特許情報を専門用語の知識源として活用するシステム

藤井 敦
筑波大学大学院図書館情報メディア研究科
fujii@slis.tsukuba.ac.jp

1 はじめに

近年，知的な創造の成果を活用して産業の国際競争力を強化する動きがある．知的財産権の一つである特許権は，高度な発明の保護を目的としている．日本では年間約 40 万件の特許が出願され，多様な専門分野に関する発明が蓄積されている．特許に内在する人間の英知を体系化し，活用することができれば，今日の高度情報化社会において産業上の価値が高い．

特許には発明に関する新語や専門用語が多く含まれている．本研究は,「言葉」に関する知識に焦点を当て，特許情報から用語辞典的なコンテンツを自動構築することを目的とする．当コンテンツは，各見出し語について，説明，分野，関連語を記載する．

筆者は，Web から言葉や事柄に関する説明情報を抽出し，さらに複数の説明情報を組織化することで，百科事典的なコンテンツを自動構築する研究を行ってきた [1, 2, 5, 6] ．さらに，構築したコンテンツに対して，見出し語，関連語，質問文，可視化グラフなどの多様な手段によって検索する機能を開発した．当該研究成果は，事典検索システム Cyclone で一般に公開している[1]．

World Wide Web の発展に伴い，Web 上のポータルサイト，検索エンジン，ソーシャルネットワーキングなどのツールを使って，様々な調べ物をすることが日常的になっている．しかし，どのような専門用語でも Web 上で調べることができるわけではない．情報技術との関連性が低いような技術分野では，Web による情報発信が盛んではない場合がある．その結果，調べたい専門用語に関する情報が Web 上にないことがある．仮に，専門用語に関する情報が Web 上にあったとしても，論文や特許に記述されているような詳細な情報が得られるとは限らない．

以下に挙げる専門用語は，Web 上では説明を検索することができなかったのに対して，特許情報からは説明を見つけることができた例である．

- 「感光性平版印刷版」
  感光性平版印刷版とは、一般に、適当な表面処理を施したアルミニウム、紙あるいはプラスチックなどの支持体の表面に、感光性化合物を含有する感光層を設けたものである。
  （特開 2002-40631 より抜粋）
- 「トラッキング誤差信号」
  トラッキング誤差信号とは、対物レンズが光ディスクの半径方向に移動する場合、記録されたピットの中心で０となり、ピットの中心からずれるに従って、値が大きくなる信号である。
  （特開 2001-52347 より抜粋）
- 「塩基プレカーサー」
  塩基プレカーサーとは、加熱下で塩基を遊離する化合物をいい、塩基と有機酸の塩等が挙げられる。塩基プレカーサーを構成している塩基としては、前記塩基で例示したものが好ましい。
  （特開 2001-89475 より抜粋）
- 「マゼンタカプラー」
  マゼンタカプラーとは、Ｎ－エチル－Ｎ－（β－メタンスルホンアミドエチル）－３－メチル－４－アミノアニリン硫酸塩（ＣＤ－３）との酸化的カップリングによって生成する色素が、メタノール中での極大吸収波長が５００～６００ｎｍの範囲にあるカプラーをいい、
  （特開平 11-160840 より抜粋）

以上の背景から，本研究は Cyclone の研究で蓄積された技術やノウハウを応用して，特許情報から用語辞典的なコンテンツを自動構築し，専門用語の知識として活用するシステムを提案する．言葉に関する知識の抽出や質問応答などの調査型検索に関する研究において，特許情報が情報源として使用されることはない．この点において，本研究は既存の研究と差別化される．

