製品情報のプレスリリースとその製品特徴に関連した特許文書との対応付け

藤村 真太郎 †　野中 尋史 ‡　酒井 浩之 §　増山 繁 §
豊橋技術科学大学大学院 知識情報工学専攻 †
豊橋技術科学大学大学院 電子・情報工学専攻 ‡
豊橋技術科学大学大学院 情報・知能工学専攻 §
{fujimura, nonaka, sakai}@smlab.tutkie.tut.ac.jp masuyama@tut.jp

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となる．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は以下の点で有用となる．

1. 他社の人気製品の特徴分析とさらにそれを実現する特許の把握により，人気製品の重要な特徴に使用されている重要特許が特定できる
2. 他社の製品に関しその特徴毎のパテントポートフォリオ（特許集合）を分析でき，特許レベルにおいての詳細な製品開発戦略の特定等が可能となり，自社の開発戦略の意思決定に役立つ
3. 特許の特徴と製品の特徴をリンクさせることで，特許レベルの特徴の製品群への寄与が分かり，様々な製品に使われる基本特許の特定が容易になる

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，特許と製品レベルの関連付け，特に特徴レベルでの関連付けを行う実用的な情報システムの開発が求められているが，実用的な手法は提案されていなかった．

そこで，本研究では，企業が発表するプレスリリースに製品の特徴情報が含まれること，さらに，特許側においても，技術の特徴となる表現が含まれることに着目し，まず，プレスリリースと特許の関連付けを行った上で，関連するプレスリリースと特許の対に対して，製品・特許のそれぞれの特徴を抽出し，その類似性を計ることで，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手法の基礎的検討を行った．

2　関連研究

特許情報処理に関連する研究としては，特許翻訳に関するもの [1] や特許情報を可視化したグラフであるパテントマップを自動生成 [2, 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し，技術動向を調査するタスクが行われた [4]．しかし，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．

酒井ら [5] は特許文書中から手がかり語を自動的に収集した上で技術課題情報を抽出する手法を提案しているが，特許明細書中の発明の効果タグに抽出対象を限定していた．本研究では，当該手法を全文に拡張して用いて，特許文書中の特徴を抽出した．

3　製品と特許の対応付け

本研究における製品特徴とは,「○×により省エネを実現しました。」や「□△脱臭装置を搭載。」のような製品に関する情報のことを示す．一方，特許文書の特徴は,「冷蔵庫内を脱臭できる。」といった効果が記載されている箇所と定義する．また，特許文書の特徴は「発明の効果」タグに記載される内容だけでなく，特許中に含まれる付随する効果もすべて含んだものとする．

本研究では，この製品特徴と特許文書中の特徴を以下の手順で対応付ける．

製品と特許の対応付け手順

Step 1. 製品発表プレスリリースの特徴ベクトルと特許文書の特徴ベクトル間の類似度を求め，閾値を越えた対を製品に関連のある特許文書として集める．

Step 2. 関連があるとされた文書対同士の特徴文の対応を取るために，製品発表プレスリリースと特許文書それぞれに対し，手がかり表現を用いて特徴文を抽出する．

Step 3. 抽出された特徴文間の類似度を求め，製品発表プレスリリース中の特徴文と特許文書中の特徴文のうち最大の類似度となり，かつ，閾値を越えた対を対応関係のある特徴文対として獲得し，製品と特許の対応付けを行う．

□

3.1 文書間の対応付け

文書間の対応付けは，製品発表プレスリリースと特許文書間の特徴ベクトルの類似度を求めることで実現する．各文書の特徴ベクトルには，各文書の全体から語を獲得し，tf-idf に基づく重みを割り当てる．このときの idf は対応付けを行う製品発表プレスリリースと同一のカテゴリである特許文書集合（今回は IPC 分類が F25D であり，かつ冷蔵庫の特許文書の集合）を用いて計算する．特徴ベクトルに使用する品詞は「名詞」と「形容詞」のみに限定する．カタカナ語に関しては,「コンプレッサー」と「コンプレッサ」のように表記揺れが見られる．そのため，これらの末尾の「ー」は取り除く．

