1K5-NFC-05b-4in

テキストデータを利用したユーザーモデリング手法の一提案

Proposing user modeling method using text data.

川島健佑 *1　　本村陽一 *1*2

*1東京工業大学大学院 Tokyo Institute of Technology　　*2産業技術総合研究所 サービス工学研究センター National Institute of Advanced Industrial Science and Technology

At web sites, such as Amazon, there are so many reviews. The text data such as review or business records for attributes and preferences, such as age and gender are reflected, it can be regarded as an important source of information. In order to analyze of such a review, we using probabilistic latent semantic structure analysis. The analysis for 82 users of who posted the 653 movie, and result to classify people and word.

1. はじめに

近年，ユーザのレビューを投稿できる Web サイトが数多く存在し，レビューが購入や視聴を行う際の判断基準となっている．Amazon や eBay のようなショッピングサイトは，アイテムに対するレビューを簡単に作成・閲覧できる機能を提供している．これらのサイトでは，実際にアイテムを購入した多くの人の意見を得ることができるためユーザにとって有益なものとなっている [岩井 2013]．レビューや業務記録といったテキストデータには年齢や性別などの属性や嗜好が反映されているため，重要な情報源として捉えることができる．吉田らは実際の映画に投稿されたレビューに基づき，ユーザーの属性に類似した投稿者が好む映画をユーザーに推薦する方法を提案している [吉田 2011]．また，サービス工学ではサービスを提供するだけでなくサービスの評価計測を行い，利用者の満足度としてシステムにフィードバックすることを提唱している．テキストデータを基にユーザをモデル化することができれば，従来見逃していた重要な知見を得られることが期待でき，業務の評価や分析，推薦システムの構築の一助となると考えた．本研究では確率的潜在構造分析を用いて，テキストデータを利用したモデル化手法の一提案を行う．

2. 提案手法

モデルの構築にあたって，確率的潜在意味構造分析 (確率的潜在意味解析 (PLSA[Hoffman 1999]) とベイジアンネットワーク [本村 2006] による構造モデル化) の結果，レビューデータに基づいた顧客像と顧客の行動を確率的モデルとして構築する．PLSA(Probabilistic Lantent Semantic Analysis) では，意味的な隠れ属性 $c_l(l = 1, 2, ..., l)$ のもと，レビュー $r_i(i = 1, 2, ..., n)$ と単語 $w_j(j = 1, 2, ..., m)$ の生起は独立と考え，r_i と w_j の同時確率 $P(r_i, w_j)$ を式 (1) のように表す．

$$P(r_i, w_j) = \sum_k P(r_i|c_l)P(w_j|c_l)P(c_l) \quad (1)$$

レビュー r_i における単語 w_j の実際の出現回数を $n(r_i, w_j)$ とすると，データの対数尤度

$$L = \sum_i \sum_j n(r_i, w_j) \log P(r_i, w_j) \quad (2)$$

連絡先: 川島健佑, 東京工業大学総合理工学研究科, 神奈川県横浜市緑区長津田町 4259, kawashima.k.ai@m.titech.ac.jp

を最大にする $P(c_l)$, $P(r_i|c_l)$, $P(w_j|c_l)$ を EM アルゴリズムで計算し最尤推定を行う．

PLSA は大量のデータから有用な潜在的なクラスタを抽出できるが，その潜在クラスタの意味がよくわからないために，実際の活用が難しいという問題がある．そこで，潜在クラスタがどんなものであるかを示す多様な説明変数との関係性をベイジアンネットワーク (図 1) によって表すことを考える．

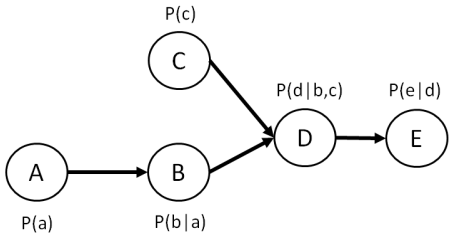

図 1: ベイジアンネットワーク

3. 実験結果

653 タイトルの映画に対してレビューに対して mecab[Kudo 2004] を使用した形態素解析を行った．形態素解析を行った後に投稿した 82 人に対して PLSA を行った結果，2 つのクラスターに分類することができた．大きく分類すると CL1 はレビューを投稿した映画に対してネガティブな意見を言うクラスター，CL2 はポジティブな意見を言うクラスターとなった (表 1)．

