明細書からの用語抽出と明細書間で共通する用語の出現頻度に基づく類似度算出
－用語精査結果の利用－

Similarity calculation of patent documents by the evaluation of the frequency of occurrence of common terms between the two documents

株式会社クレステック　情報技術部　システムコンサルタント　楠本　浩二

PROFILE　1986 年 九州工業大学大学院工学研究科修士課程情報工学専攻修了。現在、例規管理システム、法令審査支援システム、損害保険・生命保険約款チェックシステム、製造業マニュアルの文書編集・比較・日本語精査・画像類似検索システムなどの設計・開発に従事。

k-kusumoto@crestec.co.jp

1　はじめに

特許明細書（以下明細書）において正しい用語が使用されているかチェックするために、予め使用可能な用語すべてを用語辞書に定義しておく。しかし、このような辞書を手作業で一から作成することは非効率なのでプログラムで既存の代表的な複数の明細書から用語を抽出する。明細書から用語を抽出するには、形態素解析後、１つ以上の形態素から構成される主に名詞を用語として判定するためのルールを適用する [1]。文章から抽出された用語のうち、統一して利用するものをユーザーが選択して用語辞書の内容とする。

執筆した明細書の日本語精査時も同様にして明細書から用語を抽出し、その読みと出現頻度を記録した「用語リスト」を自動生成する。リスト内の用語それぞれが、上記用語辞書に登録された用語と一致しなかった場合、警告メッセージを出力する。明細書中、辞書になかった用語を修正するか、または用語として正しいと判断して新規に辞書登録するかはユーザーが判断する。また用語リストを用語の読み仮名順でソートすると、使用されている用語の表記ゆれや送り仮名の不揃い、同音異義語などを一覧できるため、明細書中、誤りの可能性がある箇所の発見が容易になる。

明細書においては曖昧性を避けるために指示代名詞ではなく、名詞を反復して使用するので、出現頻度が高い用語は明細書の話題（トピック）を反映する。したがって、ある２つの明細書で共通に出現する用語が高頻度で使用されている場合、双方の明細書に記載のトピックも関連している可能性が高い。本稿では、あるキーワードから全文検索された明細書すべてから用語リストを生成し、そのうちの１件の明細書の用語リストと、残りの明細書の用語リストとの類似性を数量化する。最後に類似度が高かった明細書のペアについて明細書のトピックの共通性の有無を調べ、その結果を評価する。

2　用語リスト

2.1　用語辞書

通常、誤った表記や文章の例が組織内の情報資産として共有されている。こうした不正な例を辞書化しておくと、文章の精査時に誤りの可能性がある箇所を警告できる。しかし、不正な用語や用語のタイプミスは、そのパターンすべてを網羅することが難しい。そこで用語精査に関しては、使用可能な正しい用語だけを辞書に登録しておき、適切な用語が使用されているかチェックする。こうした辞書の作成では、一から用語を登録することは非効率なので、まず既存の明細書の集合から用語をすべて抽出し、その中から正しい用語だけを登録する。

2.2　MeCab による形態素解析

本章では説明のために、明細書の例文テキストの文章を形態素解析エンジン MeCab で解析して用語を抽出する。文章例の一部「枠体（14）と、該枠体の上部に設置された数センチ長の軸受け部などによって」の解析結果を表 1 に示す。

用語は名詞から構成されるので、表 1 の「品詞」列

表１　文章例の一部の形態素解析結果 (抜粋)

No	単語	品詞	細分類 1	細分類 2
1	枠体	名詞	一般	*
2	(	記号	括弧開	*
3	14	名詞	数	*
4	)	記号	括弧閉	*
5	と	助詞	格助詞	引用
6	、	記号	読点	*
7	該	名詞	一般	*
8	枠体	名詞	一般	*
9	の	助詞	連体化	*
10	上部	名詞	一般	*
11	に	助詞	格助詞	一般
12	設置	名詞	サ変接続	*
13	さ	動詞	自立	*
14	れ	動詞	接尾	*
15	た	助動詞	*	*
16	数	名詞	数	*
17	センチ	名詞	接尾	助数詞
18	長	名詞	一般	*
19	の	助詞	連体化	*
20	軸受け	名詞	一般	*
21	部	名詞	接尾	一般
22	など	名詞	接尾	一般
23	によって	助詞	格助詞	連語

