一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
MVE2020-17(2020-09)

俯瞰支援のための文書彩色方法の検討

—単語から連想される色に基づく文書内容の可視化—

松平　茅隼† カストナーマークアウレル††,† 井手　一郎†††,† 川西　康友† 平山　高嗣†††† 道満　恵介†††††,†出口　大輔† 村瀬　洋†

† 名古屋大学 大学院情報学研究科 〒464-8601 愛知県名古屋市千種区不老町
†† 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
††† 名古屋大学 数理・データ科学教育研究センター 〒464-8601 愛知県名古屋市千種区不老町
†††† 名古屋大学 未来社会創造機構 〒464-8601 愛知県名古屋市千種区不老町
††††† 中京大学 工学部 〒470-0393 愛知県豊田市貝津町床立 101
E-mail: † matsuhirac@murase.is.i.nagoya-u.ac.jp, †† mkastner@nii.ac.jp,
††† ide@i.nagoya-u.jp, † {kawanishi, ddeguchi, murase}@i.nagoya-u.ac.jp
†††† takatsugu.hirayama@nagoya-u.jp, ††††† kdoman@sist.chukyo-u.ac.jp

あらまし　本報告では，小説などの書籍を対象とし，ページ単位でその文書内容に応じた彩色を行なう手法を検討する．このような彩色を施すことにより，書籍をわざわざ開かずとも，読者がその内容の変遷を俯瞰的に知ることができるようになると期待される．提案手法では，文書中に出現する各単語に対して「人間が単語から連想する色」に関する知見を用いて色を割り当てる．そして，各ページにおいて出現確率が高い k 色の積み上げ棒グラフを作成し，それらをページ順に並べて表示することで，文書内容の可視化を試みる．最後に，実在する小説を対象にして本手法を適用して定性的な分析を行ない，実際の文書内容に対応付けて本手法の彩色結果を説明できることを確認した．
キーワード　文書検索，色彩連想語，可視化

A study on a coloring method for document overviews

—Visualizing the document structure using word-color associations—

Chihaya MATSUHIRA†, Marc A. KASTNER††,†, Ichiro IDE†††,†, Yasutomo KAWANISHI†, Takatsugu HIRAYAMA††††, Keisuke DOMAN†††††,†, Daisuke DEGUCHI†, and Hiroshi MURASE†

† Graduate School of Informatics, Nagoya University
†† National Institute of Informatics
††† Mathematical and Data Science Center, Nagoya University
†††† Institutes of Inovation for Future Society, Nagoya University
††††† School of Engineering, Chukyo University
E-mail: † matsuhirac@murase.is.i.nagoya-u.ac.jp, †† mkastner@nii.ac.jp,
††† ide@i.nagoya-u.jp, † {kawanishi, ddeguchi, murase}@i.nagoya-u.ac.jp
†††† takatsugu.hirayama@nagoya-u.jp, ††††† kdoman@sist.chukyo-u.ac.jp

Abstract　In this report, we study a method to automatically color each page of a book according to its contents. This allows readers to easily understand the transition of the contents without opening the book. In the proposed method, a representing color is selected for each word based on an existing dictionary on mental word-color associations. Then we generate a 2D map that visualizes the top-k representative colors for each page. In this way, we can visualize the document structure. Finally, we applied this method to existing novels to showcase its applicability.
Key words　document retrieval, word-color association, visualization

1. はじめに

一般に小説などの書籍には，ページ単位で内容を反映した見出しや索引は付与されていない．そのため，読者にとって特定の事象が発生するページを検索することは必ずしも容易ではない．そこで本報告では，書籍の各ページにその内容に応じた色付けを行ない，内容を類推する手がかりとして色を付与することで，読者によるこのような文書内容の俯瞰作業の支援を試みる．このような手法により生成される色特徴は，文書検索時のキーとしての利用や，文書要約への応用が期待される．

