ユーザレビューを用いた服素材の定量的可視化
Quantitative Visualization of Clothing Materials Using User Reviews

江上　碧海
Aomi Egami
法政大学情報科学部ディジタルメディア学科
E-mail: aomi.egami.2g@stu.hosei.ac.jp

Abstract

Recently, clothing is being made of various materials such as cotton, hemp and polyester. People can choose clothing materials for their own purposes. However, to choose appropriate materials, people need to know their features. For example, cupra is similar to, but thinner than rayon. It is difficult to remember their characteristics, and also to get to know their right uses. In addition, it is difficult to select clothes by using only the physical features of their materials, because people may have different feelings when actually wearing clothes. This paper proposes a method for visualizing features of clothing materials by analyzing user reviews of clothes. The method quantitatively and visually presents for which category each clothing materials is suitable. Categories indicate people's uses of clothes such as spring, summer and sports. Such categories are assigned to materials by using characteristic words extracted from user reviews of clothes. This paper presents the results of experiments on the quantitative visualization of the system using user reviews about 3000 clothing items obtained from the Web. The experimental results indicate that the method quantitatively and visually shows suitable categories of clothing materials.

1. はじめに

近年，様々な素材から服が作成されている．代表的なものとして，綿や麻，ポリエステルなどがある．用途に合わせて服の素材を選択することができる．しかし，用途に適した素材を選択するためには，それぞれの特性などを知らなければいけない．例えばキュプラは，レーヨンとよく似ているが，より湿潤状態でも強度の低下が少ないという特性がある．このような特性を全て覚えることは難しく，また適した用途がどれかもわかりにくい．加えて，実際に着用してみると，人によって感じ方が異なるため，物理的特性だけで服を選択することも難しい．

本研究では，服のユーザレビューから得られた素材の特徴を定量的に可視化する手法を提案する．本手法では素材の特徴を定量化するために，用途などのジャンルを用いる．具体的には，服のユーザレビューから特徴語を抽出し，特徴語をジャンルに対応させ，Tf-Idf 値を用いて各素材がどのジャンルにどれだけ適しているかを定量化する．さらにその結果に対して Zoomable Sunburst を用いて各ジャンルの割合を提示することで，素材の特徴を定量的に可視化する．

この手法を応用し，服を評価したユーザレビューを Web から取得して，素材の特徴を可視化するシステムを作成した．また，Web 上の約 3000 商品の服のユーザレビューを使用し，システムの可視化結果を評価する実験を行った．実験の結果，各素材が適した状況を本手法が定量的に可視化していることを確認できた．

2. 関連研究

川井ら [1]はユーザレビューから評判情報を抽出し，それらの肯定・否定情報と合わせて可視化を行った．機械学習によって語句を分類し，肯定語・否定語情報の辞書を作成した．この辞書に基づいて，HK Graph により未知のユーザレビューの評判情報を可視化した．

打田ら [2]はユーザレビューをもとに，製品などに対する評価解析を行った．多次元尺度法によりユーザレビュー間の関係性を把握し，HK Graph を用いてレビュー内容の時系列変化・類似性の視覚的な把握支援を行った．

松田ら [3]は服地と服地画像で視覚から受ける印象の違いに関して実験を行った．また服地の色による心理効果にも着目した．

3. 準備

3.1. Tf-Idf

Tf-Idf は文章中の単語に関する重みの一種である．文書d内の単語tの出現頻度$\mathrm{tf}(t,d)$，単語tの逆文書頻度$\mathrm{idf}(t)$の 2 つの指標に基づいて，Tf-Idf 値$\mathrm{tfidf}(t,d)$を以下の通り計算する．

$$\mathrm{tfidf}(t,d) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t)$$

$$\mathrm{tf}(t,d) = \frac{n_{t,d}}{\sum_{k \in T_d} n_{k,d}}, \qquad \mathrm{idf}(t) = \log \frac{N}{\mathrm{df}(t)} + 1$$

ただし，$n_{t,d}$は文書 d 内での単語tの出現頻度，$\sum_{k \in T_d} n_{k,d}$は文書d内の単語の集合T_dの出現頻度の和，Nは全文書数，$\mathrm{df}(t)$は単語tが出現する文書の数である．

3.2. Zoomable Sunburst

D3.js [4]の Zoomable Sunburst は，Stasko ら [5]によって提案された Sunburst グラフにズーム機能を加えたものである．Sunburst は階層化されたデータをドーナッツ状のグラフで表す．図 1 は高校 1 年生のクラス内の文系・理系の選択者についての例である．内側のドーナッツ状グラフで各クラスの人数比がわかる．濃いオレンジ色が最も大きいため，1 年 A 組が最も人数の多いクラスである．

