確率的トピックモデルを用いた評判文書における意外な評価視点の発見とそれに基づく情報推薦

古橋 慎之介[1]， 内田 理[2]

[1] 東海大学大学院工学研究科情報理工学専攻
4bdrm019@mail.tokai-u.jp

[2] 東海大学情報理工学部情報科学科
o-uchida@tokai.ac.jp

1. はじめに

ユーザの嗜好にできる限り適合した情報を推薦する目的で，様々な情報推薦アルゴリズムが提案されてきた．代表的なものとしては，コンテンツベースフィルタリングや協調フィルタリングなどが挙げられる．しかし，嗜好への過度な適合により，ユーザが新規性を感じられない情報が推薦されてしまうという問題点が指摘されている．この問題に対応するために，ユーザにとって魅力的な情報や意外性のある情報の抽出，及び推薦を目的としたアルゴリズムに関する研究が注目されている[1]．ユーザに対してこれらのような情報を提供することは，ユーザの興味の幅を広げ，人生をより豊かなものにするきっかけとして有用であると考えられる．

本稿では，レビュー文のような評判文書から製品について意外性のある情報を抽出し，ユーザに提供する手法について検討する．製品について意外性のある情報をユーザに提供することで，ユーザの製品に対する興味を獲得し，消費行動につなげられるのではないかと考えられる．そこで，本研究では，レビュー文書における意外な情報は，他の評価視点と言及内容の異なる評価視点を持つレビュー文に存在すると仮定し，意外な評価視点の発見，及びそれに基づく情報推薦手法を提案し，その妥当性について検討する．

2. 関連研究

Tsukuda ら[2]は，Wikipedia の記事における見出し語に対して意外な情報となり得る単語を，その見出し語の記事中から抽出する手法を提案している．Tsukuda らの手法では，非典型度と認知度という二つの指標を用いて意外性を評価している．非典型度と認知度はそれぞれ，見出し語との概念的な距離，及び記事集合における記事の重要度に基づき算出される．Tsukuda らは，評価実験の結果をもとにこれら二つの指標が情報の意外性をモデリングする上で有効であると主張している．

3. 提案手法

ある製品のレビュー文書群における 1 センテンスのレビュー文の意外度を評価し，それを基に推薦リスト（推薦レビュー一覧）を作成することが，提案手法の目的である（以降，レビュー中の 1 センテンスを，単に「レビュー文」と表記する）．提案手法の流れは以下の通りである．まず，評価視点を抽出する．具体的には，レビュー文書群に対して確率的トピックモデルを適用し，評価視点となるトピックを得る．次に，得られたトピック情報からあるトピックとその他のトピックとの内容的・概念的な相違度を評価することで，他のトピックと明らかに言及内容が異なる(孤立性の高い)トピックを抽出する．そして，得られた相違度を基にレビュー文の意外性を評価し，意外度の高いレビュー文を用いて推薦リストを作成する．

3.1. 評価視点の抽出

本研究では，評価視点の抽出に Titov ら[3]が提案した Multi-Grain LDA(MG-LDA)を用いる．MG-LDA とは，Blei ら[4]によって提案された確率的生成モデルを用いた文書モデル化手法として知られる Latent Dirichlet Allocation(LDA)を拡張した手法であり，評判文書のモデル化を目的としている．MG-LDA では，製品特徴を表現するグローバルトピックと評価視点を表現するローカルトピックを推定する．グローバルトピックは，通常の LDA で推定されるようなトピックであり，ローカルトピックは，ウィンドウと呼ばれる隣接センテンス n 件（ウィンドウ幅）をまとめて 1 文書と見立てた時に推定されるトピックである．本研究では，高精度な評価視点の抽出を行うために，推薦リストを作成する対象となる製品のレビュー文書の他に，その製品と同種，同カテゴリのレビュー文書も合わせて入力文書とする．また，MG-LDA の素性として，MeCab[5]による形態素解析の結果より，一般名詞，固有名詞，一般動詞，自立動詞，ナイ形容詞語幹，形容動詞語幹を利用する．ただし，「ある」や「こと」などの一部の形態素は，評価視点を抽出する上で特徴になりにくいことから，