の「名詞」と認識された単語が用語の候補となる。ここでは MeCab の日本語辞書として IPADIC を利用している。明細書において、IPADIC に未定義の語句が出現していた場合、それを名詞として処理し、新規の用語として検出できる。

2.3　用語抽出ルール

表 1 に網かけで示した名詞のうち「14」「該」「数」「センチ」「長」「部」「など」は独立した用語ではない。また名詞「設置」は動詞化されるサ変名詞なのでここでは用語と判定しない。そこで表 2 に示すようなルールを定義する。

表 2 には用語として「対象の形態素」を「包含する (include)」かまたは「除外する (exclude)」かを「ルール」列に指定し、上から下に順に適用する。また「対象の形態素」の前後に接続する形態素の条件を「直前の形態素」列と「直後の形態素」列にそれぞれ指定する。「*」は、形態素解析の結果を特に限定しない任意の値でよいことを示す。

表 2 のルールを適用すると、表 1 に示した形態素解析結果の「14」はルール 1、「該」はルール 4、「設置」はルール 5、「数センチ長」はルール 1, 2, 3、によってそれぞれ用語から除外される。解析結果の「軸受け」「部」は、ルール 8, 9 によって含められるが、「など」はルール 7 によって除外される。こうして、以下の網掛けした部分に示すように、専門用語だけでなく「ドア」や「上部」のような一般名詞を含む独立した用語が文章から抽出される。

> 本願の第１の発明によれば、スライド式ドア（A）に設置された、ペットが安全に出入りすることを可能にするためのくぐり戸であって、該くぐり戸が、扉体（１２）と、枠体（１４）と、該枠体の上部に設置された数センチ長の軸受け部などによって支承された、扉体を揺動自在に軸支する回転軸（１３）と、からなっていて、この回転軸の一端部が延設され、この延設された軸部に、ストッパー（2）が軸支され、そのストッパーが、上記扉体と連動し、スライド式ドアの動きを制限するようにドア内の収納部から出没自在になっていることを特徴とするペット用くぐり戸が提供される。

2.4　用語リストの例

上記の例文から抽出された用語と用語の読みと出現頻度を記録した「用語リスト」を表 3 に示す。

表２　形態素解析結果から用語として判定するためのルール（抜粋）

No	ルール	対象の形態素				直前の形態素				直後の形態素			
		単語	品詞	細分類 1	細分類 2	単語	品詞	細分類 1	細分類 2	単語	品詞	細分類 1	細分類 2
1	exclude	*	名詞	数	*	*	*	*	*	*	*	*	*
2	exclude		名詞	接尾	助数詞	*	*	*	*	*	*	*	*
3	exclude	*	*	*	*	*	名詞	接尾	助数詞	*	*	*	*
4	exclude	該	名詞	*	*	*	*	*	*	*	*	*	*
5	exclude	*	名詞	サ変接続	*	*	*	*	*	*	動詞	*	*
6	include	*	名詞	サ変接続	*	*	*	*	*	*	名詞	*	*
7	exclude	など	名詞	接尾	*	*	*	*	*	*	*	*	*
8	include	*	名詞	接尾	*	*	*	*	*	*	*	*	*
9	include	*	名詞	一般	*	*	*	*	*	*	*	*	*

