2G3-1

出現単語とメタな文章構造に基づく商品説明文のマイニング

Mining Sales Item Descriptions on the Internet Shopping Sites
Based on Term Occurrences and Sentence Structures

白井 康之[*1*3] 櫻井 祐子[*2*1] 鶴間 浩二[*1] 小山 聡[*3*1]
Yasuyuki SHIRAI Yuko SAKURAI Koji TSURUMA Satoshi OYAMA

[*1]独立行政法人科学技術振興機構 ERATO 湊離散構造処理系プロジェクト
JST ERATO Minato Discrete Structure Manipulation System Project

[*2]九州大学大学院システム情報科学研究院
Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University

[*3]北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

In this article, we discuss the characteristics of the shopping item descriptions that have lots of review comments, such as term occurrences and other structural features, based on the emerging pattern mining techniques. These results are expected to be used by the participants of the Internet shopping site who aim to get more number of visitors. We also discuss the recommendation functions to show the items or features to be added to the current description, using the Zero-suppressed Binary Decision Diagram (ZDD) as a future direction.

1. はじめに

インターネット上でのショッピングは，総務省による統計（ネットショッピングの状況）[Soumu] に見られるように，近年大きな伸びを示しており，旅行の予約といった趣味・娯楽分野から衣類や日常品といった一般的な消費財への拡大，また従来ネットショッピングの主たるユーザであった若年層から壮年層への拡大といったように，量だけではなく質的な拡大も見られている．

こうした状況において，インターネット上に出店する各店舗では，どのような商品説明あるいは商品アピールをしていけばより多くの消費者を取り込めるかが大きなカギとなっている．特に，ユーザは，検索によって必要とする商品を探していくと仮定すると，どのようなキーワードを商品説明として付与していけばより多くのユーザの閲覧を得ることができるのかといった点が重要である．

そこで，本研究では，楽天が公開している楽天データセット [Rakuten 10] を用いて，商品説明文にどのような特徴があればより多くの消費者の目に触れるか（商品レビュー数を増やすことができるか）といった点に着目して分析を行った．具体的には，レビュー数が多い説明文と少ない説明文をもとに，その差異を顕在パタン（Emerging Patterns）[Dong 99a, Dong 99b, Hamad 06, Morita 11] として抽出し，さらに，得られた顕在パタン集合から，より多くのレビューを得るための推薦を行うための方法論を議論する．

分析においては，主に 3 つの商品分野（メンズファッション，レディースファッション，日本酒・アルコール）を対象とした．

以下，2. では，楽天データセットの概要を述べ，3. で分析方法について，また，4. で分析結果を記す．

連絡先: 白井康之，(独) 科学技術振興機構 ERATO 湊離散構造処理系プロジェクト, 札幌市北区北 14 条西 9 丁目 北海道大学情報科学研究科, shirai@erato.ist.hokudai.ac.jp

表 1: 楽天市場公開データの例 [Rakuten 10]

Column	Sample
店舗コード	rakutenstore
商品 ID	12345678
商品名	【Rakuten T-Shirt】楽天ロゴ入り　着心地満点の T シャツ　体を締め付けない伸びる記事とデザイン
商品説明文	上質の素材を使用し，シルエットにも気を使ったデザインです．ストリートだけでなく，有名百貨店でも取り扱われるようになりました．人気急上昇の T シャツです．ベーシックなシルエットでありながら，年齢，性別を問わず楽しんでいただけるデザインになります．オンにもオフにも，チノにもジーンズにもコーディネイトに最適です．動きやすいだけでなく，着心地にもこだわりました．
商品 URL	http://www.rakuten.co.jp/rakutenshop/12345/98765
商品価格	4,500
ジャンル ID	403862

2. 楽天データセットの概要

楽天データセットは，株式会社楽天が保有する様々なデータを研究目的のために公開しているもので [Rakuten 10]，主に大学を中心とする研究機関が利用している．

昨年度の公開では，以下の 3 種類のデータが対象となっている．

- 楽天市場の全商品データ
- 楽天トラベルの施設データならびにレビューデータ
- 楽天 GORA（ゴルフ）の施設データならびにレビューデータ

本分析では，楽天市場の全商品データ（約 5000 万商品）を利用する．

3. 分析方法

3.1 分析データの作成

分析対象データは，表 1 で示したような商品説明文ならびに各商品につけられたレビュー件数である（レビュー結果そのものについては，本分析の対象外）．すなわち，どのような商品説明文を付与すると，ユーザの目に触れやすくなるかを検討することとし，ユーザの目に触れたか否かをレビュー件数によって代替するものとする．また，商品カテゴリ別に傾向が異なることが想定されるため（たとえば，楽天市場には，ファッション関係のほか，嗜好品，日常品，家具等極めて広範囲なものが存在している），商品シソーラスに従って，対象商品のデータのみを抽出した．今回の分析では，メンズファッション，レディースファッションならびに日本酒のカテゴリを対象とすることとし，別途用意された商品シソーラスにしたがって，データ抽出を行った．

