レビュー観点の推移パターンに基づく商品属性の抽出手法

村松　直哉†　　佐藤　哲司††　　伏見　卓恭††

† 筑波大学 情報学群 知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s1613130@u.tsukuba.ac.jp, ††{satoh,fushimi}@ce.slis.tsukuba.ac.jp

あらまし　本研究では，インターネット上で展開されている商品レビューサイトにおける，商品レビュートピックの変動を分析する．トピックモデルを用いた，商品レビューに関する手法の多くが，主にレビュー文書の特徴に注目したものであるのに対して，本研究では，トピックの時系列変化に注目したものである．商品が，どの程度，主要なトピックを持つかは，商品の販売戦略において非常に有効であると考えられる．レビュートピックの推移パターンを分析すると，商品レビューのトピックの変化は，3 つのグループに分類できることがわかった．それぞれのグループでは，トピックの推移パターンが違い，また時間に対するトピック数の増加傾向も違う．本論文では，LDA を用いて推定したトピックに対して，分析を行うことで，時系列変化とレビュートピックの傾向変化が関係することを明らかにした．

キーワード　トピック, 時系列, レビューサイト

1. はじめに

近年，インターネットの発達に伴い，Web サイトを通じて商品に関する意見を発信する機会が増えている．そのような商品に関する意見（以下，レビューと呼ぶ）は，個人が商品を購入する際に非常に有用である．しかし，商品のレビュー数が膨大なので，レビューの内容を 1 つずつ分析する作業は非常に難しい．したがって，膨大な量のレビュー文からの商品に関する情報を抽出し，比較する方法が必要とされている．

これまでの研究は，各レビューそれぞれについてのトピックの抽出に注目したものが一般的である．落合ら [5] は，語彙間の依存構造を反映した素性を用いてトピックモデルを生成することで，表層的な表現にとらわれないトピックの抽出をした．

一方で，文書集合に対する時系列分析の研究が行われている．張ら [4] は，ニューストピックと入力時系列データの関連の有無を判断する手法を提案し，2 つが関連があると判断できる実例を上げた．水落ら [6] は，新聞記事集合に対して，一日単位で記事をトピックに分類し，時系列でトピックの関連付けを行う手法を提案した．この手法では，個別にトピックを取り出すだけでなく，それぞれトピックを時系列ごとに関連付けることにより，精度の向上を図っている．短時間に特定のトピックが多く出現する，トピックのバーストに関する研究として，高橋ら [7] の研究がある．この研究では，トピックモデルとして DTM(dynamic topic model) を用い推定したトピックに対して，トピックごとの各キーワードの条件付き確率とその日におけるキーワードのバースト度との積の和を求めることでバースト度を付与する手法を提案している．また，トピック単位のバーストが検出可能であることを示している．

本研究では，商品トピックを対象として時系列分析を行うことを目的とする．LDA によって，分析期間におけるレビューのトピックを抽出し，トピックの時間経過による変化を解析することで，トピック数の増加の傾向が 3 つのグループに分けることができ，レビューの投稿間隔とトピックの傾向変化に関係性があることがわかった．

2. トピック抽出

本研究では，LDA を用いてトピックを抽出する．LDA の処理対象として，形態素解析器 MeCab [1] で抽出されたユーザーレビューの名詞と形容詞を扱った．

2.1 潜在的ディリクレ配分法（LDA）

潜在的ディリクレ配分法：Latent Dirichlet Allocation（以下，LDA）について説明する．LDA は，Blei ら [2] によって提案されたマルチトピック抽出モデルの 1 つである．LDA は，特定の特徴的な単語の分布である「トピックの混合分布」から文書が生成されるという考えに基づいている．トピックモデルでは，トピックの混合分布に基づいて文書のトピックが生成され，各トピックの単語分布に基づいて単語が生成される．

この生成のプロセスに焦点を当て，例えば，商品のデザインに関して多く生成された文書は，デザインに関した話題を中心としたレビューであることを期待した．尚，LDA のトピック分布を求める手法には Blei らの変分ベイズと呼ばれるベイズ事後分布から近似して求める手法があるが，今回は Griffiths ら [3] が提案し，実用性が高いとされているギブスサンプリングの手法を用いた推定手法を用いる．

