

TETDM 統合環境と R によるテキストマイニング

Text Mining Using R and TETDM

徳永 秀和

Tokunaga Hidekazu

香川高等専門学校

Kagawa National College of Technolog

In this study, we develop a Web retrieval support system that suggests the word. The word is keyword ideas which used by next search, so that the users do not have to read a snippet. The Web retrieval support system is using the words in the snippet, and indicates the subject of the search results to the users. Then determine the keyword that the result of selecting the topic of user demands. We measured the time to reach the purpose site between the Web retrieval support system, "predictive search suggestions for Google" and site re-ranking "Rerank.jp". As a result, Google search and Rerank.jp was not able to reach the purpose site. The Web retrieval support system can be obtained a search guideline without knowledge of the user .

1. はじめに

Web から情報を得るために、関連したキーワードを用いて検索エンジンにより目的のページを得る方法が使われる。しかし、得たい情報についての知識が少ない状態では、適切なキーワードを用いた検索クエリを作り出すことができず、検索結果のスニペットやリンク先のページを読み、次の検索クエリを探し出すことが行われる。このため、検索者は自分の得たい情報とは無関係な情報も多く読むことになる。

本論文ではユーザが、検索結果のスニペットを読むことなく、次の検索に用いるキーワードを選択できるように、単語を提示する Web 検索支援システムを提案する。本システムは、スニペットに含まれる単語を利用して、検索結果がどのような話題のページを含んでいるかをユーザに示す。そして、ユーザが自分の求めている話題を選択することによって、推薦するキーワードを決定し、提示するというものである。

システム開発を、TETDM[砂山 2011]と統計計算ソフト R[豊田 2008]とデータマイニングソフト Weka[IAN 1999]を利用することによって、他のテキストマイニングツールとの統合、統計処理、データマイニングツールの利用などによる機能拡張が容易なシステムを実現する。

2. SOMと決定木による推薦キーワード抽出

(1) 検索エンジンに検索ワードを入力し検索を行い、検索結果を得る。得られた検索結果のスニペットから名詞のみを抽出する。この名詞を推薦キーワード候補の単語とする。

(2) すべての抽出された名詞に対して、その名詞がどの URL のスニペットに出現していたかを属性とし、出現していれば1、出現していなければ0という属性値を与える。これにより、各単語の特徴ベクトルを作成する。

(3) 単語の特徴ベクトルを自己組織化マップ(SOM)[コホネン 1996]によって、類似した特徴ベクトルの単語が集まったポジショニングマップを作成する。

(4) ユーザが単語のポジショニングマップを興味のある単語集合 A と興味のない単語集合 B に分割する。

(5) 各 URL に、スニペットに出現した単語を属性とし、出現していれば1、出現していなければ0という属性値を与える。これ

により各URLの特徴ベクトルを作成する。

(6) URL のクラスの決定方法は、URL の特徴ベクトルの属性値が1である単語の属する集合が A か B かを調べる。集合 A に属する単語の方が多ければ興味のある URL とし、classA を与える。集合 B に属する単語の方が多ければ興味のない URL とし、classB を与える。集合 A と B に属する単語数が同じときは、classA か B をランダムに与える

(7) クラス属性を追加した URL の特徴ベクトルを決定木学習 [キンラン 1995]する。この決定木の根に近い単語ほど、興味がある URL と興味のない URL を効率よく分類することができることになる。

(8) ユーザが決定木をみて、興味を持ったキーワードを追加し、再検索を行う。そして、(1)から繰り返すことによって、ページの絞込みや新たな興味を持ったキーワードについて検索範囲の拡大などを行い、情報を収集する。

3. TETDM 統合環境

TETDM プロジェクトの統合環境は、図1のような構成となっており、開発言語は Java である。簡単な処理の流れは次のようになる。解析対象のテキストを統合環境インターフェース部に入力する。テキストは日本語形態素解析ソフト ChaSen により形態素解析され、テキストマイニング処理に必要な様々なデータを取り出すインターフェースを持つ Java のオブジェクト「text」を生成する。パネル内で、マイニング処理が制御処理部から必要な情報を取り出し、テキストマイニングを行って、可視化 IF によりユーザに結果を提示する。

