研究論文

基底選択を用いた話題性の評価
Topic Estimation Method Using Base Selection

輪島 幸治 † 古川 利博 †† 佐藤 哲司 ‡
Koji WAJIMA Toshihiro FURUKAWA Tetsuji SATOH

† 筑波大学 図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba
〒 305-8550 茨城県つくば市春日 1-2
E-mail: kwajima@ce.slis.tsukuba.ac.jp
†† 東京理科大学 工学部情報工学科
〒 125-5846 東京都葛飾区新宿 6-3-1
Dept.Information and Computer Technology., Science University of Tokyo
‡ 筑波大学 図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba
〒 305-8550 茨城県つくば市春日 1-2
E-mail: satoh@ce.slis.tsukuba.ac.jp

近年，ユーザが情報発信する CGM(Consumer Generated Media) やソーシャルメディアが台頭する情報化社会へと移行してきている．商品やサービスの利用者であるユーザが情報交換を行うオンラインコミュニティは誰もが質問・回答できる利点がある．オンラインコミュニティは，誰もが閲覧できる環境であることから投稿された質問記事に対して，適切な対応をとり続けることが，炎上などの社会的な影響を回避する上で不可欠となっている．本論文では，オンラインコミュニティに投稿された質問記事から，記事が言及している話題性を評価するのに有効な基底の選択手法を提案する．提案手法は，非負値行列因子分解 (NMF) による特徴量変換および基底選択で構成されている．提案手法では，既存研究のテキスト情報の特徴量である表層情報，語種や品詞，文末表現など 2,000 次元を越える特徴量を用いる．提案手法はを複数のオンラインコミュニティ記事に適用し，サポートベクタ回帰モデル (SVR) と分類器を用いて評価した結果，基底選択の有効性が確認できたので報告する．

The social concern with communication between a consumer and maker has been growing for the last several years. A large number of unspecified users are investigating Consumer Generated Media's communication on the Internet. The claim has the tremendous impact on the purchasing behavior in Consumer Generated Media. This paper is intended as an investigation of significant feature quantity for topic forecast. In this Paper, We investigate online media's topic forecast on Online Community article. We comprehensively extracted text information's feature quantity using existing research. Feature quantity is used, for 31 types and 2,071 dimensions in text information. Proposed Method consists of feature transformation and the base on evaluation using Nonnegative Matrix Factorization (NMF). The validity of the Proposed Method is verified by Support Vector Regression (SVR) and Classifiers. It was found from the evaluative result that have an impact on view count. We report that research result.

キーワード: NMF，LSI，LDA，SVR，質問記事

1 はじめに

近年，CGM(Consumer Generated Media) やソーシャルメディアが台頭する情報化社会へと移行してきている．例えば，ソーシャルメディアの情報は，企業の評判分析など多様な分野で用いられている[1]．

[1]Oracle Help Center - Cloud Documentation : https://docs.oracle.com/en/cloud/saas/index.html

CGM やソーシャルメディアなどのオンラインコミュニティは誰もが気軽に閲覧・質問・回答が行える．このため，情報収集や気楽なコミュニケーション，時間をかけた議論など目的に応じた様々な利点が示されている．また，商品やサービスの利用者が，内容や使い勝手などの情報交換も盛んに行われている．そこで共有･共感された価値ある情報は，情報伝播や購買行動に影響を与えることが大きい．共感に基づいた生活者消費行動モデルは，SIPS（Sympathize Identify Participate Share & Spread）[1] [2] [3] と呼ばれる．一方で，クレームや批判的なコメントも急激に広がる．このため，企業は話題性やクレームがある質問記事を検知・予測し，早急に対処することが欠かせない．

本研究では，質問記事における話題性の多寡を推定することを目的に基底選択を用いた質問記事の評価手法を提案する．本研究における質問記事の話題性は閲覧数が多い質問記事とする．閲覧数が多いことは，多くの利用者が，共感・関心を持って質問記事を閲覧している状態だと言える．このため，質問記事の閲覧数に基づいて評価する．異常や変化の兆しに基づく閲覧数に基づいた特徴が明らかになることで，コンテンツの内容で，投稿後に話題になる質問記事を検知・予測することが期待できる．

提案手法では，質問記事からできる限り多くの特徴量を抽出し，非負値行列因子分解（NMF）によって，特徴量を変換する．特徴量の変換後に得られる基底を質問記事集合の特性に基づいて評価し，基底選択する．本研究では，各質問記事で，コンテンツが異なる特性に着目した．質問記事を高次元なベクトル化した場合，0 の多いスパースなベクトルになる傾向にある．したがって，各コンテンツで，観測ベクトルは大きく異なる．そこで，提案手法では，変換特徴量である基底に対する各質問記事の重みに基づいて，閲覧数が多いコンテンツにおける基底を評価した．提案手法では，既存研究におけるテキスト情報の特徴量を網羅的に抽出した 31 個，2,071 次元の特徴量を用いている．

提案手法の有効性はサポートベクタ回帰モデル (SVR) と分類器を用いて評価する．SVR を用いた閲覧数の予測では平均絶対誤差 (MAE) および平均二乗平方根誤差 (RMSE) の評価指標で有効な結果が得られた．また，分類器を用いた評価では，適合率，再現率，F 値の評価指標で有効な結果が得られた．

本論文では 2 章で関連研究に関して述べる．3 章で提案手法である基底選択の評価手法を詳述する．4 章では，提案手法の実装と評価対象について述べ，5 章で実験結果を示す．6 章では，得られた結果に対する考察を論じる．最後に 7 章でまとめと今後の課題を示す．

2 関連研究

本研究では，オンラインコミュニティのコンテンツである質問記事のテキスト情報から話題性の評価を試みた．2.1 節および 2.2 節で，オンラインコミュニティおよび話題性に関する既存研究を概観する．2.3 節，2.4 節，2.5 節で，本研究で評価要素として用いる特徴量に関する研究を述べる．

2.1 cQA サイトを対象とした研究

本研究の対象は，情報発信メディアの一つで，ユーザが相互に質問・回答を行うオンラインコミュニティの cQA サイト (Community based question-answering service) である．cQA サイトを対象にした研究には，類似質問の同定，回答の評価，要約の作成など数多くの研究がある [2]．質問記事内から特徴量を抽出し，ベストアンサーを推定する研究 [3]，質問記事の文体に着目し，最適な回答を提示する研究 [4]，質問記事に教師付き機械学習アルゴリズムを適用させ，既存のナレッジベースから類似の質問記事を検索する研究 [5] などがある．テキスト情報を用いたオンラインコミュニティの場合，各オンラインコミュニティで投稿者が異なる．したがって，投稿内容や特徴的な単語などは異なる．本研究では，オンラインコミュニティの話題性が評価対象である．

2.2 話題性

オンライン上のメディアを対象とした話題や動向の評価は，多くの評価方法がある．既存研究では，

[2] SIPS では，生活者消費行動を S(Sympathize：共感する)，I(Identify：確認する)，P(Participate：参加する)，S(Share & Spread：共有・拡散する) とモデル化している．SIPS モデルでは，共感に次ぐ重要な要素は，参加してもらうことである．参加はエバンジェリスト (伝道者)，ロイヤルカスタマー (支援者)，ファン (応援者)，パーティシパント (参加者) など，企業やブランドの生涯顧客価値を高めていく過程と重なっている．

[3] SIPS：http://www.dentsu.co.jp/sips/index.html

質問記事のコンテンツであるテキスト情報から，出現頻度や推移に着目した評価が行われることが多い．評価では統計量や重要語などの要素が用いられる [6,7]．また，ソーシャルブックマーク[4]を用いたブックマークの周期性の分析や検索の時期や検索結果のランキング手法の研究 [8,9]，ユーザーの潜在変数などのメタ情報を用いて情報伝播を分析する方法などがある [10]．加えて，話題性の流行や人気に影響を与える要素には，コンテンツに加え，クチコミ [11] [5]や共感 [1] [6]，技術のハイプ・サイクル (Gartner Hype Cycle) [12] [7][8] などもある．本研究では，オンラインコミュニティである cQA サイト上の質問記事のテキスト情報を用いて，特徴量を抽出・特徴量変換し，話題性の評価に適用する．

テキスト情報を用いた話題性評価の研究には，トレンドキーワード (流行語) やトピックを用いた手法がある．トレンドキーワードとは，検索エンジンやソーシャルメディアで，利用者から興味関心が高いキーワードである．検索エンジンの検索語であるクエリの頻度などで抽出が行われる．トレンドキーワードのウェブリソース間の振る舞いに関する研究や，コミュニティにおける発言割合の研究などがある [13,14]．2.4 節で後述するテキスト情報が言及する話題 (トピック) を用いた手法では，アルゴリズムでテキスト情報からトピックと呼ばれる単語集合を抽出し，時系列や種別に基づいて流行や人気を評価する [15,16]．話題性の評価基準には，トピックのバースト[9]度合い [16] や盛り上がりの早さや平均返信数 [17]，ユーザ行動やコミュニティの成長率 [18] などが用いられている．

2.3 テキスト情報の特徴量に関する研究

本研究で使用する既存研究のテキスト情報の特徴を表 1 に示す．表 1 の特徴はいずれも文章の表層的な特徴であることから，これらの特徴を表層情報と称する．本研究の文字種は，ひらがな，カタカナ，漢字，アルファベット，数字，半角記号，空白記号，全角記号の 8 個である．

表 1: 表層情報

項番	特徴名	次元数	値の定義/例
1	文数	1	[。][？][！] の合計頻度
2	読点	1	[、] の頻度
3	句点	1	[。] の頻度
4	文長	1	文字数
5	文字種 (1)	8	各文字種の頻度*
6	文字種 (2)	8	各文字種の比率*
7	読点間距離	1	読点間の距離 [19]
8	漢字含有率	1	漢字の包含率 [19]

2.4 話題抽出に関する研究

テキスト情報の特徴量に関する研究に，話題抽出がある．話題抽出で用いられる基本的なアルゴリズムには，潜在意味解析 (LSI : Latent Semantic Indexing) とトピックモデル (LDA : Latent Dirichlet Allocation) がある [15]．潜在意味解析やトピックモデルを適用し，得られる結果は類似した語彙の集合であり，本研究で述べる話題に相当する．以下，特に断らない場合は，本研究では話題抽出アルゴリズムの結果をトピックとし，話題あるいは話題抽出を言う．潜在意味解析やトピックモデルの詳細は，文献 [15] にゆずり，ここでは，概略を述べるに留めることにする．

