汎用連想計算エンジンGETAを用いた特許連想検索システム

株式会社日立製作所　中央研究所
東京工業大学精密工学研究所准教授
岩山　真

PROFILE　1992年(株)日立製作所入社。以来、文書検索、文書分類、自然言語処理等の研究に従事。また、NTCIRにおいて特許検索用テストコレクションの作成に携わる。

makoto.iwayama.nw@hitachi.com

株式会社日立製作所　中央研究所
今一　修

PROFILE　1998年(株)日立製作所入社。以来、文書検索、自然言語処理等の研究、および、汎用連想計算エンジンGETAを用いた連想検索システムの開発に従事。

osamu.imaichi.xc@hitachi.com

1 連想検索とは

最近、連想検索という聞きなれない検索法を耳にした方がいるかもしれない。連想検索とは、複数の文書を丸投げしてそれらに関係する文書を芋づる式に集めたり、検索結果を特徴付ける単語群から次の検索のための新たなキーワードを見つけたりするための検索法である。概念検索に近いともいえるが、ユーザとやりとりしながら検索結果をより良いものにしていくという意味では、レレバンス（適合性）フィードバック [3] という手法に近い。同種の商用サービスもある※1。連想検索の目的と意義については、昨年のJapio Year Bookに掲載された論文 [1] に詳しいのでそちらを参照されたい。

Japioと我々は昨年度、(独)工業所有権情報・研修館の委託を受けて、連想検索を特許検索に適用した「特許連想検索（試験）システム」を開発した [2]。このシステムは、大学等における研究利用を前提としているため、ソースコードが全て公開されており、利用者は自由にシステムを改造することができる。本稿では、特許連想検索システムおよびその基盤となった汎用連想計算エンジンGETA（Generic Engine for Transposable Association）の一端を紹介する。

※1　例えばリコーテクノシステムズ(株)のRIPWAY

2 特許連想検索システム

図1は特許連想検索システムの検索画面である。本システムの検索インターフェイスはWebブラウザ上で動作する。上段が一致検索、下段が概念検索に相当する。つまり、文章がそのまま入力できる。図の例では「質問に直接答える検索システム」と入力している。

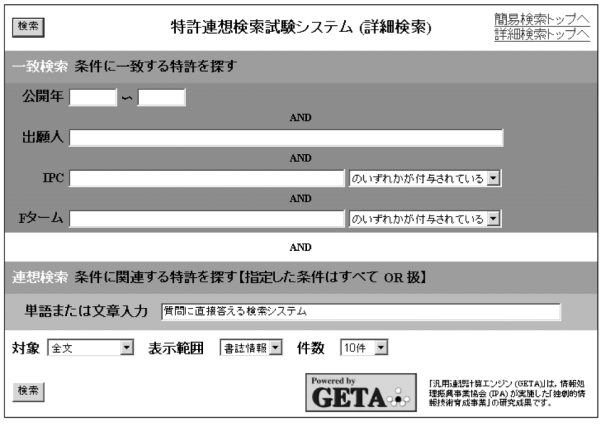


図1：検索画面

検索ボタンを押すと、1993年から2005年までの公開特許公報を対象に検索を実行する。図2が検索結果画

面である。概念検索と同じく検索結果の文書は適合度を示すスコア順に表示される。スコアの計算式は公開されていて、利用者が独自のスコア計算式を定義して組み込むこともできる。

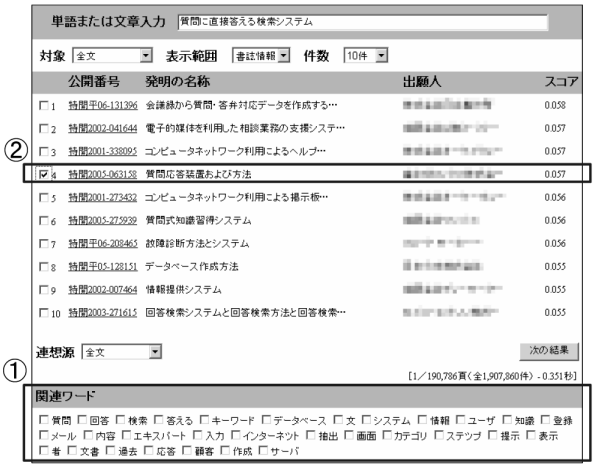


図2：検索結果画面

検索結果を見てみると、会議録から質問答弁データを作成する方法や、故障診断のために質問応答を行う方法など、検索意図からは離れた公報も多い。このような時、普通は検索式を修正して（キーワードや文章の追加や削除）再検索をする。しかし、検索式をどのように修正すればよいのかわからないことも多い。本システムでは、こういった状況で、検索結果をより良いものに改善していくための仕組みを2つ提供する。

