F-018

クラスタ抽出によるWebの検索結果の分類
Classification of Web Search Result by Cluster Extraction

持田 広志† 大町 真一郎† 阿曽 弘具†
Hiroshi Mochida Shinichiro Omachi Hirotomo Aso

1. はじめに

近年インターネットの普及(図1)により誰でもどこでもいつでもインターネットに接続することが可能になりつつある．また，それとともにインターネット上に存在するウェブページの数は爆発的に増加し，内容も多岐にわたっている．これらにより，インターネット上で検索エンジンを利用して調べ物をすることが日常的になってきている．しかし，情報が多いがゆえに検索結果は玉石混淆で，有用な情報を得られるページもあればまったく意味のないページも得られてしまうのが現状である．

普段われわれが利用している検索エンジンは入力したキーワードを含むページをWeb上から集めてきて索引を作るもので，とくにページに編集や加工を加えるものではない．Google[3]では検索した結果を多くのページからリンクされている順に表示するようになっており，また各ページの要約を載せることによりユーザーの検索の支援をしている．しかし，それでも検索結果から目的のページを見つけ出すのはユーザーの役目で，多くの負担を強いられている．検索エンジンでキーワードを検索した際にキーワードが複数の意味や使われ方をしているとその複数のページの検索結果が混在することとなり，目的のページを見つけ出すことがより困難になる．検索の際，複数のキーワードや文章をキーワードに用いて検索する方法もあるが，どうすれば適切に検索できるかを考えることはユーザーにとっては悩ましいことが多い．

インターネットでの検索の利便性の向上のためにキーワードの意味や用法ごとにウェブページをクラスタリングすることが有用であり，過去の研究としてZamirらの研究[2]やMSNのウェブページクラスタリング[5],vivisimo[6]などがある．ウェブページを内容でクラスタリングする際，一つのクラスタに一つのトピックのウェブページのみが分類されるわけではなく，違う内容のウェブページが一緒に分類されていることがある．よってこうした内容の違うウェブページをクラスタから除去する必要がある．本研究ではこのような内容の違うページを除去する方法を提案する．

2. Webページのクラスタリング

まず，ウェブページを，含まれている単語の情報を用いてクラスタリングする．次にトピックの違うページを除去することで目的のページ群を得る．具体的な手法は以下の通りである．

1. Googleで目的の単語を検索する．
2. 検索結果の上位M件のウェブページを収集する．
3. 各ウェブページ内の単語の出現頻度から特徴ベクトルを作成する．

†東北大学大学院工学研究科

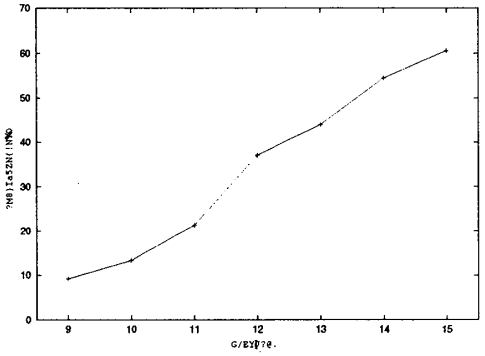

図1: 日本のインターネット人口普及率

4. 得られた特徴ベクトルを用いてK-means法[1]によりウェブページのクラスタリングを行う．
5. クラスタごとに，クラスタの中心からの距離をもとにトピックの異なるページを除去する．

2.1 特徴ベクトル作成法

ウェブページにはその内容に固有の単語が使われている．従って，ページ内の単語の出現頻度を調べることにより内容やトピックを分類できると考えられる．そこで，単語の出現頻度を要素とするベクトルをページの特徴ベクトルとする．その作成法は次の通りである．

1. HTMLから不要なタグやスクリプト，ヘッダ部分を除去する．
2. 残った文章を茶筌[4]を用いて単語に分割する．その際，名詞や動詞を残し，助詞や助動詞は削除する．
3. ウェブページごとに各単語の出現する頻度を調べ，特徴ベクトルとする．
4. 単語の出現頻度の上位の累計頻度が全単語の数の60％となる単語を残し他の単語は除去する．
5. 特徴ベクトルを正規化する．

タグやヘッダを除去する際にすべてを除去するのではなく，TITLEタグなどウェブページの内容と関係の深い部分はタグたけ除去し，単語は残しておく．茶筌を用いて文章を単語に分割した際，英語も単語に分割しておく．

4番目の作業について詳しく説明する．単語の出現頻度で全出現単語をソートし，その上位からの累計頻度を計算する．累計頻度が全単語数(最下位までの単語の累計頻度)の60％となる上位の単語を残し，残りの出現頻度が少ない単語を除去する．これにより単語数を減らすことができ，特徴ベクトルの次元数を減少させることができる．なお６０％という値は実験的にクラスタリングに影響がでない割合として求めたものである．

2.2 K-means法

K-means法はベクトルをクラスタリングする代表的な方法である. 具体的なクラスタリングの流れは次のようになる.