表３　例文から生成された用語リスト

No	用語	読み	頻度	辞書
1	安全	アンゼン	1	○
2	一端部	イッタンブ	1	○
3	動き	ウゴキ	1	○
4	回転軸	カイテンジク	2	○
5	可能	カノウ	1	○
6	軸受け部	ジクウケブ	1	○
7	軸部	ジクブ	1	○
8	収納部	シュウノウブ	1	○
9	出没自在	シュツボツジザイ	1	×
10	上部	ジョウブ	1	○
11	ストッパー	ストッパー	2	○
12	スライド式ドア	スライドシキドア	2	○
13	ドア	ドア	1	○
14	特徴	トクチョウ	1	○
15	発明	ハツメイ	1	○
16	扉体	ヒタイ	3	○
17	くぐり戸	クグリド	3	○
18	ペット	ペット	2	○
19	本願	ホンガン	1	○
20	揺動自在	ヨウドウジザイ	1	×
21	枠体	ワクタイ	2	○
計			30	2

3　用語リストの利用

3.1　用語辞書との照合

新規に執筆した明細書を用語精査するとき、生成された用語リストにある用語すべてが既存の用語辞書にある用語と一致するかチェックする。一致しなかった場合、未知の用語が見つかったことを警告するメッセージを出力する。例えば用語辞書に「出没自在」「揺動自在」が登録されていなかった場合、表３の「辞書」列には用語辞書との照合結果として×を記録し、用語精査の結果として下記の警告メッセージを表示する。この例では、メッセージを確認したユーザーが、「出没自在」「揺動自在」を正しい用語として辞書に新規登録する。

```
WARNING-YOGO-009-「出没自在」は用語辞書にありません。
WARNING-YOGO-020-「揺動自在」は用語辞書にありません。
```

3.2　用語のゆれ

表４は、ある建設機械に関する明細書から用語リストを生成し、用語のゆれを集めたものである。

表４　用語リスト中に発見されたゆれの例（抜粋）

用語	読み	頻度
エキゾースト	エキゾースト	6
エギゾースト	エギゾースト	2
エンジン稼働状態	エンジンカドウジョウタイ	1
エンジン稼動状態	エンジンカドウジョウタイ	1
締付け	シメツケ	7
締め付け	シメツケ	1
傷害	ショウガイ	13
障害	ショウガイ	1
シリンダ固定冶具	シリンダコテイジグ	5
シリンダ固定治具	シリンダコテイチグ	1
身体	シンタイ	5
人体	ジンタイ	1
摩耗	マモウ	1
磨耗	マモウ	1

日本語辞書 IPADIC に正しい読みを登録しておくと、用語リストを正しくソートできるので例えば「傷害」と「障害」、「摩耗」と「磨耗」のような同音異義語や「固定冶具」のような誤字、「締付け」と「締め付け」のような送り仮名の不揃い、「エキゾースト」と「エギゾースト」のようなカタカナ表記のゆれなどの発見が容易になる。こうした用語のゆれを修正できると、明細書を他言語へ自動翻訳する過程において誤りの可能性を防ぐことができ、翻訳コストの削減になる。

4　用語リスト間の類似度

4.1　用語リストのサイズ

本章では、実際の明細書から生成した用語リストを評価するため、法律や条例を対象にした特許公報を例とする。ここではキーワード「条文」AND「改正」で全文テキスト検索した明細書 83 件を形態素解析し、用語を抽出した。表５には、文字数、形態素数、用語数などの平均値を算出した結果を示す。

文字数の平均 60% が形態素数である。その形態素数の 20% が用語の出現総数であり、１つの用語の出現頻度は平均４回である。したがって、１明細書中の文字

表５　検索された明細書 83 件に関する平均値

文字数	形態素数	形態素数 / 文字数	用語の出現総数	用語の出現総数 / 形態素数	用語数	用語の出現頻度
28,983	17,328	0.6	3,351	0.2	828	4

数をNとすると、1つの明細書に存在する用語数の見積もりは、N × 0.6 × 0.2 ／ 4、すなわち文字数Nの約3パーセントである。

4.2　用語リストの比較

着目している技術に関連する明細書を効率的に検索するためには、通常、検索用のキーワードや規定のFタームを利用する。学術論文や明細書では、キーワードとなる専門用語や一般用語が本文中に混在し、特に専門用語は、時間とともに変化している [2]。このため検索のための用語やキーワードの選択次第では検索結果が大きく変わることがある。一方、特許を効率的かつ高精度に分類するために、機械学習結果の分類ルールと明細書テキストを分類エンジンに入力し、明細書の分類の自動推定をする取組みもある [3]。このように、関連する特許を効率よく発見するためのハードルは依然として高い。

