2002年日本オペレーションズ・リサーチ学会
春季研究発表会

北海道観光情報の効果的提供に向けたシソーラスの構築

02103721　北海道大学大学院　＊金城　伊智子　KINJO Ichiko
1004631　北海道大学大学院　大内　東　OHUCHI Azuma

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，これから来訪客数の増大に結びつくような北海道観光情報の提供を行う必要がある．

現在，雑誌，TV，WWW等多様なメディアにおいて北海道観光情報が提供されている．これらのメディアの中でも特にWWWはその情報量，情報の最新性といった利点からより効果的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，WWW上で公開されてい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．

そこで，本研究では北海道観光情報のシソーラスを構築することにより必要な情報を効率的に収集し，その情報を効果的に提供するためのシソーラスの構築法を提案する．

2．北海道観光情報のシソーラスの構築

本章では，提案するシソーラスの構築法の詳細について述べる．本研究では，テキストマイニング技術[1][2][3]を基に，以下のような手順によって北海道観光情報のシソーラスの構築を行う．

[シソーラス構築法]

1．HTMLタグに基づくテキスト抽出

2．形態素解析の適用

3．名詞頻度ベクトルの作成

4．各単語間の類似度の算出

5．類似度に基づく単語のクラスタリング

以下では，その詳細を説明する．

2．1　HTMLタグに基づくテキスト抽出

Webページの内容を用い，効果的な北海道観光情報の提供を行うためには，WWWのようなノイズを含む情報においてWebページが表す内容を的確に把握する必要がある．したがって，Webページの内容を的確に把握するためにHTML言語において用いられるHTMLタグに基づくWebページの内容推定を行う．本研究では，このHTMLタグに基づくWebページ内容推定において以下の二種類のタグを採用する．

- <TITLE>タグ
- <HREF>タグ

<TITLE>タグに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と考えられる．また，<HREF>タグに囲まれるテキストには，そのWebページの具体的な内容を示していると考えられる．

本研究では，この二つのタグに囲まれるテキスト情報を個別に抽出し，以下の類似度算出およびクラスタリングにおいて用いる．

2．2　形態素解析の適用

類似度の算出において，HTMLタグに基づき抽出されたテキスト情報全てを利用した場合には計算コストが膨大であると考えられる．また，テキストにはノイズが含まれると考えられるため，適切な類似度の算出が困難である．したがって，本研究ではHTMLタグに基づき抽出したテキスト情報に対して形態素解析を適用すること

によってこれらの問題の解決を行う．本研究では形態素解析のために「茶筅」を用いる．まず「茶筅」によって分解されたテキスト情報の中から名詞句の単語のみを抽出する．

2．3　名詞頻度ベクトルの作成

次に，各名詞句の単語の出現頻度を算出し，以下のような名詞頻度ベクトルを作成する．

Webページをdとし，m個のdの集合を

$$D=\{d_1,d_2,\cdots,d_i,\cdots,d_m\} \qquad (1)$$

とする．また，Dに含まれる名詞句の単語を単語kとし，その集合を

$$K=\{k_1,k_2,\cdots,k_j,\cdots,k_n\} \qquad (2)$$

とする．ただし，nはDに含まれる単語の総数である．

このとき，あるWebページd_iに含まれる単語k_jの出現頻度をw_i^jとすると，単語k_jに対する名詞頻度ベクトルW^jは

$$W^j=\{w_1^j,w_2^j,\cdots,w_i^j,\cdots,w_m^j\} \qquad (3)$$

となる．

この名詞頻度ベクトルは出現頻度順にソートし，類似度の算出においては，名詞頻度ベクトルにおいて高頻度の名詞句を用いる．高頻度の名詞句のみを類似度算出において利用することによりノイズに対してロバストであり，かつ計算コストを抑えた類似度算出が可能となる．

2．4　各単語間の類似度の算出

形態素解析に基づき作成された名詞頻度ベクトルを用いて各単語間の類似度を算出する．この類似度の算出においては，<TITLE>タグおよび<HREF>タグ，二つのタグに基づく名詞頻度ベクトルを用いる．以下に名詞頻度ベクトルに基づく類似度の算出方法を示す．

ある単語k_pとk_qの類似度R_{pq}をそれらの名詞頻度ベクトル間の内積

$$R_{pq}=W^p\cdot W^q \qquad (4)$$

により算出する．したがって，R_{pq}は単語k_p，k_q間の共起関係を表すものとなる．

2．5　単語のクラスタリング

続いて，算出された類似度に基づき各単語のクラスタリングを行う．クラスタリング方法としては類似度が最大の単語を結合する最短距離法を採用する．

このような単語のクラスタリングを行うことによって，北海道観光情報のシソーラスを構築する．利用者は生成された北海道観光情報のシソーラスに基づき，WWW 上の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観光情報の提供が可能となる．

3．おわりに

本研究では，北海道観光情報の効果的な提供を行うことを目的とし，北海道観光情報のシソーラスの構築法を提案した．その具体的な結果を当日発表する．

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）．

[2]河野浩之，川原稔：Web 検索におけるテキストマイニング，人工知能学会誌，Vol.16，No.2，pp.212-218（2001）．

[3]坂本比呂志，有村博紀：Web マイニング，人工知能学会誌，Vol.16，No.2，pp.233-238（2001）．