2007－EIP－38 (1)
2007／12／1

4段階に重み付けした適合性フィードバックを用いた特許公報のグループリーディング支援

阿部 仁
東北大学大学院工学研究科技術社会システム専攻

開発者が発明創出プロセスの中で，どの項目に工数を費やしているのかを分析すると，公報の読解プロセスに最も時間を掛けている事が判明し，更に公報読解プロセスでは，同一テーマに関してグループメンバーで協働して，特定分野の先行特許の読解を進めていくと言うユースケースが見られることがわかった．

そこで本稿では，グループでの公報精読を支援する為に，開発現場で従来行われていた4段階の特許公報の分類を生かし，4段階に重み付けした適合性フィードバックを利用した特許公報のグループリーディング支援技術を提案する．

Group patent reading supported by 4-level relevance feedback

Hitoshi Abe
Graduate School of Engineering
Management of Science and Technology
TOHOKU University

On what the developer had spent the man-hour in the invention creation process was analyzed. And I came to know that the developer spend the most of time in the comprehension process of the patent. Then I found that there is some use case that they read the patent with a group member in the comprehension process of the patent for the same theme.

Therefore, the classification of the patent at four levels done on the development so far is made the best use of to support the careful reading of the patent of the group, and it proposes the group reading support technology using the relevence feedback that has four priority levels in this thesis.

1. はじめに

日本が知財立国を目指す中で，知財の創造・保護・活用に関しての政府の様々な政策が進められており，ますます企業における知財活動の重要性が高まってきている．

しかし一方で，研究・開発現場での特許出願に関しては，まだまだ質・量ともに十分な特許出願ができているとは限らず，改善の余地が大きく残されているのが現状であると思われる．

特に，発明創出時に必ず実施しなければならない先行特許調査に関して見てみると，図1に示すように，世の中の検索能力の向上に反比例するかの様に，特許庁における特許査定率が年々低くなってきている．この為，自己の発明に関連する先行特許を，限られた工数の中で効率良く発見し，特許庁の審査官によって容易に検索し得る様な先行発明は，発明者自身が予め発見して排除しておく事が，有効な特許出願をする為の前提として，重要になってきている．

