Vol.2017-DPS-169 No.12
2017/1/20

トピックモデルによる書籍から Web コンテンツのクロスドメイン推薦方式の実装

石塚 大貴[1]　中沢 実[1]

概要：本論文では，図書館の書籍貸出システムの貸出履歴から各ユーザの嗜好の傾向を分析し，ユーザが興味を示しそうな最新情報が掲載された Web コンテンツを推薦する方法を示す．嗜好の分析には，「Wikipedia の記事文章をトピックモデル化したデータ」,「ユーザごとに貸出書籍をまとめたデータ」,「最新情報が掲載された Web コンテンツのデータ」の３つのデータを使用する．以下３つのデータは，トピックモデル，書籍データ，Web コンテンツと略す．まず，書籍データとトピックモデルの間で単語ヒストグラムの内積を取り，ユーザの嗜好がどのトピックに近いのかを計算する．同じように，トピックモデルと各 Web コンテンツの間でも単語ヒストグラムの内積を取り，各 Web コンテンツがどのトピックに近いのかを計算する．その結果，書籍データと Web コンテンツの間の類似度が分かる．また，書籍データと Web コンテンツの間での中間にトピックモデルを使用することで，書籍データと Web コンテンツの間での単なる類似度ではなく，ユーザの嗜好を考慮した推定ができる．

キーワード：トピックモデル，推薦システム，テキストマイニング，嗜好分析

Implementation of Cross-Domain Recommendation Method of Web Content from Book Data using Topic Model

Daiki Ishizuka[1]　Minoru Nakazawa[1]

Abstract: In this paper, we analyze the preference tendency of user from the history of the book lending system of the library, and then we introduce how to recommend web content of latest information that users are likely to be interested in. For the analysis of user preference, three pieces of data are used: ”data in which a Wikipedia content is modeled as a topic model”, ”data that summarizes lending books for each user” and ”data of the latest Web content”. The three data below are abbreviated as topic model, book data, Web content. At first, an inner product of the word histogram is calculated between the book data and the topic model. Then, which topic is close to the user’s preference is calculated. Next, similarly, the inner product of the word histogram is calculated between the topic model and Web content. Then, which topic is closer to the Web content is calculated. As a result, the degree of similarity between the book data and the Web content is calculated. Furthermore, using a topic model between books and WEB content, It can estimate not only the similarity between books and WEB content but also the user’s preferences.

Keywords: Topic Model, Recommender System, Text Mining , Liking Analysis

1. はじめに

書籍は調べたい事柄に対して詳細な知識を効率よく収集できる素晴らしいツールである．しかし，出版されるまでの間，内容の正確さを確認され，伝わりやすい文章になっているか否か検証されるなど，出版までである程度の時間を要する．そのため，読者に渡るまでに時間がかかり，特に専門分野の内容については新規性に失う場合が多い．また，読書をするための目的が曖昧である場合，本を開くとユーザにとって興味のない内容であることも多々ある．そ

[1] 金沢工業大学
Kanazawa Institute of Technology

のため，本研究ではユーザに読まれている書籍を分析し，そのユーザの嗜好に合った情報が掲載された Web の記事を推薦する方法を提案する．

本システムでは，ユーザが読んでいる書籍の題材として，大学図書館の貸出システムから学部・学科ごとに読まれている書籍の傾向を読み取り，嗜好を Wikipedia のトピックモデルを用いて分類・推定している．また，「この書籍から推定した嗜好の分類」と「事前に Web サイトからスクレイピングした記事の分類」を照合して，各ユーザにおすすめと思われる情報を推薦することを可能にした．

2. 関連研究

近年，トピックモデル [1] による推薦システムの手法に関する研究がよく行われている．

伊藤ら [2] は，Twitter の投稿内容と天気・時期の関係性を表すトピックモデルを提案し，ユーザの嗜好が天気から影響を受けることを示した．しかし，ユーザごとのコンテキストから受ける影響度を調べていないため，全体としての天気・時期による嗜好については把握しているが，ユーザごとの嗜好は考慮されていない．

