Vol.2013-DBS-157 No.10
Vol.2013-IFAT-111 No.10
2013/7/22

特長表現に注目した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、特長表現を網羅的に集めるための手法を提案するとともに、効果とその対象に注目し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．インターネットや周辺機器の発達により情報を入手・処理できる基盤は拡大を続けているのに対し，情報を利用する側は，手に入る情報を活用しきれていないのが現実である．

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．

また，特許に限らず，技術情報を体系化し，該当分野における技術開発の方向性を予測することは，直接技術を利用・開発する立場の企業や研究機関だけでなく，国や機関による科学技術戦略の決定にも重要である．以上のような背景から，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，Advantage Phrase）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．特長表現は，定型的に記述されることが多く，「〜が向上する」のような手がかり句を用いて抽出することができる．

本研究は，上記の特長表現を用いて，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について論ずる．

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

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．記載例は，特許庁ホームページの「出願の手続き」[3] における作成例 [4] から引用した．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．例えば，**図 1** は出願年ごとの出願件数を示した特許マップで，

特許の一つ一つに付与されている書誌情報（出願日，特許分類コードなど）を利用すれば比較的容易に作成できる．

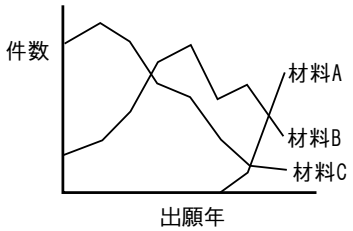

図 1　件数推移マップ

図 2 はマトリクス表示マップと呼ばれ，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」などがあり，本研究では，特許文書中の「特長表現」と定義される記述に注目して，マトリクス表示マップを半自動的に生成する方法を提案する．

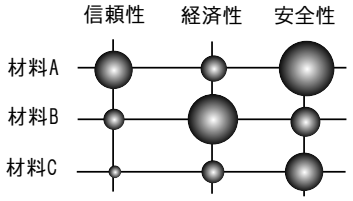

図 2　マトリクス表示マップ

3. 特長表現とその抽出方法

3.1 特長表現の手がかり句

西山らは，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．

特長表現は，増強クラス (Enhancement class) と改善クラス (Amelioration class) 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．本稿では「向上する」のような特長表現を同定するのに使用するフレーズを手がかり句と呼ぶことにするが，西山らは人手で作成した手がかり句を用いて特長表現を抽出する方法を採用していた．本研究では，より網羅的に特長表現を抽出できるようにするため，次に説明する手法を用いて手がかり句のさらなる獲得を行った．

3.2 Espresso アルゴリズムによる特長表現の抽出

酒井らは特許明細書から技術課題情報の抽出を行うために，技術課題情報の手がかり句を bootstrapping により獲得する研究を行なっている [2]．酒井らの研究における技術課題とは，本研究で扱う特長表現とほぼ同じであるが，増強クラスと改善クラスという区別はしていないため，酒井らの手法をそのまま 2 クラスの特長表現の収集に適用すると，後述する「意味ドリフト」が起こりやすかった．今回は特長表現に特有の，2 つのクラスの手がかり句を別々に収集するため，bootstrapping の代表的な手法である Espresso アルゴリズムを使用した．

bootstrapping とは，集めたいインスタンスに共起するパターンの収集と、パターンに適合するインスタンスの収集を再帰的に繰り返し，少数の正解例から、同種のものを順次獲得し増やしていくための方法である．しかし，bootstrapping の過程で，適切でないパターンが入り込んでしまうと，集めたいものと異なる集合のインスタンスが獲得されてしまう．これを「意味ドリフト」と呼び，意味ドリフトを抑えて bootstrapping を行う代表的な手法が Pantel[5] による Espresso アルゴリズムである．このアルゴリズムは，信頼度の高いパターンから得られたインスタンス候補は高得点となるようなスコアリングを行うもので，逆も同様（信頼度の高いインスタンスから得られたパターンは高得点）である．