2.2 レビューとトピックの対応付け

本研究では，一つのレビューごとに，トピックを一対一で割り当てることにより，レビューをトピックによって表す．

あるレビューにおけるトピック数を K，1 つのレビューを r とすると，その代表トピック $z_n(n = 1, ..., K)$ は以下の式で表される．

$$z_n = \underset{z_u(u=1,\ldots,K)}{\operatorname{argmax}} \; p(z_u \mid r)$$

ここで，LDA により生成したトピックモデルを用いて，各レビュー r におけるトピック z_n の確率分布 $p(z_u \mid r)(n = 1, ..., K)$

を推定する．これはつまり，レビュー r におけるトピック分布において，確率が最大のトピックをレビュー r に割り当てていることになる．

ただし本研究では，レビューのトピック推定に用いるトピックモデルは，商品ごとのトピックの推移をみるために，商品ジャンルごとに違うものを使っている．

3. 分　　析

3.1 使用データ

対象とするレビュー文書には，2010 年 1 月 1 日 ~2015 年 7 月 9 日の楽天株式会社のネットショッピングサイト 楽天市場のレビュー文書 (注1) を用いる．各商品に含まれる属性情報は，商品名，商品コード，商品価格，商品説明文，販売方法別説明文，商品 URL，商品画像 URL，レビュー件数，レビュー平均，店舗コード，ジャンル ID である．各レビューに含まれる属性情報は，投稿者，年齢，性別，商品コード，商品名，店舗名，商品 URL，商品ジャンル ID，商品価格，購入フラグ，内容，目的，頻度，評価ポイント，レビュータイトル，レビュー内容，レビュー登録時間である． 今回使用したデータの詳細を以下に示す．分析の対象とした商品は，最初のレビュー登録から最後のレビュー登録が 30ヶ月以上 40ヶ月未満，かつ各商品レビュー件数が 100 件以上の 263 個の商品である．

3.2 レビュー数の時系列変化

分析対象である楽天市場のユーザーレビューについて，時間経過によるレビュー数の増加傾向を調査する．これらの調査を各商品ごとに行う．

分析結果を図 1 に示す．ここでは，ランダムに抽出した 6 つの商品に対して解析を行った．横軸は，その商品の最初のレビュー登録時間からの経過時間で，縦軸は，ある時点におけるレビュー数の累積である．また，横軸と縦軸は，それぞれ正規化されている．

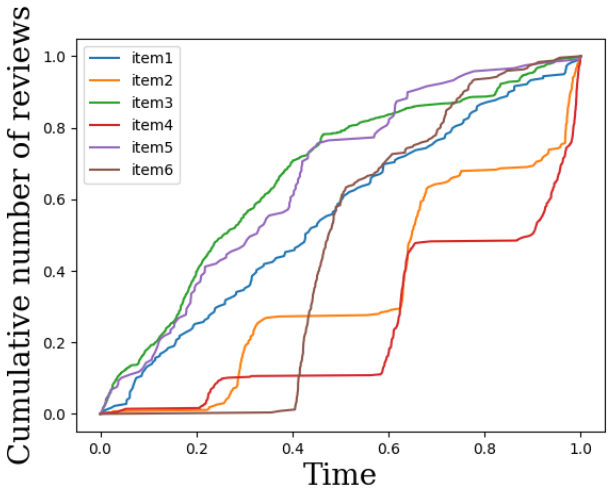


図 1　レビュー数の時系列変化

(注1)：https://rit.rakuten.co.jp/opendataj.html

3.3 レビュー増加によるトピック数の変化

まずは LDA を用いて，各レビューからトピックの抽出を行う．トピックの抽出に使うトピックモデルは，商品ジャンルごとに生成し，商品のジャンルに対応するトピックモデルによって，トピックの抽出を行った．

時間経過とともに増加するレビュー数に対する出現トピック数を図 2 に示す．横軸は，登録時間順にしたレビューである．縦軸は，その時点における出現トピックの種類の累積である．対象の商品の傾向を比べるために，各商品のレビュー数を 100 個に分割した．

まず商品のレビュー数 K_r に対して，$\lfloor K_r/100 \rfloor$ という値 k を取る．このとき，1 つのレビューを $r_i (i = 1, ..., k)$ とすると，期間 j の代表トピック $z_{k_j} (j = 1, ..., \lfloor K_r/k \rfloor)$ は以下の式で表される．

$$z_{k_j} = \underset{z_u (u=1,...,K)}{\operatorname{argmax}} \sum_{i=1}^{k} p(z_u \mid r_i)$$