そこで本稿では，企業の研究・開発現場における発明創出時の，特許調査から発明創出に至

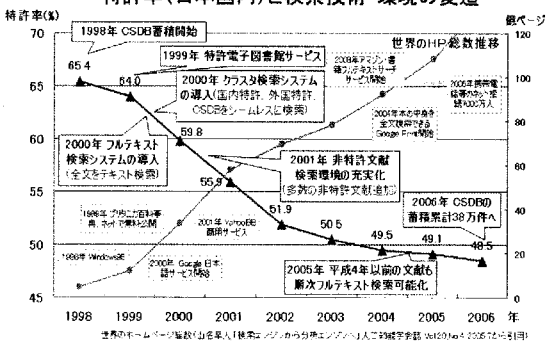

図1 特許率と検索能力との相関

るまでのプロセスを効率化するために，4段階に重み付けした適合性フィードバックによる特許公報のグループリーディング支援技術を提案する．

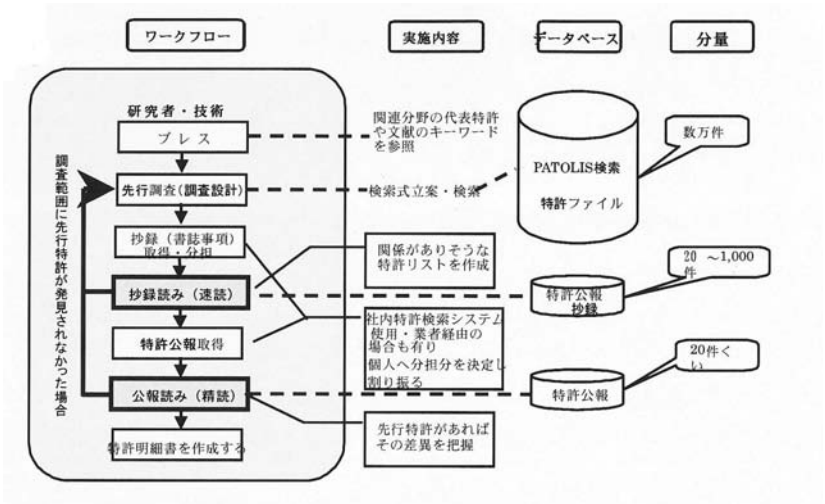

図 2 従来の特許明細書作成ワークフロー

2．特許出願までのワークフロー

組織により若干の相違はあると思われるが，現状で，ある企業における開発部門での特許明細書作成までのワークフローは，図２に示すようなものであった．

まず，技術内容の整理を行い，関連分野での代表的な特許や文献をリストアップし，検索時のキーワードとしてどの様なものが適当であるのかを抜き出す．

次に調査設計を行なう．ここでは，検索式を立案し，立案された検索式に基づき特許公報の抄録の検索を実施する．検索には PATOLIS などの商用データベースを用いて，数万件の特許公報の中から 200 件～1000 件程度の特許公報の抄録を抽出する．

取得された抄録の内容を速読によって読み進め，関連のありそうな特許を絞り込み，絞り込んだ特許公報の全文を取得する．この過程で 200 件～1000 件の公報を，20 件程度に絞り込んでいく．

更にその後，取得された特許公報全文を精読する事で先行特許を理解し，自己の発明との相違点を明確にした上で特許明細書を作成する．

3．ワークフロー上の課題と解決案

前記の従来の特許出願までのワークフローにおいて、最終的に精読した結果の中に，関連のありそうな先行発明が発見される場合には問題無いが，発見されなかった場合には，再度調査設計からやり直す必要がある．

この調査設計のやり直しに関する試行錯誤のプロセス部分が，研究・開発者への負担として重くのしかかっている．

現状のワークフローでは，スキルが低く調査設計の検索式が不適切でも，再び調査設計し直すまで、大量に公報を読まねばならない．

そこで、この部分をＩＴにより支援し、できるだけ早く再調査に移る仕組みが望まれる．

またもう一点，このワークフロー自体が，個人作業ベースでのワークフローになっており，同一の研究テーマに対して，先行特許調査を行なおうとする共同発明者との共同作業を支援する仕組みが無い．その為，協働して調査を進める場合には，公報を取り寄せたのちに、人数分に分担した公報の束を各自読み進める様な形式を取っており，この為，例えば誰かが先に，非常に近い類似先行技術を発見していたとしても，それを知らない他のメンバーは自己の分担分が完了するまで公報を読み続けていた．

これらの課題に対する解決策として，現状の様なバッチ的な処理では無く，公報を読み進めるたびに結果をシステムにフィードバックして検索式を洗練していく様な技術が有効ではないかと考える．

4．従来技術としての適合性フィードバック技術

上記の様な考え方に基づいた検索技術としては，Rocchio[2]による適合性フィードバック技術［1］が既に存在しており，この適合性フィード

バック技術が，上記の課題を解決するのに最も適した方式であるかどうかを検討した．

ここで，Rocchio による古典的な適合性フィードバック技術をまず記載する．

Q を前回の検索時に用いられた検索ベクトル，Q' を新しく生成される検索ベクトル，R_i を適合特許公報に対する特徴ベクトル，S_i を不適合特許公報に対する特徴ベクトル，n_1 が適合する特許公報数，n_2 が不適合な特許公報数，β および γ をヒューリスティックスにより与えられる重みとすると、Rocchio の式は（1）式の様に表される．

$$Q' = Q + \beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2} \quad \cdots（1）式$$

5．4段階に重み付けした適合性フィードバック

上記の Rocchio の式では，適合・不適合のみの 2 段階で適合性の判断を行なっているが，ここで調査した企業における先行特許調査の様な場合には，通常，特許公報を 4 段階に分類して関連度を比較している．すなわち，「重要公報」，「注目公報」，「参考公報」，「無関係公報」などの 4 段階に振り分けて利用しているため，この分類をそのまま活かして，更に効率的な適合性フィードバックを行なう様な手法を検討した．

元の検索ベクトル Q から，利用者の反応により修正された検索ベクトル Q'を求める際に，$R_1,\ldots,R_{Nimp}$ の重要公報，$S_1,\ldots,S_{Nnotice}$ の注目公報，$T_1,\ldots,T_{Nref}$の参考公報，$U_1,\ldots,U_{Nirrevent}$ の無関係公報とにユーザが振り分けた判断結果を利用して検索ベクトルを修正するものとすると，前記 Rocchio の式を（2）式の様に表現する．

$$Q' = Q + \frac{\alpha}{Nimp} \sum_{j=1}^{Nimp} R_j + \frac{\beta}{Nnotice} \sum_{j=1}^{Nnotice} S_j + \frac{\gamma}{Nref} \sum_{j=1}^{Nref} T_j - \frac{\delta}{Nirrevent} \sum_{j=1}^{Nirrevent} U_j$$

