D-043

Web検索解析によるクラスタリング手法の研究
Research of Clustering for Web Search Results

丸山 謙志† 王 冠超‡ 徳山 豪†
Kenji Maruyama Guanchao Wang Takeshi Tokuyama

1. はじめに

1990年代にWebサーチエンジンが登場して以来,膨大なWebデータベースからいかに高速に,有効な情報を取り出すことができるかが研究されて来た.これらgoogleをはじめ多くの検索エンジンでは,ユーザが所望する情報に関連するキーワードを入力すると,キーワードを含むWebページがヒットし,それらはPage Rankなどのアルゴリズムによってソートされ,ユーザに一次元的なリストを提示する.この仕組みが有効に働くのは,ユーザが検索意図を十分絞ったキーワードで検索を行い,閲覧するのに適当なページ数を得られる場合である.しかしながら,検索エンジンに入力される大多数のキーワードは経験上1語から3語の長さが一般的で,このような場合,数百から数十万のページがヒットし,単純なランキングではユーザの意図を十分反映した結果が得られない.そのような大量な検索結果に対して,ユーザの閲覧効率を高めるために,類似したページのクラスタリングが必要となる.

Webページクラスタリングの従来技術としては,ニュース記事を対象にしたScatter/Gatter [1]やWebページ間のリンク関係を利用したWang[2]らの方法などがあるが,いずれも検索結果(タイトルとサマリー)を利用し,Webページを非排他的なグループ分けをするわれわれの手法とは異なる.また,最近,ViViSimoをはじめ,検索結果を自動分類するポータルサイトが登場しているが,残念ながら,手法の詳細が公開されていない.

本研究は,Web検索結果を話題に基づいて非排他的クラスタに分類し,各クラスタには自動的にラベル付けし,さらにはクラスタ内のランキングを行うことで,利用者のブラウジングを支援するシステムを構築した.なお,クラスタリングの手法としては,ベクトル空間モデルに基づく特異値分解(SVD)を用いた.

2. 本システムの仕組み

本システムの特徴は次の三つである.(1)非排他的クラスタの生成により,文書のもつ多様性をとらえる.(2)文書クラスタと双対関係にある特徴語クラスタからクラスタラベルを選定することで,文書クラスタを特徴づける.(3)クラスタへの適合度情報を用いてクラスタ内ランキングを行う.

構築したシステムは,下記に示す4つのフェーズで構成されている.なお,キーワード検索をし,検索エンジンより返された結果を入力とする.

2.1 データクリーニング

HTML文から,HTMLタグや非文字記号を除去し,句読点によってフレーズに分割する.分割された文書集合を$D = \{d_1, d_2, ..., d_m\}$とする.(ただし,$d_i$はそれぞれ分割されたフレーズの集合)

2.2 特徴語抽出

文書集合Dから「茶筌」[3]で形態素解析を行い,名詞を抽出してtf-idf法で重み付けし,閾値k以上の重みをもつ特徴語集合$T = \{t_1, t_2, ..., t_n\}$を抽出する.

2.3 特徴語・文書行列作成

文書集合と特徴語集合との関係をベクトル空間モデルを用いて表現する.具体的には,特徴語t_i・文書d_jを用いて,以下のように特徴語・文書行列を構築する.文書d_jは,次のようなベクトルで表される

$$\text{文書ベクトル：} \boldsymbol{d_j} = \begin{pmatrix} a_{1,j} \\ a_{2,j} \\ \vdots \\ a_{m,j} \end{pmatrix} \quad (1)$$

ここで,$a_{i,j}$は特徴語t_iの文書d_jにおける重みである.なお,$a_{i,j}$を以下のように定義する.

$$a_{i,j} = \begin{cases} 1 & \text{if } t_i \ occurs \ in \ d_j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

また,文書集合全体は,次のような$m \times n$の特徴語・文書行列Aによって表現することができる.

$$A = [\boldsymbol{d_1}, \boldsymbol{d_2}, \cdots, \boldsymbol{d_n}] = \begin{array}{c} \\ \boldsymbol{t_1} \\ \boldsymbol{t_2} \\ \vdots \\ \boldsymbol{t_m} \end{array} \begin{array}{c} \begin{array}{cccc} \boldsymbol{d_1} & \boldsymbol{d_2} & \cdots & \boldsymbol{d_n} \end{array} \\ \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \end{array} \quad (3)$$

特徴語・文書行列の第j列はj番目の文書に関する情報を表し,同様に,特徴語・文書行列の第i行はi番目の特徴語に関する情報を表しているベクトルである.

2.4 特異値分解によるクラスタリング

ベクトル空間における特徴語ベクトルの次元は,文書から抽出される特徴語の総数と等しいので,文書数が増えると,特徴語の数も膨大となり,計算機のメモリに入り切れなくなるだけでなく,文書に含まれる不必要な特徴語がノイズにもなる.そこで,本研究では,特徴語・文書行列Aを特異値分解することで,特徴語ベクトルの次元を削減し,文書と特徴語を表現するための最適な非排他的クラスタを求める.

