様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１２　日現在

機関番号：　１２３０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０３６３

研究課題名（和文）類似文字列検索と分類型検索を活用した調理用語の体系化に関する研究

研究課題名（英文）A Study on Systematization of Culinary Vocabulary by Approximate String Matching and Related Terms Clustering

研究代表者

安川　美智子（Yasukawa, Michiko）

群馬大学・大学院理工学府・助教

研究者番号：７０３６１３８４

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：本研究では、申請者のこれまでの研究成果をさらに発展させ、特に、食や調理に関する用語を収集し、分類することを目的とし、大規模な料理レシピコーパスから調理に関する特徴語を抽出し、分類型の検索方式（分類型検索）と組み合わせ、ユーザにとってわかりやすい関連語を提示する手法、および、その情報検索への応用について検討した。提案法により、具体的な検索語（料理名、食材名、調理法を表す用語）がわからない場合でも、検索結果を効率よく絞り込むことができ、食に関する情報の流通を促進することが期待できる。

研究成果の概要（英文）：In this research, we investigated systematization of food related words and phrases. Our method uses two different approaches for related search: (1) approximate string matching and (2) related terms clustering. Specifically, our method extracts feature words and phrases in an arbitrary length from a large recipe corpus, groups them by attribute types, and presents the searched related words and phrases according to the user's search intent. An exploratory search with the proposed method is expected to improve usefulness of searching systems on food related information by suggesting dish names, ingredient names, and cooking methods even if users do not know effective query words beforehand.

研究分野：　情報検索

キーワード：　情報組織化

１．研究開始当初の背景

申請者は、研究開始当初までの研究開発で、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、多言語対応の分類型検索システムの開発に取り組んだ。本研究では、これまでの研究成果をさらに発展させ、特に、食や調理に関する用語を収集し、分類する手法を検討した。

一般に、何かを探索中に検索語の関連語を選択することは、具体的な関連語を考えることよりも容易である。このことから、システムが情報を整理してユーザに提示し、ユーザが選択肢の一覧の中から興味のあるものを選ぶ「ナビゲーション型の情報探索」の方が、情報要求を表現する適切な単語をユーザが考える「アドホック検索」よりも、ユーザにとって楽な作業であると言える。キーワードによるアドホック検索と情報を整理したカテゴリ体系によるナビゲーション型の情報探索を組み合わせることで、選択肢の絞り込みと拡大を円滑に行えるようにしているが、カテゴリ体系を用いたナビゲーション型の情報探索は、カテゴリを人手で管理する手間が大きいという問題がある。文書クラスタリングは、文書の自動分類によりカテゴリ管理の手間を省いたナビゲーション型の情報探索を可能にしているが、分類の整合性や一貫性の点で、ユーザにとって直感的でない情報提示がされるという欠点がある。興味のあるカテゴリを見つけるまでに時間がかかりすぎる場合、ユーザは求める情報に到達できなくなる。ナビゲーション型の情報探索において、ユーザにとって直感的なカテゴリ体系を自動構築できることが必要である。

２．研究の目的

本研究では、大規模な料理レシピコーパスから調理に関する特徴語の抽出を行い、文書分類型の検索方式（分類型検索）と組み合わせ、ユーザにとってわかりやすい関連語の提示と情報検索について検討する。多様な内容を含む情報の検索が仕事や日常生活に欠かせないものとなっており、「Search Diversity(検索結果の多様性)」が近年、学術的に重要な研究課題となっている。「検索結果の多様性」を考慮していない従来型の情報検索方式では、検索結果の曖昧性や多義性を排除するために、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った。この問題を解決する一つの手法として多様な情報を分類して表示する分類型検索がある。分類型検索においては、検索ユーザが検索対象をよく知らなくても、関連語のグループを一瞥するだけで、検索結果にどのような情報が含まれているのかを把握できる。本研究では、特に、食や調理に関する用語を収集し、分類する手法を検討する。本研究は、国民の毎日の食生活に深く関わる調理用語を分類し、電子化辞書の構築を行うという点に、学術的な特色がある。本研究の目的は、具体的な検索語（料理名、食材名、調理法を表す用語）がわからない場合でも、検索結果を効率よく絞り込むことを可能とし、食に関する情報の流通を促進することである。

３．研究の方法

本研究は、調理用語を分類し、電子化辞書を構築することを主たる目的としている。電子化辞書は、記録媒体として DVD メディアや HDD などの電子的手段を用い、機械可読な形式で構造化された辞書データである。調理用語を体系化し、電子的なデータとして整備することで、食に関する情報を含む文書検索システムや栄養計算ソフトウェアのユーザビリティの向上が期待できる。本研究では、研究期間内に以下の方法で研究を行った。

［検索タスクの設計と評価実験］大規模なユーザ生成コンテント(User Generated Content)を検索対象とするアドホック(ad hoc)検索タスクと献立作成を簡易化したタスク(recipe paring)のタスク設計と評価実験、及び、研究者コミュニティの共通の評価基盤となるテストコレクションの構築を行った。