文書の内容に応じてページの色付けを行なう方法としては，さまざまなものが考えられる．本報告では，人間が単語から連想する色に関する知見を利用した文書の彩色により，その内容・構造を可視化する方法について検討する．

単語から連想される色というテーマについては，既に Mohammad により研究が行なわれている [3] ．この研究では，被験者実験を通して約 1 万の英単語について連想される色に関するデータを収集し，それを辞書としている．この辞書では各単語について，その単語の用法ごとに，basic color terms [1] と呼ばれる 11 色（白・黒・赤・緑・黄・青・茶・桃・紫・橙・灰）のうちで最も連想されやすい色が割り当てられている．また，色の割当に対して，投票のばらつきに応じた確信度が与えられている．例えば，単語「mean」の名詞としての用法について緑色と投票した被験者が 5 人中 2 人で最多であれば，その確信度は $\frac{2}{5} = 0.4$ となる．一方で，形容詞としての用法について赤色と投票した被験者が 5 人中 4 人で最多であれば，その確信度は $\frac{4}{5} = 0.8$ となる．

また Kim らは，この単語色辞書の内容可視化のためのデモの作成に付随して，文書の色特徴化を試みている [2]．具体的には，この辞書を用いて入力テキスト全体に出現する単語の色情報を集計し，各色の割合を出力するテキストエディタを作成している．しかし，入力テキストをページや一定語数単位で分割するような処理は行なっていない．

時系列データの彩色及び内容の可視化というテーマについては，Morr らにより類似した試みが行なわれている(注1)．彼らは文書ではなく映像を対象として，各フレームを幅 1 画素の彩色された線に縮退し，それを時系列的に横に繋げて帯状にすることで映像内容を可視化している．

以降，2. で提案手法の詳細について述べる．次に，3. では提案手法による文書彩色の結果とその応用例について述べる．最後に 4. で本報告をむすび，今後の課題について述べる．

2. 文書の彩色手法

本報告では，文書をページ単位で彩色することにより文書の内容・構造を可視化する手法を提案する．

2.1 ページ単位での色付け

具体的な色付け方法は以下の通りである．まず事前処理として，単語色辞書 [3] 中の各単語 w に対する，確信度が r_v であるような色 c_v への投票 $v = (c_v, r_v)$ の集合 $V[w]$ を用いて，単語 w に対する色の確率分布 $\boldsymbol{c}(w)$ を，以下の式により 11 次元のベクトルとして算出しておく．ただし，Z_1 は正規化定数であり，$\boldsymbol{c}_v$ は色 c_v に関して 11 次元のワンホットベクトル化したものである．

$$\boldsymbol{c}(w) = \frac{1}{Z_1} \sum_{v \in V[w]} r_v \, \boldsymbol{c}_v \tag{1}$$

次に，入力テキストに形態素解析を施すことで単語単位に分かち書き・基本形化し，これをページ単位で N 語ずつにまとめる．本研究においては，彩色単位となるページは書籍における実際のページではなく，単語数（1 ページ当たり $N = 500$（英文）または 100 単語（和文））に基づいて決定した．これは単に使用したデータの都合であるため，ページ境界さえ与えられれば，手法自体は実際の書籍におけるページ単位で処理することも可能である．

最後に，各ページ $P = \{w_1, w_2, \ldots, w_N\}$ に対して，先に算出しておいた単語に対する色の確率分布 $\boldsymbol{c}(w)$ に基づき，ページに対する色の確率分布 $\boldsymbol{c}_P$ を以下のように算出する．ただし，Z_2 は正規化定数である．

$$\boldsymbol{c}_P = \frac{1}{Z_2} \sum_{w \in P} \boldsymbol{c}(w) \tag{2}$$

2.2 可視化の方法

前項で求めたページに対する色の確率分布 $\boldsymbol{c}_P$ に基づき，文書内容を可視化する．

本研究では 1. で紹介した Morr らの手法を参考にし，ページ単位の最頻 k 色を積み上げ棒グラフとして描画した後，それをページ順に横に並べた帯の形で文書内容の可視化を試みる．図 1 に可視化の例を示す．

