特許文献中の重要語を用いたFターム自動付与

佐々木 深　　綱川 隆司　　西田 昌史　　西村 雅史
静岡大学大学院総合科学技術研究科情報学専攻
gs15021@s.inf.shizuoka.ac.jp , {tuna, nishida, nisimura}@inf.shizuoka.ac.jp

1 はじめに

特許出願の際には，先行特許に関する調査が出願者や審査官によって行われるが，そこでは関連する特許文献の効率的な検索方法が必要とされている．特許検索は，キーワード検索のようなテキストによる検索と特許文献に付与される特許分類によるインデックス検索の 2 つに大別できる．前者は，単語の入力のみであるため検索者が使いやすく，最新の技術用語にも対応可能という利点があるが，表記揺れや多義性を持つ単語への対応が必要であり，ノイズや検索漏れが多いという欠点がある．一方後者は，人間が文献を精読して付与した IPC（国際特許分類），FI，F ターム等の特許分類の情報を利用することで前者よりも高精度での検索が可能である[1].

特許分類の付与作業においては，効率化のために最初に粗い分類を自動的に行っているが[2]，より細かい分類にも対応した高精度な自動分類システムが望まれる．既に，既存の特許文献に付与された分類を学習データとして用いる教師あり学習に基づく自動分類方法が提案されている[3-6]．しかし，特許分類によっては学習データ数が不十分あるいは存在しないといった課題があり，付与精度も向上の余地が残されている．

本稿では，従来の機械学習手法に加え，特許文献中の重要語に着目して F ターム自動付与の精度向上を図る．また，学習データが不足する特許分類に対し，各 F タームと特許文献間の類似度に基づく方法を提案する．

2 Fターム

日本で用いられている特許分類は，分類の粗い順に IPC，FI，F タームがある[1]．IPC は 1975 年から用いられている国際的に定められた世界共通の分類であり，階層的な体系を成している．全技術分野を段階的にセクション，クラス，サブクラス，メイングループ，サブグループの順に細分化されており，およそ 7 万の分類が存在する．

IPC は国際的に通用する分類であるため，各国の特許を効率的に検索できる一方で，国ごとに開発が活発な分野が異なる等の理由で，一部の分野に文献が集中し検索などに不都合が生じる場合がある．このことから，日本では IPC をさらに展開した索引として FI および F タームが用いられている．

FI は IPC に付加する形でサブグループの下位分類として展開記号・分冊識別記号を付与したものである．一方，F タームはさまざまな技術的観点（目的，用途，構造，材料等）によって分類する特徴を持ち，複数観点を組合せることで関連特許をより効率的に絞り込むことを目的に定められている． FI で定められる一定の技術範囲ごとに区分され，区分された各技術範囲は「テーマ」と呼ばれる．F タームは，各テーマで定義された複数の観点に対して割り当てられ，全テーマ約 2600 のうちおよそ 7 割にあたる約 1800 において F タームが作成されている．F タームは，テーマを表すテーマコード（英数字 5 桁），観点（英字 2 桁），数字（2 桁）で構成される．たとえば，テーマ“ハードウェアの冗長性”のテーマコードは 5B034 であり，4 つの観点“受動的冗長”，“能動的冗長”，“冗長回路”，“機能・構成”を持つ．各観点において展開される F タームを表 1 に示す．各 F タームの先頭のドット（・）は階層の深さを表しており，ドットの数が多いほど下位の階層を示す．表 1 で形成される階層構造を図 1 に示す．F タームは組合せて検索されることが想定されており，1 つの特許文献に対して同一テーマコードから展開される複数の F タームが付与される．

3 関連研究

NTCIR-5 および NTCIR-6 の特許検索タスクにおいて，F タームの分類に焦点を当てた特許文献の自動分類タスクが設けられ，テストコレクションが公開されている[7, 8]．NTCIR-6 分類タスクは，1993～1997 年に公開された日本公開特許公報の特許文献全文を学習データとし，1998～1999 年の特許文献 21606 件に対して F タームを付与する課題であり，6 グループが参加した．Li et al. [3] は特許文献の bag-of-words を素性として用いた SVM による識別器を用い，完全一致による評価で F 値 0.4125 の精度で最高性能を得た．Fujino and Isozaki [4] は，特許文献の各要素（発明の名称，出願人・発明者，要約書，特許範囲，明細書）について bag-of-words によるナイーブ

