4B-2

ノード属性を用いた特徴的リンク傾向分析

伏見　卓恭†　　斉藤　和巳†　　風間　一洋‡
†静岡県立大学大学院経営情報イノベーション研究科　　‡和歌山大学システム工学部

1　はじめに

Web 技術の進展により，ユーザ（消費者）は購入したアイテム（商品）に関する評価・感想をレビューサイトへ投稿する機会が増加している．それに伴い，豊富なユーザ嗜好に関する情報を入手可能になり，ユーザの嗜好に応じた推薦サービスが注目を浴びている．

本研究では，レビューサイトにおけるユーザとアイテムをノードとしたレビュー関係ネットワークから，特徴的なリンク傾向を抽出・分析することを試みる．特に，ユーザとアイテムの属性に着目し，どのような属性値（カテゴリ）を有するユーザがどのようなアイテムに高い評価をするかについて定量化する．すなわち，特徴的なリンク傾向とは，ユーザおよびアイテムのカテゴリ間に際立って，かつ，偏って存在するリンクのことを指す．単純な方法として，ある属性を持つユーザからどの属性のアイテムへのレビュー数により定量化し，数の多い属性ペアを特徴的なリンクとして抽出する手法が考えられる．しかし，レビュー数の少ないアイテムやユーザはランキング上位に浮上することは少なく，見落とされがちである．そこで，アイテムおよびユーザの属性により Mixing Matrix を構築し，その値が多項分布に従うとし，ランダム性を仮定したモデルによる期待値および標準偏差により Z スコアを計算し，定量化する．算出した Z スコアの値によりユーザ属性とアイテム属性のペアをランキングし，特徴的なリンク傾向を抽出・分析する．

2　提案手法

ユーザノード集合 $\mathcal{U} = \{1, \ldots, U\}$，アイテムノード集合 $\mathcal{I} = \{1, \ldots, I\}$，ユーザノードとアイテムノード間のリンク集合 $E \subset \mathcal{U} \times \mathcal{I}$ からなる 2 部グラフ $G = (\mathcal{U}, \mathcal{I}, E)$ として扱う．

1. リンク集合 E から Mixing Matrix c_{ij} を構築；
2. Mixing Matrix の要素が多項分布に従うと仮定し，Z スコア z_{ij} を計算；
3. z_{ij} を降順ソートし，上位のリンク (i, j) を抽出；

Analysis of Characteristic Link Tendency using Node Attributes
†Takayasu FUSHIMI　†Kazumi SAITO　††Kazuhiro KAZAMA
†Graduate School of Management and Information of Innovation, University of Shizuoka

2.1　Mixing Matrix 構築

Mixing Matrix は，あるカテゴリに属するノード群と，別のカテゴリーのノード群の間のリンク存在確率を要素とする行列である [1]．形式的には，ユーザ $u \in \mathcal{U}$ がカテゴリ $f(u) = i \in \{1, \ldots, K\}$ に属し，アイテム $v \in \mathcal{I}$ がカテゴリ $g(v) = j \in \{1, \ldots, H\}$ に属するとする．この時，$i \in \{1, \ldots, K\}, j \in \{1, \ldots, H\}$ に対し，$c_{ij} = |\{(u, v) \in E; u \in \mathcal{U}, v \in \mathcal{I}, f(u) = i, g(v) = j\}|/|E|$ により Mixing Matrix $\mathbf{C} = \{c_{ij}\}$ を計算する．

この値により，カテゴリ間の依存度などの関係の強さがわかる．また，行と列それぞれの周辺確率分布を $a_i = \sum_{j=1}^{H} c_{ij}, b_j = \sum_{i=1}^{K} c_{ij}$ とする．

2.2　Z スコア計算

Mixing Matrix の各要素の値 c_{ij} に対して Z スコアを考える．すなわち，周辺確率分布を固定して，ランダムに $|E|$ 本のリンクを生成するとき，カテゴリ i と j 間に張られるリンク数の期待値は $|E|e_{ij} = |E|a_i b_j$ となる．$K \times H$ の要素が多項分布に従い生成されると仮定し，ランダムなモデルと比較してリンクがどの程度有意に多くまたは少なく存在するかを表す Z スコアを期待値および標準偏差から計算する．

$$z_{ij} = \frac{|E|c_{ij} - |E|e_{ij}}{\sqrt{|E|e_{ij}(1 - e_{ij})}} \tag{1}$$

Z スコアが正で大きいほど，カテゴリ i のノードとカテゴリ j のノード間のリンク数は統計的有意に多く存在するといえる．逆に Z スコアが負で絶対値が大きいほど，有意に少ないといえる．Z スコアを用いることにより，出現頻度の少ないリンクであっても，ランダムな場合に比べて特徴的なリンクの場合は大きな値となり，規模の格差により隠れてしまうカテゴリ間の関係の強さを表現できる．

Z スコアの値 z_{ij} により，KH 個のカテゴリペアをランキングする．

3　評価実験

化粧品レビューサイトの@COSME のレビュー情報を用いて，提案手法を評価する．2009 年 12 月に収集したレビューデータから，レビュー数の多いアイテム 1,000 件を対象とした．用いるデータのユーザ数は $U =$