唐澤ら [3] は，ユーザの入力したキーワードから，トピックモデルを使って新たな類似したキーワードを検索エンジンに与え返されたページ群に対して，更に閲覧履歴によってページの順位付けを行っている．しかし，新しく推薦されるキーワードの精度が低いため，ユーザの嗜好をうまく反映できていない．

冨士谷ら [4] は，テレビ番組に適した書籍の推薦を TF-IDF とトピックモデルを組み合わせることで精度の向上を行っている．この研究ではテレビ番組から書籍という意味で，複数のメディアをまたぐというクロスドメイン推薦を可能にしている．しかし，ユーザの番組視聴履歴などを用いてユーザの嗜好を考慮していない．

3. 推薦のために必要な 3 つのデータ

本システムでは，「Wikipedia の記事文章をトピックモデル化したデータ」「各ユーザごとに貸し出し書籍をまとめたデータ」「最新情報が掲載された Web コンテンツをスクレイピングしたデータ」の 3 つのデータを使用する．3 つのデータは，以降それぞれ “ トピックモデル ”, “ 書籍データ ”, “ Web コンテンツ ” と省略する．

また，日本語では英語のように “ 分かち書き ” されておらず単語の分割が必要なため，随所で MeCab [5] と呼ばれる形態素解析ツールを使用する．

3.1 トピックモデル

ユーザの嗜好の分類にはトピックモデルを用いる．図 1 にトピックモデルの概要を示す．

複数の文書からトピックモデルを生成すると「文書ごとのトピック分布」と「トピックの単語分布」を推定できる．これにより，1 文書に対して複数のトピックを当てはめることができるため，各文章からトピック（傾向）を抽出することができる．また，生成されたモデルの “ トピックの単語分布 ” を用いることにより，トピック間での相関関係や単語同士の関連性を見つけ出すことができる．さらに，トピックモデルでは生成に使用した文書だけでなく，新しい文書に当てはめて応用することができるため，関連性を見つける上で注目されている．

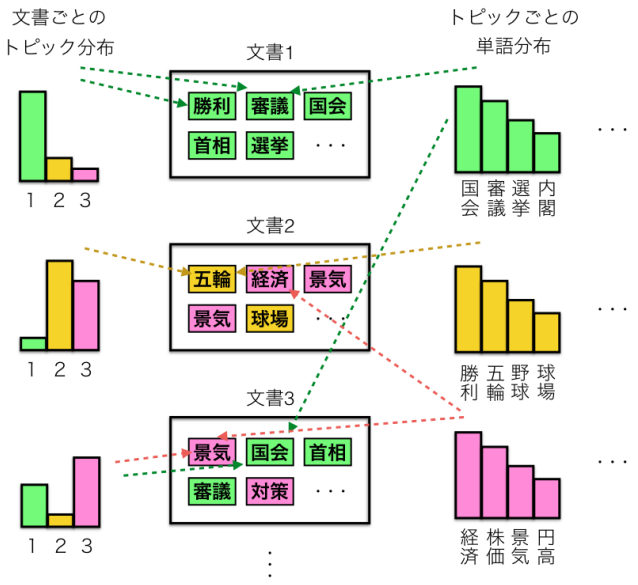

図 1 トピックモデルの概要
（この図は，文献 [1]，p.57，図 4.2 より一部修正して転載してある）

本研究では，あらゆる網羅されている Wikipedia の各記事を文書としてトピックモデルを生成する．生成されたモデルを新しく取得した Web コンテンツに適用することで，ユーザの登録した書籍と Web コンテンツの間での関連性を見つけ出すことを目指す．

まず，トピックモデル作成の具体的な方法として，まず，Wikipedia の記事を 6 万件ほどにサンプリングされた jawiki-latest-abstract.xml [*1] に対して XML のマークアップの除去を行い記事の本文のみを抽出する．得られた各記事の本文をデータセットとして gensim [6] にインプットしトピックモデルを生成する．また，今回のモデル生成では，計算対象の単語を名詞のみに加工し，トピック数が 300 にパラメータを与えた．表 1 に生成されたトピックの一例を示す．

今後はトピック T_k のとき単語 w が生成される確率を ϕ_{kw} とする．

[*1] 全記事データを扱うと計算に膨大な時間を要するため，今回は Wikipedia の記事が既にサンプリングされている https://dumps.wikimedia.org/jawiki/latest/jawiki-latest-abstract.xml から分析用データを作成した．