また，一般的な製品発表プレスリリースや特許文書に高頻出の語（例：発明，前記，手段，販売）は対応付けに寄与しないノイズとなることが考えられるため，ストップワードに登録する．このストップワードリストの生成には，タイトルに「販売」や「発売」を含む製品発表プレスリリース[1]から無作為に抽出された 10,000 件から構成されるプレスリリース文書集合と，2002 年に公開された特許文書から無作為に抽出された 10,000 件から構成される特許文書集合を用い，その中でエントロピーが一定値以上のものをストップワードに加える．

以上の過程を経て生成される製品情報のプレスリリースの特徴ベクトル $\boldsymbol{q}$ と特許文書の特徴ベクトル $\boldsymbol{d}$ 間の次式で表される cos 類似度を求める．この類似度が閾値を越えたものを製品情報のプレスリリースに関連する特許文書とする．

$$sim(\boldsymbol{q}, \boldsymbol{d}) = \frac{\boldsymbol{q} \quad \boldsymbol{d}}{\|\boldsymbol{q}\|\|\boldsymbol{d}\|} \tag{1}$$

3.2 特徴文間の対応付け

文書間で対応付けられた製品発表プレスリリースと特許文書の対に含まれる特徴文の対応付けを行うために，製品発表プレスリリースと特許文書の特徴文の抽出を行う．製品特徴を含む文（以下，製品特徴文）の抽出には，酒井らの製品発表プレスリリース中から製品特徴を抽出する手法 [6] で獲得された手がかり表現である，文末用手がかり表現，文中用手がかり表現，共通頻出表現をすべて用い，製品特徴文を抽出する．

また，特許文書の特徴を含む文（以下，特許特徴文）の抽出には，特許文書中から技術課題情報を抽出する手法 [5] で獲得された手がかり表現を利用する．酒井ら [5] は「発明の効果」タグから技術課題情報の抽出を行っていたが，本研究では，抽出対象を特許文書全体とする.「発明の効果」に記載されるような効果だけでは，例えば，脱臭装置の製造方法に関する特許では「発明の効果」タグのみを対象とする場合,「製造時のコスト削減ができる」のような特徴しか獲得できない．しかしながら，本研究においては,「発明の効果」以外に現れる「脱臭装置を用いることで，庫内を清潔に保つことができる」といった表現が重要になるため，抽出対象を特許全体とした．

文書間の対応関係があるとされた文書対はそれぞれの特徴文同士の cos 類似度を求め，製品特徴文 $\boldsymbol{q_i}$ と特許特徴文 $\boldsymbol{d_j}$ の最大の cos 類似度をとり，かつ閾値を越えた対を対応関係のある特徴文とする．

$$sim(\boldsymbol{q_i}, \boldsymbol{d_j}) = \frac{\boldsymbol{q_i} \quad \boldsymbol{d_j}}{\|\boldsymbol{q_i}\|\|\boldsymbol{d_j}\|} \tag{2}$$

また，文間の特徴ベクトル生成時には，製品特徴抽出用の手がかり表現と特許文書中の特徴を抽出する手がかり表現に含まれる語を文書間の対応付けの際に作られたストップワードに追加する．

[1] 日経プレスリリースから取得．http://release.nikkei.co.jp/

4 実験と結果

4.1 実験用コーパス

家電メーカー A 社と B 社の冷蔵庫の製品発表のプレスリリースと特許文書を対象に実験用コーパスの作成を行った．製品発表のプレスリリースは，プレスリリースのタイトルに「冷蔵庫」が含まれており，プレスリリースの発表日から遡って 120 日以内に出願され，かつ，1999 年から 2006 年に公開された特許文書を持つものを収集した．このとき，冷蔵庫の製造工場の立ち上げなどに関するプレスリリースも得られたが，本研究の目的からは無関係であるため，除去した．