表 1　F タームリストの例

5B034	ハードウェアの冗長性						
観点	F ターム						
AA	AA00	AA01	AA02	AA03	AA04	AA05	…
	受動的冗長	・二重化	・・照合	・・・圧縮照合	・多重化	・・多数決	…
BB	BB00	BB01	BB02	BB03	BB04	BB05	…
	受動的冗長	・切替	・・予備切替	・・・共通予備	・・選択	・・・信頼度	…
		BB11	BB12	BB13		BB15	…
		・再構成	・・緊急制御回路	・・機番変更		・切離し	…
…	…	…	…	…	…	…	…

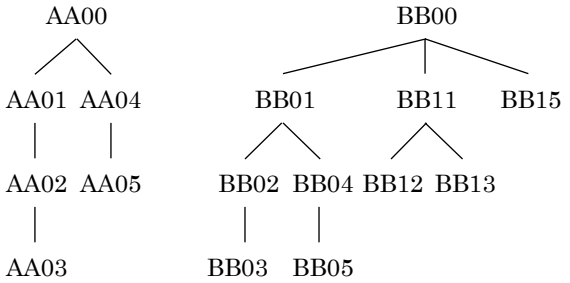

図 1　F タームリスト階層の例

ベイズの識別器を作り, これらを最大エントロピー法に基づき組合せる方法を提案した. Murata et al. [5] は k-NN 法に基づく方法を提案し，特許文献間の類似度として SMART [9] や BM25 [10] を用いて評価を行った.

小林 [6] は，F タームの付与根拠データ[1]を用いた結果をもとに, 付与根拠データが少ない分野に対して機械学習に基づき精度を向上させるため，tf-idf および分類の階層構造を利用した分類推定方法を提案した.

4　提案方法

本稿では, 特許文献に付与されるべき F タームを自動付与するため，既存の特許文献に人手で付与された F タームの情報を学習データとする教師あり学習手法を用いる. 関連研究において有効であった SVM（サポートベクタマシン）を用い[3, 8]，特許文献の bag-of-words に加え，特許文献に出現する重要語を素性として用いる. また，学習データが不足する F タームについて，word2vec を用いて各 F タームの名称・説明文と特許文献間との類似度を算出する.

4.1　用いる素性

(1)　出現語の bag-of-words（正規化 tf-idf）

形態素解析ソフト MeCab[2]を使用し，特許文献から名詞（非自立名詞，固有名詞，数は除く），自立動詞，自立形容詞，未知語を抽出する. これらに対し tf-idf で重みづけしたベクトルを求め, 正規化して各出現語に対する素性値とする.

(2)　特許文献を特徴づける重要語の出現有無

特許文献中には属する技術分野の専門用語が多く含まれており, 形態素解析によってこれらの単語を分割すると本来の意味が反映されないケースが想定される. そこで，重要語自動抽出モジュール TermExtract[3] [11] を使用し，特許文献を特徴づける重要語の抽出を行う. 抽出対象の重要語は，TermExtract によって計算される重要度が閾値以上のものとし，本研究では閾値を 20 とする. 以上によって得られた重要語の出現有無を素性とする. ここで, (1) の素性値とのバランスをとるため，(1)で求めた正規化ベクトルの各要素の平均値を算出し, その重要語が出現した場合はその平均値，出現しない場合は 0 を素性値とする.

4.2　学習アルゴリズム

1つの特許文献に対し複数の F タームが付与されるため，学習データに付与されているテーマコードに対し, 展開される F タームそれぞれに対する識別器をテーマコード別に学習する. すなわち，ある F タームの識別器を学習する際，対象 F タームが付与された文献を正例，展開元のテーマコードが付与され，かつ対象 F タームが付与されていない文献を負例とする. このとき，負例が圧倒的に多くアンバランスな学習データになるため, 正例はすべて採用し, 負例は最大で正例数の 2 倍の数だけランダムに選択する[3, 12]. また，カーネル関数にはシグモイドカーネルを使用する[3].

[1] 特許文献への分類付与者が，付与することとなった根拠箇所を明細書等から抽出したもの.