表 1 生成されたトピックの一例

トピック 1		トピック 2		トピック 3	
単語	生起確率	単語	生起確率	単語	生起確率
大統領	0.01460	関数	0.02526	音楽	0.09380
政府	0.01057	数	0.01867	音	0.04094
アメリカ	0.00777	次	0.00958	演奏	0.03637
軍	0.00632	変数	0.00934	作曲	0.02660
人	0.00609	定義	0.00875	ピアノ	0.00849

3.2 書籍データ

本システムでは，ユーザの嗜好を推定する材料となる書籍データをバーコードリーダでスキャンし，データベースに登録する．また，書籍データの中身である「書籍のタイトル」と「目次」は予め Hanmoto API [7] を用いて取得しておく．

本来は，ユーザの書籍を用いた推薦を行いたいが，読書をする人数が少なくデータ量が圧倒的に不足しているため，今回の研究では，本学図書館の書籍貸出データを用いて，各学科ごとの傾向をみていくことにした．

3.3 Web コンテンツ

ユーザに対しておすすめの記事を紹介するためには，あらかじめ Web からコンテンツを取得して記事の内容を解析する必要がある．そこで記事を収集するためスクレイピングツールを作成した．このツールは，登録した RSS にある記事を自動でスクレイピングしてデータベースに保存してくれるツールである．

今回は新規性の高い Web コンテンツを中心に分析を行いたいため，RSS には比較的新しい情報が手に入りやすそうなサイトに厳選している．例として，はてなブックマーク，主要なニュースサイト群，各ブログサービスの新着記事が含まれる RSS などが挙げられる．

また，ここで注意すべき事柄は，同一ドメイン内の記事の単語群に関して分析すると，類似した単語の頻出が極めて大きいことである．そのため，同一ドメイン内の記事に対していえば，コンテンツそのものの内容よりもドメイン全体の特徴が現れてしまい，記事本来の特徴を考慮した推薦ができない．これを解決するために，筆者は予め記事のコンテキストのみを抽出 [8] し記事本来の特徴だけが推薦にデータ加工を行った．

今回の実験で使用する記事数は 4,769 個である．本研究では，この中から各ユーザが興味のもちそうなものを抽出し推薦することを試みるものである．

4. システム設計と提案手法

4.1 システム設計

本システムでは，3 章で説明した 3 つのデータを用いてユーザの嗜好に合った Web コンテンツを推薦する．本システムの構成を図 2 に示す．

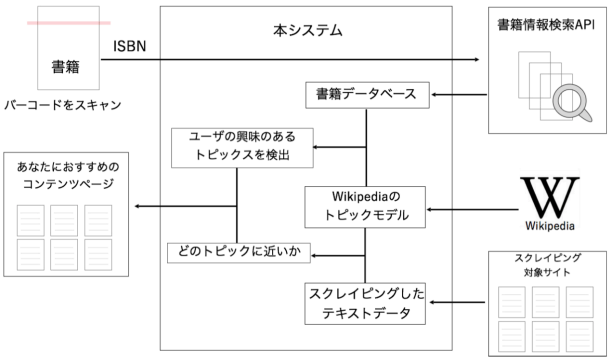

図 2 システム設計

ユーザは読んだ書籍をバーコードリーダーでスキャンして書籍データベースに蓄える．すると，本システムがユーザの書籍データを用いて嗜好の傾向を分析し，その人に合ったおすすめの Web コンテンツ一覧を提供する．予め RSS から取得した最新の Web コンテンツを取得して分析を行うため，新しい情報の場合が多く，その人が興味を示しそうな Web コンテンツを推薦するため，効率的な知識の取得を手助けできる．

本システムでの Web コンテンツ推薦の流れは以下のような流れで行う．

(1) 書籍データとトピックモデルの間で単語ヒストグラムのコサイン類似度を用い，ユーザの嗜好がどのトピックに近いのか計算する．
(2) トピックモデルと各 Web コンテンツの間で単語ヒストグラムのコサイン類似度を用いて，各 Web コンテンツがどのトピックに近いのかを計算する．
(3) 上記 2 つの結果に基づき書籍データから Web コンテンツを推薦する．

