製品特徴に基づく製品発表プレスリリースと特許との関連性の判定

酒井浩之 [1*]　増山繁 [2]

[1] 成蹊大学 理工学部 情報科学科
[2] 豊橋技術科学大学大学院 工学研究科 情報・知能工学専攻

1　はじめに

企業にとって製品開発を行う際，あるいは，技術開発完了後に新規の特許出願を行う際に，他社特許を調査することは非常に重要である．このような特許調査において，単純に関連技術分野の特許を検索するだけでなく，他社の競合特許と他社との競合製品の関係に着目し，製品に関連している特許の特定，さらには，それぞれの特徴レベルでの関連の特定を行うことが必要となる．しかしながら，製品の機能を把握し，大量の特許文書の中から，その製品の機能に関連する特許を見つけだすことは容易ではない．そこで，製品とそれに使用されている特許の判定，特に特徴レベルでの関連性の判定を行うシステムの開発が求められている．例えば,「エチレン強力分解で野菜の新鮮保存を可能にした冷蔵庫の発売」に関する製品発表プレスリリースにおいて，製品特徴を示す情報として「新製品は，当社が独自開発した光プラズマ強力脱臭・抗菌装置を搭載しており，野菜の劣化を進める大きな要因であるエチレンガスを紫外線と光触媒の作用によって分解し，従来の約１００倍の能力で減少させることができます。」という文が含まれている．この情報から「放電手段を構成するメッシュ状電極を光触媒組立のケースに固定し、放電手段から安定した放電を得ることができ、脱臭性能や酸化分解性能の向上を図ることができる。」という文を含む「冷気の循環経路内に脱臭装置を配置して庫内を脱臭するようにした冷蔵庫に関する」特許と関連性が高いことが分かる．本研究は，この関連性の判定を自動的に行うものである．本研究では，まず，製品発表プレスリリースから，その製品の製品特徴となるキーワードを抽出する．上記の「エチレン強力分解で野菜の新鮮保存を可能にした冷蔵庫の発売」の例では「光触媒」「光プラズマ」「脱臭」「抗菌」といったキーワードを抽出する．そして，特許明細書において，特許発明を使用することによりもたらされる効果を示す記述が出現している可能性が高い「請求項」「発明の属する技術分野」「発明が解決しようとする課題」「課題を解決するための手段」「発明の効果」を対象とし，抽出した製品特徴キーワードを使用して特許との関連性を判定する．

*連絡先：成蹊大学 理工学部 情報科学科
〒 180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: h-sakai@st.seikei.ac.jp

2　関連研究

特許情報処理に関連する研究としては，特許翻訳や，特許情報を技術と効果を軸として平面上に可視化したパテントマップを自動生成 [3] するものなどが中心となっていた．一方，NTCIR-3 では新聞記事から記事内容に関連する特許を検索するという技術動向の調査を模倣するタスクが行われた．このタスクでは，Itoh らは新聞と特許という 2 種類のコーパスで単語の出現頻度が異なるという性質を利用した Term Distillation を提案している [1]．更に，この結果を Web 検索タスクに対しても適用している．しかしながら，NTCIR-3 のタスクは，単純に新聞記事とその関連特許の間のリンク付けを行うものであり，製品・特許の特徴レベルに踏み込む形で対応付けを行うものではなかった．それらに対して本研究では，製品と特許それぞれの特徴レベルでの関連性を自動的に判定することができる点が異なる．

3　製品発表プレスリリースからの製品特徴の抽出

3.1　手がかり表現の自動抽出手法

製品発表プレスリリースからの製品特徴を示すキーワードの抽出は,「に優れた」「を搭載」「が可能」といった手がかり表現を使用して抽出する．手がかり表現は，我々が以前行った業績発表記事からの業績要因表現の抽出手法 [4] を用いて行った．本節では，この手がかり表現の抽出手法を簡潔に説明する．

以下に，手がかり表現自動抽出手法の概要を示す.