つまり，各ユーザーレビューのトピックを求める際に，登録時間で並べた k 個分のレビューの確率分布を合計し，確率が最も高かったトピックを期間 j のトピック z_{k_j} とした．この方法だとレビューの余りが発生するが，全体の傾向をみることが目的であため，余ったレビューは分析対象には加えないものとした．

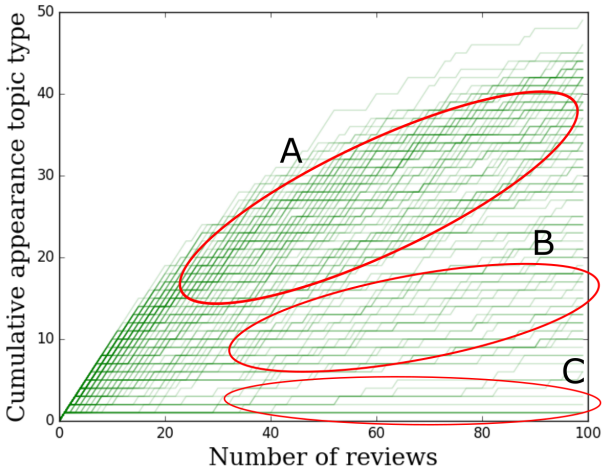


図 2　レビュー増加によるトピック数の変化

3.4 トピック数と匿名投稿者数の関係

実際の web レビューを確認すると，人為的に書かれたと思われるレビューが発見された．それらのレビューは似た内容を書かれる傾向があり，なおかつ，プロフィール（ニックネーム・性別・年齢）の表示が非公開になっている（匿名投稿者）と考えた．そこで，各商品のレビュー全体に現れるトピック数と匿名投稿者の数に相関を調べた．

商品ごとのトピック数と全体の投稿者数に対する匿名投稿者数の割合の関係を図 3 に示す．横軸は，商品ごとのトピック数である．縦軸は，匿名投稿者数の割合である．2 つの変数の相関係数は，−0.3051 であった．

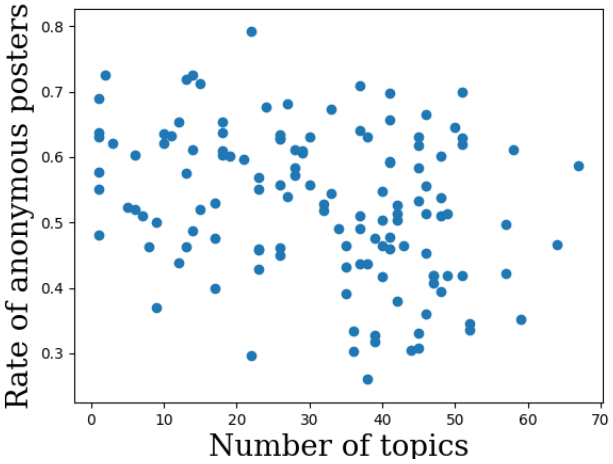


図 3　トピック数と匿名投稿者数の関係

4. 考　　察

4.1　レビュー数の時系列変化

図 1 は，ごく一部の結果だが，他の結果も観察するとレビュー数の増加傾向は，大きく 3 種類に分けられると考えられる．

1 つは，”item2”のようにほぼ同じペースで増加するような商品である．これは生活用品など，販売個数に季節的な偏りのない商品であると考えられる．実際に”item2”は,「メンテナンス」というウサギ用の飼料である．

2 つ目は，”item4”のように，グラフが階段状に変化するような商品である．これは季節性の商品で，販売個数が時期ごとに偏りがあるような商品であると考えられる．実際に”item4”は,「ボジョレヌーボー」という赤ワインであり，これは 11 月に解禁される新酒であることが知られている．

3 つ目は，”item5”のように，途中から一気にレビュー数が増加するような商品である．何かのきっかけで，販売個数が伸びるような商品であると考えられる．実際に”item5”は,「簡易トイレレスキュー 200X」という非常用の簡易トイレであり，調べると 3 月 16 日以降に急激にレビュー数が増えることがわかった．これは，3.11 東日本大震災の影響で商品販売数が増加したためであると考えられる．

以上の結果から，ユーザーレビューを登録時間で評価すると，それぞれの商品の特性が現れることがわかった．ユーザーレビューのトピックに注目して，分析を行う際には，各商品の登録時間ではなく，登録順にすることで，商品の季節性の影響を抑えられると考える．

