多観点分析に基づくトピック推定と特許マップへの応用

小野寺 大輝　吉岡 真治

北海道大学 情報科学研究科
onodera@kb.ist.hokudai.ac.jp yoshioka@ist.hokudai.ac.jp

概要

我々は、特許文書中の特許の特長を表す表現から、技術課題とその課題を解決するために注目する対象の情報を抽出し、類似した技術課題や対象を整理したマトリクス型特許マップの自動生成手法を提案してきた。しかし、従来の手法では、技術課題として扱える語の種類が限られていたため、技術課題をうまく抽出できない特許が多いという問題と、技術課題や対象を整理するためのトピックモデルの枠組についての検討が不十分であるという問題があった。本稿では、前者の問題に対し、新しい機能の提案を技術課題として抽出する方法を提案し、後者に対し、facet-based LDA を提案するとともに、2001,2002 年の公開特許公報中における航行装置に関する特許に対して分析を行った結果について報告する。

1　はじめに

特許文書の多くは、その分野における技術課題とその解決方法について述べている。我々は、これらの特許に現れる技術課題（課題語）とその課題を解消するための対象となる対象物（対象語）を軸としたマトリクス型特許マップの自動生成手法を提案している。[4] しかし、これまでの手法では、特長表現 [5] に注目した課題語のみを扱っていたため、技術課題をうまく抽出できない特許が多かった。また、収集した課題語と対象語のクラスタリングに用いるトピックモデルの枠組についての検討が不十分であった。そこで、本研究では、課題語と対象語の収集において、特長表現に注目するだけでなく、動詞＋目的語の形で表現される機能表現 [3] も扱うことで、より広範な課題語の収集を行った。また、これらの多数存在する課題語と対象語のクラスタリングの手法として、文書を複数のトピックの組み合わせとして表現可能なトピックモデルを利用し、特許文書が課題とその関連語のトピック、対象とその関連語のトピック、一般の語のトピックの組み合わせから成っていると仮定し、facet-based LDA(fbLDA) を提案する。本モデルを用いることにより、課題と対象に関する語が同時にクラスタリングされ、共起語の情報も利用することで、まとまりのあるクラスタを作成することが可能になる。また、本システムを用いて 2001,2002 年の公開特許公報を分析した結果について実際に作成した特許マップと作成されたクラスタの評価を行った。

2　課題と対象に注目した特許マップ

2.1　技術動向把握のための特許マップ

特許マップの形態は多く存在し、特許情報から読み取りたい情報によって様々な形となる。我々は既に特許に現れる特長表現 [5] の多くに、特定分野の技術課題である課題語（例えばコスト、頑健性）とそれをどの部分により実現するかを示した対象（例えば液晶パネル、LED 照明の 2 つを含むことから、この組み合わせによる特許マップの生成手法を提案してきた [4]。このような技術課題と対象を軸とした特許マップを利用することにより、どのような技術課題が存在するのか、また、どのような対象に注目して特許が提案されているのかといった技術動向の把握に役立つことが期待される。

2.2 課題語と対象語抽出のための記述パターン

従来手法では、既存の技術の改良や問題点の改善を行う特許に多く存在する特長表現に注目して、その課題語と対象語の抽出を行っていた。しかし、様々な分野の特許について分析をしている中で、特許には、このような既存の技術の改良や問題点の改善だけではなく、これまでにはなかったような機能の提案を行っているような特許が数多く存在し、これらの特許について、従来手法では、適切な課題語、対象語の抽出を行うことができなかった。そこで、本研究では、特許文書における課題語の記述の形式について、以下の 2 つのパターンに分けて考える。

1. 機能の特定部分の性能を改善、向上させる技術
2. これまでにないような機能を実現する技術

1. については、特長表現に注目した課題であるため、従来手法である以下のパターンを利用する。

- [対象] の [課題] を向上
- [対象] の [課題] を高める
- [対象] の [課題] を抑制
- [対象] の [課題] を防止
- [対象] の [課題] を低減

2. については、機能の表現の多くが「動詞＋目的語」の形で表現される [3] ことに注目し、これらの機能を実現したことを説明する文書を分析し、以下のパターンを用いて機能に関連する語の抽出を行う。