明細書においては、専門用語、一般用語に限らず用語の繰り返しに特徴がある。出願した明細書または関心のある明細書から自動生成された1件の用語リストと、キーワード検索された明細書から生成された複数の用語リストそれぞれとを比較すると、2つの用語リスト間の類似度が高いものは、その明細書の内容も類似性があると予想した。またこのような用語リスト同士の比較の場合、比較回数や明細書のサイズが大きくなっても、用語数の規模から計算時間は大きな問題とならない利点もある。

本稿では、テキスト検索した法律や条例を対象にした明細書83件の中から1件の明細書P0に着目する。このP0から抽出された用語リストL0と、残りの明細書P1 ～明細書P82それぞれから生成された用語リストL1 ～ L82それぞれとの類似度を算出する。

単語あるいは用語の重みづけの算出方法は過去30年間、数十に上っているが、いずれも「偏在性の数量化」という共通の原理に従っている [4]。本稿では2つの明細書に共通に出現する用語を対象に以下3通りの重みづけによる数量化を試みた。

4.3　共通した用語の割合に基づく類似度

P0とPi (i=1,…,82) それぞれとの類似度を求めるために、2組ごとの用語リストを対象に、P0とPi双方共に存在する用語が用語全体の何割を占めるかを類似度 $Similarity_{P0,Pi}$ として、以下の式で求める。

この算出結果を昇順にグラフ化したもの図1に示す。横軸の数字は明細書番号iを示している。縦軸はP0とPiの類似度である。筆者が明細書の内容に目を通し、実際に明細書P0と関連していると判断したものには（後述するグラフも同様）黒に着色した棒グラフで示した。

図1に示すように、比較結果の多くが20%前後の類似度となり、類似度 $Similarity_{P0,Pi}$ と実際の内容の関連性とは相関を示さなかった。その理由として以下の2点が挙げられる。

(1) 共通に出現する用語には、明細書の内容を特徴付けないものも多いため、これがノイズとなって類似度を

$$Similarity_{P0,Pi} = \frac{\text{Piにも存在するP0の用語数} + \text{P0にも存在するPiの用語数}}{\text{P0の用語数} + \text{Piの用語数}} \times 100$$

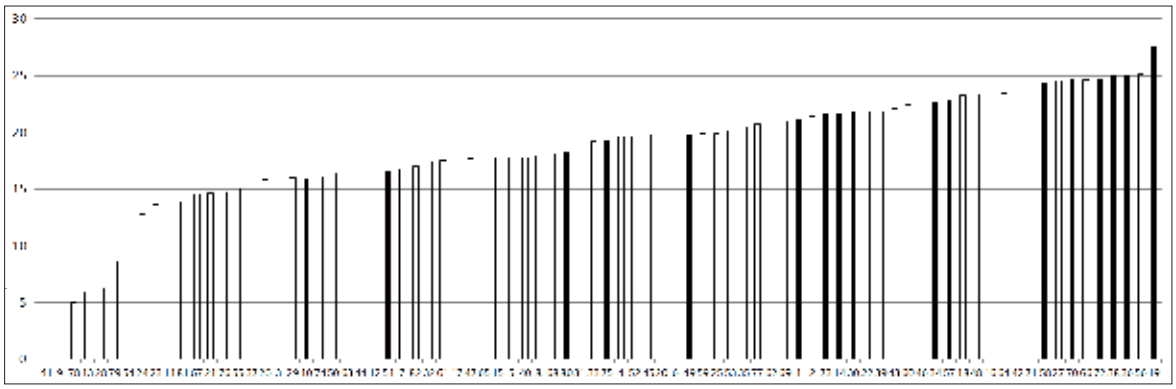

図1　文書間に共通して出現する用語の割合に基づく類似度

大きくしている。

(2) 共通に出現する用語それぞれの出現頻度を反映していない。

4.4 出現頻度に基づく類似度

上記 (1) のノイズを低減するために、明細書において汎用的に使用される用語は、内容を特徴付けないものとして集合 ***G*** の要素とし、これらは常に類似度の算出対象から除外する。集合 ***G*** の一部を表 6 に示す。

