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

白井 康之 [1]　櫻井祐子 [2]　鶴間浩二 [1]　小山聡 [3]

[1] JST ERATO 湊離散構造処理系プロジェクト

[2] 九州大学大学院システム情報科学研究院

[3] 北海道大学大学院情報科学研究科

2011 年 6 月

はじめに

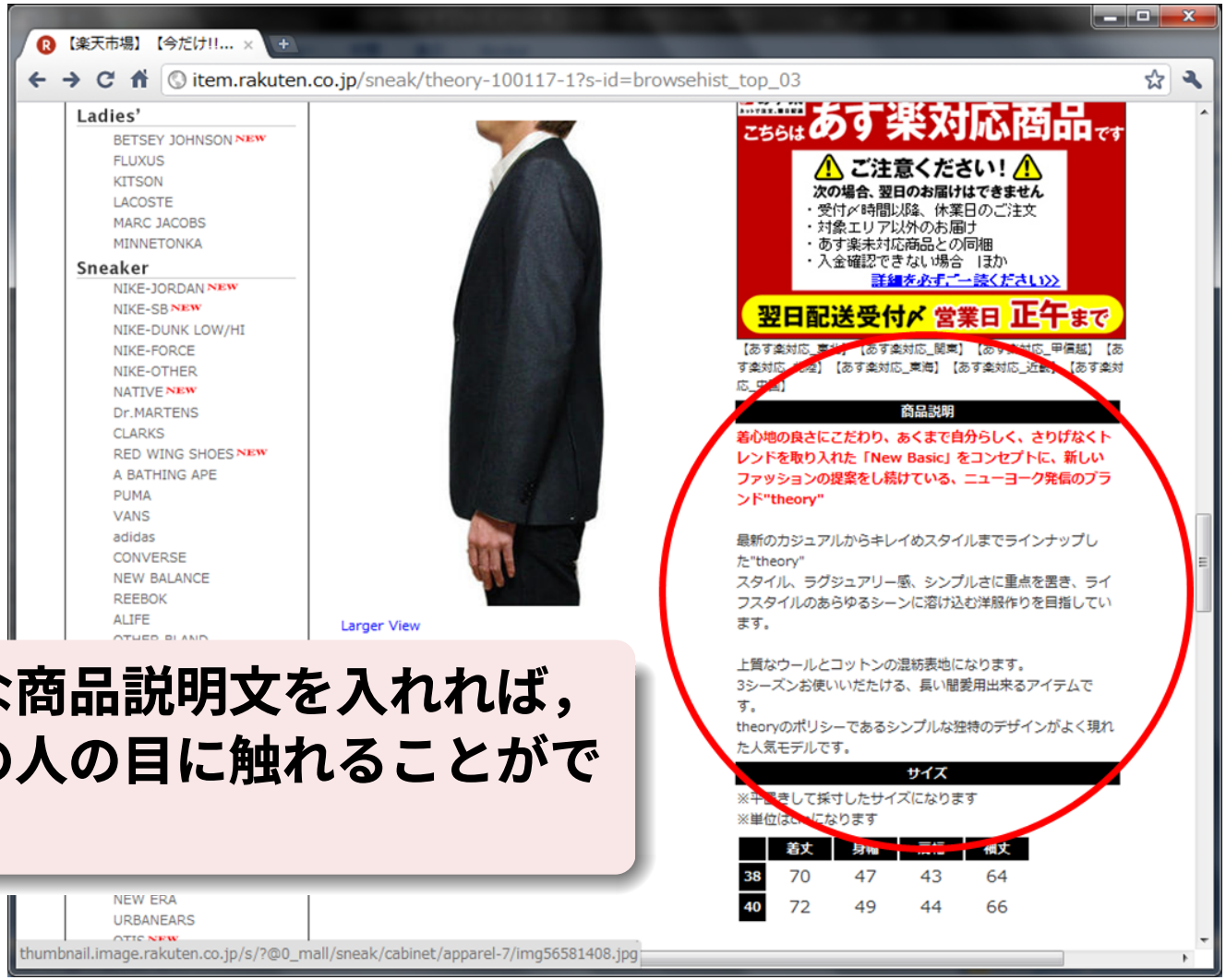
背景・目的

- インターネットショッピングの量的・質的拡大
- どのような商品をどのように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

実施概要

- 楽天データセット (http://rit.rakuten.co.jp/rdr/index.html) を使ったマイニング実験.
- 商品説明文の特徴と商品レビュー数から，レビュー数の差異の要因を顕在パタンとして抽出.
- より多くのレビューを得るためのリコメンデーション機能の検討.
- 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