4.2 書籍データからユーザーの嗜好の特徴を抽出

はじめに，ユーザ u が本システムに登録した全ての書籍データから「タイトル」と「目次」を抜き出して結合する．結合したテキストを 1 つの文書 B_u として扱う．また，B_u 内での単語 w の生成確率を B_{uw} とする．また，全ての単語集合を W とおく． ここで式 (1) のように，コサイン類似度を用いて計算し，ユーザの嗜好がトピックモデルのトピック T_k にどのくらい近いのかを分析する．

$$cos(B_u, T_k) = \frac{\sum_{i \in W} \phi_{ki} B_{ui}}{\sqrt{\sum_{i \in W} \phi_{ki}^2} \sqrt{\sum_{i \in W} B_{ui}^2}} = \mu_{uk} \quad (1)$$

コサイン類似度の性質より， μ_{uk} はトピック T_k とユーザ u の登録している書籍データ B_u が類似しているほど値が大きくなる．そのため，μ_{uk} が高いほどユーザ u が Wikipedia のトピック T_k に対して嗜好を示しやすいと推定できる．

4.3 スクレイピングした記事データの特徴を抽出

次に，スクレイピングした記事 a の本文を D_a とおき，D_a 内での単語 w の生成確率を D_{aw} とする．書籍データと同じようにコサイン類似度を用いて，Web コンテンツがトピックモデルのトピック T_k にどれくらい近いのかを式 (2) を用いて計算する．

$$cos(D_a, T_k) = \frac{\sum_{i \in W} \phi_{ki} D_{ai}}{\sqrt{\sum_{i \in W} \phi_{ki}^2} \sqrt{\sum_{i \in W} D_{ai}^2}} = \theta_{ak} \quad (2)$$

4.4 記事データから Web コンテンツを推薦

式 (1), (2) から得られた μ_{uk} と θ_{ak} を用いて記事データから Web コンテンツの推薦を行う．まず，rank 関数を定義する． $rank_i(\alpha_i, n)$ は， i を変化させた時に生じる α の全要素について降順に並べたとき上位から n 番目の要素を示す．推薦する手順は二段階に分けられる．

第一段階では，ユーザ u に対して嗜好の度合いを表す μ_{uk} から 1 〜 N 位までのトピックを抜き出し各トピックの生成確率を正規化する．正規化された要素のうち s 番目の値を式 (3) で表す．

$$\lambda_{us} = \frac{rank_t(\mu_{ut}, s)}{\sum_{i=1}^{N} rank_t(\mu_{ut}, i)} \quad (3)$$

例えば，N=3 とし，上位 3 位を表す値がそれぞれ $[\mu_{12u}, \mu_{98u}, \mu_{27u}] = [0.3, 0.1, 0.02]$ とすると，$\lambda_{u1}, \lambda_{u2}, \lambda_{u3}$ のそれぞれの値は $[0.3, 0.1, 0.02]/0.42$ で約 $[0.714, 0.238, 0.0476]$ となる．また，第二段階で数式を分かりやすく示すために，λ_s のトピックの識別子を， γ_s とする．上記の例では $[\gamma_{u1}, \gamma_{u2}, \gamma_{u3}] = [12, 98, 27]$ である．

第二段階では，式 (3) で得られた $\boldsymbol{\lambda}$ を用い，ユーザ u に対して Web コンテンツ a の推薦度を式 (4) で計算する．

$$Recommend(u, a) = \sum_{i}^{N} (\lambda_{u\gamma_i})(\theta_{a\gamma_i}) \quad (4)$$

第 5 章では，式 (4) を用いて Web コンテンツの推薦を行う．

5. 推薦結果

5.1 推薦に不適切なトピックを除外

筆者は，Web コンテンツを推薦するにあたって除外すべきトピックがあると考えた．理由は，式 (2)〜式 (4) を使ってそのまま計算した場合，各学科で共通するトピックが複数存在するため，学科に特化した推定ができなかった．そこで，本研究では特定のトピックを除外することで推薦結果を向上させた．

