特許概念検索結果の理解支援に関する考察

―特許検索における概念検索の活用促進に向けて―

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、自然言語検索などとも言う）が普及してきている。一般の概念検索では、入力文章に出現する特徴語集合と、検索対象文書中の特徴語集合を比較して類似度を算出し、類似度の高い文書から順に出力する。

概念検索の長所は、（1）検索条件として複雑な検索論理式を入力する代わりに然言語文章を入力するので、検索条件を作成する作業負荷が少ないことと、（2）入力文章との類似度が高い文書から順にランキング出力するので、所望の文書にいち早く到達できることである。

しかし一方、概念検索では、入力文章に出現する数多くの特徴語を用いて類似度を総合的に算出するので、検索結果としてなぜその文書がその順位で出力されたのか、入力文章とどこがどのくらい類似しているかなどの検索根拠を、利用者が直感的に理解するのが難しいという短所がある。

本稿では、この短所を克服し、概念検索の活用をさらに促進すべく、概念検索結果の理解支援の実現可能性について考察する。まず２章では、利用者が概念検索結果を理解するとはどういうことかについて私見を述べる。次に３章では、理解支援の一例として、「特徴語－文書マトリクス」の活用について述べ、その有効性について考察する。

2　概念検索結果を理解するとはどういうことか？

本稿では、「利用者が概念検索結果を理解できる」とは、「検索がうまくいったのか」「なぜうまくいかないのか」「どうしたら検索がうまくいくのか」のすべてを理解できることであると定義する。そして、概念検索結果を理解できると、図１に示すように、必要に応じて概念検索をスパイラルに繰り返すことで、所望の文書を効率的にかつ高精度に検索できるようになると考える。

（1）結果理解：検索がうまくいったかを理解できる

概念検索結果の上位Ｎ件に所望の文書が含まれているか（含まれていそうか）を理解するプロセスである（Ｎの値は検索業務内容にも依存するが、10 ～ 50 程度であろう）。

検索論理式（AND ／ OR ／ NOT）による検索でも同様であるが、検索結果の中に所望の文書が含まれているかを判定するためには、Ｎ件の文書を読んで内容をチェックしなければならない。しかし、文書を１件１件

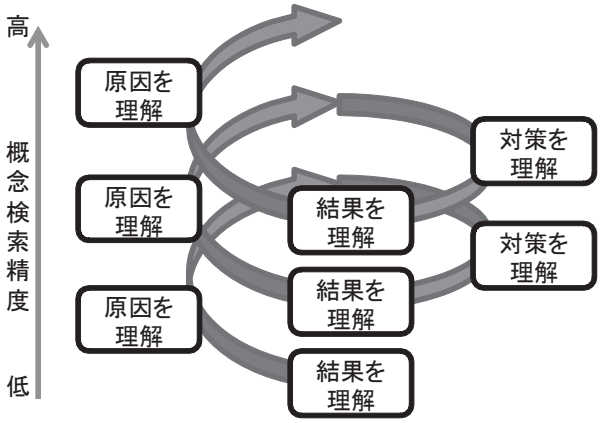


図１　スパイラルな概念検索プロセス

読解する作業は、非常に労力がかかる。そこで、N件の検索結果文書の中に所望の文書が含まれているか（各々の文書を読解して内容をチェックするに値する良好な検索結果集合であるか）を、各々の文書を読解する前に判定できると、文書を無駄に読む必要がなくなり、検索業務における概念検索の活用が促進すると考える。

（2）原因理解：なぜうまくいかないかを理解できる

上記フェーズ（1）で、概念検索結果がうまくいっていないと分かった場合に、その原因がどこにあるのかを理解するフェーズである。

一般に、概念検索がうまくいかない主な原因として、（a）入力文章の特徴語の抽出（重み付け）精度が悪い、（b）特徴語の照合精度が悪い、（c）検索対象文書集合の量的・質的特性の３つが挙げられる。