Supervisor: Prof. Hiroshi Hosobe

また 1 年 A 組から時計回りの最後にある 1 年 C 組が最も人数の少ないクラスである．外側のドーナッツ状グラフで同様に見ていくと，1 年 A 組は理系，1 年 B 組は文系が多く，1 年 C 組は文系のみであることがわかる．

図 1　Sunburst の例

4. 提案手法

本研究では，服のユーザレビューから得られた素材の特徴を定量的に可視化する手法を提案する．具体的には，素材に対応する服商品のユーザレビューからその特徴を表す語を抽出した後，それぞれの語の特徴量を求め，さらにジャンルごとに集計した結果を可視化する．

4.1. 単語の抽出

ユーザレビューから単語を抽出し，商品の素材ごとにまとめて，各単語の出現頻度を求める．最初に，ユーザレビューが 5 件以上ある服商品に限定し，その素材とユーザレビューを取得する．次にユーザレビューから単語の抽出を行う．最後に，抽出した単語の出現頻度を求め，素材ごとにまとめる．

4.2. 分類辞書の作成

素材の特徴を表す特徴語とそのジャンルを登録した辞書を作成する．本研究で設定するジャンルは，春，夏，秋，冬，スポーツ，カジュアル，フォーマル，その他，の 8 種類とする．その他には質感や価格などの素材の特性を表す特徴語が属する．また特徴語によっては 1 種類だけではなく複数のジャンルに属する場合もある．

特徴語の選択とジャンルへの対応付けは手動で行う．選択する特徴語は下記の条件を満たすものに限定する．

- 品詞が名詞，形容詞，形容動詞であり，記号や数字のみで構成されていない．
- 取得したユーザレビューを素材ごとに分類したとき，出現頻度が多く，8 ジャンルと関係がある．

4.3. 素材の特徴の定量化

素材の特徴を定量化するために，素材ごとに各特徴語の Tf-Idf 値を算出して用いる．具体的には，Tf-Idf 値 $\mathrm{tfidf}(t, d_m)$ の算出において，ある素材 m に関するユーザレビューを統合したものを文書 d_m とし，特徴語を単語 t とする．

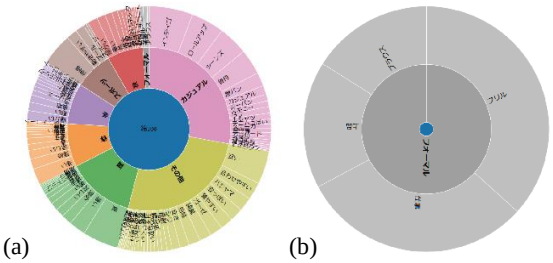
素材 m，ジャンル g の特徴量を $q(m, g)$ とすると，g に対応する特徴語の集合 T_g に対して，$q(m, g)$ は以下の式で算出される．

$$q(m, g) = \sum_{t \in T_g} \mathrm{tfidf}(t, d_m)$$

4.4. Zoomable Sunburst による可視化

定量化した素材の特徴を Zoomable Sunburst によって可視化する．中央の円で素材を，内側のドーナッツ状グラフでジャンルを，外側のドーナッツ状グラフで特徴語を表現する．Zoomable Sunburst を使う理由は，定量的な比較が視覚的にわかりやすく，階層的なデータが可視化できるためである．Sunburst ではなく Zoomable Sunburst である理由は，割合が少なく文字が読めないような単語も拡大することによって読むことが可能になるからである．

図 2(a)は素材が綿 100%である服商品のユーザレビューを可視化した例である．内側のグラフにあるその他やカジュアルなどがジャンルを表し，外側のグラフにある「安い」や「インディゴ」などが特徴語を表している．例えばカジュアルは「インディゴ」，「ロールアップ」，の順に Tf-Idf 値が高いことを示している．外側のグラフでは「インディゴ」が最も広いことから，最も Tf-Idf 値が高い特徴語であることが読み取れる．関連するジャンルとして，その他を除くとカジュアル，夏，春の順に高いことが読み取れる．図 2(b)は図 2(a)のフォーマルをズームした場合である．図 2(a)ではフォーマルは割合が少ないため，文字の重なり合いによって読みにくかったり，同ジャンル間での特徴語の割合の差がわかりにくかったりしている．図 2(b)ではズームすることで，各特徴語や，ジャンル間での特徴語の差を確認することができる．



(a)　(b)

図 2　提案手法による可視化の例

5. 実装

4 節に述べた提案手法に基づき，服商品のレビュー文を取得し，素材ごとに Tf-Idf 値を算出した特徴語とそのジャンルを Zoomable Sunburst で可視化するシステムを実装した．服商品のデータ収集，及びレビュー文の形態素解析，Tf-Idf 値の算出，単語とジャンルの辞書による対応付けを Python で，Zoomable Sunburst での可視化を JavaScript で行った．ベルメゾン[1]と SHOPLIST[2]から服商品のユーザレビューと素材のデータを収集した．レビュー文の形態素解析には MeCab [6]を使用した．

