印象による Web ページのクラスタリング

中山 記男† 江口 浩二†‡ 神門 典子†‡
総合研究大学院大学情報学専攻† 国立情報学研究所‡

抄録
本稿では，Web ページの評判情報に着目したクラスタリング手法についての提案を行う．本研究では，Web 上のテキストに対して，テキスト中で感情や感性などの印象をあらわす語から連想される色で索引付けし，それに基づいてクラスタリングを行った．これにより，印象によって Web ページを分類することを目指す．実際に Web 上の書評 487 件に対して本手法を適用し，その精度を確認した．

1. はじめに

本稿では，Web より収集した書評を対象として，書評に含まれる語から連想された色を用いる索引付けによって Web ページのクラスタリングを行った．

Blog と呼ばれる Web サイトの形態が流行したことにより，様々な事象や製品に対して，多くの個人の意見が Web を通じて得られるようになった．この Web サイト上の意見には利用者側，開発者側ともに注目しており，実際に購入物の決定や製品開発などに利用され始めている．このような情報を評判情報と呼び，情報抽出にはそれを専門に扱う研究がある．主に意見や見方といった評判情報がどこに書かれているかをテキストから検索及び抽出し，それを分析することにより対象物自体の評価を行う研究である．

本研究では評判情報の中でも文書中にあらわれた感情や感性などの印象を対象とした．また印象を色として表現することを提案している．印象は非常に抽象的な概念であり，特定の言葉で表現することはとても難しい．本稿では，印象を示す語から連想される色を用いることによって，印象を言葉に変換して扱うよりも直感的な処理が出来るのではないかと考え，それによるクラスタリング手法を提案する．

2. システム

2.1. 適用データ

杉田ら[5]が収集した Web 上の書評に対し，本研究のシステムを適用した．文書は主に書籍等に対しての書評で構成されており，単一の書籍に対してだけではなく複数の書籍に対しての書評も含んでいる．本研究では当該データに対し，単一の書籍に関する書評のみを含む文書を手作業で選別した．最終的にデータとして扱う文書は 487 件である．

2.2. 辞書

事前に印象を示す語を登録した辞書を作成した．語は長町[6]の収集した感性ワード集と筆頭著者が収集した語からなる 1171 件である．各語には，連想される色を肯定と否定の場合に各々１色ずつ対応付けている．表１は辞書の一部である．

表 1 作成した辞書の一部

語	肯定の時の色	否定のときの色
熱い	#FF0000(明るい赤)	#FF0000(明るい青)
楽しい	#FFCCFF(明るい黄)	#333300(薄暗い緑)

2.3. 色

本研究で利用した色は，小林[7]が定義した，人間がイメージとして感じる色 130 色から類似色を排した 36 色である．また，語と色との対応付けは，36 色のカラーテーブルと語を提示し，一番妥当であると感じた 1 色を選んだ[8].

2.4. 索引付け

2.1 にて選別を行った文書 487 件から，2.2 で作成した辞書に基づいて索引を作成した．この際，索引付けは辞書に登録されている語との完全なマッチングのみにて行われ，形態素解析などは行っていない．また，索引付した語の前後 40 バイトに「ない」という表現が出現した場合にのみ，否定表現として語を扱っている．その後，同じく 2.2 の辞書を参照して，語に対応する色で文書ごとに索引付けした．

2.5. 文書間類似度の計算

2.4 の索引付けから，色数に基づき 36 次元のベクトルでコサイン類似性尺度を用いて各文書間の類似度を求めた．その際，後述のように *tf* と *idf* の計算に倣って *cf*idf* を求め，文書内での

Web Page Clustering using Impression of Documents
Norio Nakayama†, Koji Eguchi†‡, Noriko Kando†‡
† The Graduate University for Advanced Studies
‡ National Institute of Informatics

色の出現頻度に基づく重み w を計算している．

$$cf_{it} = \frac{\text{文書 } D_i \text{ における色} C_{it} \text{ の延べ出現回数}}{\text{文書 } D_i \text{ 中の全ての色の総出現回数}}$$

$$idf_t = \log \frac{N}{\text{色 } C_t \text{ が出現する文書数}}$$

$$w_{it} = cf_{it} \times idf_t$$

2.6. クラスタリング

2.5 で求めた文書間類似度に基づいて，文書クラスタリングを行った．文書クラスタリングのアルゴリズムに関しては a)単一リンク法， b)完全リンク法， c)グループ平均法 を用いた．本稿では後述の平均精度が最も高かった完全リンク法について結果を述べる．クラスタリング処理に関しては，各クラスタを併合する文書類似度の閾値を定め，併合するクラスタが生まれなくなった時点で処理を停止した．

