プレスリリースを用いた製品関連特許の IPC 推定

大田 仁克[†]　太田 貴久[‡]　酒井 浩之[*]　増山 繁[‡]

[†] 豊橋技術科学大学 工学部 知識情報工学課程

[‡] 豊橋技術科学大学大学院 工学研究科 情報・知能工学専攻

[*] 成蹊大学 理工学部 情報科学科

ohta@la.cs.tut.ac.jp

1　はじめに

企業が行う特許調査は，知的財産の管理・運用や，研究開発の戦略立案のために必要不可欠である．特許調査では，自社の保有する知的財産と関連する特許や，競合他社の開発した製品に使用されている特許を検索する．特に，製品に対して行われる特許調査では，他社の持つ権利を把握するために，実際に製品に使用されている特許だけでなく，製品に関連のある特許も検索対象となる．ここで，関連のある特許とは，製品に使用されている特許と同一の効果を持つ特許などの，製品に用いられる可能性のある特許である．一般的な特許調査では，発明内容を端的に表すキーワードとその関連語，及び，IPC コードなどの，発明に用いられる技術によって特許を分類した特許分類コードを用いて特許を検索する [1].

しかしながら，1 つの製品には複数の技術分野にまたがる特許が数多く関連しており，これらの特許をすべて検索するための検索式を構築することは専門家であっても労力を要する．そのため，製品情報から特許分類コードを推定することができれば，特許調査に必要な検索式を構築することができ，製品に関連のある特許の検索が容易となると考えられる．そこで，本研究では，製品情報としてプレスリリースを用い，そこから特許分類コードの 1 つである IPC(International Patent Classi cation) コードを推定する手法の提案と検討を行う．

2　関連研究

特許情報処理の研究は，特許翻訳に関するもの [2] 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するもの [3] などが中心となっていた．一方，NTCIR-3 では，新聞記事に掲載された技術や商品について，関連する特許の検索を行う特許検索タスク [4] が行われた．このタスクにおいて，Itoh ら [5] は，新聞と特許という 2 つのコーパス間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．本研究では，文書間の対応付けを行わず，プレスリリースから IPC コードの推定を行う．

酒井ら [6] は,「ができる」のような手がかり表現を使用することで,「3D 画像を容易に作成することができる」といった特許の技術特徴を表す文を特許明細書から抽出する手法を提案している．また，酒井ら [7] は，[6] の手法を改良して,「高画質を実現しました」といった製品特徴を表す文をプレスリリースから抽出する手法を提案している．これらの手法を用いて，藤村 [8] は，プレスリリースと特許の特徴レベルでの対応付けを行っている．しかしながら，[8] は，対象製品を冷蔵庫に限定した上で，なお，対応付けの精度が F 値 0.5 以下であり，プレスリリースと特許の対応付けが困難であることを示している．本研究では，特許分類コードである IPC コードを推定することで，プレスリリースと特許の対応付けのための検索式の構築を目指す．

3　IPC

IPC コードとは，国際的に統一された特許の技術内容による分類コードであり，全ての特許に 1 つ以上付与されている．また，1 つの特許に複数の IPC コードを付与することが可能である．分類の構造を図 1 に示す．IPC は，技術内容を「セクション」「クラス」「サブクラス」「メイングループ」「サブグループ」と階層的に分類する．例えば IPC コード「G10C1/04」が現わす技術分野は,「音を発する装置であるピアノの全体構造の中でも，グランドピアノのもの」である．本研究では，IPC サブクラスのコードまでの推定を行う．

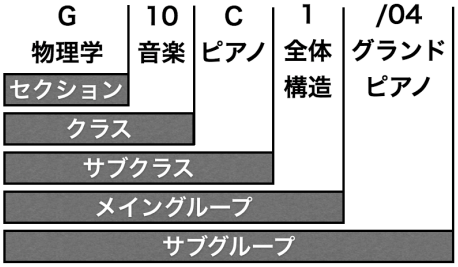

図 1: IPC の構造

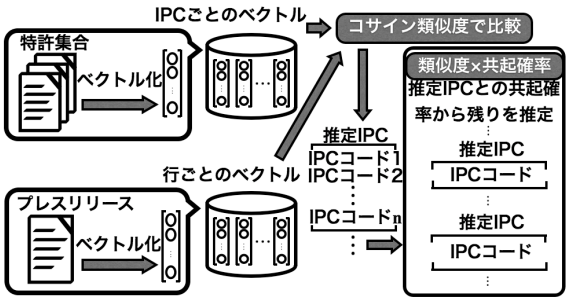

図 2: 提案手法

4 プレスリリース

プレスリリースとは，企業から報道機関向けに発表される製品や技術動向などの情報のことである．プレスリリースは，新製品に関する情報の中で，最も早く入手可能なものであり，製品特徴がまとまって記述されている．本研究では，製品の特徴を表す文に含まれる表現を利用して，IPC コードを推定する．

