# 過去の利用者の製本履歴に基づくブログ製本デザイン推薦

佐野 和広† 木村 文則†† 田名辺 健人‡ 前田 亮†††

†,††,†††立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1
‡欧文印刷株式会社 〒113-8484 東京都文京区本郷 1-17-2
E-mail: †is020083@ed.ritsumei.ac.jp, ††fkimura@is.ritsumei.ac.jp, †††amaeda@media.ritsumei.ac.jp

‡tanabe-tak@obun.jp

**あらまし** 近年，ブログサービス利用者が増加する中で自分自身のブログを製本化したいというニーズが高まっている.それに伴いブログを製本化するサービスが広がりつつある.しかし製本化する際に書体や表紙デザイン等，ユーザが設定しなければならない項目が多くあり，それらがユーザの手間となっている．本研究ではブログ製本サービスにおける過去の製本履歴をもとに，ユーザに対して製本化する際のデザインを推薦する手法を提案する．ブログの本文テキストから$\chi^2$値を用いて各ユーザの特徴量選択を行い，k-means 法を用いユーザ単位のクラスタリングを行う．これらのクラスタの中から，新たに製本化するユーザに最も類似するクラスタを求め，そのクラスタ内のユーザが使用したデザインを推薦する.

**キーワード** ブログ，クラスタリング

# Design Recommendation for Blog Binding Service Based on Past Users Results

Kazuhiro SANO† Fuminori KIMURA†† Takehito TANABE‡ Akira MAEDA†††

†, ††,†††College of Information Science and Engineering, Ritsumeikan University 1-1-1 Noji-Higashi, Kusatsu, Shiga, 525-8577 Japan
‡ Obun Printing Company, Inc. 1-17-2 Hongo, Bunkyo-ku, Tokyo, 113-8484 Japan
E-mail: †is020083@ed.ritsumei.ac.jp, ††fkimura@is.ritsumei.ac.jp, †††amaeda@media.ritsumei.ac.jp

‡tanabe-tak@obun.jp

**Abstract** Recently, demands of blog binding service are increasing owing to an increase of blog service users. However, users must input many settings, for example typeface and cover design etc., when they use a blog binding service for their own blog. In this research, our proposed system recommends the proper design for the users binding blog based on the past results of other users' bindings. The proposed system estimates each user's feature based on chi-square statistic of terms in their blog sentences, and clusters users by k-means method. The proposed system selects the proper cluster for new binding user's blog from these clusters, and recommends designs from used ones in the selected cluster.

**Keyword** Blog, Clustering

## 1. はじめに

近年，インターネットの普及に伴い，ブログやTwitterといったSNSが急激に普及している.その中でもブログサービスは，芸能人の情報発信ツールとして使用されることや，アバターやゲーム要素などを盛り込むなどの多角的なサービスの展開も後押しして，ブログ登録者数は増加し続け，2009年時点で日本国内だけでおよそ2695万人と言われている[1]．手軽にブログを作成できるサービスが提供されているため，専門的な知識を有していないユーザでもブログサービスを利用して，読むだけではなくブログを書くユーザの増加に繋がっている．また，以前はブログを書く手段がパソコンだけに限定されていたが，携帯やスマートフォンなどの普及により，時・場所を選ばずブログを書く・読むことがより気軽に行えることもユーザ数増加の要因の一つとして挙げられる．またブログから発信される情報はマスメディアの情報発信と異なり，書き手の趣味や興味の内容が大きく反映されるという特徴がある.

ブログを書くユーザが増加していく中で，自分自身

が作成したブログを形ある「本」として残したいというニーズが出てきている．それらを受け，ブログの製本化サービスがブログサービスと連携して展開され普及しつつある．ブログを製本化する際，書体や表紙デザインなどユーザ自身が指定するのが一般的となっているが，設定しなければならない項目が多くユーザの手間となっている．そこで本研究では，日本における代表的なブログ製本サービスの一つである「MyBooks.jp」から提供を受けたブログ製本の過去の履歴を用いて，ユーザに対して製本化する際のデザインを推薦する手法を提案する．本手法では，ブログの本文テキストから$\chi^2$値を用いて各ユーザの特徴量選択を行う．k-means 法によりユーザ単位でクラスタリングを行い，似ているユーザ同士のクラスタを生成する．このクラスタの中から新たに製本化するユーザに最も類似するクラスタを求め，そのクラスタ内のユーザが使用したデザインを推薦する．