ストップワードとして素性から除外した（ストップワードの設定は手動で行なった）．

3.2. 評価視点トピックの孤立性の評価

MG-LDA で得られたローカルトピックを基に，各トピック間での相違度を算出し，相違度行列を作成する．本研究では，二つの集合の類似度を測る指標である Jaccard 係数に基づき相違度を算出する．任意の二つのローカルトピック t_i, t_j 間の Jaccard 係数は，式(1)で算出される．

$$Jaccard(t_i,t_j)=\frac{A}{B} \tag{1}$$

ここで，A，B はそれぞれ t_i，t_j における語彙集合の積集合，和集合であり，各ローカルトピックで出現頻度が 2 以上の単語を語彙集合とする．Jaccard 係数は，二つトピック間の類似度が最も高い場合に最大値 1 となるため，任意の二つのローカルトピック間の相違度 *Diff* を式(2)で与える．

$$Diff(t_i,t_j)=1-Jaccard(t_i,t_j) \tag{2}$$

次に，得られた相違度 *Diff* を基に，各ローカルトピックの孤立性を評価する．本研究では，Brin ら[6]によって提案された PageRank の考え方を用いる．PageRank とは，Web のリンク構造から Web ページの重要度を算出するアルゴリズムである．本研究では，相違度を PageRank における推移確率とみなして PageRank 値を求める．これにより，他のトピックと相違度の高いトピックほど PageRank 値が高くなることから，トピックの孤立性を評価することができる．PageRank を計算するために，まず，式(2)によって得られた相違度行列について式(3)を用いてローカルトピック t_i から t_j への推移確率 *Diff′* を算出し，推移確率行列を求める．

$$Diff'(t_i,t_j)=\frac{Diff(t_i,t_j)}{\sum_{t_k\in L(t_i)} Diff(t_i,t_k)} \tag{3}$$

ここで，$L(t_i)$は t_i の隣接トピックである．求められた推移確率行列から PageRank を算出し，それを用いて各ローカルトピックの孤立度 I を算出する．

$$I(t_i)=\frac{1-d}{N}+d\sum_{t_j\in L(t_i)} I(t_j)Diff'(t_j,t_i) \tag{4}$$

ここで，d はダンピングファクターである（本研究では Brin ら[6]と同様，0.85 に設定した）．ローカルトピックの孤立度 I は，他のローカルトピックとの相違度が大きい場合，または孤立度の高いローカルトピックとの相違度が低い場合に大きくなる．

3.3. レビュー文の意外性の評価

レビュー文は，トピック分布で表現されることから，評価視点を一意に決定することは困難である．そのため，ローカルトピックの孤立度を算出しただけではレビュー文の意外性を評価することはできない．そこで，ローカルトピック分布を用いてレビュー文の意外度を求める．本研究では，前処理として，半数以上の形態素にグローバルトピックが割り当てられている，または，レビュー文の長さが平均長以下のレビュー文を除外しておく．次に，レビュー文のトピック分布を作成するために，クエリと文書の関連性を測る指標である Okapi BM25[7]により，各ローカルトピックの単語に対してスコアリングを行う．ここで，逆文書頻度の算出には，tf·idf 法における IDF を採用することとした．ローカルトピック t_i の単語 w のスコア $score(w, t_i)$は式(5), (6)を用いて算出する．

$$score(w,t_i)=IDF(w)\cdot\frac{freq(w,t_i)(k+1)}{freq(w,t_i)+k\left\{(1-b)+b\dfrac{length(t_i)}{avelen}\right\}} \tag{5}$$

$$IDF(w)=\log_2\frac{|T_{loc}|}{DF(w)}+1 \tag{6}$$

ここで，$freq(w, t_i)$は，t_iにおける w の頻度，$length(t_i)$は t_i の単語数，*avelen* はローカルトピックの平均単語数である．T_{loc} はローカルトピック集合，$DF(w)$は w を含むローカルトピック数である．また，k, b はパラメータであり，本研究では $k=2, b=0.75$ とした．各レビュー文に対して，ローカルトピック毎に単語スコアの総和を算出することにより，トピック分布を求め，得られたローカルトピック分布からレビュー文 s の意外度 *Srd* を式(7)で算出する．

$$Srd(s)=\log_2(|M_s|)\sum_{u\in T_{loc}} I'(u)\cdot\frac{sum_{u,s}}{\sum_{v\in T_{loc}} sum_{v,s}} \tag{7}$$

$$sum_{t,s}=\sum_{w_{t,s}} score(w_{t,s},t) \tag{8}$$

$$I'(t_i')=\begin{cases} I(t_i') & if \quad i=1 \\ I(t_1')\prod_{j=2}^{i}\alpha\dfrac{I(t_j')}{I(t_{j-1}')} & otherwise \end{cases} \tag{9}$$

ここで，M_s は s における形態素集合，$w_{t,s}$ は s におけるローカルトピック t ($t \in T_{loc}$)が割り当てられた形態素を表す．ローカルトピック分布中で孤立度の高いローカルトピックの割合が大きいほど，意外度は大きくなる．また，t_i'は T_{loc} 中で I について降順でソートされたローカルトピックである．α はパラメータであり，本研究では α=1.5 とする．α の値が大きくなるほど，孤立度の大きいローカルトピックを含むレビュー文の *Srd* は相対的に大きくなる．