1. ランダムに選んだk個のベクトルを初期クラスタ中心(seeds)とする.
2. 各ベクトルを, クラスタ中心との距離が最も小さいクラスタに属させる.
3. 各クラスタに属するベクトルを用いて, そのクラスタのクラスタ中心を再計算する.
4. クラスタ中心の位置がほぼ定まり, 収束判定条件を満たすまで, ステップ2から 繰り返す.

2.3 トピックの違うウェブページの除去

トピックが違うウェブページは内容が大きく違っているはずで, その内容の違いは特徴ベクトルに現れると考えられる. すなわち, トピックの違うページの特徴ベクトルは同一クラスタ内の他のウェブページの特徴ベクトルとは離れている可能性が高い. そこで, クラスタの重心から遠く離れている特徴ベクトルをもつページをトピックが違うページとして除去する.

1. クラスタの重心から一番遠い特徴ベクトルをクラスタから除去する.
2. 残ったベクトルで改めて重心を計算しなおす.
3. 重心を更新した際の重心の移動距離を求める.
4. 重心の移動距離があるしきい値以下になったら直前に除去した特徴ベクトルをクラスタに加えて終了する. しきい値以上の場合はこの作業をしきい値以下になるまでくり返す.

3. 実験

3.1 手法

提案手法の有効性を確かめるため実験を行った. 実験データとして,「ジャガー」という単語をGoogleで検索した結果の上位200件を用いた. 200件のウェブページに対して特徴ベクトルの作成を行い,k-means法によりウェブページを5つのクラスタに分類した. その後, 各クラスタに対して中心から遠いページを除去した.

実験結果の評価においてページの内容は人の目で見ておおよそどういうトピックかということを判断している.

3.2 結果

「ジャガー」の検索結果でクラスタリングを行った結果, 5つのうち2つのクラスタには半数以上同じトピックのウェブページが含まれていた. 残りの3つのクラスタはどのトピックが有力ということは言えなかった. 以下では上の2つのクラスタを対象に考察する.

ページ除去前のクラスタの内訳と除去後の内訳を表1に示す. 1つのクラスタには「ジャガー」という車に関係したウェブページが多く含まれていたが半分近くは車とは関係ない漫画や音楽のウェブページが混ざっていた. (表1(a)). ここからウェブページの除去を行った結果,30ページが除去された. そのうち8ページが車に関係したウェブページで残りの22ページがその他の内容であった.

表1: ウェブページ除去前と除去後のクラスタの内訳

検索語	クラスタ内訳	除去前	除去後
ジャガー (a)	車	42	34
	それ以外	42	20
ジャガー (b)	少年野球	12	10
	それ以外	9	1

もう1つのクラスタは少年野球のチームに関係したウェブページが多く含まれていた. (表1(b)).(チーム名がジャガーズであった).

ここからウェブページの除去を行った結果,10ページが除去された. そのうち2ページが少年野球のチームに関係したウェブページで残りの8ページがその他の内容であった.

本来除去されるべきではない車に関係したウェブページや少年野球のチームに関係したウェブページまで除去されている. しかしこのウェブページの除去を行ったことによりトピックの違うウェブページのクラスタ内に占める割合が除去する前は50 %と43 %だったのに対し, 除去後は37 %と9 %まで減ったので各クラスタを人の目で見た場合, クラスタリングされているトピックはわかりやすくなったといえる.

4. まとめ

ウェブページのクラスタからトピックの違うページを除去する手法を提案した。実験には一つのクラスタに複数のトピックのウェブページがクラスタリングされた場合に違うトピックのウェブページを除去できることを示している. ウェブページを除去することでひとつのクラスタにひとつのトピックのページが集まる割合が高くなり情報探索の際に有効であることがわかった.

同一クラスタの中でページ数が多いトピックのページを抽出できたとも言える. あまりページ数が多くないトピックのウェブページの抽出も必要であり, 除去したページを新たにクラスタリングするなど, 少数のトピックのウェブページの扱いを考える必要がある.

参考文献

[1] Duda, R. O., Hart, P. E., and Stork, D. G. "Pattern Classification," John Wiley and Sons (2000).

[2] Zamir O., Etzioni O. "Web Document Clustering: A Feasibility Demonstration," SIGIR'98 (1998).

[3] Google search engine "www.google.co.jp"(2005).

[4] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原 正幸, "日本語形態素解析システム『茶筌』version 2.2.1 使用説明書", (2000).

[5] H. J. Zeng, Q. C. He, Z. Chen, W. Y. Ma, J. Ma "Learning to Cluster Web Search Results," SIGIR'04 (2004).

[6] vivisimo "http://vivisimo.com"(2005).