表 5 に，式 (1) より，各学科の嗜好として推定されたトピックの上位 15 位を示す．表をみると，どの学科にも高い類似度で該当するトピックが存在する．例として，トピック T_{17} は，全ての学科にて類似度が 1 位と推定されているため，もし，式 (2) より推定された任意の Web コンテンツのトピックが T_{17} に近い場合，その Web コンテンツがどの学科にも推薦されてしまう．これでは，ユーザの嗜好を考慮した推薦ができているとは言えない．

表 2 トピックを除外する前の各学科に対して推定されたトピック

	1 位	2 位	3 位	4 位	5 位
機械工	T_{17}	T_{122}	T_{180}	T_{209}	T_{124}
航空システム	T_{17}	T_{171}	T_{84}	T_{124}	T_{205}
ロボティクス	T_{17}	T_{30}	T_{176}	T_{228}	T_{171}
電気電子	T_{17}	T_{287}	T_{228}	T_{57}	T_{209}
情報工	T_{17}	T_{228}	T_{176}	T_{102}	T_{84}
	6 位	7 位	8 位	9 位	10 位
機械工	T_{84}	T_{38}	T_{87}	T_{120}	T_{203}
航空システム	T_{180}	T_{38}	T_{18}	T_{87}	T_{65}
ロボティクス	T_{84}	T_{209}	T_{180}	T_{38}	T_{124}
電気電子	T_{124}	T_{180}	T_{84}	T_{205}	T_{38}
情報工	T_{180}	T_{209}	T_{87}	T_{124}	T_{38}
	11 位	12 位	13 位	14 位	15 位
機械工	T_{88}	T_{127}	T_{287}	T_{228}	T_{98}
航空システム	T_{209}	T_{120}	T_{122}	T_{127}	T_{197}
ロボティクス	T_{127}	T_{87}	T_{205}	T_{273}	T_{120}
電気電子	T_{87}	T_{260}	T_{120}	T_{203}	T_{127}
情報工	T_{205}	T_{120}	T_{273}	T_{127}	T_{28}

今回の推薦ではそれぞれのユーザに対して適切かつある程度特徴のある Web コンテンツの推薦を行いたいため，各ユーザのトピック群から共通のトピックを除去することにした．具体的には，学科それぞれのトピックの上位 15 位の中で 4 学科以上がマッチしているトピックを推薦から除外することで共通な関心を反映させないことにした．表 3 に各学科にて上位 15 位トピックについて 4 学科以上該当したトピックの一覧を挙げる．

表 3 各学科で共通のトピック一覧

トピック	数	トピック内の単語の生起確率
T_{17}	5	章，旗，国旗，紋章，制定
T_{38}	5	大学，研究，学，学科
T_{84}	5	人，日本，文化，歴史，日本人
T_{87}	5	社会，主義，政治，経済，学
T_{120}	5	学，性，意味，理論，研究
T_{124}	5	日本，研究，研究所，会，年
T_{127}	5	書，ユダヤ，聖書，イスラエル，人
T_{180}	5	者，利用，削除，ロード，アップ
T_{209}	5	システム，化，技術，処理，開発
T_{205}	4	版，ページ，日本語，英語，編集
T_{228}	4	情報，通信，ネットワーク，インターネット，機器

表の左側は対象のトピック，真ん中は各学科の上位 15 位のトピック中で該当する学科数，右側がトピック内の単語の生起確率を示す．トピック内の単語の正規確率は左側から順に上位 5 単語を表示している．上位 15 位の中で 4 学科以上が該当するトピックを推薦から除外することで精度を向上させることができた．筆者は，表 3 の共通のトピック一覧から，5 つカテゴリに分けることできると考える．表

4 に共通したトピックをカテゴリに分類したものを載せる．

表 4 共通したトピックのカテゴリ分け

カテゴリ	トピック
書籍が含まれるトピック	T_{17}
Wikipedia が含まれるトピック	T_{180}, T_{205}
大学や研究が含まれるトピック	T_{38}, T_{120}, T_{124}
人文系科目が含まれるトピック	T_{84}, T_{87}
工学系用語が含まれるトピック	T_{209}，T_{228}