4.2　レビュー増加によるトピック数の変化

図 2 のように，3 つのグループに分かれているように見える．

グループ A は，レビュー数の増加とともにトピックが単調に増えていくグループである．このグループは，様々なトピックが次々に現れるようなレビューが多くある商品であると考えられる．つまり，レビューごとにある程度内容が違い，また，一定の間隔で新しいトピックが増えることから，用途や評価される特徴の多い商品であることが考えられる．

グループ B は，一定レビュー数出現後に現れるトピック数が飽和するグループである．特に，主要なトピックが複数あるようなものである．つまり，新しいトピックが増えにくいような性質を持っており，用途などが限られた商品であることが考えられる．

グループ C は，レビュー数が増加してもトピックが増えないグループである．どのレビューも同じようなトピックを示しており，ほとんどのレビューが似たことを議論しているものと考えられる．

4.3　商品ごとのトピック推移

図 2 のように，図 4 に示す，それぞれのグループの典型的商品のトピック推移を分析する．グループ A からはトピック数が一番多い商品 a を，グループ B からは商品 b_1 と商品 b_2 を，グループ C からはトピック数が一番少ない商品 c を対象とした．図 5 は，横軸に登録時間順のレビュー r_t をとり，縦軸にトピック z とレビュー間の時間間隔の逆数 f を示す．

図 5 をみると，グループ A から B，B から C に移ると，トピックの分散が少しずつ減っていく様子がわかる．これはグループ A の方が商品の全トピック数が多く，グループ C の方が全トピック数が少ないからである．

図 5(a) では，トピックが全体的に分布していることがわかる．レビュー間の時間間隔の変化 f を見ると，レビュー $r_t = 68$ 以降には，12 ∼ 23 番目のトピックが出現しない．これは f のピークが来たタイミング，つまり短期間にレビューが投稿されたときに，トピックの傾向が変化したものであると考えられる．

図 5(b_1) では，いくつかのトピックに各期間のトピックが集中していることがわかる．全体を通して，30 番目のトピックに集中している．ここで，レビュー $r_t = 52$ で現れる f の山がある．比較的短期間にレビューが連続したことを示しているが，これらのレビューのトピック $z = 54$ に固まっている．これは全体から見ても，局所的にトピックが集中しており,「一部の話題についての議論が活発化している」,「限られたトピックを繰り返し投稿するユーザの存在」などの原因が考えられる．

図 5(b_2) でも，2, 39, 57, 65 番目のトピックに各期間のトピックが集中していることがわかる．また出現トピックの種類が商品 b_1 に比べて少ないため，特定のトピック z に集中していることがはっきりとしている．局所的に見たとき，レビュー $r_t = 66$ 周辺で，f のピークが来て，短期間にレビューが投稿されたことがわかる．$r_t > 66$ では，$z = 65$ のトピックが出現していない．これは $r_t = 66$ 周辺の短期間にレビューが投稿された後，トピックの傾向が変わったものと考えられる．

図 5(c) の結果から分かる通り，商品 C は全てのレビューが $z = 49$ のトピックを示している．商品の特性上，限られたトピックについてのみ注目するレビューが多くなることが原因であると考えられる．また限られたトピックを繰り返し投稿するユーザの存在も考えれる．

図 5(b_2),(c) について，f が全体として，右肩下がりの傾向にある．これはレビューの投稿間隔が少しずつ大きくなっており，商品寿命と関係があるものと考えられる．

図 5(b_1),(b_2), (c) には，主要なトピックがあることがわかる．

これらのトピックには，その商品ユーザーレビューが特に注目しているポイントがあると考えられる．そのため，商品戦略において有効な判断材料になりえると考える．

図 5(a),(b_1),(b_2) においては，短期間にレビュー投稿があると，トピックの傾向が変化する様子が観測できた．レビューの投稿間隔とトピックの変化に，何らかの関係性があるものと考えられ，詳しい解析が必要になる．

