技術動向調査と無効資料調査を対象とした特許検索の技術動向

東京工業大学大学院情報理工学研究科准教授 **藤井　敦**

PROFILE

1998 年東京工業大学大学院博士課程修了。博士（工学）。筑波大学大学院准教授等を経て、2009 年より現職。自然言語処理、情報検索、Web マイニング、特許情報処理の研究に従事。2009 ～ 2010 年度、特許版産業日本語委員会委員。

1 はじめに

筆者らが NTCIR ワークショップ（国立情報学研究所主催）において特許検索タスクを始めてから 10 年が過ぎた。2001 ～ 2010 年にかけて特許情報の検索、分類、分析、翻訳に関するテストコレクションを構築し [1]、一連の成果は国内外における特許情報処理の研究開発に波及してきた。欧州の評価ワークショップの CLEF では、NTCIR 特許検索タスクの方法論を踏襲して，特許検索用テストコレクションが構築された [2]。本稿は、NTCIR を通して提案された特許検索の技術動向について解説する。

NTCIR 特許検索タスクでは、技術動向調査と無効資料調査を対象として、ベンチマークテスト用のテストコレクションを構築した。技術動向調査では、技術系の新聞記事を検索質問とし、当該技術に関する公開特許公報を検索対象とした。無効資料調査では、ある公開特許公報中の請求項 1 つを検索質問として、そこで請求されている権利を無効化できる別の公開特許公報を検索した。

NTCIR における 1 回のワークショップは，1 年半をかけて行われる。本稿では，第 n 回のワークショップを NTCIR-n と略記する。

2 特許検索を題材とした情報検索技術の比較評価

NTCIR のような評価ワークショップでは、参加者ごとのシステムを比較することが容易である。しかし、システムを構成する要素技術の単位で比較することは容易でない。Iwayama ら [3] は、NTCIR の技術動向調査用テストコレクションを用いて、既存の汎用的な情報検索手法を要素技術の単位で網羅的に比較評価した。以下、当実験によって得られた知見についてまとめる。

- 検索質問の種類： ユーザの情報要求を丁寧に文章化する方が新聞記事で代用するよりも有効であった。
- 索引付けの対象： 要約や請求項だけでなく公報全文の索引付けが有効であった。特許公報は他の文書に比べて長いため、索引付けのコストも大きくなる。しかし、コストをかけて公報全文を索引付けすることに意義があることが分かった。
- 索引語の単位： 文字バイグラムよりも単語に基づく索引付けが有効であった。さらに、文字バイグラムと単語を最適な重みで混合すると、個別に使用した場合よりも有効であった。
- 索引語の重み付け： TF.IDF の単純な計算式よりも、SMART や Okapi BM25 といった先端のシステムで使用されている計算式が有効であった。

Fujita [4] は、NTCIR の特許検索テストコレクションを用いて、擬似適合性フィードバックの有効性について考察した。技術動向調査では擬似適合性フィードバックが有効だったのに対して、無効資料調査では有効性が確認されなかった。擬似適合性フィードバックでは初回の検索結果に適合文書が高順位で含まれることを前提としている。しかし、無効資料調査では適合文書の件数が少ないため、この前提が成り立ちにくいことが原因である。

3 索引語の重み付け

NTCIR の技術動向調査タスクにおいて、Itoh ら [5] は Term Distillation という索引語の重み付け手法を提案した。検索質問の新聞記事と検索対象の特許公報では索引語の出現頻度分布が異なる。そのため、「一般的な用語ほど重要ではない」という考えに基づいて単語の重要度を計算する TF.IDF 等の手法を用いると、同じ単語でも新聞記事と特許では重要度が異なる。Itoh らは、単語の新聞記事集合中での出現頻度と特許中での出現頻度の違いを考慮して重み付けを行った。

Mase ら [6] は、速度や温度などの数量表現に大きな重みを与えた。「度」や「長」といった接尾辞に基づいて、「密度」や「波長」などの数量表現を含む辞書を作り、この辞書を用いて特許公報中の数量表現を特定し、さらに文中の修飾関係によって数量表現の対象語を特定する。数量表現と対象語の両方に対して大きな重みを与える。例えば、「用紙の搬送速度を制御する」という記述では、「速度」が数量表現として抽出され、「用紙」と「搬送」が対象語として特定される。