［検索有効性の評価］　形態素と文字 n-gram（長さ n の文字列）の両方を索引語とする手法を活用して、形態素解析が困難な未知語（辞書に登録されていない長い文字列）や複合語（二つ以上の形態素が結合し新たに一つの語としての意味をもつもの）の検索有効性の評価を行った。

［関連語推薦と検索質問拡張］　大規模なユーザ生成コンテント(User Generated Content)を検索対象として、連想検索と類似文字列検索の組み合わせによる分類型検索の拡張を行い、以下のことを検討した。
●検索対象を文書とする場合(単語→文書、文書→文書)
単語、フレーズ（複数の単語)、文書を検索クエリとして文書群を検索する。
●検索対象を単語とする場合(文書→単語、単語→単語)
単語、フレーズ（複数の単語)、文書を検索クエリとして単語群を検索する。

４．研究成果

本研究の主な成果は以下の通りである。

（１）　平成 26 年度の研究成果
文部科学省科学技術・学術審議会資源調査分科会報告 「日本食品標準成分表２０１０」 に収載されている食品名(Food and description)を抽出し、機械可読な形式で、単語のリストを作成した。また、大規模なユーザ生成コンテント(User Generated Content)に含まれる食品名の表記揺れを、類似文字列検索の手法を用いて機械的に抽出し、同じ食品名を表す表記が異なる単語のリストを半自動的に作成する手法について検討した。さらに、NTCIR-11 ワークショップに参加し、調理や食材についての情報を扱う情報アクセスのための技術について、議論と意見交換を行った。具体的には、ソーシャル・ネットワーキング・サービス(social networking service)に投稿された日本語の料理レシピと、情報推薦(recommendation) と 個 別 化(personalization)を行うウェブアプリケーション（Web application）でクローリングされた英語のレシピを検索対象とする料理レシピ検索のアドホック（ad hoc）検索タスクと、献立作成を簡易化したレシピ・ペアリング(recipe pairing)のタスクを設計し、検索結果データの作成、人手での適合判定、公式評価結果の出力と結果に基づく分析検討を行った。

（２）　平成 27 年度の研究成果
日本語の形態素(morpheme; 意味を持つ最小の言語単位) と文字 n-gram(長さ n の文字列)を組み合わせた検索手法の検索有効性を評価する実験を行い、得られた知見を 2015 年 12 月 8 日から 12 月 9 日の期間にシドニー(オーストラリア)で開催された「 20th Australasian Document Computing Symposium (ADCS2015)」において口頭発表およびポスター発表を行ない、「Best Paper Award」を受賞した。実施した評価実験では、日本語の検索システム評価用のテストコレクションを使用し、検索対象となるコーパスに日本語の形態素解析器 ChaSen、MeCab、KaKaSi、Juman、KyTea を適用して、5 種類の単語の索引を作成した。また、コーパスに含まれるテキ

ストデータを入力として、文字列の長さを1〜3に変化させたn-gramの索引を作成した。単語の索引を用いた検索は、n-gramの索引を用いた検索に対して、ほとんどの検索トピックにおいて検索有効性が高いが、形態素解析が困難な未知語（辞書に登録されていない長い文字列）や複合語（二つ以上の形態素が結合し新たに一つの語としての意味をもつもの）の検索では、n-gramの索引に劣る検索トピックが存在し、その数は無視できない程度存在することを確認した。異なる索引からの検索結果を組み合わせるデータ融合により、単語の索引を用いた検索が失敗する検索トピックにおいて、n-gram の索引を用いた検索が有効に機能し、これにより、適合文書の検索漏れを防ぐことができる。実験の結果、提案手法はベースライン（単語の索引を用いた検索）に対して検索有効性の向上が統計的に有意であることが確認できた。

Algorithm 1 Boolean-search to calculate $\mathrm{S}(d_i)$.

$\mathrm{S}(d_i) \leftarrow 0.0$
for $t_j \in d_i$ **do**
  if t_j is included in Q_{OR} **then**
    $\mathrm{S}(d_i) \leftarrow \mathrm{S}(d_i) + 1.0$
  end if
end for
if any $t_j \in Q_{\mathrm{AND}}$ is not included in d_i **then**
  $\mathrm{S}(d_i) \leftarrow 0.0$
end if
if any $t_j \in Q_{\mathrm{NOT}}$ is included in d_i **then**
  $\mathrm{S}(d_i) \leftarrow 0.0$
end if

図１：検索範囲の絞り込み

Algorithm 2 Cross-search for related terms.

for $t_j \in \mathrm{INDEX}_{\mathrm{TARGET}}$ **do**
  $\mathrm{DF}(t_j) \leftarrow 0.0$
end for
for $d_i \in \mathrm{INDEX}_{\mathrm{TARGET}}$ **do**
  calculate $\mathrm{S}(d_i)$ in $\mathrm{INDEX}_{\mathrm{SOURCE}}$ with query Q
  if $\mathrm{S}(d_i) \geq 1.0$ **then**
    for $t_j \in d_i$ in $\mathrm{INDEX}_{\mathrm{TARGET}}$ **do**
      $\mathrm{DF}(t_j) \leftarrow \mathrm{DF}(t_j) + 1.0$
    end for
  end if
end for

