5ZB-1

文末の表現に着目した閲覧者が受ける印象による Web 文書のクラスタリング

包 直也† 松本 章代† 鈴木 雅人†
東京工業高等専門学校†

1. はじめに

現在 Web 上には膨大な量の情報が存在しており、必要とされる情報を選別したり検索する技術が数多く開発されている。既存の方法としては、カテゴリを作りページを登録して情報を分類する方法や、ニュース検索やブログ検索など、特定のタイプの情報を選別して自動的に登録する方法などがある。本研究では、既存の分類方法とは違う"ページの雰囲気"や"意見の多さ"、"解説の多さ"といった、閲覧者が受ける印象を指標とした情報のクラスタリング方法の開発を目指す。このシステムの応用としては、Web 検索の際に読みやすい砕けた雰囲気のページを特定する機能や、ブログの記事の意見性または解説性の強さを自動的に判断して表示する機能などが考えられる。本稿では、このクラスタリングの指標として、日本語の文末の表現に着目する。

2. 閲覧者の受ける印象と文末の表現

本研究で、閲覧者が受ける印象の主観的な判断の基準として設定するのは、A.雰囲気がどの程度砕けているか、または硬いか、B.筆者の意見がどのくらい含まれているか、C.解説がどのくらい含まれているか、の 3 つである。これらの情報は、対象の文書中の文末の表現に特徴として現れることが予想される。例としては、次のような場合が挙げられる。

A.～じゃん。～だよね。 (砕けている表現)
～である。～でございます。(硬い表現)
B.～ではないだろうか。～でしょ。(意見を述べている場合)
C.～である。～している。(物事を解説している場合)

よって、対象となる文書内の文章を既存の形態素解析器を用い解析した結果から、単語とその品詞情報に分解して文末の付属語についてのデータを解析することで、Web 文書をクラスタリングすることが可能であると考えられる。また、これらの文末の表現は、文章で述べられているテーマや種類によらないことから、A-C の閲覧者が受ける印象を文書で述べられる事柄によらずクラスタリングできるものと考えられる。

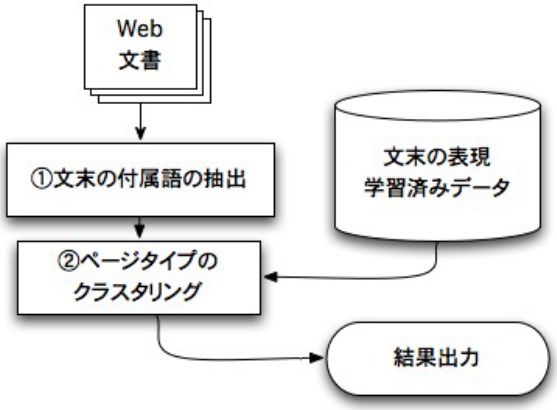

図 1：システムの処理の概要

3. システムの概要

図 1 にシステムの処理の概要を図示する。

まず、①では、Web ページのソースの取得、HTML タグの除去文の抽出、形態素解析器を用いた文末の付属語の抽出を行う。文末の付属語とは、文末から遡って自立語が現れる直前までの１つ、または複数の付属語である。形態素解析器として MeCab を、MeCab 用の辞書として IPA 辞書を使用した。

閲覧者が受ける印象を基準としたクラスタリングでは、実際に人間が Web ページを見て評価したデータを元に機能を実装する必要がある。②では文書の閲覧者の主観的な印象をデータとして学習させた、文末の付属語のデータが格納された辞書を用い、前節で述べた項目について

A readers' impression based web documents clustering using the expression of sentences' ending

Naoya Tsutsumi † Akiyo Matsumoto † Masato Suzuki †
† Department of Computer Science, Tokyo National College of Technology

クラスタリングを行う。辞書の作成に関連する事項について 4 節で詳しく述べる。

表 1：評価項目

雰囲気の硬さ
文章の丁寧さ
解説の多さ
意見の多さ
ニュースの多さ
個人的な記録・感想の多さ

4．学習辞書の作成

4.1．サンプル文書の収集と評価

閲覧者が受ける印象をコンピュータに学習させるため、まずサンプルとなる Web ページを収集した。NTCIR-4WEB の検索クエリ 100 件を用い、Google 検索エンジンでそれぞれについて検索した結果得られた検索結果の上位 100 件、述べ 10,000 件のページを収集した。そのうち 1,000 件を無作為に選び、東京高専情報工学科 5 年の日本人学生 8 人で、それぞれ 250 ページずつ、1 ページあたり 2 人で評価し、平均を取った。評価項目は表 1 に示す 6 項目である。各項目について 4 段階で評価した。

1,000 件中 80 件は正しく評価できず、920 件を評価データとして使用した。評価できなかった理由については次の項目が挙げられる。

- ページのリンク切れ
- 文章が含まれていない
- リンク集など、他の引用の集まり

4.2．文末の付属語に現れる特徴

評価されたデータから、文末の付属語の表現に着目して、クラスタリングのためのアルゴリズムを実装する必要がある。

クラスタリングの前段階として、評価された各項目について、文末の付属語にそれぞれ特徴が現れるかどうかを確かめる。また、それぞれの項目について、特徴的な語を特定することで、クラスタリングの際に参考にする。

まず、Web 文書から日本語の文の文末の表現を抽出する。これは 3 節の図 1 に示した①の工程と共通である。

次に、抽出されたデータと、評価されたデータを用いて、文末の表現に特徴が現れるかを検証する。ここでは、特徴を表す指標として TF-IDF 値を用いる。

まず、表 1 のそれぞれの指標について 4 段階で評価されたデータの中から、最も高い評価を受けたページの集合と、最も低い評価を受けた集合に分けた。次にそれぞれの集合から付属語を抽出しそれぞれの頻度を tf 値とし、4.1 節で述べたサンプルデータ 10,000 件からそれぞれの付属語について、出現する文書数を df 値とし、TF-IDF 値を計算した。

表 2：TF-IDF 値の上位 10 件

	雰囲気が硬い	雰囲気が砕けている
1	ます	よ
2	です	だ
3	た	ね
4	だ	か
5	ね	た
6	いる	な
7	か	です
8	ん	ます
9	う	う
10	よ	てる

表 2 は、雰囲気が硬い、または砕けている、と判断された集合中の付属語の TF-IDF 値を降順に並べたときの上位 10 件である。なお、雰囲気が硬いと評価されたページは 920 件中 579 件であり、雰囲気が砕けていると判断されたのは 134 件であった。

雰囲気の評価が高いものと低いもので TF-IDF 値が高い特徴的な語に違いが見られる。例えば、”よ”に関して見ると、1 ページあたりの出現個数は、雰囲気が硬い集合では 1.45 個であるのに対し、雰囲気が砕けた集合では、13.4 個であった。従って各指標は文末の付属語の表現が関係しているのではないかと考えられる。また、意見、解説についても同様の特徴の違いが見られた。

5．まとめ

本稿では閲覧者が受ける印象という新しい指標を用いた Web 文書のクラスタリングを実現するための方法について、文末の表現に着目することで実現する方法を提案した。更に本手法の実現性について実験的に確認した。

今後、具体的なクラスタリングの手段の考案と実装、応用システムの開発を行っていく。