Step 1: 少数の手がかり表現を人手で与え，それに係る節を取得する．

Step 2: 取得した節の集合から，その中で共通して頻繁に出現する表現を「共通頻出表現」と定義し，それを抽出する[1]．

Step 3: 共通頻出表現が係る節を取得し，その中から新たな手がかり表現を抽出する．

Step 4: 獲得した手がかり表現から，それに係る節を取得する．

Step 5: Step 2 から Step 4 を，新たな手がかり表現と共通頻出表現が獲得されなくなる，もしくは，予め定めた回数まで繰り返す（図 1 を参照）．□

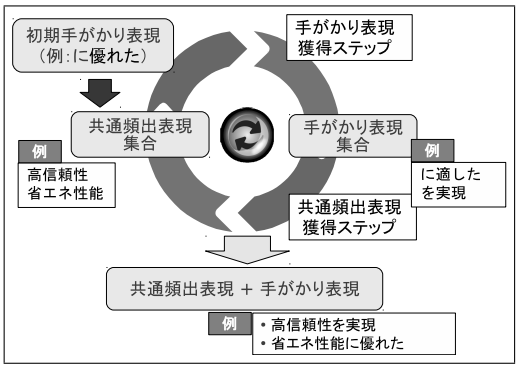

図 1: 手がかり表現自動抽出手法の概要

Step 2 では，適切な共通頻出表現を選別する．具体的には，様々な手がかり表現に係っている共通頻出表現は適切であるという仮定に基づき，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを式 1 で求め，その値がある閾値以上の共通頻出表現を選別する．

$$H(e) = -\sum_{s \in S(e)} P(e,s) \log_2 P(e,s) \qquad (1)$$

ただし，製品発表プレスリリースの集合において，$P(e,s)$ は共通頻出表現 e が手がかり表現 s に係る確率，$S(e)$ は共通頻出表現 e が係る手がかり表現の集合である．Step 3 にて，その選別した共通頻出表現から新たな手がかり表現を獲得するが，ここでも，様々な共通頻出表現が係っている手がかり表現は適切であるという仮定に基づき，手がかり表現候補に対して共通頻出表現が係る確率に基づくエントロピーを求め，閾値以上の候補を手がかり表現として抽出する．

3.2 手がかり表現自動抽出手法の実装

本研究では，上記の手法を実装し，さらに，ノイズとなる手がかり表現を排除する機能を追加したツール[2]を用いて，製品発表プレスリリースからの手がかり表現の自動抽出を行った．製品発表プレスリリースは，日経プレスリリース[3]にて発表される製品発表プレスリリースを使用した．本ツールでは，ポジティブリストにて初期手がかり表現を，ネガティブリストにて排除する表現を指定することができる．本手法にて入力した初期手がかり表現を示したポジティブリストを図 2 に示す[4]．初期手がかり表現は「[Particle] タグで指定された格助詞リスト＋ [Clue] タグの表現」となる．また，本手法

```
[Particle] が, を, に
[Common] 性能, 機能
[Clue] 搭載，図れます。，楽しむ，優れた，図りました。，向上，図る，対応した，お楽しみいただけます。，楽しめる，備えた，図ります。，楽しめます。，対応，可能，実現，高い，高め
```

図 2: 本手法におけるポジティブリスト

にて入力した排除する表現を示したネガティブリストを図 3 に示す．[Particle] タグで指定された格助詞を先頭にもつ手がかり表現は全て排除される．また，「全ての格助詞＋ [Clue] タグで指定された表現」の手がかり表現を排除する．本ツールにより，1974 個の製品発表プレスリリースから 103 個の手がかり表現を得た．初期手がかり表現に含まれる表現以外では，「に適した」，「を強化する」，「を活用した」といった手がかり表現が新たに獲得された．

```
[Particle] と
[Common] 事が，ことが，こと，ことなど
[Clue] した，向けて，向けた，なります，行い，行う，はじめ，行います，する，なっています，し、，なる，あります
```

図 3: 本手法におけるネガティブリスト

3.3 手がかり表現を使用した製品特徴の抽出

手がかり表現を使用して製品特徴を表すキーワードを抽出する．ここでは，手がかり表現の先頭の助詞を削除した文字列（「に優れた」ならば「優れた」）に係っている文節を取得し，文節の最後尾の助詞が「を」「が」「に」「で」である場合に，その助詞を除去した文字列をキーワードとして抽出した．例えば，「新開発「スッキリするん棚」で冷蔵室の収納性を向上」という文からは「を向上」という手がかり表現を使用して「スッキリするん棚」，「収納性」を抽出し，また，「独自の省

