D-027

Web 検索結果のクラスタリングと特徴語の抽出

Clustering of Web Search Result and Feature Term Extraction

安川 美智子† 山田 篤‡
Michiko Yasukawa Atsushi Yamada

1．はじめに

Web 検索を行うユーザは，過去に行った Web 検索の記憶をもとに，検索対象についての大まかな概念を頭の中に描く事は出来るが，検索に役立つ具体的な語を思いつくことが出来ないという場合がある．このため，ユーザが検索を行う際に，ユーザの検索に役立つ語を提示できることが望ましい．本論文では，Web 検索結果として得られる Web ページ群に対して自己組織化マップ(Self-Organizing Map: SOM)を用いたクラスタリングを行い，生成されるクラスタに含まれる語の中から Web ページの検索に役立つ特徴語を抽出する手法を提案する．

2．Web 検索における問題点

Web 検索・閲覧を行っているユーザは，いくつかの関連するトピック（主題）についての Web ページを検索するにつれて，そのトピックについての理解が深まっていくとともに，トピックに特徴的な Web ページのグループ（Web ページの概念カテゴリ）を見出していくと考えられる．そのようなユーザは，検索対象には何らかの概念カテゴリがあることは分かったとしても，その概念を特徴付ける語を必ずしも分かるわけではなく，また，ユーザが検索しようとする対象には，そのような固有の表層語（Web ページ中に出現している語）が存在しない場合もある．

たとえば図 1 のような野菜（「エンダイブ」，「コールラビ」，「トレビス」）に関する Web ページを検索しているユーザは，これらの野菜名称で検索される Web ページは，「料理」「ガーデニング」「栄養素」などいくつかのグループに分類できるということが分かってくる．このようなユーザが，たとえばエンダイブを使った「料理」についての Web ページを検索したいと考えたとき，

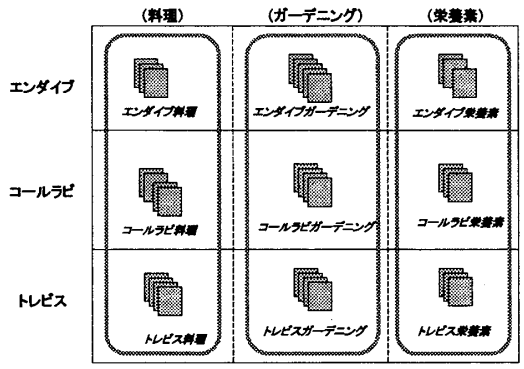

図 1 Web 検索結果に対する概念構造の例

「エンダイブ料理」のような固有の表層語にユーザの検索対象が結びついていれば，ユーザの検索は容易に行える．しかし，そのような固有の表層語が存在しない場合，ユーザの検索要求を，検索語に結び付けるための支援が必要になる．検索語の候補になり得る語をユーザに表示する仕組みとして Google サジェスト[9]がある．これは，ユーザが検索窓に入力する文字列をリアルタイムで予測して，人気が高いと思われる語をユーザに提示するものであり，検索語入力の手間を省き，ユーザにとって新しく興味深い語を表示できる利点がある．しかし，ユーザ個人の過去の検索履歴は考慮されていないため，個々のユーザが持つ Web ページの概念カテゴリに対応した検索語の提示がされない．本研究では，ユーザが過去に行った Web 検索結果から個々のユーザの検索に役立つ特徴語を抽出し，抽出した特徴語を検索支援に用いることを提案する．

3．提案手法

3．1 Web 検索結果のクラスタリング

Web ページのクラスタリングの手法として，SOM[1]を用いる手法が提案されている．Lagus らは SOM を用いて，高次元で表現された多数の文書群を二次元マップ上に射影することで検索や閲覧を支援する WEBSOM を提案している[2]．また，波多野らは SOM と既存の検索エンジンを用いた Web 文書の動的な分類と，ユーザの視点を反映した操作を対話的に可能にする Web 文書の分類ビュー機構を提案している[3]．SOM は幅広い分野に応用されている教師なし競合学習アルゴリズムであり，使いやすく整備されたソフトウエアが公開されていることから，本研究においても SOM を Web ページの分類器として利用する．Web 検索結果のクラスタリングの概要を以下に説明する．

（1）前処理

Web ページ群から茶筌[4]を用いて語を抽出し，SOM に入力する特徴ベクトルとして，語—文書ベクトル(term-document matrix)を作成する．語の重み付けには tf/idf を用いる．特徴ベクトルの次元を削減し，また，Web ページ中のトピックと無関係な語を軽減するために，使用する語に対して以下の制限を行う．

- 品詞が助詞，助動詞，記号，名詞－数である語はページの特徴を表現する上で雑音になるため除外し，「ひらがな」と「カタカナ」の一文字の語は，語の抽出に失敗している場合が多いため除外する．
- ユーザが Web 検索を行った際の検索語の近傍はページのトピックに関連が強いが，検索語から離れた位置にある語は，ページ中の別のトピックや広告等に含まれる場合が多いため除外する．
- Web ページ群全体での語の出現頻度に閾値を設け，全体的に出現頻度の低い低頻度語を除外する．

† 群馬大学工学部情報工学科, Gunma university
‡ 財団法人京都高度技術研究所, ASTEM RI.

