様式 C-19

# 科学研究費補助金研究成果報告書

平成２２年　５月２８日現在

研究種目：基盤研究（Ｃ）
研究期間：2007～2009
課題番号：１９５００９６
研究課題名（和文）　問い合わせ意図を抽出する機能を有する検索システムの構築

研究課題名（英文）　The study on constructing an information retrieval system with the mechanism of extracting users' retrieval intention

研究代表者
陳　幸生（CHEN YUKIO）
神奈川工科大学・情報学部・教授
研究者番号：１０２８２３４４

研究成果の概要（和文）：
情報システムの使用者の急激な増加と伴い、デジタル化した情報の爆発的な増加になっている。高精度の情報検索システムの開発が急務になってくる。特に、利用者の検索特徴を学習しながら検索精度を向上できる検索システムが望まれている。本研究では、本来人間同士が話すとき、“潜在的会話背景”を無意識に利用し、会話を進めると同様な原理で、利用者の検索意図に沿う高精度の検索システムの開発を目的として、研究開発を行った。利用者の検索意図の抽出方法、利用者検索意図に沿う検索方法、および、検索精度を向上するための検索意図学習方法を研究成果として得られて、その有効性を、実験を通して確認した。

研究成果の概要（英文）：
With the rapid increase in utilization of information systems, digitalized information has increased dramatically. It is important to develop information retrieval system with high retrieval accuracy. In particular, it is desired to develop the information retrieval system with the mechanism to increase the retrieval accuracy by learning the retrieval characteristic of users during the retrieving processing. In this research, study and development are performed in order to develop a retrieval system with high retrieval accuracy along users' retrieval intentions based on the principle referred to as “potential conversation background” that is unconsciously used when people talking to each other. As the research results, we obtained the method of extracting users' retrieval intention, the method of performing information retrieval based on user's retrieval intention and the method of learning user's retrieval intention in order to increasing retrieval accuracy. In the same time, the effectiveness of these methods was demonstrated on the developed experimental system.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| ２００７年度 | 1,800,000 | 540,000 | 2,340,000 |
| ２００８年度 | 600,000 | 180,000 | 780,000 |
| ２００９年度 | 1,100,000 | 330,000 | 1,430,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 3,500,000 | 1,050,000 | 4,550,000 |

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：情報検索

１．研究開始当初の背景

近年、多くの検索手法、検索エンジンが開発され、文字情報の検索はもちろん実現できるし、画像や音楽などのマルチメディア情報の検索も実現でき、検索精度も上がりつつある。

多数の検索手法を２つに分類できる。第１の手法は、パターンマッチングによる検索手法に基づき検索システムを構築する。第２の手法は、非パターンマッチング手法に基づき検索システムを構築する。これらの手法には次のメリット／デメリットがある。第１の方法は、検索者が検索問い合わせ意図をはっきり反映する文字パタンや画像パタンなどを入力すると、高精度の検索結果を得ることが可能である。しかし、その精度の向上が利用者に高精度のパタンの入力を要求する。第２の方法は、高精度のパタンが入力されなくても、第１の方法により検出できない情報の検出が可能である。しかし、検索問い合わせ意図に沿わない情報、つまり、ごみ情報を大量出力する場合がある。これらの方法は、いずれも検索結果の精度が利用者の入力した検索問い合わせの質に作用されるとの共通の欠点がある。

デジタル化した情報の利用者の爆発的な増加により、質が劣る問い合わせが検索システムに送られるのが一般的となってきた。また、携帯端末や、携帯電話や、パソコンなどの使用が個人専用の情報機器として検索問い合わせが検索システムに送信されるとなっている。このため、個人専用の情報機器から検索者の検索問い合わせ意図を探れる手法が期待される。

情報検索の分野では、利用者の問い合わせ、および検索の対象となるデータ群をベクトルとして表現し、それらのベクトル間の類似性に基づいて相関量を計算し、検索対象のランキング順位を決めるベクトル空間方式が広く利用されている。ベクトル空間モデルを用いた文書検索方式では、問い合わせと文書群の内容的な類似性の比較を行う文書検索に対して有効であると確認されている。しかし、文書間や単語間の関連性を計量するための特徴ベクトル空間の次元数増加に伴い、問い合わせと文書群の相関量の計算コストが大きくなり、また、文書中に含まれる不必要な単語群がノイズとして影響を及ぼすことにより、検索精度が低下する傾向がある。このため、特徴ベクトル空間の次元の縮小により、検索時の計算処理コストを下げることや、ベクトル空間を構成する適切な特徴単語群の選択により、利用者の検索意図や目的に適った文書群を検索結果の上位にランキングさせ、利用者が、自分自身の嗜好に合致する文書獲得を可能とする方式の実現が重要な研究課題となっている。