表 6 明細書において汎用的に使用される用語の集合 *G*（抜粋）

システム	関係	作業	詳細	適用	複数
ステップ	機能	作用	情報	動作	方向
データ	記憶部	参考	条件	特徴	方式
ファイル	記録	参照	状態	内容	方法
フロー	技術	実施	図	媒体	目的
フロー図	技術分野	実施形態	図形	範囲	問題
プログラム	形式	実施例	図示	必要性	矢印
ブロック図	形態	手順	図面	表	要約
可能性	効果	手続き	説明図	表示	流れ
課題	効果的	手段	選択図	表示画面	力
画面	構成	順序	操作	不図示	例
解決手段	構成要素	処理	装置	符号	例示
概念図	構造	処理手順	態様	部分	・・・

更に上記 (2) を解決するために、出現頻度の重みづけとして、明細書 Pi 内の用語 t の出現頻度 $n_{t, Pi}$ の対数（底 10）をとる。ただし、明細書 Pi に 1 度しか出現しない用語 t は除外したいため、$\log n_{t, Pi}$ に 1 を加算しない。そうすると、条件 $t \notin G$ であって $\log n_{t, P0} > 0$ かつ $\log n_{t, Pi} > 0$ のとき、明細書 P0 と明細書 Pi に共通する用語 t の出現頻度の重みづけの値として $\log n_{t, P0} + \log n_{t, Pi}$ を計算する。類似度の母数は、明細書 P0 の用語すべての出現頻度の対数と明細書 Pi の用語すべての出現頻度の対数の和とし、類似度 $Similarity'_{P0, Pi}$ を以下の式で求める。

上記の算出結果を昇順にグラフ化したものを図 2 に示す。図 2 では図 1 よりも類似度の差異が明確に現われており、明細書の内容が実際に関連しているものが高い類似度の領域に集まっている。

4.5 出現頻度の差を考慮した類似度

前述の条件 $\log n_{t, P0} > 0$ かつ $\log n_{t, Pi} > 0$ に着目すると、例えば明細書 P0 内での用語 t の出現頻度が 100、明細書 Pi 内での出現頻度が 2 の場合、双方の明細書において単語 t は同等の重要性を持っていない。そこで出現頻度差に閾値を設定し、出現頻度に一定以上の開きがない場合に限って、類似度として加算するように以下の条件を加える。

$$\log n_{t, P0} \geqq \frac{MAX(\log n_{t, P0}, \log n_{t, Pi})}{M} \quad \text{かつ}$$

$$\log n_{t, Pi} \geqq \frac{MAX(\log n_{t, P0}, \log n_{t, Pi})}{M}$$

ここでは M=2 とする。この条件で求められた結果を図 3 に示す。用語の出現頻度の差に閾値を設定しない図 2 と比較すると、類似度の差異が一層鮮明に現れ、明細書の内容が実際に関連しているものもより高い類似

$$Similarity'_{P0, Pi} = \frac{\sum_{t \in P0 \cap t \in Pi} (\log n_{t, P0} + \log n_{t, Pi})}{\sum_{s \in P0} \log n_{s, P0} + \sum_{u \in Pi} \log n_{u, Pi}} \times 100$$