まずは、検索キーワードの思いつき（連想）を支援するために、検索結果（上位数十件）に特徴的にあらわれる単語を「関連ワード」として表示する（図2の①）。ユーザは、ここの単語をみることで、指定し忘れていたキーワードや思いつかなかったキーワードなどを発見できる。例の場合、「答える」の他にも「回答」「応答」といったキーワードがあることがわかる。当たりをつけたキーワードをチェックすればそのまま次回の検索に反映できる。検索結果のスコアと同様に、単語の特徴度を計算する式も公開されていて、独自のスコア計算式に基づいて関連ワードをリストアップすることもできる。関連ワードは検索結果の概要を把握するのにも役立つ。

検索結果を改善するためのもう1つの方法は、検索結果にひとつでも当たりの文書があれば、それを種文書にして関連する文書を芋づる式に集める方法である。この例では、4位の公報が、検索入力である「質問に直接答える検索システム」に関連している（図2の②）。この公報を種にして、似ている別の公報を自動的に集めるためには、種とする公報をチェックして検索ボタンを押すだけでよい。種文書は何個でも指定できる。

図3が、4位の公報を種に芋づる式検索した結果である。先の結果とは異なり、上位の多くが「質問に直接答える検索システム」に関する公報になっている。この中から複数の関連文書を選びもう一度芋づる式検索を行うと検索結果は更にシャープになる。種文書を増やしながら芋づる式検索を数回行うと効果があることが経験的にわかっている。

また、芋づる式検索の結果である図3では、関連ワードに「構文」「解析」など質問応答システムに必要な要素技術があらわれてくる。ここからこれらのキーワードを使って検索結果を絞り込んでいくこともできる。検索結果が変わるとそれに応じて関連ワードも再表示される点に注意されたい。

図3：芋づる式検索の結果画面

以上、特許連想検索システムでは、関連ワードの表示と芋づる式検索により、ユーザとシステムが対話しながら検索結果をより良いものにしていくことができる。本

システムと同様の機能を持つ一般公開サービスとして、国立情報学研究所の大学図書館等蔵書検索サービスWebcat Plus[※2]があるので、連想検索に興味をもった方はまずそちらで試していただきたい。

肝心の検索精度に関しては、特にチューニングを行っていないため、特許検索用のテストコレクション（NTCIR-4 Patent）[4]で測定すると平均レベルになっている。上述したようにランキングのスコア計算式は公開されていてかつ自由に修正することができるため、本システムをベースラインにして検索精度を向上させることもできる。

作成した試験システムはXeon 5050が2個搭載されたメインメモリ16GBの計算機で動作している。OSはRedHat Linuxである。また、システムには、検索プログラムだけでなく、データ蓄積プログラムも含まれている。エンジンも含め、使用した全てのミドルウェアはオープンソース形式のものである。

次節では、特許連想検索システムのエンジンである汎用連想計算エンジンGETAについて詳しく説明する。

3 汎用連想計算エンジンGETA

3.1 特徴

GETAは情報検索や自然言語処理のための研究ツールとして開発が始まった。2002年からは、オープンソース形式で無償配布[※3]されている。GETAのコアは巨大な疎行列に効率よくアクセスするための汎用ライブラリであり、検索には特化していない。そのため、単語の共起解析や文書クラスタリング等、検索以外のプログラムも容易に実現することができる。文書検索では、行が文書、列が単語に対応する行列（図4）を作成し、列要素を指定してそれを含む行要素を集計することになる（転置インデックスの検索に相当）。

※2 http://webcatplus.nii.ac.jp/
※3 http://geta.ex.ac.jp/

GETAの基本ライブラリはC言語で実装されているがそのほとんどはperlからも呼べる。また、巨大な行列を複数サーバに分散して配置し検索することもできる。

3.2 連想のしくみ

GETAが扱う行列はWAM（Word Article Matrix）と呼ばれる。文書検索では、行が文書、列が索引語の巨大な行列になる。索引語には形態素を用いることもできるし、Nグラム文字列を用いることもできる。行列の要素は整数値であり、文書検索では文書に含まれる索引語の頻度になる。WAMの概念図を図4に示す。

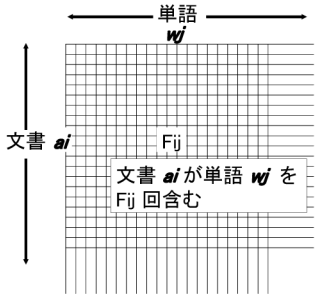


図4：WAM（Word Article Matrix）

WAM上で指定した複数の行要素（列要素）に含まれる列要素（行要素）を集計しソートするのがGETAの基本関数である。ソートの際のスコア計算式は独立したファイルで定義されているため、いろいろなスコアを差替えて使うことができる（後述）。以降、行から列のアクセスを文書－単語連想、列から行のアクセスを単語－文書連想と呼ぶ。いずれも、まったく同じ関数で実現されている。これに限らずGETAの行列アクセスはすべて行と列の双対性が保証されている。