今回の bootstrapping において，パターンは手がかり句 (インスタンス) との係り関係とする（図 3）．

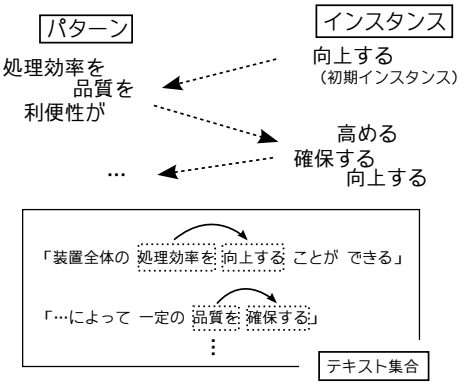

図 3　bootstrapping

Vol.2013-DBS-157 No.10
Vol.2013-IFAT-111 No.10
2013/7/22

4. 特許マップの自動生成

4.1 対象-観点マップ

本研究では，特許文書中の特長表現から「対象」と「観点」を抜き出し，それらを二軸に配置したマトリクス表示の特許マップの生成を行う．まず対象と観点の例を挙げると，「磁気記憶装置の耐障害性を高める」という特長表現があったとき，対象は「磁気記憶装置」，観点は「耐障害性」である．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 **4**）．

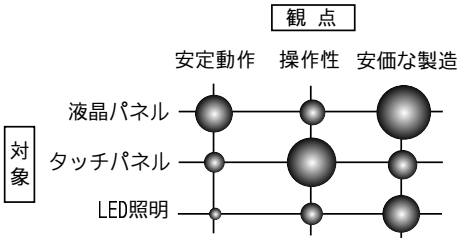

図 **4** [対象]-[観点] マップ

既存のマトリクス表示マップと比べて，特長表現を用いることで得られると思われる利点は大きく次の二点である．

- **ユーザーにとって分かりやすい**
  よく扱われる「技術課題-解決手段」の組み合わせと比べて，「対象-観点」マップは最終的な発明の効果に着目しているので，該当分野に詳しくない者やユーザー側から扱いやすいマップが生成できる．
- **マップ視認性の向上**
  特長表現は，良い所を伸ばす表現と悪い所を減らす表現を，それぞれ増強クラスと改善クラスというように，明確に区別している．それによって，正反対の観点が混ざらないようにマップを作ることができ，対象と観点の関係が見やすくなる．

4.2 [対象] と [観点] の抽出方法

特長表現は，「コンテナの断熱性を向上する」などのような，「[対象] の [観点] を [手がかり句]」という順番で構成されていることが多い．これをもとに，[対象] と [観点] を抽出する．

まず，観点の抽出方法について述べる．

(1) 特長表現の手がかり句に係っている文節の中で，最後の文字が「が」「を」「も」であり，最も手がかり句に近い位置にある文節を C_1 とする．
(2) $S_{Aspect} := C_1$
(3) C_1 のひとつ前の文節 C_2 が C_1 に係っていないなら，終了．観点は S_{Aspect}.
(4) C_2 の末尾が「の」である場合，$C_a := C_2$．観点は S_{Aspect} として終了．
(5) $S_{Aspect} := C_2 + S_{Aspect}$
(6) $C_1 := C_2$, (3) に戻る

観点の抽出処理の終了後，対象の抽出を行う．対象は手がかり句や観点と離れた位置に存在したり，そもそも明示的に書かれていないことも多い．よって，対象の抽出方法は次のようにした．

(1) C_a が未定義の場合，「発明の名称」を対象とする．定義されている場合，(2) へ
(2) $S_{Target} := C_a$
(3) C_a のひとつ前の文節 C_b が C_a に係っていないなら，終了．対象は S_{Target}.
(4) $S_{Target} := C_b + S_{Target}$
(5) $C_a := C_b$，(3) に戻る

5. 実験と考察

5.1 実験内容

特長表現の手がかり句を bootstrapping で収集し，獲得した手がかり句を用いて特長表現を抽出する．次に，特長表現から [対象] と [観点] を取り出す．

実験の特許文書セットは，国立情報学研究所によって作成された NTCIR-5 PATENT [6] の公開特許公報全文データ中の 2002 年前半の特許明細書から，ランダムに選んだ 1861 件を使用した．特長表現は，特許明細書の「発明の効果」セクションから抽出する．特許明細書において，「発明の効果」セクションは必須ではないが，実験に使用した 1861 件中 1651 件に「発明の効果」セクションが存在した．また，形態素解析器は mecab 0.993，係り受け解析には CaboCha 0.66 を使用した．

5.2 結果 1: 手がかり句獲得

5.2.1 手法

Espresso アルゴリズムを用いて，新たな特長表現の手がかり句を収集する．少数の正解例（初期インスタンス）をまず人手で与える必要があるが，増強クラスの初期インスタンスには，「向上する」「可能となる」「実現する」を与え，改善クラスの初期インスタンスには，「防止する」「抑制する」「低減する」を与えた．