２．研究の目的

利用者が自らの検索特徴を適応し学習により検索精度を向上しながら使用していく検索システムが望まれていることに答え、利用者の検索問い合わせ意図をベクトル空間上に射影し、ベクトル空間の構成軸と検索問い合わせとの関連性を研究し、検索者の検索問い合わせ意図に沿う高精度の検索システムを研究開発していることを研究の目的とする。

本研究の目標は、検索者の問い合わせ意図に沿う高精度のサーチエンジンの研究開発である。そのポイントは、個人専用の情報機器、例えば、携帯電話などの携帯端末や、個人専用のパソコンから検索者の個人“潜在的検索問い合わせ意図”の抽出機能の実現方法と、検索問い合わせ意図に合致した検索空間の生成機能の実現方法と、“検索問い合わせ意図” メタレベル知識ベースシステムの構築方法である。また、各機能を連携し、ユーザ問い合わせの入力から検索結果の出力までのシステム全体の機能を、実験レベルで実現することである。

３．研究の方法

本研究では、研究代表者らが提案した文書群中に出現する単語群の特徴を抽出し、小さな計算コストで、低次元のベクトル空間を作成する方式（Feature Extraction Model、以下 FEM 方式と略）に基づき、利用者の検索意図や目的に適った検索方法の提案を行っている。特徴抽出モデルは、初めに、文書の意味的相関特徴の抽出するため提出され（陳幸生、清木 康、 “A query-meaning recognition method with a learning mechanism for document information retrieval,” Information Modelling and Knowledge Bases XV (IOS Press), Vol.105,

pp. 37-54)。FEM 方式では、意味や内容に基づいて分類された異なる文書群により、利用者の検索意図や目的と緊密な関連を有する特徴単語を抽出し、その特徴単語を利用し、特徴ベクトル空間を生成する。生成した特徴ベクトル空間は、利用者の検索意図や目的を反映できる特徴がある。

利用者の問い合わせ意図を抽出ために、本研究では、利用者の検索意図に従って複数の文書を、特徴ベクトル空間を生成するためのサンプル文書群として準備し、意味や内容に基づいてサンプル文書群を分類する。分類された異なる利用者の意図サンプル文書群の中に出現する単語群の特徴を利用し、低次元のベクトル空間を作成する。特徴ベクトル空間は、サンプル文書群から抽出した特徴単語を用いて生成される。特徴ベクトル空間は、分類された文書群により、利用者の検索意図や目的と緊密な関連を有する特徴単語に基づき作成されたため、利用者の検索意図や目的を反映できる特徴がある。特徴ベクトル空間は、サンプル文書群から抽出した特徴単語を用いて生成され、この段階において“潜在的検索問い合わせ意図”は反映されている。

また、利用者同士の“潜在的検索問い合わせ意図”を共有する機能を実現するため、共通サンプル文書群を準備し、共通特徴単語を抽出し、共有ベクトル空間を生成する。

更に、検索精度を向上するために、FEM 方式により生成された特徴ベクトル空間を対象として、その特徴ベクトル空間を形成する特徴単語を、利用者の意図や関心に基づいて修正することにより、利用者の意図や関心を特徴ベクトル空間に反映させる学習方式を導入する。本学習方式では、特徴ベクトル空間を構成する特徴単語群を対象として、検索処理における相関量の算出には寄与していないが、検索意図において必要な単語群を、逆に、相関量の算出に寄与している。検索意図に合致する必要な単語群、合致しない不要な単語群を、利用者の検索意図に基づいて分析することによって抽出し、それらの単語群の追加、削除、重み付けといった操作により、特徴ベクトル空間の最適化を行う。データベースに格納されている文書群と問い合わせは、最適化された特徴ベクトル空間上に射影されることにより、利用者の問い合わせ意図や関心に沿った最適な文書ベクトル群と問い合わせベクトルとして表現される。最適化された特徴ベクトル空間を利用した検索において、検索精度の向上が実現可能となる。

