2M-5

相関ルールに基づく外れ値検出手法を用いたユーザレビュー情報の分析

高橋　毅[†]　　天笠　俊之[‡]　　北川　博之[‡]

筑波大学第三学群情報学類[†]　　筑波大学大学院システム情報工学研究科[‡]

1．はじめに

インターネットの普及に伴い，web 上では様々な情報が氾濫している．その中でもレビューサイトにおいては各個人が任意にコメントを書き込む事が出来るものが多く，そのコメントを全ての人が自由に閲覧することができる．これらの情報は内容の信憑性に注意を払わず公開される事が少なくないため，利用者は自身でその信頼性を判断する事が重要である．また，多数の意見を取り出しても，必ずしも信頼できるとは限らず，少数の意見こそが重要である場合もある．そこで本研究では，トランザクションデータに対する外れ値検出手法[1]をレビューデータに適用し，ユーザにとって有益であると思われるレビューの抽出を行う．

2．レビューデータ

本研究はレビューデータに対して外れ値検出手法が有効であるかどうかを調べるとともに，外れ値の中から特異なデータを抽出することを目的とする．今回は楽天技術研究所から提供されている「楽天トラベル」のレビューデータ[2]を利用した．データ内容は以下の通りである．

施設番号	施設毎に固有の数値
投稿本文	ユーザのコメント
投稿番号	投稿毎にユニークな数値
分類	投稿の分類を示す文字列
プラン ID	プラン ID の数値
プランタイトル	プランタイトルの文字列
部屋種類	「s」「s1」等の文字列
部屋名前	部屋の名前の文字列
施設回答本文	施設側からのコメント
施設名	ホテル名の文字列
ニックネーム	マスクされたユーザ名
目的	旅行目的を示す文字列
同伴者	同伴者を示す文字列
参考になった数	他のユーザに支持された回数
参考にならなかった数	他のユーザに不支持とされた回数
評価 1(立地)	0-5 の 6 段階評価
評価 2(部屋)	0-5 の 6 段階評価
評価 3(食事)	0-5 の 6 段階評価
評価 4(風呂)	0-5 の 6 段階評価
評価 5(サービス)	0-5 の 6 段階評価
評価 6(設備・アメニティ)	0-5 の 6 段階評価
評価 7(総合評価)	0-5 の 6 段階評価

このデータに対して次の処理を行った：1)施設名に対して Google Maps API から提供されている逆ジオコーディングを適用し，住所から都道府県名を抽出，2)部屋名前から「シングル」，「和室」等の単語を抽出，3)プラン付ならば「プラン有」そうでなければ「プラン無」を追加，4)投稿本文にセンチメント分析を適用し，「嬉しい⇔怒り」，「楽しい⇔悲しい」，「のどか⇔緊迫」の 3 次元それぞれに対する感情値を得る．値に応じて各次元に応じた感情を示す文字列に変換し，その文字列を追加，5)評価 1～7 の値を，値が 3 以下ならば「悪」，4 以上ならば「良」に変換．

以上の処理により得られるレコードをデータとした．

3．トランザクションデータに対する外れ値検出

本研究で用いる外れ値検出手法[1]はトランザクションデータベースを対象としている．

トランザクションデータベースとは，「アイテムの集合を一つのデータとしたトランザクションの集まり」である．具体例を以下に示す．

TID	アイテムの集合
001	{パン，ジャム，牛乳}
002	{ベーコン，コーン，ジャム，牛乳}
003	{ベーコン，玉子，ジャム，牛乳}
004	{ベーコン，パン，玉子，牛乳}

トランザクションデータベースを T，トランザクション t∈T とする．外れ値検出ではアイテムのサポート値を用いる．サポート値は T 中で，アイテム X を集合に含むトランザクションの割合を示し，sup(ベーコン)は 75%となる．また，「牛乳を含むときにベーコンも含む」という相関ルール R:{牛乳}→{ベーコン}を満たすトランザクションの割合は(牛乳とベーコンを含むトランザクションの数)/(牛乳を含むトランザクションの数)で表され，例では 75%となる．これを確信度という．このとき，TID001 のトランザクションは，高い確信度を持つ相関ルール R を満たさないため，その他トランザクションに比べ，稀であると考えられる．このことから，TID001 のトランザクションを外れ値として検出する．

成田らの手法[1]では，上記の外れ値検出を行うために外れ値度という指標を提案している．外れ値度は，

A User Review Analysis using Outlier-Detection over Association Rules
Tsuyoshi Takahashi[†](zeppeli@kde.cs.tsukuba.ac.jp)
Toshiyuki Amagasa[‡](amagasa@kde.cs.tsukuba.ac.jp)
Hiroyuki Kitagawa[‡](kitagawa@kde.cs.tsukuba.ac.jp)
†College of information Sciences,University of Tsukuba
‡Graduate School of Systems and Information Engineering, University of Tsukuba

