Vol.2014-IFAT-114 No.9
Vol.2014-DD-93 No.9
2014/3/29

特長表現に注目した対象-観点型特許マップの自動生成

岸 桂太[1,a)]　吉岡 真治[1]

概要：多くの特許では、既存の製品の機能向上や好ましくない点の抑制といった効果を目指している。これらの効果に関する記述は、定型的に表現されることが多く、これを特長表現と呼ぶ。本研究では、特長表現から対象と観点を抽出し、それらを用いた特許マップの生成手法を提案する。

1. 研究の背景と目的

今日，各分野において様々な技術が新しく生み出され，蓄積されている．技術情報の多くはテキストデータとして電子化されており，新技術の活用のためには，膨大なデータから必要な情報を迅速に獲得することが求められる．

広く一般に公開される技術情報として公開特許公報，科学技術論文などが存在し，特に特許に関しては，現在日本国内で年間 35 万件近い申請があり，そのうち認可され特許として認められるものだけでも 20 万件に及ぶ．

そのような大量の特許情報を視覚化したものを特許マップと呼び，特許の出願や利用などの特許実務には不可欠なものとなっている．しかし，この特許マップを作成するためには，専門家による作業が必要であり，多大なコストがかかる．

また，注目している技術分野において有効な技術を発見することを支援するために，技術文書から技術の特長を示す表現（特長表現）を抽出し，整理された情報を利用者に提供しようという研究 [1] がある．

本研究では，上記の特長表現に多く含まれる〈対象〉と，その〈対象〉の機能向上等の特長を表す〈観点〉に注目し，特許情報の分析のために作られる特許マップの自動生成を目的とする．

2. 特許と特許マップ

2.1 特許明細書

特許とは，発明の保護及び利用を図るために国が発明者に権利を与えるものであり，公開特許公報によって，出願から 1 年半経過した特許情報が公開される．特許文書は書式がある程度決まっており，出願人はその書式に従った形で発明の詳細（特許明細書）を記述する．特許明細書中には「発明の効果」という項目が存在し，そこには発明によってどのようなことが可能になるかが簡潔に記されていることが多く，従来の技術と比べて有利な点を素早く把握できる．「発明の効果」の記載例を以下に示す．太字部分が，最終的な効果を述べている箇所であり，後に詳細を説明する「特長表現」である．

【発明の名称】ハンドスキャナ
...
【発明の効果】
本発明のハンドスキャナは、ハウジング上部から斜めの光軸を通して 1 次元イメージセンサで走査するため、センサの視野すなわち入力位置を、直接あるいは近傍で常に観測確認できるので、**入力対象の綴じ込み条件や操作方法に応じて左右の側端部を使い分けられる**という利点がある。

2.2 特許マップ

特許マップとは，大量の特許情報を分析するために作られるグラフや表のことである．特に決まった形式はなく，調査対象や目的によって多種多様な形式が存在する．

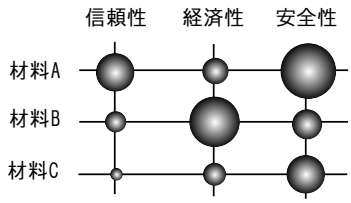


図 1　マトリクスマップ

図 1 はマトリクスマップと呼ばれ，二軸の組み合わせ次第で，該当分野の技術開発の濃淡を多角的に分析することができる．組み合わせの例として，「技術分野-企業」「技術課題-解決手段」などがある．

[1] 北海道大学
Hokkaido University, N14W9, Kita-ku, Sapporo-shi, Hokkaido, 060-0814, Japan
[a)] famksn-fe12@ec.hokudai.ac.jp

2.3 特長表現とその抽出方法

西山ら [1] は，特長表現を，「当該技術の新たな長所を示した表現」と定義している．また，特長表現は，増強クラスと改善クラスの 2 種類に分けられる．増強クラスの特長表現は技術が持つ属性の中で高めるべきものを高めること，または備わっていることが望ましい性質を実現することで，従来技術との差分とすることを示す．対して改善クラスの特長表現は，技術が持つ属性の中で抑えるべきものを抑えること，または備わっていることが望ましくない性質を抑えることで，従来技術との差分とすることを示す．例えば，携帯電話に関する特長表現として