話題抽出のアルゴリズムを適用する文書集合は，1 行が 1 文書，各列は語彙に相当し，要素は語彙 V の出現頻度を持つ行列 N である．LSI は，文書集合を低ランク行列の積 $U^{T}H$ にそれぞれの行列の要素を二乗し，総和をとったものが最小になるように行列分解する手法である．$U = (u_1, \cdots, u_D)$ は，K 行 D 列の重み付き係数行列である．$H = (h_1, \cdots, h_V)$ は，K 行 V 列の語彙行列である．重み付き係数行列における U の要素は，文書に対する次元 K の重みに相当する．本研究の LSI の特徴量は，U の要素を用いた．

[4]ソーシャルブックマークとは，編集したブックマークをインターネット上に公開できるウェブサイトである．

[5]インターネットでは，各種商品やサービスなどに関する利用者 (消費者側) の評価・体験談の投稿など．

[6]他人の体験する感情を自分の体験のように感じること．

[7]新しいテクノロジが時間の経過と共にどのように発展していくのかのトレンドを示す先進テクノロジの成熟度と採用率のグラフ．

[8]Gartner - Research & Advisory Overview : https://www.gartner.com/en/research/methodologies

[9]特定の話題が起因して，トピックの頻度が急上昇することで検出される現象．

*Unicode 10.0 Character Code Charts http://www.unicode.org/charts/

LDA は，文書 w_d が低ランク行列 ϕ と θ をパラメータとして持つカテゴリ分布から生成されると仮定する手法である．パラメータ θ_d はトピック分布，パラメータ Φ は単語分布集合である．本研究の LDA の特徴量は，文書集合 N に対し，LDA を適用し得られるトピック K の各文書の割り当て確率であるトピック分布 θ である．

LSI における低ランク行列の次元数 K および LDA におけるトピック数 K が話題に相当する．LSI の次元数や LDA のトピック数は任意に設定できるが，4.1 節で詳述するが，本論文では十分に大きな値 300 とした．話題の特徴量を表 2 に示す．

表 2: 話題の特徴量の詳細

項番	特徴名	次元数	文献
1	LSI (重み付き係数 u_D)	300	[15, 40]
2	LDA (トピック分布 θ_d)	300	[15, 40]

2.5 意味情報に関する研究

意味情報は，個々の研究で言語や形態論に基づき，異なる特徴が用いられている [20]．本研究においては，社会システム理論のゼマンティクという概念に基づき，既存研究の辞書を用いる [21]．ゼマンティクとは，高度に一般化され，状況に依存せずに使用できる意味である．既存研究の辞書を用いることで，状況に依存せずに意味情報が特徴量として抽出できる．本研究では，学術目的の言語資源として，従来研究で作成された日本語の辞書のうち，言語資源の利用事例や応用研究がある既存辞書 21 個を選定した．本研究で用いる 21 個の辞書を表 3 に示す．

表 3 の，項番 15 の拡張固有表現の辞書は，関根の拡張固有表現階層の定義に基づき，Wikipedia の見出し語に対し，固有表現クラスを付与した辞書である．また，項番 18-20 で用いている単語感情極性対応表では，日本語の辞書の見出し語と読みを用いた．

加えて，表 3 の項番 1-7 の辞書は調査研究目的に作成された辞書に基づいている．項番 1，3，6 の辞書は意味分類体語彙表，項番 2，5 の辞書は日本語教育基本語彙，項番 4，7 の辞書は分類項目一覧表である．

本研究では，下記に記載の基準で，辞書の加工および見出し語を選定した．

1. 見出し語の加工
空白記号,「-」「0」「など」を除去

2. 対象外の見出し語
「」「-」「〔」「→」「/」「(」「」「その他」を含む語

表 3: 既存研究の特徴量の詳細

項番	特徴名	次元数	文献
1	語種	7	†, [22]
2	基本語 (1)	2	†, [23]
3	基本語 (2)	6	†, [23]
4	基本語 (3)	2	†, [23]
5	意味分類 (1)	233	†, [24]
6	意味分類 (2)	487	†, [24]
7	意味分類 (3)	307	†, [24]
8	機能表現	122	‡, [25]
9	文末モダリティ	32	[26]
10	質問文末表現	38	[27]
11	IPA 品詞	14	§, [28]
12	固有名詞	4	§, [28]
13	名詞比率	1	[29]
14	MVR	1	[29]
15	拡張固有表現	132	¶, [30]
16	評価極性情報（用言編）	4	‡, [31]
17	（名詞編）	51	‡, [32]
18	感情極性（頻度）	2	[33] ‖, [33, 34]
19	（比率）	2	‖, [33, 34]
20	（平均値）	1	‖, [33, 34]
21	評価値表現	1	**, [35]

†『日本語教育のための基本語彙調査』データ
http://mmsrv.ninjal.ac.jp/bvjsl84/

‡機能表現タグ付与コーパス，日本語評価極性辞書
http://www.cl.ecei.tohoku.ac.jp/index.php

§ipadic version 2.7.0 ユーザーズマニュアル
http://chasen.naist.jp/snapshot/ipadic/ipadic/doc/ipadic-ja.pdf

¶NAIST Japanese ENE Dictionary on Wikipedia
https://github.com/masayu-a/NAIST-JENE

‖単語感情極性対応表
http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html

**評価値表現辞書

本研究における意味情報は，表 3 で示した項番 1 の語種，項番 2-4 の基本語，項番 5-7 意味属性，項番 8 の言語表現，項番 9 の文末表現，項番 11-14 の品詞，項番 15 の固有表現，項番 16-21 の評価表現の 8 個を評価に用いる．

3 提案手法

3.1 概要

オンラインコミュニティでは，多くのユーザが閲覧するため，社会的な影響も大きい．閲覧数が多い質問記事は，多くのユーザの疑問を解消，あるいは興味と合致した重要な質問記事であると言える．提案手法では，重要な質問の判別に有効な基底選択の手法を提案する．

本研究ではグレゴリー・ベイトソン[10] の情報の定義に着目している．ベイトソンは情報を「“違い” を生む “違い”」であると定義している [36]．“違い” を土地と地図の差異に例えた場合，土地には，高低，建造物，人口の分布など，多様な要素がある [37]．また，土地が違えば地図も異なる．地図には，高低図，街の配置図，人口分布図などがある．地図は土地の特定の要素を選択した結果である．地図と他の地図との差異は，差異に基づく「“違い” を生む “違い”」である．

ここで，地図を変換特徴量とみなすことで，ベイトソンの情報の定義とみなすことができる．ゆえに，変換特徴量に対する係数値の差で，変換特徴量の差異を表すことができると言える．そこで，基底に対する係数値の差に着目した．

提案手法では，特徴量の変換を行い，重要な基底を評価する．特徴量変換には，主成分分析や独立成分分析など，これまで多くの多変量解析手法を用いた手法が提案されている．

本研究では，他の特徴量変換を用いた次元削減手法と比較し，優れた利点を持つ非負値行列因子分解 (NMF:Non-negative Matrix Factorization) を用いた特徴量変換を行う [38] [39]．3.4 節で詳述するが，NMF の結果である基底は，特徴量の共起成分がグルーピングされた結果である [38]．

各基底では，特徴量は基底に対して，寄与率を持つ．基底における寄与率上位の特徴量は特徴量をグルーピングした性質に対して寄与が大きい特徴量である．基底を選択することで，グルーピングした性質である “違い” を表すことができると言える．

提案手法の概要を図 1 に示す．図 1 の (1) から (4) は，提案手法の手順を示すステップである．

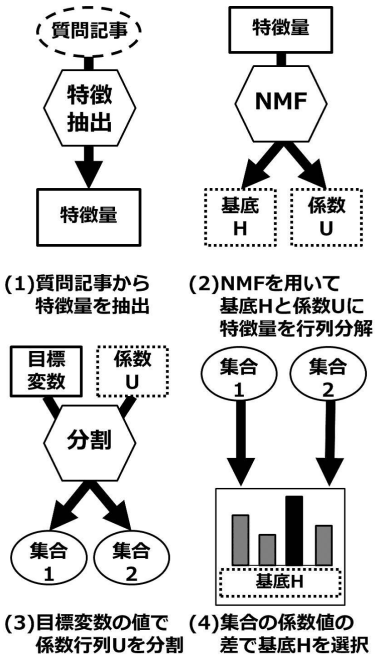

図 1: 提案手法の概要

図 1 の特徴量の抽出では 2.3 節，2.4 節，2.5 節に示した特徴量を抽出している．本研究は，31 個の特徴量を用いており，既存研究のテキスト情報を網羅的に抽出し，有効性を論じているところに特徴がある．各特徴量を水平方向に連結した際の次元数は 2,071 次元である．網羅的に特徴を抽出することで，詳細な観測ベクトルとする．質問記事から抽出する特徴量は，従来研究で言語資源として用いられているテキスト情報の特徴量である．

そしてを図 1(2) から (4) が，本論文で提案する，目的変数に基づいて重要な基底を選択する手順である．5.2 節で後述するが，本研究の目的変数には，話題性の評価を目的に質問記事の閲覧数を用いた．

本章では，まず，3.2 節で特徴量の変換手法を示す．そして，3.3 節で分解結果を目的変数で分割する方法を述べる．最後に 3.4 節で目的変数に基づく重要な基底を評価する方法を示す．

---

http://www.syncha.org/evaluative_expressions.html

[10] アメリカの人類学者 (1904-1980)，民族誌「ナベン」「バリ」や人間関係論におけるダブル・バインド論など，文化とパーソナリティ，コミュニケーションに関する理論で，学際的な功績を残した．

3.2 NMF を用いた特徴量の変換

特徴量の変換に NMF を用いる．NMF は，観測行列 Y を基底行列 H と係数行列 U の積に分解するアルゴリズムである．以後，文書の特徴量である観測ベクトルを並べた行列を観測データ行列と見なし，データをベクトルで表記する．また，本研究で NMF を適用する観測行列は，2.3 節，2.4 節，2.5 節の各特徴量を質問記事より抽出し，水平方向に連結した行列である．

