特許概念検索における特徴語抽出に関する評価と考察

統計的解析から発明概念に基づく意味的解析へ

株式会社日立製作所　中央研究所　**間瀬　久雄**

PROFILE

1990年に株式会社日立製作所入社。システム開発研究所に配属、2011年度より中央研究所にて勤務。特許や新聞記事、Webページ等を対象とした、分類自動付与、検索、文章要約、テキストマイニング等の日本語処理の研究に従事。2007年度から特許版産業日本語委員会委員。

hisao.mase.qw@hitachi.com

1　はじめに

任意の自然言語文章を入力して、内容の類似する文書を検索する概念検索（類似文書検索、自然言語検索とも言う）が普及してきている。概念検索は、複雑な検索論理式を入力する代わりに、検索したい発明内容を文章で入力することによって、その内容に類似する文書を検索できることと、検索結果が類似する度合いの高い順に出力されるので、所望の文書に早く到達できることが特長である。

一般の概念検索ではまず、入力文章からその内容を特徴付ける語（特徴語）を抽出し、その重要度を表す重みを付与する。次に、検索対象となる文書の各々から予め抽出しておいた重み付き特徴語を格納した検索インデクスと照合し、共通する特徴語の重みに基づいて、入力文章と各文書との間の類似度を算出する。最後に、類似度の高い順に文書を並べ替えて検索結果を出力する。

概念検索の精度を向上させるためには、入力文章から適切な特徴語を選定し、適切な重みを付与することが必須となる。概念検索システムの多くでは、いわゆる「統計的解析」によって、特徴語の重みを算出している。すなわち、入力文章における出現頻度（Term Frequency，以下 TF と呼ぶ）が高い特徴語、または、検索対象文書集合においてその特徴語が出現する文書数の逆数（Inverted Document Frequency，以下 IDF と呼ぶ）が大きい特徴語に対して、重みを高くしている。

一方で筆者は、Japio YEAR BOOK 2010 において、特許を対象とした概念検索の技術課題を整理した[1]。その中で、概念検索精度を低下させる要因の一つとして、「計算機が発明概念を正確に特定できないため、適切な特徴語を抽出できていない」ことを挙げた。ここで、発明概念とは、「要するにこの発明のポイントは何なのかを端的に記述した内容」である。すなわち筆者は、統計的解析だけでなく、意味的解析によって適切な特徴語を選定する必要があると主張している。

そこで本稿では、特許を対象とした概念検索システムにおいて、まず「統計的解析」によって抽出される特徴語がどのくらい妥当であるかを、発明概念を考慮した「意味的解析」によって選定される特徴語と比較することによって評価する。次に、意味的解析によって選定された特徴語を概念検索に使用した場合の有効性について評価・考察する。

2　統計的解析による特徴語抽出の妥当性評価

2.1　評価方法

情報検索に関するタスク型国際ワークショップ NTCIR-5 特許検索タスク[2][3]で使用されたフォーマルラン課題データ 619 件のうち、2002 年に公開された 107 件の公開特許公報を入力文章（以下、クエリと呼ぶ）として、これらから抽出される特徴語の妥当性を評価した。検索対象特許は、公開特許公報、特許公報など 19 年分（1993 年～ 2011 年、約 940 万件）とした（検索対象特許件数は、IDF の算出に影響すること

に留意されたい）。

妥当性評価は以下の方法で行った。

（1）クエリの全文から統計的解析によって、重みの高い上位 10 個の特徴語を抽出する。

（2）筆者がクエリの【要約】を読んで、「発明概念に相当する記載箇所」を手作業で特定し、そこに含まれる特徴語を抽出する。

（3）上記（1）と上記（2）の特徴語に共通する特徴語の数の割合を算出する。

【課題】自然言語文章を入力とする文書検索において、どのような視点で文章を入力したかという観点を考慮した検索がされていないため、入力文章の特性を十分に活かした検索ができておらず、検索精度も十分に高いものとなっていない。
【解決手段】入力文章の観点集合を利用者に提示し、観点毎にターム抽出方法、重み付け方法、検索範囲を予め定義した観点テーブルを参照して、選択された観点に応じてターム抽出、重み付け、検索範囲を動的に変更して検索を実行する。

