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

小野寺 大輝　　吉岡真治
北海道大学 情報科学研究科

{onodera,yoshioka}@kb.ist.hokudai.ac.jp

概 要

特許文書には多数の技術情報が含まれており、その内容を俯瞰的に見るための方法として特許マップが提案されている。我々は、既に特許に含まれている定型的な効果に関する表現である特長表現に注目して、特長の持つ対象と観点という 2 つの次元から構成される特許マップの自動生成手法を提案している。本研究では、特許マップの対象と観点を表す語のクラスタリングを行うための facet-biased トピックモデルを提案すると共に、このモデルで作成した特許マップの作成手法を提案する。

1　研究の背景と目的

特許は開発した技術の知的財産権を守るためのものであり、出願者は請求範囲を広く取るために抽象的な特許独特の表現を多様する傾向にある。そのため、特許を読むことは非専門家にとって障壁が高いものであった。そこで特許の情報を分かりやすく示した特許マップ [1] がある。

我々は既に特許マップの自動生成のために、西山ら [2] の研究で定義される特長表現を更に対象と観点に分けた特許マップの自動生成を提案してきた [3]。しかし、この手法では特許文書中に含まれる対象と観点を表す語をクラスタリングする必要があるが、その手法の検討が十分ではなかった。

そこで、本研究では各文書が対象語と観点語、その他の語から構成されるトピック群から生成されると考え、代表的な対象・観点トピックを用いてマップを作成する。そこでトピックモデルにおいて生成されるトピックについて対象と観点を代表させるトピックという考え方を導入した facet-biased トピックモデルを提案すると共に、このモデルで作成した特許マップの作成手法を提案する。

2　特許マップと関連研究

特許の情報を分かりやすく図にして表したのが特許マップ [1] である。本節では、一般的な特許マップを紹介すると共に、特許情報の実用的活用を目指した関連研究を示す。

2.1　特許マップ

特許マップとは大量の特許情報を分かりやすく、図にまとめたものである。様々な形式の特許マップがあり、調査目的によって利用の仕方も異なる。例としては縦と横の組み合わせで技術を 2 つの側面から観察し、技術の穴場や動向を把握できるマトリクスマップ型や特許技術情報を体系化し、技術ごとのツリーを作成することで技術の成熟度や技術の集中部分等を把握できるツリー型などがある (図 1)。

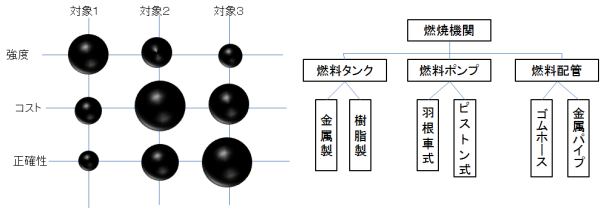

図 1: 特許マップの例

2.2　特長表現に注目したトピックモデル

多くの技術文書には、その技術が既存の技術に比べてどのような利点があるのかといった記述が存在する。西山ら [2] はこのような表現を特長表現と呼び、「当該技術の新たな長所を示した表現」と定義した。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

などがプラスの極性を持つ特長表現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

などがマイナスの極性を持つ特長表現の例として挙げられる．西山らはこれらの表現を構文パターンを用いて解析し、更に斬新な技術応用を発見するためにランキング付けを行うことで新たな技術文書のまとめ上げとして提案した。

岸ら [3] は、特長表現の多くに、評価の基準である観点（例えば安定動作、操作性）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（図 2）の生成手法を提案してきた。この手法では、文書中からパターンを用いて対象語、観点語の文書クラスタリングを行うことで、特許マップの作成を行う。このクラスタリングを行う方法として、対象語と観点語の各々について、それらの語を含む特許文書の全てを集めた代表文書を作成し、トピックモデル [4] で次元圧縮をした後、Ward 法でクラスタリングを行った。しかし、代表文書の作り方などに

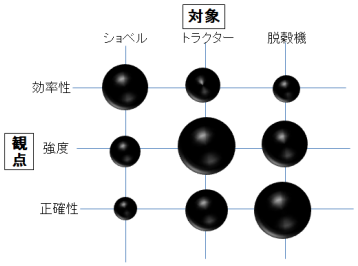

図 2: 〈対象〉-〈観点〉の特許マップ

ついての検討などが不十分であり、手法についての再検討が必要であった。

以下で岸が用いたシステムの流れを示す。

1. 特定分野の特許を集める
2. 特許文書中の特長表現から対象語と観点語のペアを抽出
3. 軸にする対象語と観点語をクラスタリング
4. 特許マップの作成

3 facet-biasedトピックモデルに基づく特許マップの生成

前節で述べた特許マップ生成手法では、対象語と観点語のそれぞれの代表文書とそれを用いたクラスタリングであり、各々の語を含む全ての特許文書を利用しているため、複数の対象語や観点語を含むような特許を扱う際に、関係の薄い記述を代表文書に含めてしまうという問題があった。