NMF の簡略化した分解表現を式 (1) に示す．

$$Y \simeq HU \tag{1}$$

式 (1) の観測行列 Y は，文書数 $\big(i = 1, \ldots, N\big)$，特徴量の次元数 $\big(j = 1, \ldots, K\big)$ から構成される N 行 K 列の長方行列である．N 行は文書数である．次元数 K は特徴量の数であり 2,071 である．したがって，本研究の観測行列 Y は，N 行 2,071 列の長方行列である．ここで，観測行列 Y は文書の特徴量である観測ベクトルを並べた行列である．また，基底行列 H は行列分解後の基底ベクトルを並べた行列である．

NMF では，観測行列 Y の次元数 K よりも，基底数 M を小さく設定することで，特徴量変換が行える．基底行列 H の基底数を $\big(m = 1, \ldots, M\big)$ とした際の NMF による観測行列の分解を式 (2) に示す．

$$\big(y_{i,j}\big)_{NK} \simeq \Sigma_{m=1}^{M} h_{j,m} u_{m,i} \tag{2}$$

式 (2) の $(y_{i,j})_{NK}$ は観測行列 Y を表す．また，$h_{j,m}$ は基底行列 H の成分でありベクトルである．$u_{m,i}$ は係数行列 U の成分を表す．ここで，$\Sigma_{m=1}^{M} h_{j,m} u_{m,i}$ は，$h_{j,m}^{T} u_{m,i}$ に相当する．したがって，$h_{j,m}^{T} u_{m,i}$ は，観測行列 $(y_{i,j})_{NK}$ と等しくなるべき値である．しかし，行列分解では一般に誤差が発生する．

ゆえに，NMF の行列分解は，観測ベクトルを並べた行列である観測行列 Y を規準の定義に応じて H，U の誤差を最小化する行列 H，U を求める最適化問題に帰着する．最適解が所望の解であるためには，背後にある観測行列 Y の生成プロセスに合った適切な規準が必要である．NMF では，乖離度規準を定義し，規準に基づく更新式の反復で最適解を求める [38]．本研究では，観測行列 Y の生成プロセスに，非負の整数の確率分布である Poisson 分布を仮定し，乖離度規準には一般化 Kullback-Leibler ダイバージェンス [38] を採用した．

観測行列 Y と行列 H，U の乖離度規準 $D(Y, HU)$ を式 (3) に，乖離度規準に基づく最適化する更新式を式 (4) に示す．

$$D(Y, HU) = y_{i,j} log \frac{y_{i,j}}{h_{j,m}^{T} u_{m,i}} - y_{i,j} + h_{j,m}^{T} u_{m,i} \tag{3}$$

$$h_{j,m} \leftarrow h_{j,m} \frac{\Sigma_i y_{i,j} u_{m,i} / x_{i,j}}{\Sigma_i u_{m,i}}$$
$$u_{m,i} \leftarrow u_{m,i} \frac{\Sigma_i y_{i,j} h_{j,m} / x_{i,j}}{\Sigma_i h_{j,m}} \tag{4}$$

本研究の NMF では，ランダムな非負値で初期化した行列 H，U に式 (4) の更新式を収束するまで繰り返し適用する [38]．更新の収束で，行列分解結果の基底行列 H，係数行列 U が得られる．

3.3 係数行列の分割

NMF は，$M < min(K, N)$ のとき，観測行列 Y が，低ランク行列の積で近似することに相当する．したがって，基底行列 H と係数行列 U の線形結合で表現できる．線形結合の例を式 (5) に示す．

$$h_{j,1} \begin{pmatrix} u_{1,i} \\ u_{2,i} \\ \vdots \\ u_{M,i} \end{pmatrix} + \cdots h_{j,M} \begin{pmatrix} u_{1,i} \\ u_{2,i} \\ \vdots \\ u_{M,i} \end{pmatrix} \tag{5}$$

式 (5) の係数行列 U の成分 $u_{M,i}$ は文書 i の基底 M への重みであり，スカラー[11]である．また，基底行列 H の成分 $h_{j,M}$ は，基底 M への特徴量 j の寄与率であり，ベクトル[12]である．

基底 M は，観測行列 Y の特徴量の共起成分がグルーピングされた結果である．NMF の行列分解では，係数行列 U の非負性の制約で，係数行列 U の要素が 0 の多いスパースになる傾向がある [38]．係数行列 U の要素がスパースであることは，成分である $u_{M,i}$ が文書 i のコンテンツによって，重みがある基底と重みが 0 の基底の差が明瞭であることに相当する．したがって，特性に基づく 2 種類の集合がある場合，集合の一方でのみ基底の重みが高い基底は，集合の特性を表す基底である．

[11] 温度のように大きさなど 1 つの数値だけで表せる量．
[12] 大きさと方向を持った量．速度・力・加速度など．平面上・空間上の有向線分 (方向を決めた線分) で表す．

基底の評価が行えた場合，寄与率を用いて目的変数に影響の大きい特徴量の評価が行える．観測行列 Y に NMF を適用した結果，質問記事の各基底の重み付きは係数行列 U で表される．そこで本研究では，目的変数に基づき，係数行列 U の行ベクトルにラベル付けする．そして，ラベル付けに基づき，係数行列 U を 2 種類の集合に分割する．本研究では，目的変数は正規分布にしたがうと仮定し，平均値以上を 1，平均値未満を 0 としてラベル付けを行う．そして，対応する係数行列 U の行ベクトルをベクトル集合 L_1，L_0 に分割する．

目的変数の集合 X を式 (6) に，ラベル付けの方法を式 (7) に，ラベル付けに基づく係数行列 U の分割の方法を式 (8) に示す．式 (7) は，x_i が平均値以上の場合に 1，平均値未満の場合に 0 に，ラベル付けされることを表している．また，x_i は文書 i の閲覧数，u_i は文書 i の係数行列 U の行ベクトルである．

$$X = \left[\begin{array}{cccc} x_1, x_2 & \ldots & x_N \end{array} \right]^T \tag{6}$$

$$x_i = \begin{cases} \frac{\Sigma_{j=1}^{N} x_j}{N} \leq x_i, x_i \in Set & 1 \\ \frac{\Sigma_{j=1}^{N} x_j}{N} > x_i, x_i \in Set & 0 \end{cases} \tag{7}$$

$$u_i = \begin{cases} x_i = 1, \quad u_i \in Set & L_1 \\ x_i = 0, \quad u_i \in Set & L_0 \end{cases} \tag{8}$$

3.4 基底選択

3.3 節で分割したベクトル集合 L_1 および L_0 を基に，集合における基底 m の係数の平均値を算出し，ベクトル集合における基底の重みを算出する．ベクトル集合はそれぞれ目的変数が多い集合，目標変数が少ない集合である．

ここで，それぞれの集合において，基底の重みが大きい場合は，重みが大きい重要な基底だが，特性を表す基底ではないと解釈できる．そこで本研究では，各集合の基底 m における係数の重みの差で，集合における基底 m の特性を評価する．集合における各基底の重みの算出式を式 (9) 示す．ラベル付けされたベクトル集合は L_j であり，j は 1 もしくは 0 である．$L_{j,m}$ は L_j における基底 m の係数の重みである．

$$L_{j,m} = \frac{\Sigma_{i=1}^{N} u_{m,i}}{N} \tag{9}$$

式 (9) における $u_{m,i}$ はベクトル集合 L_j にラベル付けされた行ベクトル u_i の基底 m への重みである．式 (9) の結果を基にした基底の特性を表す集合を S とし，本研究の基底 m の評価方法を式 (10) に示す．

$$S_m = L_{1,m} - L_{0,m} \tag{10}$$

上述の 3.2 節，3.3 節，3.4 節のとおり，提案手法は，テキスト情報から抽出した特徴量を NMF で行列分解 (3.2) を行い，分解結果を目的変数で分割 (3.3)，そして文書の特性を表す基底を選択 (3.4) する，基底の評価手法である．

本研究では，特徴量である観測ベクトルを並べた行列を観測データ行列と見なし，データをベクトルで表記している．NMF を観測行列 Y に適用した場合，基底行列 H と係数行列 U に行列分解される．NMF における行列分解の結果得られる基底行列の要素は，ベクトルであり，係数行列 U の要素はスカラーである．基底では，特徴量は基底の特性に基づいた，寄与率を持つ．提案手法で得られる基底は，2 種類の集合のうち，一方の集合で，係数値が大きい基底である．係数値はスカラーである．一方で，スカラーが大きい基底は，集合の特性を表す基底に相当する．したがって，提案手法は，スカラーで基底を評価し，ベクトルである特徴量の寄与率で，特徴量を評価する手法である．提案手法の有効性の評価は，非線形回帰手法を用いた閲覧数の予測と文書分類で行う．次章で，実装方法と閲覧数の予測と文書分類の評価方法について述べる．

4 実装

4.1 実験環境

本研究の実装は，プログラム言語 Python [13]を用いた．単語分割，品詞判定は，MeCab[14]を用いた．アルゴリズムの実装は，Gensim[15]，scikit-learn[16] [17] を用いた．NMF の観測行列 Y 作成は，Pandas[18] および Numpy[19]を用いた．特徴量の値は 0 から 1 の間に正規化し，各変数の計測尺度の違いを考慮するため，L1 正則化による制約補正を実施した．

[13]Welcome to Python.org : https://www.python.org
[14]MeCab : http://taku910.github.io/mecab/
[15]gensim : https://radimrehurek.com/gensim/
[16]scikit-learn : http://scikit-learn.org/stable/
[17]scikit-learn : https://github.com/scikit-learn
[18]Pandas : http://pandas.pydata.org/
[19]NumPy : http://www.numpy.org

提案手法の評価で用いる回帰分析では，学習用のデータ (75%) でモデルを作成し，評価用のデータ (25%) で評価を行った．また，回帰分析では，特徴量の値は，計測尺度の違いを考慮するため，正規化および L1 正則化を実施した．加えて，閲覧数および回帰分析の結果の予測値は，対数変換を実施した．5.2 節で後述するが，提案手法の評価に用いる MAE，RMSE の算出には，scikit-learn および Numpy を用いた．また，本研究で話題の特徴量抽出に用いる LSI の次元数や LDA のトピック数は，任意で値を設定する．LSI の K の最適値は 300 から 500 の範囲が提案されている [40]．本研究では，K=300 を設定した．LDA に関しても同様に K=300 を設定した．