5.1. 単語の抽出

Python で作成したプログラムによってユーザレビューの収集と単語の抽出を行った．Web 上から服商品の素材とユーザレビューを 3090 商品について収集した．全商品

[1] http://www.bellemaison.jp/

[2] http://shop-list.com/

の形態素解析を行い，素材の組合せごとに単語の出現頻度をまとめ，テキストファイルに保存した．割合に応じた素材の組合せは 15 商品以上になるように手動で分け，59 組にまとめた．ただし，アルパカ 100%やレーヨンと麻の組合せなど，15 商品以上にならない素材の組合せは含まない．組合せの一部を表 1 に示す．

表 1　素材組合せの例

素材
綿 80～99%，ポリエステル 1～20%
綿 66～79%，ポリエステル 21～34%
ポリエステル 100%

5.2. 分類辞書の作成

ジャンルと特徴語を関連付けるための分類辞書を手動で作成した．5.1 節でまとめた 59 組の素材の組合せごとのユーザレビュー文のそれぞれで出現頻度が上位 500 語，合わせて 3588 語を自動で選出した．この 3588 語からジャンルと関連する特徴語 319 語を手動で選出し，ジャンルと合わせてテキストファイルに保存した．各特徴語は 1 つのジャンル，または複数のジャンルに分類を行った．辞書の一部を表 2 に示す．

表 2　分類辞書の例

特徴語	ジャンル
安い	その他
薄い	春，夏
あたたかい	秋，冬
フリース	冬

5.3. 素材の特徴の定量化

Python で作成したプログラムによって，各素材の特徴を定量化するために特徴語の Tf-Idf 値を算出し，可視化時に使用する JSON ファイルを作成した．JSON ファイルは，分類辞書に登録した各特徴語の Tf-Idf 値と属するジャンルが記述されている．Tf-Idf 値は，5.1 節で作成したテキストファイルから 59 組の素材ごとの各単語の出現頻度を読み込み，特徴語のみを使用して算出した．JSON ファイル作成の際，複数のジャンルに含まれる特徴語であっても値を変えずに使用した．

5.4. Zoomable Sunburst による可視化

JavaScript で作成したプログラムによって可視化用の JSON ファイルを読み込み，Zoomable Sunburst で定量的に可視化する．5.3 節で作成した Tf-Idf 値とジャンルが記された JSON ファイルを読み込む．D3.js のライブラリを利用して，Zoomable Sunburst で可視化する．

6. 実験

本手法の有用性を示すために，まず一般的にジャンルの想像が容易な素材の可視化を行い，次に一般的ではない素材が含まれる場合の可視化を行う．図 3 は一般的な素材である麻 100%の場合のジャンルと特徴語を可視化した結果である．ジャンルの中で最も割合が高いものは夏であることから，麻 100%は夏に適していることがわかる．反対に寒い秋や冬の割合が低いことから，これらの季節には適していないことが読み取れる．

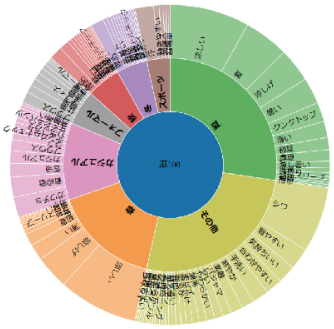


図 3　麻 100%

図 4 は素材が毛 100%の場合のジャンルと特徴語を可視化した結果である．麻 100%の場合とは反対に冬の割合が高くなっているが，夏の割合は低くなっている．また季節に関するジャンルの割合が半分以上を占めていることから，季節によって選択されることが多いと読み取れる．

図 4　毛 100%

図 5 は素材が綿とポリウレタンであるもののうち，綿が 87～95%，ポリウレタンが 5～13%の割合であるものを可視化した結果である．図 2 の綿 100%を可視化した結果と比べるとスポーツの割合が増えている．またスポーツのうち，「動きやすい」と「伸縮」が大部分を占めていることから，ポリウレタンは伸縮性があることがわかる．

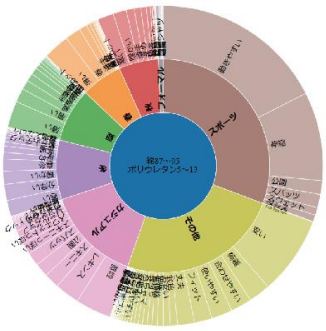


図 5　綿 87～95%，ポリウレタン 5～13%

図 6 は，一般的ではない素材であるモダール 50%，綿 45%，ポリウレタン 5%を可視化した結果である．本実験では，新たに楽天市場[3]から，この素材に関して，レビューが 15 件以上(最大で約 5000 件)である 6 商品のレビューを取得し，前述のデータと合わせた上で Tf-Idf 値を算出し直して用いた．6 商品しかないためにそれぞれの服商品の特徴が出やすくなってしまっており，図 5 に比べると，春や夏の割合が増えている．また，その他の中で「なめらか」や「やわらかい」などの質感に関する単語が増えている．これはモダールが加えられたためであると考えられる．