5 提案手法

提案手法は，製品発表に関するプレスリリースから，製品に使用されていると思われる特許（公開特許公報）の IPC サブクラスのコードまでを推定する．以下，特許中の【＊＊＊】タグが付与された文書を【＊＊＊】と略記する．

本手法では，プレスリリース全文と，特許の【発明の名称】及び【発明の効果】との比較を行う．【発明の名称】は発明内容を簡潔に表した特許のタイトルであり，製品名を含む可能性が高い．また，【発明の効果】には，特許の技術的特徴とその技術によって得られる優れた効果が記述されており，これらの形態素はプレスリリースにも含まれる可能性が高い．なお，名詞のみを形態素として使用する．

前処理として，IPC サブクラスのコードごとに特許をまとめ，特許集合を作成する．次に，特許集合中の特許から【発明の名称】及び【発明の効果】を抽出し，形態素解析を行う．形態素解析結果から，形態素の和集合を用いて特許集合ごとに TF-IDF ベクトルを作成する．

本手法の具体的な手順を以下に，概要図を図 2 に示す．

Step 1. プレスリリース全文を形態素解析し，プレスリリースの TF 値と特許集合での IDF 値を用いて，プレスリリースの文ごとに TF-IDF ベクトルを作成する．

Step 2. プレスリリースの m 文目 (初期値 m=1) のベクトルと各特許集合のベクトル間のコサイン類似度を求め，しきい値 θ_1 以上の IPC コードを取得する．

Step 3. 取得した IPC コードのうち，類似度の上位 n 件の IPC コードを推定 IPC コードとし，これらと特許中で共起している IPC コードの共起確率を求める．

Step 4. 取得した IPC コードのうち，n+1 件以降の IPC コードのコサイン類似度と共起確率の積を求め，しきい値 θ_2 以上の IPC コードを推定 IPC コードとする．

Step 5. m=m+1 とし **Step 7.** に戻る．m 文目が末文の場合，推定 IPC コードを出力して終了する．

Step 1. では，プレスリリースに出現した形態素の TF 値と，特許集合での IDF 値を用いてベクトルを作成している．すなわち，プレスリリースと特許集合の双方に出現する形態素のみに，IDF による重みが付与される．これは，プレスリリースのみに現れる特徴的な語 (企業名や商標) の影響を抑え，発明に関する語のみに焦点を絞るためである．

Step 2. では，プレスリリースの文ごとに TF-IDF ベクトルを作成している．プレスリリース中には，製品の特徴を表す文が記述されているため，文ごとにベクトルを作成することで，特徴ごとに特許との比較を行うことができる．

Step 3. 及び **Step 4.** では，コサイン類似度の上位 n 件の IPC コードが正解していると仮定して，それらの IPC コードと特許中で共起している IPC コードを併用して IPC コードを推定する．予備的検討として，プレスリリース全文の TF-IDF ベクトルと特許集合のベクトルを比較したところ，コサイン類似度の最上位

は 97.8%の精度で正解の IPC コードであった．積算に用いる共起確率は，類似度がより高い IPC コードでの値を用いる．これにより，n+1 件以降の IPC コードにフィルタをかけ，適合率の向上を図る．

6 実験と結果

6.1 学習用コーパス

1996 年から 2006 年までに出願された特許 3,875,712 件から抽出した【発明の名称】及び【発明の効果】を，IPC コードごとにまとめ特許集合を作成した．また，日本経済新聞電子版[1]から，2011 年 10 月から 2012 年 5 月までに発表されたプレスリリース 13,528 件を取得した．

6.2 正解データ

取得したプレスリリースから選択した製品発表に関するプレスリリース 50 件に対し，製品に関連する特許の IPC コードを人手で付与し正解データを作成した．IPC コードは，公開特許公報に製品名で検索をかけた結果に基づき，プレスリリース 1 件当り平均 8.7 個付与した．また，正解データの検証は工学部学生 1 名で行った．

6.3 評価実験

提案手法の評価実験を行った．システム開発は Ruby で行い，形態素解析器には MeCab[2]を使用した．また，MeCab のシステム辞書には Unidic[3]を使用した．取得した IPC コードのうち，類似度が最上位の IPC コードを推定 IPC コードとした．各しきい値での実験結果を表 1，表 2，表 3 に示す．θ_2 の値に関わらず，θ_1 の上昇に比例して，再現率が低下し適合率が向上した．また，F 値は，表 1 と表 2 の実験で θ_1=0.09 のとき最大となり，表 3 の実験で θ_1=0.08 のとき最大となった．

表 1: 実験結果 (θ_2=0.0)