いま,特徴語・文書行列Aの特異値分解：$A = U\Sigma V^T$は計算済み,行列U, Vが得られたとすると,行列Uの最

†東北大学大学院情報科学研究科
‡日立製作所

初の k 個 ($k = rank(A)$) の列ベクトル $\{u_1, u_2, \cdots, u_k\}$ は文書 (検索結果) クラスタであり，行列 V の最初の k 個の列ベクトル $\{v_1, v_2, \cdots, v_k\}$ は特徴語クラスタである．また，文書クラスタ u_i と特徴語クラスタ v_i は互いに双対であり，同じトピックに関するクラスタである．いま，u_i, v_i をそれぞれ以下のようなベクトルとすると，

$$u_i = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{pmatrix} \qquad v_i = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{pmatrix}$$

クラスタ i の作成に当たって，以下の処理を行う．

- **クラスタメンバの決定およびクラスタ内文書ランキング**：文書クラスタベクトル $u_i = (\delta_1, \delta_2, \cdots, \delta_n)^T$ はすべての検索結果 $\{d_1, d_2, \cdots, d_n\}$ がクラスタ i における適合性情報を表しているので，この内，このクラスタに属さないもの，すなわち $\delta_i \leq 0.0, 1 \leq i \leq n$ となるすべての検索結果をこのクラスタから取り除かなければならない．さらに，残されたものに対して，クラスタ内ランキングを行わなければならない．具体的には，次のような処理を行えば良い．

  step1: $\{\delta_1, \delta_2, \cdots, \delta_n\}$ を降順にソートし，それを集合 Φ とする．

  step2: バイナリ・リサーチを用いて，集合 Φ の中で，最初に出現した ≤ 0.0 の要素の位置を見つけ，それとそれ以降のものをすべて削除する．よって，残されたものはランキング済みのクラスタメンバーである．

- **クラスタラベルの決定**：ベクトル v_i の要素をチェックすることによって，特徴語集合 $\{t_1, t_2, \cdots, t_m\}$ の中に，うえで作成したクラスタと適合性が最大となる特徴語 t_j を判明できる．なお，本システムでは，適合性が 1 位から 3 位までの特徴語をクラスタのラベルとして用いる．

上述した処理を k 回繰り返すことによって，見つかったすべてのクラスタを生成する．

3. 実験結果および考察

いくつかキーワードを検索エンジン Google に入力し，出力されたデータのクラスタリング処理を行った．以下一例として，キーワード { **牛タン**} に対する 200 件のクラスタリング結果を示す．

表 1: 生成されたクラスタ

検索キー	クラスタ数	見つかったクラスタの例
牛タン	20	仙台名物・牛タン店 情報誌・価格・牛タン業界 牛カルビ・牛タン骨付き

表 2: クラスタのラベルとメンバー

クラスタのラベル	上位 5 件のメンバー
仙台名物・牛タン店	ののほんのにしかた.... [ぐるなび] ヘルシーな牛... 牛タンのチャーハン 岩手日報のニュース 牛タン利休
情報誌・価格・牛タン業界	牛タンのつけ焼き レンジで牛タン 山陽新聞社ホームページ 牛タンなど対日輸出急増... Pocket Warmer(絵日記)

表 3: google とクラスタ内とのランキングの比較

牛タンに関する検索結果	Google	クラスタ内
"ののほんのにしかた...."	54	1
"[ぐるなび] ヘルシーな牛..."	37	2
"牛タンのチャーハン"	189	3
"岩手日報のニュース"	102	4
"牛タン利休"	124	5

上の実験例では,20 個のクラスタが生成され，それぞれのクラスタに含まれる検索リストの長さの平均は 10 である．また，いくつかのキーワードによる実験後,google とのランク付けの比較により，本システムではいくつかのクラスタに含まれるページが,google では分散的に配置されていることを確認した．

4. まとめ

本システムを用いると，膨大な検索リストが短縮され，また，複数のクラスタラベルの表示やクラスタ内のランキング導入により，ユーザの閲覧効率が向上したことを確認した．今後の検討課題として，計算時間が全体の 5 割以上を要する特徴語抽出部分の改良である．

参考文献

[1] D.Cutting,D.karger,J,Pedersen,J.W.Tukey. Scatter/Gather:A Cluster-based Approach to Browsing Large Document Collections. *Proceedings of the 15th Annual International ACM/SIGIR Conference*, 1992.

[2] Yitong Wang, Masaru Kitsuregawa. On Combining Link and Contents Information for Web Page Clustering. *13th International Conference on Database and Expert Systems Applications*, page 902–913,2002.

[3] 日本語形態素解析システム Chasen 「茶筌」 <http://chasen.aist-nara.ac.jp>.