なお，一般に商品説明文には，サイズに関する記述や購入手続きに関する情報など，さまざまな不要な文章が数多く含まれている．このため，商品を説明する，あるいはアピールするセンテンスのみを抜き出す必要がある．ここでは，テキストの構成や使用されているキーワードや言い回しを元にして，各センテンスに対して以下のようなタイプ分類を行った．

- 商品仕様情報 (サイズや価格など)
- アピールポイント
- 購入手続きに関する情報

例えば,「送料」や「分割配送」などのキーワードを含むセンテンスは，購入手続きに関する情報であり，商品サイズをあらわす数字を含むものは，商品仕様情報を表していると考えられる．一方，アピールポイントに属するセンテンスは例えば以下のようなものである．

- “荒々しさが魅力の麦 100 パーセント焼酎．原酒の味わいをぜひ “
- “厳選した材料を精米歩合 65%まで磨き上げ，独自の発酵技術でじっくりと熟成”
- “お祝いの席での一献一献に．ぜいたくな金箔の彩りとすっきりとした淡麗辛口”

本実験においては，アピール文のみを対象とすることとした．ファッション系では約半数のみがアピール文であったのに対し (サイズや扱い方に関する記述が多いため)，日本酒のカテゴリでは，80% がアピール文であった．特に，ファッション系の分析においてはノイズとなるために注意が必要である．

次に，各商品説明文に対して茶筅[*1] を用いた形態素解析によって品詞分解を行った．また，連続する名詞句に対しては，n-gram 連結 (最終的に $n = 3$) を行い，複合語を抽出した上で，得られたキーワード集合に対して，$tf \times idf$ を算出し，上位キーワードを選別した．抽出されたキーワードは,「レディースファッション」で 8275 語,「メンズファッション」で 15694 語,「日本酒」で 10190 語であった.「日本酒」カテゴリにおけるキーワード抽出結果の例を表 2 に示す．

またこれに加えて，商品説明文における特徴として，センテンス数，センテンスの長さ (平均)，アピール文の全体における比率の 3 つを算出し，特徴データとして付与した（一般に，レビューの多い商品説明においてはセンテンス全体の長さが長いことがあらかじめわかっていたため）．以上，特徴データとしては，出現する単語集合ならびにアピール文における特徴ををあわせて，分析用データを作成した．

なお，目的変数であるレビュー件数については，各分析データにおいてレビューのないデータが相対的に多くなっている (例えば，日本酒・焼酎カテゴリの例では，レビュー無しが 1,347,336 件に対して，レビュー有りが 354,421 件となっている)．訓練データでは，レビューの多いデータと少ないデータを分類するパタンを見つけるため，レビュー数のあるデータとレビューのないデータが一定の比率以上で混在している必要がある．このため，各カテゴリにおいて，レビューのあるデータは全件訓練データとして含め，ほぼそれに該当する量のレビューのないデータをランダムに選択した上で含めることとした．

表 2: 品詞抽出結果の例

単語	商品数	出現回数	tf・idf
米麹	26508	29853	0.1223
酸度	22861	25308	0.1136
蔵元	21151	29012	0.1091
熟成	21081	29103	0.1089
コク	20605	23956	0.1075
ロック	18078	20499	0.1000
産地	17913	18824	0.0995
醸造	17803	23239	0.0992
吟醸	16814	29698	0.0960
風味	16119	19296	0.0936
楽しみ	15702	17285	0.0922
旨み	15393	18477	0.0911
原酒	15161	22562	0.0903
逸品	14567	16294	0.0882
製造元	14378	15437	0.0875
口当たり	13535	14826	0.0843
辛口	13488	18413	0.0841
米酒	13349	23025	0.0836
甘み	13242	15956	0.0832
酒類	12718	13245	0.0811
杜氏	12601	20687	0.0807
伝統	11673	14245	0.0769
黒麹	11568	16589	0.0764

3.2 各クラスに顕著な特徴の抽出

上記で整備された訓練データは，説明変数として出現単語集合ならびに特徴データ，また目的変数に相当するものとしてレビュー件数をデータとして持っている．ここから，レビューの多い商品（以下ポジデータ）の特徴，逆にレビューの少ない商品（以下ネガデータ）の特徴を抽出することが目的である．

