社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－DD－48 (5)
2005／1／28

教員連想検索システム

関 隆宏* 廣川 佐千男†

要旨. 近年の大学情報公開の流れのなかで，教員情報を提供するために教員検索機能を持ったホームページを開設する大学が増加している．しかし利用者が要求する情報を効率よく検索できるとは限らないのが実情である．本稿は検索要求だけでなく，検索対象における語の重要性を考慮に入れた教員検索を行い，さらに検索結果を利用して再検索できる教員連想検索システムについて述べる．

Teacher Associative Search Engine

Takahiro Seki* Sachio Hirokawa†

Abstract. Information of university tend to be opened recently, and many universities have opened homepages with a teacher retrieving function in order to offer teachers' information. However, the demanded information cannot be necessary searched efficiently in actuality. This paper describes the teacher associative search engine. The engine can search teachers based on not only user's demand but also importance of words in the referred target, and moreover re-search by using the results.

1 はじめに

「大学の情報公開」や「社会に開かれた大学」といった言葉が聞かれて久しい．教員の研究内容や研究業績，教育内容といった教員情報を印刷物の形ばかりでなく，インターネット上に公開する大学が多くなっている．そしてこの教員情報をインターネット上に公開する場合に，検索機能を持たせる大学もある．この検索機能は例えば，ある教員に興味を持つ人がその教員について詳しく知りたい場合や，企業などが大学の教員と共同研究を行うためにはどの教員が適当なのか知りたい場合に利用されている．一般にこの検索機能では一致検索が採用されているため，前者の場合は特に問題はないが，後者の場合に検索結果が利用者にとって都合のよい形で出力されるとはいいがたい面がある．本稿はこの問題を解決する検索機能を持つ教員連想検索システムの構築について述べる．

大学のホームページの多くには「教員検索機能」を有するページがある．筆者らが所属する九州大学では「九州大学研究者情報」([3]) にその機能がある．これを例にすると，ユーザがあるテーマの研究をしている教員を探そうとするならば，ユーザがキーワードを入力し，教員検索機能により入力したキーワードを含む教員を教員のデータベースから検索し，該当する教員の一覧が出力される．ほとんどの大学で用いている教員検索機能も同様である．そこで，このような教員検索機能を「既存の教員検索機能」と呼ぶことにしよう．

社会連携あるいは産学官連携を例に考えると，企業との共同研究や受託研究，あるいは社員に高度な技術を学ばせるための社会人入学などが近年盛んに行われている．これらは教員との直接のつながりのある企業でしばしば行われてきたが，教員とのつながりのない企業でもこうした要求が高まりつつある．企業がこれらを行う際，適当な教員を探すために，大学の教員検索機能を用いて検索することもあるだろうし，大学の産学官連携窓口に問い合わせて教員を紹介してもらうこともあるだろう．特に後者の場合，問い合わせのあった窓

*九州大学大学評価情報室・Office for Information of University Evaluation, Kyushu University
†九州大学情報基盤センター・Computing and Communications Center, Kyushu University

口では，専属の教職員がまず大学の教員検索機能を利用して紹介すべき教員を探すところから始まる．いずれにしても，「既存の教員検索機能」を用いる点は同じである．そして，その後の絞り込み作業はほぼ同様である．この絞り込み作業において，作業能率の差こそあれ，時間がかかりすぎることが問題になっている．この問題の原因として考えられるのは，「既存の教員検索機能」では検索内容に合致する結果をすべて出力することに主眼がおかれていることである．確かに検索結果は参考にはなるが，ユーザの細かい要求とずれがあるのが普通であり，要求に近い教員を見つけるために検索結果のすべてをチェックしたり，似たような検索条件に変えて検索したりせざるを得ず，そのために時間がかかってしまうことになる．そうすると，この問題を軽減するようなシステムがユーザから求められているといえる．

さて，時間がかかってしまう原因のうちのいくつかは，そのまま「既存の教員検索機能」の問題点ということになる．この問題点としては次のようなものが挙げられる．

1. 検索結果がどの程度合致しているのか分からない．
2. ユーザが検索内容を上手に設定しないと検索がうまくいかない．
3. 再検索するには最初からやり直さなくてはならない．
4. 検索結果のそれぞれをチェックして目的とする結果を得なければならない．