$\cdots（2）式$

ここで Nimp は重要公報数，Nnotice は注目公報数，Nref は参考公報数，Nirrevent は無関係公報数，α，β，γ，δ は，それぞれに対するヒューリステックな重み付け係数である．

（1）式と（2）式の相違が最も顕著な場合の例を図 3 に示す．

ここで，文書群中の検索語で構成されるｎ次元の文書特徴ベクトル空間の中で，求める文書特徴ベクトルに対して，適合，不適合の 2 値で分類を行なった場合には、図 3 で示した☆の値を持つようなフィードバック文書に対し，不適合と言う評価がなされる為，結果的に破線で示した前回の検索ベクトルに対して，☆から遠ざかる方向にベクトルの修正が行なわれる．

しかし，この修正は，実際には求める文書特徴ベクトルからも離れる方向に働く為，適切な修正とは言えない．

それに対して，参考文書としてフィードバックする事で，求める文書特徴ベクトルに引寄せる方向に修正される今回の提案は，この様な例においては，Rocchio の式のまま実現するよりも効率的な検索結果を得られる可能性が高く，研究・開発現場でのこれまでの分類結果がそのまま，最大限に活用される事が期待されうる．

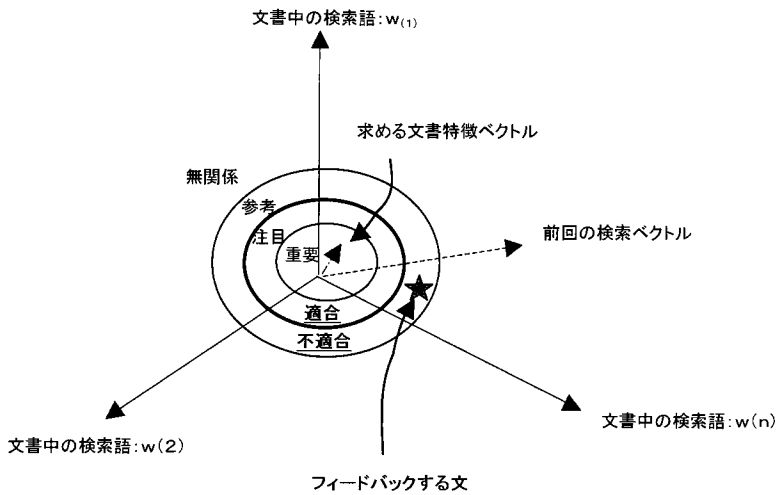

図 3　4段階分類の効果

6. グループリーディングの実現

次に，もう一つ課題である個人ベースでのワークフローを改めて，先行特許公報の評価結果を，逐次，共同発明者間で共有する為に，前記の適合性フィードバックをグループ内で共有する為のフレームワークを検討した．

ここではサーバ・クライアントモデルによってシステムを実装する事で，検索エンジンに対して Web を経由してアクセスする事により，同時に複数人からの検索エンジンに対するアクセスと，結果の共有を実現する．

図４に示す様に，各グループメンバ（共同発明者）は，ポータルとしての検索インタフェースのトップページを経由して，検索 CGI を起動し検索を行なう．検索結果はサーバ内部で中間ファイルとして蓄積され，次に同一の利用者によって適合性フィードバックによる検索が行われる場合や，他のグループメンバによってトップページにアクセスされる場合の最新の検索結果の表示の為に利用される．

読んだ結果を，③で「重要」，「注目」，「参考」，「無関係」の 4 段階に分類評価し，結果を検索サーバに入力する．

検索サーバは分類された結果に基づいて、4 段階に重み付けした適合性フィードバックによって分類した特許公報から算出された検索ベクトルにより，類似度の高い順に特許公報をグループメンバー全員に対して再表示する（④）．

同様に、検索サーバからの特許公報の表示，読解，そして分類を，他のメンバーも実施し，「重要」，「注目」，「参考」，「無関係」のどの分類評価結果になるかを⑤で検索サーバに返す．

