ポータルサイト自動作成のための用語説明獲得

菅井 俊介 (210046)

北陸先端科学技術大学院大学 情報科学研究科

2003 年 2 月 13 日

キーワード: ポータルサイト, 用語説明獲得, Web 探索支援, 情報抽出.

ポータルサイトとは、ユーザーがあるテーマに関連した事柄を調べるときに最初に訪れることを目的に作られ、そのテーマに関する様々な情報を集約した Web サイトである。その主な構成要素は、検索エンジン、リンク集、用語集などである。現在、ポータルサイトの多くは人手により作成されているため、多くの労力を要する。そのため、与えられたテーマに沿ったポータルサイトを自動作成する技術の確立が期待される。

そこで、本研究はポータルサイトのうち用語集の作成を目的とし、用語の説明文を自動的に獲得する。用語の説明は Web 文書から動的に獲得する。なぜなら Web 上には既存の辞典や辞書には掲載されていない専門性の高い用語や、造語、新語等の説明文や定義文の存在が期待されるからである。また、Web は辞書や辞典に比べ、頻繁に更新されるという特徴を持っているため利用価値が高い。一般に、ユーザーが知りたい用語を調べたり、用語集を作る場合、検索エンジンで用語説明を得ようとすることがある。しかし、膨大な検索結果の中から用語の説明が記述されているページを見つけることは困難である。これに対し、本研究は Web ページから用語説明を自動的に獲得し、ユーザーに提示することを目指す。

また、同じ語でも分野により意味（語義）が異なる場合がある。そのため、目標となるポータルサイトのテーマに合致する語義を有する説明文を自動的に選別し、ユーザーにスコアの高い順に提示するシステムを提案する。具体的には、ポータルサイトのテーマを表わす分野に特有の名詞を多く含む用語説明文を見つけることにより、ポータルサイトのテーマと関連のない語義に対応した説明文を排除する。また、ポータルサイトのテーマは動的に変化するため、分野コーパスも動的に獲得し、用語説明の関連度を求める点に本研究の特徴がある。

本システムは、ポータルサイトのテーマと用語を入力とし、その用語の説明を Web から獲得する。これは以下の手続きから成る。

1. 用語の説明があると思われる候補ページの取得

用語集に掲載する用語 X を与える。システムは検索クエリを「X とは」、「X は」に分け、これを検索エンジン goo に入力し、Web ページを取得する。網羅性を重視し、最大 2000 ページの Web ページを収集する。同時に文字コードを EUC に統一する。

2. 候補ページからの不必要なタグの除去、ページ整形

文字装飾タグやスタイルタグ等のリストを用意し、１で取得した候補ページから除去する。次に、HTML タグとそれ以外の文章が一行毎に交互に現れるように、タグの前後に改行を入れ、タグ同士が続く場合は間に空行を入れ、ページ最後に終端記号を付与する。

3. HTML タグを利用した用語説明箇所の抽出

HTML タグを利用し、２で整形した Web ページから用語説明の書かれている箇所を抽出する。ここでは、用語が見出し語となっていて、その後に用語説明が書かれている場合と、用語が文自体に内包されている場合を考慮し、それぞれについて用語説明箇所抽出アルゴリズムを作成した。また、形態素解析を行い、後続する文が連体詞や接続詞で始まる場合、その文も用語説明として抽出する。また、抽出の際、用語説明箇所を複数の文型パターンに分類する。

4. ふさわしくない用語説明の除去

特定の文型パターン、長すぎる説明文、用語が複合語の一部となっている場合の 3 つについて、ふさわしくない用語説明を除去する。

5. 用語説明のスコア付け

ポータルサイトのテーマとの関連の深さを表わすスコアの付与。用語説明 E へのスコア付けは以下の式で行う。

$$Score(E) = \frac{1}{|E|} \sum_{n \in E} score(n) \tag{1}$$

ここで E は用語説明文、n は説明文中の名詞、$score(n)$ は名詞 n のスコア（テーマとの関連度）を表わす。$score(n)$ を求めるため、テーマをクエリとして検索を行い、テーマに関する Web 文書群を収集する。$score(n)$ として、単語の頻度や文書頻度をもとにした TF-IDF と、文書頻度をもとにした RDF の 2 通りの定義を考えた。いずれにせよ、そのテーマに特有の名詞に対し高いスコアを与えることを目的とする。最終的に、用語説明を $score(E)$ の降順に並べ、その上位の用語説明を出力する。

本システムの有効性を測るため、評価実験を行った。実験では、求めたい用語に対し、語義が１つの用語、複数の語義をもつ用語の 2 種類について用語説明を獲得した。また、

それぞれの用語に対し、TF-IDF によるスコア付けと RDF によるスコア付けの両方を試した。

得られた用語説明のスコア上位 10 件について、正しい用語説明が得られたかどうかを評価した。実験の結果、ほぼ全ての用語で、スコア上位 10 件にふさわしい用語説明が含まれていた。また、スコア付けは TF-IDF より RDF のほうが優れていることがわかった。語義を複数持つ用語の場合、語義が 1 つの用語に比べ、ふさわしい用語説明を抽出することが困難であった。これは多くの語義に対応した説明が多数抽出されたことや、用語説明とテーマとの関連度を測る手法がうまく働かなかったことが原因として考えられる。これを解決するため、本システムの手続き 3 で得られた文型パターンをスコア付けに反映させる方法が考えられる。