商品説明文の例（画面）



どのような商品説明文を入れれば，より多くの人の目に触れることができるか？

http://item.rakuten.co.jp

楽天説明文の例（公開データ）

Table: **楽天市場公開データの例**

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．**
商品価格	4,500
ジャンル ID	403862

分析方法：分析データの作成

- 商品説明文から，アピール文のみを抜き出す．(その他のサイズ記述や手続き的な記載は除外)
- アピール文に属するセンテンスの例
  - “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”
- ファッション系では約半数，日本酒カテゴリでは，80% がアピール文．
- 分析対象データは，**商品説明文ならびに各商品につけられたレビュー件数**とする．

分析方法：特徴抽出 (単語抽出)

- 各商品説明文（アピール文）に対して茶筅を用いて形態素解析を行い，名詞，形容詞，動詞，副詞を抽出．
- 連続する名詞句は n-gram 連結 (最終的に $n = 3$) し，連語を生成．
  ex) 本醸造酒, 健康食品, 独自製法, 柔らか素材, 特殊樹脂, メイドインジャパン, ...
- $tf \times idf$ から上位キーワードを選別．
- 「レディースファッション」で 8275 語,「メンズファッション」で 10190 語,「日本酒」で 15694 語．

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
醸造	17803	23239	0.0992
吟醸	16814	29698	0.0960
風味	16119	19296	0.0936

単語	商品数	出現回数	tf・idf
楽しみ	15702	17285	0.0922
旨み	15393	18477	0.0911
逸品	14567	16294	0.0882
口当たり	13535	14826	0.0843
辛口	13488	18413	0.0841
甘み	13242	15956	0.0832
杜氏	12601	20687	0.0807
伝統	11673	14245	0.0769
黒麹	11568	16589	0.0764

分析方法：特徴抽出と分析データ作成

単語出現以外の特徴抽出

- 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

分析データの作成

- レビューありデータ，レビューなしデータをほぼ同数含め，訓練データを作成．
- 目的変数は，レビューの多さ（ある (2 件以上)・ないで分類）．
- 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．

分析方法：各クラスに顕著な特徴の抽出

一般的な頻出パタンマイニングでは，頻度の高い項目の組み合わせを発見することは容易だが，ポジデータ，ネガデータに共通するパタンを多く見つけてしまうことが多い．

頻度	キーワード
3507	味わい
3425	香り
2380	米
1795	香り 味わい
1602	米麹
1601	産地
1360	ロック
1287	特徴
1276	熟成
1261	口

頻度	キーワード
1254	度数
1239	蔵元
1217	酸度
1202	コク
1116	楽しみ
1031	仕込み
994	風味
989	味わい 米
939	醸造
933	量

頻度	キーワード
905	感じ
879	原酒
875	蔵
858	こだわり
855	旨み
851	産地 味わい
842	本格焼酎
834	香り 米
818	ロック 味わい
815	旨味

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** [Dong 99a, Dong 99b, Morita 11, Hamad 06] (Emerging Pattern Mining) を行うこととした．

分析方法：顕在パタンマイニング [Dong 99a, Dong 99b]

- 各クラスに特徴的に出現するパタンの抽出
- パタン π のクラス A における Growth Rate $GR_A(\pi)$ は以下のように定義される．

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

  - $SP(\pi, A)$ は，クラス A におけるパタン π のサポート値（比率）
  - $SP(\pi, A^c)$ は，A 以外のクラスにおけるパタン π のサポート値

- 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 1 を超えるパタンは，そのクラスにおいて顕在パタンといえる．

分析結果：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 (一部抜粋)

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
pos	0.034	8.010	贈答 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	0.034	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	0.031	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	0.037	4.424	年賀 結婚祝い 内祝い	センテンス長

分析結果に関する考察

- 日本酒カテゴリ
  - 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.
  - ネガデータでは,「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ユーザのニーズを考えれば，十分に納得できる.
- ファッション（メンズ，レディース）
  - ファッションに関しては，抽象的なアピールのみで具体的な言及がないものはレビューも少ない.
  - 一方，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い.
- 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.

分析結果（参考）：分類精度

- s をあるインスタンス，集合 $E(\mathcal{C})$ をクラス $\mathcal{C}$ の訓練集合に含まれる顕在パタン集合としたとき，s のクラス $\mathcal{C}$ に対するスコア $score(s, \mathcal{C})$ を以下のように定める [Dong 99a, Dong 99b]．

$$
\begin{aligned}
score(s, \mathcal{C}) &= \sum_{\pi \subseteq s, \pi \in E(\mathcal{C})} \frac{GR_{\mathcal{C}}(\pi)}{GR_{\mathcal{C}}(\pi) + 1} \times SP(\pi, \mathcal{C}) \\
&= \sum_{\pi \subseteq s, \pi \in E(\mathcal{C})} \frac{SP(\pi, \mathcal{C})}{SP(\pi, \mathcal{C}) + SP(\pi, \mathcal{C}^c)} \times SP(\pi, \mathcal{C})
\end{aligned}
$$