～：発明対象　＿：発明前提　＝：課題効果　……：解決手段

図 1　発明概念特定の例 1（特開 2007-102723）

【目的】文章の構成、分野／内容の違いによって加工精度が低下しないための文章加工方法および装置を提供することにある。
【構成】記憶装置に、単語辞書 4 と、文法辞書 5 と、複数の属性項目からなる文章属性データと、文章解析ルール 6 と、複数の文章構成タイプ識別ルール 9 と、複数の文章内容タイプ識別ルール 12 と、複数の加工方法設定テーブル 15 と、加工ルール 18 を設定し、文章解析部 3 で、入力電子文書を前記 4、5、6 と文章属性データに基づき解析し、文章解析テーブル 7 を生成し、文章構成タイプ推定処理部 10 で前記識別ルール 9 と文章解析テーブル 7 の内容に基づき入力電子文書の文章構成タイプを推定し、同様に文章構成タイプ推定処理部 13 で文章内容タイプを推定し、加工方法設定処理部 16 で前記推定したタイプの組に対応する加工方法設定テーブル 15 を選択し、該テーブル 15 と加工ルール 18 に基づき入力電子文書を加工する。

～：発明対象　＿：発明前提　＝：課題効果　……：解決手段

図 2　発明概念特定の例 2（特開平 8-272826）

表 1　抽出された特徴語の字種別割合

1 文字目の字種	該当特徴語数	割合
漢字	627	58.6%
カタカナ	340	31.8%
英文字	96	9.0%
ひらがな	3	0.3%
その他	4	0.4%
合計	1070	100.0%

表 2　抽出特徴語の比較 1（特開 2007-102723）

統計的解析による特徴語 10 語（重み順）（明細書全文を形態素解析）（数字は出現頻度）		意味的解析による特徴語 24 語（出現順）（図 1 下線を形態素解析）
ターム	148	自然　言語　文章　入力
該観	7	文書　検索　観点　考慮
文章	129	精度　十分　高い　文章
インデクス	12	集合　利用　者　提示
ブックエンド	5	選択　応じる　ターム
IDF	7	抽出　重み付け　範囲
重み付け	62	動的　変更
文書	104	
品詞	9	
単語	24	

※共通する特徴語はターム、文章、重み付け、文書

表 3　抽出特徴語の比較 2（特開平 8-272826）

統計的解析による特徴語 10 語（重み順）（明細書全文を形態素解析）（数字は出現頻度）		意味的解析による特徴語 16 語（出現順）（図 2 下線を形態素解析）
文章	628	加工　精度　低下　文章
単語	69	入力　電子　構成
ルール	163	タイプ　推定　内容
NUMOFSENTS	2	組　対応　方法　設定
記事	42	テーブル　選択
為替	8	
属性	128	
文	60	
文法	12	
辞書	25	

※共通する特徴語は、文章のみ

ここで、「発明概念に相当する記載箇所」は、公開特許公報の【要約】において、以下の 4 種類のいずれかに相当する記載箇所とした。

①発明対象（発明内容が作用する対象物）

②発明前提（発明内容が作用する状況）

③課題効果（発明内容が解決する課題・効果）

④解決手段（課題解決のポイントとなる手段）

発明概念の特定結果の例を図１および図２に示す。下線が施された記載箇所が発明概念に相当する。

２.２　結果と考察

表１は、107 件のクエリの各々から抽出された上位 10 個の特徴語（合計延べ 1070 個）を字種別にまとめた時の割合を示している [1)]。特徴語全体の 58.6% を漢字特徴語が、31.8% をカタカナ特徴語が、9.0% を英文字特徴語が占めている。カタカナ特徴語および英文字特徴語は、漢字特徴語に比べて固有度が高く、IDF が比較的高くなり、その結果として重みが高くなると考えられる。

