

北海道観光情報のための Web データ分析に関する研究

金城 伊智子, 大内 東
北海道大学 大学院工学研究科

現在、多様なメディアにおいて北海道の観光情報が提供されている。中でも特に、WWW はその情報量、情報の最新性といった利点から他のメディアと比較して、より有効な観光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、WWW 上で公開されている情報は必要とされない情報、すなわち多量のノイズを含む情報である。更に、WWW 上の情報を記述するための HTML 言語は視覚的な構造は定義できるものの、意味的な構造は定義することができない。そこで本研究では、このような問題に対して HTML 言語によって記述される情報の視覚的な構造と意味的な構造の関係に関する調査、分析を行う。

Web data analysis for Hokkaido tourism information

Ichiko Kinjo, Azuma Ohuchi
Graduate School of Engineering, Hokkaido University

Recently, Hokkaido tourism information is offered in various media. In particular WWW can offer more effective tourism information compared with other media. However, information shown on WWW includes information that is not needed, i.e. information to have a large quantity of noise. Furthermore, the HTML language to describe information on WWW, can define structure of visual, but cannot define structure of meaning. So, in this study, we do analysis about relation of structure of visual and structure of meaning.

1. はじめに

観光がすでに基幹産業となっている北海道では、これから北海道への来訪客数の増大に結びつくような観光情報の提供を行う必要がある。

現在、多様なメディアにおいて北海道の観光情報が提供されている。中でも特に、WWWはその情報量、情報の最新性といった利点から他のメディアと比較して、より有効な観光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、WWWにおける情報は必要とされない情報、例えば広告等の情報を多く含む。すなわち、多量のノイズを含む情報である。

したがって、Webページに含まれる全ての情報を分析した場合、そのノイズにより悪影響が出ると考えられる。更に、WWW上の情報を記述するためのHTML言語は視覚的な構造は定義できるものの意味的な構造は定義することができない。そこで、本研究ではこのような問題に対してHTML言語により記述される情報の視覚的な構造と意味的な構造の関係を調査、分析することによって、北海道の観光情報の特徴の抽出を行う。

2. 北海道観光情報

世界観光機構は、観光に対する需要と供給の動向や他の産業分野への影響などを調査、分析し、今世紀の初頭は世界中で10億人以上が旅行を楽しむ「大観光時代」になると予測している。観光がすでに基幹産業となっている北海道では、このような分析結果から北海道への来訪客数の増大に結びつくような観光情報の提供を行っていくなければならないといえる。

現在、北海道の観光情報が様々なメディアにおいて提供されている。その代表的なメディア

としては雑誌、TV、WWW等が挙げられる。

以下に、代表的なメディアとそのメリット・デメリットを示す。

	メリット	デメリット
雑誌	定期的な発行 視覚的に効果的	発行日が限定 情報量が限定
無料ガイド	無料 視覚的に効果的	情報量が少ない 入手できる場所が限定される
新聞	視覚的に効果的	情報量が少ない 不定期
TV	全国配信が可能	不定期 情報が限定
WWW	情報量が多い 最新情報が獲得可能	無駄な情報が多い 情報を効果的に収集することが困難

表1 各メディアのメリット・デメリット

これらのメディアの中でも特に、WWWは情報を効果的に収集することが困難であり、その収集した情報も必要とする情報なのか否かを判断するのに手間がかかるといった問題点がある一方、情報量が多く、時間や場所に依存しないといった利点がある。また、観光客がWWW上の掲示板等に実際に訪れた立場からの情報を掲載しているため、提供側の立場からの情報だけでなく、利用者側の立場からの情報もまた多く提供されている。

このようにWWW上には多種多様な情報が存在するため、WWW上に存在する観光情報を分析することによって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、より有効な情報の提供を行うためには、第一に北海道の観光情報そのものを明確にする必要がある。そこで、本研究ではWebデータを分

析することにより、北海道の観光情報の明確化に関する研究を行う。

3. Web データからの情報抽出

北海道の観光情報を明確化するためには、観光情報を提供している Web ページ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 ページの多くは、HTML 言語によって記述されている。

HTML 言語は、不等号で囲んだ命令であるタグと呼ばれる記述方式に基づき、テキストを開始タグと開始タグにスラッシュタグを付加した終了タグで囲んで記述することによって文章や画像などの表示方法を指定する。つまり、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 上の北海道の観光情報を分析するためには Web ページ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し、Web ページにおける視覚的構造と意味的構造の関係を調査・分析することによって、そこに現れる北海道の観光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の観光情報を明確化することができると考えられる。

