社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－82 (11)
2006－ＤＤ－54 (11)
2006／3／22

特許文書中のタームの出願人別使用傾向の分析と類似特許文書検索精度への影響評価

間瀬 久雄* 大西 昇**
* （株）日立製作所 システム開発研究所
** 名古屋大学大学院 情報科学研究科

出願特許の請求項文章を入力としてその発明内容を無効化する特許文書を検索する類似特許文書検索において，ある検索精度向上方式の有効性を評価する際には，出願特許とその発明内容を無効化する特許の間の出願人の同一性を考慮すべきであることを提唱する．まず，特許文書中のタームの使用傾向を統計的に分析し，一般語を含むタームの出願人別使用傾向に偏りがあることを示す．次に，検索精度向上方式として不要語除去を採り上げ，出願特許の発明内容を無効化する特許の出願人が元の出願特許の出願人と同じ場合と違う場合で，検索精度の振る舞いが異なることを実験で立証し，出願人の同一性が検索精度に与える影響が強い場合があることを示す．

An Analysis of Term Usage in Patent Documents by Applicant and an Evaluation of its Effect to Patent Retrieval Accuracy

Hisao Mase* Noboru Ohnishi**
* Systems Development Laboratory, Hitachi, Ltd.
** Department of Media Science, Graduate School of Information Science, Nagoya University

In the relevant patent document retrieval using a patent claim text as input, we should take the identity of the patent applicants into account when we evaluate the effect of a patent retrieval method. First, we show that term usage depends on author's affiliation. Then, we show that in the term-based relevant patent document retrieval, stopword deletion, one of the possible retrieval methods to improve retrieval accuracy, is less effective to retrieve relevant patents whose applicants are the same as those of the query patent, and is more effective when the applicants of relevant patents are different from those of the query.

1．はじめに

特許出願された発明内容を特許として認めるべきかを判定するために，特許庁の審査官はその発明内容に類似する過去の特許を入念に検索するが，審査期間の一層の短縮が求められている現在，必要な特許を迅速かつ高精度に検索する技術が強く求められている．

一方，文書検索の一形態として，自然言語文を入力としてその文章内容と類似する文書を検索する類似文書検索技術が普及している．この種の検索では，入力文章の中から内容を端的に表すタームを抽出し，入力文章中での出現頻度（Term Frequency，以下 TF）及び検索対象文書群での出現文書数（Document Frequency，以下 DF）から各タームの重要度を算出してタームとその重要度を要素とするタームベクトルを生成し，タームベクトル間の内積や余弦を類似度として算出する方式[1]が主流である．

特許文書は文書構造がタグで規定されており，記載すべき項目が決まっているが，使用する語や言い回し，構文などに関する執筆スタイルは執筆者毎に異なっている．この執筆スタイルの同一性／相違性が類似特許文書検索精度に大きく影響すると我々は考えている．その一方で，ある出願特許の発明内容を拒絶するために審査官が引用した類似特許が，出願特許と同一の発明者または出願人による特許である場合が全体の実に約 22%も占めているという事実がある．

これらを踏まえ本稿では，検索精度向上方式の有効性を評価する際には特許文書の出願人の同一性を考慮すべきであることを提唱する．まず，「同一の執筆者が記載する文章には共通する言語的特徴がある」という計量文体学における知見を拡張し，「特許文書では執筆者のレベルだけでなく，出願人のレベルでもタームの使用傾向に偏りがある」ことを，特許文書中のタームの出願人別使用傾向を統計的に分析することにより示す．次に，出願特許の請求項文章を入力としたタームベースの類似特許文書検索では，ある出願特許の発明内容を無効化する特許の出願人が元の出願特許の出願人と同じ場合と違う場合で，ある検索精度向上方式の適用が検索精度に与える影響が大きく異なる場合があることを精度評価実験によって示す．

以下，2．では，特許文書におけるタームの使用傾向に関して我々が立てた仮説及びその妥当性を検証するために行った出願人別ターム使用傾向の分析結果について述べる．3．では，検索精度向上方式の一例として不要語除去を採り上げ，不要語除去が類似特許文書検索精度に与える影響について，精度評価実験結果を元に出願人の同一性の観点から考察する．

