1B2-NFC3-1

各種ツールを統合したWeb検索支援システムの開発

Development of Web Retrieval Support Ststem which Unified Various Tools

徳永　秀和
Tokunaga Hidekazu

香川高等専門学校
Kgawa National College of Technology

A user gets the target page by a search engine with the related key word, in order to acquire information from the Internet.However, when there is little a user's information, it is difficult to choose a suitable key word.Therefore, it is necessary to read many unrelated homepages.In order to solve this problem, key word selection support, re-ranking, etc. are performed.The basic technology is the clustering of a page and the keyword extraction of a class which were searched.We proposed the approach of supporting search using SOM and a decision tree.This method is implemented as a system compatible with the integrated environment proposed in Total Environment for Text Data Mining of Challenge for Realizing Early Profits.WebAPI, MeCab, and R are used for implementation.

1. はじめに

Web から知識を得るために，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，知識が少ない状態では，適切なキーワードを選択することができず，無関係な検索結果ページも多く閲覧することになる。この問題の解決のために，検索キーワード選択支援[大澤 1999][若木 2006]や検索結果ページの分類[佐野 1998]や再ランキング[山本 2008]などの研究が行われている。

このような検索支援の基本は，検索結果ページのクラスタリングと各々のクラスのキーワードを抽出すること，そしてそれらを見やすく表示することである。我々は検索結果のスニペットを利用し，自己組織化マップ(SOM)とユーザの意思によりクラスタリングを行い，決定木によりキーワードを抽出することにより，検索支援を実現できる可能性を見出した。[松原 2010]

この手法をシステム化するためには，WebAPI，形態素解析，データマイニング，テキストマイニング，GUIなどの様々な既存のツールを統合して利用することが必要である。また，人工知能学会近未来チャレンジ Total Environment for Text Data Mining（TETDM）においてテキストマイニングの統合環境の構築を目指している[砂山 2010]。この統合環境を考慮した各種ツールを統合した Web 検索システムの構成案を示す。

2. SOMと決定木によるWeb検索支援

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に，形態素解析を行い名詞のみを抽出する．抽出された名詞が，どの URL に含まれているかを基準とし図1の単語ベクトルを作成する。単語が URL に含まれている場合は 1，含まれていない場合は 0 としている．単語ベクトルの次元は URL の数となる．作成した単語ベクトルを自己組織化マップによってポジショニングマップを作成する．そして，図2のようにユーザがポジショニングマップ上に境界線を引くことによって，クラス分けを行う．ユーザが境界線を引き文房具関連の単語をクラス A，パソコン関連の単語をクラス B に分けている．

次にクラス A とクラス B を分ける際に影響力が強い単語はどれかということを決定木によって抽出する．これにより，抽出された単語はユーザ自らがクラス分けをしたクラスの特徴語としてユーザに提示することができる．決定木学習を行うにあたって図3のような URL の属性ベクトルを作成する．このベクトルは URL に含まれる単語と属性を持っている．URL の属性の判断基準は，各 URL のスニペットにどのクラスに属する単語が多く含まれているかである．図3を用いて決定木学習を行った結果を図4に示す．決定木学習は根に近い分岐を生じさせている変数が基準変数に対して強い影響力を持っていると解釈できる．図4の例では「PC」が，クラス A，クラス B を分ける際に最も影響力が強い特徴語ということが分かる．このようにユーザとシステムの対話的なキーワード抽出を行うことができる．

