様式Ｃ－１９

# 科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月２０日現在

機関番号：１２３０１
研究種目：若手研究（Ｂ）
研究期間：2009～2012
課題番号：２１７００２７３
研究課題名（和文）　分類型検索システムの多言語対応に関する研究

研究課題名（英文）　A Study on Multilingualization of Web Search Clustering

研究代表者
安川　美智子（YASUKAWA MICHIKO）
群馬大学・大学院工学研究科・助教
研究者番号：70361384

研究成果の概要（和文）：
本研究では、検索語の関連語を用いて Web 検索結果をクラスタリングする、文書分類型の検索方式（分類型検索）の研究開発を行うことを目的とし、特に実用的な検索機能を持つ、多言語対応の分類型検索システムの開発に取り組んだ。研究期間中に、日本語以外の言語に対応できるように分類型検索システムを拡張し、情報検索システムの評価用テストコレクションを用いて、多言語対応の分類型検索が言語に依らず有用であることを定量的に確認した。

研究成果の概要（英文）：
In this study, we achieved multilingualization of web search clustering, which has remarkable ability to generate document clusters according to related terms of web search terms. One of our achievements is extending the prototype system of clustering search system from Japanese to other languages. We performed a series of evaluation experiments to confirm our achievement in increasing search effectiveness.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2009 年度 | 900,000 | 270,000 | 1,170,000 |
| 2010 年度 | 800,000 | 240,000 | 1,040,000 |
| 2011 年度 | 700,000 | 210,000 | 910,000 |
| 2012 年度 | 700,000 | 210,000 | 910,000 |
| 年度 | | | |
| 総　計 | 3,100,000 | 930,000 | 4,030,000 |

研究分野：情報検索
科研費の分科・細目：情報学・図書館情報学・人文社会情報学
キーワード：情報組織化

## １．研究開始当初の背景

Web 検索エンジンを用いた情報の検索や収集が日常に欠かせないものとなっている。たとえば、流行している物や現象、人、企業、商品、サービス、テレビ番組などが多数のユーザにより検索されているが、従来型の検索方式には、ユーザが具体的かつ詳細な検索クエリを入力しなければならないという問題がある。

検索対象(検索したいこと）について知るための検索であるにもかかわらず、検索対象についての詳細な説明を入力させる従

来型の Web 検索エンジンでは、目的とする情報に辿りつく前に検索を断念せざるを得ない場合も生じてしまう。この問題を解決するため、本研究では、検索語の関連語を用いて Web 検索結果をクラスタリングする、文書分類型の検索方式（分類型検索）についての検討を行った。

分類型検索とは、文書を分類することによってユーザビリティを向上する情報検索技術である。従来から、文書を特徴量の類似性により分類する試みがなされ、小規模な文書群など、ノイズがほとんど含まれない文書群では、ある程度の実用性が認められてきた。

しかしながら、多様なトピックを含む大規模な文書群には分類の精度を悪化させる不要な特徴量が多く含まれる。このため、文書群に含まれる不要な特徴量を効果的に除去し、ユーザが一瞥して理解できるような、分かりやすい分類提示を行う情報検索技術の開発が早急な課題となっている。

２．研究の目的

本研究では検索語の関連語を用いた分類型検索システムの研究開発を目的とし、特に実用的な検索性能を持つ、多言語対応の分類型検索システムの検討を行うことを主たる目標としている。検索エンジンに検索語と共に入力される関連語は、一般性が高く、ユーザにとって馴染みのある語が多いという特徴がある。本研究では、研究期間内に以下のことに取り組んだ。

[多言語対応] 日本語、および、日本語以外の言語(具体的には、まず英語と中国語) にも対応できるようにシステムを拡張する。

[検索性能評価] 情報検索システムの評価用テストコレクションを用いて、多言語対応の分類型検索が言語に依らず有用であることを定量的に確認する。

３．研究の方法

本研究において分類型検索システムを多言語対応に拡張することで、日本語以外の言語においても提案手法が適用可能となり、開発システムは国内外の多数のユーザに利用され得る。本研究では、まず、日本語文書を検索する場合（日本語→日本語）の評価実験を行った。日本語における有効性を確認し、日本語以外の言語に対応できるようにシステムを拡張した。

次に、英語での提案手法の有用性を確認し、その他の言語（マレー語）についても言語資源を収集し、提案手法の有用性を考察した。評価実験には、過去の NTCIR ワークショップで構築されたテストコレクションを活用し、検索有効性の評価を行った。