表 1: レビュー数ランキング

順位	$\|E\|c_{ij}$	i	j
1	1151	混合肌	ラッシュ
2	943	混合肌	ルナソル
3	851	乾燥肌	ラッシュ
4	744	混合肌	オルビス
5	732	混合肌	マジョリカ M

表 2: 全レビュー Z スコアランキング

順位	z_{ij}	i	j
1	6.38E+00	乾燥肌	ディオール
2	4.63E+00	アトピー	ルベル
3	4.41E+00	脂性肌	ケイト
4	4.08E+00	脂性肌	マジョリカ M
5	3.90E+00	アトピー	ザ・ボディショップ

表 3: ポジティブレビュー Z スコアランキング

順位	$z_{ij}^{(+)}$	i	j
1	5.23E+00	アトピー	ルベル
2	4.51E+00	乾燥肌	ディオール
3	3.93E+00	脂性肌	キス
4	3.68E+00	脂性肌	ビューティー N
5	3.68E+00	脂性肌	ＶＯ５

表 4: ネガティブレビュー Z スコアランキング

順位	$z_{ij}^{(-)}$	i	j
1	4.66E+00	アトピー	ネピア
2	4.50E+00	乾燥肌	ディオール
3	3.77E+00	普通肌	ルナソル
4	3.66E+00	アトピー	日本薬局方
5	3.46E+00	脂性肌	キャンメイク

10,403，アイテム数は $I = 1,000$，総レビュー数は $|E| = 93,015$ である．本稿では，ユーザカテゴリとして肌質（$K = 6$），アイテムカテゴリとしてブランド（$H = 332$）を用いる．各肌質属性ユーザのレビュー数分布は，混合肌：32,063，乾燥肌：22,654，敏感肌：16,116，普通肌：13,611，脂性肌：5,789，アトピー：2,782 である．

表 1 の単純なレビュー数 $|E|c_{ij}$ でランキングした結果を見ると，一般的にユーザからのレビュー数の多いブランドである「ラッシュ」や「ルナソル」,「オルビス」などが上位になっている．さらに，レビュー数分布の多い「混合肌」,「乾燥肌」が上位になっている．1 位の「混合肌–ラッシュ」のペアは，実際にレビュー数が多いことを確認したが,「ラッシュ」も「混合肌」も含まれるノード数が多いため，当然の結果であるといえる．

表 2 の Z スコア z_{ij} でランキングした結果を見ると,「乾燥肌–ディオール」のペアが 1 位にランクインしている.「ディオール」のアイテムにレビューしたユーザの肌質分布をみると，確かに「乾燥肌」ユーザが多くレビューしており，妥当な結果であると言える．2 位の「ルベル」のアイテムにレビューしたユーザの肌質分布を確認すると，レビュー数は全体の分布と同様に「混合肌」,「乾燥肌」が多く，それらに比べると「アトピー」は少ない．しかし Z スコアが高いということは，全体の分布と比較すると「アトピー」は有意に多く「ルベル」にレビューしていることが示唆される．このように単純なレビュー数は多くなくても，全体の分布に比べて有意に多くリンクが存在する特徴的なリンクが抽出可能である．

表 3，表 4 にポジティブ，ネガティブレビューに関する Z スコア $z_{ij}^{(+)}$，$z_{ij}^{(-)}$ でランキングした結果を示す．ユーザ i のレビューしたアイテム集合を $R_i \subset \mathcal{I}$，アイテムノード $j \in R_i$ へのレビュー評点を r_{ij} とすると，評点平均は $\mu_i = \frac{1}{|R_i|}\sum_{j \in R_i} r_{ij}$ と計算できる．本研究では，各ユーザノードに関して自身の評点平均 μ_i より高い（低い）レビューのことをポジティブ（ネガティブ）レビューと呼ぶ．

表 3 と表 4 を見ると,「乾燥肌–ディオール」がどちらも 2 位にランクインしている．実際にレビューを見てみると，年配のユーザが日常的に使用しており，高い評点を付けている．一方，ブランド名に過度な期待を持って購入した中年のユーザが，性能が悪くなくても，高い価格には見合わないとして悪い評点を付けているレビューが目立った．これらの事実を反映し，両方で上位にランクインし，特徴的なリンクとして抽出された．すなわち，意見が割れていることになり，評点の高低は肌質以外にあることが示唆される．

4 おわりに

本研究では，Z スコアを用いてユーザ属性とアイテム属性のペアをランキングし，特徴的なリンク傾向を抽出する手法を提案し，抽出結果を評価した．評価実験より，ある程度妥当な特徴的リンクを抽出できることを示した．今後は，本研究の結果を導入したユーザ評点行動のモデル化を目指していくつもりである．

謝辞 本研究は，科研費 (No.25・10411) の補助を受けた．

参考文献

[1] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, Vol. 67, No. 2, pp. 026126+, Feb 2003.