## 2. 本研究で扱うブログデータ

本研究で使用したブログデータは「MyBooks.jp」[2]から提供を受けたデータを使用している．ユーザ 1 人分のデータに格納されている情報は，記事数分のブログ記事タイトル，記事数分のブログ本文テキスト（個人情報が分からないように，あらかじめ MeCab による形態素解析が行われ出現順序がランダムに格納されており，それにより本文を復元できないようになっている），製本の際にユーザが実際に選んだデザインの 3 つである．本研究の実験の際に用いたのは，636 人分のブログ本文とデザインのペアである．デザインの種類は 45 種類（カラーバリエーションを含めると 89 種類）存在する．

ブログ製本サービス「MyBooks.jp」とは，イースト株式会社および欧文印刷株式会社が共同事業として提供するサービスである．ユーザのブログを印刷・製本して届けるサービスで，アメーバブログ，Yahoo!ブログ，livedoor ブログなど国内大手のブログサービス 20 社と提携し（2011 年 8 月時点）サービスを展開している．従来面倒だった製本化する際の修正作業（改行やレイアウトの変更）をプレビューとして見ながら作業を行えるようにするなど，簡単に編集できるように工夫している．育児ブログやペットブログ，旅ブログ，結婚ブログなどのブログを製本化して，記念として手元に残したい，家族や友人にプレゼントしたい，という女性ブロガーを中心に人気がある．

## 3. 関連研究

1 章で述べた通り，ブログは書き手の趣味や興味の内容，商品に対する評判情報などが反映されやすく，リアルタイム性に優れるという特徴がある．それらの特徴に着目し，流行しているものが何か等の情報を取り出すことを目的とした研究や，それに伴いブログを分類する必要性があり研究が行われている．その中にはスパムブログ除去のための分類やブログの記事内容からどのような属性を持つ人が作成したかに基づく分類（[3][4][5]）等がある．池田ら[3]はブログ著者の性別をブログから素性（機能語+一人称+形態素）選択を行い Support Vector Machine（SVM）により分類している．大倉ら[4]は性別・居住域・年齢層の属性を対象とし，素性選択を行ったうえで Complement Naïve Bayes（CNB）により分類を行っている．[3][4]ともに形態素の素性選択には$\chi^2$値（詳細は 4 章で説明）を利用して選択を行っている．この手法による素性選択の有用性についても[3][4][5]で述べられており，機能語・一人称・形態素の素性の組み合わせによる比較実験も行われ，形態素の素性選択が有効であることが示されている．池田ら[6]は副分類機による特徴ベクトルから半教師有り学習を行う手法の評価を行っている．

これらの研究と本研究との違いは大きく 2 つある．1 つ目は研究に用いるブログデータの性質が大きく異なることである．[3]～[6]で実験に用いているブログデータはインターネット上から収集したものであることと，学習に用いるユーザのプロファイル情報（性別・居住域・年齢）を取得していることが挙げられる．本研究では企業から提供を受けたブログデータを用いていること，ユーザが実際に製本化する際に選んだ表紙デザインがデータの内容として含まれていることに違いがある．2 つ目は SVM や CNB のような学習・分類を行うことが主目的ではなく，似ているユーザをクラスタリングしブログの製本デザインを推薦することが目的ということである．製本化したいユーザに対して 1 つのデザインを推薦するのではなく，似ているユーザが選択したデザインも推薦候補として出力する価値があると思われる．なぜならデザインを選択する際に他のデザインと迷ったかもしれず，他のデザインの方がよかったという可能性も考えられる．このような考えから最適と思われる製本デザインのみを推薦するのではなく，適切であると思われる複数のデザインを推薦する．

## 4. 提案手法

### 4.1. 概要

全体の流れを示したシステム全体図を図 1 に示す．新たに製本化したいユーザのブログを入力データとして用いる．過去に製本化したブログユーザ群のデータをあらかじめクラスタリングしておく．入力したユーザの座標と各クラスタの重心座標のコサイン類似度を

求め,入力ユーザがどのクラスタに属するかを求める.入力ユーザに最も類似するクラスタ内の製本デザインを推薦結果として出力する.クラスタリングの際に必要となる①特徴量選択は[3]～[5]で用いられている$\chi^2$値を用いる.選択した特徴量は各ユーザの特徴ベクトルの値として用いられる.②クラスタリングにはk-means 法を用いる.以下,それぞれについて説明する.

