特許情報からの文章のベクトル化を用いた3D有向グラフによる発想

2020032　平井遥斗　　情報システム工学科　指導教員 奥原浩之

要約

近年人工知能(AI)などの情報処理技術は目覚ましく発展しており，記憶や認識，データ収集やその処理などは人間を凌駕しつつある．特許情報は過去の情報をアーカイブしたいわば発明の保管庫的なデータであり，それを活用することで経営戦略・技術的発展等広く社会に役立てることができる．しかし，現状の特許プラットフォームは人手で少数の特許事例を調べるのには必要充分であるが，ビッグデータとして特許全体の分析を行いたい場合には整理されているとはいいがたい．

キーワード： 自然言語処理，業務推進系システム，特許情報処理，テキストマイニング

1 はじめに

2 特許情報処理と発想支援

2.1 発想支援と表現手法

心理学者ギルフォードは，頭の動きを,「認知」,「記憶」,「発散的思考」,「収束的思考」,「評価」の5つに分けている．発想支援システムには,「収束的思考」を支援するための研究と「発散的思考」を支援するための研究がある，これらは個人の発想支援をするための研究と，複数人で発想支援するための研究とに分類できる．

「発散的思考」さまざまな方向に指向が動くことによって，いろいろな発想を生み出すことといえる．試行錯誤的な思考が行われるときには，必ず働くともいえるだろう．つまり，与えられた条件から，多種多様な発想を生み出す思考である．発散的思考は，問題把握の時は事実を，問題解決の時はアイディアを出すのに用いられる

「収束的思考」正しい答えをもたらす働き，つまり与えられた条件から唯一の解答を導き出す思考を言う．重要な関連語に絞り3D有向グラフとして描画する．

2.2 特許情報処理

公開番号とは、個々の公開特許公報に付与される番号をいう。出願番号と同様に公開年が識別できるような書式で付与される。例えば、「特開平10-123456」。ただし、2000年以降は、「特開2006-123456」の書式。同じく、公表特許公報には「特表平10-123456」、「特表2006-123456」のような公報番号が付与されるが、再公表特許公報には「WO2006/123456」のような国際公開番号がそのまま公報番号として付与される。

国際特許分類（International Patent Classification：IPC）は、特許文献（特許内容を掲載した文献。公開特許公報などが該当する。 ）の国際的な利用の円滑化を目的に作成された世界共通の特許分類です。特許文献の「Int.Cl.」の項に記載されています。2023年9月現在、IPC第8版（2006年1月発効）が最新の分類となっていますが、技術の進展に柔軟に対応するため、適宜改正が行われています。

我が国では独自の特許分類FI(File Index)も導入している．FIはIPCを細分化したものであり，わが国で特許出願が多い分野の細分化がIPCでは十分ではなく，IPCで特許検索を行っても十分な絞り込みができないことが多い．そのため，特許庁はIPCを我が国の技術に合わせて再展開したFIを用意している．

F ターム（File Forming Term）は，特許審査のための先行技術調査を迅速に行うために開発された検索インデックスで，関連先行技術を効率的に絞り込むことを目指している