提案手法で用いる NMF の観測行列 Y は，$\big(j = 1, \ldots, K\big)$ から構成される N 行 K 列の長方行列である．ここで，特徴量の次元数 K は，表層情報の次元数 22，アルゴリズムの次元数 600，辞書の次元数 1,449 の合計値 2,071 である．

テキスト情報の前処理では，文字コードを UTF-8 に変換，空白記号・改行コードを除去した．バイト列の欠損行は対象外とした．また，等価な文字は Normalization Form KC (NFKC) で正規化した．加えて，形態素解析では，活用形を標準形に変換し，1 文字単語などを削除する前処理を実施した．本研究では，2 種類のオンラインコミュニティのデータセットを用いて提案手法の評価を行う．

データセット 1 は Apple Inc.[20]が提供している Apple サポートコミュニティ [41] に 2008 年 10 月 1 日から 2014 年 1 月 24 日に投稿された質問記事 10,391 件を評価対象に用いる．したがって，データセット 1 の文書数 N は 10,391 である．このため，NMF を適用する観測行列 Y は，10,391 行 2,071 列の長方行列である．

データセット 2 は Stack Exchange, Inc.[21]が提供している 2018 年 5 月 5 日までに投稿された Stack Exchange Data Dump [42] のうち，stackover flow のデータ・セットである ja.stackoverflow.com の質問記事 35,945 件を評価対象に用いる．したがって，データセット 2 の文書数 N は 35,945 である．このため，NMF を適用する観測行列 Y は，35,945 行 2,071 列の長方行列である．

[20]Apple : https://www.apple.com
[21]Stack Exchange: Hot Questions : https://stackexchange.com/

提案手法の有効性はオンラインコミュニティの話題性で評価する．話題性を評価する場合，重要な要素は話題が取り上げられるか，ユーザの興味と合致しているかである．既存研究の話題性の評価基準には盛り上がりの早さや平均返信数 [17]，ユーザ行動やコミュニティの成長率 [18] などがある．

本研究では，評価対象が質問記事であることから，話題性を判断する要素にオンラインコミュニティの質問記事の閲覧数を用いる．

4.2 閲覧数予測における評価方法

本研究では，話題性の評価の一つを閲覧数の予測で行う．閲覧数の予測では，非線形回帰手法であるサポートベクタ回帰モデル (SVR : Support Vector Regression) [43] を用いる．SVR では，入力空間から，カーネル関数を用いて高次元の特徴空間へ写像する．そして，特徴空間で線形回帰を行う非線形回帰手法である．

SVR は，汎化能力が高い回帰モデルであることが知られている [44]．SVR の回帰関数を式 (11) に示す．

$$f(x) = \sum_{i=1}^{n} \big(\alpha_i - \alpha_i^*\big) K(x_i, x) + bias \quad (11)$$

$K(x_i, x)$ は入力 x_i を特徴空間へ写像するカーネル関数である．本研究では，カーネル関数に RBF カーネルを用いた．α_i,α_i^*,$bias$ などの詳細は文献 [45] などを参照していただきたい．SVR の評価には，予測値と実測値との予測誤差を評価指標に用いる．本研究では，MAE(Mean Absolute Error) および RMSE(Root Mean Squared Error) を用いる．MAE を式 (12) に RMSE を式 (13) に示す．n は予測対象のデータ数，y_i は実績値，$\widehat{y_i}$ は予測値である．

$$MAE = \frac{1}{n} \Sigma_{i=1}^{n} \big| y_i - \widehat{y_i} \big| \quad (12)$$

$$RMSE = \sqrt{\frac{1}{n} \Sigma_{i=1}^{n} \big(y_i - \widehat{y_i}\big)^2} \quad (13)$$

本研究では，提案手法で得られた基底の特徴量を SVR の入力 x_i に用いる．提案手法の結果，得られた基底に対する特徴量を寄与率の順に SVR の入力 x_i とし，予測誤差を算出する．

ここで，目的変数の予測誤差が小さい結果である場合，基底選択が有効であり，基底に基づいた寄与率で，少数の特徴量を特徴量選択する提案手法が有効であることに相当する．したがって，本研究のSVR では実績値に話題性の評価基準とした質問記事の閲覧数，予測値に回帰分析を行った結果を用いて，予測誤差を算出し，話題性の評価を行う．

4.3 文書分類における評価方法

本研究では，閲覧数の予測に加えて，文書分類で話題性を評価する．文書分類では，分類の基準となるラベルは，質問記事の閲覧数の平均値を基に話題性がある質問記事と話題性がない質問記事をラベリングした．提案手法で得られた基底の寄与率上位の特徴量を用いて文書分類を行う．

ここで，文書分類で話題性があるとラベリングした質問記事が分類できた場合，提案手法の基底選択および特徴量選択が有効であると言える．

本研究における文書分類の分類器は，AdaBoost，RandomForest，MLP，K-NN を用いる．各文書分類の手法について簡単な説明を行う．

AdaBoost は，ブースティング方式の分類手法である．ブースティングは，精度が低い分類器である弱分類器 (Weak classifier) を複数組み合わせることで，高精度の分類器 (Strong classifier) を構築するアルゴリズムである．分類誤り率から，適応的 (adaptive) に仮説に対する重みと次のラウンドの訓練データに対する重みを決定することから AdaBoost(adaptive boosting) と呼ばれる [46].

Randam Forests は，複数の決定木 (decision tree) を用いた分類手法である [47]．決定木は，あるデータ集合とその属性で木構造を構成し，識別ルールを構築する手法である．

MLP(Multi-Layer Perceptron) は，入力信号を，出力信号に変換する (無記憶) 非線形の神経回路網モデル (ニューラルネットワークモデル) である [48]．また，MLP は多層パーセプトロンとも呼ばれ，入力層，出力層およびいくつかの中間層 (hidden 層) からなり，入力から出力の方向にいくつかの層間の結合がある．

K-NN 判別 (K-Nearest Neighbours 判別) は正しく分類が行われている既存のデータから，判別を行いたい個体に最も「近い」個体 (最近傍点) を k 個選び出し，それらの個体が最も多く属しているクラスに当該固体を分類する分類器である [49].

文書分類の評価実験は，評価指標は「適合率」「再現率」「F 値」を用いた．指標算出のための分割表を図 2 に，適合率を式 (14)，再現率を式 (15)，F-measure を式 (16) に示す．

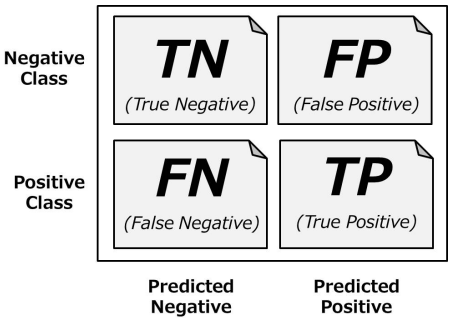

図 2: Contingency Table

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{16}$$

適合率は分類器の評価指標であり，不正解データを正解データと判定しないようにする指標である．再現率は，分類器がすべての正解データを判定する指標である．また，F-measure は適合率と再現率の調和平均値である．

提案手法の有効性評価では，特徴選択を行わず，すべての特徴量を用いて文書分類した場合と，既存の特徴量選択手法を用いて比較する．比較対象とする既存の特徴量選択手法には，分散分析 (ANOVA F-value) による単変量特徴量選択とモデルベースによる再帰的特徴量選択を用いて比較する．本研究では，モデルベースの再帰的特徴量選択では，Randam Forests を用いた．

次章で，話題性の評価基準に用いた評価対象の閲覧数の分布と評価対象に提案手法を適用し，評価した結果を示す．

5　評価実験

5.1　オンラインコミュニティの話題性

本研究では，オンラインコミュニティの話題性を評価する．話題性を評価する場合，重要な要素は話題が取り上げられるか，ユーザの興味と合致しているかである．したがって，本研究では，閲覧数をオンラインコミュニティの話題性を評価する指標に用いている．評価実験では，2 種類のオンラインコミュニティを評価する．オンラインコミュニティ1 は Apple サポートコミュニティ，オンラインコミュニティ2 は stackover flow のコミュニティである．オンラインコミュニティ1 における閲覧数に対する返信数の散布図を図 3 に，オンラインコミュニティ2 における閲覧数に対する返信数の散布図を図 4 に示す．

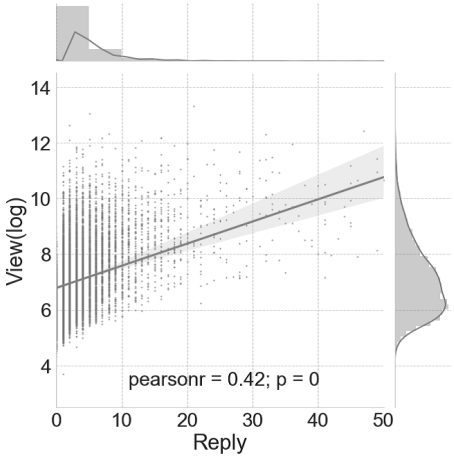

図 3: 閲覧数と返信数 (1)

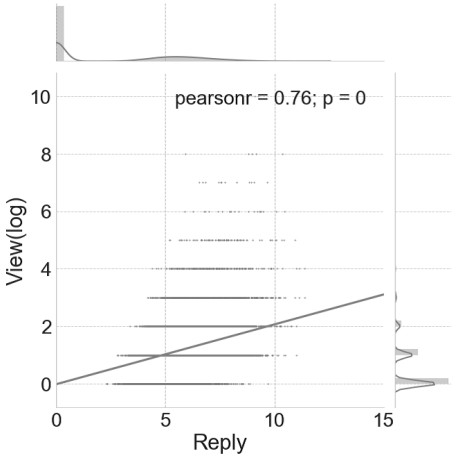

図 4: 閲覧数と返信数 (2)

図 3 では，質問記事の閲覧数は正規分布の傾向が確認できた．返信数は返信がない質問記事も多い一方で，質問記事の 1 件に対し，40 件から 50 件程の返信がある場合もある．したがって，オンラインコミュニティ1 は閲覧に対し，返信のバリエーションが多いコミュニティであると言える．閲覧数に対する返信数の相関係数は，0.42 を示した．

図 4 では，閲覧数が少ない質問記事が多い一方で，閲覧された質問記事に対しては，返信される傾向がある．したがって，オンラインコミュニティ2 は閲覧に対し，均質な返信が行われるコミュニティであると言える．閲覧数に対する返信数の相関係数は，0.76 という高い相関が明らかになった．