図２：単語重要度の計算

（３）平成 28 年度の研究成果

英語の大規模レシピコーパスを用いて、料理レシピの文書構造に注目した関連語推薦と対話的な検索質問拡張への応用についての検討を行い、得られた知見を 2016 年 11 月 10 日に慶応義塾大学日吉キャンパス(神奈川県横浜市)で開催された「人工知能学会第 14 回 インタラクティブ情報アクセスと可視化マイニング研究会」において口頭発表を行い、「2016 年度 研究会優秀賞（JSAI SIG Research Award 2016）」を受賞した。提案手法では、検索ユーザにとって直感的な関連語推薦を行うために、「分類の粒度」と「分類の規則性」の 2 つの点を考慮した関連語推薦を行う。NTICR-11 の料理レシピ検索タスクのテストコレクション（英語の大規模レシピコーパス、アドホック検索クエリ、適合判定ファイル）を用いた評価実験を行い、提案法による検索有効性の向上を確認した。また、提案手法による単語検索（図１、図２）を対話的な検索質問拡張に応用し、Google の関連語推薦と比較した。Google の関連語推薦では、例えば「ピザを作るときに使うチーズは？」という検索を行うための検索クエリとして「recipe pizza cheese」と入力した場合に、チーズの名称 1 件（具体例：goat cheese）が推薦されたが、他は「pizza」との強い関連性を持つ、「crust」や「tomato」などの関連語であった。また、「オーブンできつね色に調理する料理は？」という検索を行うための検索クエリとして「recipe oven golden brown」と入力すると関連語はゼロとなり、「recipe oven golden」で推薦される料理は「fried chicken」の 1 件のみであった。これに対して、提案法では、大規模レシピの検索範囲を「pizza AND cheese」で絞り込み、材料における接尾辞タイプの単語 2-gram(単語 2 つの複合語の後半部分に指定の単語を含むもの)を検索すること

で、「mozzarella cheese」「parmesan cheese」「cheddar cheese」など、ピザ用の材料として使用されている多種多様なチーズを提示できた。また、大規模レシピに含まれる「chocolate chip cookies」「chicken pot pie」「french onion soup」など、任意の粒度(単語 n-gram、この例では単語 3-gram)で、「オーブンできつね色に調理する料理」が提示できた。以上のように提案法では、ユーザの任意の検索クエリに対して、コーパス中の多種多様な関連語を任意の粒度で推薦でき、実応用における有用性が示唆された。

今後の展開としては、複雑なブール検索クエリを簡単に入力し、ユーザが検索要求を詳細に入力できるようにするための直感的な検索インタフェースを検討していく予定である。

５．主な発表論文等

〔学会発表〕（計 6 件）

(1) "料理レシピの文書構造に注目した関連語推薦と対話的な検索質問拡張への応用," 安川美智子, 人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会第 14 回, SIG-AM-14-03, 2016 年 11 月 10 日, 慶応義塾大学日吉キャンパス(神奈川県横浜市)

(2) "表記の微細な差異を含む内容が同一である料理レシピの効率的検索," 安川美智子, じんもんこん 2015 論文集, 2015 年 12 月 19 日, 同志社大学京田辺キャンパス(京都府京田辺市)

(3) "NTCIR-11 料理レシピ検索タスクと群馬大学における研究事例の紹介," 安川美智子, 電子情報通信学会ヒューマンコミュニケーショングループ(HCG) シンポジウム食メディア研究会(CEA) 企画セッション IV 講演・パネル討論大規模料理レシピデータを用いた研究の最前線, 2015 年 12 月 18 日, 富山国際会議場（富山県富山市）

(4) "Data Fusion for Japanese Term and Character N-gram Search," M. Yasukawa, J. S. Culpepper, F. Scholer, Proc. the 20th Australasian Document Computing Symposium (ADCS2015), 2015 年 12 月 9 日, Sydney (Australia)

(5) "Overview of the NTCIR-11 Cooking Recipe Search Task," M. Yasukawa, F. Diaz, G. Druck, N. Tsukada, Proc. NTCIR-11 Workshop Meeting, 2014 年 12 月 9 日～12 月 12 日, 学術総合センター(東京都千代田区)

(6) "Gunma University, Kiryu University, and RMIT University at the NTCIR-11 Cooking Recipe Search Task," M. Yasukawa, H. Ishii, F. Scholer, Proc. NTCIR-11 Workshop Meeting, 2014 年 12 月 9 日～12 月 12 日, 学術総合センター(東京都千代田区)

〔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・大学院理工学府・助教

研究者番号：70361384

(2)研究分担者

無し

(3)連携研究者

無し

(4)研究協力者

F. Scholer (RMIT University)