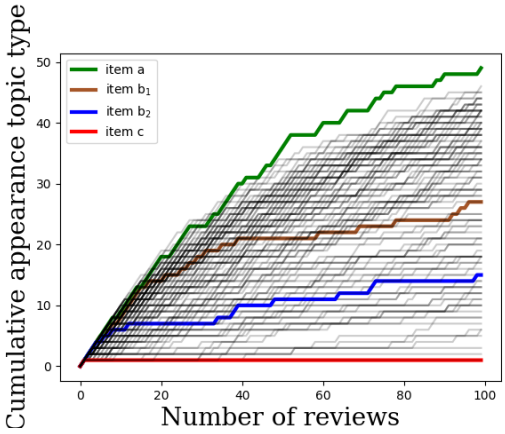


図 4　レビュー数の増加によるトピック数の変化

4.4　トピック数と匿名投稿者数の関係

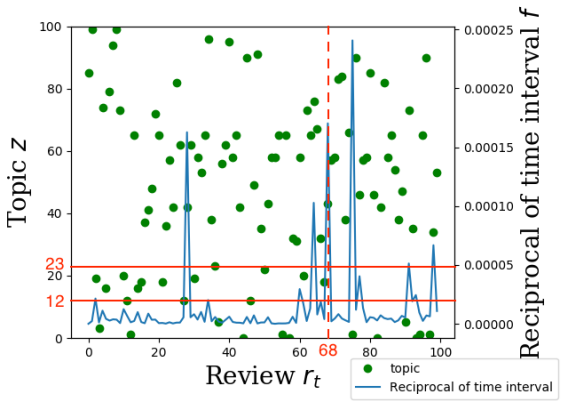
レビューの投稿者が，匿名になっている場合，それらのレビューが同一人物によってなされたものか判定することは困難である．しかし，全ての匿名投稿者が同一人物であるとは考え難く，ほとんどの匿名投稿が別々の人物によるものと考えられる．つまり理想的なレビュー群では，匿名投稿者数とレビューの出現トピックには，相関が無いと期待される．また，匿名であることにより，心理的な障壁が少なくなり，自由な意見を書き込めると考えると，相関係数が正に傾くことが期待できる．

一方，図 3 の結果から，商品ごとのトピック数と匿名投稿者数の割合は，負の弱い相関があることがわかった．つまり，レビューに出現するトピック数が減ると，匿名投稿者の数が上がる傾向がある．匿名で限られたトピックを繰り返し投稿するユーザの存在が考えられる．この結果は，4.3 での考えとも一致する．

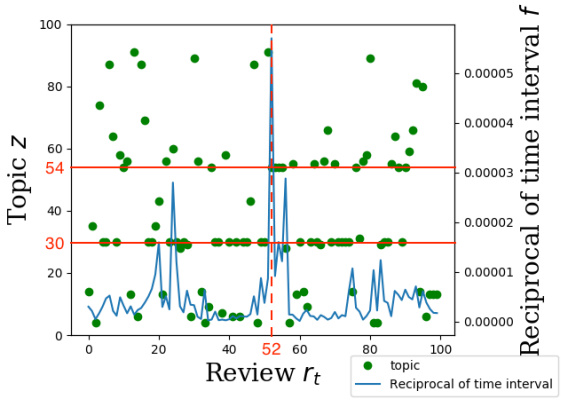
5.　関 連 研 究

トピックモデルを用いたレビュー文書の分析の研究として，落合ら [5] は，語彙間の依存構造を反映した素性を用いてトピックモデルを生成することで，表層的な表現にとらわれないトピックの抽出を可能にした．

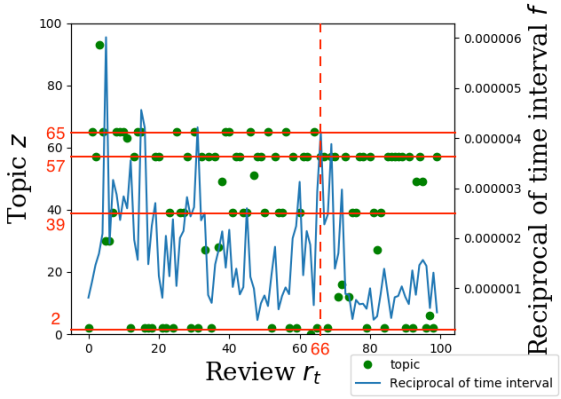
トピックモデルによる文書集合に対する時系列分析として，ニュースや新聞記事を用いた研究が広く行われている．張ら [4] は，ニューストピックと入力時系列データの関連の有無を判断する手法を提案し，2 つが関連があると判断できる実例を上げた．水落ら [6] は，新聞記事集合に対して，一日単位で記事をトピックに分類し，時系列でトピックの関連付けを行う手法を提案した．この手法では，個別にトピックを取り出すだけでなく，それぞれトピックを時系列ごとに関連付けることにより，精度