（a）入力文章の特徴語の抽出精度

入力文章の内容を端的に表す特徴語を抽出できていない、あるいは、明らかに特徴語でない語を特徴語として誤抽出してしまうなど、入力文章からの特徴語の抽出精度が低いと、概念検索はうまくいかない。また、概念検索では、抽出した特徴語に対してその重要度に相当する重みを配分するが、重みの値が実際の重要度と整合していない場合も、概念検索精度は低下する。

（b）特徴語の照合精度

入力文章から質の高い特徴語を抽出できても、その特徴語が所望の文書に数多くヒットしなければ、所望の文書を概念検索結果の上位に出力できない。特徴語がヒットしない原因としては、表記の異なる同義語・類義語の場合（例：「検索」と「サーチ」）や、構文などの表現方法の相違（例：「文書検索（名詞句表現）」と「欲しい文書を探して見つける（文表現）」）などが挙げられる。

（c）検索対象文書集合の量的・質的特性

入力文章に類似する（個々の特徴語がヒットする）文書が検索対象文書集合の中にどのくらい含まれているかによって、概念検索精度は大きく変化する。例えば、検索対象文書が公開特許公報である場合、計算機ソフトウェア分野に関する特許の出願件数は、食品分野に関する特許に比べてはるかに多い。そのため、計算機ソフトウェアに関する入力文章に類似する（個々の特徴語がヒットする）特許件数は、食品に関する入力文章のそれよりも多くなり、所望の文書でないノイズ文書が混入する確率が高くなる。

また、食品分野では、「にんじん」「調味料」など、発明内容を端的に表す技術分野固有の特徴語が、他の発明内容の特徴語に比べて多いため、類似する特許を絞り込んで識別することが比較的容易である。しかし、計算機ソフトウェア分野では、発明内容を端的に表す特徴語の多くが、他の特許の特徴語と共通していることが多く、特徴語のレベルで類似する特許を識別することが困難な場合が多い。

概念検索がうまくいっていない原因が、これら３種類の原因のうちのどれに相当するかを、利用者が簡単に特定できるための仕掛けが必要となる。

（3）対策理解：検索がうまくいく方法を理解できる

上記フェーズ（2）で、概念検索がうまくいかなかった原因を特定できた場合、その原因を解消してより良い概念検索結果を得るためには、何をどうすればよいかという対策方法を理解するフェーズである。

上述した３種類の原因のうち、入力文章の特徴語の抽出精度が悪い場合（原因（a））には、不必要な特徴語を削除し、抽出漏れした特徴語を追加し、重みが妥当でない特徴語の重みをチューニングするといった対策が必要となる。また、重要な特徴語であるが、概念検索結果上位N件にヒットした件数が少ない場合には、その特徴語の表記が特異である場合（原因（b））があるため、同義語を追加するという対策が必要となる。

ただし、特徴語を適切にチューニングしたからといって、必ずしも概念検索結果が改善されるとは限らないことが経験的に分かっている。また、特徴語のチューニングは、単語レベルの処置となるが、どの特徴語が必要／不必要であるかを、利用者が容易に判定できない場合も多い。チューニング作業に時間と負荷がかかるようだと、本稿の冒頭で述べた「検索条件の作成の負荷が少ない」という概念検索の長所がなくなってしまう。したがって、特徴語と入力文章との対応付けなど、チューニング作業の効率を向上させる仕掛けが必要となる。

次章では、特に上記フェーズ（1）（2）にかかる理解

を促進するための一手法として、「特徴語－文書マトリクス」を採り上げ、概念検索結果の理解支援の観点からの有効性について考察する。

3 特徴語－文書マトリクスによる理解支援

3.1 マトリクスの概要

特徴語－文書マトリクス（以下、マトリクス）は、縦軸に入力文章から抽出される特徴語 Ti を並べ、横軸に概念検索結果として出力される上位N件の文書 Dj を並べ、マトリクス値 Vij として、特徴語 Ti の検索結果文書 Dj における重要度（重み値）を持つマトリクスである。図2にマトリクスの一例を示す。図2では、マトリクス値 Vij が大きいほどセルを濃く塗りつぶしてビジュアル表示することで、マトリクスの視認性を高めている。