表 1: クラスター分析結果

クラスター	所属するワード (一例)
CL1(ネガティブ)	えぐい，がんばれ，単純
CL2(ポジティブ)	満足，楽しみ，うまい

分類したクラスターをもとに，ベイジアンネットワークの構築を行った (図 2)．モデルの構築はベイジアンネットワーク構築ソフトウェアである Bayonet[本村 2003] を使用し，構造の探索アルゴリズムは Greedy Search Algorithm を用いる．Greedy Search Algorithm とは近似アルゴリズムの Greedy Algorithm(欲張り法) を利用した探索アルゴリズムであり，各

ノードに対して親の組み合わせを Greedy Search によって決定し，有向グラフを構築する．説明変数には投稿されたレビューの映画のジャンル (表 2) を算出して使用した．構築されたベイジアンネットワークは，目的変数に直結するノード (説明変数) のみでは，何がクラスターの分類に大きく関わっていくのか判断しにくい．したがって，Bayonet の確率推論を行うことで，影響の強い変数を見つけていく (表 3)．

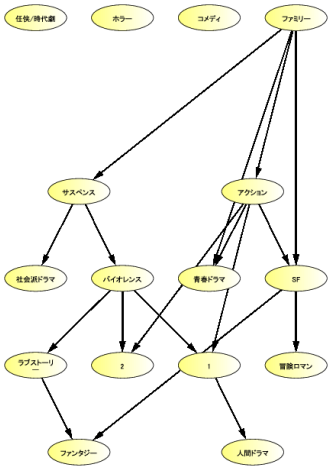

図 2: 構築されたベイジアンネットワーク

表 2: 顧客行動のベイジアンネットワーク作成に使用した映画ジャンル

SF
ラブストーリー
アクション
冒険ロマン
サスペンス
バイオレンス
ファンタジー
社会派ドラマ
人間ドラマ
青春ドラマ
ホラー
ファミリー
任侠/時代劇
コメディ

表 3: 影響の強い説明変数

CL 群	顧客の特徴
CL1	アクション
CL2	なし

4. 考察

確率推論を行った結果，アクション映画を観た 60%のユーザーはネガティブなレビューを投稿していることが判明した．理由として考えられるのはアクションはインパクトが大きいため，マイナスイメージも強くユーザーの心に残ったのではないかと考えられる．また，ポジティブなレビューに対して影響をあたえる変数は見つからなかった．

5. おわりに

本研究では，映画に関するレビューテキストデータに対して確率的潜在意味構造解析を実行することでレビューテキストとユーザーを 2 つのセグメントに分類した．さらにそのセグメントを説明する変数を探索するためにベイジアンネットを用いてモデル化し、そのモデルの上で確率推論を実行することでネガティブなレビューとユーザーはアクション映画に関するものが多いという結果が得られた．今回はポジティブなレビュー、ユーザーに対しては顕著な説明変数が得られなかった。今後、形態素解析を改良し，レビューテキストに多い話し言葉に対応した形態素解析や，レビューテキストを追加することで，新たなセグメントを抽出し分析することも必要である．また，アンケートデータやユーザーの属性を変数に加えることで，新たな説明変数を加えた確率的潜在意味構造モデルを作成することによっても，新たな知見を抽出することも今後の課題である．

参考文献

[岩井 2013] 岩井秀成, 池田郁, 土方嘉徳, 西田正吾: レビュー文を対象としたあらすじ分類手法の提案とあらすじ非表示システムの開発, 電子情報通信学会論文誌. D, 情報・システム J96-D(5), 1222-1234 (2013).

[吉田 2011] 吉田真, 本村陽一, 梅津文, 横井健: レビューデータの分析に基づく映画推薦手法の提案, 情報処理学会全国大会講演論文集 2011(1), 629-631 (2011).

[Hoffman 1999] Thomas Hofmann: Probabilistic Lantent Semantic Analysis, Proc. UAI ’99, pp.289-296, 1999

[本村 2006] 本村陽一, 岩崎弘利: ベイジアンネットワーク技術, 東京電機大学出版局 (2006)

[Kudo 2004] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 (2004).

[本村 2003] 本村陽一: ベイジアンネットソフトウェア Bayonet, 計測と制御, Vol,42, No.8, pp. 693-694 (2003).