D-046

Webコンテンツマイニング によるページ間の類似性の判定ツール

大薮 永† 小柳 滋†

†立命館大学大学院 理工学研究科

1. はじめに

Webコンテンツ内のテキストデータより、テキストマイニングの手法を用いて重要単語を抽出し、コンテンツ間の相関関係による類似性を判定するツールの開発をおこなった。データの前処理の部分では、フリーの形態素解析ソフト"茶筌"を用い、重要単語の抽出にはTF/IDF法という重み付け手法を用いた。類似性に関しては、凝集法というクラスタリングの手法を用いた。また、このマイニング結果をGUI上で確認することができるように、VC++を用いたツールを作成した。これにより、任意のコンテンツ集合よりコンテンツ間の類似性やカテゴリ分けが可能になると考えられる。

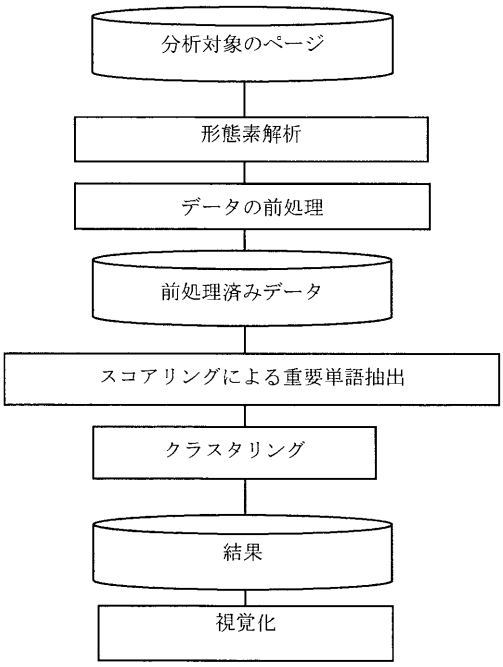

図1 処理の流れ

A Tool for Measuring Similarity between Web Pages by Web Contents Mining
Hisashi Ohyabu† Sigeru Oyanagi†
† Graduate School of Sience and Engineering Ritsumeikan University

2. システムの概要

Webページにおけるテキストデータを用いて単語の重要度より話題抽出を実現すること、そして、各ページの単語の重要度によるドキュメント間の類似性の検証を目的とした。また、クラスタリングにより、類似するWebコミュニティの発見とカテゴリ分けも目的とする。システム全体の処理の流れを図1に示す。

分析対象のWebページを選択する。テキストデータに対して、茶筌を用いた形態素解析を行い、その結果のデータに対して前処理を行う。そして、前処理済みデータに対して、スコアリングによる単語抽出・各ページ間のクラスタリングの処理を行い結果を視覚化する。

3. 形態素解析

索引語としては適当でない単語は不要語（stop word）と呼ばれる。まずは、不要語を省く必要がある。不要語とは、日本語の助詞などはきわめて頻繁に使われる単語であり、ほとんど全ての文章に高い頻度で出現してしまう。また、名詞句でもたくさんのページに出現し、かつそのページ内でも頻出する単語も不要語とする。これらでは、文章の特徴を抽出する評価としては使用できないので、不要語とする。

今回の前処理では、選択したデータに、茶筌を実行して、実行結果（単語、品詞）より名詞句、未知語句のみを抽出し、単語、その文書中における出現回数のカウント、その単語の出現ページを書き出す。これにより不要語を識別し除外して前処理済みデータを抽出した。

4. 重要単語の抽出

スコアリングによる重要単語の抽出には、TF/IDF法を用いた。

- TF（Term Frequency）法
  - ➢ 同一文書で繰り返し出現する単語が重要
- IDF（Inverted Document Frequency）法
  - ➢ 出現する文書数が少ない単語は文書の絞り込みに役立つから重要

TF/IDF値
＝ 頻度 × 1000 ×(log（総文書数 / 出現文書数）+ 1)