3. 文書の彩色結果

いくつかの著名な文献をテキスト形式で提案手法に入力し，その出力を確認した．ここで，簡略化のために，提案手法の入力は挿絵・目次・引用符等を除いたテキストのみを対象として分析した．また，彩色の妥当性を確認するために，各ページ中に出現する単語のうち，ページ中の最頻色 c_l と同色でかつその確率 $c_l(w)$ が高いものを付記した．

3.1 彩色例

Project Gutenberg(注2)から入手した英文の著名な小説に対して，提案手法を適用した例を図 2 に示す．ここで，英単語の基本形化には，PyPI の Inflection ライブラリ(注3)中の関数 singularize を使用した．

また，単語色辞書 [3] 内の全単語を Google 翻訳 API(注4)を用いて日本語に翻訳することで，日本語版の辞書も作成した．これを用いて，青空文庫(注5)から入手した和文の著名な小説に対

(注1)：https://timelens.io/

(注2)：http://www.gutenberg.org/

(注3)：https://pypi.org/project/inflection/

(注4)：https://pypi.org/project/googletrans/

(注5)：https://www.aozora.gr.jp/

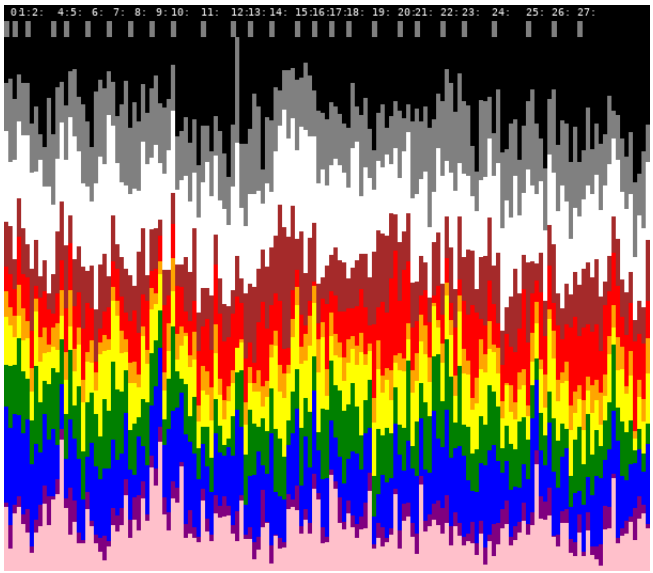

図 1　Mary Shelley による小説 “Frankenstein; Or, The Modern Prometheus” [4] に対する提案手法（$k = 11$）の出力結果．（上段の数字は，各章の開始ページを表す．）

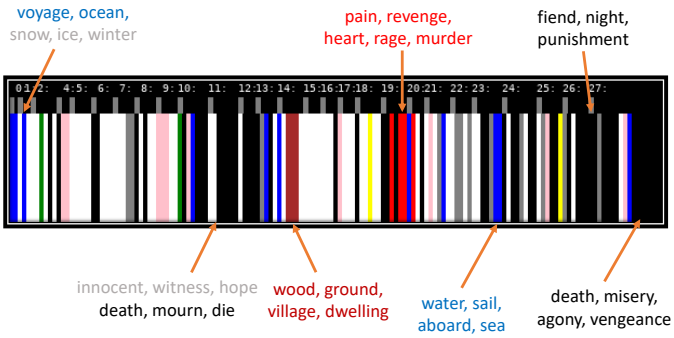

図 2　Mary Shelley による小説 “Frankenstein; Or, The Modern Prometheus” [4] に対する提案手法（$k = 1$）の出力結果．（上段の数字は，各章の開始ページを表す．）