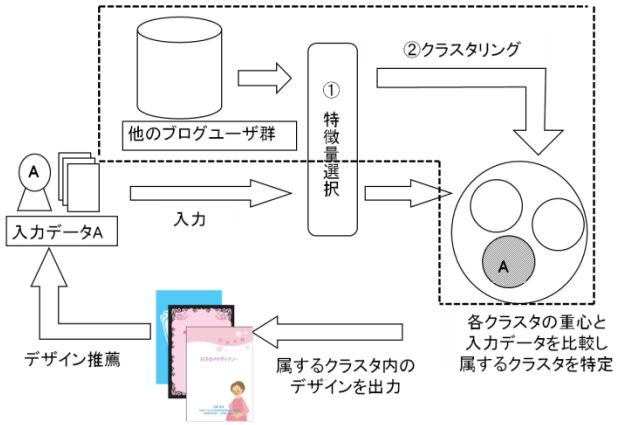


図 1：システム全体図

### 4.2. $\chi^2$値を利用した特徴語抽出

$\chi^2$値は有意検定をする際によく使用される値で,ある集合に対し,複数の分割がどの程度存在し影響を与えているかを示す指標である.1 ユーザに,ある単語 $t$ が含まれているか否かによる分割と,ユーザのある属性値が $c$ であるか否かの分割,計 4 分割がどの程度一致するかを測るものである.この値が大きいほど単語 $t$ が属性値推定に有用であるといえる.

ここで,$A$ を単語 $t$ を含み属性値が $c$ であるユーザ数,$B$ を単語 $t$ を含み属性値が $c$ でないユーザ数,$C$ を単語 $t$ を含まず属性値が $c$ であるユーザ数,$D$ を単語 $t$ を含まず属性値が $c$ でないユーザ数としたとき,$\chi^2$値は,

$$\chi^2 = \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

と計算される.

本研究の場合,属性値 $c$ とは特定の製本デザインを指し,そのデザインを使用したかしていないかである.ある単語 $t$ が特定のデザインに多く使われ,他のデザインであまり使われていない場合に$\chi^2$値が高くなる.

$$\chi^2_{max}(t) = \max_c \chi^2(t,c)$$

この計算により,$\chi^2_{max}(t)$の値が大きい順に特徴語として選択する.しかし単純に値が高いものを使用してしまうと一般性に欠ける単語が上位に来てしまうなどの問題がある.この問題に対する処理や利用する単語の数と選出した特徴語については 5 章で示す.

### 4.3. k-means 法

k-means 法は非階層クラスタリングの代表的手法であり,クラスタの数 k をあらかじめ決定しておいて,データを k 個のグループに分類する方法である.アルゴリズムは,ランダムに各サンプルを k 個のクラスタに初期配置した後,各クラスタにおける代表座標とそれに属するサンプルの重心との距離を短くするよう,機械学習によりクラスタ座標を変化させる.これらを繰り返し変化が無ければ終了し,その時点のクラスタが分類結果となる.計算コストが少ないため,対象データのサンプル数が膨大で,必ずしも疎ではないベクトルデータに対して直接クラスタリングを行う手法としては,SVM 等の手法より適していると考えられる.しかしながら,k-means はもともと局所解を求めるアルゴリズムであるため,最初にランダムに決められる初期配置に依存する.そのため,k-means を有効に利用するには,初期値を工夫するか,学習過程における初期値依存性を低くする,複数回の試行を行った際に,いずれが妥当であるかを判断できる材料を提供することが必要となる.

## 5. 評価実験

### 5.1. 評価実験内容

入力データとする 45 人分を除いた 591 人分のデータを用いる.正解データにはユーザが選択した製本表紙デザインを用いる.ブログデータで用いられた単語を特徴量（特徴ベクトル）として k-means 法によるクラスタリングを行う.各ユーザのベクトルの次元数は単語の種類であり,各次元の値はブログ内での各単語の出現頻度である.$\chi^2$値による特徴語選択を行ったものと行っていないものの 2 種類で実験を行う.