2 辞典コンテンツの構築手法

本研究の基礎となる Cyclone は，以下の手順によって Web からコンテンツを構築する．

(1) 新語抽出
　Web から新しい見出し語を抽出する．
(2) ページ検索
　見出し語を含むページを Web から検索する．
(3) 説明抽出

[1] http:cyclone.slis.tsukuba.ac.jp/

検索されたページ集合から，見出し語について説明している可能性が高いテキスト部分を抽出する．

(4) 組織化

抽出された複数のテキストに対して，説明としての尤度に基づいてスコアを計算し順位を付ける．さらに分野に分類することで多義や観点による意味の違いを区別する．

(5) 関連語抽出

見出し語の説明によく使われる言葉を抽出する．

(6) 要約

説明テキストから，対象の見出し語を説明する観点ごとに代表文を選択する．その結果，ユーザは短い記述で種々の観点から見出し語の説明を読むことが可能になる．

以上の手順によって，ある見出し語に関する説明と関連語が抽出される．本研究では，上記の (1)〜(5) を特許情報に適応させた．

手順 (1) は，Web 用のプログラムをそのまま使用すると，見出し語として不適切な語も含めて，多数の見出し語候補が抽出されてしまう．まず，特許に頻出する言葉(「本発明」や「請求項」など）を見出し語の候補から削除するために，不要語リストを作成した．また，特許情報では，見出し語の先頭や末尾に特許特有の接辞を伴うことがある．例えば，先頭には「該」や「前記」，末尾には「等」や「近傍」などの接辞が付く．また，末尾には構成要素の番号を示す英数字が付くことがある．そこで，辞書や規則を用いて，見出し語候補の先頭や末尾から不要な文字列を削除する．

手順 (2) では，公開特許公報 1993〜2005 年発行分（文書数は約 456 万件）を対象に検索を行う．特許検索のエンジンは独自に開発した．索引付けは単語単位で行い，検索モデルとして Okapi BM25 [3] を用いる．具体的には，見出し語を含む特許公報を検索し，適合度のスコアを計算し，ソートする．

手順 (3) では，特許公報のレイアウトを解析して段落を抽出する.【請求項】や【発明の詳細な説明】などの墨付括弧で区切られた領域を段落として抽出する．ただし，予備調査の結果，要約と請求項には用語の説明がほとんど存在しなかったため，要約と請求項は段落として抽出しない．また，検索する特許公報の件数や抽出する段落の件数を制限しないと後続の処理に時間がかかる．現在は，検索する特許公報の件数は最大で 500 件とし，各公報から抽出する段落の件数は最大 10 件としている．段落が 10 件を超える場合は，特許公報中の出現順に基づいて先頭から 10 件を抽出する．

手順 (4) では，Web 用の手法が HTML タグやリンク構造など Web 固有の情報を用いているのに対して，特許情報ではこれらの情報を得ることができない．そこで，手順 (2) で計算した「Okapi BM25 のスコア」と「表現に基づくスコア」を統合して段落のスコアを計算する．

Cyclone を開発し拡張する過程で，用語説明に使われる表現やスコアを経験的に決めてきたので，これらを用いて「表現に基づくスコア」を計算する．説明固有の表現として格助詞と係助詞が連結した「とは」がある．しかし，説明以外の目的で使われる「とは」も存在する．例えば,「X とは異なり」や「X や Y とは Z で連結される」は，X や Y を説明するための表現ではない．そこで,「とは」を含む文が説明文か否かを判定するために，教師事例に基づく機械学習を用いる．具体的には,「とは」を含む文を Web と特許から 1 万件ずつ抽出し，人手で説明として使われているかどうかを判定して，正例と負例を作成した．これらの教師事例を用いて，サポートベクターマシン（SVM）によって 2 値分類器を学習した．当該分類器によって,「X とは」を含む文を X の説明とそれ以外に分類する精度は 93%である．SVM は分類のスコアを計算するので，当スコアを「表現に基づくスコア」に導入する．

分野の分類は Web 用の手法と同じである．具体的には，機械翻訳用の日英対訳辞書から 18 分野に関する統計モデルを構築し，分類に利用する．

手順 (5) では，Web 用の手法と同じように，語構成に関する規則によって段落から単語や複合語を抽出し，関連語としてのスコアを計算する [5]．ただし，特許情報に適用するために，手順 (1) の新語抽出と同じ手法によって，特許に頻出する言葉や特許特有の接尾辞を削除する．