ここで，一般的な頻出パタンマイニングを用いれば，頻度の高い項目の組み合わせを発見することは容易である．しかしながら，ポジデータ，ネガデータに共通するパタンもまた多く見つけてしまい，膨大な数に及ぶ頻出パタンの中に，本当に意味のあるパタンが埋もれてしまうことが多い．

本実験では，ポジデータ，ネガデータにそれぞれ特徴的に出現するパタンを見つける必要があるため，一般的な頻出パタンマイニングではなく，顕在パタンマイニングを行うこととした．顕在パタンマイニングは，1999 年に G. Dong らによって提唱されたもので，近年，データマイニングの分野では，実用的なパタンを抽出するための技術として広く活用されている [Dong 99a, Dong 99b, Morita 11, Hamad 06]．顕在パタンマイニングの概要は以下の通りである．

∗1 http://chasen.naist.jp

表 3: 抽出されたクラス別の特徴 (日本酒カテゴリから抜粋)

class	Support	GR	Pattern			
neg	0.031	6.878	本格焼酎 産地 種別			センテンス少
neg	0.037	3.842	製造元 産地		センテンス短	センテンス少
neg	0.033	3.436	アミノ酸	アピール文少	センテンス短	
neg	0.031	2.995	米 酸度			センテンス少
neg	0.034	2.656	味わい 酸度			センテンス少
neg	0.037	2.278	コク		センテンス短	センテンス少
neg	0.048	2.084	吟醸			センテンス少
neg	0.031	2.068	甘味	アピール文多		センテンス少
pos	0.032	8.029	旨 米吟醸			センテンス多
pos	0.034	8.010	贈答 旨			センテンス多
pos	0.035	7.964	旨 地酒			センテンス多
pos	0.032	7.552	冷やし 旨			センテンス多
pos	0.030	7.427	米麹 注文			センテンス多
pos	0.030	7.170	味わい プレゼント 父			
pos	0.031	7.079	米吟醸 地酒	アピール文少		
pos	0.031	6.922	味わい 母 父 お歳暮 内祝い			
pos	0.034	6.452	米 吟醸酒			センテンス多
pos	0.034	6.278	米 贈答			センテンス多
pos	0.036	6.139	結婚祝い	アピール文少		
pos	0.033	6.098	記念 誕生 お歳暮	アピール文少	センテンス長	
pos	0.033	5.899	父 年賀 結婚祝い	アピール文少		
pos	0.035	5.430	プレゼント 父 記念			
pos	0.033	5.425	味わい 常温			センテンス多
pos	0.031	5.309	記念 お歳暮			センテンス多
pos	0.035	5.290	父 還暦祝			センテンス多
pos	0.033	5.128	プレゼント お中元 見舞い			
pos	0.031	5.056	地酒 淡			
pos	0.032	4.609	味わい 手			センテンス多
pos	0.038	4.501	おすすめ ロック		センテンス長	センテンス多
pos	0.037	4.424	年賀 結婚祝い 内祝い		センテンス長	
pos	0.033	4.382	味わい 淡			センテンス多
pos	0.032	4.336	米 冷やし 香り			
pos	0.036	3.933	プレゼント 誕生 お歳暮 年賀			
pos	0.047	3.912	記念 お歳暮 御歳暮			
pos	0.035	3.889	香り 余韻			センテンス多
pos	0.052	3.789	記念 誕生			
pos	0.034	3.782	プレゼント お歳暮 年賀		センテンス長	
pos	0.032	3.738	水割り こだわり			
pos	0.041	2.989	ギフト	アピール文多		センテンス多

【顕在パタンマイニングアルゴリズムの概要】[Dong 99a, Dong 99b]
与えられたデータセット（アイテム集合ならびに各レコードに対応するクラス）から，各クラスに特徴的に出現するパタンを抽出する．今，パタン π のクラス A における Growth Rate (GR) を以下のように定義する．ここで，$SP(\pi, A)$ は，クラス A におけるパタン π のサポート値（比率），また，$SP(\pi, A^c)$ は，A 以外のクラスにおけるパタン π のサポート値を表わすとする．

$$GR_A(\pi) = \frac{SP(\pi, A)}{SP(\pi, A^c)}$$

顕在パタンは，上記の指標 (Growth Rate) にしたがって抽出される（指標が大きいほど顕在 (emerging) である）．たとえば，クラスが 2 つの要素から構成される場合には，Growth Rate が 0.5 を超えるパタンは，そのクラスにおいて顕在パタンであるということができる．

