では、「A などの B」という手掛かりパターンを用いて、「フロッピーディスクなどの磁気ディスク」という文章から、「磁気ディスク」の下位語が「フロッピーディスク」であることを抽出している。ただし、ここでも、上位・下位関係以外での一致を防ぐために、観点抽出の場合と同様の手法でノイズ対策を行う。

近年は、このような語と語との関係を雪だるま式に自動抽出する試みも多い。この手法はブートストラップ法と呼ばれる [4]。図 6 にブートストラップ法のイメージ図を示す。

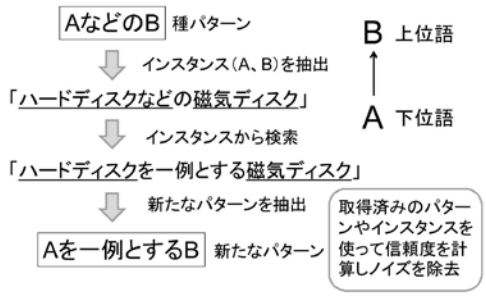

図 6　ブートストラップ法のイメージ図

ブートストラップ法では、まず種となるパターンを入力する。例では、「A などの B」を入力している。次に、このパターンで出現する A と B のペア（インスタンスと呼ぶ）をコーパスから収集する。例えば、「フロッピーディスクなどの磁気ディスク」という表現がコーパス内で多く現れる場合、「ハードディスク」と「磁気ディスク」がインスタンスとして抽出される。今度は、これらのインスタンスが近接して現れる別の場所をコーパス内で検索する。その結果、例えば、「ハードディスクを一例とする磁気ディスク」という表現が大量に検索出来たとすると、「A を一例とする B」という新たなパターンが取得できる。このサイクルを繰り返すことによって、手掛かりパターンと上位・下位語の双方を増やしていくことができる。ブートストラップ法により、観点抽出のための手掛かりパターンを収集することもできる。興味のある読者は、論文 [7] を参照されたい。

（2）同義語の自動収集

上位・下位語は、同じ明細書に近接して現れることが多いため、手掛かりパターンで収集することができた。一方、同義関係にある語は、同じ明細書内で同時には現れにくい。例えば、「計算機」と「コンピュータ」を同じ明細書で使い分けることは少ない。また、使ったにせよ、手掛かりパターンのような決まったパターンで使うことはまずない。

そこで、同義語は、その出現パターンの類似性で抽出することが多い [1]。図7にイメージ図を示す。

パターン	計算機	コンピュータ	キーボード
＊を起動する	10	9	0
＊を接続する	8	4	12
＊の管理	3	5	0
サーバー＊	1	3	0

図7　同義語抽出のイメージ図

例えば、「計算機」は「＊を起動する」「＊を接続する」「サーバ＊」というパターンで現れやすい。同様に、「コンピュータ」もこれらのパターンで現れやすい。一方、「キーボード」は「＊を接続する」というパターンでは頻繁に使われるが、「＊を起動する」というパターンではほとんど使われない。よって、各出現パターンを要素としたベクトルを作成し、このベクトル間の類似度（例えば余弦）を計算すれば、単語の意味の近さが計算できる。例の場合、「計算機」と「コンピュータ」は意味的に近いが、「計算機」と「キーボード」は関連してはいるものの、意味的には別のものであると判定することができる。本手法は、パターンの種類、ベクトル要素の重み付け、更にベクトル間の類似度の定義に関して、様々な研究がおこなわれている。近年では、機械学習によりこれらのチューニングを自動的に行う研究もある [11]。

6 おわりに

本稿では、特許マイニングのための文書処理技術として、明細書を構造化する際に使われる技術を紹介した。これらの中で、観点を自動抽出する技術については、国立情報学研究所が主催する NTCIR 特許マイニングサブタスクで実際に評価されている。同時に訓練データも研究用途で公開されている。まだまだ、訓練データの量は少ないものの、研究環境は整いつつある。今後は、観点自動抽出の更なる精度向上に加え、抽出した観点の同一化についても研究を進めていく必要がある。本稿では、そのための要素技術として、上位・下位語や同義語を自動的に収集する技術についても紹介した。

[参考文献]

[1] Hindle, “Noun classification from predicate-argument structure”, ACL-90, 1990.

[2] Nanba, Kondo, Takezawa, “Hiroshima City University at NTCIR-8 Patent Mining Task”, NTCIR-8, 2010.

[3] Nishiyama, Tsuboi, Unno, Takeuchi, “Feature-Rich Information Extraction for the Technical Trend-Map Creation, NTCIR-8, 2010.

[4] Pantel, “Espresso: Leveraging generic patterns for automatically harvesting semantic relations”, COLING/ACL-06, 2006.

[5] Sato, Iwayama, Experiments for NTCIR-8 Technical Trend Map Creation Subtask at Hitachi, NTCIR-8, 2010.

[6] 岩山，藤井，神門，丸川，「特許検索の諸相」，言語処理学会第９回年次大会，2003.

[7] 坂地，野中，酒井，増山，「Cross-Bootstrapping: 特許文書からの課題・効果表現対の自動抽出手法」，電子情報通信学会論文誌 D, Vol.J93-D, No.6, pp.742-755, 2010.

[8] 高村，「言語処理のための機械学習入門」，コロナ社，2010.

[9] 難波，奥村，新森，谷川，鈴木，「特許データベースからのシソーラスの自動構築」，言語処理学会第１３回年次大会，2007.

[10] 間瀬，辻，絹川，石原，「特許テーマ分類方式の提案とその評価実験」，情報処理学会論文誌，Vol.39, No.7, pp.2207-2216, 1998.

[11] 森本，柳井，岩山，「文脈類似度と表記類似度を用いた教師あり同義語抽出」，言語処理学会第１６回年次大会，2010.