そこで本研究では、トピックモデル [4] を拡張し、対象語に関係するトピック、観点語に関するトピック、その他の語に関するトピックという、3 つのタイプのトピックの混合として特許文書が生成されると考えたfacet-biased トピックモデルを提案し、クラスタリングに利用する。

facet-biased トピックとは、一般のトピックモデルのトピックと異なり、特定の facet（特許マップにおける対象と観点）に関連する語を主に生成するためのトピックである。本節では、facet-biased トピックモデルの提案を行うと共に、facet-biased トピックモデルの結果を用いた特許マップの作成を行う。

3.1 facet-biasedトピックモデルによるクラスタリング

トピックモデルとは、1 つの文書は複数のトピックから確率的に生成されていると考え、トピックを生成するモデルである。代表的には LDA[5] がある。LDA は単語間に現れる相関共起の情報を用いるので、文書群を特徴付けるような重要な語があれば、その後に関連する語も同じトピックに帰属する可能性が高い。しかし、このトピックモデルでは、本研究で想定しているようなトピックの種類という考え方がなく、結果として対象語や観点語に関係するトピックが生成される場合もあるが、特許マップ作成のための情報としては不十分である。そこで本研究では、facet-biased トピックモデルを提案する。

このトピックモデルでは、単語群が facet に対応する形で分類されていることを仮定する。また、既存のトピックモデルと異なり、特定の facet に属する単語を主に生成するトピックとして facet-biased トピックというトピックのタイプを定義する。また、このモデルでは各々の facet に属する単語は facet-biased トピックのみから生成されると仮定する。この考え方に基づいて生成するトピックと、そのトピックによる単語の発生確率の関係を表 1 に示す。

表 1: 単語の発生確率の設定

	対象語	観点語	その他の語
対象 (facet1)	推定値	0	推定値
観点 (facet2)	0	推定値	推定値
その他	0	0	推定値

表 1 の中で推定値は既存の LDA と同様の手法による推定値を用いる。0 と示した部分は、facet に属する語は facet-biased トピックからのみ生成されることを示している。ただし、トピックに関係する文書に存在しない語と同様に 0 に近い確率を与える。この表からわかるように、本手法では各々の facet 毎については、独立性を仮定してトピックを生成し、その他の語については facet-biased トピックに関連性が強い場合には、これらのトピックに、そうでない一般的な語（例えば、特許一般で用いられる語）は、その他のトピックに含まれることが期待される。

しかし、予備実験を行ったところ（詳細は他の実験と合わせて後述）、対象語や観点語の種類が少なく、またその頻度がその他の語と比べて少ないため、本来なら facet 語を中心としたトピックになるはずの facet-biased トピックの重要語にその他の語が多く存在し、多くの観点語が複数（3〜5）のトピック割り当てを持ち、クラスタリングとして役に立たないという問題があった。そこで、各文書のトピック割合を決める際の単語の重みを設定するパラメータ f_wを導入する。具体的には LDA のモデルから文書のトピックを決定する lda inference において、ファセットに属する語については、全体のトピックに寄与する割合を変更するために、単語の出現頻度ではなく、出現頻度に一定の重み（facet 単語重み f_w）を乗じたものを出現頻度として利用する。

3.2 対象語・観点語リストの作成

岸ら [3] の研究における構文パターンのみによる対象語と観点語の収集には網羅性の問題が生じており、上述したように対象語や観点語を重要語とするためには対象語や観点語の数と、その頻度を増やす必要があった。そこで本研究では対象語に注目して特許の性質上、抽象的な複合語の組み合わせで現れることが多い専門用語を対象語として獲得するために専門用語抽出ツールの利用を考える。専門用語を抽出するツールとしては単語の成長可能性に注目した TermExtract[6] が存在する。例えば「IC チップ」という語は「IC」と「チップ」の複合語である。TermExtract はこれらの複合語に注目して、頻繁に現れる語の組み合わせを重要な複合語だと判断する。つまり、「IC チップ」の用語性を高いと判断したならば、「ベア IC チップ」などのように用語性が高い用語を構成要素としている複合語も用語性が高いと判断する。本研究においては、観点語は「〜性」や「〜能」など現れる語彙の種類が限定的であることが多数見受けられたため、抽出精度の高い岸の構文パターンを利用して収集した。対象語は複合語の組み合わせで現れる語が頻出し、その種類も多岐に渡ることから TermExtract を用いて収集した。

