4ZE-01　WWW 空間の弱組織化とエリアビューにおけるコアページ抽出法の改良*

猪股 健太郎　大澤 幸生　伊庭 斉志　石塚 満

東京大学大学院工学系研究科電子情報工学専攻

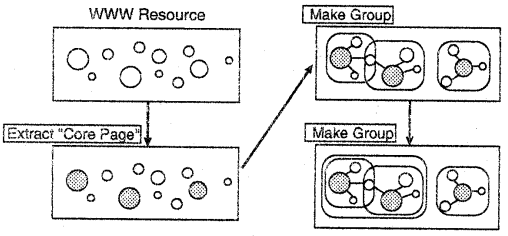

図 1: 弱い構造化

1　はじめに

我々の研究室では、コアページ抽出という方法により、WWW 情報空間を弱い枠組みで構造化し、それによって実現するエリアビューという機能を開発した。本稿では、コアページ抽出の方法を改良することで精度を上げる手法を報告する。

2　弱い構造化

WWW 情報空間は WWW ページ間がリンクで結ばれたネットワーク構造をなしている。多数のリンクにより参照されている WWW ページは、その参照元のページの作成者が自分以外である場合、多くの人に既知でそして意味のあるページであると考えられる。そこで、我々はその意味のあるページ（コアページ）を抽出し、コアページを基に WWW 情報空間上のページをグループ化する手法を開発した。一つの WWW ページが複数のグループに属することもあるため、この手法を WWW 情報空間の弱い構造化と呼ぶ (図 1)。

コアページの抽出

各 WWW ページは、参照リンク数に応じて重みを付け、重みの大きい WWW ページをコアページとして抽出する。具体的には、他のサーバへのリンク（外部リンク）が大きくなるようにリンクに重みをつけ、各 WWW ページは参照されているリンクの重みの和をそのページの重みとする。

グループ化

まず、抽出されたコアページが参照しているページと、コアページから参照されているページでグループを形成する。次に、どのグループにも属さないページは、関連度の高いグループに分類する。関連度は各 WWW ページ内から抽出した単語の特徴ベクトルを用いて算出される。

グループ間の関連度

グループ間の関連度は、WWW ページとグループとの間の関連度と同様に、単語の特徴ベクトルを用いて計算される。グループの特徴ベクトルはそのグループに属する各 WWW ページの特徴ベクトルの和で定義される。

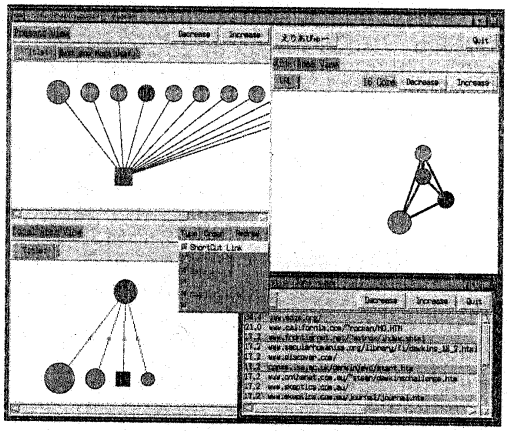

図 2: システムの概観図

*Improvement of extracting core pages on Weak organization and Area-view of WWW space.

Kentaro Inomata, Yukio Osawa, Hitoshi Iba, Mitsuru Ishizuka
University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan
inomata@miv.t.u-tokyo.ac.jp

3 コアページ抽出法の改良

エリアビューシステムにおけるコアページの抽出法は、リンク情報を解析・分類しそれに重みづけをするものであった。しかし、リンクの分類が粗かったため、適切な重み付けができていたとは言い難かった。

そこで、コアページ抽出の制度を上げるために、ページが表している内容を考慮したリンクの重みづけを行う。具体的には、アンカータグの近くにある文に注目し、ユーザが着目している分野における重要単語が現れている場合には、そのリンクの重みを大きくする。

3.1 単語の重要度の決定

単語の重要度の決定には $TF \cdot IDF$ 法を応用する。ユーザが着目している分野の文書群 (sub) と、それを含む文書群 (sup) を考える。$TF \cdot IDF$ 法では、ある語 t がある文書 d を弁別する能力を表すが、今回の場合は、ユーザが着目する文書群を弁別する能力を表す必要がある。したがって、TF をユーザが着目する分野での文書頻度、IDF をそれを含む文書群での IDF とする。

$$
\begin{aligned}
TF' &= \frac{\text{単語が出現する文書数 (sub)}}{\text{総文書数 (sub)}} \\
IDF' &= \log \frac{\text{総文書数 (sup)}}{\text{単語が出現する文書数 (sup)}} \\
\text{重要度} &= TF' \cdot IDF'
\end{aligned}
$$

3.2 リンクの重み

アンカータグの前後の文とリンク先のタイトルを取り出し、その中で使われている単語の重要度を調べ、最大のものをそのリンクの重要度とする。これは、アンカータグに挟まれているテキストだけでは、リンク先の URL や、「戻る」など、情報の少ないものも多いためである。

そして、エリアビューシステムでリンクの重み付けを行う際、リンクの種類別による重みと今回の重要度の積をとり、新しい重みとした。

4 実験

今回の実験では、ユーザが興味を持った分野を人工知能についてとした。Yahoo!の当該カテゴリ中のサイトの HTML 文書と、Yahoo!における上位カテゴリ中のサイトの HTML 文書を収集した。

リンクの種類別による重み

同様なページへのリンク	1
ショートカット	1
詳細情報の提示	0.01
語彙説明	0.01
リンク集へのリンク	0.01
その他外部リンク	0.5

5 おわりに

今回の実験では、テストデータとして、Yahoo! に掲載されているサイトを使用したが、いったん単語の重要度のデータベースを作成してしまえば、そのあとに解析する HTML 文書は必ずしも同じものである必要はないと思われる。よって今後は、より多くの HTML 文書を収集し、解析してみたい。

また、このような「コアページ」という新しい概念は、今の段階では計算機によって評価できないため、評価部分は人手に頼らざるを得ない。膨大な数の HTML 文書を一つ一つ人手で評価していく作業はたいへんコストのかかる作業であるので、コアページの表か手法をも新たに考案する必要がある。

また、この成果を生かし、「ユーザが希望する分野を入力すると、その分野のコアページが数十個ほど抽出され、ユーザはそれを読み進めるだけで、その分野の概観を得ることができる」というシステムを完成させるのが今後の目標である。

参考文献

[1] http://www.yahoo.com/

[2] 福島, 小野田, 石塚 : WWW 情報空間におけるハイパーリンクの意味理解とリンク構造の視覚化, 人工知能学会全国大会 '98.

[3] 福島, 小野田, 石塚 : WWW情報空間におけるコアページの抽出と弱い構造化, 情報処理学会第 57 回全国大会, No.3-398–399, 1998.

[4] 福島, 石塚 : WWW 情報空間の弱い構造化とエリアビュー機能, 情報処理学会第 58 回全国大会, 1999.

[5] 福島, 石塚 : WWW 情報空間のリンク構造を用いた弱い構造化. 電子情報通信学会研究報告会人工知能と知識処理, 1999.