5.2.2 獲得した手がかり句（増強クラス）

「向上させる」「向上して」「向上できる。」などの初期インスタンスと同じ単語が入っているもの以外には，「高める」「確保する」「期待する」「提供する」「達成する」などの表現が得られた．

5.2.3 獲得した手がかり句（改善クラス）

初期インスタンスと同じ単語が入っているもの以外には，「防ぐ」「除去する」「少なくする」などの表現が得られた．

5.2.4 考察

増強クラスの新しく得られた手がかり句は正しいと思われるものが多かったが，改善クラスの新しく得られた手が

かり句は,「与える」「生じ､」「捉える」などの手がかり句にしては一般的すぎる表現が散見された．パターン・インスタンス収集の反復回数を増やすか，使用するテキストデータ量を増やせば改善されるかもしれない．

5.3 結果 2: [対象]-[観点] の抽出

結果 1 で得られた手がかり句により特長表現を抽出し，そこから [対象]-[観点] として，以下のようなものが取得できた．ただし，特長表現内から [対象] が見つからず,「発明の名称」を [対象] として代用しているペアの場合，その場合の [対象] 部はカッコで囲んである．

5.3.1 増強クラスのペア

- ソフトハンドオーバー中の移動局-通信品質
- 原稿台上で-作業性
- データ取得-省力化
- 発熱性の電気部品-放熱効果
- 作業性-向上
- (反応器)-反応率
- 汚染土壌-浄化
- 製品-歩留り
- 印刷機械-稼働率
- 画像出力-品質
- (車両用ステアリング装置)-長寿命化
- エーテル化反応工程後に得られるセルロースエーテルの水溶液-透明度
- 軸受部材-固定精度

5.3.2 改善クラスのペア

- 電気部品-過熱
- (無線基地局ネットワークシステム、統括局、信号処理方法、及びハンドオーバー制御方法)- 相互に干渉すること
- 歪みが大きくなる等の操作性-低下
- (端末装置、中継装置、通信方法及びその通信プログラムを記録した記録媒体)- 無駄に中継すること
- MR 素子-静電破壊
- 弾性コーナー部材-脱落
- 触媒上に吸着した反応種による反応率-低下
- (排水処理システム)-2 次流量調整槽が溢れるような不具合
- 低温腐食-発生
- (起動スイッチ及びこれを備えた電動機)-消費電力
- 燃料電池-損傷

5.3.3 考察

「発明の効果」が記載されている 1651 件の特許から，増強クラスの [対象]-[観点] ペアは 1199 件，改善クラスの同ペアは 599 件抽出できた．明確な正解データは用意できていないため，絶対的な正解率を算出することはできないが，筆者の基準で判断すると，正解率は 6，7 割である．ただし，[対象]-[観点] のペアとして意味的には正解であったとしても，表現が長すぎたりして，実際にマトリクスマップにそのまま使用することは出来ないものが多いため，シソーラスなどを使って他の [対象] や [観点] と意味的に統合する必要がある．

うまく抽出出来なかったペアを見ると，抽出元が，今回の [対象]-[観点] ペア抽出アルゴリズムの前提となっている,「[対象] の [観点] を [手がかり句]」という構成の特長表現ではなかったことからの失敗が多いが，そもそも特長表現の手がかり句として不適当なものが特長表現の収集に使われてしまっていることも，抽出がうまく行われていない原因である．

6. まとめと今後の課題

[対象] と [観点] を抽出したあと，そのままマトリクスの形にするのではなく，実用のためには，使用する特許データの分野を限定した上で，同じものや似た [対象] や [観点] を統合する必要がある．また，今回はまずマトリクス表示マップを生成することを目標に研究を進めてきたため，各段階において結果の評価やアルゴリズムの調整が十分にできていない．これからは，正解データを用意するなどして実験結果の確実な評価を行い，抽出手法の改善を図りたい．

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] 酒井浩之，野中尋史，増山繁：特許明細書からの技術課題情報の抽出，人工知能学会論文誌，Vol.24, No.6, pp. 531-540 (2009).
[3] 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).
[4] 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).
[5] Pantel, Patrick and Pennacchiotti, Marco.: *Espresso: leveraging generic patterns for automatically harvesting semantic relations*, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 113-120, (2006).
[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).