このようなグループメンバによる評価の結果，最初のユーザの場合と同様に 4 段階に重み付けられた適合性フィードバックがそれぞれ実施され，結果を④でグループメンバー全員に再度伝える事となり，このプロセスが繰り返される．

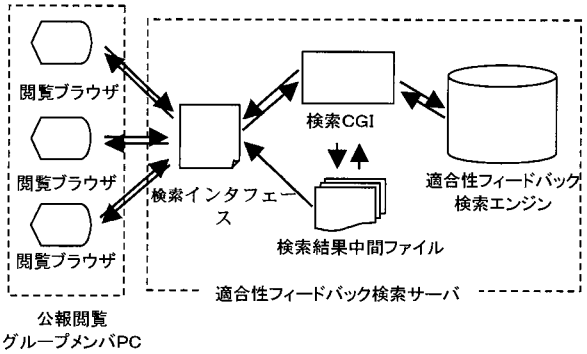

図４　検索サイト

次に、この検索システムを使用した時のシステムの動作を，図５にユースケース図として示す．

まず①で，最初に簡単な検索式を用いて検索サーバから，検索した結果の特許公報リストを表示する．

利用者は，②の公報読解端末で、検索サーバから表示された特許公報リストを眺め，任意の公報を選択し，内容を読んでいく．

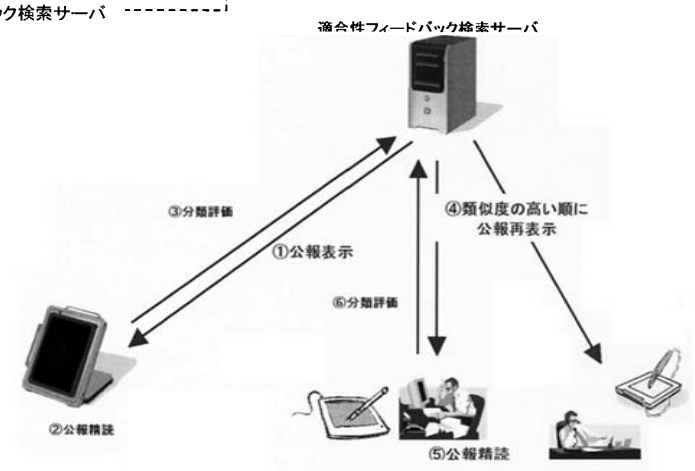

図５　適合性フィードバック検索システム

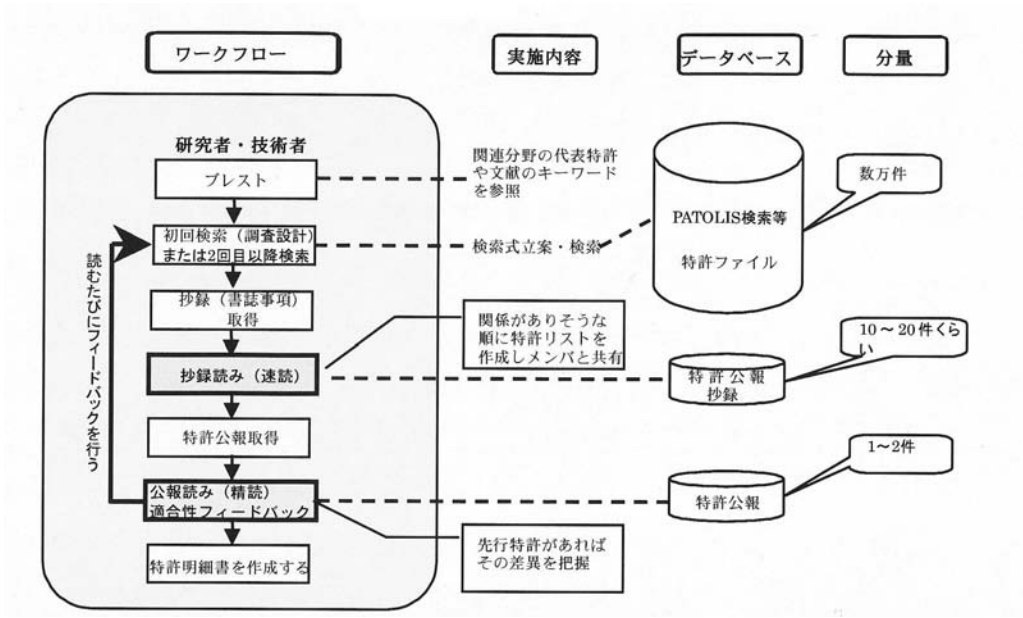

図 6　提案システムによる特許出願ワークフロー

7. 提案システムによるワークフローの改善

図 6 に，本提案により変化する出願ワークフローを記載する．

最初に，技術内容の整理を行い，関連分野での代表的な特許や文献をリストアップし，検索時のキーワードとしてどの様なものが適当であるのかを抜き出すところまでは，従来のワークフローとほぼ同じである．

次に調査設計を行なうが，ここでは，初回の検索と 2 回目以降で若干アプローチが異なる．

初回に関しては従来とほぼ同じだが，抄録ベースで読む公報は多くて 20 件程度までで構わない．数万件の特許公報の中から 20 件程度の特許公報の抄録を抽出したら，抄録の内容を速読によって読み進め，関連のありそうな特許を絞り込み，絞り込んだ特許公報の全文を取得する．ただし，ここでも 1～2 件読んで次のステップに移る．次のステップではここで全文を取得して精読した特許公報に関して，重要度の評価を行い，評価結果とともに、システムにフィードバックを行う．

フィードバックされた特許公報を用いて 4 段階に重み付けした適合性フィードバック検索を行い，再び数十件の特許公報に絞り込んで抄録を取り寄せる．以下，同様のプロセスが続く．

結果的に，精読した特許公報の中に，自分の発明に非常に近い先行発明が見つかれば，それを基にして，自己の発明との相違点を明確にした上で特許明細書を作成する．

以上の過程において，グループで協働してひとつのテーマについての特許調査を行なっているため，ワークフロー上には“抄録読み”のフェーズでの実施内容に記載したように，作成された抄録特許リストはメンバー間で共有しており，途中から参加する事も，途中までの他のメンバーの調査結果を確認する事も容易である．

