4Q-03

HTML文書構造を利用したWeb検索結果クラスタリング手法の有効性について

阿部　寛之[†]　松原　雅文[†]　Goutam Chakraborty[†]　馬淵　浩司[†]

岩手県立大学ソフトウェア情報学部[†]

1. はじめに

知らない単語の語義を調べる際に用いる国語辞典などには新出単語や専門用語が収録されていないことが多い．しかし，インターネットにはそういった情報でも多く流通しているため，検索エンジンを用いた調査は日常化している．

しかし，個人からの情報発信が容易になりインターネットの規模が拡大しているため，目的とする情報に到達することが困難になってきている．特に，ある単語の語義を調べる際に，複数のWebページが該当した場合，そのいずれかを選択し，ページをスクロールしたりリンクを辿ったりして閲覧しなければ，目的の情報を得ることができないことがある．そのため，余計な時間を割くことになってしまい，検索効率が低下する．

本研究では，Webページ内のリンク数などのHTML構造を利用したクラスタリングを行い，対象とする検索クエリについてスニペットを自動生成することで，情報検索効率を向上させることを目的とする．この時，もとから存在する検索クエリを辞書的に定義しているWikipediaのようなWebページは，他のWebページと比べ情報の質が異なるため，別クラスタに分類することも目的としている．

本稿では，複数の語義をもつ検索クエリに対し，X-means法とHTML構造を利用したハードクラスタリング手法を提案し，行った評価実験の結果から提案手法の有効性と今後の展望について述べる．

2. 先行研究

2.1. X-means法

クラスタリング方法の1つであるK-means法はクラスタ内に含まれる要素の座標の平均を用い，指定したクラスタ数をK個に分類する手法として知られている．K-means法には問題点の1つにクラスタの個数Kを指定しなければならない点がある．

そこでPelleg and Moore[1)]は，K-means法の逐次繰り返し処理とベイジアン情報量推定による分割停止基準を用いて最適なクラスタ数を決定している．この手法は前提とする情報がない場合に，発見的な方法に頼ることなく最適なクラスタ数を求めることができる．

3. 提案手法

3.1. 概要

K-means法におけるクラスタの個数をあらかじめ指定しなければならない問題を踏まえ，検索結果のクラスタリングにはX-means法を採用する．提案手法のアルゴリズムを3.2.〜3.5.に示す．

3.2. Webページ取得

ユーザが入力したクエリについて，Google検索結果上位N件分のWebページを取得する．

3.3. Webページから本文を抽出

取得したWebページを構成するHTMLを解析し，Webページの記事部分からタグなどを除去しながら本文データを抽出する．同時に，HTMLの記事部分のみに出現するリンクタグの数をカウントする．

3.4. 文章間類似度の計算

本文データ内の名詞について，全文書の1割以上出現する名詞に対し，tf-idf値を計算する．計算したtf-idf値を用いて，文章間のユークリッド距離を求め，文章間類似度とする．

3.5. クラスタリング

X-means法でクラスタリングし求めた各クラスタ内の文書について，文書内の名詞数に対するリンク数の割合が50%以上のWebページは別のクラスタとして分類する．

このようにして，ユーザが入力した検索クエリに関するWebページを語義ごとにハードクラスタリングする．

Effectiveness of Web Clustering Method using HTML Document Structure

Hiroyuki ABE[†], Masafumi MATSUHARA[†], Goutam CHAKRABORTY[†], Hiroshi MABUCHI[†]

[†]Faculty of Software and Information Science, Iwate Prefectural University

表 1: Web ページのクラスタリング結果と再現率

検索クエリ	クラスタ数		再現率 (%)
	正解	結果	
リーダー	3	3	89.66
エージェント	2	2	92.25
ライター	2	2	88.75
トランプ	2	2	95.50
コード	2	2	90.50
リード	4	4	90.50
レバー	2	2	89.50
ラバー	2	2	89.50
フレイム	4	4	87.25
カート	2	2	95.50
再現率の算術平均値 (%)			90.89

4. 実験

4.1. 実験条件

本実験では，2016 年 11 月 28 日における Google 検索結果上位 50 件分の Web ページ本文データを用いる．この時，Google 検索オプションのパーソナライズ検索[1]は無効にしている．検索クエリは，複数の語義をもつ「リーダー」,「エージェント」,「ライター」,「トランプ」,「コード」,「リード」,「レバー」,「ラバー」,「フレイム」，そして「カート」の 10 単語を設定した．開発言語には Python 3.5.2，形態素解析器に MeCab，そして形態素解析用の辞書には mecab-ipadic-Neologed[2]を使用し，実験を行った．

4.2. 評価方法

クラスタリング結果の評価には，各クラスタごとに求めた再現率の調和平均値を用いる．

4.3. 実験結果と考察

HTML 構造を利用せず，X-means 法のみでクラスタリングを行った場合において得られたクラスタ数と再現率の結果，そして全クエリの再現率の算術平均値をまとめたものを表 1 に示す．

再現率の算術平均値から検索クエリ 10 単語すべてにおいて高い再現率が得られた．また，検索クエリが「トランプ」と「カート」の場合については 95%以上の結果が得られ，ほぼ正しくクラスタリングを行えた．

しかし，目的の 1 つとしていた検索クエリについて辞書的に定義している Web ページについては，記述されている文章量によっていずれかに分類され，別クラスタに分類することができなかった．

表 2: リンク数の割合を取り入れたクラスタリング結果

検索クエリ	クラスタ数		再現率（%）
	正解	結果	
リーダー	4	4	95.50
エージェント	3	3	100.00
ライター	3	3	95.50

そこで，提案手法に従い Web ページから本文を抽出する際に同時にカウントしたリンク数の割合を用いたクラスタリングを行う．クエリについて辞書的に定義している Web ページやショッピングサイトのような Web ページは，Web ページ内のすべての名詞数に対するリンク数の割合が半分以上になる傾向があるからである．

表 2 は，リンク数の割合が 50%以上だった場合，新たなクラスタを生成し，別のクラスタとして分類した結果である．X-means 法のみでクラスタリングを行った表 1 の結果と比べ，すべての検索クエリで再現率が向上している．特に，検索クエリが「エージェント」の場合については再現率 100%という結果が得られ，本手法によって正確にクラスタリングが行えることを示せた．

2 つの実験結果から，X-means 法と Web ページ内に名詞に対するリンク数の割合によるクラスタリングが有効であることが確認された．

5. おわりに

本稿では，検索結果の Web 文書をハードクラスタリングした後に，名詞に対するリンク数の割合を用いて再分類する手法を提案し，実験結果から複数の語義を含む検索クエリに対して有効であることが示された．

実験で用いた検索クエリは，複数の語義をもつ単語として設定したものであるため，今後はユーザが日常的に入力する検索クエリに対しても本手法が有効であることを示すため，実験を行っていく予定である．

謝辞

本研究の一部は JSPS 科研費 15K00155 の助成を受けたものである．

参考文献

1) Dan Pelleg and Andrew Moore. "X-means: Extending K-means with Efficient Estimation of the Number of Clusters". ICML-2000, pp. 727-734, 2000.

[1] https://support.google.com/websearch/answer/1710607?hl=ja

[2] https://github.com/neologd/mecab-ipadic-neologd