IGVcaep[*2][Morita 11] は，上記のような顕在パタンの抽出ならびにこれに基づく分類機能を頻出パタンマイニングアルゴリズム LCM[Uno 03] を用いて高速化したものである．本実験では，IGVcaep を利用してポジデータ，ネガデータに特徴的に出現するパタン抽出を行った．

4. 分析結果

以下，分析結果を記す．

4.1 特徴抽出結果

日本酒カテゴリにおいて抽出された顕在パタンを表 3 に示す．ここで，`Support` は各クラスに対するサポート値，また，`GR` は上記で定義した Growth Rate を表わす．

∗2 http://kgmod.jp/mcmd/index.php?igvcaep.rb

一般的な頻出パタンマイニング機能では，Growth Rate によらず頻出なパタンを抽出するため，全般的に識別にあたっては意味のないパタンが抽出されがちである．これに対して，表 3 に示されたパタンは，ポジデータ，ネガデータの特徴をよく表わしているといえる．

たとえば，ポジデータでは,「プレゼント」,「贈答」,「記念日」といったイベントに関連するキーワードが多数出現していることがわかる．一方，ネガデータには,「コク」や「吟醸」といったいわゆる一般的に日本酒のアピールに適していると思われるキーワードが多い．この差異は，ネットショッピングを行うユーザのニーズを考えれば，十分に納得のできるものであろう．すなわち，自分自身が日常的に嗜好するためのものを購入するのではなく，結婚式や記念日等のイベントやプレゼントとしてネットショップを利用することが多いと想定される．したがって，商品説明文を作成する立場からいえば，単に販売している商品の特徴をうたうだけでなく，プレゼントや贈答品として適当であるということをアピールすべきである．

また，文構造の特徴に関していえば，ネガデータに関しては，一般に個々のアピール文のセンテンスが短く，かつ数も少なくなっているが，アピール文の比率に関しては，必ずしも多い/少ないことによる優劣は言い難い結果となっている．記述内容により，しっかり説明をした方がよいのか，あるいは完結に記載するのみの方がよいのかが変わってくるものと思われるが，この部分はさらなる詳細な解析も必要であろう．

一方，ファッションに関しては，ブランドや素材に関して抽象的なアピールのみで具体的な言及がなく，かつ長い文章で語られているものや，素材やサイズなどの客観的な記述のみになってるものはコメントも少ないという結果であった．一方，通販サイトで購入した際に多くの人が気がかりになるであろうほつれや継ぎ目，品質，素材，あるいは製造工程に関して具体的な記載があるものはコメントも多く得られている．以上のように，一般的なリアルのショップと比較して，通販サイトならではの特徴が見て取れる結果となっている．

5. 情報推薦機能の検討

本研究の最終目標は，多くのレビューを集められるような（すなわち多くの閲覧を得られるような）商品説明文を作成するための情報推薦機能を提供することにある．一般に，ソーシャルレコメンデーションの場合には，過去の蓄積データに基づき，類似したアイテムの推薦を行っている．しかしこれらの推薦されたアイテムを含めることによって，どのような利得が得られるかは明らかではない．たとえば，商品の推薦で，ある商品が推薦されたとすると，確かにそれら商品が同時に購入されることが多いとはいえ，実際に購入することで利得が得られるかどうかは別である．

一方，本研究で検討を進めている情報推薦は，実際にそれらを追加することで，閲覧自体が増えることが想定されるものでなければならない．実際，本分析データでは 8 割以上の商品がレビュー 0 となっており，単に数が多いパタンではなく，むしろ多くレビューを集めるサイトとの差分が提示されるべきである．つまり，与えられたパタン集合に対して，どのような情報を付与するとリターンの期待値が増大するかを考えなくてはならない．こうした評価結果に基づき差分を提示するような推薦技法は，本研究で対象としているような商品説明文のみならず，評価結果を付随するデータを対象とした一般的なソーシャルレコメンデーションすべてに適用可能である．

このような推薦機能は，概念的には，現在の特徴に対して，

それに最も近くかつ高い評価が得られる可能性があるものへの移行を推薦する形で実現できる．例えば，表 3 の結果からは，「味わい」,「プレゼント」が含まれるアピール文に対して，さらに具体的に「父」「母」あるいは「還暦祝」「記念」といったキーワードを付与することで，より多くのレビューを集めることが期待できる．あるいは,「米」「酸度」が含まれるアピール文に対しては，検索はされにくいであろう「酸度」という言葉よりは,「吟醸」や「地酒」といったキーワードの方が適切であろうと想定される．