8. 評価実験

本提案の有効性を検証するために，4 段階に重み付けした適合性フィードバック検索サーバのプロトタイプを試作した．

検索エンジン部には，GETA(連想検索エンジン)第 2 版を利用し，その上に適合性フィードバック部を実装する事でシステムを実現した．

使用した環境は，以下の様なものである．

- ハードウエア：　IBM ThinkPad R50
- OS: VineLinux2.1
- perl: perl, version 5.005_03
- jperl: jperl5.005_03-990822
- 茶筌: ChaSen version 2.0b6
- nkf: Network Kanji Filter Version 1.62
- libae:　GETA(連想検索エンジン)第 2 版

いずれも、GETA［3］の動作確認済みの環境を使用した.

8.1. 評価用データ生成

○評価用データの内容

4 段階の分類では，まず求めるべきターゲット公報を一つ設定し，それとの類似性により，重要公報，注目公報、参考公報，無関係公報を、各１０文書ずつ計４０文書を評価用に収集し，上記の順番に文書番号１～１０、１１～２０、２１～３０、３１～４０とした.

適合性フィードバックが有効に機能する目安としては、文書中の索引語が 25 以上であるべき事が経験的に知られており[4]、この文書数であれば、その条件を満たすものと考える。

2 段階の重み付けでは、4 段階の重み付けで使用した文書のうち、重要公報と 注 目 公 報とを適合文書、参考文書と無関係文書を不適合文書として扱った.

○文書特徴ベクトルの生成

形態素解析された各単語とその出現頻度を，文書毎にあらかじめ生成しておき，フィードバック時に検索ベクトルと比較できる様に準備した．これは本来はスクリプト言語などを使って自動化するべき所であるが，今回は簡易的に KH Coder を利用し，その出力結果を若干手作業で編集することで生成している.

8.2. 評価

この実験で試作したシステムを用いて，フィードバックが起きやすい初期検索文章で，適合性フィードバックを行った場合の適合率と再現率とを比較した。

Rocchio の式では「重要特許」および「注目特許」を“適合”，「参考文書」及び「無関係文書」を“不適合”として計算している.

初期検索文章を入力し，類似度の高い順に１０文書を表示した.