４．研究成果

（１） 平成 21 年度の研究成果

文書分類型システムのプロトタイプを拡張し、単語の文字列長を考慮した分類型検索を提案した。また、日本語と英語の地図情報検索を目的とした関連語による分類型検索システムの検索性能の評価を行った。提案法により携帯電話などの小さな端末でも表示画面を有効活用した分類結果の提示により、効果的かつ効率的に地図情報を検索できることが確認できた。また本年度は、日本語と英語以外の他の言語に対応するための準備としてマレー語のステマーを開発した。開発したステマーは、マレー語の語幹と派生語に対する過剰な接辞処理を抑制するため語幹辞書と派生語辞書を参照する。開発したステマーを、マレー語の文書自動分類に応用し、過剰な接辞処理が効果的に抑制されていることを確認した。

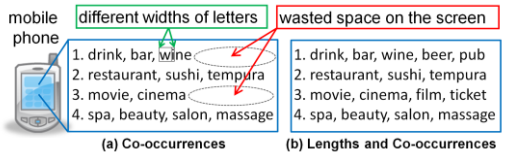


図１：単語の文字列長を考慮した携帯電話向けの分類型検索

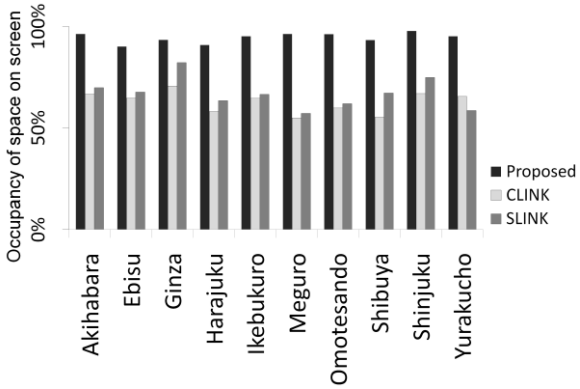


図２：画面上に表示される文字の表示効率（クラスタリングの手法ごとの比較）

（2） 平成 22 年度の研究成果
日本語の特許文書群から複合語を抽出し、文書分類のための単語リストを生成する手法について検討した。特許庁の審査官は審査の過程で新規性のない発明を拒絶するために、拒絶の根拠となる別の特許を過去の膨大な特許文書群の中から検索しなければならない。この検索を支援するためには、種々の検索要素技術を組み合わせて精度、再現率を高める工夫が必要となる。本年度の研究では、そのような要素技術の一つとして、特許文書に含まれる、漢字・カタカナからなる複合語を抽出する手法を検討した。特許文書は従来にない新規の技術を説明する技術文書であり、フォーマルな文体で記述されることから、ひらがな表記の単語の使用頻度は低く、カタカナや漢字の専門用語が多数使用されている。たとえば、特許文書には、化学物質、原料、医薬品、食品、農業器具、工業製品の名称などの専門用語が多数含まれる。このため、特許文書群においては、特に漢字・カタカナの単語に対する形態素解析の誤りが多数発生し、もともと意味のあった文字列が無意味な短い文字列に分割されるという問題が生じる。この問題に対して、本年度の研究では、まず、すべての特許文書に対して形態素解析を行い、文字種を手がかりとして過剰に分割された形態素列から複合形態素の候補を生成する手法を検討した。さらに、これまでの研究で取り組んだマレー語の接辞処理の手法を発展させ、日本語の漢字・カタカナの複合形態素の候補から、無意味な接辞部分を取り除く手法を検討した。接辞を除去した複合形態素を文書分類のための単語リストに含めることで、重要な複合形態素のみからなる決定則の抽出が期待できる。文書分類のための決定則抽出のソフトウエアと特許検索タスクテストコレクションを用いた評価実験の結果、提案法により文書分類の精度（F-measure）が向上することを確認した。

（3） 平成 23 年度の研究成果
前年度までの研究では表記された文書の意味的な類似性に基づく分類型の文書検索の研究開発を進めてきたが、本年度は、これまでの研究をさらに発展させ、日本語の文書が発音された際の音声の類似性の観点から文書検索を行えるように検索システムを拡張する研究開発を行った。具体的には、英語で標準的に用いられている手法を拡張し、日本語の音声を考慮した発音照合を行うアルゴリズムを開発し、評価を行い、得られた研究成果を情報処理学会自然言語処理研究会において報告した。さらに、本年度は、海外の研究者グループとの共同研究を行い、NTCIR-9 の検索タスク GeoTime と検索タスク INTENT に参加し、言語に依存しない文書索引付の手法とランク付けの手法について研究開発を行った。参加した情報検索システムの評価型ワークショップ NTCIR-9 では、GeoTime(地理的・時間的情報の検索) タスクの英語と日本語の新聞記事データ、および、INTENT(検索意図)タスクの日本語の Web 文書データに対して、言語に依存しない文書索引付の手法の実現可能性を検証できた。また、文字列の曖昧性や多義性を解消するため、日本語版の Wikipedia を用いた類義語の定義の手法を検討し、GeoTime (地理的・時間的情報の検索) タスクと INTENT(検索意図)タスクにおいて、類義語を用いた検索質問拡張について検討した。