単語－文書連想は通常のキーワード検索に相当する。文書－単語連想は、指定した文書群に含まれる単語をランキングすることに相当する。2節で説明した関連ワード抽出のことである。同じく2節で説明した芋づる式検索では、まず文書－単語連想により種文書に含まれる特徴単語を抽出し、次に単語－文書連想により関連する文書を検索する。この過程を図5に示す。複雑な検索が2つの連想の合成で見通し良く実現できる点に注意されたい。例えば、それぞれの連想過程で独立にスコア計算式

のチューニングが行える。

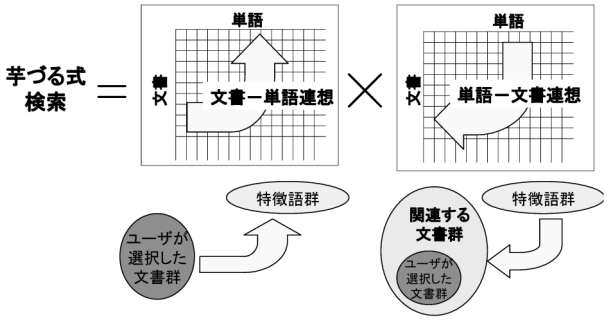
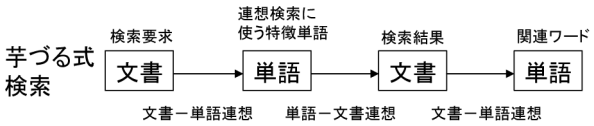


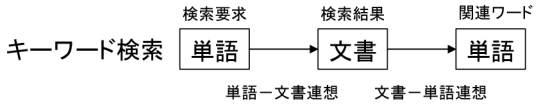
図5：芋づる式検索のしくみ

3.3 様々な連想

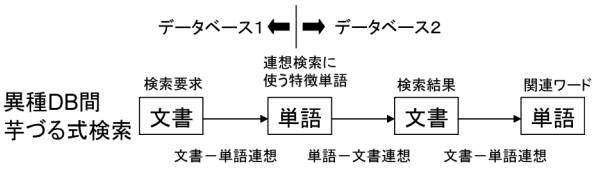
前節の2種類の連想を組み合わせることで様々な検索が実現できる。特許連想検索システムの芋づる式検索は、実際には下図のような組み合わせからなっている。



キーワード検索は、以下のとおりである。



それぞれの連想を別のデータベースで実行することもできるため、特許から論文を直接検索するといった異種DB間の芋づる式検索も全く同じ流れでプログラムの変更無しに実現できる（下図参照）。



3.4 スコア計算式のカスタマイズ

何度か述べたように、GETAではランキングに用いるスコア計算式が自由に定義できる。デフォルトで定義されている式は、SMARTという有名な検索システムで採用されている式である[5]。この他にも、BM25と呼ばれる確率的な式[6]を定義することもできる。上記2つの式は、TRECやNTCIRという文書検索の評価コンテストでも常に上位を占めている。GETA上でこれらの尺度を使い特許検索の精度を比較した研究事例もある[7]。

4 おわりに

本稿では、連想検索を特許検索に適用した特許連想検索システムを紹介した。また、そこで使われている検索エンジンである、汎用連想計算エンジンGETAについて簡単に紹介した。

これらのシステムおよびエンジンは、大学等における改良を前提にソースコードが全て公開されている。今後、本システムを土台にした新たな検索システムが提案されることを期待したい。

参考文献

[1] 高野明彦：情報をひらめきに変える連想エンジン，Japio 2006 Year Book, pp.94-97（2006）.
[2] (財)日本特許情報機構, 大学等における効率的な特許情報利用に関する調査研究報告書(平成18年度(独)工業所有権情報・研修館 委託事業), (2007)
[3] Rocchio, J. J.:Relevance feedback in information retrieval, The SMART Retrieval System, Salton, G.(Ed.), Prentice Hall, pp.313-323(1971).
[4] Fujii, A.,他.:Overview of Patent Retrieval Task at NTCIR-4, Proceedings of NTCIR-4,(2004).
[5] Singhal, A.,他：Pivoted document length normalization, SIGIR-96, pp.21-29(1996).
[6] Robertson, S.,他：Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, SIGIR-94, pp.232-241(1994).
[7] Iwayama, M.,他：Evaluating patent retrieval in the third NTCIR workshop, Information Processing & Management, 41(1), pp.207-221(2006).