従来の研究では、利用者のフィードバックによる検索精度の向上を実現する方式として、適合フィードバック等が提案されている。適合フィードバックは、検索結果から正解文書や不正解文書を指定し、それに基づいて検索質問を修正する方式である。また、利用者の意図に合った情報を効率よく抽出するための、構造化されたレコード抽出方式も提案されている。検索精度を向上するために、利用者のフィードバックを利用している。

これらに対して、本研究の学習方式は、FEM 方式により生成された低次元の特徴ベクトル空間上に射影された文書ベクトルに着目し、文書ベクトル分布や、その文書ベクトルの形成に必要な特徴単語成分を分析した結果に基づいて、元の特徴ベクトル空間を再構成する方法を実現するものである。再構成された特徴ベクトル空間上では、文書ベクトル分布が改善されるため、FEM 方式による文書検索システムの検索精度を向上することが実現できる。

本研究の学習方式による、特徴ベクトル空間の最適化方式は、次の手順により実現される。

Step-1. FEM 方式により、選択・分類されたサンプル文書群から特徴単語群を抽出し、それらを用いて特徴ベクトル空間を生成する。

Step-2. Step-1 により生成された特徴ベクトル空間上におけるそれらの単語群の重要度を視覚化したグラフを、検索目的や意図の分析道具として提供し、利用者は、システムを学習させるために、修正フィードバックを入力する。

Step-3. Step-2 における分析結果に基づき、生成した特徴ベクトル空間を利用者のフィードバックにより修正する

Step-1 では、文書の意味や内容に基づいて分類した文書群の中から、利用者の検索意図に従って複数の文書を選択・分類する。このようにして準備された文書群から抽出した特徴単語を用いて、FEM 方式により特徴ベクトル空間を生成する。この段階でも利用者の検索意図は特徴ベクトル空間に反映される。しかし、サンプル文書群には、利用者以外の人間が作成した文書が含まれる可能性がある。また、特徴ベクトル空間の生成過程において、利用者の意図に反して重要な単語が特徴単語から除外されてしまう等により、利用者の検索意図を反映しきれない場合がある。従って、Step-2 によって特徴ベクトル空間に利用者の検索意図が反映されているか、また、反映されていないのであればどのようなフィードバックを行う必要があるのかを分析する。分析結果に基づいて、Step-3 において利用者のフィードバックにより特徴ベクトル空間を修正する。これにより、利用者の意図や関心を特徴ベクトル空間へ反映する。

４．研究成果

本研究では、検索者の個人特徴を抽出し、それに基づき、問い合わせ／検索対象の各次元軸上の射影値の計量方法、検索者の検索問い合わせ意図に沿う問い合わせ／検索対象の各次元軸上の正確な射影位置を求める方法を研究結果として得られた。

本研究では、提案方式を適用することにより、Web 文書群を対象として、利用者の検索意図や関心を反映するための学習機構を伴った実験用検索システムを構築した。この実験用検索システムは、Web ブラウザに登録されているブックマーク情報を、利用者の意図や関心を示すプロファイル情報の一つと捉え、このブックマーク情報を利用して、利用者の嗜好に沿った特徴ベクトル空間を生成する機能を有する。また、利用者の登録している Web 上の RSS（Rich Site Summary）により配信された記事文書をデータベースに格納し、これらの記事文書群を生成した特徴ベクトル空間上へ射影する機能を備えている。

利用者は、本実験システムにより、ブラウザのブックマーク情報から、Web 文書を選択・分類することにより、自分の嗜好に合致する記事文書獲得を可能とする特徴ベクトル空間を生成することができる。また、利用者は、自分の期待する文書群が獲得できない場合においても、学習操作により、特徴ベクトル空間を修正し、検索意図や目的に応じた文書獲得を行うことができる。

本研究では、実装した実験検索システムを用いて、京都に関する Web 文書群、および、利用者が購読登録をしている Web 上の RSS により配信された記事文書群を対象とした検索実験を行い、提案方式において利用者の嗜好に沿った文書群の獲得が可能であることを明らかにした。

（1）“潜在的検索問い合わせ意図”の抽出

個人専用の情報機器に貯蓄された過去の閲覧情報、検索情報などの動的に変化するデータからデータ間の相関特徴を特徴抽出 FEM モデルにより、実現した。本研究では、貯蓄された過去の閲覧情報、検索情報に特徴抽出モデルを適用し、その中から利用者個人の“潜在的検索問い合わせ意図”の抽出機能を実現した。