上 2 つのカテゴリは，主に Web コンテント本来の内容を推薦することが難しいトピック群である．書籍が含まれるトピックは T_{17} のみであり，単語の生起確率の中の第 1 位に「章」が存在する．本システムでは書籍データの中で扱う材料として書籍の目次を扱っていため，各ユーザが登録した書籍により「章」を含むトピック T_{17} が推定されやすいと考えられる．また，Wikipedia が含まれるトピックには，T_{180} と T_{205} がある．トピック T_{180} では，「利用者」，「削除」，「アップロード」などの単語，トピック T_{205} では，「日本語」，「英語」，「ページ」，「編集」などの単語が属する．これらのトピックは，Wikipedia の利用方法のページなどのトピックが推定されたと考えられる．以上より，この 2 つのカテゴリでは，Web コンテント本来の情報を取得することはできないため，対象から除外すべきである．

下 3 つのトピックのカテゴリは，主に各学科共通のトピック群といえる．大学や研究が含まれるトピックでは，全ての学科に当てはまる大学や研究に対してのトピックが存在する．人文系科目に含まれるトピックでは，全学科で必修とされている科目群があり，参考文献を用いて授業に臨むことが求められるため，全学科の共通のトピックとして現れると考えられる．また，工学系用語が含まれるトピックでは，工学系の大学であるため推定されていると考えられる．以上より，各学科に対しての嗜好を考慮して Web コンテンツを推定するためには，この 3 つのカテゴリについても除外すべきだと考えられる．

実際に本システムでは，5 つのカテゴリに属するトピックを除外して Web コンテンツの推薦を行った．

5.2 各学科に対して推定されたトピック

表 5 に各学科に対して推定されたトピック一覧を示す．これは，式 (3) により推定された結果である．ただし，rank 関数では，5.1 において除外したトピックを含めない結果を返すものとする．表 5 より共通トピックを除去した後の推定結果は，ある程度，学科の傾向を踏まえていると言える．

5.3 学科ごとに推薦された Web コンテンツ

5.2 で推定した結果を式 (4) に用い，学科に対しての各記事の推薦度を計算した．表 6 に各学科に対して推薦された Web コンテンツのタイトルの例を示す．

全体的に見ると，ある程度，各学科の嗜好を考慮した Web コンテンツを推薦できていることが分かる．

5.4 推薦に関する問題

これまでの流れでユーザの嗜好を考慮した Web コンテンツを推薦することができた．しかし，問題点がいくつか存在する．

まず，機械工学科と電気電子学科に対しての推薦をみると“スコットランド初の「潮力発電所」が稼働へ”というタイトルが両学科に 1 番おすすめとして推薦されている．両学科では普段異なる分野を学習しており，嗜好が異なるとかんがえられるため，あまり適切な推薦とは言えない．

また，ロボティクス学科に注目すると，推薦された Web コンテンツの 1 位と 3 位で“J-deite RIDE”についての内容が推薦されている．また，Web コンテンツの内容に関してもかなり似通っており，推薦される側としては助長だと感じてしまう．ここでは，重複した記事を推薦しないように抑制する仕組みが必要である．

この他にも，Web コンテンツの内容の文字数が少ない場合には，非常に高い推薦度で Web コンテンツを推薦してしまう問題点があるため，文字数による抑制の仕組みも必要である．また，画像ばかりを多様している Web コンテンツに対してもうまく推薦できない．これは，本システムでは Web コンテンツの内容の中でもテキストのみしか考慮できないために生じる問題である．画像や動画等も考慮するべき仕組みを作る必要があると考えられる．

6. まとめ

本論文では，ユーザが登録した書籍から嗜好の傾向を見つけ出し，おすすめの Web コンテンツを推薦する方法を提案した．この手法では Wikipedia からトピックモデルを生成し，このモデルを用いて書籍データと Web コンテンツの間の関係性を築いて推薦を行った．書籍データと Web コンテンツの中間にトピックモデルを用いることで単なる類似度ではなく，ある程度，ユーザの嗜好を考慮した推定ができることが分かった．

