3P-7

各単語と検索語の出現位置を考慮したウェブページのクラスタリング

秋山　龍太郎† 金盛　克俊† 大和田　勇人†
東京理科大学理工学部経営工学科†

1. はじめに

検索エンジンによる検索結果のウェブページ群をページの内容ごとにまとめることで，ユーザーが目的のページを探すことが容易になる．これを行うために，検索結果のウェブページ群をベクトル空間法[1]を用いてベクトル化し，類似度を計算してクラスタリングする手法[2,3]が提案されている．しかし，一般的なベクトル空間法を使用するだけでは3つの問題点が挙げられる．

1つ目はクラスタリングの精度が低いことである．ベクトル空間法を使用する際に単語の重要度を計算する必要があるが，一般的に使用されるtf-idf法では精度が低い．新たな重要度を用いてクラスタリングの精度を上げることが望ましい．

2つ目は計算コストが大きいことである．全てのページで1回でも出現した単語全部をベクトル空間法に用いると単語数が膨大になってしまう．単語数を減らすことで，計算コストを下げることが望ましい．

3つ目はユーザーが各クラスタの内容を知るには，クラスタ内の数ページをチェックした後に判断しなければならないことである．単にクラスタリングするだけでは各クラスタの内容はわからないので，各クラスタの内容を表す単語を自動的に提示するのが望ましい．[3]

本研究ではこれらの問題点を改善する手法を提案し，検索結果のウェブページ群を内容別に自動分類するシステムを構築する．

2. 提案手法

本研究では，クラスタリングの精度が低い問題に対して，検索語の近くに出現する単語ほど検索語の意味を表わしていて重要であると考え，tf-idf 法にテキスト上での検索語までの距離を組み合わせた単語の重要度の指標を使うことで解決を図る．

Clustering of the web page considering the appearance position of each word and the search word
†Ryutaro Akiyama, Katsutoshi Kanamori, and Hayato Ohwada・Dept. Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

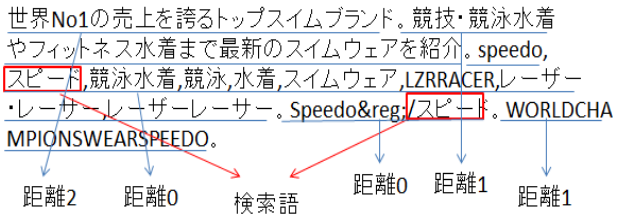

図1　検索語「スピード」における距離の例

また，計算コストが大きい問題に対して，検索語の近くに出現する単語のみでもページの内容を表わすことができると考え，使用する単語を検索語の近くのもののみに限定することで，クラスタリングの精度をあまり下げることなく計算コストを下げる．

さらに，ユーザーが各クラスタの内容を判断するのに手間がかかるという問題に対して解決を図る．各クラスタを表わす単語はクラスタ内の半分以上のページに出現していると考え，クラスタリング後に各クラスタ内で半分以上のページに出現し，かつ重要度が最も高い単語数語を提示する．

本研究で提案するシステムは5つの処理から構成される．各処理について順に各項で説明する．

2.1 ウェブページの取得

日本語であり，かつ意味的な曖昧性を持つ検索語による検索結果の上位100件の各ウェブページのHTML形式のソースコードを取得し，HTMLタグを外すことでテキストにする．

2.2 不要語の削除や距離の計算

まず，各テキストに対して形態素解析を行い，名詞以外の単語を不要語として削除する．

次に，検索語までの距離を計算する．距離は文ごとに計算する．検索語を含む文に含まれる単語は距離0，検索語を含む文の隣にある文に含まれる単語は距離1，さらにその隣にある文に含まれる単語は距離2とする．図1に検索語「スピード」における距離の例を示す．

最後に，距離が遠い単語を削除する．

クラスタ1: テスト 回線 測定
クラスタ2: 映画 ヤン ボン
クラスタ3: 富士 スピードウェイ レーシング
クラスタ4: トランプ ゲーム カード
クラスタ5: その他