2．タームの使用傾向の分析

2．1　出願人別ターム使用傾向に関する仮説

我々は，特許文書中のタームの使用傾向に関する以下の仮説を立てた．

【仮説 1-1】特許文書中のターム使用傾向は「出願人」レベルでも異なる．

【仮説 1-2】仮説 1-1 は，発明内容に直接関係しない一般語についても成り立つ．

上記仮説の定性的根拠は以下の 3 点である．

(1)特許文書の構造と執筆者の執筆スタイル

特許文書の記載項目は決まっており，特許を何度も書いている執筆者は特許執筆スタイル（使用する語や言い回し，構文など）が発明内容に依存せずに固まっていることが多い．特に請求項の記載方法は執筆者に大きく依存していると考えられる．

(2)執筆ノウハウの共有及び文書の再利用

上記特許執筆スタイルは，特許教育を受け，所属部署の先輩が執筆した特許を参考にして特許を実際に執筆する中で徐々に固まってい

く．したがって，特許執筆スタイルは同じ出願人組織の中で書かれた特許文書に強く影響されると考える．

(3)発明者と執筆者の関係

一般に企業では特許文書を執筆するのはその発明の考案に最も寄与した技術者であるが，執筆された特許文書に対して企業の知財部の専門家が加筆修正したり，社外の特許事務所などに発注して特許文書を作成・修正させたりする場合もある．

2．2　出願人別のターム出現傾向の比較分析

仮説1-1及び仮説1-2の妥当性を検証すべく，ここでは技術分野を限定した上で，タームが出現する特許文書数の割合が，出願人によってどの程度偏っているかについて分析する．

2．2．1　分析内容

2002年に公開された特許374,550件に含まれる「検索技術」に関する（公開特許公報に記載される分類コードであるテーマコードが5B075「検索装置」である）特許5,330件のうち，出願件数の多い上位10社による特許1,925件を分析対象とする．ターム抽出範囲は「請求項1」のみとし，形態素解析ツール「茶筌[2]」によってタームを自動抽出する．

出願人別ターム使用傾向の評価尺度として，分析対象となる総文書数に対する各タームの出現文書数の割合を出願人別に算出し，その値を10社で比較する．そしてこの割合値の最大値と最小値の差（以下「割合格差」と呼ぶ）が大きいタームほど出願人別ターム使用傾向に偏りがあると判定する．