3.3 特許マップの作成

facet-biased トピックモデルを用いることにより、特許文書中の全ての対象語と観点語についてトピックが割り当てられる。これらの facet-biased トピックに含まれる重要語（発生確率の高い facet 語）により、各々のトピックが特徴付けられると考えて、特許マップを作成する。従来の手法と異なり、多義を持つと考えられる語については、複数の項目と関連付けられることがある。また、岸らの手法と同様の手法で抽出した特長表現に含まれる対象語、観点語のトピック割り当てを用いて、その特長表現のマップ上の対応箇所を決定する。対応する特長表現の数を表すことにより、特許マップとする。

4 実験とその評価

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[7] 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 件を用いた。facet-biased トピックモデルの実装には [5] の著者である blei 氏が公開している LDA の implement である lda-c[8] を利用させていただいた。

4.1 実験結果

今回用いた IC タグ分野の特許 1972 件においては対象語と観点語、その他の語の出現頻度は表 2 の結果となった。

表 2: それぞれの語の出現回数と異なり語総数

	異なり語数	総出現回数
対象語	45	4270
観点語	50	2347
その他の語	1813	52001

表 2 を見ると分かるように、対象語と観点語の出現頻度とその他の語の出現頻度には大きな差が存在している。また、本実験では、10 × 10 の特許マップを作成することを前提に対象や観点、その他のトピックを各々10（計 30 トピック）と設定し、facet-biased トピックモデルの生成を行った。最初に、生成された facet-biased トピックモデルがクラスタリングに有用かどうかを検討するために対象語や観点語がそれぞれどれくらいそのトピック内での支配的要素になっているかという評価について、対象語と観点語の割り当てられたトピックに関するエントロピーを計算し、どれだけバラツキがあるかを求めた。エントロピーが高すぎる場合は、対象語や観点語がそのトピックで支配的な要素となっていないために、その他の語の存在に強く影響を受けている可能性が高い。ただし、多義語については複数のトピックに割り振られることが適切であるため、エントロピーが低ければ低いほど良いということではないことに注意が必要である。エントロピーの計算にはそれぞれの対象語が、どのトピックに平均で何回割り振られたかという情報を用いて行った。図 3 に対象語と観点語のエントロピーと facet-biased トピックモデルに用いた単語の重みパラメータの関係を示した。

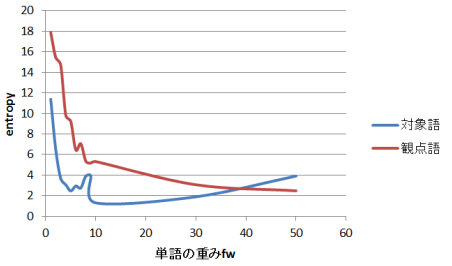

図 3: 対象語、観点語とパラメータの関係

対象語についてはその他の語に対して約 12 倍、観点語については約 22 倍の総出現回数の割合が異なるが、共に facet 単語重み f_w を増やしていくにつれて、最初は減少傾向を示すが、その後増えたり減ったりと不安定になる。値を大きくしすぎると、対象語と観点語のトピックが全体の文書を構成するトピックとして支配的になりすぎ、その他のトピックの要素が減ってしまう。その結果、対象語については $f_w = 10$ を超えてから、その他の語のうち共起する語の情報が減少していくのでエントロピーは上昇してしまっている。一方で観点語については対象語に比べて各単語の出現頻度は低いものの、種類は対象語よりも多いため、f_w を増やすとトピックの決定に利用できる観点語が増えて、エントロピーが減少していくと考えられる。

更に、モデルと特許文書の適合性の評価には perplexity を用いた。perplexity は $PPL = exp\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\}$ であり、文書中に現れる各単語における出現確率の逆数の幾何平均になっている。図 4 に facet-biased トピックモデルに用いた単語の重みのパラメータと perplexity の関係を示した。

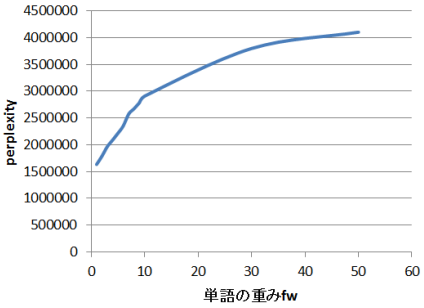

図 4: perplexity とパラメータの関係

perplexity は単語の重みパラメータ f_w を増やすと増大した。また、f_w を大きくしすぎると、トピックの推定に有用なその他の語が利用できなくなるために perplexity が増大していると考えている。なお、同データをトピック数 30 で既存のトピックモデルに適用したときの perplexity は 1422209 で、これは生成するトピックに制約を与えていることから当然の結果である。

facet-biased トピックモデルによるクラスタリング結果を表 3 と 表 4 に示す。この表の作成には $f_w = 10$ での結果を用いて行った。ラベルは筆者がクラスタの要素を見て判定した。

