2002年日本オペレーションズ・リサーチ学会
秋季研究発表会

Webマイニングによる北海道観光情報に関するシソーラスの構築

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

現在，北海道観光情報が様々なメディアにおいて提供されている．特に，WWWは情報を効率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，最新情報が獲得可能であるといった利点がある．

本研究では，このようなWWW上の情報をテキストマイニング技術[1]により分析し，そこから得られる単語の出現頻度や出現頻度のパターンから単語間の意味的な関係を推測，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に関するシソーラスの構築を行う．

2. 北海道観光情報に関するシソーラス

例えば，北海道へ旅行したいと考え，北海道に関する観光情報を収集しようと検索エンジンを利用するユーザがいるとする．ユーザが検索エンジンに対するクエリとして「とうきび」という単語を用いた場合，北海道観光情報に関するシソーラスを用いた検索結果としては「コーン」，「トウモロコシ」，「とうもろこし」といった通常のシソーラスを用いた検索結果が提示されるのに加えて，「じゃがいも」や「カニ」，「うに」といった北海道の特産物の単語を含む検索結果も提示されることが望ましい（図1）．

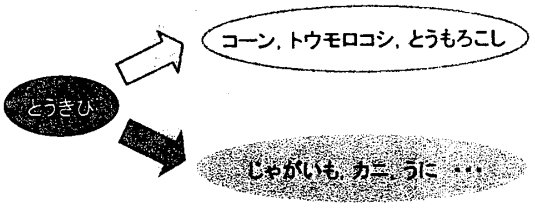

図1　北海道観光情報に関するシソーラス

ここで，「じゃがいも」，「カニ」，「うに」，・・・を「とうきび」のシソーラスであるとみなす．ユーザは生成されたシソーラスを用いることによって，「とうきび」に関する情報だけでなく，北海道の特産物等のグルメ情報をも入手することが可能となる．

Webページは検索したいと考える場所や地域，また季節によって時々刻々と変化するものである．このようなWebページに対応させるためには，シソーラスもまた常に更新する必要性がある．そこで，本研究ではシソーラスの構築法を自動化することよってWebページの最新情報に常に対応できるようにする．

3 実験

3.1 Webデータ

実験に用いるWebデータとして，現在WWW上で公開されている北海道に関する観光情報を用いる．まず，WWW上における情報収集の一般的な方法である検索エンジンを用いて北海道に関する観光情報の収集を行う．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するためである．次に，検索結果のうち北海道観光情報に関するシソーラスとして妥当な単語がクラスタリングされるようなWebページを10ページ用いて実験を行う．ここで，このようなWebページを用いるのは，単語のクラスタリングを行うシステムが的確なシソーラスを構築できるかどうか評価実験を行うためである．

3.2 HTMLタグに基づくテキスト情報抽出

WWW上の情報は広告等多量のノイズを含む情報である．このようなノイズを含む情報を用い，効率的な北海道観光情報の収集を行うためには，Webページが表す内容を的確に把握する必要がある．そこで，HTMLタグに基づくテキスト情報を用いることにより Webページの内容把握を行う．

本研究では，<TITLE>タグと<HREF>タグ 2種類のタグを採用する．<TITLE>タグに囲まれるテキストにはそのWebページの概要的な内容が，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が示されていると考えられるからである．本研究では，この2種類のタグに基づくテキスト情報を用いて単語間の類似度算出およびクラスタリングを行う．

3.3 名詞頻度ベクトルの作成

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[2]を適用する．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みを抽出する．次に，各名詞句の各ページに対する出現頻度を算出し，名詞頻度ベクトルを作成する．ここで，名詞頻度ベクトル作成において使用する，主な記号と定義について記述する．

定義1　Webページ集合D

HTML文書で記述されたWebページをdとする．また，m個のdの集合を

$$D=\{d_1,d_2,\cdots,d_i,\cdots,d_m\}$$

とする．

定義2　名詞句集合K

Dに含まれる名詞句を単語kとし，その集合を

$$K=\{k_1,k_2,\cdots,k_j,\cdots,k_n\}$$

とする．ただし，nはDに含まれる単語の総数である．

定義3　名詞頻度ベクトルW

このとき，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

$$W^j=(w_1^j,w_2^j,\cdots,w_i^j,\cdots,w_m^j)$$

となる．

この操作によって，各Webページにおける<TITLE>タグに基づく名詞頻度ベクトルおよび<HREF>タグに基づく名詞頻度ベクトルが作成される．

3.4 単語のクラスタリング

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．ここで，類似度算出において使用する，主な記号と定義について記述する．

定義4　類似度

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の内積で表し，

$$R_{pq}=W^p\cdot W^q$$

とする．したがって，R_{pq}は単語k_p，k_q間の共起関係に対応する．

最後に，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．

4. 結果と考察

検索結果のうち，北海道観光情報に関するシソーラスとして妥当な単語がクラスタリングされるような「函館」という単語を含むWebページを10ページ用いて実験を行った．その結果を以下に示す．

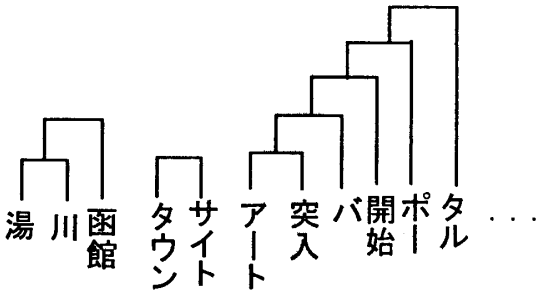

図2　単語のクラスタリング結果

「湯」と「川」は「函館」にある温泉街の「湯の川」を示していると思われるが，意味辞書に「湯の川」を登録していなかったために別々の単語としてクラスタリングされてしまったと考えられる．しかしながら，「函館」と関連のある単語がクラスタリングされていることが分かる．また，このような関連のある単語もクラスタリングされている一方，関連性のあまりない単語もクラスタリングされている．今後，その原因と考えられる類似度の算出方法，採用するタグ，Webページの選択方法の改善を行わなければならない．

5. おわりに

本研究では，現在公開されているWWW上の北海道観光情報を効率的に収集し，それを用いてより効果的な情報の提供を行うことを目的とした北海道観光情報に関するシソーラスの構築法を提案し，その評価実験を行った．

参考文献

[1]那須川　哲哉：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，Vol.16, No.2, pp.219-225（2001）

[2] 松本　祐治　他："形態素解析システム『茶筅』version 2.2.6 仕様説明書"，奈良先端科学技術大学院大学　松本研究室（2001）．