3. 1 Web データ

分析に用いる Web データには、現在公開されている WWW 上の北海道の観光情報を用いる。そこで、まず WWW における情報収集の方法として一般的である検索エンジンを用いて北海道の観光情報の収集を行う。

検索エンジンには、ディレクトリ型検索エンジンとロボット型検索エンジンという 2 種類が存在する。今回、その 2 種類の検索エンジンにより結果に違いが出るかを確認するため、デ

ィレクトリ型検索エンジンである Yahoo!JAPAN とロボット型検索エンジンである goo という 2 つの検索エンジンを用いて情報収集を行う。また、その検索エンジンに対するクエリとして、「北海道」と「観光」という 2 つのキーワードを用いた。これは、収集する情報が観光に関するもので、なおかつそれが北海道のものであることに限定したかったためである。

検索結果として、Yahoo!JAPAN では約 124000 件、goo では 103401 件がヒットした。このうち、各検索エンジンによって表示された Web ページの上位 100 件づつを用いてタグ分析を行う。ここで、上位 100 件としたのは、検索エンジンにより示される結果の中でも上位の Web ページのほうが、より与えた検索キーワードと関連が深いものが示されていると考えられること、また、検索エンジンごとに表示件数が上位数百件と定められているためである。

3. 2 タグ情報

このようにして得られた Web ページにおけるタグ情報の分析を行う。

Web ページには様々なタグ情報が含まれているが、今回は、分析対象として

<TITLE> : タイトル

<H1> : 見出し（レベル 1）

<HREF> : リンク

<COLOR> : 色指定

という 4 つのタグ情報を収集した。この 4 つは、タグ情報の中でも検索結果として得られた Web ページにおいて、提供者の意図が反映されやすく、Web ページの内容と関連が深いものと考えられる。

次に、HTML の規則に従って記述された

Web ページのソースからこれらのタグによって囲まれたテキストを抽出し、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査することにより、データ全体における各語句の出現傾向の分析を行う。

4. 分析の結果と考察

タグの数

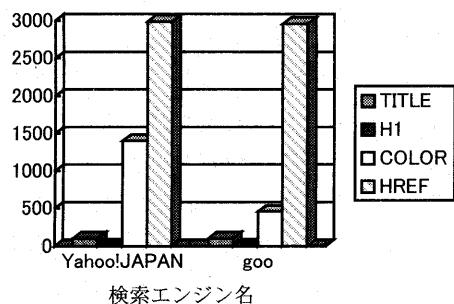


図1 各タグの出現頻度

まず、2種類の検索エンジンによるタグの数の比較を行った。

検索エンジン間における、それぞれのタグ数はほぼ一致している。したがって、ディレクトリ型検索エンジンとロボット型検索エンジンという種類の違いによるタグの出現頻度の差は無いものと考えられる。

ここで、**<TITLE>**タグは各ページにおいて一度しか記述されないため、頻度的には少ないものとなっている。また、**<H1>**タグもまたほとんど出現していないことが分かる。

今回、分析対象としている4つのタグのうち**<HREF>**タグが最も多く、そのWebページの情報量として多くのものをもっていると考えられる。しかしながら、Webページの情報は、それぞれのタグによって囲まれているテキスト自体に依存する。したがって、これらのタグ

によって囲まれるテキストについて分析を行う。

まず、**<TITLE>**タグによって囲まれているテキストを分析した。

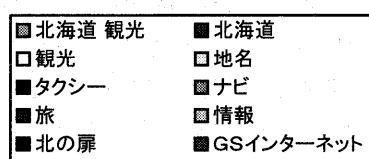


図2 <TITLE>

これは、**<TITLE>**タグによって囲まれているテキストを語句単位に分解し、各語句についての出現頻度を示したものである。この図2から検索に用いたキーワードである、「北海道」と「観光」という語句を含むことが多く、他の語句を含むことがほとんどないことが明らかとなった。このことから、**<TITLE>**タグによって囲まれるテキストは、そのWebページの概要的な内容を表しているものと考えられる。

次に、**<HREF>**タグによって囲まれているテキストについて分析を行った。

ここで、**<HREF>**タグの場合にはタグの間にURLやgifファイルといった、テキストではないものが含まれることが考えられるため、**<HREF>**タグにテキストが含まれない頻度について調査した。

その結果が図3である。グラフの上部の斜線部分がgifファイル等の画像ファイルによってリンク付けされるためにテキストの分析が不可能となっている<HREF>タグの数である。