最後のものは，検索をコンピュータに任せている以上避けられない問題であるが，各検索結果に要約をつけることによってユーザの負担を軽減することも考えられる．一方，1 に対しては出現頻度順に検索結果が出力される事例もあるし，3 に対しては絞り込み検索を可能にしている事例もある．

本稿では先述の問題に対する一つの解決として，国立情報学研究所で開発された汎用連想検索エンジン（GETA）を利用して構築した教員連想検索システムについて述べる．教員連想検索システムの強みは，検索内容に関連する結果を関連の強い順に出力される点と，検索結果を基に再検索できる点にある．さらに，この教員連想検索システムでは検索結果の要約を出力する機能も加えてある．

2 連想検索とGETA

一般に検索といえば，キーワードの有無により検索を行う一致検索を指す．しかし，検索要求をうまく設定しなかったために，求めている情報が検索されなかったり，求めていない情報が検索されるといった問題点が指摘される．連想検索では，キーワード (群) あるいは文書 (群) の検索対象文書内における語の重みを考慮して検索することにより，入力キーワード (群) あるいは入力文書 (群) と類似の文書を検索することを可能にしている．したがって，連想検索は一致検索に比べ，おおまかな検索要求でもユーザが求めている情報が得やすいと考えられる．

連想検索 —特に文書における連想検索— は基本的に

- 文書（群）からその文書（群）の特徴語
- 単語（群）から関連文書（群）

の 2 つの検索から成る．前者であれば，文書（群）のデータから「連想」される語，すなわち特徴語を検索し，後者であれば，単語（群）から「連想」される文書（群），すなわち関連文書（群）を検索する．したがって，これらを組み合わせることにより文書（群）から関連文書を検索したり，単語（群）から関連語を検索することができる．上記の「文書」を教員データに見立てたものが教員連想検索である．

教員検索の場面を考えると，探している教員が明確な場合，氏名や所属，研究業績などのキーワードを入力することにより検索する一致検索は有効なものであるが，これはまさに既存の教員検索機能が実現しているものである．しかし，そうでない場合，すなわち探している教員がある程度規定できる場合には，検索キーワードから関連性の高い単語を抽出し，それを含む教員をもれなく探し出す検索方法である連想検索の方がより有効であると考えられる．

汎用連想計算エンジン（Generic Engine for Transposable Association；以下，GETA と呼ぶ）は，文書検索における頻度付き索引データを行列化したもの（Word-Article-Matrix；WAM と呼ばれる）を対象に，行と行あるいは列と列（具体的には文書間および単語間）の類似度を高速計算するツールで，国立情報学研究所で開発された．GETA

には大規模文書における連想検索，文書分類，単語間類似度計算などの大規模文書分析に必要な要素技術がサポートされている．なお，GETA についての詳細は [6] を参照のこと．これを利用した検索システムの例として，図書の連想検索を行う国立情報学研究所の Webcat Plus ([2]) がある．

3 教員連想検索システムの特徴

先に述べた既存の教員検索機能の問題点を解決し，GETA の連想検索機能を生かした教員連想検索システムの実装を考える．本稿で述べる教員連想検索システム (以下「本システム」と呼ぶ) は Perl で記述した CGI プログラムにより実装されている．本システムの特徴は以下の 3 点にまとめることができる．

1. 関連度順に検索結果を出力

   検索結果を関連度順に出力することにより，ユーザが必要としている情報が得やすくなる．多くの場合，代表的教員すなわち関連度の高い教員を知りたくて検索するので，必要とされる情報は検索結果の上位にくる教員である．逆に，検索対象をよく知っている場合にはわざと多く検索結果を出力させることにより，意外性のある結果を得ることも可能になる．

2. 関連結果も出力

   本システムでは「合致する」結果ばかりではなく，「関連する」結果も得られる．したがって，検索条件とはほとんど合致しないように思われても，検索条件との関連が強いとされれば検索結果として出力される．また，検索結果として教員の一覧が得られるのが普通であるが，本システムではこれらの教員から連想される語 (以下，「関連語」と呼ぶ) を関連結果として出力する．さらに，検索結果として得られる教員のそれぞれに対してその教員から連想される語を「研究の重要語」として，そして再検索時に教員をキーとして検索した場合にキーにした教員の研究の重要語が現れたらそれらを「指定した教員との主要共通語」として表示できるように選択できる．関連結果を出力することにより，見落としていた情報や新たな情報の発見が可能になる．例えば，検索要求が実際の検索で有効でなかったにしても，関連結果を見ることにより適当な検索要求を発見することも可能である．さらに，関連結果は検索結果の要約と考えることができるので，関連結果を見ることにより，検索結果がユーザの意図する結果であったかどうか分かる．