手順 (6) では，Web 用の手法 [1] をそのまま使用する．

本稿執筆当時までに，専門用語集などから収集した技術用語を対象に，約 50 万語を見出し語として辞典コンテンツを構築した．構築した辞典コンテンツは，Web 用の Cyclone と同じ検索インタフェースを使用して，多様な手法によって検索することが可能である．図 1 に，見出し語「トラッキング誤差信号」に対する辞典コンテンツの検索結果を示す．図 1 の画面上部には，見出し語に関する「分野」,「関連語」,「複合語」が提示されている．ここでいう「複合語」とは，関連語のうち，見出し語を含む語である．さらに，画面下部には，見出し語に関する説明テキストが 2 件提示されている．

3 評価実験

3.1 実験方法

本研究で提案したシステムは複数の要素で構成されており，様々な観点からの評価が必要である．しかし，本稿では，抽出された用語説明の段落を順位付ける精度を評価した．当該機能は提案システムの最も重要な機能である．評価用の見出し語は情報処理用語辞典 [4] から抽出した．当該辞典には見出し語が 1408 語収録されている．また，各見出し語について説明があるため，正解判定の参考にした．

本研究の主旨は，Web では見つからないような用語の説明を検索する点にある．しかし，そのような用語は専門性が高く正解判定が困難であるため，今回は書籍として出版されている情報処理用語辞典から評価用の見出

図 1: 見出し語「トラッキング誤差信号」に対する辞典コンテンツの検索結果

し語を選択した．

1408 語のうち，説明の段落が 1 件以上抽出された見出し語は 1291 語あった．これら全てを評価対象にすると正解判定のコストが高い．そこで，以下の手順で評価用の見出し語を選択した．

まず，用語説明に「とは」が使用されている場合は，高い精度で用語説明を抽出できることが経験的に分かっている．他方において，「とは」を含む文が特許公報中に存在しない見出し語も存在する．抽出された段落に「（見出し語）とは」を含む見出し語と含まない見出し語の内訳を調べた結果，それぞれ 537 語と 754 語であった．そこで，各グループから評価用の見出し語を選ぶ必要があると考えた．

各グループから 20 語ずつ選択して評価用の見出し語とした．このときに，抽出される段落の件数が精度に影響する可能性がある．例えば，段落が 10 件しかない見出し語の場合は，順位付けの効果があまりない．そこで，抽出された段落の件数によって，100〜1000 まで 100 刻みで 10 個の階級に見出し語を分けて，各階級の中で段落が少ない方から 2 語ずつ選択した．以上をまとめると，「とは」を含むグループと含まないグループから 20 語ずつ，合計 40 語の見出し語を選択し，評価に使用した．

順位付けのスコアに Okapi BM25 のスコアだけを用いた手法（手法 1）と本手法（手法 2）を比較した．両手法の違いは，本手法では Okapi BM25 のスコアに加えて，表現に基づくスコアを用いている点にある．

両手法が個別に順位付けた段落から上位 10 件ずつを抽出して用語説明としての適否を評価した．このときに，どちらの手法が出力した段落か分かってしまうと正解判定に偏りが出る可能性がある．そこで，両手法が出力した段落を併合し，文字コードでソートした上で正解判定を行った．

正解の段階として，（A）単独で見出し語の説明になっている，（B）説明の一部にはなっているものの単独では不十分，（C）見出し語の説明になっていない，を用意した．（B）は，見出し語に関する内包的な定義がなく，具体例，機能，問題点など定義以外の観点から説明されている場合である．

評価尺度として，順位付けされた段落の上位 10 件に対する MRR（Mean Reciprocal Rank）を用いた．MRR は，見出し語ごとに，正解が最初に見つかった順位の逆数（Reciprocal Rank）を計算し，全見出し語の RR を平均した値である．正解が 1 位で見つかった見出し語の RR は 1 になり，正解が 2 位で見つかった見出し語の RR は 0.5 まで落ちる．上位 10 件に正解がなかった見出し語の RR は 0 にする．MRR は精度重視の質問応答などに対する評価に用いられる．しかし，MRR は最初に見つかった正解しか考慮しないため，上位 10 件に含まれる正解の総数も評価した．