3.4. レビュー文の提示

推薦リストの作成にあたり，本手法では，レビュー文の意外性を単語ベースで評価しているため，単純に *Srd* の値が上位 10 件のレビュー文を取り出した場合，内容に重複が起きる可能性がある．そこで，Carbonell ら[8]によって提案された情報検索における内容の網羅性と冗長性の削減を考慮した指標である MMR(Maximal Marginal Relevance)を応用した指標を利用することで，内容に重複のない推薦リストを作成する．本研究では，レビュー文 s の *Srd* に基づく MMR である MMR_{Srd} を式(10)で算出することとした．

$$MMR_{Srd}(s) = \arg\max_{s \in R \setminus S}[\lambda Srd'(s) - (1-\lambda)\max_{s' \in S}(Sim(s,s'))] \tag{10}$$

ここで，R は *Srd* で降順にソートされたレビュー文集合，S は抽出済みのレビュー文集合を表す．λ は各項の重みを決定するパラメータであり，本研究では λ=0.3 と設定した．また，*Srd'* は *Srd* を最大値が 1，最小値が 0 となるように正規化された値であり，*Sim*(*s* ,*s'*)は s と s'とのコサイン類似度である．集合 $R \setminus S$ 中で計算された MMR_{Srd} が最も大きいレビュー文を集合 S に加え，その都度 $R \setminus S$ 中のレビュー文の MMR_{Srd} の値を更新する．$R \setminus S$ が空集合になるまで抽出を繰り返すことで，再ランキングリスト S を得る．そして，リスト S から上位 10 件をユーザに提示するレビュー文の推薦リストとする．なお，「購入経緯」や「配送」に関連する内容である，または，1 センテンスでは意味の通らないレビュー文は手動で除外する．

4. 評価実験

4.1. 実験設定

本手法の有用性を検証するために，16 名の被験者による評価実験を行った．被験者の内訳は 20 代の男性 15 名，女性 1 名である．提案手法により作成されたレビュー文 10 件から成る推薦リストを被験者に提示し，各レビュー文に対して意外性を 5 段階で評価をしてもらった．

本実験では，楽天市場[9]における Android タブレット関連のレビュー文書 10634 件を用い，レビュー文書 995 件から成る「Zenithink C93」を推薦リスト作成の対象とした． MG-LDA の設定は，グローバルトピック数を 40，ローカルトピック数を 20，ウィンドウ幅を 3 とした．また，モデルのパラメータ推定には崩壊型ギブスサンプリングを用い，反復回数は 800 回とした．

4.2. 評価方法

被験者の評価に基づく評価方法として，NDCG(Normalized Discounted Cumulative Gain)[10]を用いた．NDCG とは，システムが出力したランキング結果の理想的なランキング結果への近さを示す指標であり，k 位までの順位付けを行ったときの *NDCG@k* は式(11)で算出される．

$$NDCG@k = Z\left(v_1 + \sum_{i=2}^{k} \frac{v_i}{\log_2 i}\right) \tag{11}$$

ここで，v_i は順位 i の評価値である．また，Z は理想の順位付けが出力されたとき NDCG の値を 1 とするための正規化項である．

4.3. 実験結果と考察

孤立度の高い順に，ローカルトピック（評価視点）と単語スコアが上位の語句を表 1 に示す．また，作成された推薦リストのレビュー文について，意外度の高い順にレビュー文と被験者による評価値の平均，及び標準偏差を表 2 に示す．さらに，評価者毎に算出した NDCG@5 と NDCG@10 の最大値，最小値，平均値と標準偏差を表 3 に示す．ここで，本実験における評価者間の評価値の一致度を示す重み付き κ 係数は，0.544～0.956（平均値 0.806）であった．

表 1 より，孤立度の高いローカルトピックとして，他のローカルトピックとは言及内容の異なる「購入経緯」や「配送」などのトピックが上位にランキングされたことは，レビュー文書における内容的・概念的な局所性を評価するという意味で，仮定を肯定する結果が得られたと考えられる．なお，これらの評価視点の内容を含むレビュー文は，推薦リスト作成時に除外している．また，表 3 より NDCG@5 のとき低い値を示していることから，推薦リスト上位のレビュー文に対して，あまり意外と感じていない評価者が多いことが分かる．しかし，NDCG@10 では，NDCG@5 に比べ大きな値となっており，順位に大きな差があるレビュー文については，提案手法に基づく順序関係が，評価者の感覚とある程度適合しているのではないかと考えられる．

5. まとめと今後の課題

本研究では，確率的トピックモデルによる意外な評価視点の発見とそれに基づく情報推薦手法を提案した．

今後の課題として，グローバルトピックの情報を基に製品固有の意外な情報を抽出することや，メーカーが公開していない製品に関するネガティブな情報に意外性を感じる評価者が多かったことから，レビュー文の評価極性を利用することなどが挙げられる．また，グローバルトピックを推定せずにローカルトピックを抽出するトピックモデルである Local LDA[11]や STM[12]の利用を検討している．さらに，商品ページの情報の利用することによる意外な情報の抽出手法を検討したい．

参考文献

[1] 奥健太, “セレンディピティ指向情報推薦の研究動向”, 知能と情報（日本知能情報ファジィ学会誌）, Vol.25, No.1, pp.2-10, 2013.

