Webからの関連用語の自動獲得

星 正人 (210083)

北陸先端科学技術大学院大学 情報科学研究科

2004年2月13日

キーワード: ポータルサイト, 用語集, 用語抽出, 検索エンジン, Web.

Web上で利用可能な膨大な文書の中から必要な情報を探し出す手段として、ポータルサイトの利用が考えられる。ポータルサイトとは、あるテーマに関する基礎的な情報や、そのテーマに関連する用語集、リンク集などのコンテンツからなるサイトのことである。しかし、自分が知りたいと思っているテーマに関するポータルサイトが存在するとは限らない。したがって、テーマを入力するだけでそのテーマに関するポータルサイトを自動的に作成することができれば便利である。本研究ではポータルサイトのコンテンツのうち、用語集に注目する。さらに用語集の見出し語を抽出することに焦点を当て、用語集の見出しとなるようなポータルサイトのテーマに対する関連用語を自動的に獲得することを目的とする。

関連用語をWeb上から自動的に獲得するという処理は、大きく分けると「関連文書収集」、「用語候補抽出」、「関連度計算」の3段階の処理からなる。「関連文書収集」は、ポータルサイトのテーマに関連する文書を集めることであり、検索エンジンを用いてテーマを含む文書を収集する。「用語候補抽出」は、関連文書の中で特に良く現れると思われる用語を選び出す処理である。最後の「関連度計算」は、選び出された用語候補の中からテーマと関連性が深いと思われるものを出力する処理である。この処理は、検索エンジンを用い、テーマと用語候補のWeb上での共起確率を求め、関連度とする。具体的には、「用語候補抽出」によって出力された用語が単体で現れる文書数と、テーマと共に現れる文書数の比をとることで、その用語がどれだけテーマに特有の用語であるかということを測る。

3段階の処理のうち、2番目の「用語候補抽出」処理は、関連研究である専門用語抽出などで研究されている分野であり、本研究でもそれらを参考にしている。具体的には、名詞句を特定し、それらのスコアを計算することで用語候補にふさわしい名詞句を選ぶ処理である。そして、この段階の処理の違いにより最終的な出力が大きく変わるため、本研究ではこの段階の処理に関して3種類の手法を実装し、それぞれから得られた結果を比較し、考察した。3つの手法とは、「出現頻度に基づく手法」、「造語能力に基づく手法」、「組み合わせによる手法」である。

「出現頻度に基づく手法」は、名詞句の出現回数を数え、同じ要素数からなる名詞句の

出現回数の平均で割ることによって、その名詞句が他の名詞句と比べてどれだけ多く現れているかということを求める。

「造語能力に基づく手法」は、専門用語抽出分野でよい成績を出しているとされる評価基準を用いた手法である。その評価基準とは、複合名詞の構成要素となりやすい単語ほど「造語能力」が高いとするものである。そして「造語能力」の高い単語によって構成されている用語が、その文書にとって重要であるという考えに基づき、スコアを計算する。

３つめの手法、「組み合わせによる手法」は、「出現頻度に基づく手法」と「造語能力による手法」の出力を組み合わせるものである。この段階の処理だけをみると非常に単純なものだが、「造語能力に基づく手法」、「出現頻度に基づく手法」それぞれの出力には異なる特徴があり、さらに最終的な処理である「関連度計算」では前段階までのスコアに依存せずに関連度を求めるため、うまくいけば両手法のよい結果のみを最終的な出力とすることができる。

実験結果から、「用語候補抽出」段階の手法の違いにより、最終的な出力にもかなりの違いがみられた。単純に正解数だけを比較すると「出現頻度に基づく手法」が最も多く、「造語能力による手法」が最も少なかった。しかし、各手法が候補とする用語には特徴があり、正解数だけでは手法の優劣はつけられないことがわかった。具体的には、「出現頻度に基づく手法」は人名や製品の型番などを多く取り出す傾向があり、「造語能力に基づく手法」はテーマを部分文字列として含むものを関連用語候補として挙げやすいという傾向がみられた。後者は、本研究における関連度計算の手法にも一因がある。「造語能力に基づく手法」が候補としやすいテーマを部分文字列とする用語は、関連度の定義から、次の「関連度計算」の段階で決まって高い評価値をもつ。したがって、テーマを部分文字列とする用語が関連用語としてふさわしくないような分野に対しては、「造語能力による手法」の最終的な出力は不正解となる用語が多い。また、「関連度計算」段階の処理は共通なので、テーマを部分文字列とする用語に関する問題は「組み合わせによる手法」にも影響が現れている。しかし、この問題さえ克服できれば、「造語能力による手法」の正解率も向上し、「組み合わせによる手法」が２つの手法の異なる性質を持つ用語候補の両方を抽出できるようになると思われる。