（2）SOMの実行とクラスタの生成

まず，ユーザが閲覧したトピックにおける共通の暗黙的な概念構造に対応するクラスタリングを行ってから，次に，生成されたクラスタ中の Web ページを検索語でさらに分類することで，図1のようなユーザの概念構造を構築する．具体的には，特徴語ベクトル中の検索語をマスク（他の文字列で置き換え）して，SOM による Web ページ集合のクラスタリングを行い，クラスタが生成された後で，マスクした検索語を復元させて，検索語ごとに Web ページを分類し，Web ページのサブクラスタを生成する．

3．2　特徴語の抽出

Web ページからの特徴語抽出の手法として，相互情報量に基づく語のトピックに対する寄与度を計算する手法[5]や，語の共起の統計情報を用いてある文書で特徴的な共起をする語をキーワードとして抽出する手法[6]が提案されている．また，WEBSOM の手法によりマッピングされたノードに，情報検索の立場からではなく，データ可視化の観点からランドマークとなるキーワードを割り当てる手法[7]が提案されている．本研究では，自動分類された Web ページのクラスタから特徴語を抽出する際に，個々のクラスタにおける語の重要度だけでなく，Web ページ集合全体（コレクション）における語の重要度も考慮する必要があることから，語の抽出において，クラスタ j における単語 w の重み(term weight)として次のような指標 $G(w, j)$ を用いる．

$$G(w,j) = F_j^{cluster}(w) \cdot F_j^{collection}(w)$$

$$F_j^{cluster}(w) = F_j(w)$$

$$F_j^{collection}(w) = F_j(w) / \sum_i Fi(w)$$

ここで $Fj(w) = fj(w) / \sum_v fj(v)$ であり，$fj(w)$ はクラスタ j における単語 w の出現頻度，$F_j(w)$ はクラスタのサイズで正規化された単語 w の相対頻度を表す．

4．評価実験

テストデータ用の Web ページ群収集のため Web 検索エンジン[8]を用いた Web 検索を行った．検索クエリとして，電子辞典[10]に掲載されている野菜名称（『スーパーでみかける「新顔野菜学」話題学』）の見出し語 62 個を用いた．検索エンジンが返す検索結果総数で検索セッションをソートして，極端に少ないもの（300 件以下）を除外した．また，検索結果総数が極端に多いもの（10 万件以上）についても，クエリ文字列が野菜名称としてではなく，同音異義語やクエリ文字列を含む人名や店名が多数混在するため除外した．検索結果総数の規模が適度であるものをさらに3つのグループ（10万件以下，1万件以下，1千件以下）に分け，それぞれのグループに対して，検索結果上位 20 件の Web ページの URL にアクセスして検索結果 Web ページを収集した（表1）．各グループには16個の野菜名称（検索セッションに対応）が含まれている．収集した Web ページ群に対して前節で述べた Web 検索結果ページのクラスタリングと特徴語の抽出を行った．生成されたクラスタと抽出された特徴語を調べてみると，「料理」と「ガーデニング」に対応するクラスタが生成され，クラスタの概念に関連のある特徴語が抽出された．

抽出された特徴語の例を表 2 に示す．「栄養素」については，単独のトピックとして成立することが少なく，「料理」や「ガーデニング」のページで同時に言及されること

データセット	野菜名称	検索総数	取得ページ数合計
(除外)	ビート	798000件	0件
(除外)	香菜	207000件	0件
under100000	ルッコラ	80900件	320件
	(略)	(略)	(16クエリ×20件)
under10000	リーフレタス	9580件	320件
	(略)	(略)	(16クエリ×20件)
under1000	パクチョイ	968件	320件
	(略)	(略)	(16クエリ×20件)
(除外)	西洋種ナバナ	1件	0件

表1　実験データ（抜粋）

データ集合	野菜名称	特徴語
under100000	レホール	ソース，肉，鶏，入り，スライス，分，山，ボール，材料，枚
under10000	ロメインレタス	レシピ，オリーブ，オイル，サラダ，レタス，生，枚，時間，ドレッシング，大さじ
under1000	豆苗	炒め，ベーコン，えんどう，中国，薄切り，軽く，レシピ，料理，半分

表2　Webページクラスタから抽出した特徴語の例

が多いため，独立したクラスタとしては生成されにくかったと推察される．このようなトピック混在型の Web ページからの特徴語抽出について今後検討を行っていく予定である．

5．おわりに

本論文では SOM を用いた Web 検索結果のクラスタリングとクラスタからの特徴語抽出の手法を提案し，抽出した特徴語の例を示した．特徴語抽出の精度を向上させることが今後の課題である．

参考文献

[1] T. Kohonen 著, 徳高平蔵ほか訳, "自己組織化マップ", シュプリンガー・フェアラーク東京, 1996.

[2] K. Lagus, "Text Mining with the WEBSOM", D.Sc.(Tech) Thesis, Helsinki University of Technology, 2000.

[3] 波多野賢治, 佐野綾一, 段一為, 田中克己, "自己組織化マップと検索エンジンを用いた Web 文書の分類ビュー機構", IPSJ(TOD), Vol.40 No.SIG03, pp.47-59, 1999.

[4] 茶筌, http://chasen.naist.jp/hiki/ChaSen/

[5] 松尾 豊, 石塚 満, "語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム", 人工知能学会論文誌, Vol.17, No.3, pp.217-223, 2002.

[6] 吉岡 真治, 原口 誠, "検索語の網羅性に注目した汎化概念により検索語選択支援を行う情報検索システムの研究", 人工知能学会論文誌, Vol. 20, No. 4, pp.270-280, 2005.

[7] K. Lagus and S. Kaski, "Keyword selection method for characterizing text document maps", Proc of ICANN99, Vol. 1, pp.371-376, 1999.

[8] Google サジェスト, http://www.google.co.jp/webhp?complete =1&hl=ja

[9] Google, http://www.google.co.jp

[10] 現代用語の基礎知識 2003, ロゴヴィスタ株式会社