3. 結果と評価

完全リンク法にて閾値を 0.5～1.0 まで 1/1000 刻みで変化させ，結果を求めた．クラスタ数に大きな変動があった 6 箇所の閾値での各クラスタ内文書に対して精度を判断した．精度の判定は，まず筆頭著者が各クラスタ内文書を全て読み，各クラスタの典型的な印象を理解した後に，各クラスタ内の各文書の持つ印象がそれに適しているかどうかを判断することでクラスタ精度 P を求めた．下式は精度を求めた際の式である．

$$p_{sk} = \frac{\text{閾値 s でのクラスタ}_{\mathrm{Lsk}} \text{ 内文書適合数}}{\text{閾値 s でのクラスタ}_{\mathrm{Lsk}} \text{ の文書数}}$$

その閾値での全クラスタの精度を求めた後に，クラスタ全体の平均精度を得た（表 2）．ここで全クラスタ内文書数は，文書数 2 以上のクラスタ内に存在する文書数の総計とし，同様にクラスタ数は文書数 2 以上のクラスタの数とする．

表 2　精度の結果

閾値	平均精度 (%)	全クラスタ内文書数	クラスタ数
0.771	11.49213	238	62
0.818	12.95122	126	41
0.849	19.62121	69	22
0.874	31.52338	44	13
0.912	52.38095	23	7
0.955	66.66667	10	3

閾値が高くなり全クラスタ内文書が減少するほどクラスタ精度が上昇していることが表 2 に見られ，文書間類似度が高ければ高いほど正しく同種の印象を持つ文書が集まっていることが観察された．また，クラスタ数が最大になる閾値（0.771）では分類が非常に粗く，精度は低かった．精度が高かったクラスタに含まれる文書の多くでは，その印象がはっきりしており，多くの語が索引付けの対象となっていた．

4. 関連研究

評判情報を扱う研究はいくつか行われている．1 つは立石ら[1]や Turney ら[2]のように，機械学習を用いて文書が対象物に対して好意的あるいは否定的どちらに記述されているかを判断する研究である．また，Liu ら[3]は常識知と統計的手法を用いて単一の語あるいは文をあらかじめ決められた感情のカテゴリーに分類している．

5. おわりに

本稿は，テキスト中に表された印象という曖昧な情報に基づいて文書をクラスタリングする手法として，印象を表す語から連想される色を用いて文書の索引付けを行う手法を提案した．話題に基づく従来のクラスタリング手法や検索と組み合わせることによる有効性も期待される．今後は話題に基づくクラスタリングとの比較や組み合わせ，また印象を表す語の中でも感情語に注目したクラスタリングとの比較によって手法の分析を行うことで，特徴の明確化や精度の改善などを目指す予定である． なお，連想された色を利用した場合には，個人差が問題となり得る．これに関して，現在本研究で用いた辞書の妥当性や色の連想の個人差を調査するための実験を行っている．

参考文献

[1] 立石健二, 石黒義英, 福島俊一, “インターネットからの評判情報検索, ”情報処理学会研究報告 自然言語処理, vol.2001, No.69, pp.75-82, 2001.

[2] Turney P.D., Thumbs up or thumbs down? Sentiment Classification using Machine Learning Techniques, In Proceedings 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), pp.417-424, 2002.

[3] Liu H., Lieberman H., Selker T., A Model of Textual Affect Sensing using Real-World Knowledge, To Appear in Proceedings of IUI 2003, Miami, Florida, January 2003.

[4] 杉田茂樹, 江口浩二,“目録データベースと Web コンテンツの統合的利用方式, ”情報処理学会研究報告 情報学基礎, Vol.2001, Num．20, pp.153-158, 2001.

[5] 長町三生, 感性工学のおはなし, pp191-208, 日本規格協会, 1995.

[6] 小林重順, カラーリスト, 日本カラーデザイン研究所, 講談社, 1994.

[7] 中山記男, 大倉典子, “感性情報を用いたテキスト分類手法の検討, ”電子情報通信学会 2003 年総合大会, 2003.