θ_1	再現率	適合率	F 値
0.03	0.941	0.120	0.209
0.04	0.901	0.171	0.277
0.05	0.816	0.211	0.322
0.06	0.742	0.254	0.360
0.07	0.691	0.301	0.396
0.08	0.602	0.348	0.410
0.09	0.527	0.411	0.421
0.10	0.433	0.471	0.405

表 2: 実験結果 (θ_2=0.0001)

θ_1	再現率	適合率	F 値
0.03	0.903	0.167	0.275
0.04	0.863	0.212	0.327
0.05	0.787	0.245	0.358
0.06	0.722	0.281	0.385
0.07	0.673	0.318	0.409
0.08	0.590	0.365	0.422
0.09	0.513	0.431	0.429
0.10	0.424	0.482	0.407

表 3: 実験結果 (θ_2=0.001)

θ_1	再現率	適合率	F 値
0.03	0.772	0.222	0.332
0.04	0.758	0.253	0.362
0.05	0.696	0.273	0.373
0.06	0.653	0.314	0.401
0.07	0.610	0.359	0.420
0.08	0.545	0.399	0.426
0.09	0.485	0.441	0.424
0.10	0.403	0.511	0.417

7 考察

7.1 実験結果に関する考察

実験結果から，θ_2 の上昇に比例して，再現率が向上し適合率が低下することが分かる．これは，θ_1 にも言

[1]http://www.nikkei.com/
[2]http://mecab.sourceforge.net/
[3]http://www.tokuteicorpus.jp/

えることであり，しきい値の上昇に伴い推定 IPC コードの個数が減少したことが原因だと思われる．

表 1 の実験では，コサイン類似度の最上位は 95%以上の確率で正解していた．これは，例えば冷蔵庫に関するプレスリリースでは，形態素「冷蔵庫」の出現頻度が高く，冷蔵庫に対応した IPC コード F25D「冷蔵庫，冷凍室，アイスボックス」がよく推定されたことが原因だと思われる．表 1 の実験のエラー解析を行ったところ，例えば家庭用エアコンに関するプレスリリースでは，IPC コード B60H「車両の暖房，冷房，換気に関する装置」がコサイン類似度の上位 2 件目に出力された．これは，IPC コード B60H でまとめられた特許から抽出した形態素「エアコン」の TF-IDF による重みが，比較的大きいことが原因と思われる．このような現象は，表 2 と表 3 の実験にも見られた．

表 2 と表 3 の実験では，取得した IPC コードのうち，類似度が最上位の IPC コードを推定 IPC コードとした．表 1 の実験結果から，類似度が最上位の IPC コードはほぼ正解している．しかしながら，表 1 の実験結果と比較して，再現率が大きく低下していることが分かる．そのため，しきい値 θ_2 によって正解 IPC コードが除外されていると言える．言い換えると，類似度が最上位の IPC コードを推定 IPC コードとした場合においては，特許中でよく共起している IPC コードが必ずしも製品に関連のあるものであるとは言えない．

7.2 正解データに関する考察

正解データの検証を行った結果，同じ製品に関するプレスリリースでは，発表元の企業が異なる場合においても，正解の IPC コードが多数競合した．製品に固有の技術的特徴がある場合，その機能を実現するために使用されていると思われる特許が検索される場合が多かった．しかしながら，検証を工学部学生 1 名で行っているため，知財に明るい者に再検証を依頼する必要があると思われる．

8 おわりに

本研究では，製品発表に関するプレスリリースから，製品に使用されている特許の IPC コードを推定する手法の提案及び検証を行った．検証の結果，コサイン類似度が最上位に出力される IPC コードが正解となる確率が高いことがわかった．また，特許中でよく共起している IPC コードが必ずしも製品に関連のあるものであるとは言えないことがわかった．今後は類似度の上位 2 件以降を推定 IPC コードとした場合における提案手法の評価を行うと共に，正解データの再検証を行っていきたい．

参考文献

[1] 増満 光, 谷川 英和, 渡辺 俊規, “外国語検索式の作成支援システムの提案”, 日本知財学会 第 10 回年次学術研究発表会, 1B4, 2012.

[2] 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.

[3] Hirofumi Nonaka, Akio Kobayashi, Hiroki Sakaji, Yusuke Suzuki, Hiroyuki Sakai, Shigeru Masuyama, “Extraction of the effect and the technology terms from a patent document”, The 40th International Conference on Computers & Industrial Engineering, Awaji Island, Japan, 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] Hideo Itoh, Hiroko Mano, Yasushi Ogawa, “Term Distillation for Cross-DB Retrieval”, Working Notes of the 3rd NTCIR Workshop Meeting, Part III : Patent Retrieval Task, pp.11-14, 2002.

[6] 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.

[7] 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 回年次大会発表論文集, pp.528-531, 2011.

[8] 藤村 真太郎, “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の研究”, 豊橋技術科学大学大学院 工学研究科 知識情報工学専攻修士論文, 2011.