（2）ベクトル検索空間の作成

ベクトル方式を用い情報検索機能を実現する場合、問い合わせ／検索対象の単語出現頻度や、正規化単語出現頻度などの特徴を多次元ベクトル空間の各次元軸上の射影値として扱い、問い合わせ／検索対象を多次元のベクトルとして表現する。各特徴が互い直交と設定すると、各特徴の値が１：１の比例で空間の各軸上に射影できる。しかし、利用者の検索意図により、各特徴間の関係が直交として扱えない場合がある。例えば、“ハードウェアー”と“ソフトウェア”との２つの単語は、利用者の検索意図により、互い直交にならない場合がある。図 1 で示すように、単語“ソフトウェア”を、１：x（0<x<1）の比例で軸 qj 上に射影しなければならない。しかし、互いに意味が異なると定義したら、単語“ソフトウェア”を軸 qj に１：１の比例で射影できる。

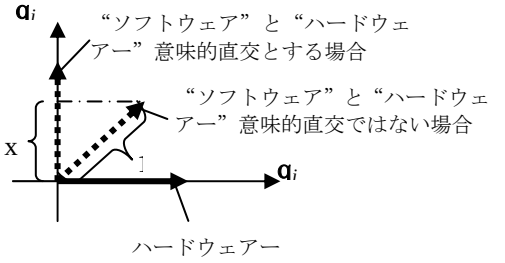


図 1. 問い合わせ／検索対象の特徴値をベクトル空間上に射影する例

本研究では、“潜在的検索問い合わせ意図”の抽出機能を使用し、特徴の空間軸への射影比例値を求めることにより、検索問い合わせ意図に沿うベクトル検索空間作成機能を実現した。図２は、実験システムにより抽出した特徴単語毎の寄与値を算出した結果を示している。それぞれ「八つ橋」、「生八つ橋」および「舞妓」、「お茶屋」など、利用者の意図に応じて選択したサンプル文書に基づいた特徴単語の寄与により、座標値が算出されていることがわかる。

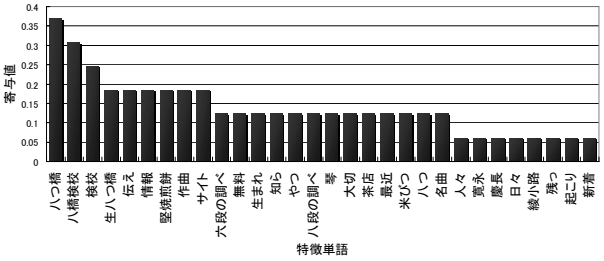


図２　実験システムにより抽出した特徴単語毎の寄与値を算出した結果

（３）　問い合わせ／検索対象の空間射影、及び相関量の計量

問い合わせ／検索対象の空間射影機能は、検索エンジン側の問い合わせ／検索対象の空間射影機能と、ユーザ側の問い合わせ／検索対象の空間射影の機能により実現された。

図３は、実験システムの特徴ベクトル空間上に、検索対象文書群を射影した場合の実験

結果である。空間上では、C1、C6 および C7 の各軸上で、サンプル文書と同じ分類の文書群が、高い座標値を示しており、また、サンプル文書が属さない分類の文書群が、小さい座標値を示しており、利用者の検索意図に基づいた適切な検索対象の空間分布を実現していることが確認できる。

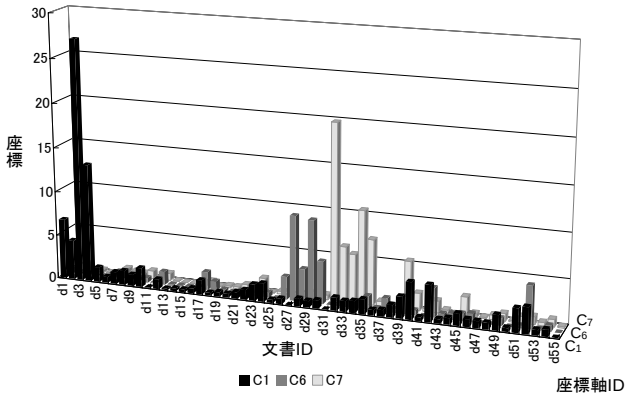


図３　実験システムの特徴ベクトル空間上に、検索対象文書群を射影した場合の実験結果

本研究で提案した学習方式は、FEM 方式により生成された特徴ベクトル空間を対象として、利用者のフィードバックによる学習操作により、最適な特徴ベクトル空間の構成ができると確認した。