4 検索質問の構造解析

NTCIR の無効資料調査タスクでは、検索質問として使用する請求項の構造を解析して、特許検索に利用する手法が提案された。

Mase ら [6] は、ジェプソン形式に基づく請求項の構造に着目した。ジェプソン形式の請求項は、従来技術に関する「前提部」と本発明に関する「本質部」で構成される。無効資料調査の目的を考えれば、前提部よりも本質部の方が検索に有効な情報である。前提部よりも本質部を重視するために、Mase らは、まず請求項全てを検索質問として網羅性を重視した検索を行い、次に本質部だけを検索質問として有望な公報に絞り込む 2 段階検索の手法を提案した。

高木ら [7] は、構成要素に基づく請求項の構造に着目して、重要な構成要素を重視する検索手法を提案した。まず、人手で作成した規則を用いて請求項を構成要素に分割する。次に、情報理論に基づいて索引語の重要度を計算し、索引語の重要度に基づいて構成要素の重要度を計算する。さらに、構成要素ごとに検索質問を構成して個別に特許公報を検索し、構成要素の数だけ文書の順位付きリストを作る。最後に、構成要素の重要度を考慮して複数の文書リストを 1 つに統合する。1 つの文書は構成要素ごとの文書リストにおいて異なるスコアを持つため、重要な構成要素によって高いスコアが与えられた文書には高いスコアを与える。高木らは、構成要素の重要度を計算する際に、Mase らと同様にジェプソン形式の請求項構造も利用している。すなわち、人手で作成した規則を用いて請求項を前提部と本質部に分割し、前提部に含まれる構成要素の重要度を小さくする。

5 検索質問拡張

NTCIR 特許検索タスクの無効資料調査では、従来とは異なる発想に基づく検索質問拡張の手法が提案された。NTCIR の無効資料調査では、検索質問は特許公報から抽出された請求項である。特許公報には請求項以外の記述も含まれるため、それらの記述から索引語を抽出し、検索質問を拡張する手法が考えられる。ここでは説明の都合上、検索質問の請求項を抽出した特許公報を「元公報」と呼ぶ。

Konishi ら [8] と Fujii ら [9] は、請求項に特有の性質を利用して検索質問拡張を行った。請求項では権利の範囲を広げるために「移動体」のような抽象的な用語が使われるのに対して、実施例では「自動車」や「飛行機」のような具体的な用語が使われる傾向にある。そこで、請求項中の用語だけを用いても所望の公報を的確に検索できるとは限らない。より具体的な索引語によって検索質問を拡張するために、請求項の内容に対応した記述を元公報において特定する手法が必要である。

Konishi らは、人手で作成したパターンに基づいて、実施例から請求項中の抽象的な用語と具体的な用語を対

応付ける文を特定した。例えば、「フラッシュメモリやROM等の記憶手段」という記述では「記憶手段」が請求項中の用語であれば、「フラッシュメモリ」や「ROM」が具体的な用語であることが分かる。

Fujiiらは、元公報を段落に分割し、検索質問の請求項を用いて元公報内の段落を検索した。その結果、Konishiらのように人手で作成したパターンを用いることなく、検索質問の請求項と索引語を共有する記述を元公報から抽出した。

6 メタ情報の利用

特許公報中のテキスト情報だけでなく、分類コードや引用関係などのメタ情報を利用する検索手法について説明する。

特許公報に付与された国際特許分類（IPC）等の分類コードに基づいて複数の文書をまとめることで、個別の文書から得られる疎な統計情報を平滑化する効果が期待できる。Kangら[10]は、言語モデルに基づく検索モデルにおいて、本来は自動的に作られる文書クラスタの代わりにIPCの分類コードを利用した。Konishiら[8]は、索引語の重み付けにおいて、文書の代わりに分類コードを単位としたTF.IDFを計算し、文書に基づくTF.IDFと統合した。

Fujii[11]は、米国特許公報の引用関係をウェブページのリンクと見なして、リンクに基づくインターネットの検索モデルを特許検索に応用した。具体的には、本文に基づく文書スコアとリンクに基づくスコアを統合した。しかし、検索質問と無関係に文書集合全体の引用関係から計算されるPageRank値よりも、まず本文を用いて公報を検索し、検索された公報間の引用関係だけを用いた方が有効であった。