- 通話音質を向上する
- 片手による操作を可能にする

などが増強クラスの例として挙げられ，

- 通話時のノイズを抑制する
- 落水による故障を防止する

などが改善クラスの例として挙げられる．

増強クラスの特長表現と改善クラスの特長表現は共に，特定の用言で表現が終わることが多いとされている．例えば増強クラスの例として挙げた，「通話音質を向上する」という表現は主に「向上する」という用言によって，増強クラスの特長表現であることが分かる．

3. トピックモデル

トピックモデルでは，文書（群）は 1 つ以上の「トピック」から構成されていると仮定される．トピック分布やそれぞれのトピックにおける単語分布を用いて文書を生成する枠組み・方法をトピックモデルと呼ぶ．トピックモデルを用いて，文書がどのようなトピックから構成されているかを推定したり，文書同士の類似度を算出することができる．本研究では，LDA(Latent Dirichlet Allocation)[4] というトピックモデルを使用してクラスタリングを行う．

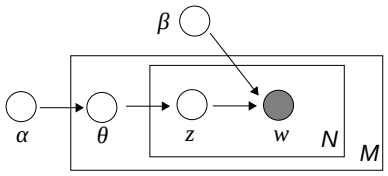


図 2 Latent Dirichlet Allocation

LDA では，文書は次の過程を経て生成される（図 2）．

- 単語数 $N \sim \mathrm{Poisson}(\xi)$ の選択
- トピック分布 $\theta \sim$ (ディリクレ分布 $\mathrm{Dir}(\alpha)$) の選択
- N 個の単語 w_n を生成:
  - トピック $z_n \sim \mathrm{Multinomial}(\theta)$ の選択
  - トピック毎の単語生成確率 β，トピック z_n から単語 w_n を生成する

各パラメータの推定には，本研究ではギブスサンプリングを使用する．ギブスサンプリングとは，マルコフ連鎖モンテカルロ法（MCMC）の一種であり，実際にモデルからサンプリングを行い，モデルのパラメータを推定する手法である．

4. 特長表現に注目した特許マップ

今回は，「特長表現」を用いたマップを作成する．「発明の効果」自体に焦点が当てられるので，該当分野にあまり詳しくない者でも親しみやすいマップになると考えられる．また，マップを形成する際に，特長表現は〈対象〉-〈観点〉のペアに分割される．これにより，向上・改善の対象物が複数存在する技術分野において，特許の分析が行いやすくなる．以下，最初に〈対象〉と〈観点〉の定義を述べた後，実際に自動生成したい特許マップについて説明する．

4.1 対象と観点

まず，〈観点〉の定義について先に説明する．〈観点〉とは，発明や新しい技術によって向上および改善されるものである．例としては，「安定動作」「耐障害性」「操作性」などの，発明や新しい技術によって，より品質を高められた属性や，「生産コスト」「ノイズ」「騒音」などの，従来と比較して改善・抑制がなされた属性が〈観点〉となる．

〈対象〉とは，前述の〈観点〉の対象である．「電子回路」「自動車」「エンジン」など，製品やその部品が〈対象〉となる．上述したような〈対象〉と〈観点〉のペアが 1 つの特長表現の中に存在するとき，そのペアを特許マップの要素として使用する．

4.2 生成したい特許マップ

本研究で自動生成したい特許マップは，特許マップの中でもマトリクスマップと呼ばれるもので，2 つの軸が存在する．今回は，特許文書中の特長表現から〈対象〉と〈観点〉のペアを抜き出し，それらを 2 軸に配置したマトリクスマップの生成を行う．

〈対象〉-〈観点〉ペアの例を挙げると，「低コストでネットワークを構築することができる」という特長表現があったとき，対象が「ネットワーク」，観点が「低コスト」のペアが得られる．発明の対象物を「対象」の軸，対象物のどのような観点が増強/改善されたかを「観点」の軸で表し，マトリクスを形成する（図 3）．