図４は、特徴空間上に検索意図に従い、「平安京」という特徴単語を追加した場合の文書分布である。

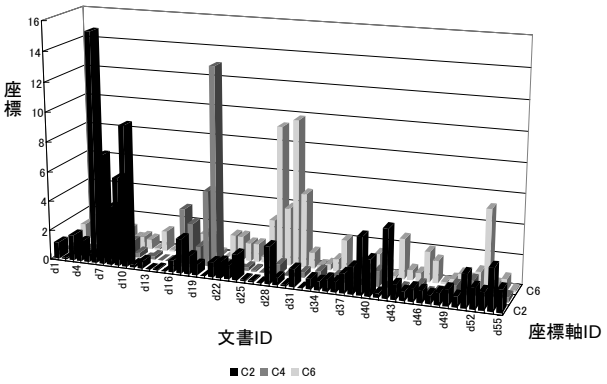


図４　特徴空間上に検索意図に従い、特徴単語を追加した場合の検索対象の特徴空間上の分布様子

図５は、特徴空間上の検索対象 d20 の特徴単語の寄与値を示しており、特徴単語「平安京」の寄与値が大きくなっていることがわかる。このように、特徴空間では、特徴ベクトル空間上における利用者の「平安京」という意図が強調されたため、検索対象 d16～d20 の座標値が大きくなり、利用者の意図に基づいて文書分布が改善されていることがわかる。

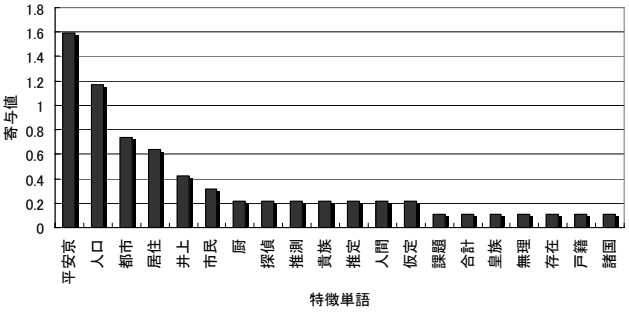


図５　特徴空間上の検索対象 d20 の特徴単語の寄与値

図６は、検索意図に基づき、学習操作を行い、検索対象 d42 や d46 の特徴単語を削除した後、検索対象の特徴空間上の分布結果である。図６で示した特徴空間では、文書 d42 や d46 とともに、これらの文書と同じ分類に属する文書 d45、d47、d49 などの座標値が減少しているが、関連性のある文書 d16～d20 の座標値も減少していることが確認できる。

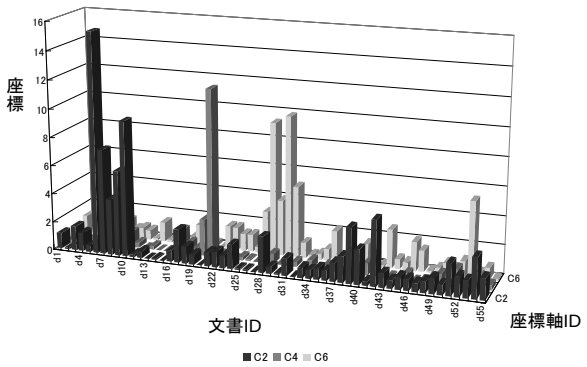


図６　検索意図に基づき、学習操作を行い、徴単語を削除した後、検索対象の特徴空間上の分布結果

図７は、実験システムにおいて行った問い合わせの再現率・適合率の関係を示している。学習操作は、検索意図に従い特徴空間の修正を行った。修正後の特徴空間S”$_2$では各再現率における不適合文書の割合が、修正前の特徴空間$S_2$よりも改善され、検索精度が向上していることが確認できる。