しても提案手法の応用を試みた．その結果例を図 3 に示す．ここで，和文の分かち書き及び単語の基本形化には日本語形態素解析エンジン MeCab(注6) を使用した．また，助詞・助動詞・その他一部の頻出単語（「その」や「する」など）はその出現頻度による影響が大きかったため，ストップワードと考えて無色とした．

3.2　彩色結果に対する考察

前項の彩色結果の妥当性について考察する．この際，可視化結果の解釈容易性が最も高かったことから，本節においては $k = 1$ として分析を行なう．

図 2 中の色変化が顕著なページ，及び併記された単語に注目すると，まず物語の最初の部分で海や雪に関するシーンがあることが推測できる．実際にこの部分は，原文において探検隊が北極で主人公の Frankenstein を発見するシーンに対応している．さらに，第 19 章の位置で，復讐による殺人が発生していることが推測できる．原文においてこの部分は，醜い見た目に

(注6)：https://taku910.github.io/mecab/

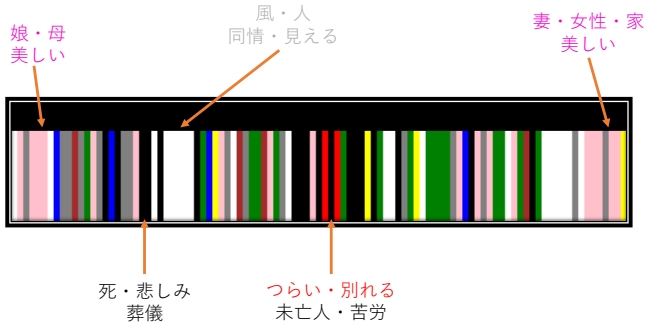

図 3　紫式部による小説「源氏物語第 1 帖：桐壺」（與謝野晶子訳）[5] に対する提案手法（$k = 1$）の出力結果．

造られた怪物が復讐心から創造者である Frankenstein の親族を殺害するシーンに対応している．

また，図 3 についても同様に彩色結果を分析すると，まず物語の前半部分に女性に関する描写及び人の死に関する描写が見受けられる．実際にこの部分は，原文において主人公である光源氏の実母である桐壺の描写及び彼女が死去するシーンに対応している．さらに，物語の最後では女性に関する描写が再び見受けられる．原文においてこの部分は，桐壺によく似た女性である藤壺の登場，そして光源氏の彼女に対する好意を描くシーンに対応している．

以上のように色変化が顕著なページを見ると，その彩色結果と小説内容に顕著な対応関係が存在していることがわかる．よって，その小説のあらすじを知っている人であれば，出力結果の顕著な色変化から内容の変遷を俯瞰することができると考えられる．

4.　ま　と　め

本報告では，文書のページ単位での彩色により，文書に含まれる事象の変遷を可視化する手法を検討した．そして定性的な分析により，提案手法により出力されたページの彩色結果を，実際の文書内容に対応付けて説明できることを確認した．

本手法の応用例として，文書検索や文書要約，文書内容の自動評価などが考えられる．

謝辞　本研究の一部は，科学研究費補助金による．

文　　献

[1] Brent Berlin and Paul Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, 1969.

[2] Chris Kim, Uta Hinrichs, Saif Mohammad, and Christopher Collins. Lexichrome: Text construction and lexical discovery with word-color associations using interactive visualization. In *Proc. 2020 ACM Designing Interactive Systems Conference*, pp. 477–488, Jul. 2020.

[3] Saif Mohammad. Colourful language: Measuring word-colour associations. In *Proc. 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pp. 97–106, Portland, Oregon, USA, Jun. 2011.

[4] Mary Wollstonecraft Shelley. *Frankenstein; Or, The Modern Prometheus*. Lackington, Hughes, Harding, Mavor & Jones, 1818.

[5] 與謝野晶子．全訳源氏物語 上巻．角川文庫，1971.