1T-8

MCによるWeb検索結果のクラスタリング手法

宮本　悠生†　　小柳　滋†

立命館大学大学院　理工学研究科†

1. はじめに

近年，インターネットの爆発的な普及により，Web 上には膨大な量の情報が氾濫している．このためユーザは Web 上の情報を入手するために検索エンジンを利用している．しかし，従来の検索エンジンでは大量の検索結果の一次元リストをユーザに返しているだけであり，又，検索結果のランクが近いページ同士の間には必ずしも関連性がある訳ではない．

そこで本研究では検索結果の閲覧性の問題に着目し，Web 検索結果のクラスタリングと視覚化により，ユーザの閲覧効率を改善するシステムの構築を試みた．マトリクスクラスタリング（以下 MC とする）を用いることで，現存の Web 検索結果クラスタリングアルゴリズムが持つ，クラスタラベルの重複やラベルに無関係な文書の混在という欠点を回避することができる．又，生成されたラベル付きクラスタを視覚化することにより，閲覧による問題を解消することが期待される．

2. MC の特徴

MC とは，大規模な疎行列から，非ゼロの割合の高い部分行列，つまり密な部分行列を抽出する手法である．これの目的は，関連する行と列の集合を同時に求めることである．MC のアルゴリズムにはいくつかある[1][2].

ピンポン法は任意の行又は列を指定し「行から列の選択」と「列から行の選択」により，不要な列・行を切り捨てることで，部分行列を抽出する手法である[1].

DMC は行と列を指定し，類似度の低い行又は列を段階的に削除することで，密な部分行列の抽出を行う手法である[2]．しかし，ピンポン法は初期行が部分行列に含まれない問題がある．DMC には行と列の双方を指定する必要があるため，検索結果のクラスタリングの様に初期行又は列の指定を行えない場合は適用できないといった問題がある．

3. ラベル付きクラスタリング

3.1 特徴語・文書行列の作成

検索結果のページに対して，MC を行うための前処理として，特徴語・文書行列の作成を行う．この行列は特徴語集合と文書集合の関係を表した行列である．

特徴語集合を得るために，検索結果のページの集合である $P=\{p_1,p_2,\cdots,p_n\}$（p_i は検索結果のあるページ）に対し，茶筌による形態素解析を行う．

形態素解析の結果，品詞が名詞句，未知語句と判定された単語を特徴語とする．しかしながら，これでは単語数が膨大な量となるため，これらの特徴語候補に対し，TF（文書中の単語の出現頻度）の小さい単語をノイズと見なし，切り捨てる．その後に TF・IDF 法による重み付けを行う．これは TF と IDF（単語が含まれる文書数の逆数）との積で，この値が大きいほど，特徴語としての評価が高いことを意味している．これにより特徴語集合 $T=\{t_1,t_2,\cdots,t_n\}$ が決定される．さらに文書長による重みの影響を無くすために，重みの正規化を行う．これによって行をページ集合，列を特徴語集合，要素をこの重みの値とした特徴語・文書行列が得られる．

3.2 密部分行列の選択手法

前章で述べた MC アルゴリズムの問題点を踏まえた上で，ラベル付きクラスタリングの要求を満たす新たな MC アルゴリズムを考案した．これは，ページのクラスタリングという観点からあるページに類似したページの集合の抽出，ラベル語の貼付という観点から重みの要素の値が大きい特徴語の抽出という 2 つの要求を満たすものである．又，前者の類似性の観点がより重要であると考えたため，初期指定はページ集合である行のみとし，行と列それぞれに異なる計算式を設けた．

●行のスコアの計算手法

初期指定したページに類似している程スコアが大きくなる様，共分散による計算式を用いた．rw_i は行のスコア，$m_{i,j}$ は行列の要素，p_a は初期行，p_i はその他の行，$\overline{p}_a$ と $\overline{p}_i$ は各々の行の相加平均，X は列の要素数である．

$$rw_i=\sum_{j=1}^{X}\left(m_{a,j}-\overline{p}_a\right)\left(m_{i,j}-\overline{p}_i\right) \qquad \text{式（3.2.1）}$$

※最大値 $rw_{\max}$ の 10%以下の値を持つ行は切り捨て

●列のスコアの計算方法

初期指定した行の重みの値が大きい程，スコアが大きくなる様な計算手法を採る．cw_j は列のスコアであり，Y は行の要素数である．

$$cw_j=\sum_{i=1}^{Y}m_{a,j}\left(m_{i,j}-\overline{p}_i\right) \qquad \text{式（3.2.2）}$$

※最大値 $cw_{\max}$ の 10%以下の値を持つ列は切り捨て

これらの式を用いて 2 章で述べた DMC の様に不要な行・列の切り捨てを行うが，本手法は DMC と異なり初期行を指定するだけよい．但し，この計算手法を用いると停止時には初期指定した行とその行の中の最大の要素

A Clustering Technique of Search Engine's Result by Matrix Clustering Algorithm

† Yuki Miyamoto, † Shigeru Oyanagi

† Department of Computer Science, Graduate School of Science and Engineering, Ritsumeikan University