- [A] が可能となる
- [A] を実現する
- [A] できる

ここで取り出した A の部分について係り受け解析を行うことによって何が可能になったのか、何を実現したのか、などといった技術課題を抽出することが可能になる。対象語については、この機能を実現するために関連する対象語を選ぶ必要があるが、本研究では、係り受け関係にない句から名詞や連名詞を抽出することで、それらを対象語として用いることで、複数の対象を組み合わせた技術についても抽出することとした。

これらの手法を、特許文書の必須項目であり、その特許が取り組んだ技術課題が記述される特許の要約書内の課題項から抽出することにより、特許を特徴付ける課題語と対象語の組として利用する。

3 facet-based LDA による分類

前節で収集した課題語と対象語をそのままマトリクスマップを作成するとほとんどの値が 0 となってしまい、視認性の悪い特許マップになる。そこで本研究で提案する facet-based LDA を用いて単語をクラスタリングすることで似た意味をもつ単語が一つのクラスタを形成し、それらのクラスタによってマトリクスマップを構成することができれば視認性の良い特許マップが作成できると考えた。

3.1 facet-based LDA の提案と特許文書群への応用

本研究では、LDA(Latent Dirichlet Allocation)[1] における文書がトピックの組み合わせにより構成されているという考え方を更に拡張し、特許文書を対象・課題という二つの軸から分析するために、主に対象に関連するトピック、課題に関連するトピックといったトピックのタイプを考え、その組み合わせにより特許文書が作成されると考えたトピックモデルを考える。特許文書群が課題トピックと対象トピック、それ以外の一般トピックから構成されると考え、LDA を目的に合うように改良した形で特許文書群に適用することを考える。このように、分類したい軸に応じたトピックを考えることにより、一つの特許文書に対応する主な対象に関連するトピック、課題に関連するトピックの情報を用いることにより、特許がマトリクス上に配置可能となり、1 回のトピックモデルの構築で、対象と課題について、同時にクラスタリングが行われることになる。このトピックを作成するためのトピックモデルとして以下の生成モデルを仮定する。コーパス D について

(1) 単語分布 Φ_{yz} を各 facet y と各 topic z に対して dirichlet パラメーター β に基づいて作成
(2) 全ての文書 $d \in D$ について
(3) 全ての単語 $w \in d$ について
(4) facet y を収集した課題語, 対象語, その他の語

の数に基づく多項分布からサンプリング
(5)　topic z を Dir(α) からサンプリング
(6)　word w を Φ_{yz} からサンプリング
前節で作成した課題語と対象語のリストに対してそれぞれ facet を割り当てることによって、それぞれの語がその facet を持つトピックに分類され、関連する一般語についても同じトピックに割り当てられることが期待される。この生成モデルに基づくグラフィカルモデルは図 1として表すことができる。

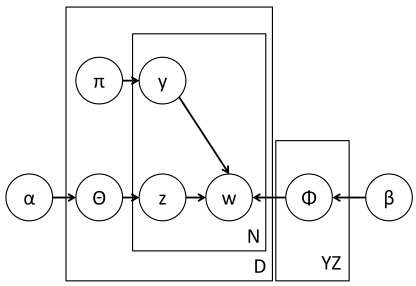

図1　facet-based LDA の graphical model

図 1中の π については実際に出現した課題語、対象語、その他の語の比率に応じた多項分布とした。このような制約を用いることで、課題 facet に属する topic からは課題語とその関連語、対象 facet に属する topic からは対象語とその関連語、一般 facet に属する topic からはその他の語が生成されるトピックモデルを構築することが可能となる。

facet-based LDA を適用することで特許文書群中の単語をトピックに分割し、特許マップ作成に役立てるために、対象を中心としたトピック、課題を中心としたトピック、その他一般トピックを作成し、作成されたトピックをマトリクスマップの軸要素として配置したい。更にこのトピックのトピックという概念、つまり今回で言えば対象と課題がファセットと呼ばれ、このファセットに基づいてトピックが形成されるために facet-based LDA とした。はじめにパターン検索によって得られた課題語と対象語のリストを作成し、特許文書群を課題トピック、対象トピック、その他一般トピックから構成されると仮定して制約を導入することで、課題トピックには課題語と課題語と関連のある語が分布の上位に、対象トピックには対象語と対象語と関連のある語が分布の上位に来るようにしており、それらの作成された課題トピック、対象トピックのラベル付けを手動で行うことによって、それらをマトリクスマップの軸要素とし、実際の要素としては課題と対象が現れた特許の数が記録されることとなる。