結果，2 種類のオンラインコミュニティで閲覧数と返信数には相関があり，閲覧数はコミュニティのコミュニケーションを増加させる要素であることが明らかになった．このため，話題性の多寡を評価する基準に閲覧数を用いることは妥当であると言える．

5.2　オンラインコミュニティ1 への適用

オンラインコミュニティ1 に対して，NMF の基底数を $M = 50$ に設定し，提案手法の適用を行った．提案手法による基底の評価結果を図 5 に示す．

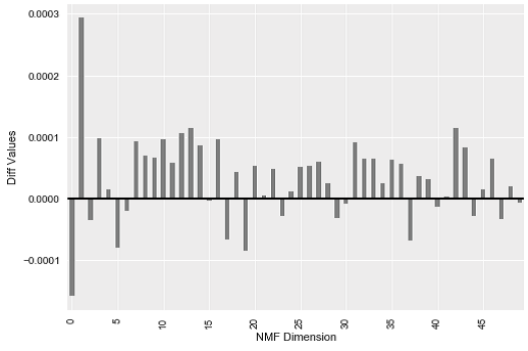

図 5: 閲覧数に寄与率が高い基底

図 5 の横軸が基底 m の特性を表す集合 S_m の要素である．また，縦軸が集合間における係数の平均値の差である．縦軸の値が非負値の基底が，閲覧数が平均値以上のベクトル集合である L_1 の特性を表す基底である．結果，基底 1 が非負値で最も差異が大きいと評価された．したがって，基底 1 が閲覧数が平均値以上の集合 L_1 の特性を表す基底である．

得られた結果の基底 1 を基底選択して，SVR を用いて回帰分析を行い，閲覧数の予測を行い，予測誤差である MAE を算出した結果を図 7 および図 6 に示す．

また同様に，RMSE を算出した結果を図 9 および図 8 に示す．図の Input Feature は，回帰分析で用いる際の特徴量の数である．回帰分析の特徴量は，基底 1 に対する寄与率の順に用いている．

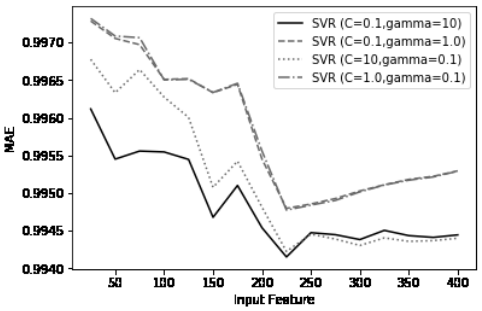

図 6: 回帰分析の結果を用いた MAE(1)

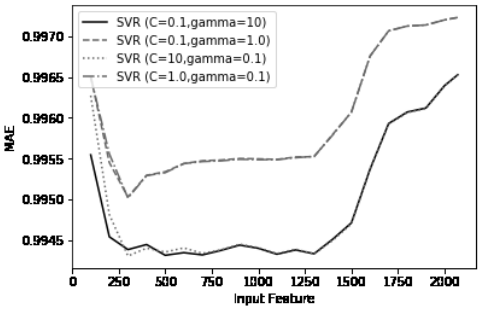

図 7: 回帰分析の結果を用いた MAE(2)

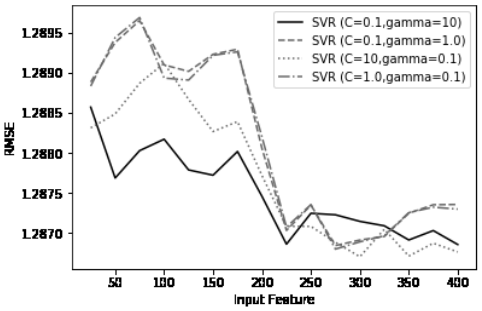

図 8: 特徴選択数と回帰分析の RMSE(1)

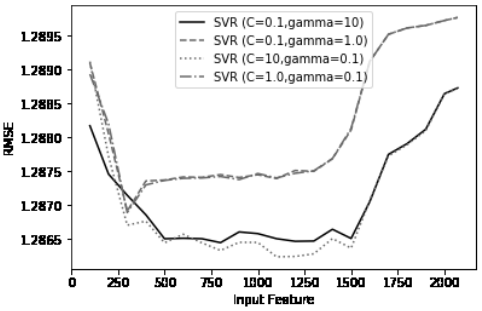

図 9: 特徴選択数と回帰分析の RMSE(2)

MAE の場合，入力特徴選択数が 250 個前後が最も精度が良い (図 6)．一方で，1250-1500 個以上の入力特徴量がある場合，精度が下がる (図 7)．結果，すべての特徴量を用いた場合よりも，提案手法で得られた基底 1 に寄与率の高い少数の特徴量を特徴選択した場合の方が，精度が高い．RMSE の場合，入力特徴選択数が 200-250 個が最も精度が良い (図 8)．一方で，1500 個以上の入力特徴量がある場合，精度が下がる (図 9)．結果，MAE と同様に，すべての特徴量を用いた場合よりも，提案手法で得られた基底 1 に寄与率の高い少数の特徴量を特徴選択した場合の方が，精度が高い．したがって，MAE および RMSE を用いた，非線形回帰において，基底選択に基づ特徴量の選択は妥当であることがデータで示された．最後に，基底 1 に対する特徴量の寄与率上位の 10 個を表 4 に示す．

表 4: 基底 1 の寄与率上位 10 個の特徴量

	特徴量名		特徴量名
1	固有 (一般)	6	機能 (I-目的)
2	機能 (I-順接仮定)	7	ひらがな (頻度)
3	品詞 (助詞)	8	機能 (I-不可能)
4	意味 (2)(4.112)	9	機能 (I-依頼)
5	機能 (I-様態)	10	機能 (B-自然発生)

表 4 の「固有 (一般) 」は「固有名詞」,「機能 (*)」は「機能表現タグ」,「意味」は「意味分類コード」である．意味分類コード (2)(4.112) の単語は,「その他の類 (展開)」の「ですから」「まして」「だから」などである．結果，選択基底は，固有名詞やひらがなの頻度の寄与が高く，機能表現タグの寄与が多い，基底であることが明らかになった．

5.3　オンラインコミュニティ1の文書分類

オンラインコミュニティ1より得られた特徴量を用いて文書分類実験を行った結果を表5から表8に示す.

表 5: 全特徴量

分類器	適合率	再現率	F 値
AdaBoost	0.45	0.29	0.35
RandomForest	0.44	0.26	0.33
MLP	0.00	0.00	0.00
K-NN	0.42	0.37	0.39

表 6: 特徴量選択 (提案手法 基底の寄与率)

分類器	適合率	再現率	F 値
AdaBoost	0.43	0.23	0.30
RandomForest	0.45	0.29	0.35
MLP	0.40	0.04	0.08
K-NN	0.44	0.42	0.43

表 7: 特徴量選択 (単変量特徴量選択)

分類器	適合率	再現率	F 値
AdaBoost	0.46	0.28	0.35
RandomForest	0.45	0.28	0.34
MLP	0.00	0.00	0.00
K-NN	0.42	0.38	0.40

表5は，得られたすべての特徴量を用いて，文書分類した結果である．表6が，提案手法で得られた基底1の寄与率上位100個の特徴量を特徴選択し，文書分類した結果である．表7は，既存手法である単変量特徴量選択を用いて特徴量選択した結果であり，表8は，既存手法である再帰的特徴量削減を用いて特徴量選択し，文書分類した結果である．すべての特徴量およびいずれの特徴量選択を用いた場合でも，オンラインコミュニティ1の質問記事では，特徴量選択手法に関わらず，文書分類の精度は十分ではない．

表 8: 特徴量選択 (再帰的特徴量削減)

分類器	適合率	再現率	F 値
AdaBoost	0.44	0.29	0.35
RandomForest	0.45	0.27	0.34
MLP	0.00	0.00	0.00
K-NN	0.43	0.38	0.40

結果から，提案手法ではなく，閲覧数の平均値を基に行った文書分類のラベリングの評価基準が妥当ではなかったと判断できる．

5.4　オンラインコミュニティ2への適用

評価では，NMF の基底数を $M = 50$ に設定し，提案手法の評価を行った．提案手法による基底選択の結果を図10に示す.

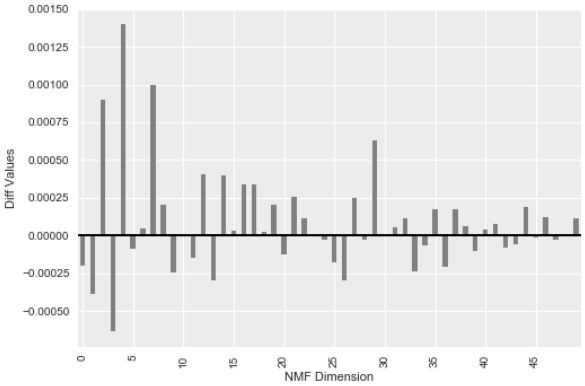

図 10: 閲覧数に寄与率が高い特徴量と状況変数

NMF の基底数を M = 50 に設定し，提案手法の評価を行った．横軸が基底 m の特性を表す集合 Sm の要素である．また，縦軸が集合間における係数の平均値の差である．縦軸の値が非負値の基底が，閲覧数が平均値以上のベクトル集合である L1 の特性を表す基底である．追加実験では，基底 4 が非負値で最も差異が大きいと評価された．

基底 4 を用いて，SVR を用いて回帰分析を行い，閲覧数の予測を行い，予測誤差である MAE を算出した結果を図 12 および図 11 に示す．また同様に，RMSE を算出した結果を図 14 および図 13 に示す.