（4） 平成 24 年度の研究成果
前年度までに予備的検討を進めてきた日本語の発音照合の手法を改良し、オープンソースソフトウエアのデータベース管理システム PostgreSQL のユーザ定義関数として実装を行い、評価を行った。発音照合は、類似文字列検索の手法のひとつであり、文字列の綴りではなく、文字列が発音された音声の類似性に基づく検索の手法である。英語における先行研究では、文字列編集距離と発音照合を組み合わせた情報検索の有効性が明らかにされていることから、本研究では、英語で提案されている手法の日本語化と日本語の文書検索における発音照合の有効性評価に取り組んだ。具体的には、まず、昨年度までの研究で開発してきた日本語の発音照合の符号化表の拡張と改良を行い、発音照合の手法をデータベース管理システムの動的ロード可能オブジェクトとして実装し、動作の検証を行った。さらに、昨年度までの研究で文書検索の評価に使用した情報検索システムのテストコレクションを使

用して、文字列編集距離と日本語の発音照合を組み合わせた手法の有効性評価を行い、得られた知見をSIGIR2012 Workshop (OSIR2012)で報告した。また、海外の共同研究者とともに予備的検討を進めてきた文字n-gramを用いた言語に依存しない文書索引付の手法の検索効率性と検索有効性を検証するための評価実験を行い、得られた結果に基づく検討を行った。具体的には、過去の情報検索システムの評価型ワークショップNTCIR7/8 IR4QAで構築されたテストコレクションを用いて、言語に依存しない文書索引付の手法の評価方法を検討し、次に、種々の文書類似度の尺度を用いて、文字n-gramを用いた検索手法と、転置索引を用いた従来法との比較を行った。

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　1件）

(1) M. Yasukawa, H. T. Lim, H. Yokoo, “Stemming Malay Text and Its Application in Automatic Text Categorization,” IEICE Transactions on Information and Systems, Vol.E92-D, No.12, pp.2351-2359, 2009.（査読有り）

〔学会発表〕（計　8件）

(1) M. Yasukawa, J. S. Culpepper, F. Scholer, “Phonetic Matching in Japanese,” Proc. SIGIR2012 Workshop on Open Source Information Retrieval (OSIR2012), pp.68-71, 2012.（査読有り）
(2) J. S. Culpepper, M. Yasukawa, F. Scholer, “Language Independent Ranked Retrieval with NeWT,” Proc. the 16th Australasian Document Computing Symposium (ADCS2011), pp.18-25, 2011.（査読有り）
(3) M. Yasukawa, J. S. Culpepper, F. Scholer, M. Petri, “RMIT and Gunma University at NTCIR-9 GeoTime Task,” Proc. NTCIR-9 Workshop Meeting, pp.69-74, 2011.（査読無し）
(4) M. Yasukawa, J. S. Culpepper, F. Scholer, M. Petri, “RMIT and Gunma University at NTCIR-9 Intent Task,” Proc. NTCIR-9 Workshop Meeting, pp.143-149, 2011.（査読無し）
(5) 安川美智子, 横尾英俊, “発音照合アルゴリズムを用いた早口言葉の検索,” 情報処理学会研究報告自然言語処理, 2011-NL-204(1), pp.1-8, 2011.（査読無し）
(6) 鶴巻有香, 安川美智子, 横尾英俊, “子音に注目した早口言葉の検索,” 情報処理学会研究報告自然言語処理, 2011-NL-201(14), pp.1-6, 2011.（査読無し）
(7) M. Yasukawa, H. Yokoo, “Composition and Decomposition of Japanese Katakana and Kanji Morphemes for Decision Rule Induction from Patent Documents,” Proc. the 15th Australasian Document Computing Symposium (ADCS2010), pp.28-35, 2010.（査読有り）
(8) M. Yasukawa, H. Yokoo, “Term Clustering based on Lengths and Co-occurrences of Terms,” Proc. the 14th Australasian Document Computing Symposium (ADCS2009), pp.126-128, 2009.（査読有り）

〔その他〕
ホームページ等
無し

６．研究組織
(1)研究代表者
安川　美智子（YASUKAWA MICHIKO）
群馬大学・大学院工学研究科・助教
研究者番号：70361384

(2)研究分担者
無し

(3)連携研究者
無し