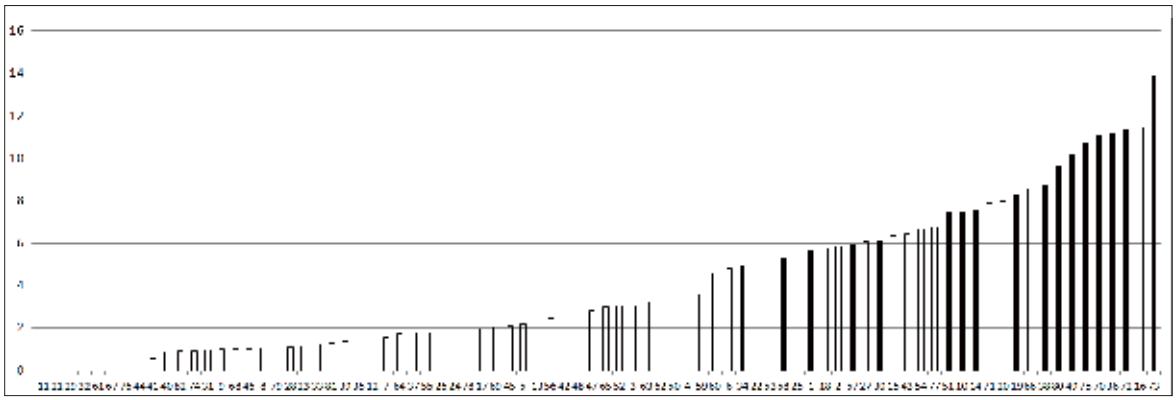

図 2　明細書を特徴付ける用語の出現頻度に基づく類似度

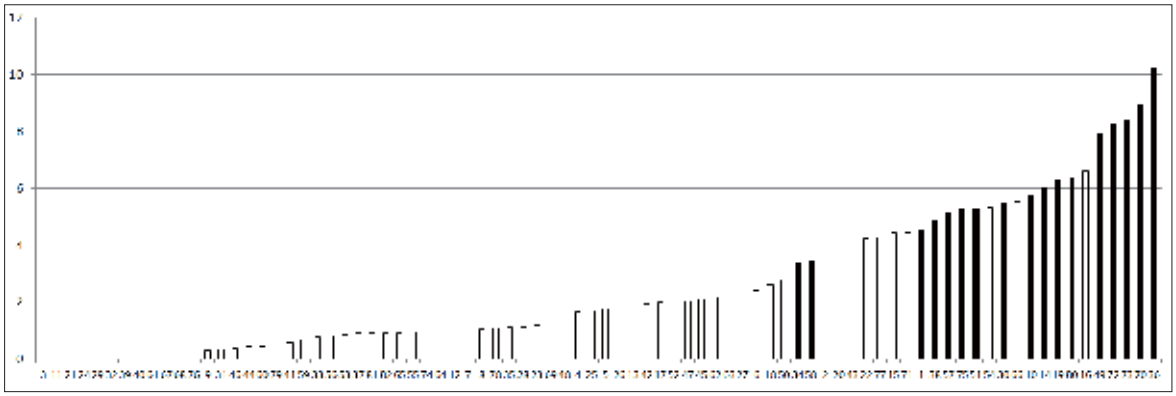

図３　用語の出現頻度の差に閾値を設定した類似度

度の領域に集中した。一方、類似度が十分に低いもの（本例では３未満）は、内容に関連性がない明細書だった。

5 考察

キーワードで全文検索された明細書の中には、着目した P0 との類似度が極めて低く算出されたものがあった。これらは内容が関連しないものとして、着目すべき対象から除外すると、P0 の内容と関連する可能性がある明細書は半数以下となる。しかし、実際に明細書の内容を確認すると P0 との類似度が高くても内容が関連しない明細書があった。本章では類似度が高く内容も関連していた P72 および P73 と、類似度が比較的高かった (5.5 以上) にもかかわらず、内容が関連していない P16 および P66 を考察する。