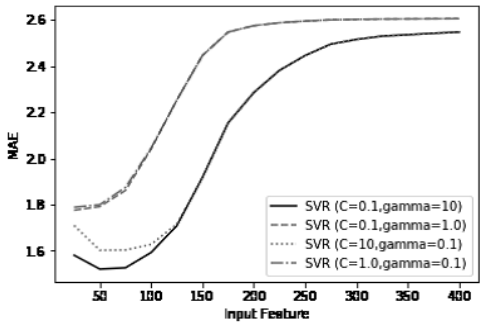

図 11: 回帰分析の結果を用いた MAE(1)

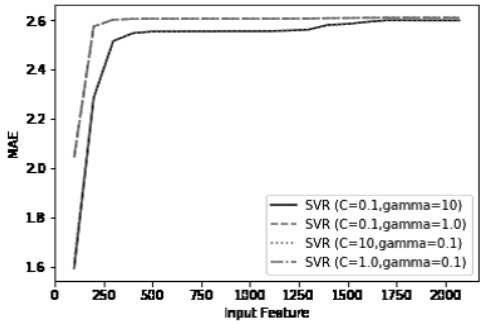

図 12: 回帰分析の結果を用いた MAE(2)

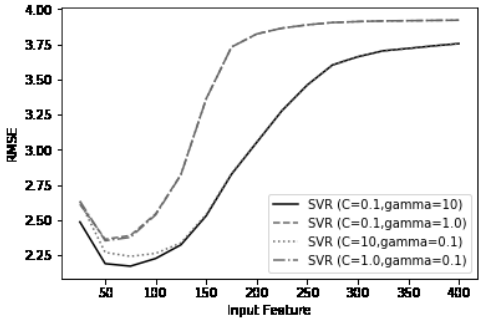

図 13: 特徴選択数と回帰分析の RMSE(1)

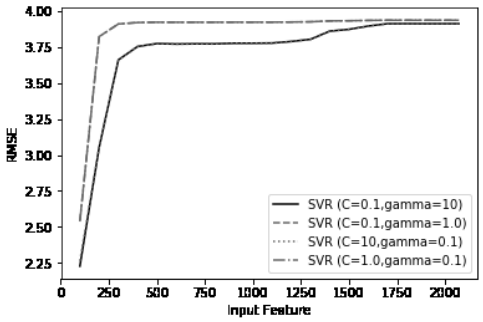

図 14: 特徴選択数と回帰分析の RMSE(2)

図の Input Feature は，回帰分析で用いる際の特徴量の数である．また，回帰分析の特徴量は，基底 4 に対する寄与率の順に用いている．MAE の場合，入力特徴選択数が 50-100 個前後が最も精度が良い (図 11)．一方で，100 個以上の入力特徴量がある場合，精度が下がる (図 12)．入力特徴選択数が 50 個である場合と 250 個以上である場合は，予測精度の差は明らかである．結果，すべての特徴量を用いた場合よりも，提案手法で得られた基底 4 に寄与率の高い少数の特徴量を特徴選択した場合の方が，精度が高い．

RMSE の場合，入力特徴選択数が 50-100 個が最も精度が良い (図 13)．入力特徴選択数が 50 個である場合と 250 個以上である場合は，予測精度の差は明らかである (図 14)．結果，すべての特徴量を用いた場合よりも，提案手法で得られた基底 4 に寄与率の高い少数の特徴量を特徴選択した場合の方が，回帰分析の予測精度が高い．したがって，MAE および RMSE を用いた，非線形回帰において，基底選択に基づく寄与率による特徴量の選択は妥当であり，有効であることがデータで示された．

最後に，基底 4 に対する特徴量の寄与率上位の 10 個を表 9 に示す．

表 9: 基底 4 の寄与率上位 10 個の特徴量

	特徴量名		辞書の単語
1	TOPIC(LDA)	6	意味 (2)(1.353)
2	意味 (2)(2.304)	7	意味 (2)(4.35)
3	意味 (2)(1.304)	8	意味 (2)(1.564)
4	意味 (2)(1.366)	9	意味 (2)(3.133)
5	意味 (2)(4.314)	10	TOPIC(LDA)

表 9 の「意味」は「意味分類コード」である．結果，基底 4 は，意味分類コードの寄与が高い基底である．意味分類コードは語を意味に基づいて分類した語彙表による特徴量である．語彙は，ある言語体系，地域，分野などで用いられる語の総体であり，意味とは，言葉や記号などで表現され，理解される一定の内容である．ゆえに，理解されやすい特徴量の寄与が高い基底であることが明らかになった．

5.5 オンラインコミュニティ2の文書分類

オンラインコミュニティ2より得られた特徴量を用いて文書分類実験を行った結果を表10から表13に示す．

表 10: 全特徴量

分類器	適合率	再現率	F 値
AdaBoost	0.89	0.87	0.88
RandomForest	0.87	0.78	0.82
MLP	0.92	0.88	0.90
K-NN	0.83	0.75	0.78

表 11: 特徴量選択 (提案手法 基底の寄与率)

分類器	適合率	再現率	F 値
AdaBoost	0.87	0.83	0.85
RandomForest	0.87	0.82	0.84
MLP	0.87	0.87	0.87
K-NN	0.85	0.80	0.83

表 12: 特徴量選択 (単変量特徴量選択)

分類器	適合率	再現率	F 値
AdaBoost	0.88	0.85	0.87
RandomForest	0.88	0.83	0.85
MLP	0.00	0.00	0.00
K-NN	0.87	0.83	0.85

表10は，得られたすべての特徴量を用いて，文書分類した結果である．表11が，提案手法で得られた基底1の寄与率上位100個の特徴量を特徴選択し，文書分類した結果である．表12は，既存手法である単変量特徴量選択を用いて特徴量選択した結果であり，表13は，既存手法である再帰的特徴量削減を用いて特徴量選択し，文書分類した結果である．結果，オンラインコミュニティ2では，正解クラスへの分類結果は0.8以上の高い分類精度が得られた．

表 13: 特徴量選択 (再帰的特徴量削減)

分類器	適合率	再現率	F 値
AdaBoost	0.89	0.86	0.87
RandomForest	0.88	0.84	0.86
MLP	0.00	0.00	0.00
K-NN	0.88	0.82	0.85

したがって，提案手法で得られた特徴量は，文書分類に有効であることが明らかである．加えて，オンラインコミュニティ2においては，閲覧数の平均値を基に行った文書分類のラベリングの評価基準が妥当であることが明らかになった．

6 考察

6.1 話題性の予測

評価に用いたオンラインコミュニティの考察から，始める．まず，オンラインコミュニティ1は，Appleサポートコミュニティのデータセットである．Appleサポートコミュニティは，コンシューマ[22]が，直接質問記事を投稿する．また，利用者層の幅は非常に広い．したがって，メディアとして考察した場合は，ゼネラル・メディア[23] に相当する性質を持つ．

一方で，オンラインコミュニティ2は，stackoverflowのデータセットである．stack overflowはStack Exchangeの Q&A コミュニティの一つであり開発者向けのコミュニティである．stack overflowは専門家や熟練者など，高度な知識を持つ，エキスパートが利用者である．このため，メディアとして考察した場合は，クラス・メディア[24]に近い性質を持つ．

5.2節の結果から，各オンラインコミュニティで，閲覧数と経過日数の相関係数が大きく異なることが明らかになった．また，閲覧数と返信数の相関係数においても，オンラインコミュニティ1で0.42，オンラインコミュニティ2で0.76である．

[22]商品やサービスの最終的な利用者，消費者のこと．消費者という概念はある商品やサービスを直接利用する人であるが，購買行為を決定する人も含めて消費者と呼ぶこともある [50].

[23]年齢，教育程度，職業，ライフスタイルなど，あらゆる社会層の人々を普遍的にオーディエンスとしている媒体．マス・メディアをオーディエンスの態様から分類した概念．一般日刊紙やテレビなど [50].

[24]特定の社会層もしくは集団を対象にした媒体．雑誌やラジオなど [50]．また，業界誌や専門雑誌，ダイレクト・メールなど限定された対象を相手にする媒体も含まれる．

ゆえに，質問記事を閲覧した際の返信度合いは，オンラインコミュニティのメディアの性質で異なると推定される．

ここで，オンラインコミュニティの一つである不満買取センター[25] の不満カテゴリ辞書データを用いて，オンラインコミュニティのコンテンツを考察する．不満カテゴリ辞書は，2015 年 3 月 18 日から，2016 年 12 月 1 日までのノイズを排除した投稿記事 3,527,336 件から作成されている [51]．辞書のエントリ数は 953,776 件であり，複数カテゴリに登録されている重複エントリを除いたエントリ数は 110,866 件である．不満買取センターに不満投稿が投稿されるカテゴリと，特徴的な単語の例を表 14 に示す．単語は，不満カテゴリ辞書の TF-IDF [26] におけるスコア上位エントリの名詞である．

表 14: 不満買取センターのカテゴリと単語の例 [51]

カテゴリ名	単語の例
暮らし・住まい	布団，雨，枕
ファッション	腕時計，靴
趣味・エンタメ	映画，CD
食品・飲料	グミ，ワイン
外食・店舗	弁当，居酒屋
医療・福祉	整体，介護，薬
アウトドア・スポーツ	バイク，自転車
デジタル・家電	録画，デジカメ
宿泊・観光・レジャー	ホテル，部屋
公共・環境	バス，飛行機
教育	幼稚園，保育園
国際・文化	留学，日本
政治・行政	選挙，政治家
人間関係	離婚，結婚，人
仕事	転職，仕事，面接
ペット	ペットショップ

ところで，不満カテゴリ辞書データの特徴的な単語は，登録カテゴリで異なる．また辞書では，飲食物の商品名など，商品の固有名詞も登録されており，商品名は食品・飲料や，宿泊・観光・レジャーなど，関連性のある複数カテゴリで登録されている．

一方で，他のカテゴリで登録されている特徴的な単語であっても，関連性がないカテゴリでは登録されていない．

結果から，不満買取センターにおける不満投稿では，関連性がないカテゴリでは単語の登録がないことが明らかになった．カテゴリにおいて，特徴的な単語が異なることは，各カテゴリが，異なるメディアあるいは，コミュニティであると言える．ゆえに，コンテンツが広範なゼネラル・メディアのオンラインコミュニティの場合，オンラインコミュニティの各カテゴリが，クラス･メディアに相当すると推定される．したがって，同一のオンラインコミュニティであっても，メディアの性質で異なると推定される．評価実験の話題性の予測では，オンラインコミュニティ1 およびオンラインコミュニティ2 で，特徴量選択は予測誤差の精度は有効な結果であった．しかし，オンラインコミュニティ2 の予測精度の向上は明らかである一方で，オンラインコミュニティ1 の予測精度の向上は限定的であった．5.2 節の結果から，オンラインコミュニティ1 は，ゼネラル・メディアに相当すると推定される．ゼネラル・メディアの性質を持つオンラインコミュニティの場合，コンテンツの広範であり，カテゴリが多い．したがって，表 14 のように，カテゴリが異なった場合，特徴的な単語は異なると推定される．