このマトリクスを操作しながら鳥瞰することで、概念検索結果に関する以下の傾向を理解できる。

（1）入力文章から抽出された特徴語の妥当性

入力文章の特徴語として、どのような語が抽出されて概念検索に使用されたのかを参照することによって、入力文章の特徴語の妥当性を判別できる。

図2は、地図情報を用いて移動体の位置と目的地までの距離を求め、蓄積されている電力量と移動に必要な電力量を比較した結果を報知するという出願特許について、明細書全文を入力文章とした場合の概念検索結果から得られるマトリクスの一例を示している。筆者が判断するに、この発明内容を特徴づける重要な特徴語として、「移動体」「位置」「目的地」「地図」「距離」「比較」「電力量」などが挙げられる。しかし、このマトリクスの縦軸を上から順に（重みの高い順に）見ると、「地図」以外の特徴語は、重みの高い特徴語として抽出できていないことが分かる。逆に、「Ｆｆ」「Ｋａ」などの変数名を表す語が、重みの高い特徴語として誤抽出されていることが分かる（実際に、所望の文書（拒絶引用特許）の検索順位は 152 位という低い順位になっている）。したがってこの場合では、入力文章の特徴語をチューニング（追加・削除・重み値修正）する必要があることが理解できる。

（2）概念検索に使われた特徴語のヒットの度合

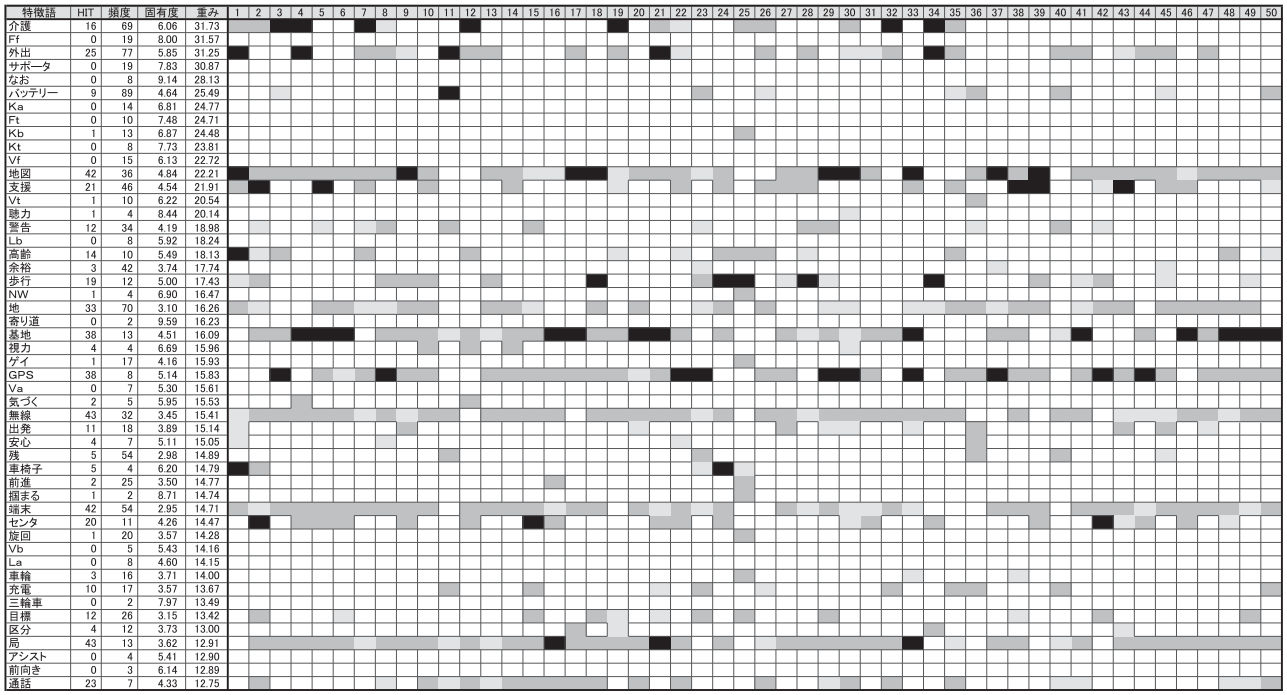
図2において、ヒット文書件数（HIT）の多い（塗りつぶされたセルが多い）特徴語である「無線」「局」「地図」「端末」「基地」などは、概念検索結果上位 50 件が出力された根拠となる特徴語（50 件を特徴づける代表的な特徴語）であることが分かる。もし、これらの特徴語を削除して再検索すると、検索結果として上位に出力される文書は大きく変わるだろう。