表２は、図１に相当するクエリから、統計的解析および意味的解析によって抽出された特徴語の一覧を示している。統計的解析によって抽出された特徴語は、TF が非常に高いものと低いものが混在している。「ターム」「文章」「重み付け」などは TF が比較的高く、逆に、「ブックエンド」「 IDF」などは IDF が比較的高いために、重みが高くなっている。一方、意味的解析による特徴語と比較すると、共通する特徴語は４語に留まっており、両者には差が見られる。統計的解析による特徴語だけでは、発明のポイントを十分に記述できていないと考えられる。

表３は、図２に相当するクエリから、統計的解析および意味的解析によって抽出された特徴語の一覧を示している。共通する特徴語は「文章」のみであり、特徴語に大きな差が見られる。統計的解析による特徴語のうち、「NUMOFSENTS」「記事」「為替」は【実施例】にのみ現れる語であり、発明概念を表していないノイズ語である。一方で、この発明のポイントである「構成」「内容」などの特徴語は、統計的解析では抽出できていない。

このように、統計的解析によって上位 10 語に入る特徴語をミクロに見てみると、特徴語は以下の二つに大別できる。

（１）他の文書に現れにくい（IDF が高い）語

化学物質や人名、地名などの固有名詞に加え、出願人が明細書の中で独自に定義した造語や、英単語、異体字表記や誤字・脱字などは、他の文書に現れにくいため、IDF が非常に高くなり、重みが高くなる。この傾向は、検索対象文書数が多くなるほど顕著に現れる。したがって、表３の「NUMOFSENTS」のように、特徴語としては非常に違和感のあるもの（ノイズ）となる。しかし、概念検索精度の観点から見ると、これらの特徴語にヒットする文書数が少ないので、さほど精度低下には影響しないと考えられる。

（２）クエリに何度も現れる語

発明内容を端的に表す語や、発明内容が適用される技術分野に関係する語は、クエリの明細書の中で何度も出現することが多い。したがって、TF の値が非常に大きくなり、重みが高くなる傾向にある。しかし、これらの単語が発明内容を表す語としてふさわしくないノイズ語で、かつ、他の文書にも出現しやすい（IDF が小さい）語であると、検索精度に大きな悪影響を及ぼすことになる。例えば、表２の中の特徴語「記事」は、【実施例】で文書タイプの一つとして使われている特徴語であり、発明概念を表しているとは言えないノイズ語であるが、他の文書にも比較的現れやすい特徴語であるので、多くのノイズ文書を出力させてしまっている可能性が高い。

図３は、各クエリから抽出された特徴語 10 語（以下、クエリ特徴語と呼ぶ）に占める、発明概念に含まれる特徴語（以下、発明概念特徴語と呼ぶ）の割合の分布を示している。発明概念特徴語は、１クエリあたり平均 19.5 語抽出された。１件のクエリ特徴語に含まれる発明概念特徴語の割合の平均は、33.9% に留まった。残りの 66.1% の特徴語には、発明概念とは関係の薄い特徴語が多く含まれており、これらが検索精度を低下させている恐れがある。

図４は、発明概念特徴語に占める、クエリ特徴語の割合の分布を示している。その平均値は、18.3% に留まった。残りの 81.7% の発明概念特徴語は、特徴語として重視されていない（重みが低い）、あるいは特徴語として抽出されていないため、これらの特徴語の抽出漏れが