[1] 共通頻出表現は手がかり表現と同様，手法を適用するタスクによって異なるが，本タスクでは「安全性」「高信頼性」「精度」といった表現が共通頻出表現となる．

[2] http://www.ci.seikei.ac.jp/sakai/clupes.html

[3] http://release.nikkei.co.jp/

[4] 実際はコンマ区切りではなく，改行で区切る．

エネ技術「ｅーＣＯＯＬシステム」で高い省エネ性を実現する」という文からは「を実現」という手がかり表現から「ｅーＣＯＯＬシステム」,「省エネ性」を抽出する．

4 製品発表プレスリリースと特許との関連性の判定

製品発表プレスリリースから抽出した製品特徴を表すキーワードを使用して，その製品と関連性のある特許を検索する．具体的には，特許明細書の「請求項」「発明の属する技術分野」「発明が解決しようとする課題」「課題を解決するための手段」「発明の効果」を対象として，以下の式 2 で，製品発表プレスリリース p と特許明細書 a との関連度 $Re(p,a)$ を求める．そして，関連度が高い特許明細書を，その製品発表プレスリリースと関連性があると判定する．

$$Re(p,a) = -\sum_{k \in K(p)} W(k,p) \cdot W(k,a) \qquad (2)$$

ここで，$K(p)$ は製品発表プレスリリース p から抽出した製品特徴を示すキーワード集合，$W(k,x)$ は $x = \{p \text{ or } a\}$ においてのキーワード k の重みで，以下の式 3 で計算される．

$$W(k,x) = tf(k,x)df(k,A(p))\log(|N|/adf(k,N)) \quad (3)$$

ここで，

$tf(k,x)$ ：x において，キーワード k が出現する頻度．

$df(k,A(p))$ ：p から抽出した製品特徴を示すキーワードを，いずれか 1 つ以上含む特許明細書の集合 $A(p)$ において，キーワード k を含む特許明細書の数．

$adf(k,N)$ ：全体の特許明細書集合 N において，キーワード k を含む特許明細書の数．

上記の式 3 によって，特許明細書 a に多く出現し，p から抽出した製品特徴を示すキーワードをいずれか 1 つ以上含む特許明細書の集合 $A(p)$ において多く出現し，さらに，全体の特許集合においては多く出現していないキーワードに対して高い重みを付与する．

5 実装

本手法を実装し，製品発表プレスリリースを入力として，その製品発表プレスリリースと特許との関連度を計算するシステムを作成した．本システムでは，図 4 のように，対象とする企業名，対象とする製品，プレスリリース本文を入力する．システムは，まず，入力された製品発表プレスリリースから製品特徴を示すキーワードを抽出する．次に，対象企業が出願した特許のうち，対象とする製品が記述されている特許を検索する．検索された特許のなかで，製品特徴を示すキーワードを使用して関連度を計算する．図 5 に，冷蔵庫のプレスリリース「シャープ、「プラズマクラスター見守り運転」搭載の冷蔵庫 3 機種を発売」から検索された特許の一部を示す.「省エネ性」,「収納性」等のキーワードがプレスリリースから抽出され，そのキーワードを使用して計算されたスコア（青文字）が関連度を表す．現在のところ，計算された関連度の上位 20 の特許を出力する．実装にあたり，形態素解析器として MeCab[5]，係り受け解析器として CaboCha[2] を使用した．

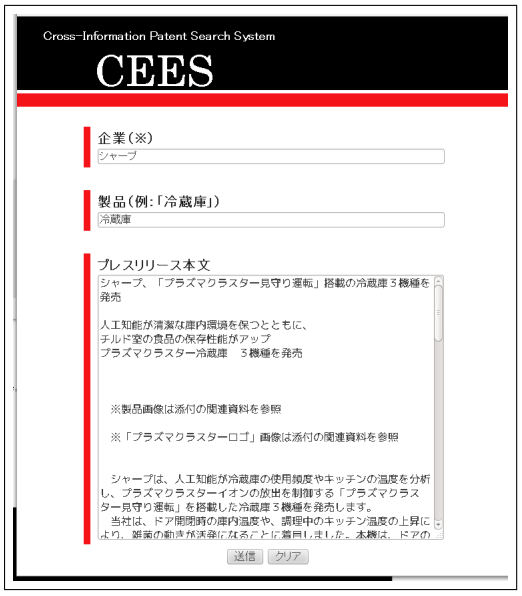

図 4: 製品発表プレスリリースからの特許検索システム

6 評価

本手法の評価を行った．評価では，特許出願の特に多い 21 の企業の特許を対象とし，システムに対して実際に製品発表プレスリリース（P1〜P7）を入力して出力される上位 n 件の特許を，入力した製品発表プレスリリースと関連がある特許と判定し，その場合における精度を測定した．以下に，評価に使用した製品発表プレスリリースの一部を示す．