4.3 クラスタリングの必要性

特長表現から抽出された〈対象〉や〈観点〉は表記ゆれや同義語，類義語が多く存在し，それらをそのままマトリクスマップの要素としてしまうと，マップの一覧性が大きく損なわれる．そのため，〈対象〉と〈観点〉それぞれに対しクラスタリングが必要となる．

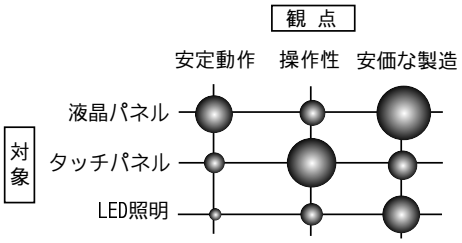


図 3 [対象]-[観点] マップ

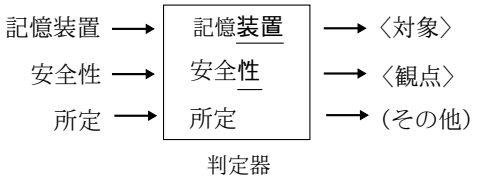
5. 実験

5.1 〈対象〉-〈観点〉ペアの収集

〈対象〉-〈観点〉ペアは特長表現から抜き出すことになるが，〈対象〉と〈観点〉を抽出できない表現はマップ作成に利用できない．よって，今回の実験では特長表現を抽出する西山らの抽出方法とは異なり，〈対象〉と〈観点〉を含む特長表現を取り出す方法を考えた．

特長表現からの〈対象〉-〈観点〉ペアの収集は，精度と再現率を高めるために，図 4 のような 2 段階の処理を行う．

1. 対象語・観点語の判定器の作成



2. ペアの収集

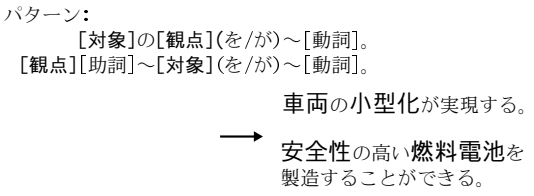


図 4 〈対象〉-〈観点〉ペアの収集方法

以下，1. と 2. それぞれの処理について説明する．

5.1.1 対象語・観点語の判定器を作成

〈対象〉部，〈観点〉部の誤獲得を防ぐため，対象語・観点語の判定器を作成する．

まずは，対象らしい及び観点らしい語の収集を行う．対象らしい及び観点らしい語のリストを作成するために，〈対象〉部と〈観点〉部の抽出精度が高い文パターンを定義し，特許明細書の「発明の効果」セクションを対象に，その文パターンを適用する．

対象語リスト，観点語リストを作成すれば，それらのリストを使用して，単語列から〈対象〉らしい表現，〈観点〉らしい表現を判定する判定器を作成することができる．判定器は単語列の末尾の語に注目し，判定を行う．例えば，観点語リストの語は「安全性」「コスト性」「静音性」など，「～性」で終わる表現が多い．よって，「～性」という表現は，〈観点〉を述べている表現と判断することができる．

5.1.2 ペアの抽出

前ステップで作成された対象語・観点語のリストを使って，実際に〈対象〉と〈観点〉のペアを抽出する．前ステップでは，〈対象〉部と〈観点〉部を正確に獲得するために，再現率は低いが精度の高いパターンを使用した．本ステップでは，カバー率の高い文パターンを定義し，実際に〈対象〉と〈観点〉のペアを抽出する．定義したカバー率の高い文パターンは次の 2 つである．

- 〈対象〉の〈観点〉(を/が)～[動詞]。
- 〈観点〉[助詞]～〈対象〉(を/が)～[動詞]。

上記の 2 つめのパターンには，例えば「**安全性**の高い**燃料電池**を製造することができる。」がマッチし，〈対象〉が「燃料電池」，〈観点〉が「安全性」のペアが得られる．

この 2 つのパターンも前ステップと同様に，特許明細書の「発明の効果」セクションに適用する．この方法が上手く機能する仮定として，次のようなものがある．

- 「発明の効果」セクションでは基本的に発明の良い点を述べている
- 一文の中でも，最終的な発明の効果は文末に記される．
- 〈対象〉語の近くに〈観点〉語が存在する場合，両者間に関係がある