本研究では，こうした推薦機能を Zero-Suppressed Binary Decision Diagram（以下 ZDD と略す）[Minato 93, Minato 06] を用いて実現することを検討している．ZDD は，疎なデータ構造に対して効率的なデータ表現方法を提供するとともに，効率的な演算実装が可能な二分決定木（Binary Decision Diagram）である．頻出パタンでは，一般に数万あるいは数十万のオーダでパタンが存在するが，これらの多くは類似した部分構造を保持しているため，フラットなデータベース構造やツリー構造で保持するのと比較して，相対的な優位性がある（詳細は上記文献参照のこと）．

具体的な実行例は以下の通りである．今，共起アイテムを積で表すとすると，ポジデータ，ネガデータの集合は，共起アイテムの和集合としてあらわすことができる．なお，以下で，`vsop>` は ZDD の処理系である VSOP[Minato 06] のプロンプトを表す．また，%以下はコメントを表す．

```
vsop> P = 2 a b c + a b d + b c d
vsop> N = b f + a b f
vsop> print (P-N).Restrict(a b) > 0
 2 a b c + a b d - a b f
% アイテム a b を含み，ポジデータのみに含まれる組み合わせ

vsop> print (P-N).Restrict(a+b) > 0
 2 a b c + a b d - a b f + b c d - b f
% アイテム a または b を含み，ポジデータのみに含まれる組み合わせ

vsop> print (P-N).Restrict(a+b) < 0
 a b f + b f
% アイテム a または b を含み，ネガデータのみに含まれる組み合わせ

vsop> print P/(a b)%d
 2 c
% アイテム a, b をともに含み，d を含まないポジデータの組み合わせ
```

上記のような演算処理により，すでに述べた各カテゴリにおけるポジデータとネガデータの差分の詳細把握や，具体的なアイテム追加に関する情報推薦を行うことが可能である．ただし，実際の情報推薦においては，アイテムを削除あるいは追加することに伴う制約，コストも発生すると思われる．たとえば,「プレゼント」や「お祝い」といったキーワードは追加すること自体は問題がないが，販売の仕組みとして対応可能かどうかといった問題も存在する．また,「米麹」などの材料に関する文言は，事実に反するものを含めることはできない．

6. まとめと今後の課題

以上，本論文では，ショッピングサイトにおける商品推薦文に対して，レビューデータ数を目的変数とした顕在パタンのマイニング結果について述べた．レビューデータを増やすことは，いわば検索機能によるヒットしやすさを表わしているともいえ，結果的にその商品がどのような評価を得るにしてもネットショッピングの出店者にとっては本質的に重要である．

得られた顕在パタン集合に基づき，どのようなアイテムあるいは特徴を追加することにより，さらに人目につきやすい説明文を作成することができるか，実用上価値のある推薦機能を実現していくことが今後の課題である．

謝辞

本分析においては，楽天株式会社から研究目的のために公開されたデータを利用している [Rakuten 10]．また，大阪府立大学森田裕之教授，関西学院大学羽室行信准教授には顕在パターンマイニング抽出プログラムのご提供ならびにご助言を頂いた．ここに記して感謝する．

参考文献

[Dong 99a] G. Dong , X. Zhang , L. Wong and J. Li : CAEP: Classification by Aggregating Emerging Patterns, Proc. of Second International Conference on Discovery Science, LNCS Vol. 1721, 1999

[Dong 99b] G. Dong, J. Li : Efficient Mining of Emerging Patterns : Discovering Trends and Differences, Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999

[Hamad 06] H. Alhammady and K. Ramamohanarao : Using Emerging Patterns to Construct Weighted Decision Trees, IEEE Transactions on Knowledge and Data Engineering, Vol.18, 2006

[Rakuten 10] `http://rit.rakuten.co.jp/rdr/index.html` (楽天データ公開)

[Soumu] `http://www.stat.go.jp/data/topics/topi33.htm` (ネットショッピングの状況/総務省統計局)

[Minato 93] S. Minato : Zero-Suppressed BDDs for Set Manipulation in Combinatorial Problems, In Proc. of 30th ACM/IEEE Design Automation Conference (DAC'93), 1993.

[Minato 06] S. Minato : VSOP (Valued-Sum-of-Products) Calculator for Knowledge Processing Based on Zero-Suppressed BDDs, In K. P. Jantke, et al. editors, Federation over the Web, LNAI 3847, 2006.

[Morita 11] H. Morita, Y. Hamuro : A classification model using emerging patterns incorporating item taxonomy, International Conference on Data Engineering and Internet Technology, 2011

[Uno 03] T. Uno, T. Asai, Y. Uchida, and H. Arimura: LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets, Proc. of Workshop on Frequent Itemset Mining Implementations (FIMI03), 2003.