また、文書の長さによる影響をなくすために、長い文書中の索引語の重みを文書長に応じて小さくする処理として"コサイン正規化"を用いた。

$$\text{コサイン正規化数} = \sqrt{\sum_{i=0}^{m}(ligi)^2}$$

以上より、重要単語の抽出は

単語のスコア= TF/IDF 値/コサイン正規化数

によって計算した。

5. クラスタリング

クラスタリングは探索的知識獲得、教師なし学習と呼ばれるデータマイニング手法のひとつである。互いに類似するレコードのグループ、すなわち、クラスタを探索する。

今回、クラスタリングに際しては、WWW ページをレコードとして凝集法を用いた。凝集法では、各レコードが合併して、１つの大きなクラスタに集められるまで、しだいにクラスタを合併するという操作を繰り返す。

最初のステップは、類似度行列をつくることである。類似度行列は、すべてのレコード間の組み合わせに対する距離の表である。

距離は、2 つのページ間のそれぞれの単語のスコアの内積をとる。

クラスタの合併の全履歴により木構造が形成されるので、事例にもっとも適したクラスタリングのレベルを選択できる。

6. 視覚化

分析対象となる前処理済みデータを選択すると、クラスタリングの結果の tree が表示される。

例えば、全 5 ページのデータが(((1 2) 3)(4 5))のように合併した場合のクラスタリング結果は、図２のように表示される。

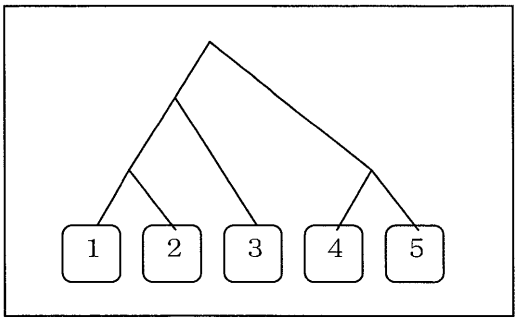

図２　クラスタリング結果の表示

7. 評価データ

データは、立命館大学の理工学部の以下の５学系の各研究室のページを対象にした。

数学物理系・１２ページ
応用化学系・１３ページ
機械システム系・１８ページ
電気システム系・１５ページ
情報系・２６ページ

データは、テキストデータがある程度含まれるページを対象として選別した。

また、本システムによる分類と、その各ページの所属する学科の分類を比較することにより評価を行った。

8. 結果と分析

クラスタリングの結果は、大きく分けて３つのカテゴリとその他に分かれた。

① 応用化学系と数学物理系が混在するクラスタが形成された。
共通する重要単語は、全体としては、研究、分子が多く、他には、たんぱく質、イオン、分析、物質、溶液、構造、現象などが共通していた。
研究や分子という比較的どこにでも使われるような単語の重みが大きくなってしまったことにより応用科学系と数学物理系を識別するクラスタが形成されなかったと考えられる。

② 機械系のクラスタが形成された。（電気電子系 x ２ 応用化学系 x １）
共通した重要単語はマイクロ、ロボット、シリコン、センサなどであった。

③ 情報系のクラスタが形成された。（電気電子系 x ６ 数学物理系 x １）
全体として、共通していたのは情報、システム、コンピュータ、ネットワークなどであった。
クラスタ②や③では、そのカテゴリの特徴を表し、他のカテゴリでは、出現しないであろう単語の重みが大きくなる結果となったので、うまくクラスタが形成されたと考えられる。５つの学系のうち電気電子系のクラスタが形成されなかったのは、情報系と機械システム系に分断されてしまったからである。

9. おわりに

Web コンテンツ内のテキストデータより、重要単語を抽出し、コンテンツ間の類似性を判定するツールの開発をおこなった。それを実際のホームページに適用し、ある程度正しく分類できることを確認した。更なる精度の向上のためには、重要単語の抽出の精度を上げること、およびクラスタリングにおけるクラスタ形成の制御を行うことが考えられる。

参考文献

[1]マイケル J.A. ベリー/ゴードンリノフ著：データマイニング手法、1999
[2]林晴比古著：新 Visual C++6.0 入門シニア編、ソフトバンクパブリッシング、2001
[3]GeorgeChang/MarcusJ.Healey/JamesA.M.McHugh/JasonT.L.Wang 著：Web マイニング、2004
[4]北 研二/津田 和彦/獅々掘 正幹著：情報検索アルゴリズム、2002