5.2 クラスタリング手法

以上の方法によって，〈対象〉-〈観点〉ペアが集められる．続いて，〈対象〉と〈観点〉それぞれ別のトピックモデルを生成し，クラスタリングを行う．特許（トピックモデルでいうドキュメント）集合からは，〈対象〉トピックと〈観点〉トピックから構成されるトピックモデルが作られることが理想である．しかし，〈対象〉と〈観点〉はある程度，単語の依存関係があることに加え，トピックモデリング時には，〈対象〉や〈観点〉ではないまた別のトピックも生成されてしまうと考えられるため，「特許」をドキュメントにしたトピックモデルの使用は，〈対象〉-〈観点〉マトリクスマップの作成には適さない．そこで本実験では，〈対象〉と〈観点〉それぞれ別のトピックモデルを生成し，クラスタリングを行う方法を考案した．

〈観点〉のクラスタリングを例に，手順を以下に説明する．

(1) 集められた〈観点〉各要素に対し，該当要素を含む文を集めて 1 つのドキュメントを作成する．
(2) 作成されたドキュメント群のトピックモデルを生成する（トピック数=クラスター数はこちらで与える）．
(3) 各〈観点〉要素のトピック分布を推定する．
(4) 各要素において，最も帰属度が高いトピックがその要素のクラスタとなる．

図 5 は上記の手順を図で表したものである．

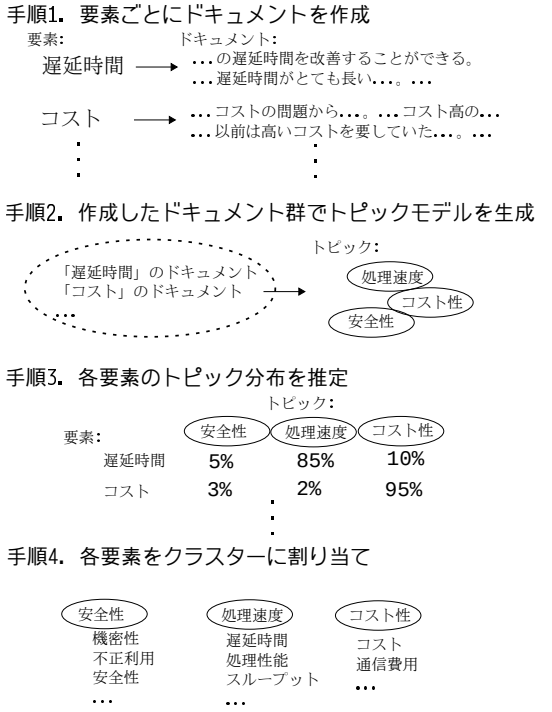


図 5 クラスタリングの方法

また，手順 1. では，名詞と動詞以外の単語，こちらで設定した一般的すぎる語（「等」「上記」「する」「こと」など）は除かれる．加えて，〈対象〉要素との依存性を抑制するために，〈対象〉要素に存在する表現も除かれる．

以上が〈観点〉要素に対してのクラスタリング方法であるが，〈対象〉に対しても同様にクラスタリングを行う．

5.3 実験環境，データ

形態素解析には MeCab，トピックモデリングツールは，自然言語処理ツール MALLET[5] の実装を使用した．

使用する特許データは，国立情報学研究所で作成された NTCIR-5 PATENT[6] の公開特許公報全文データ中から，国際特許分類 (IPC)「G06K 19/07」(主に IC タグ）分野の特許 1972 件を用いて，特許マップ（マトリクスマップ）の作成実験を行った．

5.4 実験結果: ペア抽出

〈対象〉-〈観点〉ペアの抽出を行ったところ，662 個のペアが得られた．その精度を調べるために，80 件をランダムサンプリングして調べたところ，55/80(69%) であった．また，再現率を調べるために特許 1 件あたりの特長表現の件数を調べた．25 件の特許中に 51 件の特長表現があり，平均 2.04 件存在する．よって，再現率は $(662 \times 0.69) \div (1972 \times 2.04) = 11\%$ 程度と推定している．