特許文書は，1999 年から 2006 年に公開され，IPC 分類が F25D であるもののうち，本文中に 1 回以上「冷蔵庫」が登場する語である．加えて，プレスリリースの発表者と同一の企業名で候補を限定した．今回対象とした家電メーカーでは，冷蔵庫の開発が子会社などで行われている可能性も存在するため，例えば，「○×電機」という家電メーカーであれば，グループ会社の特許文書も候補とするために，「○×」を含む社名であるか否かを基準に選別した．

冷蔵庫のような製品では，製品発表のプレスリリースが発表された日の近傍に特許出願されている可能性が高い．そこで，今回は，プレスリリースの発表日を基準日とし，120 日前までの特許文書とそのプレスリリースの組について正解データを作成した．こうして得られた実験用のデータセットの概要は以下のとおりである．

表 1: 実験用データセット

	A 社	B 社
特許文書数	780	453
プレスリリース数	10	13
平均特許候補数	32.2	15.2

4.2 評価実験

作成したコーパスを対象に実験を行った．文献 [6] における製品特徴文を抽出するための手がかり表現と文献 [5] における特許特徴文を抽出するための手がかり表現を獲得する際の閾値は 0.4 とした．また，製品特徴文の抽出に使用する文中手がかり表現は上位 20 個を使用した．形態素解析器には MeCab[2]，係り受け解析器には CaboCha[3]を使用した．

以上の設定に基づき，文書間の対応付けと特徴文間の対応付け実験の結果は以下のとおりとなった．類似度の閾値は両方のデータセットの F 値が高くなるように設定した．文書間の対応付けの閾値は 0.025，特徴文間の対応付けは 0.10 とした．

表 2: 文書間の対応付けの結果

	A 社	B 社
精度	0.606	0.600
再現率	0.518	0.622
F 値	0.558	0.609

表 3: 最終結果（文書対応と文対応両方）

	A 社	B 社
精度	0.453	0.405
再現率	0.349	0.333
F 値	0.395	0.366

5 考察

以上の実験の結果に基づき，考察を行う．文書間の対応付け結果，特徴文間の対応付け結果共に課題の残る結果となった．文書間の対応付け，特徴文間の対応付けでは共通の問題が見られた．

両者に共通した精度・再現率を下げる要因として考えられる点は，製品発表プレスリリースと特許文書間では異なる表記が行われるため，対応がある場合でも，現状では対応がないと判定されてしまうことがあるためである．例えば，特許文書では「観音開き式扉」と記載される一方で，製品発表プレスリリースでは「フレンチドア」のような記述が行われることがある．この表記の揺れは以下に分類することができる．

- 技術語（可燃性冷媒，イソブタン，ノンフロン，HC 系冷媒など）
- カタカナ語（コンプレッサ，圧縮機やドア，扉など）
- 造語（○×システム，□△テクノロジー）

文書間の対応付けの結果のうち，false negative に関して調査を行った．その他の原因とされているもの

[2]MeCab http://mecab.sourceforge.net/

[3]CaboCha http://chasen.org/ taku/software/cabocha/

表 4: 文書間対応付けで対応関係がなしとされた原因

技術語	カタカナ語	造語	その他の原因
8 文書対	8 文書対	17 文書対	24 文書対

は，特徴語が一致すると考えられるものが含まれているが，tf-idf による語の重み付けによって特徴語が過小に評価されてしまうことに起因すると考えられるもの（例えば，冷蔵庫の特許文書中では一般的な「圧縮（器）」や「製氷」などの語）や，複雑な冷蔵庫の構造を認識すること（プレスリリースでは,「冷蔵庫の左を冷凍温度ゾーンとして，右部分を冷蔵温度ゾーンとして」であり，特許文書では「冷蔵庫本体の幅方向の中間部に上下方向に延びる縦仕切を備え」のような特徴）が求められるようなことに起因しているものである．