3. 出力結果を基に再検索可能

   出力結果のうち，教員名と関連語にはチェックボックスがついており，そのチェックボックスにチェックした教員名あるいは関連語を基に再検索を行う．この再検索では，チェックされた教員から連想される教員とチェックされた関連語から連想される教員との両者に出てくる教員が検索結果として得られる．出力結果を基に再検索できることにより，ゆるい検索条件だったものをきつい検索条件に変えて再検索したり（絞り込み検索），きつい検索条件を関連結果を利用してゆるい検索条件にして再検索したりできる．つまり，対話的な利用によってユーザにとって満足のいく結果が得やすくなると考えられる．

構築した教員連想検索システムは九州大学の教員連想検索を行うものである．ここで用いるデータは 2004 年 7 月 1 日現在の「九州大学教員の研究教育活動等報告書データベース」(現「九州大学研究者情報」; 以下この記述は省略する) 中の「研究・教育・社会連携活動概要」と「研究業績」である．なお，このデータは全部で 2052 人分ある．前者は，各教員が自分の研究内容や教育内容，社会連携活動について述べたもので，基本的に日本語で書かれている．後者は，各教員が「九州大学教員の研究教育活動等報告書データベース」で公開を認めた学会発表や原著論文，著書のタイトルを用いている．外国語で書かれた論文等は外国語のまま扱う．これらのデータの抽出が可能になっているのは，「九州大学教員の研究教育活動等報告書データベース」のデータ源になっている「九州大学大学評価情報システム」に保存されているデータ形式として XML を採用しているためである ([4])．これらのデータを形態素解析器「茶筌」にかけ，名詞と品詞が分からなかった語（外国語や専門用語がほとんどである）だけを残し，それらの出現頻度を求め，WAM を作成する．これを基に Singhal

らによる方法 ([1] ; GETA では WT_SMART と指定する) により関連度を計算する．ここで Singhal らによる方法を採用したのは [5] に類似性尺度の比較で優れていると記述されていることによる．

関連度はつぎのように計算される．検索要求を q，検索対象文書を d，すべての検索対象文書に含まれている単語の集合を T とする．さらに，$*$ 中に現れる $t \in T$ の頻度を $\mathrm{Tf}(t,*)$，d 中の異なり単語数を $w(d)$，$*$ の平均を $\overline{*}$ で表すことにする．さて，q と d の関連度 $\mathrm{sim}(q,d)$ は

$$\mathrm{sim}(q,d) = \frac{1}{N}\Sigma_{t\in T}\{\mathrm{tf}(t,q)\cdot\mathrm{idf}(t)\cdot\mathrm{tf}(t,d)\}$$

により求められる．ここで，$idf(t)$ は $t \in T$ の idf 値，

$$\mathrm{tf}(t,*) = \frac{1+\log(\mathrm{Tf}(t,*))}{1+\log(\overline{\mathrm{Tf}(t,*)})}$$