本研究で用いたトピックモデルアルゴリズムは，カテゴリ分類においても有効なアルゴリズムである．提案手法の適用の結果，オンラインコミュニティ1 では，閲覧数予測においては，有効性は限定的であった．ゼネラル・メディアに近い性質を持つオンラインコミュニティ1 は，利用者が広範である．したがって，閲覧数が増加しやすい傾向や特徴的な単語が不明瞭であることなどが推定される．このため，オンラインコミュニティ1 の閲覧数予測の精度向上には，カテゴリ分類後に，閲覧数予測を行う必要があると推定される．

一方で，オンラインコミュニティ2 は 5.2 節の結果から，クラス・メディアの性質を持つと推定される．オンラインコミュニティ2 はオンラインコミュニティ1 と比較して，利用者が限られた範囲である．提案手法の適用の結果，オンラインコミュニティ2 では，閲覧数予測において，高い有効性が明らかになった．これは，クラス・メディアの性質を持つオンラインコミュニティ2 は，利用者が限られており，閲覧数が増加しやすい一定の傾向が存在することや，

[25] 不満買取センター：http://fumankaitori.com

[26] 単語の重み付け技法の一つ．文書内単語の相対的な重要性を表す正規化されたスコア．単語の頻度と文書頻度の逆数で算出．

あるいは閲覧数が増加する特徴的な単語が明瞭であることなどが要因であると推定される．加えて，コンテンツも限られた範囲であり，カテゴリが異なった場合でも，特徴的な単語は類似の傾向があると推定される．

ゆえに，SVR で回帰分析を行った際に，クラス・メディアであるオンラインコミュニティ2 は明瞭な結果となり，ゼネラル・メディアであるオンラインコミュニティ1 では有効ではあるが限定的な結果であったと推定される．

6.2 話題性の判別

分類実験では，ゼネラル・メディアに相当するオンラインコミュニティ1 とクラス・メディアに相当するオンラインコミュニティ2 で結果が大きく異なった．まず，既存手法で抽出した特徴量を用いて，文書分類を行った結果を考察する．

本研究の文書分類では，質問記事の分類の基準となるラベルは，閲覧数の平均値を基にラベリングした．オンラインコミュニティ1 の文書分類では，すべての特徴量を用いた場合や，既存研究の特徴量選択手法を用いた場合など，いずれの手法を用いた場合でも，正解クラスへの分類結果は F 値で，0.3 から 0.4 程度である．パラメータの最適化を行わない場合，MLP の分類器では分類困難である．

次に，オンラインコミュニティ1 の層化 5 分割交差検証の交差検証スコアを表 15 に示す．表 15 の RF は，RandamForests である．

表 15: 交差検証 (オンラインコミュニティ1)

分類器	1	2	3	4	5
AdaBoost	0.54	0.53	0.52	0.54	0.53
RF	0.52	0.54	0.53	0.52	0.52
MLP	0.55	0.55	0.56	0.54	0.55
K-NN	0.50	0.52	0.52	0.51	0.51

表 15 より，交差検証スコアの場合でも，結果は十分ではない．したがって，オンラインコミュニティ1 では，基底選択に関わらず，十分な文書分類が行えていないと言える．このため，オンラインコミュニティ1 では，特徴量選択ではなく，正解クラスのラベリング基準を最適値にする必要がある．

一方で，オンラインコミュニティ2 では，既存研究の特徴量選択手法を用いた場合など，いずれの手法を用いた場合でも，正解クラスへの分類結果は F 値で，0.7 から 0.9 の範囲である．

次に，オンラインコミュニティ1 の層化 5 分割交差検証の交差検証スコアを表 16 に示す．表 16 の RF は，RandamForests である．

表 16: 交差検証 (オンラインコミュニティ2)

分類器	1	2	3	4	5
AdaBoost	0.86	0.86	0.87	0.87	0.87
RF	0.86	0.85	0.87	0.86	0.87
MLP	0.89	0.88	0.89	0.89	0.90
K-NN	0.84	0.84	0.85	0.85	0.85

表 15 より，交差検証スコアの場合でも，有効な結果であった．したがって，オンラインコミュニティ2 では，閲覧数のラベリング基準は平均値が妥当であったことが明らかになった．

ここで，提案手法は，目的変数である閲覧数から，有効な基底を明らかにし，基底選択を行う手法である．文書分類で有効な結果が得られたオンラインコミュニティ2 の結果を用いて，提案手法による基底選択の有効性を考察する．

基底選択の結果，基底の寄与率上位の特徴量で，重要な特徴量が明らかになり，特徴量の評価が行える．したがって，重要な特徴量のみを用いる特徴量選択が行える．分類結果で考察した場合，既存手法である単変量特徴量選択および再帰的特徴量削減では，分類困難であった MLP による分類が行えた．したがって，分類器に依存せずに，少ない特徴量で，すべての特徴量を用いた場合に相当する結果が得られた．このため，分類器に依存せず有効な特徴量選択が行える手法である．

分類精度だが，提案手法を用いたと比較し，すべての特徴量を用いた場合や，既存手法の特徴量選択を行った場合の方が，適合率や再現率，F 値の結果は良い．しかし一方で，値の差は誤差の範囲であり，すべての特徴量を用いた場合と同程度の分類結果であり，特徴量選択手法としての性能を有している．加えて，既存の特徴量選択手法と，同程度の結果も得られた．

特徴量選択としては，基底に対する特徴量の寄与率で，統計的な関係を基にした単変量特徴量選択と同様に，関係がないノイズに相当する特徴量の除去が行える．また，計算量基準では，再帰的特徴量削減は，特徴量の次元数を指定する場合，特定の次元数になるまで，繰り返しモデルベースを適用する．したがって，非常に大きな計算量を必要とする．一方で，提案手法は基底を選択した場合に，基底のベクトルである寄与率で特徴量を評価している．したがって，計算量は，再帰的特徴量削減と比較して少ない．このため，計算量基準では，既存手法である再帰的特徴量削減と比較し，有効な特徴量選択である．ゆえに，文書分類においては，提案手法は分類器に依存せず，再帰的特徴量削減と比較して，少ない計算量で特徴量選択が行える手法であると言える．

文書分類で有効な結果が得られたオンラインコミュニティ2 で，提案手法で得られた基底の特徴量を用いた結果の Precision-recall カーブ，受信者動作特性 (ROC)，AUC の結果を図 15，図 16，図 17 に示す．AUC は，ROC のカーブ下の領域である．

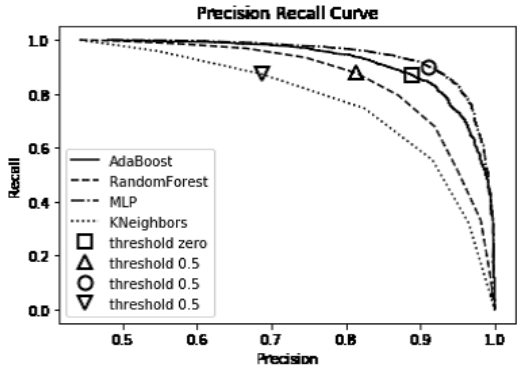

図 15: Precision-recall カーブ

表 17: AUC

Classification methods	AUC
AdaBoost	0.9608
RandomForest	0.9313
MLP	0.9725
K-NN	0.8747

結果，分類スレッショルドの最適化を行うことで，

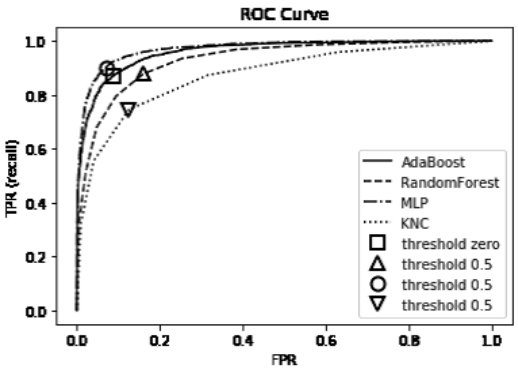

図 16: 受信者動作特性 (ROC)

適合率や再現率，F 値などの最適化が行える特徴量選択であることが明らかになった．したがって，提案手法による基底選択は，文書分類に有効な特徴量を評価する手法としての有効性があると言える．また，結果から，AUC では，MLP が最も良いスコアの分類器であることが明らかになった．

7　まとめ

本研究では，オンラインコミュニティの質問記事を対象に既存研究の表層情報，語種や品詞，文末表現などの 2,000 次元を越える特徴量に NMF を適用し，閲覧数を正解データとして基底を評価した．そして，NMF の結果である基底を係数値の差に基づいて評価し，話題性に影響の大きい基底を評価する方法を提案した．

提案手法で得られた基底に寄与率の高い特徴量を用いて，話題性の予測と文書分類で評価した．話題性の予測では，SVR で閲覧数に対する予測誤差を算出し，すべての特徴量を用いた場合比較し，MAE および RMSE の評価指標を評価基準とした．文書分類では，AdaBoost，RandomForest，MLP，K-NN の 4 種類の分類器を用いて，投稿されたコンテンツのテキスト情報から，閲覧数が多い話題性のある質問記事と，閲覧数が少ない質問記事の分類を評価基準とした．評価実験では，ゼネラル・メディアに相当する Apple サポートコミュニティと，クラス・メディアに相当する Stack Exchange の Q&A コミュニティの一つである stack overflow の 2 種類の特性の異なるオンラインコミュニティを用いて提案手法を評価した．

非線形回帰を用いた評価では，話題性の予測において，Apple サポートコミュニティでは，限定的な結果であった．一方で，stack overflow では，明瞭な有効性が得られた．分類器を用いた評価では，文書分類において，Apple サポートコミュニティでは，平均値では分類困難な結果であった．一方で，stack overflow では，F 値で 0.8 以上の，明瞭な有効性が得られた．また，すべての特徴量を用いた場合においても有効な結果が得られた．

考察では，オンラインコミュニティの考察を行い，不満買取センターに投稿される不満投稿に基づいて，コンテンツが広範なゼネラル・メディアのオンラインコミュニティの場合，オンラインコミュニティの各カテゴリが，クラス・メディアに相当することが明らかになった．そして，メディアの性質で特徴的な単語が，異なること結果であることが推定できた．ゆえに，評価実験において，ゼネラル・メディアとクラス・メディアで，結果に差が出たことは，妥当であることが明らかとなった．加えて，明瞭な結果が得られた stack overflow において，Precision-recall カーブや，ROC カーブの結果から，分類器のパラメータや分類スレッショルドを最適化することで，既存手法よりも良い分類結果を得られることが明らかになった．

今後の課題は，分類基準となる質問記事をラベリングする基準，分類器のパラメータや分類スレッショルドを最適化である．さらに，ベイジアンネットワークによる状況変数との因果関係の推定などでも基底選択の有効性を評価したい．

8 謝辞

本研究に連関した研究協力および研究助成頂いた皆様に御礼申し上げます．本研究における評価結果の考察に際しては，株式会社 Insight Tech が国立情報学研究所の協力により，研究目的で提供している「不満調査データセット」から，不満カテゴリ辞書データを利用した．ここに記してデータ提供頂いた株式会社 Insight Tech に感謝申し上げます．

参考文献

[1] 佐藤尚之, 金田育子, 京井良彦, 信澤宏至, 茂呂譲治, 橋口幸生, 宮林隆吉. SIPS ～来るべきソーシャルメディア時代の新しい生活者消費行動モデル概念～. 電通モダン・コミュニケーション・ラボ.

