4ZE-02

WWW情報空間のAreaViewシステムにおけるコアページの構造化 *

松崎 明† 大澤 幸生‡ 伊庭 斉志† 石塚 満†

† 東京大学工学部電子工学科 ‡ 筑波大学経営システム学専攻

1 はじめに

WWW(World Wide Web) は Web ページを単位として、関連する情報同士がハイパーリンクで結ばれた巨大な情報空間を形成している。この WWW 情報空間の大きな特徴は、誰もが容易に情報を発信できるというオープンな性格にあるが、この事は情報空間全体の拡大を助ける事になった反面、秩序や構造といったものを皆無にし、必要とする情報を見つけ出す事を困難にしてしまっている。

我々の AreaView システム [1] の目的は、WWW 情報空間をリンク構造を用いて弱い枠組みで構造化を図り、ユーザにとって必要な情報の検索や、閲覧する Web ページ間の関連を視覚的に支援することにある。このシステムは Web ページ内に埋め込まれたリンクを元に、個々のページ間の関連性やページの重要性を特定し、これらに単語解析を合わせてページの意味解析を行った上で、弱い構造化を実現する。

本研究においては、弱い構造化の手法の一つとして、重要性の高いページを従来のディレクトリ型検索エンジンのように分野ごとに分類し、ツリー構造化することを目的とし、これらを自動化することで、雑然とした状態の WWW 情報空間を整理する事を可能とする。

2 Webページの内容理解

WWW 情報空間のアクセス支援を行う研究において共通の課題となるのは、Web ページの多様性と、WWW 情報空間の特徴である複雑に絡み合ったリンク構造の処理である。更に、Web ページを記述する言語である HTML は、ユーザに対する視覚化を第一において開発された言語であり、コンピュータが情報の意味理解を容易に行えるようには設計されていない。しかしユーザの情報アクセスを支援する目的においては、個々のページの意味理解こそが最も重要となる。ページの内容を把握しなければ Web ページを弱構造化することは不可能である。

コアページ

AreaView システムにおいて弱構造化を行う際に WWW 空間上に存在する多くの Web ページの中から抽出される、意味があり利用価値の高いと考えられるページをコアページと呼ぶ。この抽出法 [1] では、主にリンク構造を用いて構造化を計るため、事前に領域知識を必要とせず、したがって日々拡大する WWW 情報空間において多くの領域に適用を図る場合、大きな優位性を持っている。本研究においてツリー構造を用いた分類を行う際、その分類対象となるのは、この抽出されたコアページ群である。

単語抽出

ページの意味を理解する上で最も基本的な作業は、単語を抽出することである。しかし WWW 空間に存在する Web ページは世界中の人々によって様々な言語で書かれている。このため実際の単語抽出においては、Web ページの言語による構造的な特徴の違いから複数の手法を使い分ける必要があるが、本研究においては最も比率の高い英語のページのみを対象とする。具体的には、空白などの区切り文字で区切られた単語を抽出後、stemmer[2] を用いて語幹を得る。

単語の重み計算

ページの意味解析を行うという観点から、ページ内の各単語の重要性を計るという意味で単語の重み

---

* Organization of Core Pages in AreaView System of WWW Information Space

† Akira Matsuzaki, Yukio Ohsawa, Hitosi Iba, Mitsuru Ishizuka

‡ Yukio Ohsawa

† University of Tokyo,7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

e-mail : matsu@miv.t.u-tokyo.ac.jp

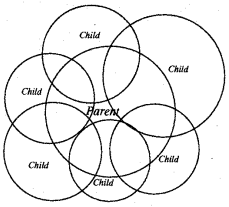

図 1: 親ページと子ページの関係

を計算する必要がある。これにはtf-idf法 [3] を用い、ページ内での単語の出現頻度と、そのページが属する分野全体中のその単語を含むページの比率の逆数の対数との積として表す。単語のページ内での出現頻度計算においては、HTML タグによる重み付けもあわせて行う。

リンク情報の利用

WWW 空間の大きな特徴となっているハイパーリンクもまた、意味解析の上では重要である。そこで、ハイパーリンクのアンカー文字列およびその周辺の単語を、リンクの参照先ページに対するページ作者の定義付けとみなして解析に利用する。

3 弱構造化と分類

本研究における弱構造化の指針となるのは、分野ごとに分類されたツリー構造を実現する事である。このことはつまり、大まかな内容を扱ったページを親に、細かい内容を扱ったページを子にして、これらを細分化するという事である。なお、ここで言うページとは、同一サイト内のある深さまでのリンク先ページを含んだものである。

コアページのレベルの概念

コアページの中にも、広く浅い内容のページと狭く深い内容のページとが存在する。そこでコアページのレベルを、広く浅い内容のページほどレベルが高く、狭く深い内容のページほどレベルが低いと定義した。また親子関係に関しては、ルートを除く各ページはおのおの親ページを一つだけ持つことが許され、同じ親を持つ子ページ群をグループと呼ぶこととした。(図 1)

レベル分けとグループ化

レベル分けのアルゴリズムは

「上位レベルのページ内に現れた専門用語をできるだけ多く包含し、かつページ数が最小となるような子ページの組を求める」

という最適化問題を解くことにある。レベルの定義とこのアルゴリズムから最上位のページは、その分野における専門用語をもっとも多く含むページとなる。この解として得られた各ページは、自分の親ページの詳細な内容のみならず、親ページに現れなかった内容も含む可能性があるため、全体として見たときに、漏れる単語の数を少なくする事ができる。実際の計算では組合せ爆発が起こらないようにするために、次の近似解法を用いる。まず、親ページ内に存在する専門用語を一定数以上持つページのみを抽出し、そのようなページの総数がある値以下になるまで、単語数の閾値を変化させ、そうして得られたある一定数のページのすべての組について式 (1) を計算し、その値が最大となる解を得る。

$$\frac{\text{子でカバーできた専門用語}}{\text{親の専門用語}} \times \frac{\text{エリアの全ページ数}}{\text{カバーに要したページ数}} \quad (1)$$

4 視覚化機能

AreaView システムの目的には、弱構造化だけではなく、ユーザに対する視覚化支援もある。実装形態としては、Web ページを閲覧するという観点から、従来のサーチエンジン同様に CGI を用いる。また、表示する情報量も重要となってくるが、基本的にはツリー構造と、グループに共通する単語のみを表示する。ただし、ユーザの要求によっては実際のリンク構造を付加することなども可能とする。

5 むすび

本研究では、非均質で巨大な WWW 情報空間のある分野を解析し、それらを視覚化する AreaView システムの更なる向上を目指す。このシステムは、主に学術分野を対象として、特定の分野で利用価値が高く重要と考えられるコアページをその内容とリンク構造という 2 つの情報からレベル分けしグループ化して、全体としてツリー構造を持たせ、ユーザに分かりやすくかつ柔軟に提示する事が可能である。これは、オープンな性格ゆえに日々複雑化する WWW 情報空間において、現在のサーチエンジンとは異なった角度からユーザのブラウジングを支援するものとして有用である。

参考文献

[1] 福島, 石塚: WWW 情報空間のリンク構造を用いた弱い構造化. 信学技報, AI98-93(1999.3).

[2] M. Porter: An Algorithm for Suffix Stripping. Program, Vol.14(3), pp.130-137, 1980.

[3] B.Salton: Term Weighting Approaches In Automatic Text retrieval. Information Processing & Management, Vol.24, No.5, pp.515-523, 1988.