クラスタリングの評価の際に用いる品詞は名詞に限定している.名詞に限定したのは生成されたクラスタから内容把握が行えると判断したためである.今回は全ユーザを通して出現回数が 10 回に満たない低出現頻度語は明らかなノイズとなるため除外している.その上で,特徴語選択を行う場合は$\chi^2$値による特徴語抽出を行っている.各デザインに対して 4.2 節の $A$ の値の条件を指定することで特徴語抽出を行い,それぞれ上位 500 語を収集し特徴ベクトルとして用いている.k=20, 40, 60 のクラスタリング結果をエントロピーと純度および k-means の評価関数で評価する.2 種類のクラスタリング結果に対する各値について比較を行う.

また k の値は様々な値で実験し，生成されるクラスタリング結果から設定した．

推薦の評価実験に用いる品詞は名詞だけではなく動詞と形容詞を用いる．3 章でも述べているが，著者の属性を捉える特徴として名詞だけではなく，動詞や形容詞も有効であると考えられる．k=20 の推薦結果について，入力データが持つ正解デザインが割り当てられたクラスタの上位 10 デザインに該当するかで評価する．

特徴語選択の際にすべての単語を対象とするとノイズとなる単語が多く出現することと，一般性に欠ける単語（使用頻度の少ない固有名詞，形態素解析が上手く行えなかった単語等）の $\chi^2$ 値が高くなってしまうことから，出現頻度が 10 回以下の単語を削除した．

### 5.2. 実験結果


図 2：デザインの一例（005, 067, 073）

表 1：図 2 における $\chi^2$ 値が大きい特徴語

| 順位 | デザイン 005 | デザイン 067 | デザイン 073 |
|---|---|---|---|
| 1 位 | 三ツ沢 | マリモ | マタニティヨガ |
| 2 位 | Ｗａｉｖｅ | ブリュレ | マタニティ |
| 3 位 | ＫＥＩＫＯ | 茶釜 | 膣 |
| 4 位 | ベルマーレ | 吉野山 | 葉酸 |
| 5 位 | バンギャ | Anniversary | 産道 |
| 6 位 | マリノス | キュッキュッ | 心音 |
| 7 位 | 善徳 | 飛騨高山 | 胎動 |
| 8 位 | ザスパ | 鰐 | 腹帯 |
| 9 位 | 保則 | 花笠 | ベビ |
| 10 位 | 前節 | 城崎 | 内診 |
| 11 位 | ベイスターズ | 勘助 | 破水 |
| 12 位 | サポ | タラバガニ | 助産 |
| 13 位 | ＤＦ | ホタルイカ | つわり |
| 14 位 | テニプリ | 天然石 | しるし |
| 15 位 | ヴィジュアル | 楊貴妃 | 心拍 |
| 16 位 | 柏レイソル | ウィンク | サラッ |
| 17 位 | ＦＷ | ジョニーデップ | ムカムカ |
| 18 位 | コーエー | だんじり | 安産 |
| 19 位 | ＤＡＩＧＯ | 豊橋 | 分娩 |
| 20 位 | オフライン | 江戸前 | 帝王切開 |

図 2 に挙げた 3 種類のデザインにおける $\chi^2$ 値の大きい特徴語を例として表 1 に，クラスタリング結果の評価を表 2 に，推薦結果の精度を表 3 に示す．また特徴語選択ありの場合の，名詞のみと名詞＋動詞＋形容詞の 2 パターンについて各クラスタのユーザ数を表 4 に示す．

表 2：クラスタリング結果の評価

| | 特徴語選択 | エントロピー | 純度 | 評価関数の値 |
|---|---|---|---|---|
| k=20 | あり | 0.2865645 | 0.7005076 | 508267259 |
| | なし | 0.2909301 | 0.6615905 | 746245176 |
| k=40 | あり | 0.2894898 | 0.6209814 | 294864182 |
| | なし | 0.3695884 | 0.4754653 | 413139617 |
| k=60 | あり | 0.3044614 | 0.5685279 | 198355172 |
| | なし | 0.3250082 | 0.5228426 | 268471316 |