2．2．2　分析結果と考察

割合格差が20ポイント以上であるターム（一部）の使用傾向を表1に示す．割合格差の大きいタームの中には，「コンテンツ」「検索」

表1　出願人別のターム使用傾向（抜粋）

	特許文書総数	5B075特許文書	特許出願件数上位10社											
			合計	A	B	C	D	E	F	G	H	I	J	平均
特許文書数	374550件	5330件	1925件	262件	218件	209件	209件	191件	190件	188件	177件	141件	140件	193件
発明者異なり数	…	4421人	1529人	180人	196人	194人	119人	150人	149人	161人	127人	129人	124人	153人
ターム見出し	そのタームが使用されている特許文書数の割合（%）													割合格差
文書	0.3	8.4	14.7	8.4	7.3	14.4	53.6	13.6	**0.0**	10.1	13.6	14.9	9.3	53.6
方法	17.4	27.4	29.4	64.5	21.6	50.2	18.2	34.0	**11.1**	11.7	20.3	29.8	15.0	53.5
備える	35.3	45.0	41.8	**17.9**	41.3	29.7	50.2	24.6	47.4	57.5	59.3	41.8	65.0	47.1
手段	25.6	52.8	54.4	**33.2**	52.8	41.2	55.0	53.9	73.7	64.9	79.1	45.4	54.3	45.9
この	14.4	12.0	12.4	17.2	13.8	3.4	12.0	42.9	7.9	5.9	**0.0**	0.7	16.4	42.9
装置	37.1	40.3	47.8	**30.5**	33.0	34.5	53.1	38.2	72.1	62.2	63.3	48.2	55.7	41.6
システム	8.7	45.6	38.0	30.9	59.6	46.4	45.5	35.6	30.0	34.0	33.3	**18.4**	39.3	41.2
前記	60.5	76.0	77.7	78.2	91.7	66.0	74.6	90.1	66.3	88.3	89.3	73.1	**51.4**	40.3
上記	9.7	9.3	11.2	9.2	**1.4**	11.5	2.4	6.3	27.4	5.9	6.2	12.1	40.7	39.3
端末	4.3	29.9	24.6	34.4	48.2	14.8	**12.9**	23.0	22.6	23.4	14.7	14.2	30.7	35.3
具備	5.6	8.3	8.7	3.8	6.4	2.4	3.8	36.1	12.1	15.4	3.4	**1.4**	**1.4**	34.7
利用	3.0	17.9	18.4	40.5	26.2	21.1	12.0	14.1	11.1	17.0	**6.8**	11.4	10.7	33.7
該	24.9	22.9	24.3	37.4	23.4	30.1	30.6	13.6	13.2	**6.9**	38.4	28.4	13.6	31.5
ある	39.3	47.2	45.7	64.5	49.5	39.7	36.8	**34.0**	42.1	45.7	58.8	42.6	34.3	30.5
特徴	82.6	80.7	84.9	82.1	88.5	90.9	90.9	91.6	82.6	**68.6**	97.2	76.6	75.7	28.6
接続	12.7	22.3	18.1	17.2	31.2	17.7	17.2	14.1	17.4	16.5	17.5	**6.4**	22.1	24.8
介す	10.1	23.0	19.4	22.9	32.6	11.5	16.8	17.3	26.8	14.9	18.1	**7.8**	20.7	24.8
情報	11.7	71.1	72.0	78.2	76.2	75.1	**55.0**	67.0	79.5	76.6	65.5	70.9	73.6	24.5
コンテンツ	0.6	7.9	10.0	19.9	7.3	6.2	3.8	**2.6**	26.8	9.6	6.8	6.4	5.7	24.2
送信	5.2	23.2	22.0	26.7	33.9	16.8	13.4	**11.0**	31.1	22.3	18.6	19.2	24.3	23.0
検索	1.2	39.4	39.3	40.1	50.9	36.4	37.3	42.9	**28.4**	31.9	42.9	34.8	47.1	22.5
情報処理	0.8	5.0	7.3	**1.9**	2.8	5.7	6.2	2.6	24.2	5.3	13.6	7.8	6.4	22.3
データベース	1.5	28.5	25.3	30.9	26.2	33.0	19.6	28.3	**14.7**	20.7	17.5	25.5	36.4	21.7

注1：網掛けされたターム見出しは一般語とみなされるもの
注2：網掛けされた数値はそのタームにおける最大値を示し、太字の数値は最小値を示す。

「データベース」など，検索分野でよく使われるタームのほかに，発明内容に関係しない一般語（表1で網掛けのターム）も含まれている．

タームを個別に見ると，ターム「文書」の割合格差は53.6ポイント（53.6%(D社)-0.0%（F社））と大きい．この大きな格差は，今回分析範囲とした技術分野が情報検索全般を網羅しており，F社では文書を対象とした検索に関する特許を出願していないために生じたと説明できる．しかし，「この（割合格差42.9ポイント）」「該（同31.5）」「ある（同30.5）」などは発明内容の観点からは格差の理由が説明できない．また，ほぼ同じ意味であるターム「前記」「上記」を見ると，B社では「前記」を専ら使用している（91.7%）のに対して，J社では「上記」を使用している特許の割合が非常に高い（40.7%）．これらのターム使用傾向の偏りは，出願人別の筆頭発明者数も多いことを考慮すると，出願人レベルの執筆スタイルの違いによるものと言える．