【結果１】

実験で用いた文章で，フィードバックし易い文章と，フィードバックしにくい文章で、それぞれ初期検索した場合を，ケース１，ケース２の例で示す．ここで数字列は、前記の用意されたテストデータの文書番号を，類似度の高い順に示したものである.

◆ケース１：入力文章（「***適、不適の評価情報を用いて，類似文書検索を応用して検索する技術***」と言う文章）

1-13-9-5-7-8-6-20-11-19

⇒上位 10 件中，重要特許が 6 件，注目特許４件,

２段階：再現率：50%，適合率：100%

４段階：再現率：33%，適合率：100%

（適合率が高いのでフィードバックしなくても調査が進められる）

◆ケース２： 入力文章（「***当初の品質を公表する***」と言う文章）

40-39-38-23

⇒ ４件のみがヒットし，参考文書が１件 ，無関係文書３件

２段階：再現率：0%，適合率：0%

４段階：再現率：3%，適合率：25%

（適合率が低いのでフィードバックすべきだが，2 段階ではフィードバックデータが得られなかった為，不可能）

【結果 2】

フィードバックし易い初期文章を入力し，類似度の高い順に１０文書までの検索結果を示すと

○初期検索文章入力（「***当初の品質を公表する***」と言う文章）

40-39-38-23 …参考に分類できる文書が現れる

２段階：再現率：0%、適合率：0%

４段階：再現率：3%、適合率：25%

○ 23 を選択し「参考」と言う評価を行う.

23-25-28-26-19-31-20-18-22-7 …重要に分類できる文書が現れる

４段階：再現率：30%、適合率：90%

○ 7 を選択し、「重要」と言う評価を行う

7-23-26-19-1-3-31-8-13-25 …4 件の「重要」文書を抽出

４段階：再現率：30%、適合率：90%

8.3. ユーザへのヒアリング

また，本プロトタイプを企業の開発メンバーに紹介し，

① 本システムの提供する機能を利用したいかどうか

② 本システムが発明創出時に有効であるかどうか

の２点についてのヒアリングを行った。その

結果①については，提案するシステムの機能を利用したい人は14名中13名であった．
あまり思わないと答えた 1 名については，「システムの挙動が難しい」と言う意見であり，検索結果の再表示の際に行われる評価関数の妥当性が、使用していて直感的にわかりづらいと言う点を欠点として挙げていた．

これに関しては、評価用に用意したデータ数の少なさにも原因があったと考える．

9. おわりに

これまでの研究は、発明から特許出願までの業務支援の為のツールの開発を目的として研究を進め，4 段階に重み付けした適合性フィードバックによるグループ読解支援技術を提案し，提案技術を実装したシステムをプロトタイピングする事で，本提案が発明創出時の業務支援ツールとして有効である見込みが得られた．

しかし，まだまだ評価データそのものが十分であるとは言えず，国立情報学研究所のテストコレクション（特に特許検索タスク）などを使った，より客観的な評価実験や，多人数・大規模な試験を行う事による負荷テストなども必要と考える．

また，企業の現場でこのツールが効率的に機能するには，信頼性や品質面を含めた，さらなる作り込みも必要になるだろう．

今後は，以下のような点を改良していく事で，開発現場での使用に耐えうるシステムにしていきたい．

・Ｄｒａｇ&Ｄｒｏｐや全文へのリンク表示などの洗練されたＷｅｂＵＩの作り込み
・適合性フィードバック検索式の係数の、更なる最適化
・セッション管理によるログインユーザー管理機能を設けることでの、人間系へのフィードバックの実現
・ＧＥＴＡに付随する inbox2freqfile の応用による、特許公報文献からの特徴ベクトル生成の自動化

以上．

参考文献

[1] **徳永健伸**: “**言語と計算 5 情報検索と言語処理**”, **辻井潤一編**, **東京大学出版会**, (1999)

[2] J. J. Rocchio: “Relevance feedback in information retrieval.”, *The SMART Retrieval System: Experiments in Automatic Document Processing*, PrenticeHall Inc., pp.313–323. (1971)

[3] **西岡真吾**, **今一修**: “**汎用連想計算エンジン GETA とそれに基づく連想検索システム**”, **情報処理学会 自然言語処理研究会 研究報告**, Jun. (2000)

[4] W.B. Frakes & R.Baeza-Yates : “Information Retrieval: Data Structures & Algorithms. ”, PrenticeHall Inc., (1992)