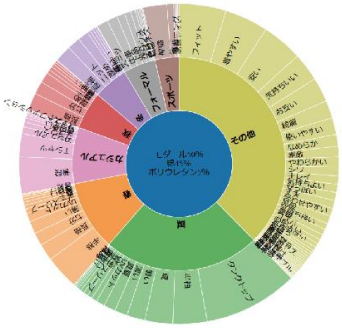


図 6　モダール 50%，綿 45%，ポリウレタン 5%

7. 議論

本手法で定量的に可視化することで，素材ごとにどのようなジャンルに適しているかが 1 つの図から確認できるようになった．またそれぞれのジャンルに含まれる特徴語も同時に見ることができるため，割合が多い特徴語からその素材の特徴がわかるようになった．

問題点として，商品数が少ないと素材の特徴を可視化することが難しく，その服商品の特徴が出やすくなってしまう点がある．その解決のためには，レビューを取得するサイトを増やし，各素材のレビューの数を多くする必要があると考えられる．

手動で分類辞書を作成したため，個人の主観の影響を受けやすく，作成者の負担が大きい問題がある．また分類辞書作成の際に使用するレビューに偏りがあると，ある素材では特徴的な単語であるが，辞書に登録されていないために可視化には現れない．しかし既存の辞書の使用は難しい．例えば一般的な連想語辞書だと，冬は「寒い」という単語が登録されているが，服商品に関するレビューには反対の意味の「あたたかい」などが出てきやすい．服のレビューに特化した辞書である必要がある．よって分類辞書を自動で作成するためには，多くのユーザレビューを集め，服商品のユーザレビューに特化するように学習などを行う必要があると考えられる．またこの際，文の肯定・否定を考慮する必要もある．

本研究では，複数のジャンルと関連がある特徴語を扱う場合に，同じ Tf-Idf 値をそれら複数のジャンルに使ってしまっている．そのため複数のジャンルと関連がある特徴語の Tf-Idf 値が素材の中で高い割合がある場合，他の特徴語を圧迫してしまう．そのため値に加工を施す必要がある．しかし単純に関連があるジャンルの数で割るだけでは，本来は高い割合を占めるはずの特徴語が，割合が低い特徴語のように見えてしまう．単純に割るだけではなく，ジャンルによって重みづけを行う必要があると考えられる．

8. おわりに

本研究では，服のユーザレビューから得られた素材の特徴を定量的に可視化する手法を提案した．本手法では素材の特徴を定量化するために，用途などのジャンルを用いた．本手法に基づくシステムを，Python と JavaScript を使用して作成した．本手法が有用であるかを確認するため，実験を行った．実験結果から，本手法は素材の適したジャンルを 1 つの図で確認でき，またジャンル内で多い割合を占める特徴語を確認できることがわかった．

データ作成にはいくつかの問題があることがわかった．まずユーザレビューの拡張が必要である．分類辞書を手動で作成しているため，学習等による自動化が必要である．また，可視化を行う際，Tf-Idf 値をジャンルに合わせて重みづけを行う必要があると考えられる．

文　献

[1] 川井康示, 吉川大弘, 古橋武, "ユーザレビューの評判情報の分類と HK Graph による可視化," 第 27 回ファジィシステムシンポジウム, pp. 906-911, 2011.

[2] 打田裕樹, 吉川大弘, 古橋武, 平尾英司, 井口浩人, "Web ユーザレビューにおける評価情報の時系列変化の可視化," 日本知能情報ファジィ学会誌, vol. 22, no. 3, pp. 377-389, 2010.

[3] 松田恵, 森本敬子, 長篤志, 木下武志, "服地と服地画像の刺激間差異が質感認知に及ぼす影響," 日本感性工学研究会論文誌, vol. 13, no. 1, pp. 91-98, 2014.

[4] M. Bostock, J. Heer and V. Ogievetsky, "Data-Driven Documents," [Online].
Available: https://d3js.org/.

[5] J. Stasko, R. Catrambone, M. Guzdial and K. Mcdonald, "An Evaluation of Space-Filling Information Visualizations for Depicting Hierarchical Structures," *Human-Computer Studies*, vol. 53, no. 1, pp. 663-694, 2000.

[6] 工藤拓, 山本薫, 松本祐治, "Conditional Random Fields を用いた日本語形態素解析," 情報処理学会研究報告：自然言語処理(NL), vol. 2004, no. 47, pp. 89-96, 2004.

[3] http://www.rakuten.co.jp/