また，本研究では，「ユーザの嗜好を傾向を見つけ出す際の精度が低いこと」や「文字数が少ない Web コンテンツを誤って推薦度を高く評価してしまうこと」，「画像や映像を考慮できていないこと」などたくさんの課題がある．今後，ユーザの嗜好分析と Web コンテンツの推薦度の精度を向上させていくためにはこれらの課題を一つずつ解決していく必要がある．

表 5 各学科に推定されたトピック

	推定されたトピック				
学科	1 位	2 位	3 位	4 位	5 位
機械工	122	203	88	287	98
	熱	式	生物	発電	光
	温度	器	学	所	使用
	エネルギー	用	環境	電力	性
航空システム	171	18	65	122	197
	宇宙	機	法則	熱	空港
	地球	型	力	温度	航空
	計画	エンジン	ソープ	エネルギー	国際
ロボティクス	30	176	171	273	88
	ロボット	画像	宇宙	精神	生物
	潜水	撮影	地球	心理	学
	ウェイ	位置	計画	科学	環境
電気電子	287	57	260	203	6
	発電	子	平成	式	物理
	所	電子	年度	器	核
	電力	軌道	号機	用	模型
情報工	176	102	273	28	263
	画像	ゲーム	精神	クラス	生物
	撮影	ソフト	心理	指向	学
	位置	コンピュータ	科学	オブジェクト	環境

表 6 各学科に対して推定された Web コンテンツのタイトルの例

学科	順位	推薦度	タイトル
機械工	1	0.019325	スコットランド初の「潮力発電所」が稼働へ
	2	0.017625	カーボンナノチューブとゴムで熱界面材料を開発
	3	0.012342	バクテリア：細胞分裂、高等生物と似た仕組み
航空システム	1	0.020294	中国、新型大型ロケット「長征 5 号」の初打ち上げに成功
	2	0.018064	遠い昔、光は今より速かった? - ICL が光速度不変の原理を覆す仮説を検証
	3	0.017724	『宇宙エレベーター　その実現性を探る』宇宙エレベーターの入門書
ロボティクス	1	0.040771	車が人型に変身する搭乗型ロボット「J-deite RIDE」、2017 年完成予定　変形所要時間は 10 秒以内を想定
	2	0.031849	ロボットが働いて、人間は遊んで暮らしたい
	3	0.025575	乗用可能な人型変形ロボット「J-deite RIDE」が来年に開発完了予定
電気電子	1	0.029807	スコットランド初の「潮力発電所」が稼働へ
	2	0.014471	アマゾン、家全体をカバー可能なネットギア製未発表ルーターを予約開始
	3	0.012949	福島第一原発 新たな「異常なし」
情報工	1	0.023590	3 分で振り返る　Nintendo Switch ニンテンドースイッチ以前の据え置きゲーム
	2	0.022159	任天堂：復刻版ファミコン、２６万台販売…発売４日間
	3	0.017586	ゲームエンジン「Frostbite Engine」に手応えを感じる EA 幹部。26 ものゲームエンジンがひとつに統合へ

参考文献

[1] 岩田 具治：トピックモデル，講談社（2015）
[2] 伊藤 拓，朱 丹丹，深澤 佑介，太田 順，“ 天気・時期コンテキストを考慮したトピックモデル ”，マルチメディア通信と分散処理学会，2015-DPS-163，No.34，2015
[3] 唐澤 貴大，中沢 実，“ トピックモデルと WEB 閲覧履歴によるユーザの意図を考慮した検索システムの開発 ”，マルチメディア通信と分散処理学会，2015-DPS-165，No.11，2015
[4] 冨士谷 康，村尾 和哉，望月 祐洋，西尾 信彦，“ コンテンツの多様性を考慮したクロスドメイン推薦 ”，DEIM Forum 2016 C2-7，2016
[5] MeCab: Yet Another Part-of-Speech and Morphological Analyzer：http://taku910.github.io/mecab/
[6] gensim: Topic modelling for humans：https://radimrehurek.com/gensim/
[7] 版元ドットコム API の概要 ：http://www.hanmoto.com/about_api
[8] Python でブログの HTML から本文抽出 2015：http://orangain.hatenablog.com/entry/content-extraction-from-html-in-python