逆に、ヒット文書件数（HIT）の少ない（白のセルが多い）特徴語である「前向き」「アシスト」「三輪車」などは、概念検索結果上位 50 件の出力に全く影響を及ぼしていない語であることが分かる。これらの語は、概念検索に不必要なノイズ語であることが多いが、表記が特異な重要語である可能性もある。後者の場合は、その技術分野で一般的に使われている語表記に置換（追加）する必要がある。

（3）特徴語の検索結果文書での重要度

図2において、特徴語「介護」「外出」などについては、ヒット文書における重要度（重み値）が高い（セルが色濃く塗りつぶされている）文書の数が多いことが分かる。すなわち、これらの特徴語がそのヒット文書の出力順位付けに大きく影響したことを意味している。

以上で述べたように、特徴語－文書マトリクスによって、どんな特徴語が抽出されて重要視されたのか、どの特徴語が検索結果に影響を与えた／与えなかったかを鳥瞰できる。また、次節で述べるように、これらのマクロな傾向から、概念検索結果の精度（検索がうまくいったのか）をある程度推定できる。



特徴語	HIT	頻度	固有度	重み
介護	16	69	6.06	31.73
Ff	0	19	8.00	31.57
外出	25	77	5.85	31.25
サポータ	0	19	7.83	30.87
なお	0	8	9.14	28.13
バッテリー	9	89	4.64	25.49
Ka	0	14	6.81	24.77
Ft	0	10	7.48	24.71
Kb	1	13	6.87	24.48
Kt	0	8	7.73	23.81
Vf	0	15	6.13	22.72
地図	42	36	4.84	22.21
支援	21	46	4.54	21.91
Vt	1	10	6.22	20.54
聴力	1	4	8.44	20.14
警告	12	34	4.19	18.98
Lb	0	8	5.92	18.24
高齢	14	10	5.49	18.13
余裕	3	42	3.74	17.74
歩行	19	12	5.00	17.43
NW	1	4	6.90	16.47
地	33	70	3.10	16.26
寄り道	0	2	9.59	16.23
基地	38	13	4.51	16.09
視力	4	4	6.69	15.96
ゲイ	1	17	4.16	15.93
GPS	38	8	5.14	15.83
Va	0	7	5.30	15.61
気づく	2	5	5.95	15.53
無線	43	32	3.45	15.41
出発	11	18	3.89	15.14
安心	4	7	5.11	15.05
残	5	54	2.98	14.89
車椅子	5	4	6.20	14.79
前進	2	25	3.50	14.77
掴まる	1	2	8.71	14.74
端末	42	54	2.95	14.71
センタ	20	11	4.26	14.47
旋回	1	20	3.57	14.28
Vb	0	5	5.43	14.16
La	0	8	4.60	14.15
車輪	3	16	3.71	14.00
充電	10	17	3.57	13.67
三輪車	0	2	7.97	13.49
目標	12	26	3.15	13.42
区分	4	12	3.73	13.00
局	43	13	3.62	12.91
アシスト	0	4	5.41	12.90
前向き	0	3	6.14	12.89
通話	23	7	4.33	12.75