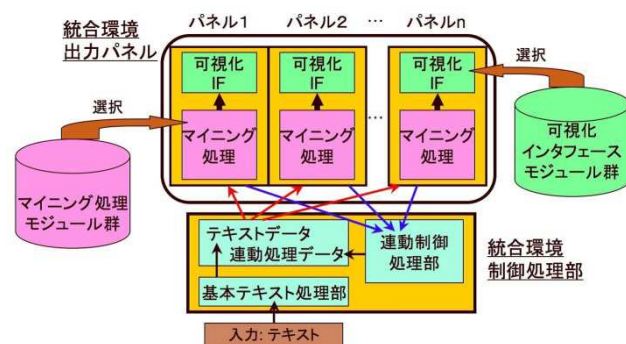


図1 TETDM 統合環境

連絡先: 徳永秀和, 香川高等専門学校 機械電子工学科,
〒761-8058 高松市勅使町355, tokunaga@t.kagawa-nct.ac.jp

4. システム構成

以下の3個のパネル(マイニング処理と可視化 IF の組)によって構成する。Rを利用して単語ベクトル, ポジショニングマップを作成する SOM 処理モジュールパネル, ポジショニングマップを表示しユーザに単語集合を作成してもらいURLベクトルを作る MAP 表示モジュールパネル, Wekaを利用して決定木を作成し推薦単語を表示する決定木処理モジュールパネル。

処理の詳細を図2のシステム構成図の記号に合わせて以下に説明する。

- HTML ページ内の javascript によりスニペットを取得し, 統合環境の入力形式のテキストファイルを作成し, 保存する。
- 統合環境の制御処理部にテキストを入力することによって, 統合環境が ChaSen による形態素解析を行い, text オブジェクトを生成する。
- SOM 処理モジュールによって, text オブジェクトを利用して単語ベクトルと URL ベクトルを作成する。
- SOM 処理モジュールにより, R を実行してポジショニングマップを作成する。
- MAP 表示モジュールにより, ポジショニングマップを表示する。SOM 処理モジュールと Display モジュールにより, ポジショニングマップの各領域に割り当てられた単語の詳細を表示する。(図3参照)
- 決定木処理モジュールにより, ユーザがクラス分けを行い, URL の特徴ベクトルにクラス属性を追加する。
- 決定木処理モジュールにより, Weka を実行して決定木を作成する。
- 決定木処理モジュールにより, 決定木のデータを取得し, 結果を整理してユーザに提示する。

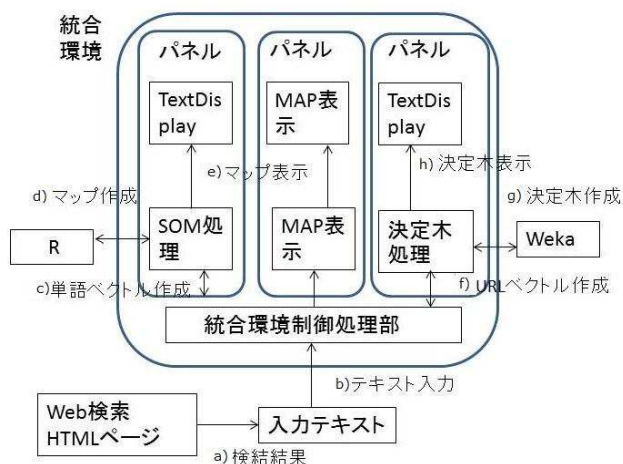


図2 システム構成図

5. 実験結果

Web 検索支援システムの有用性を調べるために, Web 検索支援システムを使用した場合と, Google の予測変換・検索候補を利用した場合, またランキングサイト「Rerank.jp」[山本 2008]を利用した場合の, 目的のサイトにたどり着くまでの所要時間の比較実験を行った。ランキングとは, ユーザの興味に合わせて検索結果のサイト順序を再ランキングして表示するというものである。