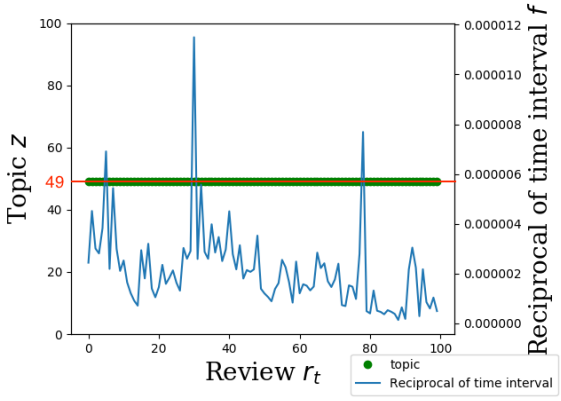
(a) 商品 a



(b_1) 商品 b_1



(b_2) 商品 b_2



(c) 商品 c

図 5　レビューの変遷と投稿間隔

の向上を図っている．

短時間に特定のトピックが多く出現する，トピックのバーストに関する研究として，高橋ら [7] の研究がある．この研究では，トピックモデルとして DTM(Dynamic Topic Model) を用い推定したトピックに対して，トピックごとの各キーワードの条件付き確率とその日におけるキーワードのバースト度との積の和を求めることでバースト度を付与する手法を提案した．また，トピック単位のバーストが検出可能であることを示した．

オンラインレビューサイトの評点に対して，時系列データを解析している研究として，山岸ら [8] の研究がある．山岸らは，ユーザーの基本評点行動として多項式分布モデルを仮定し，尤度比検定により，異常期間を検出することを特徴とする手法を提案し，その有用性を示した．

6. おわりに

本論文では，レビュー観点の推移パターンに基づく抽出手法を提案した．そこでは，商品をジャンルごとに分けで，LDA を用いてトピックモデルを作成した後，各レビューのトピックを求め，時系列での変化を解析している．

実際に運用されているオンラインショップのレビュー情報を用いて解析を行った．それにより，時間経過によって出現するトピック数の増加傾向が商品ごとに違い，それらの傾向は 3 つに分類できることが分かった．またレビューの投稿間隔と出現トピックとの関係を見ることで，短期間に投稿されるレビューとトピック傾向が変化する現象の関係が観測できた．一方で商品のトピック数と匿名投稿者数には，負の相関があった．これは限られたトピックを繰り返し投稿するユーザの存在が考えられ，このようなユーザを分析の対象から外すことでより，本質的なトピックの傾向を捉えることができる．

今後の課題として，レビュートピックの増加傾向から分けた 3 つのグループと商品ジャンルとの対応関係の調査が必要である．また，本研究では，ジャンルごとに違ったトピックモデルを用いて，トピックの確率分布を求めたが，すべての商品に対して，同じトピックモデルを使用した場合との違いも評価することで，所属ジャンルのレビューが少ない商品に対しても，分析の対象に加えることできると思われる．さらに考察で得意なレビュー傾向を持つ商品を発見することができたが，トピックによるレビュー傾向を数値化することで，異常なレビュー傾向の商品判別手法を明らかにすることが期待できる．

謝　辞

本研究は JSPS 科研費 JP16H02907 の助成を受けたものです．また，楽天株式会社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「楽天公開データ」を利用しました．ここに記して謝意を示します．

文　献

[1] Mecab: Yet another part-of-speech and morphological analyzer, 2017.

[2] David M. Blei et al. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[3] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proc. National Academy of Sciences*, Vol. 101, pp. 5228–5235, 2004.

[4] 張一萌, 何書勉, 小山聡. 時系列データに意味的に関連するニューストピックの発見. *DBSJ letters*, Vol. 5, No. 1, pp. 133–136, jun 2006.

[5] 落合恵理香, 小林一郎. 商品の評価を対象としたレビュー文書の分析. 言語処理学会 第 18 回年次大会 発表論文集, pp. 1176–1179, mar 2012.

[6] 水落大史, 井上悦子, 吉廣卓哉, 村川猛彦, 中川優. 新聞記事集合に対する時系列のトピック抽出. *DEIM Forum 2010*, pp. 133–136, 2010.

[7] 高橋佑介, 横本大輔, 宇津呂武仁, 吉岡真治. Analyzing burst of topics in news stream. 研究報告自然言語処理（NL）, Vol. 2011, No. 6, pp. 1–6, nov 2011.

[8] 山岸祐己, 大久保誠也. オンラインレビューサイトの評点時系列データからの異常検出. pp. 1176–1179, 2012.