動詞の提示による動的な検索支援システム

関屋　翔† 堀　憲太郎†† 大石　哲也†††
長谷川　隆三†††† 藤田　博†††† 越村　三幸††††

†九州大学工学部電気情報工学科
††九州大学大学院システム情報科学府知能システム学専攻
†††九州大学大学院システム情報科学研究院情報知能工学部門
††††九州大学大学院システム情報科学研究院情報学部門

1　はじめに

インターネット技術の進歩により，一般家庭からでも容易にWWW(World Wide Web)にアクセスすることができる環境になっている．そのため，日常的に物事を調べる際にも，インターネット上で検索エンジンを用いて情報を得ることが多くなっている．しかし，Web上に存在するページ数は常に増加しており，すべてを把握できないほど膨大な量となっているため，検索エンジンを用いたとしてもユーザの必要とする情報にたどりつけないこともしばしばある．検索結果を絞り込む手段として，複数のキーワードを与えるＡＮＤ検索がある．ユーザの目的に適したキーワードを追加すると，単一のキーワードの時に比べて検索結果を絞り込むことが可能である．しかし，ユーザは検索する際に必ずしも適切なキーワードを思いつくとは限らない．Web検索に慣れていないユーザにとってはなおさらである．

Web検索を支援する技術として，[1]や[2]のようなものがある．前者は，統計的に偏りなく抽出された人(パネル)を対象にURL履歴の収集を行う大域Webアクセスログ(パネルログ)を用いて，与えられた検索語に関連する検索語(関連語)群を表示し，ユーザに検索語を想起させるシステムの提案を行っている．後者は，リンク構造解析による重要文抽出と，自然言語処理を利用した解析手法を提案し，意味関係を抽出することで，Wikipediaから機械可読な概念辞書を自動的に構築することを目指している．どちらの文献も，名詞を抽出することで関連語を生成している．ところが，様々な目的が考えられる名詞がクエリとして与えられた場合，例えば「京都」というクエリであれば，京都に関連する名詞は「観光」や「グルメ」や「賃貸」等，多岐にわたり，ユーザは目的の関連語を探すことが困難である．

そこで本研究では，Web検索に慣れていないユーザが検索語をうまく作成できないという問題を解決するために，ユーザがクエリを入力した際，その語に関連した動詞を提示して検索補助を行うシステムを提案する．提示された動詞を選択してもらうことで，名詞としては表現しきれなかったユーザの目的を把握し，より適切な関連語を提示することができるようになると考えられる．

また，クエリに関連する動詞の抽出にはブログを用いる．ブログは「～へ行った」「～を食べた」等，名詞と動詞を端的に結びつけた簡明な記述が多いからである．

2　システム概要

はじめに，ユーザに元となるクエリを入力してもらう．

次に，関連語を見つけ出すための情報源として，クエリに関連した文書(以降適合文書と呼ぶ)の収集を行う．本研究では，クエリを用いてブログ検索を行い，その結果表示された上位10件のブログ本文を適合文書として用いる．

収集した適合文書を，高速形態素解析システム「MeCab」[1]を利用して形態素解析を行い，名詞と動詞を抽出する．そして適合文書に対し，関連単語抽出アルゴリズム[3]を適用し，クエリに関連する単語を関連度の高い順に抽出する．関連単語抽出アルゴリズムとは，単語間の距離に着目した独自のアルゴリズムである．そしてそれぞれの適合文書から抽出した結果のうち動詞だけを表示する．

表示された動詞の中からユーザが1つを選択する．システムはその結果を受け取り，その動詞が含まれている適合文書の抽出結果の中の名詞を，動詞との関連度が高い順に表示する．ここで表示された名詞が，クエリに関連していて，かつユーザの検索目的に合った単語となり，それらの単語をユーザが選択することにより，よりよい検索結果をクリックのみで絞り込めるようになる．