1）特徴語抽出において使用している形態素解析ツールや抽出アルゴリズムは、概念検索システムによって異なるので、抽出される特徴語はシステムによって多少異なる。

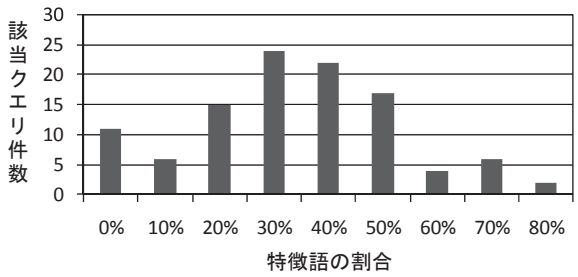

図３　クエリ特徴語に占める発明概念特徴語の割合

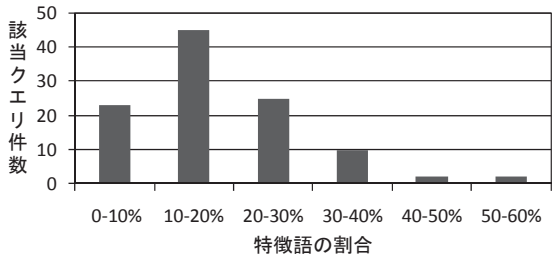

図４　発明概念特徴語に占めるクエリ特徴語の割合

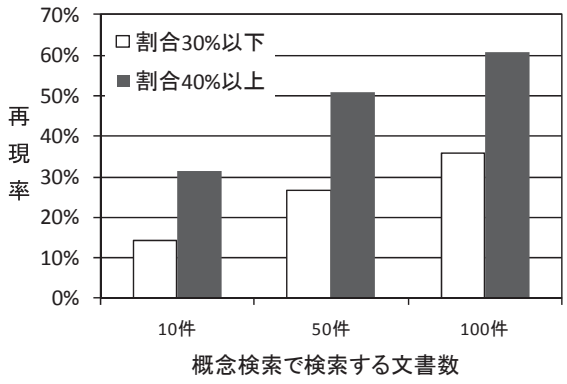

図５　発明概念特徴語の割合と概念検索再現率

検索精度を低下させている恐れがある。

図５は、図３において、クエリ特徴語に占める発明概念特徴語の割合が大きい（40% 以上）クエリ群と、割合が小さい（30% 以下）クエリ群に分けた場合に、概念検索精度がどのくらい違うかを再現率（検索結果文書Ｎ件に含まれる正解特許（審査官引例）の割合）で比較評価した結果を示している。発明概念特徴語が多く含まれているクエリ群の方が、15 ～ 20 ポイントも再現率が高くなっていることがわかる。このことから、概念検索では発明概念特徴語を積極的に使うことにより、検索精度の向上につながる可能性が高いと考えられる。

3 発明概念を考慮した概念検索の精度評価と考察

発明概念を特定できるとして、クエリから抽出した特徴語に発明概念中の特徴語を追加して特徴語の質を向上させることにより、概念検索精度がどのくらい改善するかを評価した。具体的には、２章の評価で用いた 107 件のクエリに対して、以下の方法で発明概念特徴語をクエリ特徴語に追加して、検索精度を比較した。

（1）クエリから特徴語を 30 語抽出する。

（2）クエリから発明概念特徴語を抽出する。

（3）30 語のクエリ特徴語のうち、発明概念特徴語にも含まれている特徴語の重みを２倍にする。

（4）発明概念特徴語のうち、30 語のクエリ特徴語に含まれていない特徴語をクエリ特徴語に追加する。その際、重みとして、クエリ特徴語の各々が持つ重みの最小値を与える。

なお、検索結果は、各クエリの出願日よりも前に公開された公報のみを出力している。

表４は、概念検索精度を比較した結果を示している。発明概念特徴語を追加して概念検索を行うことによって、概念検索精度が 4.7 ～ 10.3 ポイント向上していることがわかる。

表４　発明概念を考慮した場合の概念検索精度

検索件数	再現率	
	全文から抽出した特徴語 30 個	特徴語 30 個に発明概念特徴語追加
10 件	20.6%	26.2%
50 件	33.6%	38.3%
100 件	42.1%	48.6%
300 件	51.4%	61.7%

また、表４を見ると、検索件数が多い場合の方が、再現率の改善度合が大きくなっている。発明概念は、発明内容を端的に記載しているものであるので、検索件数が少ない場合の再現率がもっと改善されてもよいのではと考えるのが普通である。これを妨げている原因は、大きく２種類あると考えられる。