単語	URL1	URL2	…	URL39	URL40
CPU	0	0	…	1	0
PC	0	1	…	0	1
価格	0	0	…	0	0

図１　単語ベクトル

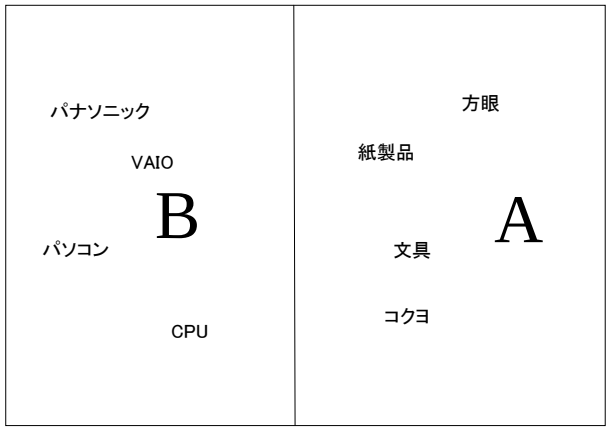

図２　ポジショニングマップ

URL	CPU	…	モール	属性
URL1	0	…	0	B
…	…	…	…	…
URL40	0	…	1	A

図３　属性ベクトル

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

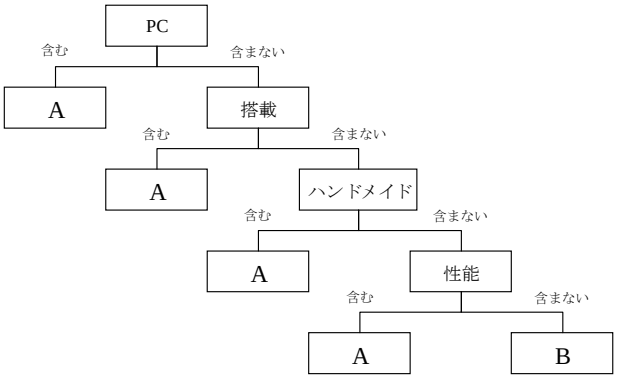

図４　B5 ノートの決定木

3. テキストマイニングの統合環境

TETDMでは，図5のような統合環境を提案している[砂山 2011]。最も単純な処理の流れを説明する。解析したい文章ファイルを統合環境インターフェース部により，形態素解析などの処理を行い，テキストマイニング処理に必要な様々なデータを出力する。モジュールでは，統合環境インターフェース部より必要なデータを取り出し，データマイニングを行い可視化インターフェースによりユーザにマイニング結果を提示する。また，TETDMでは，開発言語として java を用いることになっている。

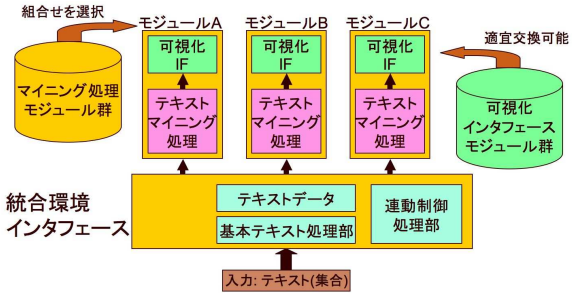

図5　TETDMの統合環境

4. Web検索支援システム

TETDMの統合環境の仕様に合わせたSOMと決定木によるWeb検索支援システムの構成を図6に示す。google などの検索サイトの WebAPI を利用して，HTML ファイル内の JavaScript により検索結果を取得する。それをサーブレットにより基本テキスト処理部に入力するテキストファイルを作成する(図8)。基本テキスト処理部は java アプリケーションとし，形態素解析 MeCab を JNI(Java Native Interface)により利用する。MeCab 用の JNI は Web 上に公開されているものを使用できる。テキストマイニング処理は java アプリケーションとし，統合環境インターフェースのテキストデータを利用し，データマイニング用のデータを作成する。データマイニング(SOMと決定木)はRを利用する。java からRを利用するためには，JRI(Java R Interface)，Sjava や Rserver などが利用できる。SOMのポジショニングマップは，Rよりデータを受け取り java のGUIにより表示し，ユーザにクラス分けしてもらう。その結果よりテキストマイニング処理が決定木用のデータを作成し，Rの決定木を利用し決定木学習を行う。学習結果の決定木の表示は R によって行う。さらに，テキストマイニングの結果より，検索結果の再表示や再検索を行う。

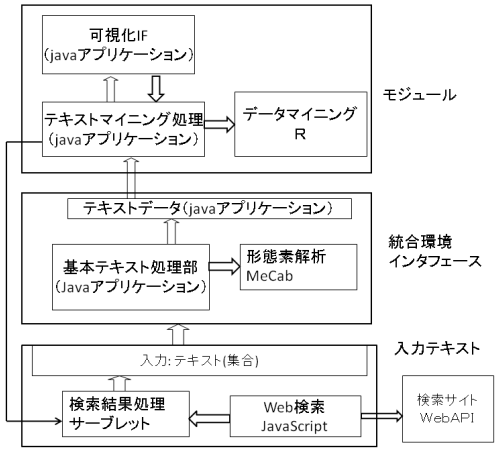

図7　Web検索支援システムの構成

図8　WebAPIを利用した検索とテキスト処理

5. おわりに

現在，java から MeCab の利用，java からRの利用，WebAPI を利用した Web 検索結果の取得とテキスト処理のテストが完了している。また，基本テキスト処理部は形態素解析に Chasen を利用した java のサンプルがTETDMにおいて作成されている。今後，可視化 IF のGUIの作成と各プログラムを統合しシステムを完成を目指す。

参考文献

[大澤 1999] 大澤幸生，N.E.Benson，谷内田正彦: "KeyGraph: 語の共起グラフの分割・統合によるキーワード抽出，電子情報通信学会論文誌 J82-D-1No.2，1999.

[佐野 1998] 佐野綾一，波多野賢治，田中克己: 自己組織化マップを用いた Web 文書の対話的分類とその視覚化，情報処理学会研究報告 Vol.98 No.57，1998.

[砂山 2008]　砂山 渡: 情報編纂と創造的テキストマイニングの融合による情報アクセス・分析環境，人工知能学会情報編纂研究会第4回，2011.

[砂山 2010]　砂山 渡: Total Environment for Text Data Mining，第24回人工知能学会全国大会 212-NFC0-1，2010.

[山本 2008] 山本岳洋，中村聡史，田中克己: Rerank-by-examples 編集操作の意図伝播によるウエブ検索結果のれランキング，情報処理学会論文誌データベース，2008.

[松原 2010] 松原弘樹，徳永秀和: 自己組織化マップと決定木による網羅的探索のためのキーワード抽出方法，人工知能学会情報編纂研究会第3回，2010.

[若木 2006] 若木裕美，正田備也，高須淳宏，安達淳: 具体性指向単語クラスタリングによる網羅的トピック発見と検索質問拡張支援，DEWS2006 2C-i4，2006.