【要約】【課題】 高齢者等が安心して外出できる支援外出支援システムを提供する。
【解決手段】 移動体本体１ａと、移動体１ａを移動させる駆動手段３０と、駆動手段３０に電力を供給するバッテリー１７と、バッテリー１７の蓄積電力量を検出するバッテリー残量センサと、人工衛星からの電波を受け、移動体１の位置を検出するＧＰＳ受信機１８と、移動目的地を設定するディスプレイ２３とを備えた移動体１と、地図情報データベースを備え、移動体本体１ａの位置と、目的地と、地図情報に基づき、移動体１から目的地までの距離を求め、これらを比較し、蓄積電力量を複数に区分して利用者、護者に通知するようにしたものである。
【特許請求の範囲】
【請求項１】移動体本体と、前記移動体本体を移動させる駆動手段と、前記駆動手段に電力を供給する電力蓄積手段と、前記電力蓄積手段に蓄積されている電力量を検出する電力量検出手段と、前記移動体本体の位置を検出する位置検出手段と、利用者の移動目的地を設定する目的地設定手段と、前記電力量検出手段及び位置検出手段の検出信号を入力すると共に、前記駆動手段を制御する制御手段とを有する移動体と、地図情報を記憶する地図情報蓄積手段と、を備え、前記位置検出手段により検出された前記移動体の位置及び前記目的地設定手段により設定された目的地並びに前記地図情報記憶手段に記憶された地図情報に基づき、前記制御手段により、前記移動体から目的地までの距離を求め、該目的地までの距離と前記電力量検出手段により検出された蓄積電力量との比較に基づいて、前記蓄積電力量を複数段階に区分して通知するように構成したことを特徴とする外出支援システム。

図２　特徴語－検索文書マトリクスの一例
（特開 2002-267463（出願人 = 日立製作所）から引用。所望の文書の出力順位 =152 位）

３.２　マトリクスの傾向と検索精度の相関分析

複数の出願特許に対する概念検索結果を分析してみると、特徴語－文書マトリクスをいくつかのタイプに分類することができる。すなわち、（1）特徴語のヒット文書件数から得られる「色つきセルの密度（全セル数に対する色つきセル数の割合）」の高低と、（2）特徴語の検索結果文書での重要度に相当する「セルの濃度（全セル数に対する色の濃いセル数の割合」との組合せによって、表１に示すように【低濃型】、【低薄型】、【高濃型】、【高薄型】の４種類のタイプに分類できる。【低薄型】では、色つきセルの分布が比較的まばらで、かつ、色の濃いセルが少なく、マトリクスが全体として薄く見える。一方、【高濃型】では、色つきセルの分布が比較的密集していて、かつ、色の濃いセルが多い。

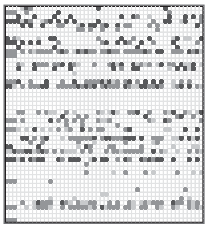
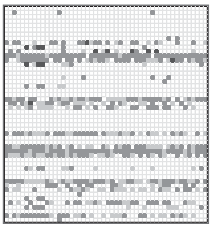
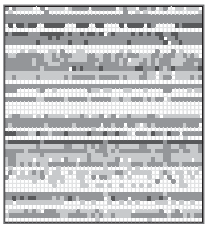
これら４種類のタイプ別の概念検索精度を比較してみた。表１に示すように、検索の漏れのなさを表す指標である再現率で見ると、概念検索結果の上位 50 件に所望の文書（拒絶引用特許）が含まれる割合（50 位再現率）が最も良いのは【低濃型】であり、上位 300 件に所望の文書が含まれる割合（300 位再現率）が最も良いのは、【低薄型】であった。色つきセルの密度が高くない（特徴語のヒット件数が多くない）ことにより、所望の文書をある程度の件数に絞り込むことができ、また、色の濃いセル（検索結果文書での特徴語の重要度が高いセル）が多いことにより、類似度にメリハリがついて、所望の文書を概念検索結果上位に押し上げて出力できていると考えられる。