[2] K. Tsukuda, H. Ohshima, M. Yamamoto, H. Iwasaki, K. Tanaka, “Discovering Unexpected Information on the Basis of Popularity/unpopularity Analysis of Coordinate Objects and Their Relationships”, Proc. of the 28th Annual ACM Symposium on Applied Computing, pp.878-885, 2013.

[3] I. Titov, R. McDonald, “Modeling Online Reviews with Multi-grain Topic Models”, Proc. of the 17th International World Wide Web Conference (WWW2008), pp.111-120, 2008.

[4] D. Blei, A. Ng, J. Michael, “Latent Dirichlet Allocation”, The Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.

[5] MeCab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[6] S. Brins, L. Page, “The Anatomy of a Large-scale Hypertextual Web Search Engine”, Proc. of the 7th International Conference on World Wide Web (WWW1998), pp.107-117, 1998.

[7] K. Jones, S. Walker, S. Robertson, “A Probabilistic Model of Information Retrieval: development and comparative experiments”, Information Processing and Management, Vol. 36, Issue 6, pp.779-808, 2000.

[8] J. Carbonell, J. Goldstein, “The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries”, Proc. of the 21th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, pp.335-336, 1998.

[9] 楽天市場, http://www.rakuten.co.jp/

[10] K. Jarvelin, J. Kekalainen, “Cumulated Gain-Based Evaluation of IR Techniques”, ACM Transactions on Information Systems, Vol.20, No.4, pp.422-446, 2002.

[11] S. Brody, N. Elhadad, “An Unsupervised Aspect-sentiment Model for Online Reviews”, Proc. of Human Language Technologies: 2010 Annual Conference of the North American Chapter of the ACL, pp.804-812, 2010.

[12] L. Du, W. Buntine, H. Jin, “A Segmented Topic Model Based on the Two-parameter Poisson-dirichlet Process”, Machine learning, Vol.81, No.1, pp.5-19, 2010.

表 1　ローカルトピックと代表語句

評価視点	代表語句
購入経緯	購入, 子供, プレゼント, 買う, 使う, 自分, 喜ぶ, 母, ゲーム
配送	届く, 注文, 商品, 早い, 到着, 発送, 梱包, 対応, 配送, 楽しみ, 次
無線設定	接続, 設定, 簡単, wifi, ネット, 出来る, 問題, 無線, つながる
付属品	キーボード, ケース, カバー, おまけ, 使う, 付く, 良い, つく
バッテリー	充電, 電源, バッテリー, 起動, 入れる, 入る, 使用, ボタン
画面・画質	画面, 大きい, 見る, 綺麗, 見やすい, きれい, 画像, 満足
入力操作	キーボード, 入力, USB, 変換, 付属, 文字, 使える, キー, 使う
値段	タブレット, 購入, 安い, 探す, 価格, 欲しい, 値段, 使う, 買う
反応速度	反応, 動作, 遅い, 悪い, 速度, タッチ, 画面, タッチパネル, 感じ
アプリ	アプリ, ダウンロード, 設定, 出来る, インストール, 入れる

表 2　推薦リスト

レビュー文	平均	標準偏差
無線 LAN につながるか心配でしたが、なんとあっさりつながってしまった	1.938	1.248
いろんなことにフル活用するヘビーユーザーにはちょっと物足りないかな	3.250	0.901
重さもキーボードを含めて重いとは感じず、なかなかいい感じです	2.750	1.199
まだ使いこなせていませんが、古いノート PC よりも起動、終了も非常に早く、当初の目的には十分です	2.563	1.368
動作も早く、また、10inch の割に軽量でびっくりしました	3.188	1.333
しかも、カバー背面に衝立が付いているのでテーブルにさっと置けて打てるのがいいですね	3.063	1.197
画面が大きいので高齢の家族も喜んでいます	2.188	1.130
音声入力により簡単に検索できるので、料理のレシピとして活躍しそうです	3.250	1.250
ＡＣアダプターは強度的に不安があるので注意する必要が有りそう	4.125	0.927
早速軽くさわってみましたが、スルスル動くし、画質も想像より綺麗	2.875	1.409

表 3　NDCG の平均値

	NDCG@5	NDCG@10
最大値	0.870	0.931
最小値	0.489	0.745
平均値	0.657	0.826
標準偏差	0.0861	0.0474