を持つ列のみが抽出されることになる．そこで終了条件は部分行列の面積とした．図 1 にその動作例を示す．

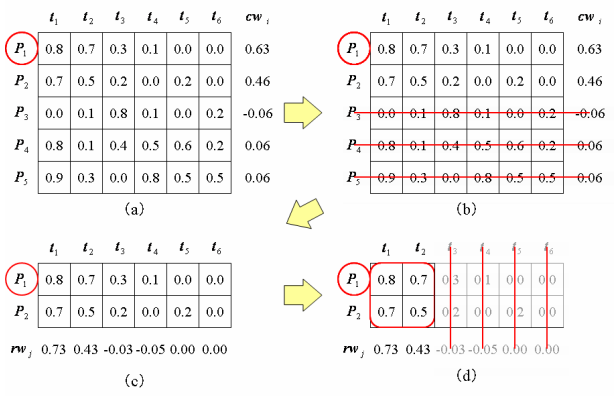

図 1： アルゴリズムの動作例

3.3 ラベリングと部分行列の統合

生成された部分行列に対して，ラベルを貼付する．MC を用いたことにより，行は初期活性化したページに類似するページ集合が，列はそれらのページが持つ，類似し且つ，重みの値が大きい特徴語集合が抽出される．各特徴語集合の重みを基にクラスタに適するラベル語を決定する．又，その際，第 2 ラベル語も決定する．なお，タイトルに含まれる単語の重みは，コンテンツ内の TF で得られる最大の重みの 1 割としている．

但し，MC は初期活性化したページによって出力行列が異なる特徴があるため，全ページに対して MC を行うと，部分行列の重複といった問題が生じる．そこで先の計算で得たラベル語を頼りに，重複するラベル語の部分行列同士を統合し，1 つのクラスタとした（図 2）．

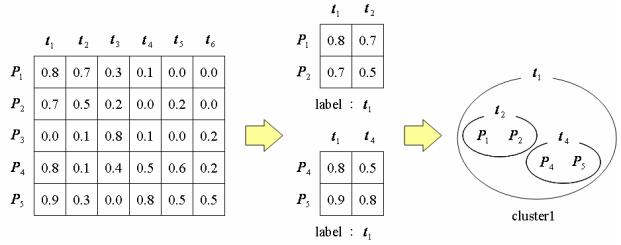

図 2： ラベルによる部分行列の統合

3.4 クラスタの視覚化

前章にて生成されたラベル付きクラスタに対して，視覚化を行う．視覚化モデルの例を図 3 に示す．

図に関しては，要素数の大きな親クラスタを中心に配置し，以下，親クラスタが含むページ数の順にその周囲に配置した．親クラスタの中の子クラスタに関しては，前節で求めた第 2 ラベル語（図 2 では t_2，t_4）を基に，親クラスタと同様の手法で配置した．子クラスタを持たない親クラスタや子クラスタの参照時には，ページのみを表示するようにした．

4. 実験と考察

実験データとして，Google で「草津」と検索した結果の上位 100 ページを取得し，3.1 節の手法により特徴語数 3073 語，密度 2.4%の行列を構成し，3 章のクラスタリングアルゴリズムを適応した結果，生成クラスタ数は 27 個となった．クラスタのラベルの例を表 1 に，「スキー」クラスタに含まれるページ群を表 2 に示す．順位は検索結果の出力順である．なお，ラベル付きクラスタリングの実行時間に関しては，Pentium Ⅳ 3.2GHz の PC 上で 0.8 秒であった．

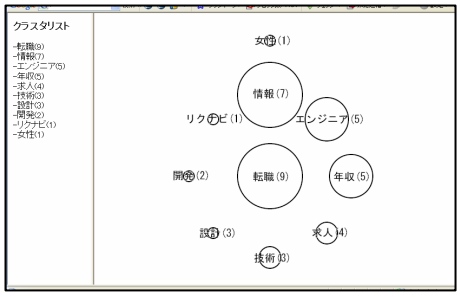

図 3 ： 視覚化モデルの例

表 1 ： 生成されたクラスタのラベルの例

スキー	リゾート	ホテル

表 2 ： 「スキー」クラスタに含まれる URL

順位	URL
5	http://www.kusatsu-kokusai.com/winter/index.html
13	http://www21.cx/kusatsu/
75	http://www.kusatsu-kokusai.com/green/index.htm
78	http://www.tokyo-np.co.jp/event/knsf/
91	http://kusatsuski.or.jp/kss/

「スキー」クラスタに関しては，100 ページの中でタイトル語に「スキー」を持つページは全て含まれていた（順位 5,75,78,91）．又，コンテンツ内に「スキー」の単語を最も多く持つページ(順位：13)も含まれていた．

表 1，2 の結果から，順位の低いページであってもラベル語を頼りに素早くアクセスできることが示される．以上の結果から提案する手法は有効であると考えられる．

5. おわりに

考案した MC アルゴリズムにより，ページ集合とラベル語を同時に求めることでラベル語の重複を解消したラベル付きクラスタリングを行った．しかしながら，閾値によって出力行列が変化するため，理想的なクラスタを生成するための閾値決定手法を考える必要がある．

又，ユーザビリティ評価手法を用いることで，検索結果のさらなる閲覧性の向上を果たすことが課題として挙げられる．

参考文献

[1] 小柳滋,久保田和人,仲瀬明彦： “Matrix Clustering : CRM 向けの新しいデータマイニング手法”， 情報処理学会論文誌， Vol42, No8, pp.2156-2166， 2001

[2]上原子正利， 小柳滋： “行列上の相互順序決定”， 情報処理学会論文誌, データベース, Vol.SIG13(TOD27), pp.16-25, 2006