表1：相関ルールの抽出結果

	ルール	Support	Confidence
1	{鳥取県米子市皆生新田}=> {鳥取県}	0.1755990	1.0000000
2	{鳥取県}=> {鳥取県米子市皆生新田}	0.1755990	0.9911032
3	{プラン有, 鳥取県米子市皆生新田}=> {鳥取県}	0.1434426	1.0000000
4	{分類:苦情}=> {プラン有}	0.09993695	0.82984293
5	{分類:苦情}=> {評価 5:悪}	0.09583859	0.79581152
6	{評価 2:良}=> {評価 7:良}	0.6320933	0.9273821
7	{プラン有}=> {評価 7:良}	0.6251576	0.7635734
8	{プラン無}=> {評価 7:良}	0.1270492	0.7020906
9	{嬉しい }=> {評価 7:良}	0.46595208	0.8147740
10	{怒り}=> {評価 7:悪}	0.14060530	0.3320923

相関ルール:{牛乳}→{ベーコン}において、{牛乳}を含んでいるトランザクションがどれだけ{ベーコン}を含まないかを表す尺度である．外れ値度は次のようにして得られる．前ページ表中のトランザクション001を元のトランザクション t とする．T 中には相関ルール{牛乳}→{ベーコン}において{牛乳}は含んでいるが{ベーコン}は含んでいない．これを違反ルールと呼ぶ．違反ルールが t 中に存在しないようにするため，{ベーコン}を t に追加し，新たなトランザクション t^+ を作る．t と t^+はそれぞれ下の表になる．

元のトランザクション t	{パン，ジャム，牛乳 }
新しく得たトランザクション t^+	{パン，ジャム，牛乳，ベーコン}

この時，t に新たに含めたアイテムの個数を，t^+が含むアイテムの個数で割った値が外れ値度となる．例では 25%となる．この外れ値度が一定の値を超えたとき，トランザクションを外れ値として検出する．

4．相関ルールの抽出

上記の外れ値検出をレビューデータに適用するための予備実験として，通常の相関ルールの抽出を適用した．データセットから 3,170 件をサンプルデータとして取り出し，apriori[4]アルゴリズムを適用した．実装には R[5]を用いた．その抽出結果の一部を表 1 に示す．

ルール 1，2 から鳥取県内での旅行先の施設として楽天トラベルが取り扱っている地域は「鳥取県米子市皆生新田」が多い事が分かる．また，ルール 3 からこの地域の施設に何らかのプランを必ず付けていると言える．

ルール 4 から，ユーザが「苦情」としているレビューの多くがプラン付きであることが分かる．これは「プラン有」が頻出なので，共起しやすくなっているためと考えられる．ルール 5，6 からは良い評価同士，悪い評価同士の結びつきが強いことが確認された．ルール 7，8 からプランの有無によってユーザの評価が変化することが分かる．ルール 9 では「嬉しい」と，良い評価との結びつきが強いことを示すが，ルール 10 では「怒り」と悪い評価との結びつきが弱いことを示している．これは「怒り」と「評価 7:悪」がどちらも頻出でないため，共起しにくい事が原因と考えられる．

5．まとめ

本論文では，レビューデータに対して相関ルールの外れ値検出手法を適用し，有用なレビューを発見する手法を提案した．また，予備実験として，レビューデータに対して通常の相関ルールを適用し，いくつかのルールが抽出できる事を示した．今後は，相関ルールの外れ値検出手法を適用し，有用なルールの抽出を試みる．また，他のデータセットに対しても同手法を適用する予定である．

謝辞

本研究に用いたレビューデータを提供して戴いた楽天技術研究所に深謝する．千葉工業大学熊本忠彦教授，京都産業大学河合由起子講師，張建偉氏には本研究のためにセンチメント分析システムをご提供戴いた．ここに深謝の意を表する．

本研究の一部は科学研究費補助金特定研究領域（#21013004）による．

参考資料

[1]成田和代，北川博之「トランザクションデータベースに対する高確信度の相関ルールを用いた外れ値検出手法」情報処理学会 2007

[2]楽天データ公開 http://rit.rakuten.co.jp/rdr/index.html

[3]Google Maps API http://code.google.com/intl/ja/apis/maps/documentation/javascript/v2/services.html#ReverseGeocoding

[4]Rakesh Agrawal, Ramakrishnan Srikant Fast Algorithms for Mining Association Rules in Large Databases. VLDB 1994.

[5]R http://www.r-project.org/