7 展望

インターネット検索では、情報検索の研究で提案された最先端の技術が次々と実用化されている。それに対して、特許調査の現場では伝統的なブーリアンモデルに基づく検索や分類コード等の非キーワード情報による検索が使われている。ここでは、その原因について考察しながら、特許検索技術のさらなる発展について展望する。

インターネット検索と特許検索では、ユーザの要望が異なる。インターネット検索では、システム主導で容易に操作ができ、誰でもそれなりの結果が得られることを目指した研究開発が多い。それに対して、特許調査の現場で求められるのは、ユーザ主導でユーザの能力に追従して検索結果の品質が向上するようなシステムである。すなわち、「素人志向」のシーズと「玄人志向」のニーズにおける齟齬を解消する必要がある。

文書に順位を付ける検索システムは、上位の文書ほど所望の情報を含んでいる可能性が高くなるように設計されている。しかし、具体的に上から何件まで調査すればよいのかが明確ではない。1回の特許調査において、専門家は100から1000の単位で文書の内容を精査する。インターネット検索のように上位10件以内に適合文書を含めることよりも、ユーザが調査する件数の範囲内に適合文書を網羅的かつ確実に含めることが重視される。すなわち、「ここまで調査すれば十分」という境界をシステムが明示しなければならない。

インターネット検索では、過去に入力された検索質問やユーザにクリックされた文書に関する履歴情報がシステムの改良に使用されている。具体的には、検索質問の入力支援、各種の計算式におけるパラメタ値の最適化、計算式自体の学習などに応用されている。特許検索においても検索履歴を活用した研究[12]が行われており、今後のさらなる発展が望まれる。

情報検索の技術が発展してシステムの仕組みが複雑になればなるほど、各文書が検索された理由がユーザにとって分かりにくくなるという問題がある。ユーザが行った調査の妥当性を第三者に対して客観的に説明することも困難になる。検索結果を分かりやすく見せる可視化に関する研究は行われている。今後は、検索の過程を可視化する技術が必要になるだろう。

参考文献

[1] 藤井敦．特許検索と特許翻訳を指向したテストコレクションの構築研究．Japio 2007 Year Book, pp.156-159, 2007.

[2] Erik Graf and Leif Azzopardi. A methodology for building a patent test collection for prior art search. Proc. EVIA, pp.60-71, 2008.

[3] Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Yuzo Marukawa. Evaluating patent retrieval in the third NTCIR workshop. Information Processing & Management, 42(1), pp.207-221, 2006.

[4] Sumio Fujita. Technology survey and invalidity search: A comparative study of different tasks for Japanese patent document retrieval. Information Processing & Management, 43(5), pp.1154-1172, 2007.

[5] Hideo Itoh, Hiroko Mano, and Yasushi Ogawa. Term distillation in patent retrieval. Proc. ACL-03 Workshop on Patent Corpus Processing, pp.41-45, 2003.

[6] Hisao Mase, Tadataka Matsubayashi, Yuichi Ogawa, Makoto Iwayama, and Tadaaki Oshio. Proposal of two-stage patent retrieval method considering the claim structure. ACM Transactions on Asian Language Information Processing, 4(2), pp.186-202, 2005.

[7] 高木徹，藤井敦，石川徹也．検索質問の主題分析に基づく類似文書検索と特許検索への応用．情報処理学会論文誌，46(4), pp.1074-1081, 2005.

[8] Kazuya Konishi, Akira Kitauchi, and Toru Takaki. Invalidity patent search system of NTT DATA. Proc. NTCIR-4, 2004.

[9] Atsushi Fujii and Tetsuya Ishikawa. Document structure analysis in associative patent retrieval. Proc. NTCIR-4, 2004.

[10] In-Su Kang, Seung-Hoon Na, Jungi Kim, and Jong-Hyeok Lee. Cluster-based patent retrieval. Information Processing & Management, 43(5), pp.1173-1182, 2007.

[11] Atsushi Fujii. Enhancing patent retrieval by citation analysis. Proc. ACM SIGIR, pp.793-794, 2007.

[12] 乾孝司，難波英嗣，橋本泰一，藤井敦，岩山真，橋田浩一．最大クリーク探索に基づく特許検索履歴の統合．言語処理学会第 17 回年次大会発表論文集, pp.1059-1062, 2011.