正しく獲得された特長表現とその〈対象〉-〈観点〉ペアは，次のようなものがある．〈対象〉は下線，〈観点〉は太字で示した．

- **送受信感度**の優れた アンテナ が得られる。
- ＩＣチップ の**回路破壊**を回避することが可能となる。
- 非接触ＩＣカード の**薄型化**を図ることができる。

また，〈対象〉要素または〈観点〉要素単体では間違っていないものでも，ペアとして見たときに間違っているものも存在した．例えば次のようなものである．

- 不揮発性メモリ の**メモリ容量**を有効に活用できる。
- **影響**が少ない 無線通信 を行うことができる。

1 つめのペアは「不揮発性メモリのメモリ容量を増やすことができる」などの特長表現から取れたペアなら良いのだが，1 つめの特長表現は「メモリ容量」を「有効に活用できる」としか述べていない（「メモリ容量」自体が向上したわけではない）ため，特長表現の〈対象〉-〈観点〉ペアとしては不適切である．

2 つめのペアは観点が「影響」であるが，これだけでは何の影響なのか分からない．元の文書では「ノイズの影響」と記載されていたか，今回の実験では〈観点〉及び〈対象〉は「名詞が連続している表現」という条件を課したので，「ノイズ」を観点に含めることができなかった．

5.5 実験結果: クラスタリング

続いて，得られた〈対象〉と〈観点〉それぞれに対してクラスタリングを行った．クラスタリングの評価指標には，Purity と Entropy を用いる．どちらも 0 から 1 の値をとり，Purity はクラスター内の正解要素の割合を表し，大きいほど良く，Entropy は正解要素の散らばり具合を表し，小さいほど良いクラスタリングといえる．

また，正解データとしては，特許庁が公開している特許出願技術動向調査報告書 [7] の IC タグの資料を参考に，こちらが作成した正解クラスターを用いる．具体的には，〈対象〉については「要素技術の範囲・分類」，〈観点〉については「発明目的の範囲・分類」から，それぞれ主要な 8 つのカテゴリーとその他のカテゴリー，合わせて 9 つのカテゴリーに手作業で分類した (表 1,2)．

表 1 対象のクラスタと代表的な語

無線 IC タグ	ＩＣタグ, 情報記憶媒体
リーダ/ライタ	ＩＣカードリーダ, 無線端末装置
IC タグの発行・管理	ＩＣカードシステム, ホストシステム
情報セキュリティ	証明書, 認証システム, 個人情報
IC タグ用アンテナ	アンテナ, コイル, アンテナ基板
応用システム	プリペイドカード, 位置検出, 印刷物
製造	基板間, 金属接合, 工程
回路構成	ＩＣ, ＲＯＭ, 外部記憶装置
その他	金額, エネルギー, カードデータ

クラスタリング時のクラスター数は 10 に設定した．これは，上記の正解データの「その他」に分類される要素が

Vol.2014-IFAT-114 No.9
Vol.2014-DD-93 No.9
2014/3/29

表 2　観点のクラスタと代表的な語

小型化・薄型化	携帯性, 厚み, 小型化
コスト低減	コスト, 運搬コスト, 導入コスト
通信機能の性能向上	スループット, 記憶容量, 書き換え
通信性・耐環境性向上	耐衝撃性, 短絡, 変形
識別性の向上	接続環境, 干渉, 鮮明性, 読み取り性
事業者における波及効果	セキュリティ改竄, 漏洩
生活向上	外観, 視認性, 操作性
製造	ばらつき, 検査時間, 歩留まり
その他	悪影響, 可能性, 回数

多く存在したため，余分なクラスターが生成されてしまうことを考慮したためである．評価結果は，表 3 のようになった．

表 3　クラスタリング評価

	Purity	Entropy
〈対象〉	0.48	0.57
〈観点〉	0.40	0.67

これまでに収集した〈対象〉-〈観点〉ペアと，クラスタリング結果を使用して，縦に〈対象〉，横に〈観点〉を並べたマトリクスマップが作成された（図 6）．マップ中の〈対象〉，〈観点〉クラスターのラベルは筆者がクラスターの要素を見て推測したものである．