図1 多目的最適化の定式化

2.3 自然言語処理

自然言語処理とは、人が書いたり話したりする言葉をコンピュータで処理する技術です。人工知能（AI）の研究分野で中核を成す要素技術の一つといえます。　自然言語処理技術は「言語理解」と「言語生成」に大きく二つに分けることができます。「言語理解」は人が書いた文章に対してなんらかの処理をする技術で、メールの自動分類、ウェブ検索などが典型的な応用になります。「言語生成」は、コンピュータに文章を生成させる技術で、文章の要約や機械翻訳などを含みます。　これまでは個別の用途ごとに技術開発が進んできましたが、ChatGPTをはじめとする最近のシステムがこの常識を変えました。高度な「言語理解」と「言語生成」が必要な質問応答もできるようになり、さまざまな作業（タスク）を一つのシステムでこなせるようになっています。　実はChatGPTが世間を騒がせる前から、自然言語処理の研究者の間では技術の急速な進歩に驚きの声が上がっていました。2018年にGoogleが発表した「BERT（バート）」というシステムでは、開発者が少し手を加えるだけでさまざまなタスクに使えるようになりました。それだけでも驚きでしたが、ChatGPTの前身である「GPT-2」や「GPT-3」では、システム自体を変更せずとも、人がシステムにあわせて入力を工夫するだけで多様なタスクの実行が可能になったのです。さらに、ChatGPTは人との対話能力が強化され、人間が人間に頼むような言葉で指示をするだけで、さまざまなタスクができるようになりました。形態素解析とは文を形態素ごとに分解する技術である[3]．自然言語処理の一つでテキストを品詞ごとに分解することである．形態素解析を行うことで取り出したい品詞を絞って分析できる．一般的に助詞や助動詞はよく使われるが，キーワードごとのテキストを分析したときに特徴が見られないと考える．

3 トピックモデル

3.1 スクレイピングによるテキストマイニング

テキストデータは,「定性データ」の代表的なもので，この「定性データ」から付加価値の高い情報を収集することがテキストマイニングの目的である．アイデア発想において人間は自然言語から思考して発想することが一般的である．そこでサイバー空間にあるテキストデータを自然言語処理することを考える．現代社会においてインターネット上の情報量は莫大になっており，今後も増え続けることが予想される．このインターネット上の情報を収集して分析することで発想支援に生かせると考える．発想支援において重要なことはキーワードからより関連度の高い単語をより多く表示させることである．そこで，より良いデータを多く収集するためにインターネットからテキストデータを収集することとする．

今回，GooglePatents複数キーワードのand検索の結果を年代ごとに取得し，そのURLからテキストを抽出しそのテキストに対し自然言語処理を行う．

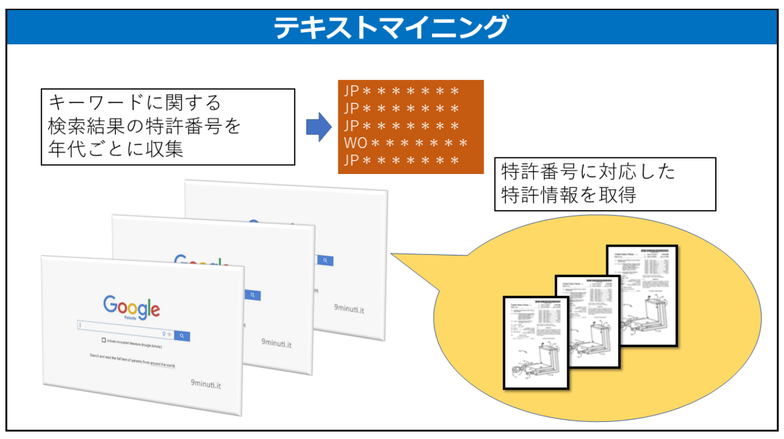

図2 データ収集の流れ

3.2 特許情報のベクトル化とクラスタリング

クラスタリング（clustering）とは、機械学習の1種でデータ間の類似度にもとづいて、データをグループ分けする手法です。この単語は機械学習や統計学の文脈以外でも使われるクラスタリングにはハードとソフトの2種類があります。それぞれのデータが単一のグループに所属するようにグルーピングするものをハードクラスタリング、それぞれが複数のグループに所属できることを許してグルーピングするものをソフトクラスタリングといいます。ことが多いため、これらの分野で用いられる際にはクラスタ分析やデータ・クラスタリングと呼ばれるのが通例となっています。

「群平均法」は、2つのクラスターに属している対象の間のすべての組み合わせの距離を求め、それらの平均値をクラスター間の距離として定める手法です。群平均法は鎖効果を防止できるメリットがあるためウォード法を実行した時に起こってしまう鎖効果（１つのクラスターに対象が１つずつ吸収されていき、新しいクラスターが作られる現象）を未然に防ぐことができます。