一つは、発明概念特徴語に付与する重みの値である。今回は、クエリ特徴語の重みの最小値を与えており、発明概念特徴語を「補足的に」追加する形で重みを設定している。クエリ特徴語に対して、発明概念特徴語の重みが相対的にもっと大きくなるように重みを設定することにより、精度が大きく変わる可能性がある。

もう一つは、同義語である。クエリにおける発明概念特徴語の表記が、正解文書での表記と異なっていると、いくら良質な発明概念特徴語を追加しても、ノイズ語になるだけである。本手法のように、概念検索に用いる特徴語をピンポイントに絞れば絞るほど、同義語が概念検索精度に与える影響が大きくなるので、同義語の吸収は必須となる。

4 おわりに

特許を対象とした概念検索で使われる特徴語の妥当性について評価した。出現頻度などの統計的解析だけでは、必要十分な特徴語を抽出するのは難しいことを述べた。また、発明概念を考慮して特徴語を抽出・重み付けをすることにより、概念検索精度を向上できることを述べた。すなわち、意味的解析による特徴語抽出は、概念検索精度の向上に有効であることを述べた。

本稿では、システムが発明概念を特定できることを前提として評価を行ったが、利用者とのインタラクションによって、利用者が発明概念の記載箇所を明示的に指定してもよいと考えている。また、発明概念特徴語の代わりに、【要約】や【請求項１】に含まれる特徴語を「すべて」クエリ特徴語に追加するという擬似的な方法でも、特徴語全体としての質は低下するものの、検索精度の改善には有効である[4]。

発明概念の自動抽出については、NTCIR-8 特許マイニングタスクでも採りあげられている[5][6]。機械学習アルゴリズムの導入により、２．１節で挙げた４種類の発明概念のうちの「課題効果」については高精度の抽出が可能である[7][8]。一方、「解決手段」については、高精度に抽出することは比較的難しい。これは、解決手段は複数の構成要素からなっており、その表記方法も多岐に亘っているため、どの構成要素が発明のポイントであるかを機械的に特定するのが難しいためである。発明概念を人間でも機械でも容易に特定・理解するための請求項の書き方については、Japio が主催している産業日本語委員会でも、特許オントロジー[9] というテーマとして議論されているので、ここでの今後の成果に期待したい。

参考文献

1）間瀬：特許を対象とした概念検索の技術課題，Japio YEAR BOOK 2010, pp.200-207, 2010.

2）N. Kando: Overview of the Fifth NTCIR Workshop, Proceedings of NTCIR Workshop 5 Meeting, 2005.

3）A. Fujii, M. Iwayama and N. Kando: Overview of Patent Retrieval Task at NTCIR-5, Proceedings of NTCIR Workshop 5 Meeting, 2005.

4）H. Mase and M. Iwayama: NTCIR-6 Patent Retrieval Experiments at Hitachi, Proceedings of NTCIR Workshop 6 Meeting, 2007.

5）N. Kando: Overview of the Eighth NTCIR Workshop, Proceedings of NTCIR Workshop 8 Meeting, 2010.
6）H. Nanba, A. Fujii, M. Iwayama and T. Hashimoto: Overview of the Patent Mining Task at the NTCIR-8 Workshop, Proceedings of NTCIR Workshop 8 Meeting, 2010.
7）R. Nishiyama, Y. Tsuboi, Y. Unno and H. Takeuchi: Feature-Rich Information Extraction for the Technical Trend-Map Creation, Proceedings of NTCIR Workshop 8 Meeting, 2010.
8）H. Nanba, T. Kondo and T. Takezawa: Hiroshima City University at NTCIR-8 Patent Mining Task, Proceedings of NTCIR Workshop 8 Meeting, 2010.
9）Japio 特許情報研究所：平成22年度特許版・産業日本語委員会報告書「産業日本語」, 第2章, 2011.