〈観点〉

〈対象〉	小型化・コスト	強度	強度	故障・劣化	歩留まり・品質	読取・書込	利用者側の利点	安全性	見た目・デザイン	(不明)
ICタグ本体	3	16	6	15	12	4	2	16	2	8
運送・物流	3	3	1	1	1	0	1	2	0	3
管理システム	2	2	5	12	1	2	3	15	2	9
製造	5	2	3	5	3	2	2	4	1	6
印刷	1	1	1	1	0	0	0	1	2	5
記憶装置	5	5	3	18	2	12	9	14	2	4
通信装置	7	4	1	4	1	0	0	3	3	3
電子回路	0	7	3	7	4	0	1	1	0	6
電子回路	3	2	2	3	1	2	0	0	0	6
(不明)	10	6	2	4	3	1	1	0	1	1

図 6　「IC タグ」のマトリクスマップ

6. まとめと今後の課題

本研究は，技術・発明の新しい長所を表す表現である「特長表現」を用いて，マトリクス型の特許マップを自動生成する方法について議論し，実験を行った．〈対象〉-〈観点〉ペアの抽出の再現率は低く，改善の余地が大きいが，〈対象〉と〈観点〉のクラスタリングにより，類似要素をまとめることで，技術の全体像を捉えることができる特許マップを生成することができた．

今後の課題は，〈対象〉-〈観点〉ペア抽出の再現率を上げることである．そのためには，大きく分けて次の 2 つの方法がある．

(1) 〈対象〉および〈観点〉表現の判定時，さらに多くの表現を取ることができるようにする

(2) 〈対象〉-〈観点〉ペア抽出の文パターンを増やす

1. の〈対象〉および〈観点〉表現の判定には，〈対象〉，〈観点〉語のリストを使うが，今回の実験では，リスト中の語の最後の形態素しか判定に使っていない．精度を保ちつつさらに多くの表現を獲得するには，リスト中の語の最後の形態素だけではなく，最後から 2 つめの形態素も使うようにするか，判定方法の見直しが必要である．

2. の文パターンは，カバー率が高いものを使用したつもりだが，対象や観点が並立して書かれている表現など，まだ取り残す表現も存在するため，いくつか増やすことが考えられる．

また，クラスターのラベリングは今回人手で行ったが，これを自動化する処理の検討もできる．

謝辞　本研究では，実験データとして国立情報学研究所で作成された NTCIR-5 PATENT の公開特許公報全文を使用した．また，本研究の一部は，科研費基盤研究 (B) 25280035 により行われた．ここに深謝する．

参考文献

[1] 西山莉紗，竹内広宜，渡辺日出雄，那須川哲哉：新技術が持つ特長に注目した技術調査支援ツール，人工知能学会論文誌，Vol. 24, No. 6, pp. 541-548, 2009.

[2] 特許庁：出願の手続き (online)，入手先 ⟨http://www.jpo.go.jp/shiryou/kijun/kijun2/syutugan_tetuzuki.htm⟩ (2010.12.10).

[3] 特許庁：特許願・特許請求の範囲・明細書・図面・要約書の具体的な作成例 (online)，入手先 ⟨ http://www.jpo.go.jp/shiryou/kijun/kijun2/pdf / syutugan_tetuzuki/02_06.pdf⟩ (2010.12.10).

[4] David M Blei, Andrew Y Ng, and Michael I Jordan.: *Latent dirichlet allocation*, the Journal of machine Learning research, Vol. 3, pp. 9931022, 2003.

[5] McCallum, Andrew Kachites. *MALLET: A Machine Learning for Language Toolkit*, 入手先 ⟨http://mallet.cs.umass.edu⟩, 2002.

[6] Fujii, Atsushi, Makoto Iwayama, and Noriko Kando.: *Overview of patent retrieval task at NTCIR-5.*, Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization. 2005.

[7] 特許庁：特許出願技術動向調査等報告 (online) 入手先 ⟨http://www.jpo.go.jp/shiryou/gidou-houkoku.htm⟩ (2005.03)