以上の結果について，統計的にまとめたものを表 4 に示す．表 4 より，語が異なるために対応付けが不可能なものが半数を占めている．そのため，このような同一の概念を示すものをまとめることが精度及び再現率の向上につながることが考えられる．ただし，今回は冷蔵庫を用いたが，同一の概念を示す語のまとめ上げには対象製品によらない汎用的な手法が望まれる．この点を踏まえ，異なるコーパス間で使用される語彙の差を埋めることは今後の課題である．

また，特徴文間の対応付けに関しては，前記の語の問題だけでなく，特徴文が抽出されないことに起因するものが多数あった．これに関しては，製品特徴の手がかり表現は対象ごとに大きく異なることが考えられる．「保存できる」や「持ち運びできる」などのサ変名詞＋できる型の手がかり表現は「できる」という表現は分野によらない表現であるため，酒井ら [6] の手法によって獲得できる．しかしながら，これが「保てる」や「持ち運べる」などの形になってしまうと「できる」という表現より獲得することが大幅に難しくなる．このため，製品特徴抽出用の手がかり表現やそれを利用するパターンの拡充が必要となる．

一方，プレスリリースごとに精度や再現率を導出したところ，大きく結果がゆらぐことがわかった．これは，幅広い権利範囲を得るために同一の技術分野の特許は類似した内容のものが同時に複数出されやすく，1 つのエラーが複数の特許とプレスリリース間の対応付けに派生しやすいためである．すなわち，プレスリリースと類似する複数の特許に関連がある場合，プレスリリース中で「ノンフロン冷蔵庫」という特徴が存在し，複数の特許文書中で同一の概念でありながら異なる表現である「可燃性冷媒」と記載されると，プレスリリースと特許間に関連がないとされ，大幅に精度，再現率が低下してしまう．逆に，1 つでも上記に掲げた問題が解決されると，同時期の複数の特許文書とプレスリリースの間の対応付けが正確になり，大幅な精度，再現率の向上が期待できるといえる．

6 おわりに

本研究では，製品発表プレスリリースと特許文書中の特徴文の対応付けを試みた．まだ，精度，再現率共に低く，多くの課題が存在することが明らかになった．1 つ目は，類義語等の同一概念の語をまとめることである．2 つ目は，製品特徴抽出における精度と再現率の向上である．今後は，本研究によって明確になった課題の解決を通し，精度，再現率の向上を目指す．

謝辞

本研究は，日本学術振興会科研費 (C)22500129, 総務省戦略的情報通信研究開発推進制度（SCOPE）地域 ICT 振興型の支援を受けた．

参考文献

[1] 村上 仁一, “ルールベース翻訳と統計翻訳を統合した特許翻訳”, 第 1 回特許情報シンポジウム, 2010.

[2] Hirofumi Nonaka, Akio Kobayashi, Hiroki Sakaji, Yusuke Suzuki, Hiroyuki Sakai, Shigeru Masuyama, “Extraction of the e ect and the technology terms from a patent document”, The 40th International Conference on Computers & Industrial Engineering, Awaji Island, Japan, July, 2010.

[3] Yusuke Suzuki, Hirofumi Nonaka, Akio Kobayashi, Hiroyuki Sakai, Shigeru Masuyama, “Extraction of Technology Terms from Patent Speci cations for Technology-E ect Type Patent Map Generation”, ITC-CSCC 2010, pp.725-728, Pattaya, Thailand, July, 2010.

[4] Makoto Iwayama, Atsushi Fujii, Noriko Kando, Akihiko Takano, “Overview of Patent. Retrieval Task at NTCIR-3”, Proceedings of the 3rd NTCIR Workshop, Tokyo, Japan, 2002.

[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] 酒井 浩之, 増山 繁, “手がかり表現自動獲得による製品発表プレスリリースからの製品特徴の抽出”, 言語処理学会 第 17 回年次大会, 2011.