表 3：推薦結果

| | 特徴語選択 | 単語数 | 精度 |
|---|---|---|---|
| 名詞 | あり | 15535 | 0.26667 |
| | なし | 50242 | 0.17778 |
| 名詞＋動詞 | あり | 15993 | 0.11111 |
| | なし | 60836 | 0.11111 |
| 名詞＋形容詞 | あり | 15888 | 0.28889 |
| | なし | 52051 | 0.15556 |
| 名詞＋形容詞＋動詞 | あり | 18256 | 0.17778 |
| | なし | 62589 | 0.13333 |

表 4：各クラスタのユーザ数

| | A | B | | A | B |
|---|---|---|---|---|---|
| クラスタ 1 | 377 | 339 | クラスタ 11 | 3 | 3 |
| クラスタ 2 | 111 | 116 | クラスタ 12 | 2 | 3 |
| クラスタ 3 | 39 | 47 | クラスタ 13 | 2 | 3 |
| クラスタ 4 | 19 | 31 | クラスタ 14 | 2 | 2 |
| クラスタ 5 | 7 | 14 | クラスタ 15 | 2 | 1 |
| クラスタ 6 | 5 | 9 | クラスタ 16 | 1 | 1 |
| クラスタ 7 | 5 | 5 | クラスタ 17 | 1 | 1 |
| クラスタ 8 | 5 | 5 | クラスタ 18 | 1 | 1 |
| クラスタ 9 | 4 | 4 | クラスタ 19 | 1 | 1 |
| クラスタ 10 | 3 | 4 | クラスタ 20 | 1 | 1 |

※A（名詞），B（名詞＋動詞＋形容詞）

### 5.3. 実験結果の妥当性の検証

表 4 から一部のクラスタに多くのユーザが割り当てられていることが確認できる．クラスタリングの偏りには，大きいクラスタにどんどん他のクラスタが吸い寄せられてしまい，偏りを持ったクラスタが形成され

てしまうチェイニング効果がある．一部のクラスタに多くのユーザが割り当てられてしまう問題を検証するために,本節では提案手法とは違う手法で実験を行い，5.2 節で得られた結果が妥当かどうか確認する．階層的クラスタリングの中でチェイニング効果が起こりにくいとされる群平均法を用いて実験を行った．クラスタリング結果のデンドログラムを図 3 に示す.また 591 ユーザの上位出現頻度語を表 5 に示す.

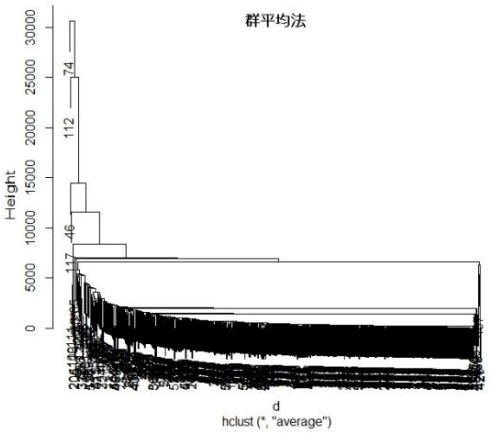


図 3：群平均法による実験結果

表 5：591 ユーザの上位出現頻度語

| 単語 | 出現回数 | 単語 | 出現回数 | 単語 | 出現回数 |
|---|---|---|---|---|---|
| さん | 91661 | 関係 | 9501 | 心配 | 8086 |
| ちゃん | 62482 | 帰り | 9485 | 買い物 | 8056 |
| ママ | 30615 | 家族 | 9416 | 息子 | 7872 |
| 好き | 23562 | 生活 | 9283 | 絵本 | 7758 |
| ♪ | 23446 | 人間 | 9280 | 学校 | 7746 |
| 子供 | 19844 | 病院 | 9064 | 料理 | 7611 |
| お母さん | 19415 | 散歩 | 9051 | あなた | 7435 |
| 友達 | 16377 | 幸せ | 8726 | 朝食 | 6557 |
| パパ | 15764 | あたし | 8619 | 旅行 | 6504 |
| お父さん | 13742 | 映画 | 8462 | プレゼント | 6338 |
| ブログ | 13625 | 久しぶり | 8441 | 携帯 | 6239 |
| ヶ月 | 9707 | 幼稚園 | 8366 | 女性 | 6172 |
| 必要 | 9687 | 誕生 | 8114 | 実家 | 6162 |
| 子ども | 9610 | 撮影 | 8106 | 成長 | 6071 |