自然言語の文章を入力して類似する文書を検索する概念検索 [5] では、入力文と検索対象の文書から特徴語を抽出し、その重みづけによって類似度を算出する。[5] では特徴語として抽出した、明細書の【実施例】における用語は、発明の概念を表していないノイズもあり得るが、逆に【要約】にある用語は、これを追加すると検索精度が改善するので特徴的な用語である可能性が高いと報告されている。そこで表７において比較元の P0 に高頻度で出現する用語のうち、【実施例】以外の【特許請求の範囲】や【要約】にも出現している用語には○を付けた。そうすると一般的な用語ではない「ｒｅｖｉｓｅ」「ｎｕｍ」「ｒｅｖｉｓｅ属性」などが【実施例】だけに出現していることがわかる。以降の表においても同様に、ノイズの可能性が少ない【特許請求の範囲】や【要約】にも出現している用語には○を付けた。

表７　明細書 P0 において頻度の重みづけが 1 以上の用語

用語 t	頻度 n	$\log n_{t,P0}$
○ 要素	130	2.11
○ 改正	89	1.95
○ 改訂	73	1.86
○ 文書	54	1.73
条	50	1.70
revise	36	1.56
○ 属性	32	1.51
num	26	1.41
○ 原文	21	1.32
in	18	1.26
○ 改訂 ID	18	1.26
revise 属性	16	1.20
○ 改正 ID	15	1.18
del	14	1.15
○ 条例	12	1.08
○ 履歴	11	1.04
○ 改正作業	10	1.00

高い類似度を示すとおり、実際に内容も関連していた明細書の結果を表８および表９に示す。双方とも【実施例】だけでなく【特許請求の範囲】や【要約】にも共通に出現している用語が多い。特に用語「改正」は、その値が大きいので共通する特徴語であることがわかる。また、ここでは動詞化されるサ変名詞は動詞の一部として用語から除外している。例えば「条例を改正する」という文章では「改正」が用語として抽出されない。しかし類似度算出のための重みづけとしてこれを含めると類似度は更に大きくなると予想できる。

更に、表８の用語「改正内容」に着目すると、例えば、「改正情報」や「改正データ」といった同義語とは無関係になる。特に明細書においてこのような用語は多様であり、出願人も異なるために用語の統一は一層困難である。このための対応として「改正内容」「改正情報」「改

表 10　同一内容の用語を統一するために追加するパターンとルール

ルール	対象の形態素				直前の形態素				直後の形態素			
	単語	品詞	細分類 1	細分類 2	単語	品詞	細分類 1	細分類 2	単語	品詞	細分類 1	細分類 2
exclude	情報	名詞	一般	*	改正	名詞	サ変接続	*	*	*	*	*
exclude	内容	名詞	一般	*	改正	名詞	サ変接続	*	*	*	*	*
exclude	データ	名詞	一般	*	改正	名詞	サ変接続	*	*	*	*	*

表 8　明細書 P0 と P72 の類似度

用語 t	$\log n_{t, P0} + \log n_{t, P72}$
○ 改正	3.36
条	3.42
○ 削除	1.73
○ 追加	1.45
○ 改正内容	1.30
○ 変更	1.48
目	1.30
差分	0.60
$Similarity'_{P0, P72}$	(14.65/177.33)*100=8.26

表 9　明細書 P0 と P73 の類似度

用語 t	$\log n_{t, P0} + \log n_{t, P73}$
○ 改正	3.44
○ 文書	3.01
○ 条	3.32
属性	2.41
条例	2.08
○ 改正作業	2.11
○ 自動的	1.75
変更	1.48
○ 改正文	0.90
○ 改正箇所	1.26
○ 施行	0.95
編集	0.78
○ 入力	0.78
$Similarity'_{P0, P73}$	(24.27/289.1)*100 = 8.39

正データ」を統一した用語として「改正」とするには表 10 のようなルールを表 2 に追加する手段がある。しかし、このような可能性がある形態素の接続パターンすべてを洗い出し、共通の用語として類似度を算出する必要がある。これに対し、任意の自然言語テキストに存在する同義語それぞれの類似度を算出し、その結果を利用する手段が最も有効と考えられるが、そのためには大規模なコーパスを対象にした解析が必要である。[6] では、このとき課題となるノイズの低減方法の提案や自然言語解析技術による精度改善が報告されている。