[2] http://taku910.github.io/mecab/

[3] http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

4.3 F ターム説明文を用いた特許文献との類似度推定

改正に伴う新設 F タームのように付与実績の少ないものは，学習データの不足という問題がある．これに対し，F タームと特許文献を直接比較し，スコアの高い F タームを付与する方法を提案する．

F タームの情報は，表 1 のような英数字の表記や名称の他に，短文の説明文がある．F タームの名称は簡略に表記されており，その文字列と特許文献とを直接対応付けするのは難しい．そのため，F タームの名称と説明文を用いて特許文献と比較する．さらに，word2vec[13]を用いて，F タームの名称･説明文の単語集合と特許文献の単語集合との類似度を算出し，スコアの高い F タームを付与する．

まず，F タームの名称・説明文および特許文献に対し MeCab による形態素解析を行い，単語集合を得る．得られた 2 つの単語集合の各要素から得られる単語対のコサイン類似度を word2vec により算出する．最後に，F タームの名称･説明文の単語集合の各要素における最大スコアの平均値を算出し，その平均値を当該 F タームと特許文献との類似度とする．

5 評価実験

評価実験として，比較的粗い分類であるテーマコードの付与，および各テーマコードから展開される F タームの付与に対し提案方法を適用した．

5.1 実験データ

NTCIR-6 のデータコレクションを使用し，1993～1997 年に発行された特許文献を学習データ，1998～1999 年に発行されたものをテストデータとする．テストデータに付与されるテーマコードは 108 種類あり，IPC のセクションレベル，すなわち最も粗い分類からランダムに選択されている．本実験では，これらのテーマコードの中から 20 個ランダムに選択し，そのテーマコードが付与された特許文献を対象とする．本実験に用いた 20 のテーマコードに含まれる文献数および 1 文献に付与された F ターム数を表 2 に示す．また，word2vec の学習には前記の学習データおよびテストデータを使用した．

5.2 テーマコード付与の実験結果

前節の実験データを使用し，テーマコード付与を行った．実験結果を表 3 に示す．使用したテーマコードは，「5B064（文字認識）」と「4L045（紡糸方法及び装置）」のように粗いレベルで異なるテーマコードが多いため識別が比較的容易であると考えられるものの，表 3 より，高精度で自動付与できることがわかる．本実験では，同一のテーマコードを使用するため，テーマコードごとに展開される各 F タームについて識別器を学習し，F タームの付与を行う．

表 2　実験に用いたデータ

訓練データの文献数		45631
テストデータの文献数		4012
1 文献に付与された F ターム数	平均	8.37
	最大	43
	最小	1

5.3 F ターム付与の実験結果

(1)　SVM を用いた教師あり学習手法

表 4 に，20 のテーマコードのいずれかを持つ特許文献全体に対する適合率，再現率，F 値を示す．素性として bag-of-words の tf-idf のみを用いた場合と比べ，重要語を素性として加えることで F 値の改善がみられた．また，20 のテーマコードのうち提案方法の F 値が最良だったものは 18 あった．

(2)　F タームと特許文献間の類似度に基づく手法

20 のテーマコードのうち「3C045（旋削加工）」を用いた予備実験を行った．当該テーマコードから展開される F タームは 197 あり，そのうち学習データにおいて正例が 5 以下の F タームは 18 あった．これら 18 の F タームのいずれかを持つ特許文献 5 件に対して，それぞれ全 197 の F タームとの類似度を算出する．表 5 に，実験対象の特許文献 5 件の中のある 1 件において本手法で得られた類似度が上位の F タームを示す．この文献には付与されるべき F タームが 25 個あり，上位 10 個の F タームには，付与すれば正しいものが 3 つ含まれており，上位 25 個では正しいものが 5 つ含まれていた．(1)での実験では，この文献にはいずれの F タームも付与されなかったため，上位数個を付与すればこの文献においては F 値が改善する．しかし，学習データにおける正例が 3 の F ターム「3C045DA20（・・多角形）」とは高い類似度を示さなかった．実験に使用した他の 4 件の特許文献においても，学習データの正例が少ない F タームは上位になかったため，F タームと特許文献を比較する際に単語のみではなく別の要素を加えて意味を反映する必要があると考えられる．