[2] 奥村学. ソーシャルメディアを対象としたテキストマイニング. 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review, Vol. 6, No. 4, pp. 285–293, 2013.

[3] 横山友也, 宝珍輝尚, 野宮浩揮, 佐藤哲司. 文章の特徴量を用いた質問回答文の印象の因子得点の推定. 日本感性工学会論文誌, Vol. 12, No. 1, pp. 15–24, 2013.

[4] 呉鍾勲, 鳥澤健太郎, 橋本力, 川田拓也, デサーガステイン, 風間淳一, 王軼謳. 意味的極性と単語クラスを用いた why 型質問応答の改善. 情報処理学会論文誌, Vol. 54, No. 7, pp. 1951–1966, jul 2013.

[5] Dongli Han, Yuhei Kato, Kazuaki Takehara, Tetsuya Yamamoto, Kazunori Sugimura, and Minoru Harada. Qa system metis based on web searching and semantic graph matching. In Zhongzhi Shi, K. Shimohara, and D. Feng, editors, *Intelligent Information Processing III*, pp. 123–133, Boston, MA, 2007. Springer US.

[6] 難波英嗣. 動向情報の抽出と要約ー動向をまとめるー. 知能と情報, Vol. 22, No. 5, pp. 549–555, 2010.

[7] 古川忠延, 松尾豊, 大向一輝, 内山幸樹, 石塚満. ブログ上での話題伝播に注目した重要語判別. 知能と情報, Vol. 21, No. 4, pp. 557–566, aug 2009.

[8] 山家雄介, 中村聡史, アダムヤトフト, 田中克己. ソーシャルブックマーキングの周期性発見と時期連動型検索ランキングへの適用. 情報処理学会論文誌データベース（TOD）, Vol. 2, No. 3, pp. 130–140, sep 2009.

[9] 山家雄介, 中村聡史, アダムヤトフト, 田中克己. ソーシャルブックマークの特性分析とそれに基づく web 検索結果の再ランキング手法. 情報処理学会論文誌データベース（TOD）, Vol. 1, No. 1, pp. 88–100, jun 2008.

[10] 吉川友也, 岩田具治, 澤田宏. ユーザの潜在特徴を考慮したソーシャルネットワーク上の情報拡散モデル. 情報処理学会論文誌データベース（TOD）, Vol. 6, No. 5, pp. 85–94, dec 2013.

[11] 濱岡豊. バズ・マーケティングの展開. *Ad Studies*, No. 20, pp. 5–10, 2007.

[12] Mike J. Walker. Hype cycle for emerging technologies, 2017. July 2017.

[13] 吉田光男, 荒瀬由紀. トレンドキーワードに関するウェブリソースの横断的分析. 情報処理学会論文誌データベース（TOD）, Vol. 9, No. 1, pp. 20–30, mar 2016.

[14] 中島伸介, 張建偉, 稲垣陽一, 中本レン. 大規模なブログ記事時系列分析に基づく流行語候補の早期発見手法. 情報処理学会論文誌データベース（TOD）, Vol. 6, No. 1, pp. 1–15, jan 2013.

[15] 岩田具治. トピックモデル (機械学習プロフェッショナルシリーズ). 講談社, 2015.

[16] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pp. 91–101, New York, NY, USA, 2002. ACM.

[17] 松村真宏, 三浦麻子, 柴内康文, 大澤幸生, 石塚満. 2 ちゃんねるが盛り上がるダイナミズム. 情報処理学会論文誌, Vol. 45, No. 3, pp. 1053–1061, mar 2004.

[18] 鳥海不二夫, 山本仁志, 諏訪博彦, 岡田勇, 和泉潔, 橋本康弘. 大量ＳＮＳサイトの比較分析. 人工知能学会論文誌, Vol. 25, No. 1, pp. 78–89, 2010.

[19] 福田誠, 吉田武尚, 吉田誠, 樫岡健史. 中学校技術・家庭科教科書電気領域の表記・表現について. 日本教科教育学会誌, Vol. 23, No. 3, pp. 37–42, 2000.

[20] 大里彩乃. 畳語の研究. 言語文化研究, No. 22, pp. 1–16, mar 2014.

[21] 高橋徹. 社会システム分化とゼマンティク : ルーマンにおける社会変動論の一視角. 社会学評論, Vol. 49, No. 4, pp. 620–634, mar 1999.

[22] 谷口永里子, 高橋真理子. 最新の女性ファッション雑誌における日本語の特徴の量的分析-年代差に焦点をあてて- . 言語処理学会第 22 回年次大会講演論文集, Vol. D5-1, , 2016.

[23] 佐藤政光. 日本語学習者の語彙習得に関する調査研究 (1) 基本語彙の問題点について. 明治大学人文科学研究所紀要, No. 44, pp. 169–180, feb 1999.

[24] 国立国語研究所. 日本語教育のための基本語彙調査, 1984.

[25] 松吉俊, 江口萌, 佐尾ちとせ, 村上浩司, 乾健太郎, 松本裕治. テキスト情報分析のための判断情報アノテーション. 電子情報通信学会論文誌. D, 情報・システム, Vol. 93, No. 6, pp. 705–713, jun 2010.

[26] 福田一雄. 日本語モダリティ覚え書き (その 1). 言語の普遍性と個別性, No. 5, pp. 1–13, mar 2014.

[27] 西原陽子, 松村真宏, 谷内田正彦. Q&a コミュニティでの質疑応答パターンの理解. 人工知能学会全国大会論文集, Vol. 22, pp. 1–4, 2008.

[28] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *In Proc. of EMNLP*, pp. 230–237, 2004.

[29] 中尾桂子. 品詞構成率に基づくテキスト分析の可能性 : メール自己紹介文、小説、作文、名大コーパスの比較から. 大妻女子大学紀要. 文系, Vol. 42, pp. 128–101, mar 2010.

[30] 関根聡, 竹内康介. 拡張固有表現オントロジー. 言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」, pp. 23–26, 2007.

[31] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203–222, 2005.

[32] 東山昌彦. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会論文集, 2008, pp. 584–587, 2008.

[33] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pp. 133–140, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[34] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–241, jul 2006.

[35] 小林のぞみ, 乾健太郎, 松本裕治. 意見情報の抽出／構造化のタスク仕様に関する考察. 情報処理学会研究報告自然言語処理（NL）, Vol. 2006, No. 1, pp. 111–118, jan 2006.

[36] 桑原（中島）尚子. 情報定義に内在する静的視座と動的視座. 日本社会情報学会全国大会研究発表論文集, Vol. 22, pp. 192–195, 2007.

[37] 伊藤守. 情報概念について：主知主義的な枠組みから解き放つために (特集社会情報学からの発信). 社会情報学, Vol. 1, No. 1, pp. 3–19, 2012.

[38] 亀岡弘和. 非負値行列因子分解. 計測と制御, Vol. 51, No. 9, pp. 835–844, sep 2012.

[39] 安川武彦. 非負値行列因子分解を用いたテキストデータ解析. 計算機統計学, Vol. 28, No. 1, pp. 41–55, 2015.

[40] Roger B. Bradford. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pp. 153–162, New York, NY, USA, 2008. ACM.

[41] Apple サポートコミュニティ. https://discussionsjapan.apple.com (最終閲覧日:2014-01-29).

[42] Stack Exchange Data Dump. https://archive.org/details/stackexchange (Publication date 2018-09-05).

[43] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, Vol. 14, No. 3, pp. 199–222, August 2004.

[44] 石川葉子, 水上雅博, 吉野幸一郎, Sakti Sakriani, 鈴木優, 中村哲. 感情表現を用いた説得対話システム. 人工知能学会論文誌, Vol. 33, No. 1, pp. DSH–B_1–9, 2018.

[45] 小林正幸, 小西康夫, 藤田貞雄, 石垣博行. サポートベクタ回帰モデルを用いた超音波モータの位置決め制御. 精密工学会誌論文集, Vol. 72, No. 5, pp. 596–601, may 2006.

[46] 永田昌明, 平博順. 情報論的学習理論とその応用：テキスト分類-学習理論の「見本市」-. 情報処理, Vol. 42, No. 1, pp. 32–37, jan 2001.

[47] 飯山将晃. 使える!統計検定・機械学習-iv : Random forests を用いたパターン認識. システム／制御／情報, Vol. 59, No. 2, pp. 71–76, 2015.

[48] 甘利俊一. 自然勾配学習法-学習空間の幾何学. 計測と制御, Vol. 40, No. 10, pp. 735–739, oct 2001.

[49] 浦田正夫. k-nearest neighbours 判別を用いたクラスター解析のバリデーション (種々のモデルの統計的解析). 数理解析研究所講究録, Vol. 1603, pp. 111–119, jun 2008.

[50] 亀井昭宏. 電通広告事典. 電通, 2008.

[51] Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, and Tomoya Mizumoto. Fkc corpus: a japanese corpus from new opinion survey service. *Proc. of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results*, pp. 11–18, 2016.