一方、類似度が高いにもかかわらず、実際には内容が

表 11　明細書 P0 と P16 の類似度

用語 t	$\log n_{t, P0} + \log n_{t, P16}$
要素	3.58
○ 改訂	2.90
○ 文書	3.19
○ 属性	2.94
削除	1.80
条文	1.65
追加	1.75
タグ	1.45
差分	0.60
入力	0.60
$Similarity'_{P0, P16}$	(20.46/309.36)*100 = 6.62

表 12　明細書 P0 と P66 の類似度

用語 t	$\log n_{t, P0} + \log n_{t, P66}$
○ 文書	3.67
削除	1.43
○ 選択的	1.91
追加	1.45
○ 変更	1.74
$Similarity'_{P0, P66}$	(10.19/185.03)*100 = 5.51

関連していなかった明細書の結果を表 11 および表 12 に示す。P0 と P16 の類似度の表 11 では【実施例】だけに出現している用語の出現頻度の重みづけの合計が 11.43 であり、重みづけの総計 20.46 の 50% を超えている。このような用語をノイズとして差し引くと実際の類似度は小さくなる。

P0 と P66 の類似度の表 12 を見ると、P0 を最も特徴付けている用語「改正」が P66 では共通の用語として表示されていない。これが大きな理由だと考えられる。

本稿では、明細書の全文を範囲として用語を抽出したため、【実施例】だけに出現するノイズも類似度として算入されている。今回の結果から明細書においては、類似度算出時に対象とする用語の出現場所を明細書の全文ではなく、特徴的な用語が出現する可能性がある範囲に限定し、更に、共通に出現する名詞だけでなく、動詞化

される名詞とその出現頻度も加えた類似度を算出することによって、関連する明細書を効率よく発見できると考えられる。

6 おわりに

明細書双方の全文から抽出した、専門用語と一般用語とを区別しない用語すべてを対象に、共通に出現する用語の頻度の重みづけから明細書間の類似度を算出した。このとき明細書において汎用的に使用される用語を登録しておき、これらの用語は類似度算出の対象外とした。類似性の数量化を評価するには明細書の内容を理解する必要があるため、今回、検証対象の明細書の数と技術分野を限定した。評価した明細書において類似度が極めて低いものは一次的に自動フィルタリングができることがわかった。今後、検証対象の明細書と技術分野をもっと増やすとともに、第５章の考察で言及した課題に取り組んだ結果を評価する予定である。

使用したソフトウェア

① 形態素解析エンジン「MeCab」, Ver. 0.99 京都大学情報学研究科―日本電信電話株式会社コミュニケーション科学研究所共同研究ユニットプロジェクト

② IPA 品詞体系日本語辞書「IPADIC」, Ver. 2.7.0 奈良先端科学技術大学　松本研究室

③ 文書比較／日本語精査／改め文生成モジュール・ソフトウェア：「やまと歌」 http://www.crestec.co.jp/yamatouta

参考文献

[1] 楠本 , 山口 , 鈴木 , 千引： 特許翻訳の品質を向上するための形態素解析結果を利用した文書比較・日本語精査ツール－歌詠と鶯－の試作 平成 24 年度 AAMT/Japio 特許翻訳研究会 第２回特許情報シンポジウム資料集 pp. 17-24.

[2] 内山 清子： 専門用語の専門性判定に関する一考察 Japio YEAR BOOK 2010, pp.152-153, 2010.

[3] 小林 英司： 特許分類の自動推定に向けた取り組み Japio YEAR BOOK 2013, pp. 234-237, 2013.

[4] 海野 敏： 出現頻度情報に基づく単語重みづけの原理 Library and Information Science No.26 pp. 67-88, 1988.

[5] 間瀬 久雄： 特許概念検索における特徴語抽出に関する評価と考察 Japio YEAR BOOK 2011, pp.166-171, 2011.

[6] 相澤 彰子： 大規模テキストコーパスを用いた語の類似度計算に関する考察　情報処理学会論文誌, Vol.49, No.3, pp.1426-1436 (2008).