$$N = \overline{w(d)} + 0.2(w(d) - \overline{w(d)})$$

である．したがって，$\mathrm{sim}(q,d)$ の値が大きいものが関連度が高いとみなされる．

この教員連想検索システムは，図 1 に記した流れで利用される．

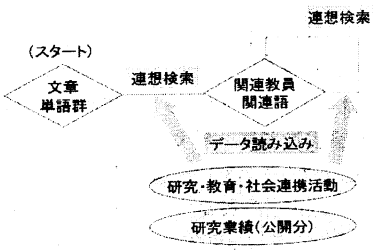

図 1: システム利用の流れ

より具体的には，以下の通りである．

1. 検索要求（文書または単語群，検索対象文書，検索人数，重要語出力の有無）の決定（図 2）

2. 検索要求に基づいて教員連想検索（検索要求と関連の強い教員の検索）および関連語検索（教員連想検索により得られた教員の重要語の検索）（図 3）

3. 検索結果から次の検索要求を決定（図 4）

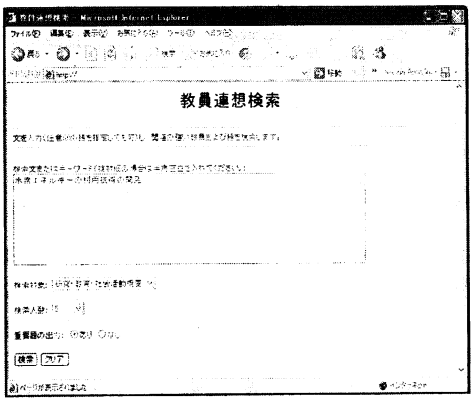

図 2: 初期画面

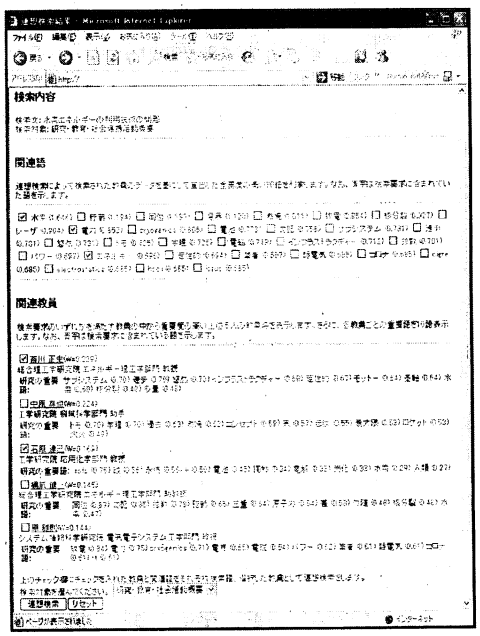

図 3: 検索結果画面

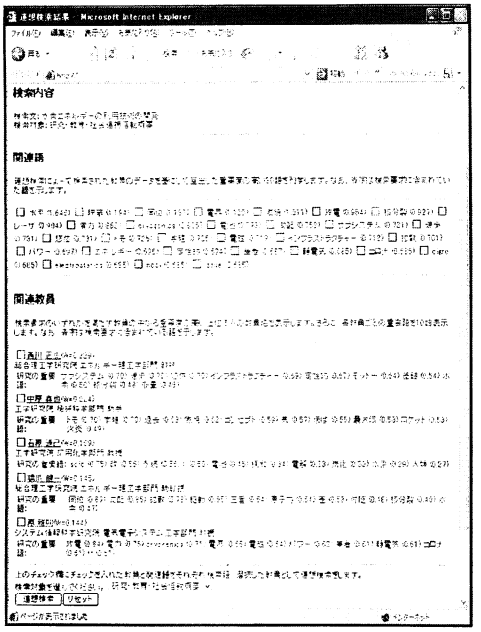

図 4: 再検索条件決定画面

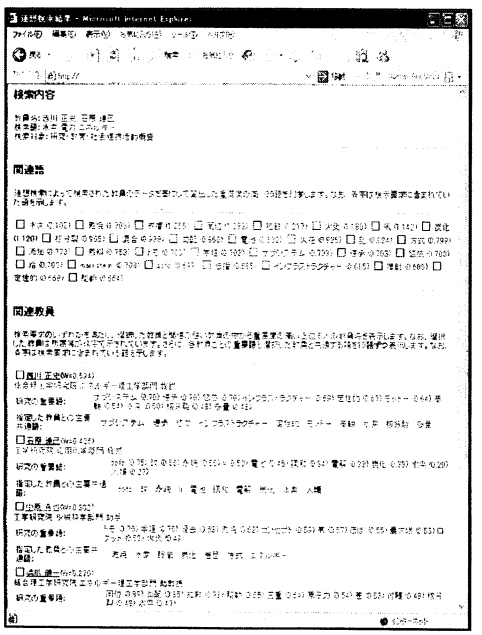

図 5: 再検索結果画面

4. 次の検索要求に基づく教員連想検索および関連語検索）（図 5）

5. 3 に戻る

4 評価実験

システムの評価をする場合に，テストコレクションを利用するのが一般的である．しかし，本システムを評価する場合，システムの性質上テストコレクションは存在しない．ここではテストコレクションの代わりに九州大学教員による特許・発明のデータを用いて評価実験を行う．ここで用いたデータは 2003 年 9 月のものである．検索要求としては，特許・発明データ中で「要約」の部分 —具体的には発明が解決しようとしている課題，解決手段，発明の効果などが書かれている— を用いる．以下，単に「発明データ」といったら，この箇所を指していることにする．「発明データ」を用いているのは特許・発明データ中の他の箇所に比べ，専門用語・技術的用語が比較的少ないことによる．各特許・発明には発明者がいるので，それを正解とみなし，教員連想検索の結果正解者が第何位に出現するかを調べる．ただし，発明者が複数いる場合は順位が高い方を正解の順位とする．正解が上位に現れていればこのシステムは良いという評価をする．