6 おわりに

本研究では，特許文献に対して特許分類 F タームを自動付与するため，特許文献中の重要語を用いた教師あり学習手法，および学習データにおける正例の少ない F タームに対して F ターム説明文を用いた F タームと特許文献間の類似度に基づく手法を提案した．評価実験の結果，教師あり学習手法では，テーマコードレベルの粗い分類であ

表 3　テーマコード付与の実験結果

適合率	再現率	F 値
0.880	0.951	0.914

表 4　SVM を用いた F ターム付与の実験結果

素性	適合率	再現率	F 値
tf-idf のみ	0.340	0.305	0.322
提案手法	0.335	0.358	0.346

表 5　F タームと特許文献比較の予備実験結果

順位	F ターム	類似度	正誤	学習正例数
1	3C045DA01	0.885	正	60
2	3C045CA17	0.878	負	6
3	3C045FE08	0.871	負	9
4	3C045CA16	0.849	負	8
5	3C045BA19	0.846	正	22
6	3C045BA15	0.844	正	15
7	3C045CA06	0.831	負	49
8	3C045GA02	0.828	負	10
9	3C045FD03	0.822	負	100
10	3C045FD05	0.819	負	34
・・・	・・・	・・・	・・・	・・・
63	3C045DA20	0.763	正	3

れば，比較的高精度な付与が可能であることを示した．また，F ターム付与では，重要語の追加によって tf-idf のみを用いる場合と比べて F 値が 0.024 改善した．

F ターム説明文を用いた F タームと特許文献間の類似度に基づく手法では，類似度の高い F タームにいくつか正解となる F タームが含まれていたが，学習データにおける正例の少ない F タームとの類似度は低く，改良の必要がある．今後の課題として，複合語に対応した word2vec のモデルを学習し，F タームと特許文献との比較方法を検討する．

参考文献

[1] 独立行政法人 工業所有権情報・研修館. (2016). 特許分類の概要とそれらを用いた先行技術調査～IPC, FI, F ターム編～（平成 28 年度版）. http://www.inpit.go.jp/content/100798564.pdf

[2] 古屋野 浩史. (2007). 特許分類等の付与精度向上への取り組み. *Japio 2007 YEAR BOOK*, pp.118-119.

[3] Li, Y., Bontcheva, K., and Cunningham, H. (2007). SVM based learning system for F-term patent classification. In *Proc. of NTCIR-6 Workshop Meeting*, pp. 396-402.

[4] Fujino, A. and Isozaki, H. (2007). Multi-label patent classification at NTT Communication Science Laboratories. In *Proc. of NTCIR-6 Workshop Meeting*, pp. 381-384.

[5] Murata, M., Kanamaru, T., Shirado, T., and Isahara, H. (2007). Using the k-nearest neighbor method and SMART weighting in the patent document categorization subtask at NTCIR-6. In *Proc. of NTCIR-6 Workshop Meeting*, pp. 407-413.

[6] 小林 英司. (2015). 特許分類の自動推定に向けた取り組み－機械学習による自動分類推定の課題と今後の展開－. *Japio YEAR BOOK 2015*, pp. 272-275.

[7] Iwayama, M., Fujii, A., and Kando, N. (2005). Overview of classification subtask at NTCIR-5 patent retrieval task. In *Proc. of NTCIR-5 Workshop Meeting*.

[8] Iwayama, M., Fujii, A., and Kando, N. (2007). Overview of classification subtask at NTCIR-6 patent retrieval task. In *Proc. of NTCIR-6 Workshop Meeting*, pp. 366-372.

[9] Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In *Proc. of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96)*, pp. 21–29.

[10] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. (1994). Okapi at TREC-3. In *Proc. of the 3rd Text REtrieval Conference (TREC-3)*, pp. 109–126.

[11] 中川 浩志, 湯本 紘彰, 森 辰則. (2003). 出現頻度と連接頻度に基づく専門用語抽出．自然言語処理, 10(1), 27-45.

[12] Li, Y. and Shawe-Taylor, J. (2003). The SVM with uneven margins and Chinese document categorisation. In *Proc. of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC 17)*, pp. 216-227.

[13] Mikolov, T., Chen, K., Corrado, G., and Dean J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781*.