「量子力学」から検索を開始して, 最終的に「統一場理論」について書かれているサイトにたどり着けるかどうかと, その過程についてまとめる。本検索支援システムでは(1)「現象, ファイル」

が推薦語として示され, (2)「量子力学, 現象」で検索し「数学, 研究所, 巨視的」が推薦語として示され, (3)「量子力学, 巨視的」で検索し「関数, 量子, 支配, 効果, 学問」が推薦語として示され, (4)「量子力学, 量子, 支配」で検索し「統一場理論」にたどりついた。かかった時間は14分である。Google 検索では, 予測変換や検索候補として出された「入門, とは, 発展, 応用, 理論」などを利用して, 各検索ワードに対して 10 ページずつ閲覧していった。その結果, システムを使用したときの時間を越えてもサイトを発見することができなかった。Rerank.jp では, 提示された「理解, 粒子, 世界の謎, 無限」などを上位に持っていく処理をしたが, キーワードがほぼ更新されず同じサイトばかり表示されるため, 途中で検索を断念した。

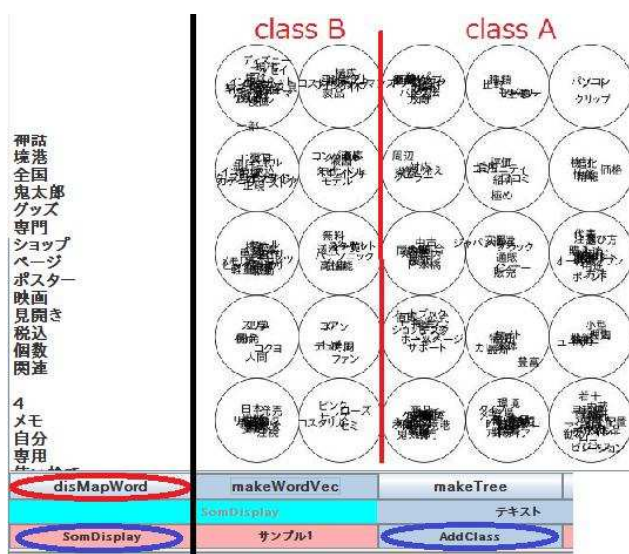


図3 ポジショニングマップの表示画面

6. おわりに

検索結果のスニペットに存在する単語を自己組織化マップによるマップを提示し, ユーザに興味のある単語集合を選択してもらうことによって, 再検索に有効な単語を推薦するシステムを提案した。実験結果より, 検索対象の知識が乏しいユーザが, より専門的な情報にたどり着けることを確認できた。多くの人が興味をもつ情報は協調フィルタリングによって効率的に得ることができるが, 少数の者しか興味を持たない情報を検索するには本システムが有用であると考えられる。

参考文献

[砂山 2011] 砂山 渡, 高間 康史, ダヌシカボレガラ, 西原 陽子, 徳永 秀和, 串間 宗夫, 松下 光範, : Total Environment for Text Data Mining, 人工知能学会論文誌, Vol.26, No.4,2011.

[豊田 2008] 豊田秀樹: データマイニング入門, 東京図書 2008.

[IAN 1999] IAN H.WITTEN,EIBE FRANK: Data Mining, MORGAN KAUFMANN PUBLISHERS ,1999.

[コホネン 1996] T.コホネン: 自己組織化マップ, シュプリンガー・フェアラーク東京, 1996.

[キンラン 1995] J.R.キンラン: AIによるデータ解析,トッパン,1995.

[山本 2008] 山本岳洋, 中村聡史, 田中克己: Rerank-By-Example: 編集操作の意図伝播によるウェブ検索結果のランキング, 情報処理学会論文誌 データベース Vol. 49, No. SIG 7,2008.