正解がある特許・発明の件数は 1094 件あるので，それらの「発明データ」を検索要求とし，「研究・教育・社会連携活動概要」を検索対象とする教員連想検索を行った．なお，先述のように教員データベースには 2052 人分のデータがあるので，正解の順位はたかだか 2052 位である．表 1 は 100 位ごとに区切った正解の順位の件数ならびに全体に対する比率（単位は％；以下「全体比」と呼び，単位は％であるとする．なお小数点第 2 位を四捨五入したので，合計が 100 にならない）を示している．検索できなかったものが 16.5 ％あるが，主な原因としては次のようなものが考えられる．

1. 「発明データ」と「研究・教育・社会連携活動概要」の用語の質的違い

   前者は特許・発明データの中でも比較的専門用語・技術的用語が少ないとはいうものの，やはり専門用語や具体的手続きによる記述が後

者に比べると断然多い．これは，後者が専門家以外でも分かるような平易な記述を主体としていることに関連している．

2. 「発明データ」における古いデータの存在

「発明データ」には10年以上前のデータも多く含まれている．そのような特許・発明の発明者が現在も同じまたは類似のテーマで研究しているとは限らない．

3. 「研究・教育・社会連携活動概要」の記述内容の質

「研究・教育・社会連携活動概要」の記述は各教員に委ねられている．したがって，丁寧に説明する教員もいれば，研究テーマを箇条書きする教員もいる．これにより，あまり記述の多くない教員が発明者だった場合に，正解として出てこなくなる可能性がある．

順位	件数	全体比
1〜100	748	68.4
101〜200	56	5.1
201〜300	28	2.6
301〜400	17	1.6
401〜500	13	1.2
501〜600	17	1.6
601〜700	12	1.1
701〜800	8	0.7
801〜900	5	0.5
901〜1000	2	0.2
1001〜	7	0.6
検索できず	181	16.5
合計	1094	(100)

表 1: 教員連想検索結果（全体）

表1で目を引くのは100位までに3分の2以上が出現することである．そこで，正解が100位以内にあったものをさらに10位ごとに区切った正解の順位の件数ならびに全体比を表2に示した．

表2からわかるのは10位までに全体の半数弱が出現することである．そこで，1位から10位までのそれぞれに正解が何件出現し，その全体比を表3に示した（小数点第2位を四捨五入したので，合計が46.0にならない）．

順位	件数	全体比
1〜10	503	46.0
11〜20	88	8.0
21〜30	39	3.6
31〜40	23	2.1
41〜50	24	2.2
51〜60	15	1.4
61〜70	19	1.7
71〜80	17	1.6
81〜90	10	0.9
91〜100	10	0.9
合計	748	(68.4)

表 2: 教員連想検索結果（上位100位まで）

順位	件数	全体比
1	186	17.0
2	98	9.0
3	60	5.5
4	38	3.5
5	40	3.7
6	26	2.4
7	21	1.9
8	15	1.4
9	9	0.8
10	10	0.9
合計	503	(46.0)

表 3: 教員連想検索結果（上位10位）

1位に全体の6分の1強の正解が現れ，2位までで全体の4分の1強の正解が現れ，5位までに全体の4割弱の正解が現れていることがわかる．この比率をどのようにとらえるかは議論の分かれるところであるが，九州大学では講座制を取るところが多く，類似する研究をする教員が数人いるという大学の特性，そして先述のように「研究・教育・社会連携活動概要」で用いられる語と「発明データ」で用いられる語の間にある質的な違い，「発明データ」が必ずしも現在の「研究・教育・社会連携活動概要」に一致するとは限らないことを考慮すると，このシステムは有効であるということができる．