[1] http://mecab.sourceforge.net/

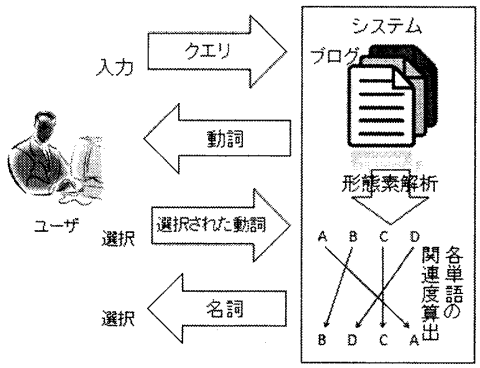

図 1: システムのイメージ

3 関連単語抽出アルゴリズム概要

このアルゴリズムは，あるキーワード群 K とそれに関するテキスト T に現れる単語から，K に関連すると思われる単語を抽出，出力するものである．単語の関連度の算出には，単語間の距離に着目し，それを元に評価を行う．

このアルゴリズムの根幹となっている考え方が，単語間の距離である．具体的には，文章中に出現する単語の順番に注目し，単語 A について A の付近に出現している単語ほど A に関連性があるという考え方である．例えば，

「今日は京都へ観光に行きました．金閣寺に行ったり，八つ橋を食べたりしました.」

という文章にクエリ「京都」で関連単語抽出アルゴリズムを適用すると,「観光，今日，行く，金閣寺，八つ橋，食べる」といったように,「京都」という単語からの距離が近いほうが関連度が高く算出される．

4 実験方法

実験には，2 語で構成されたクエリを用いる．また，提案手法と比較するために以下の手法を用意した．

1. (比較手法) 単純に名詞で抽出して並べた場合
2. (提案手法) 動詞をいったん選んでから名詞を抽出して並べた場合

まずクエリの 1 語目を用いてブログ検索を行う．そこで抽出した内容を用いて，それぞれの手法から関連語候補となる名詞を作成する．その中に 2 語目の名詞が存在するかを調査する．

例えば，クエリを「京都　嵐山」とする．手法 1 の場合,「京都」に関連する単語が選出される (金閣寺，八つ橋，ホテル，バス，等) ものの，その種類は多岐にわたるため，2 語目のクエリが見つかりにくい．手法 2 の場合，動詞をユーザに選んでもらうことで，表示される関連語をよりユーザの目的に合致したものにすることができる (動詞「行く」なら金閣寺，銀閣寺，清水寺，祇園，嵐山，等) ため，2 語目のクエリが見つかりやすい．クエリが「京都　嵐山」の場合，2 語目の「嵐山」が，手法 1 には存在しておらず，手法 2 には存在しているので，手法 2 のほうがうまく検索結果を絞り込めているということになる．

また，検索に動詞を挟んでいるため，動詞と相性が悪い関連単語も存在する．そのため特にどんなクエリについて効果が高いかを調査する．

さらに，どの程度このシステムがユーザの検索目的に合った結果を返すのかの統計を取るため，何人かに実際にシステムを使ってもらってアンケートを取る．

5 おわりに

実験結果としては，検索に動詞を挟んだことにより，名詞だけで検索した場合と違って，的確にユーザの検索目的に合った関連単語を表示し，他の Web 検索支援システムと同等かそれ以上の結果を返ってくると期待される．ただ，この実験では，ユーザが毎回動詞を選択して，結果を見て判断しなければならないため，ユーザに負担がかかることが問題である．考えられる対策としては，動詞の選択がユーザの負担とならないようなインターフェースの作成があげられる．今後の予定としては，問題点を解決しつつ，よりよい検索結果を導き出すために他のアルゴリズムとの比較を考えている．

謝辞

本研究は科研費 (21500102) の助成を受けたものである．

参考文献

[1] 大塚　真吾，喜連川　優，“大規模アクセスログを用いた検索支援システム”　DEWS2006,1B-o2

[2] 中山　浩太郎，原　隆浩，西尾　章治郎，“自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築に関する一手法”DEWS2008,A3-2

[3] 大石 哲也, 倉元 俊介, 峯 恒憲, 長谷川 隆三, 藤田 博, 越村 三幸: ”関連単語抽出アルゴリズムを用いた Web 検索クエリの生成” 電子情報通信学会 情報・システムソサイエティ和文論文誌データ工学特集号 Vol.J92-D,No.3,Mar. 2009.