2001年度日本オペレーションズ・リサーチ学会
秋季研究発表会

北海道「遊」産業情報におけるWebマイニング

北海道大学　＊金城　伊智子　KINJO Ichiko
北海道大学　長尾　光悦　NAGAO Mitsuyoshi
北海道情報大学　斎藤　一　SAITO Hajime
1004631　北海道大学　大内　東　OHUCHI Azuma

1．はじめに

現在，多様なメディアにおいて北海道「遊」産業情報が提供されている．代表的なメディアとしては，雑誌，TV，WWWなどが挙げられるが，この中でも特にWWWはその情報量，時間や場所に依存しないといった利点から他のメディアと比較して，より有効な情報の提供を行うことができると考えられる．しかしながら，有効な情報の提供を行う為には，第一に，北海道「遊」産業情報そのものを明確にする必要がある．

本研究では，WWW上のデータをテキストマイニング技術[1]により分析することによって，北海道「遊」産業情報の明確化を行う．また，WWW上における多量の北海道「遊」産業情報を収集し，効果的な情報の提供を行うためのWebマイニングシステムを提案する．

2．北海道「遊」産業情報の分析

2．1　タグ情報

北海道「遊」産業情報を明確化する為には，情報を提供しているWebサイトを分析することが必要である．

現在WWW上で公開されているWebサイトの多くは，HTML言語によって記述されている．このHTML言語はタグと呼ばれる記述方式に基づき，テキストに対して視覚的構造を付与するものである．ここで，WWW上の北海道「遊」産業情報を分析するためにはWebサイトが示す内容，すなわち，Webサイトにおいて意味的構造を獲得する必要がある．したがって，HTML言語におけるタグ情報を分析することによりWebサイトにおける意味的構造を獲得し，視覚的構造との関係を調査，分析することによって，そこに現れる北海道「遊」産業情報の特徴を抽出することができ，WWW上の情報に基づく北海道「遊」産業情報を明確化することができると考える．

2．2　タグ情報の分析

本研究ではWebサイトにおけるタグ情報を収集し，それら収集したタグ情報とWebサイトの示す内容の関係を分析することにより，タグ情報に基づくWebサイトの内容把握に対する妥当性の検討を行う．まず，WWWにおける情報収集方法として一般的であるサーチエンジンを用いて北海道「遊」産業情報を収集する．検索結果から得られたWebサイトにおいて，提供者の意図が反映されやすく，Webサイトの内容と関連が強いと考えられる<TITLE>，<H1>，<HREF>，<COLOR>という4つのタグ情報の分析を行う．これらのタグ情報に基づくテキストをWebサイトから抽出し，語句単位に分解する．そして，分解された各語句の出現頻度を調べることによりデータ全体における各語句の出現傾向の分析を行う．

<TITLE>タグに囲まれるテキストは，検索に用いたキーワードと一致する語句を多く含み，その他の語句を含むことがほとんどないことからそのWebサイトの概要的な内容を表していると考えられる．一方，<H1>，<HREF>，<COLOR>タグのそれぞれに囲まれるテキストには，検索に用いたキーワード以外にも観光と関係のある語句が多く出現するため，その Webサイトの具体

的な内容を示していると考えられる．このような分析を行うことにより，タグ情報に基づき多量のノイズを含む Web サイト情報の全体を調査することなくその内容についての把握，すなわち，視覚的構造に基づき意味的な構造を定義することが可能であることを明らかにする．

3．Web マイニングシステム

現在，WWW 上には多種多様な北海道「遊」産業情報が分散的に存在している．本研究では，この北海道「遊」産業情報を網羅的に収集し，利用者が必要とする情報を効率的に収集可能となるよう，いくつかのクラスタに Web サイトを分類し，情報提供を行うことが可能な Web マイニングシステムを提案する．

本システムの構成とデータの流れを図 1 に，インターフェースを図 2 に示す．図 1 に示されるように，本システムはメタサーチエンジンモジュールとクラスタリングモジュールから構成されている．メタサーチエンジンモジュールでは，現在の WWW 上に存在するサーチエンジン群に対して検索要求を送信し，検索結果の URL リストを受け取る．受け取った URL リストに重複サイトがある場合には削除する．その後，URL リストの Web サイトの HTML テキスト情報をクラスタリングモジュールへ送信する．クラスタリングモジュールでは，メタサーチエンジンモジュールから送られた HTML テキストに基づき的確に Web サイトが表現する内容を得るために HTML 言語におけるタグ情報の抽出を行う．このタグ情報に基づくテキストに対して形態素解析が適用され，テキスト情報における名詞頻度ベクトルが生成される．生成されたベクトルに基づき Web サイト間の類似度が算出される．算出された類似度に基づいて Web サイトのクラスタリングが行われる．このクラスタリング結果のための新たな Web サイトが生成される．

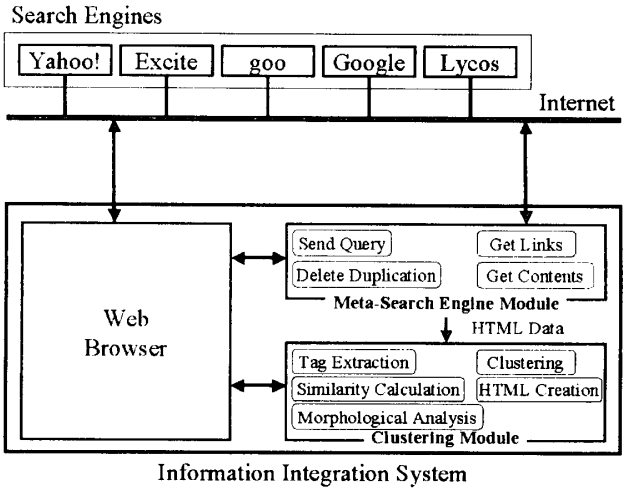

図 1：システム構成

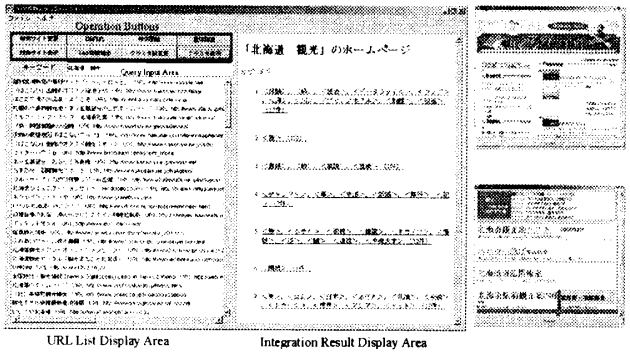

図 2：インターフェース

利用者は生成されたクラスタに基づき情報の取捨選択が可能であり，効率的かつ効果的な北海道「遊」産業情報の提供が可能となる．

4．おわりに

本研究では，北海道「遊」産業情報の明確化を目的とし，Web サイトにおけるタグ情報を収集し，その情報の分析を行った．また，Web サイトを収集し，いくつかのクラスタに分類することによって効果的な情報提供を行うことが可能な Web マイニングシステムを提案した．

参考文献

[1]那須川哲哉，河野浩之，有村博紀：テキストマイニング基盤技術，人工知能学会誌，Vol.16, No.2, pp.201-211 (2001)．