2．3　出願特許とそれを無効化する特許のターム共通性の比較分析

仮説1-1及び仮説1-2の妥当性を検証する別の方法として，ここでは出願特許とそれを無効化する特許との間に共通して使用されているターム異なり数の割合を比較する．

2．3．1　分析内容

単独の出願人によって出願され1993年から1997年に公開された特許のうち，1993年から1997年に公開された特許によって無効と判定された出願特許とその無効化特許のペア20,000組を分析対象とした．このうちの10,000組は出願人が違う特許ペアであり，残りの10,000組は出願人が同じ特許ペアである．出願人が同じ特許ペアの内訳は，筆頭発明者も同じである特許ペア2,758組，筆頭発明者は異なるが共通の発明者が一人でも存在する特許ペア2,285組，発明者が全く違う特許ペア4,957組である．なお，出願特許を無効化する特許に関するデータは，整理標準化データ[3]から抽出した．

ターム抽出範囲は「請求項1」のみとし，形態素解析ツール「茶筌」によってタームを自動抽出する．なおここでは，付属語などを含めた全てのタームを対象として分析を行った．

ターム共通性の評価尺度として，「出願特許文書中のタームの異なり数に占める，出願特許文書及び無効化特許文書に共通に出現するタームの異なり数の割合（以下，『ターム共有率』と呼ぶ）」を特許ペア毎に算出する．そして，特許ペアの出願人が同じであるか違うかによって，ターム共有率の平均値及び，ターム共有率に対する文書数分布の傾向を比較する．

2．3．2　分析結果と考察

ターム共有率に対する特許文書数の分布を図1に示す．出願人が同じ特許ペアにおける平均ターム共有率は53.1%であり，出願人が違う特許ペアの平均ターム共有率 43.0%に比べて10.1 ポイント高く，出願人が同じ場合に同じタームが好んで使用される傾向が高い．また，「筆頭発明者も同じ」である場合の平均ターム共有率が60.6%と最も高いが，「出願人は同じだが発明者が全く違う」場合でも，平均ターム共有率は48.6%であり，「出願人が違う」場合の平均ターム共有率43.0%に比べて5.6ポイント高くなっている．また，図1では，出願人が違う場合の文書数の分布曲線（太い実線）に比べて，出願人が同じ場合（太い点線，細い実線，1点鎖線，破線）の方が，分布全体が右（ターム共有率の高い方向）にずれている．

これらの結果から，発明者が全く違っていても，出願人が同じであれば，同じタームが使用される傾向があると言える．

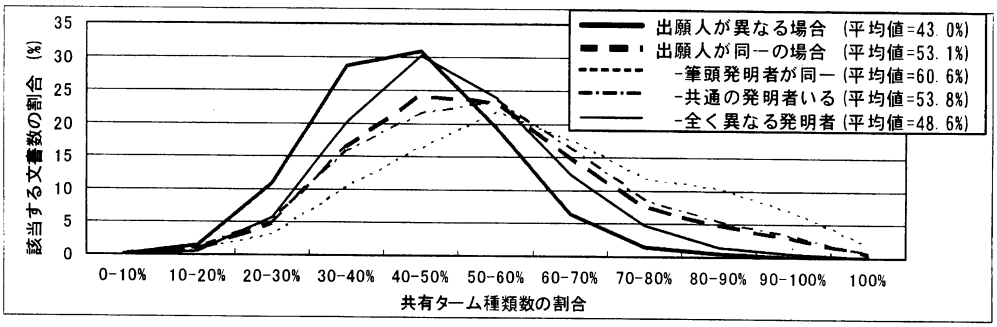

図1　出願人の同一性からみたターム共有率の分布

3．検索方式の検索精度への影響の検証

ここでは，2．で検証した仮説 1-1 及び 1-2 を踏まえ，検索精度向上方式が類似特許文書検索の検索精度に与える影響について，出願人の同一性の観点から考察する．本稿では，検索精度向上方式の一例として不要語除去を用いる．