逆に、再現率が最も悪いのは【高薄型】であった。色つきセルの密度が非常に高いため、正解文書を絞り込むことができず、かつ、濃いセルが少なく類似度にメリハリがつかないことにより、所望の文書を概念検索結果上位に押し上げることができていないと考えられる。

また、【低薄型】では、色つきセルの密度が高くないことにより、所望の文書を 300 件以内には絞り込めているが、濃いセルが少ないことにより、所望の文書を概念検索結果上位に押し上げることができていないと考えられる。

このように、マトリクスのタイプによって、概念検索精度の良し悪しが異なっており、検索結果の傾向を理解する有用な情報となることが分かった。

表１　特徴語－文書マトリクスのタイプ分類と概念検索精度との間の相関

#	タイプ	【低濃型】	【低薄型】	【高濃型】	【高薄型】
1	該当するマトリクスのイメージ例				
2	色つきセル密度（ヒット文書件数）	低	低	高	高
3	セルの濃度（特徴語の重要度）	濃	薄	濃	薄
4	再現率※（上位 50 件）	○ 45.7%（16/35）	△ 28.6%（ 4/14）	△ 30.8%（4/13）	△ 32.4%（11/34）
5	再現率※（上位 300 件）	○ 60.0%（21/35）	○ 71.4%（10/14）	△ 53.8%（7/13）	× 38.2%（13/34）
6	概念検索結果の傾向	ヒット特徴語の重要度が高く、所望の文書を特定できている	ヒット特徴語の重要度は低いが、所望の文書は絞り込めている	分野はある程度特定できているが、所望の文書を絞り込めていない	所望の文書を絞り込めていない

※ NTCIR-5 Formal Run 課題 619 件 [1][2] のうち、2002 年公開 96 件の全文が入力文章、1993-2002 年公開特許公報 10 年分が検索対象
拒絶理由通知書における引用特許が所望の文書（１課題につき１件存在）として、再現率を算出

4 おわりに

本稿では、概念検索結果を理解するとはどういうことかについて私見を述べた。また、特徴語－検索文書マトリクスを活用することによって、概念検索結果の傾向をマクロに理解できることを示した。さらに、マトリクスのタイプ分けおよび概念検索精度との相関分析により、色つきセルの密度は、所望の文書を絞り込めるか否かを表す指標に、また、セルの濃度は、所望の文書を概念検索結果上位に押し上げられるか否かを表す指標とみなすことができ、概念検索結果の傾向（うまくいったのか）を、文書内容をチェックしなくてもある程度推定できることが分かった。

しかし、個々の入力文章に対して、概念検索がうまくいっているのかいないのか（所望の文書が検索結果に含まれているか否か）を判別するためには、さらに違った観点からも含めた総合的な分析が必要であると考えている。

今後、特許に関連する業務において概念検索の活用を活性化させるためには、単に概念検索精度を向上するための方式を開発するだけでなく、２章で述べた概念検索のプロセスをうまく回していくことを支援する技術も確立していく必要である。これら両側面からの研究を今後も進めていく所存である。

参考文献

[1] N. Kando: Overview of the Fifth NTCIR Workshop, Proceedings of NTCIR Workshop 5 Meeting, 2005.
[2] A. Fujii, M. Iwayama and N. Kando: Overview of Patent Retrieval Task at NTCIR-5, Proceedings of NTCIR Workshop 5 Meeting, 2005.