表 1: MRR による評価

	（A）だけを正解とした場合			（A）と（B）を正解とした場合		
	「とは」あり	「とは」なし	合計	「とは」あり	「とは」なし	合計
手法 1	0.056	0	0.028	0.160	0.046	0.103
手法 2	0.558	0.365	0.462	0.904	0.591	0.748

表 2: 出力された正解の件数による評価

	（A）だけを正解とした場合			（A）と（B）を正解とした場合		
	「とは」あり	「とは」なし	合計	「とは」あり	「とは」なし	合計
手法 1	4	0	4	10	7	17
手法 2	66	17	83	105	49	154

3.2 結果と考察

各手法が出力した上位 10 の段落に対する MRR と正解数を表 1 と表 2 にそれぞれ示す．表 1 と表 2 から大きく 2 つのことが分かった．

1 つ目は，手法 1（Okapi BM25 のスコアだけを使用する）と手法 2（Okapi BM25 のスコアと表現に基づくスコアを併用する）を比較すると，手法 2 は手法 1 の結果を大きく上回った．手法 1 の結果から，通常の文書検索をして上位 10 件の段落を読む方法では，正解の用語説明はほとんど見つからないことが分かった．手法 1 と手法 2 の MRR における差に意味があるかどうかを確認するために t 検定を行った結果，有意水準 0.01 で有意であった．

2 つ目は，手法 2 に対する結果を見出し語のグループごとに見ると,「とは」を含む文が存在しない見出し語に対する結果は，存在する見出し語に対する結果よりも悪かった．このことから，説明文に「とは」が使用されているかどうかが用語説明抽出の精度に影響することが分かった.（A）と（B）の両方を正解と見なした場合は,「とは」が使用されない見出し語グループに対する MRR は 0.591 であり，平均して 2 位までに正解が見つかった．それに対して,「とは」が使用された見出し語グループに対する MRR は 0.904 であり，ほとんどの見出し語に対して正解が 1 位で見つかった．

正解ではない段落の傾向として，発明の特徴や構成といった発明の本質に関する記述が多かった．これらの記述では，用語の説明があったとしても対象となっている発明における特殊な意味であり，一般的な意味ではないことが多い．発明の本質に関する記述と一般的な記述を分類することができれば，用語説明抽出の精度がさらに向上する可能性がある．

4 おわりに

本研究は，特許情報から用語辞典的なコンテンツを自動構築し，多様な手法でコンテンツを検索するためのシステムを提案した．さらに，用語説明を順位付ける精度を評価し，既存の文書検索手法に対する優位性を示した．今後は，システムを構成する他の機能についても評価を行う必要がある．

構築したシステムは特許調査においても有効に機能する可能性がある．発明を特徴付ける重要語は，特許の明細書中で出願人によって定義や意味が積極的に記述されることがある．本システムを応用すると，TF.IDF といった語の統計頻度に基づく手法では区別できない用語の重要度を区別できる可能性がある．こうした応用について今後検討する価値があるだろう．

謝辞

本研究は，NEDO「産業技術研究助成事業」の助成で行った．

参考文献

[1] Atsushi Fujii and Tetsuya Ishikawa. Summarizing encyclopedic term descriptions on the Web. In *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 645–651, 2004.

[2] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Cyclone: An encyclopedic Web search site. In *Special Interest Tracks & Posters of the 14th International World Wide Web Conference*, pp. 1184–1185, 2005.

[3] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, 1994.

[4] 藤本喜弘（編）. 第二種・シスアド情報処理用語辞典. 経林書房, 1998.

[5] 藤井敦, 伊藤克亘, 石川徹也. Web マイニングによる事典的コンテンツの構築と多様なアクセス手法. 電子情報通信学会技術研究報告, DE2004-6, pp. 31–36, 2004.

[6] 藤井敦, 石川徹也. World Wide Web を用いた事典知識情報の抽出と組織化. 電子情報通信学会論文誌, Vol. J85-D-II, No. 2, pp. 300–307, 2002.