3．1　不要語除去と検索精度に関する仮説

仮説 1-1 及び仮説 1-2 を踏まえ，我々は特許検索精度に関する以下の三つの仮説を立てた．

【仮説 2-1】「ある出願特許の発明内容を無効化する特許が元の出願特許の出願人と同じ場合，不要語除去は検索精度に悪影響を及ぼす」

【仮説 2-2】「ある出願特許の発明内容を無効化する特許が元の出願特許の出願人と違う場合，不要語除去は検索精度に好影響を及ぼす」

【仮説 2-3】「仮説 2-1，2-2 より，特許検索精度向上方式の有効性評価では，出願人の同一性を考慮すべきである」

仮説 2-1 の根拠を述べる．一般のタームベースの類似文書検索では，検索に使われるすべてのタームの一致度によって類似度が算出される．出願人が同じ場合，仮説 1-2 により，発明内容に直接関係しない一般語の使われ方が類似しており（共通して使用される確率が高く），これが類似度に反映されている．この一般語を不要語として除去すればその分だけ類似度が下がり，検索順位が低下すると考えられる．

一方，仮説 2-2 の根拠は次のとおりである．出願人が異なると一般語の使用傾向も異なる．一般語は発明内容を端的に表さないので，一般語を不要語として除去しても，順位が低下して検索精度に悪影響を及ぼすことは比較的少なく，逆にノイズとなるタームがなくなることによる効果の方が検索順位の向上に貢献すると考えられる．

3．2　不要語除去の検索精度への影響評価

本節では，上記仮説 2-1，2-2 及び 2-3 の妥当性を検証すべく，不要語除去が検索精度にどのように影響するのかに関する評価実験と，その結果について述べる．

3．2．1　不要語の定義

本論文では，「不要語」を「文章の内容を特定しない語で検索精度の向上に貢献しない語」と定義する．すなわち一般には，助詞や助動詞，接続詞，形式名詞，記号類，一部の名詞・動詞などが主な構成要素となる．しかし，実際に不要語として登録されているタームは，検索システムによってまちまちで普遍的な定義がないのが現状である．また，不要語除去そのものに懐疑的な研究報告もある[4]．

表2　検索精度評価実験で使用する不要語集合

#	処理形態	着目する属性	生成条件	不要語数	不要語のサンプル
1	全自動生成	DF	DFが50%以上	10	こと/する/れる/記載/項/請求/前記/特許/特徴/範囲
2		DF	DFが10%以上	88	ある/いずれ/ため/以上/可能/含む/手段/供給/固定
3		DF	DFが 5%以上	193	1つ/およぶ/これら/データ/応じる/加熱/回転/回路
4		DF + 単語長	高頻度の1文字語	50	圧/一/下/化/外/該/間/器/基/機/型/系/後/光/孔/項
5		DF + 品詞	高頻度の動詞	100	いる/おく/くる/圧す/囲む/備える/行う/持つ/示す
6		DF + 文字種	高頻度のひらがな語	100	あける/あたり/あと/ある/いう/いく/いくつか/いずれ
7	半自動生成	DF + 特許固有性	高頻度の特許固有語	24	特許/特徴/手段/前記/当該/工程/構成/成る/設ける
8		DF + 語意	DFが1%以下の一般語	613	ここ/ところ/以来/為/一番/右記/過程/該当/我々/何
9	手動作成	語意	一般語すべて	765	場合/数々/上述/上/是非/先程/他方/対する/大体/誰

本実験では，検索精度評価のベースラインシステムとして検索エンジンGETA[1][5]を採用する．検索に使用されるタームは，名詞(代名詞，形式名詞などを含み，一文字からなるひらがな／カタカナを除く)，動詞（助動詞含む），英文字列（一文字からなる英文字は除く）のみで，他のタームは除外している．

本実験では GETA によって抽出される上記タームをベースとして，表2に示す観点及び基準によって生成された9種類の不要語集合を不要語除去で使用することにより，検索精度がどのように変わるかについて比較する．9種類のうち，6種類は出現文書数，品詞，単語長，文字種などの客観的観点に基づいて全自動で生成した不要語集合であり，残りは筆者の主観的判断を介して作成した不要語集合である．