「ウォード法」は、凝集型のクラスター分析の手法の1つで「凝集型階層的クラスタリング」とも呼ばれています。ウォード法はすでにあるクラスターの中で、1番距離の近い2つのクラスターが選ばれ、1つのクラスターに結合されていく操作を、目標のクラスター数になるまで続ける方法です。

最短距離法は単連結法とも呼ばれる、2つのクラスタ間で一番近いデータ同士の距離を、クラスタ間の距離として採用する手法です。群平均法と同様に、クラスタを構成する要素同士の距離をすべて求め、その中で一番距離の短い組み合わせを選ぶことでクラスタ間の距離として求めます。この方法のメリットはウォード法などと比較した場合に、計算量が少なくなりますが、同時に外れ値に弱いというデメリットも抱えています。

最長距離法とは、最短距離法とは逆の方法で行う計算手法です。完全連結法と呼ばれることもあります。クラスタを構成している要素同士のすべての距離の中で、最も距離が長いものをクラスタ間の距離として採用するという手法です。

非階層クラスタリングは、階層を作らずにデータをグルーピングしていく手法です。母集団の中で近いデータを収集し、指定された数のクラスタに分類します。この方法では階層クラスタリングとは対照的に、クラスタを形成した後で自由にクラスタを分けることができないため、事前にクラスタ数を指定する必要があります。

k-means法とは、非階層クラスタリングを行うためのアルゴリズムのことです。「指定されたk個のクラスタに、平均（means）を用いて分類していく」という意味が込められています。そんなk-means法は、初めに指定したクラスタの数だけ「重心」をランダムに指定して、その重心をもとにクラスタをグルーピングしていくという手法です。k-means法を活用すれば、データ間の距離を計算する必要がなくなるというメリットがあります。

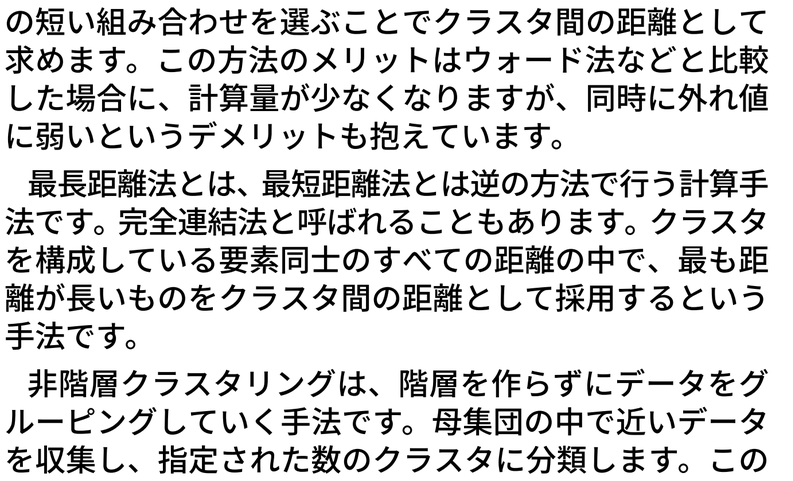

図1 データの収集と活用

3.3 共起関係と共起ネットワーク

ある単語とある単語が同時に出現することを共起するといい，文章において関係深い単語は共起することが多い．共起分析では単語同士のJaccard係数を比較したり，共起関係を持つ単語と単語を線で結んで描かれる共起ネットワークが利用される．文章また単語群に対して共起する単語をネットワークで表した共起ネットワークという．今回，キーワードごとに集めたテキストに対してそれぞれの共起ネットワークを作成する．

4 提案手法

図4 自動献立作成の流れ

5 数値実験並びに考察

図5 実験結果

6 おわりに

参考文献

[1]

[2]

[3]

[4]

[5]