4　実験結果と考察

4.1　実験データについて

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[2] の公開特許公報全文データ中から，国際特許分類 (IPC)「G01C 21」(主に航行装置）分野の特許 3117 件 (2001 年〜 2002 年) を用いた。課題語 540 個、対象語が 1628 個、その他の語が 3576 個が得られ、課題語と対象語に特定の facet を付与し、facet-based LDA を適用する。なお、トピック数は 8 で全ての facet についてトピックが存在するので合計 24 トピックで実験を行う。

4.2　実験結果

課題語と対象語の収集については従来の 1. の手法において獲得できたペアは 713 個となり、機能表現に注目した 2. の手法においては 4642 個のペアが抽出された。2. の手法においては 1 文書当たり 1 つ以上のペアを抽出できていることになり、従来手法に比べより網羅的に抽出できることが確認された。それぞれのファセットのトピックにそれぞれ推定された単語について上位 50 語の中に含まれる課題語と対象語をそれぞれ、そのトピックに含まれる課題語、対象語として考えた。これらのトピックに含まれる課題語と対象語について元々の特許文書に特定の課題語と対象語のペアが存在するとき、属するトピックのマトリクスの要素に +1 を加えることで、マトリクス型特許マップの要素数が、該当する特許文書数になると考えた。得られた特許マップを図 2に示す。これまで課題であったアドホックな仮定を導入せずに、ファセットを持ったトピックが作成可能になったが、1 つの特許文書の要約書内の課題項だけでは、1 文書当たりの単語数が少なく、出現回数が極端に少ない語が存在したため、このようにスパースなマトリクス型特許マップとなってしま

課題

対象	送受信性能	表示性能	地図の表示	経路探索の正確性	コストの低減	情報の提供	地域情報の取得	なし
経路案内	1552	710	234	96	123	18	32	0
ディスプレイ	289	78	6	12	13	0	20	0
通信システム	105	45	6	6	5	3	0	0
位置情報	25	29	0	2	6	0	0	0
地図データ	9	18	8	2	2	3	0	0
不明	0	0	0	0	0	0	0	0
入力操作	1	0	0	0	0	0	0	0
ナビゲーションシステム	0	0	0	0	0	0	0	0

図2　作成した特許マップ

ったが、多くのペアについては上手く分類することができた。

5　まとめと今後の課題

本研究では従来手法 [4] における課題語と対象語の抽出の網羅性が低い問題に対し、機能表現を含む課題の抽出を行うことにより、より網羅的な特許の分類ができるようになることを確認した。また、facet-based LDA として新しいトピックモデルの枠組を適用した結果、前回の手法でアドホックに導入したパラメータなどを用いなくても、適切にクラスタリングが行えることを確認した。今後の課題として、この特許マップの有用性の評価を行うために、特許マップを扱った経験のある知財担当者へのインタビューを行う予定である。

謝辞

本研究の一部は、科研費 25280035,26280111 の支援を受けて実施した。

参考文献

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Atsushi Fujii, Makoto Iwayama, and Noriko K. Overview of patent retrieval task at ntcir-5. In *In Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 269–277, 2005.

[3] Yasushi Umeda, Masaki Ishii, Masaharu Yoshioka, and Tetsuo Tomiyama. Supporting conceptual design based on the Function-Behavior-State modeler. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing (AIEDAM)*, 10(4):275–288, September 1996.

[4] 小野寺大輝 and 吉岡真治. 対象-観点を考慮した facet-biased トピックモデルと特許マップへの応用. 言語処理学会第 *22* 回年次大会発表論文集, (22):13–5, 3 2016.

[5] 西山莉紗. 技術文書の情報編纂: 課題・特長・手段を表す表現の抽出と利用 (人工知能学会全国大会 (第 26 回) 文化, 科学技術と未来) – (近未来チャレンジセッション「nfc-4 (卒業セッション) 情報編纂の基盤技術」). 人工知能学会全国大会論文集, 26:1–4, 2012.