3．2．2　実験内容

検索対象となる特許文書は，1993 年から1997年までの公開特許公報約170万件である．入力文章として，1999 年に公開された審査済みの特許 14,399 件の請求項1を用いた．検索されるべき正解特許として，整理標準化データに記載された無効化特許延べ 25,148 件（出願人が入力特許と同じである特許 5,132 件（20.4%）を含む）を用いた．

ターム抽出にかかる形態素解析は茶筌を用い，検索エンジンは GETA を用いた．本章で評価する内容は，検索エンジンの検索アルゴリズムに影響される部分が少なくないが，本実験ではTF及びDFを用いたオーソドックスな検索方式を採用している．

本評価実験における精度評価尺度として以下の2種類を採用する．

(a)順位の上がった正解特許件数の割合

不要語除去を適用したことにより，適用しない場合に比べて検索順位が上がった正解特許が何件あるかの割合である．ここでは，「不要語除去によって検索順位が上がった件数から下がった件数を差し引いた値を，検索順位が下がった件数で割った値（%)」を用いて比較評価する．この値が0より大きい場合，検索精度が改善されたとみなし，逆に0より小さい場合は検索精度が悪化したとみなすことができる．

(b)平均精度（MAP）

直感的には正解特許の検索順位の逆数に相当する値であり，下式によって算出される．

$$\text{平均精度} = \frac{1}{\sum_{i=1}^{N} X_i} \sum_{i=1}^{N} \left[\frac{X_i}{i} \left(1 + \sum_{k=1}^{i-1} X_k\right) \right]$$

ここで，N は出力文書の総数，Xi は出力第 i 位の文書が正解であるか否かを 1/0 の2値で

1 「汎用連想検索エンジン（GETA)」は，情報処理振興事業協会（IPA)が実施した「独創的情報技術育成事業」の研究成果である．

表3　出願人の同一性から見たベースライン検索精度

評価尺度		合計 25,148件	出願人異なる 20,016件	出願人同じ 5,132件
MAP		0.1386	0.0865	0.2998
検索順位	1	7.0%	4.0%	18.8%
	1- 10	21.0%	14.9%	45.0%
	1- 50	36.6%	29.9%	62.8%
	1- 100	44.4%	37.9%	69.8%
	1- 500	64.0%	59.1%	83.2%
	1-1000	71.9%	67.9%	87.8%
	1001-	28.1%	32.1%	12.2%

表したもので，正解である場合は1で，そうでない場合は0とする．入力文書毎にこの値を計算し，最後にその平均値を求めて全体の平均精度（MAP）とする．

3．2．3　実験結果と考察

まず，不要語除去を適用しない場合（ベースライン）の検索精度の評価結果を表3に示す．全体のMAPは0.1386であるのに対して，入力特許と出願人が同じ正解特許 5,132 件に限定するとMAPは0.2998とはるかに高くなり，逆に出願人が違う正解特許 20,016 件に限定するとMAPは0.0865とはるかに低くなる．すなわち，評価データ全体のMAPの値は，その20%を占めるに過ぎない出願人が同じ正解特許のMAPの値に大きく依存していることが分かる．また表3に示すように，上位100位以内に正解特許が出力される割合が全体で44.4%であるのに対して，出願人が同じ場合は69.8%と非常に高く，逆に出願人が違う場合は37.9%という低い値となっている．これらの結果は，出願人が同じ特許間のターム共有率が高いという 2.3 節の分析結果によるものと考えられる．

次に，9種類の不要語集合を適用した場合の検索順位の変動の割合を図2に示す．棒グラフが上に伸びているものは，不要語除去によって検索順位が向上した件数の方が悪化した件数よりも多い，すなわち検索精度が向上したことを示しており，逆に下に伸びているものは検索精度が悪化したことを示している．本評価実験で重要なのは，不要語除去によって検索精度がどれだけ向上したか（棒グラフがどのくらい長いか）ではなく，出願人が同じ場合（白い棒グラフ）と違う場合（網掛けのかかった棒グラフ）で，不要語除去が検索精度にどのくらい影響を及ぼしたかの違い（両グラフの差分）を検証することであることに留意されたい．