		対象									
		ICカード	アンテナ	カード	情報システム	メモリ	ICチップとタグ	製造	システム	利用	認証装置
観点	信頼性	61	47	49	37	41	40	19	20	0	7
	耐熱性	47	44	23	17	0	28	19	0	0	0
	利便性	35	0	15	15	21	5	24	25	19	0
	生産性	40	23	38	0	1	15	2	7	0	4
	セキュリティ1	27	16	0	34	0	14	4	0	0	0
	セキュリティ2	26	0	0	2	41	1	0	0	3	1
	安全性1	4	0	6	19	0	1	12	0	0	1
	形状	24	0	1	0	1	1	6	0	0	1
	精度	0	10	0	1	0	0	0	0	0	4
	安全性2	0	0	0	3	0	0	0	0	0	0

図 5: 本研究で作成した特許マップ

表 3: 対象を代表するトピック

クラスタ	要素
IC カード	IC カード、半導体チップ、カード表面、製造工程、システム、無線通信カード
アンテナ	アンテナ、非接触 IC カード、非接触データ送受信体、製品、IC チップ、複合 IC カード
カード	カード、システム、カード表面、情報、IC モジュール、非接触 IC カード
情報システム	情報、商品、情報処理システム、システム、物品
メモリ	データ、メモリカード、IC メモリカード、システム、情報
IC チップとタグ	IC チップ、IC モジュール、IC タグ、REID タグ、半導体装置
製造	工程、電子部品、パッケージ、製造工程、接続部
システム	物品、システム、アンテナコイル、部材、非接触情報担体
利用	利用者、サービス提供、二次電池、システム、情報、工程
認証装置	認証、PC カード、情報記憶装置、接続部、情報、無線通信カード

表 4: 観点を代表するトピック

クラスタ	要素
信頼性	信頼性、感度、剛性、損傷、短縮、分解
耐熱性	耐熱性、製造コスト、破壊、偽造、通信距離、通信特性
利便性	短縮、伝送、利便性、凹凸、再利用、簡素化、改竄
生産性	影響、品質、性能、生産性、断線、反り歩留まり
セキュリティ1	強度、書き換え、耐久性、破損、セキュリティー、損傷
セキュリティ2	容量、セキュリティ、増加、実用性、損傷ばらつき
安全性 1	発生、取引、安全性、ばらつき、不正使用
形状	外観、増加、短縮、影響、生産性、生産コスト
精度	精度、薄型化、分解、安全性、流通、耐久性
安全性 2	利用、コスト、流通、漏洩、簡略化、不正使用、小型化

対象トピックについては多くの対象語が重要語（発生確率の高い語）として利用され、類似した語のまとまりとなったが、観点語については重要語のまとまりはあるものの、特に発生確率の低い語については類似の語が別のトピックに割り当てられることも多かった。この問題についてはそもそもトピック数を 10 に分けることが適切だったのかという点も含めて再考察する必要がある。

図 5 は本研究で作成した特許マップである。このマップを用いることで、独立性を仮定した対象と観点の依存関係を読み取ることができる。

5 まとめと今後の課題

本研究では facet-biased トピックモデルによって対象語と観点語の代表トピックを作成し、特許文書を用いた特許マップの自動生成に応用した。また、facet に関する語とその他の語の発生頻度の違いを考慮した facet 単語重みを導入することで、facet 語についてのまとまりのある facet-biased トピックが作成できることを確認し、これに基づく特許マップの作成方法も提案した。

今後の課題としては、facet 単語重みや適切なトピック数などの決め方について、十分な考察ができていないため、その実験的な分析が必要である。さらに、特許マップとしての妥当性、有用性についても分析していきたい。

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 の公開特許広報全文を使用した。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。ここに記し、謝意を表す。

参考文献

[1] 藤井 敦, 谷川 英和, 岩山 真, 難波 英嗣ほか: 特許情報処理 : 言語処理的アプローチ (自然言語処理シリーズ) 奥村 学監修, コロナ社, 2012.

[2] 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.

[3] 岸 桂太, 吉岡 真治 : 特長表現に注目した対象-観点型特許マップの自動生成，情報処理学会情報基礎とアクセス技術研究会,2014-IFAT-114-9,2014.

[4] 佐藤一誠: トピックモデルによる統計的潜在意味解析（自然言語処理シリーズ）奥村 学監修, コロナ社, 2015.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan.: *Latent dirichlet allocation*, the Journal of machine Learning research, Vol. 3, pp. 993-1022, 2003.

[6] Nakagawa, H, Mori, T,: Automatic Term Recognition based on Statistics of Compound Nouns and their Components, Terminoloy, Vol.9 No.2, pp.201-209, 2003.

[7] 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.

[8] LDA-C: `http://www.cs.princeton.edu/~blei/lda-c/` (2016.01)