WWW情報空間における特徴ベクトルを用いたリンクの分類

2R－8

加賀屋 潤　福島 伸一　石塚 満　伊庭 斉志
東京大学工学部電子情報学科 石塚研究室

1. まえがき

近年、インターネットの日常生活・社会への浸透は目を見張るものがある。しかし、インターネット上における情報発信の容易さのため、爆発的な勢いで情報量も増しており、情報過多という問題も生じている。今後さらに情報が爆発的に増えることを考えると、時間を有効に利用するためにも情報の取捨選択を行うことは重要であるといえる。

当研究室では、WWW情報空間における情報収集・ブラウジングを効率化することを目的とし、ハイパーリンクの意味を解析する研究を行っている[1]。そこで、本稿ではユーザのブラウジングを支援するためにWWW情報空間の基幹をなすWWWページとページ間のリンクを特徴ベクトルを用いて分析する手法を提案する。これにより、閲覧しているページの周辺情報をユーザに提示することが可能となる。システム概要を図1に示す。

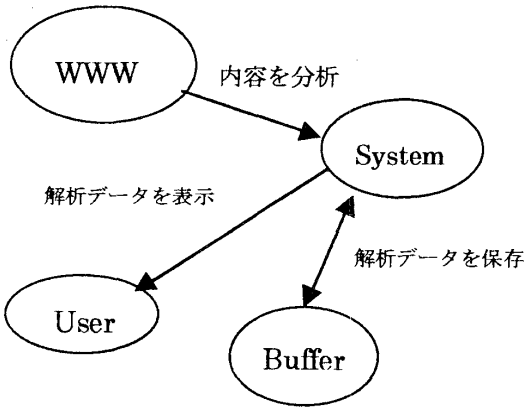

図 1　システム概要

2．WWWページの解析

2.1 HTMLによる文章分け・リンクの情報

WWWページ (html,htm file)は、html tagと呼ばれるものを使用して構成されている。html tagにはそれぞれ意味があり、それを利用してページの内容を理解し、解析する。

例えば、<head> - </head>で囲まれた部分は、ヘッダ部分であり、<body> - </body>で囲まれた部分はページの本文である。

タグを使用して、まずはページの意味解析に必要な情報のみ残す。その後、ページを段落に応じて細分化する。細分化することによって、複数の意味段落から構成されているページの内容をより詳しく分析することが可能であるからである。もちろん、この分割も役割に応じたタグを用いて行う。この一連の作業により、各ページは1個以上の段落に分割される。各段落について、以下に述べる「特徴ベクトル」を用いて分析を行う。

なお、ハイパーリンクについてのタグについては固定した形式があるので、各リンクに関してそこから得られる情報を確保する。

2.2 特徴ベクトルによる解析

特徴ベクトルとは、文書の全体的な傾向(つまり、キーワード)を数値化して表示したものである。

特徴ベクトルを使用するには、文書に含まれる各単語の重要度を導き出すことが不可欠であるが、本研究では、語の重要度を表す値として tf-idf 値

Link-Classification using "feature vector" in WWW Information Space
Jun Kagaya, Shinichi Fukushima, Mitsuru Ishizuka, Hitoshi Iba
University of Tokyo, 7-3-1 Hongo, Byukyo-ku, Tokyo 113,Japan
Jun@miv.t.u-tokyo.ac.jp

を採用した。この tf-idf 値により各文書の全体的な傾向を示した「特徴ベクトル」が導出でき、さらにこれをページ全体に応用することにより、ページ全体としての姿を見ることもできる。

これをもとにして、リンク元のページの段落の特徴ベクトルと、リンク先のページの特徴ベクトルにより近似度(関連度)が求まり、ページの周辺情報がより的確に示される。

3.周辺情報の表示

2．で示したような方法でページの分析を行った後に、ページの周辺情報を視覚的に表現する。

周辺情報の表示法には多くの方法が提唱されているが、将来性なども考慮し、なるべく小さな形にまとまるような方法を採用することにした。

その方法とは、図2のように、中心ページを円の中心に置き、中心ページからの最短リンク回数に応じて周辺ページを配置する方法である。ページの大きさ、内容によって色や大きさ・形を変化させ視覚性の高いインターフェイスを実現することを目指している。さらに、ページ間のリンクについても、関連度をリンク上に表示し、リンクの太さでその度合いを示す形になっている。

さらに、図2上のページアイコンやリンクをクリックすると、図3のようなページ情報やリンクの情報を表示する機能も有している。

このような方法をとれば、ユーザは、関連度の高いページを容易に発見することが可能となるのである。

4．むすび

本研究において、意味ベクトルを用いたページ分析システムを作成したが、結果的に、意味ベクトルによる分析に一定の成果が得られ、ページの周辺情報を図2・3のようなインターフェイスで表示することが可能になった。

本稿執筆現在の時点で、代表的なタイプの WWW ページを選択して、主観的・客観的な視点から分析した結果と、本システムを使用して分析したデータとの一致性を見ている。

本研究を今後の情報選択においてさらに役立つものにしていきたい。

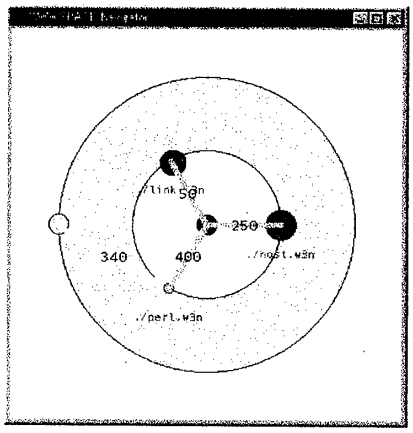

図2　周辺情報の表示

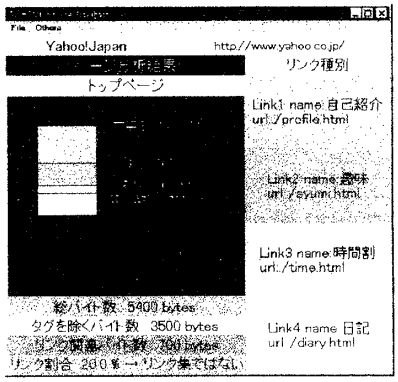

図3　ページ情報の表示

6．参考文献

[1]小野田, 土肥, 石塚 ``WWW ハイパーリンクの意味による分類とノードリンク構造の提示", 第56回情処全大, No.1Z-03, 1998.3.