DF を抽出基準とした 3 種類の不要語集合（#1-#3）及び 1 文字語からなる不要語集合（#4）においては，出願人が同じ場合も違う場合も検索順位が悪化した件数の方が多くなっており（棒グラフが下に伸びている）検索精度が悪化しているが，出願人が同じ場合の方が悪化の度合いが高くなっている（棒グラフが下により長く伸びている）ことが分かる．また，動詞からなる不要語集合（#5）及び DF1%以下の一般語からなる不要語集合（#8）については，どちらも検索順位が向上した件数の方が多くなっている（上に伸びている）が，出願人が違う場合の方が向上の度合いが高くなっている（上により長く伸びている）．更に，特許固有語からなる不要語集合（#7）及び一般語からなる不要語集合（#9）では，出願人が同じ場合は精度が悪化しているのに対して，出願人が違う場合は精度が向上しているという，全く逆の傾向が見られる．

これらの結果は，仮説 2-1 及び 2-2 が妥当であることを示している．なお，ひらがな語からなる不要語集合（#6）に関しては唯一ほかの不要語集合と異なり，出願人が違う場合の方が精度の改善傾向が小さい．これは，一般にひらがなで書かれる語は，出願人を問わずひらがなで書くことが多く，出願人別の偏りが出にくい

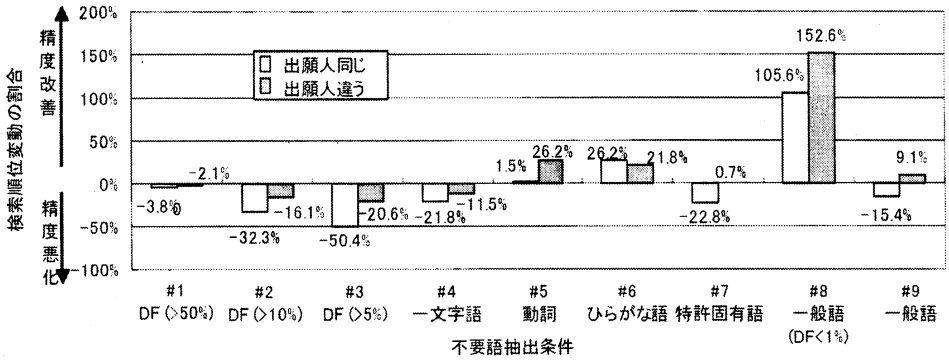

図2　不要語除去による検索順位変動

ためと考えられる．同様に，DF50%以上の不要語集合（#1）において，出願人の同一性による差異が小さいのも，この不要語は非常に出現頻度が高く，出願人を問わず使用されるタームであり，出願人別の偏りが出にくいためと考えられる．

4．むすび

本稿ではまず，同一の出願人によって記載される特許文書ではターム使用傾向が類似すること，発明内容に直接関係しない一般語についても，出願人によって使用傾向が偏っていることを示した．次に，この知見を発展させ，不要語除去処理など類似特許文書の検索精度を向上させる方式の有効性を評価する際には，出願人の同一性を考慮することが必要な場合があることを実データによる評価実験結果をもとに提唱した．

本論文では，不要語除去を題材として考察したが，同様のことは，異表記展開や同義語展開，関連語展開など，タームを追加・削除するほかの処理にもあてはまると考えている．出願人が同じ場合，これらの処理によって別のタームがノイズとして加わることとなり，検索精度が低下しやすくなると考えられる．

参考文献

[1] G. Salton and M. J. McGill, “Introduction to Modern Information Retrieval”, McGraw-Hill Book Company, 1983.

[2] 茶筌, http://chasen.naist.jp/hiki/ChaSen/.

[3] 独立行政法人工業所有権情報・研修館, “整理標準化データの提供”, http://www.ncipi.go.jp/info/standard/.

[4] Riloff, E. : Little Words Can Make a Big Difference for Text Classification, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.130-136, 1995.

[5] GETA, http://geta.ex.nii.ac.jp/.