さらに，これを裏付けるデータとして本稿で述べた教員連想検索システム (「本システム」と呼ぶ) と出現頻度に基づくシステム (「頻度によるシステム」と呼ぶ) で同じ実験を行い，それらの比較をした結果を図6に示す．頻度によるシステムは語の重みを無視し，出現頻度がその関連度であるとする方法である．簡単にいえば，出現頻度が高ければ関連度は高いと考える．

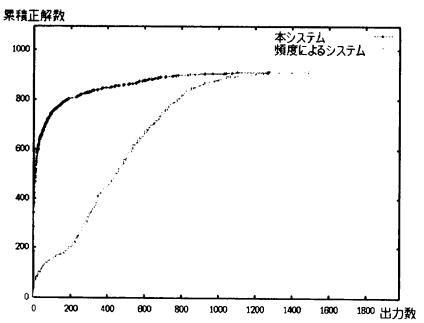

図6: 本システムと頻度によるシステムとの比較

図6は出力数を n としたときに，第 n 位までに正解が出現している特許・発明の件数をグラフにしたものである．このグラフでは，グラフが上側にきていれば少ない出力数で多くの正解を検索できると考えられるので，明らかに本システムの方がより良い結果を出力することがわかり，本システムの有効性が確認できる．

5 まとめと今後の課題

本稿では，九州大学の教員データに基づく教員連想検索システムを構築し，特許・発明データといったある種の答えがあるデータに対しての有効性も確認した．

この教員連想検索システムは2004年8月に「九州大学 教員連想検索システム」として学内のみで公開されるようになった．データ源となる「九州大学大学評価情報システム」は，各教員が自分のデータの保存・公開の作業がいつでも行えるので，最新の情報を常に公開できるという特色を持つ．そこで「九州大学 教員連想検索システム」でも最新の情報を提供できるように，毎月1回データの更新を行っている．現在，1日平均10件前後の利用があるが，研究グループを形成する必要があるプロジェクトの申請シーズンには50件弱の利用をされた日もあった．また，学内の研究戦略面で本システムが有効に利用されているという報告も受けている．

一方，以下のような問題点・要望が出されている．

1. 検索対象文書を「研究・教育・社会連携活動概要」にした場合はそれほどでもないが，それを「研究業績」にした場合，論文が英語で書かれている場合には日本語で検索した場合に望まれるであろう検索結果が得られないという問題．

2. 例えば「ユーザ」と「ユーザー」のような表記の違いにより関連なしと判定される問題．

3. ある教員が検索されたときに，その教員を特徴づけるはずの語が「研究の重要語」として出てこない場合がある問題．

4. 複雑な検索式を扱えるようにしてほしいという要望．

これらの問題の解決にあたって，1に対しては辞書機能をのせること，2に対しては表記を統一させる機能をのせることがまず考えられる．また，3の問題については，関連度の計算手法と形態素解析器の2つの問題が考えられる．前者については，単語の重み付けの方法に何らかの工夫を加えることを考えなくてはならない．後者については，多くの専門用語が「茶筌」の辞書に入っていないために生じていることが分かっているので，「茶筌」の

辞書を拡張すればよいが，その作業には相当の労力がかかるように思われる．4については，GETAの仕様上の問題であるが，うまく解決できる方法を考えたい．

また，教員連想検索システムでの関連度計算方式の比較や検索結果の分析などが，今後行うべき課題であろう．これらの諸課題の解決を進めるとともに，学外公開に向けた改善・改良を進めていきたい．

参考文献

[1] A.Singhal, C.Buckley and M.Mitra. Pivoted document length normalization, *Proceedings of SIGIR'96*, pp.21-29, 1996.

[2] Webcat Plus. http://webcatplus.nii.ac.jp/

[3] 九州大学研究者情報. http://hyoka.ofc.kyushu-u.ac.jp/search/

[4] 杉本典子，金丸玲子，池田大輔，竹田正幸，井上 仁，廣川佐千男. 九州大学自己点検・評価関連情報システム，情報処理学会 第41回デジタル・ドキュメント研究会資料, 2003.

[5] 高野明彦，西岡真吾，今一修，岩山真，丹羽芳樹，久光徹，藤尾正和，徳永健伸，奥村学，望月源，野本忠司. 汎用連想計算エンジンの開発と大規模文書分析への応用. http://geta.ex.nii.ac.jp/pdf/itx2002.pdf

[6] 汎用連想計算エンジン GETA. http://geta.ex.nii.ac.jp/