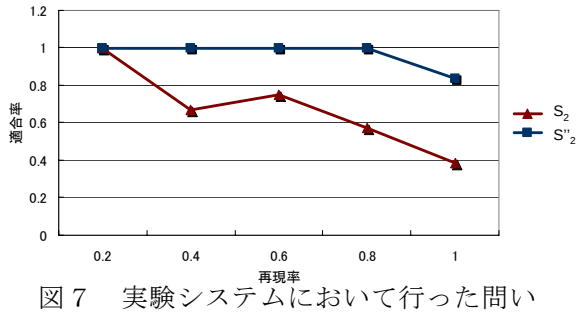


図７　実験システムにおいて行った問い合わせの再現率・適合率の関係

図８は、問い合わせ「情報流出」による検索結果の上位 20 件について、順位ごとの適合率を算出し、作成したグラフである。この実験では、検索意図に基づき特徴空間を作成し、さらに、学習方式を用い、特徴単語の追加、および、特徴単語の重み操作も行い、検索を行った。図に示した結果は、問い合わせ「情報流出」を検索語として使用して、検索意図に基づき、特徴単語「セキュリティ」を加え、更に「セキュリティ」の重みを他の特徴単語により２倍に増加させた場合の上位20件検索結果である。このグラフからも、検索意図に基づき特徴空間の操作を行うことにより、検索結果が向上していることがわかった。

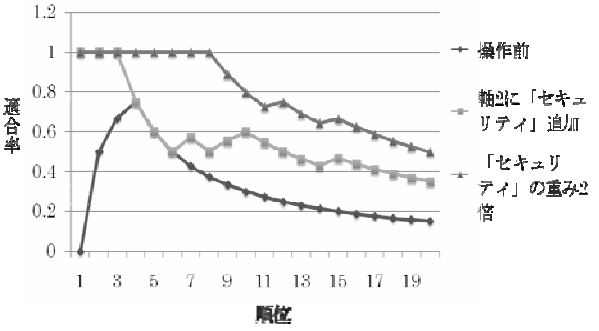


図８　検索意図に基づき特徴空間の操作を行うことにより、検索精度を向上する様子

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）


① Hannu Jaakola, Bernhard Thalhem, Yutaka Kidawara, Koji Zettsu, Xing Chen and Anneli Heimburger, Information Modelling and Global Risk Management Systems, Information Modelling and Knowledge Bases, 査読有, Vol. 20, 2009, pp. 429-445.

② 鷹野孝典、増田圭祐、陳幸生, A Framework of a Feedback Process for Analyzing and Personalizing a Document Vector Space on a Feature Extraction Model, Information Technology and Management, 査読有, Vol. 10, No. 2-3, 2009, pp. 151-176.

③ 陳幸生, 清木康, 鷹野孝典, 増田圭祐, A Semantic Space Creation Method with an Adaptive Axis Adjustment Mechanism for Media Data Retrieval, Information Modelling and Knowledge Bases, 査読有, Vol. 19, 2008, pp. 40-58.

④ 鷹野孝典, 陳幸生, 文書ベクトル分布に基づく次元削減による文書検索空間の生成方式に関する評価実験, 情報処理学会研究報告, 査読無, Vol. 88, 2008, pp. 55-60.

⑤ 鷹野孝典, 増田圭祐, 内山亜美, 陳幸生, 動的フィードバック機能をともなった特徴語抽出方式における文書ベクトル空間改善プロセスに関する評価実験, 情報処理学会研究報告, 査読無, Vol. 66, 2007, pp. 61-66.


〔学会発表〕（計５件）


① 鷹野孝典、倉林修一、陳幸生、清木康, An Adaptive Search System using Heterogeneous Document Vector Spaces, IEEE, 2009, Victoria, BC, Canada

② 金子洋平, 鷹野孝典, 陳幸生, デスクトップ環境における利用者の検索意図に基づく情報獲得システムについての評価実験, 日本データベース学会, 2009, 静岡県掛川市.

③ 鷹野孝典, 陳幸生, 増田圭祐, Experiments of a Feedback Method for a Document Vector Space Based on a Feature Extraction Method, AIS, 2007, Montréal, Québec, Canada.

④ 内山亜美, 鷹野孝典, 陳幸生, A Distinct-based Subspace Creation Method for a Semantic Document Retrieval, IEEE, 2007, Surabaya, Indonesia.

⑤ 鷹野孝典, 陳幸生, 増田圭祐, A Feedback Method of Improving the Performance of a Document Vector Space Developed on a Feature Extraction Model, IEEE, 2007, Victoria, B.C., Canada.


６．研究組織
(1)研究代表者


陳　幸生（CHEN　YUKIO）
神奈川工科大学・情報学部・教授
研究者番号：10282344