P1 ：シャープ、「プラズマクラスター見守り運転」搭載の冷蔵庫 3 機種を発売

P3 ：ＮＥＣ、必要な文書を短時間で発見できる意味検索エンジンを開発

[5]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

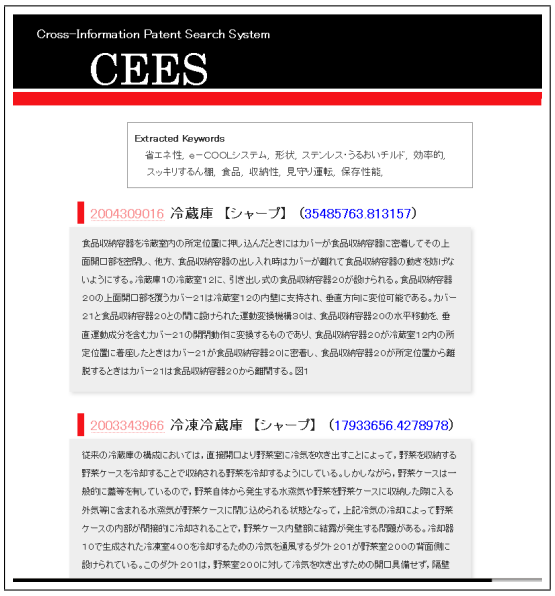

図 5: 製品発表プレスリリースから検索された特許

P7 ： 三菱電機、高精度金型加工などの幅広い用途に適応するワイヤ放電加工機を発売

表 1 に評価結果を示す.

表 1: 製品発表プレスリリースによる精度（%）

n	P1	P2	P3	P4	P5	P6	P7
1	100	100	100	100	100	100	100
3	100	100	66.7	100	100	66.7	100
5	100	60	80	80	80	60	80
10	80	40	40	40	50	60	60

7 考察

表 1 より，$n = 5$ までは，高い精度を達成することができたが，$n = 10$ では精度が低下している．現在は，関連度の上位 n 件の特許を全て出力しており，たとえ，関連度が低くても検索された特許が少ない場合は出力されてしまう．そこで，関連度に閾値を導入し，関連度が閾値より低い特許を出力しないようにする必要があると考える．

より高い精度を達成するための方法として，関連度を計算するための文の対象を絞り込む方法が考えられる．現在は，特許明細書の「請求項」「発明の属する技術分野」「発明が解決しようとする課題」「課題を解決するための手段」「発明の効果」を対象としているが，特許のもたらす効果とは無関係である記述も多い．そのような記述は，関連度を計算するための情報として排除するべきであると考える．そこで，例えば，文献 [5]，[6] の手法を特許明細書に適用して，特許発明を使用することによりもたらされる効果が記述してある表現を抽出し，この部分を対象とすることで，さらなる精度の向上が達成できると考える．

8 まとめ

本稿では，製品発表プレスリリースと，その製品と関連のある特許を特徴レベルで判定する手法について述べた．まず，製品発表プレスリリースから，その製品の製品特徴となるキーワードを抽出し，抽出した製品特徴キーワードを使用して製品発表プレスリリースと特許との関連度を計算した．ここで，キーワードは「に優れた」といった手がかり表現を使用して抽出し，その手がかり表現も初期手がかり表現から自動的に抽出した．評価の結果，関連度の上位 5 件までは高い精度を達成したが，上位 10 では高い精度を達成できなかった．そこで，文献 [5] の手法等を併用して，より高い精度の達成を目指す．

参考文献

[1] Itoh, H., Mano, H. and Ogawa, Y.: Term Distillation for Cross-DB Retrieval, *Working Notes of the 3rd NTCIR Workshop Meeting, Part III : Patent Retrieval Task*, pp. 11–14 (2002).

[2] 工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002).

[3] Nonaka, H., Kobayashi, A., Sakaji, H., Suzuki, Y., Sakai, H. and Masuyama, S.: Extraction of Effect and Technology Terms from a Patent Document, *Journal of Japan Industrial Management Association*, Vol. 63, No. 2E, pp. 105–111 (2012).

[4] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968 (2008).

[5] 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).

[6] 坂地泰紀, 野中尋史, 酒井浩之, 増山繁: Cross-Bootstrapping: 特許文書からの課題・効果表現対の自動抽出手法, 電子情報通信学会論文誌 D, Vol. J93-D, No. 6, pp. 742–755 (2010).