図2 クラスタを表す単語の例

2.3 単語の重要度の計算

単語の重要度を tf-idf にテキスト上での検索語までの距離を組み合わせた指標により計算する．以下にページPにおける単語Tの重要度の式(1)を示す．Rは検索語までの距離を表している．

$$v(P,T) = \text{tf-idf} \times \left(\frac{1}{R+1}\right)^{1.5} \quad (1)$$

2.4 ウェブページのクラスタリング

まず，計算コストを下げるために，重要度の高い単語 200 語のみを選択してベクトルの軸とし[3]，各ウェブページをベクトル化する．

次に，ウェブページ間の類似度をコサイン類似度により求め，クラスタリングを行う．類似度が高いものからまとめていき，類似度の閾値αを超える範囲でクラスタをつくる．また，ページ数が2ページ以下のクラスタは1つにまとめる．これにより，クラスタ数が多くなり過ぎず，見やすくなる．

2.5 クラスタを表わす単語の抽出

各クラスタ内で半分以上のページに出現し，かつ重要度が高い上位3つの単語をクラスタの特徴を表わす単語として抽出し，ユーザーに提示する．また，まず最も重要度が高い単語 1 語を提示し，その単語から内容が判断できないとユーザーが判断した場合に残り 2 語も提示するようにする．その理由としては，提示される単語が多いことでユーザーが内容を判断するのに混乱することを防ぐためである．また，ページ数2ページ以下のものをまとめたクラスタは「その他」と提示する．提示された単語をクリックすることで，対応するクラスタ内のページが表示される．図2にクラスタを表す単語の例を示す．

3.実験

従来手法と提案手法のクラスタリング精度を Purity/Inverse Purity 指標の F 値の最大値により比較することで，提案手法の有効性を示す．本実験の検索語は「スピード」，「トレーナー」，「ワンピース」を用いた．

従来手法は全単語を用いて単語の重要度の式に tf-idf 法を使用したものを用いた．提案手法は検索語までの距離が 20 以下の単語のみを用いて単語の重要度に式(1)を使用したものを用いた．

表1 実験結果

		従来手法	提案手法
スピード	精度	84.36	86.46
	単語数	58100	44940
トレーナー	精度	66.94	70.71
	単語数	64724	45361

表1より，提案手法の方がクラスタリングの精度が高く，単語数は「スピード」の場合は約 77%に，「トレーナー」の場合は約 70%になっていることから，提案手法の有効性が示された．

また，検索語「ワンピース」に関しては，どちらの手法でも1つのクラスタに全ページがある場合が最もF値が大きく，うまくクラスタリングできていなかった．その理由は，「マンガの ONEPIECE」に関するページが全検索ページ 71 件中 59 件であり，単語の重要度に idf 値を含むため，「マンガの ONEPIECE」の特徴を表わす単語の重要度があまり高くないからだと考えられる．

4.まとめ

本研究では，ウェブページのクラスタリングをベクトル空間法を用いて行う際に起こる問題を改善する手法を提案し，検索結果のウェブページ群を内容別に自動分類するシステムを構築した．実験から，従来手法より検索語までの距離が 20 以下の単語のみを用いて単語の重要度に検索語までの距離を考慮した式(1)を使用する手法の方がクラスタリングの精度が高く，計算コストが低いことがわかった．

今後の展望として，本研究の実験で新たに問題となった，検索結果の多くを同じ意味を表わすページが占める場合うまくクラスタリングできないことに対して，解決を図ることが挙げられる．

参考文献

[1] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
[2] 大谷紀子，″情報検索におけるベクトル空間モデルの応用″，武蔵野工業大学環境情報学部研究論文 pp.3-6，2004.
[3] 城市広大，三好力：″ベクトル空間法とファジィ推論を用いた WEB 検索結果自動分類システム″，知能と情報(日本知能情報ファジィ学会誌)Vol.18,No.2,pp.184-195(2006)