図 3 のデンドログラムを見ると，チェイニング効果が起こりにくい手法を用いても左に偏っていることなどから，実験データに偏りがあることが確認できる．また表 5 から子育てや家庭を想像する単語が高い頻度で出現している．表 5 からも実験に使用したデータに偏りが存在していることが確認でき，本提案手法で得られた 5.2 節の実験結果が妥当であると言える.

### 5.4. 考察

$\chi^2$値による特徴語選択を行った方がクラスタリングの評価，推薦の精度ともに，行わない場合に比べ良くなった．しかしながら実験結果を総合的に判断すると全体として精度が低い．理由の一つとして，似ているユーザとデザイン選択の関係性が絶対のものではないことが挙げられる．これは入力ユーザが他の似ているユーザと必ず同じデザインを選ぶとは限らないためである．表 4 から一部のクラスタに割り当てられるユーザが多くなることが確認できる．今回の実験に使用したデータには女性や主婦，妊婦と思われるデータが多く感じられた．実験に用いるデータを確認すると，やはり女性や主婦，妊婦が使うと思われる単語が非常に多く出てきている．このことから本研究に用いたデータに偏りがあることが確認できる.ただ逆に言えば，ブログを用いて似ているユーザ同士をクラスタリングするという本研究の目的に沿った結果であるといえる.また特徴量として使う単語の数をかなり制限してもクラスタリング結果に大きな変化がなかった．このことから$\chi^2$値による特徴量選択が上手く行えていることと，より少ない単語でのクラスタリングが可能であることが言える．また表 4 から動詞と形容詞を追加した方が，各クラスタ内のユーザが分散し偏りが緩くなっている．このことから名詞だけではなく動詞や形容詞が著者の属性推定に有効であることが言える．また表 1 で例として 3 デザインの特徴語を示したが，デザイン 005 はスポーツや芸能について，デザイン 067 は旅行や観光地，デザイン 073 は出産や子育てに関連性が強いと思われる単語が抽出できた．各デザインとユーザがブログで使用した単語にある程度の関係性があると考えられるデータが得られた.

## 6. まとめ

本論文では，ブログの本文テキストから製本デザインを推薦する手法を提案した．評価実験の結果から，テキスト情報と製本デザイン選択にある程度の相関が見られることが分かった.現在では 10 代をはじめ幅広い年代の男女がブログを書いていることから，幅広い年代，分野，性別等の情報が含まれるデータで実験を行ってみる価値はあると考えられる．また特徴量選択の際に，今回は形態素解析後のデータを使用しているため考慮していないが，顔文字・絵文字の使用，ひらがなが使われる割合や語尾の使い方,ブログデザイン，リンク等もユーザの属性を推定する一つの情報として利用できると考えられる．また，著者の属性推定を含め特徴量選択を行うのであれば，形態素解析以外の素

性選択も合わせて利用することが有効であると思われる．それらの要素を盛り込むことで，より類似しているユーザ同士をクラスタリングすることができると思われ，推薦の精度を向上することができると考える．

## 参 考 文 献


[1] 総務省情報通信政策研究所調査研究部，“ブログ・SNS の経済効果の推計”, http://www.soumu.go.jp/main_content/000030547.pdf , 2009.

[2] BizPal ブログ製本サービス MyBooks.jp , http://mybooks.jp/

[3] 池田大介，南野朋之，奥村学，“blog の著者の性別推定”，言語処理学会第 12 回年次大会, 2006.

[4] 大倉務，清水伸幸，中川裕志，“スケーラブルで汎用的なブログ著者属性推定”，情報処理学会研究報告,SIG-NL181,pp.1-6, 2007.

[5] 奥村学，“ブログにおける偏り補正のための書き手のプロファイリング”，人工知能学会誌，vol.23 No.6, pp.798-802, 2008.

[6] 池田大介，高村大也，奥村学，“blog 分類のための半教師有り学習”，情報処理学会研究報告，FI-89,NL-183,pp.59-66, 2008.