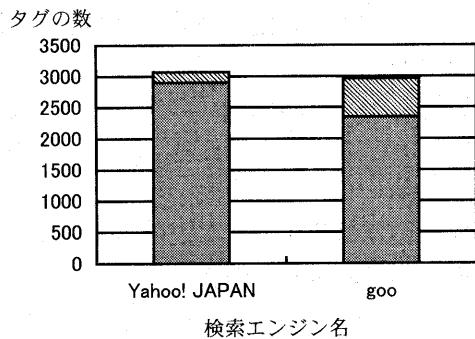


図3 テキスト<HREF>タグではない割合

<HREF>タグにテキストが含まれない頻度が全体の頻度に占める割合は比較的少ない。したがって、タグによって囲まれているのがテキストである<HREF>タグのみを扱い、そのテキストの内容についての分析を行っても問題はないものと考えられる。

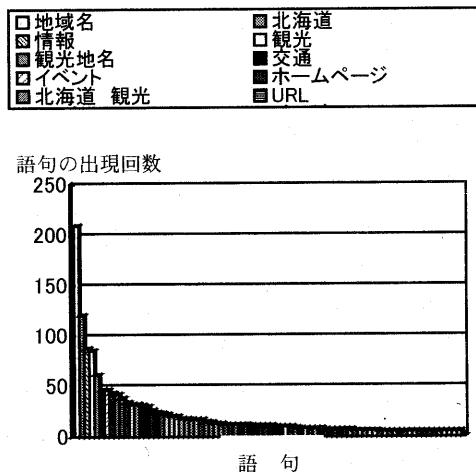


図4 <HREF>

図4に<HREF>タグによって囲まれているテキストの分析結果を示す。ここで、例として出現頻度の高い単語の上位10個を示してある。

<HREF>タグの場合は<TITLE>タグと異なり、「北海道」と「観光」という語句より「地域名」に関する語句が多く出現することが明らかになった。また、「イベント」、「URL」といった語句が多く出現している一方で、バナー広告といった観光に関係のない語句も多く出現していた。このことから、<HREF>タグによって囲まれているテキストには、観光に関連する語句が多く含まれる一方で、必要とされない情報、つまり多量のノイズをも含んでいることが分かる。

最後に、<COLOR>タグによって囲まれているテキストの分析を行い、その結果を以下に示す。

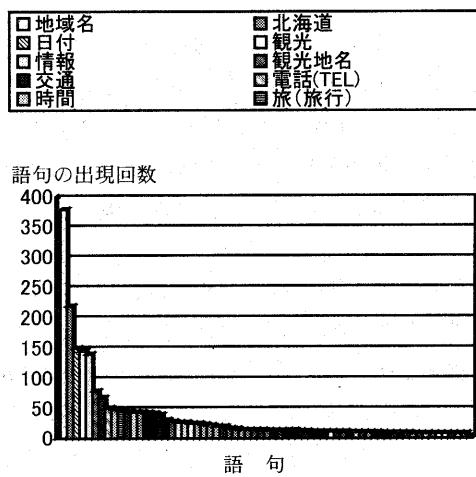


図5 <COLOR>

<COLOR>タグに関しては、「地域名」に関する語句が最も多く出現し、その他の頻度の多い語句には観光と関係する語句を多く含んで

これらのタグ情報の分析結果から以下のこととが推測される。まず、<TITLE>タグによって囲まれているテキストは検索エンジンに対して出した要求と一致するテキストが含まれているため、そのWebページの概略を表す内容を捉えているものと考えることができる。また、<HREF>タグと<COLOR>タグによって囲まれているテキストには、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していることから、そのWebページのある概念に沿った具体的な内容を示していると考えられる。

したがって、このようなタグ情報を用いることにより、Webページ全体を調査することなく、Webページ自体の内容を把握することができるようと考えられる。

5. おわりに

本研究では、北海道観光情報の明確化を目的として、Webデータにおけるタグ情報の収集を行い、その情報の分析を行った。この調査に基づき、タグ情報によってWebページの内容を把握することに対する妥当性の検討を行った。

今回行った分析では、分析対象としたタグが4種類であった。しかしながら、HTML言語にはまだ多くのタグがあり、他のタグがWebページの内容、概念に対してどのような影響を及ぼしているのかを調査する必要がある。また、今回の分析は人間による目視を必要としたため人手によって行ったが、今回の分析結果により、タグ情報のみを扱い、それを分析することによってWebページの概要を把握することができることが可能であることが示されたため、今後はタグ情報の分析の自動化を行う。また、システムを構築することにより、さらに多くの

Webページの分析を行い、統計的な有意さを明らかにしていく。

参考文献

- [1]那須川：コールセンターにおけるテキストマイニング、人工知能学会誌、16巻、2号（2001）
- [2]河野・川原：Web検索におけるテキストマイニング、人工知能学会誌、16巻、2号（2001）
- [3]坂本・有村：Webマイニング、人工知能学会誌、16巻、2号（2001）