- s は，上記スコアを最大化するクラス $\mathcal{C}$ に属するものとする．

分析結果（参考）：分類精度

日本酒カテゴリ

メタ情報なし (accuracy = 0.56)

	classified as		
	neg	pos	total
neg	715	2405	3120
pos	316	2795	3111
total	1031	5200	6231

→

メタ情報あり（accuracy = 0.61）

	classified as		
	neg	pos	total
neg	1992	1128	3120
pos	1290	1821	3111
total	3282	2949	6231

レディースファッション

メタ情報なし (accuracy = 0.56)

	classified as		
	neg	pos	total
neg	1626	1465	3091
pos	1233	1872	3105
total	2859	3337	6196

→

メタ情報あり（accuracy = 0.60）

	classified as		
	neg	pos	total
neg	1941	1175	3116
pos	1312	1799	3111
total	3253	2974	6227

メタ情報を入れることにより，ネガデータの識別力が増す結果．

情報推薦機能の検討

- 与えられたネガデータに対して，ポジになるためのアイテムの追加，削除を（出店者に対して）推薦する問題．

クラス	顕在パタン
P	{ ○, △, □ }
P	…
N	{ ○, △, ■ }
N	…

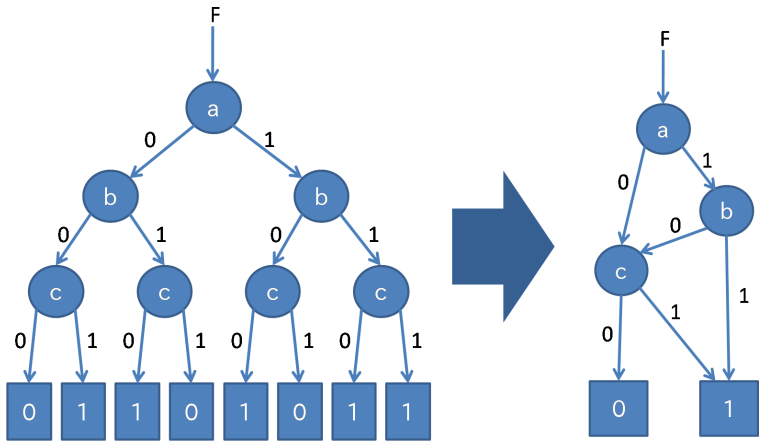
{ ○, △, ■ } のデータは N に分類される．
■→□とすると，
{ ○, △, □ } のデータは P に分類される．

- 例えば：
  - 「味わい」「プレゼント」→ 「父」「母」「還暦祝」「記念」などを付加
  - 「米」「酸度」→ 「吟醸」「地酒」に置き換え
- $\pi \in N, \pi \notin P$ なるパタン π に対する推薦 σ：
  - $\pi' = \pi\sigma,\ \pi' \in P, \pi' \notin N$
  - $cost(\sigma)$ を最小にする（あるいは閾値以下にする）置換/追加/削除 σ．
- 評価結果を付随する一般的な協調フィルタリングに適用可能．

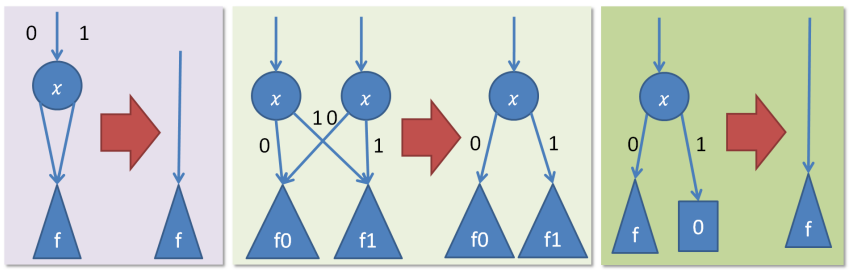
ZDD 処理系（VSOP）

BDD（Binary Decision Diagram），ZDD (Zero Suppressed BDD)

- 論理関数を効率的に表現・計算するための手法．
- 類似した共通部分集合の表現，スパースな構造の表現において圧縮効果が高い．
- 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．



二分決定木による $F = (a \wedge b) \vee c$ の表現（左）と BDD による表現（右）



BDD/ZDD による二分決定木の圧縮

情報推薦の実行例

- ZDD の処理系である VSOP（Valued Sum Of Products）[Minato 06].
- 共起アイテムの集合を積和形で表現する.
- ポジパタン，ネガパタンの集合をそれぞれ P，N とする.

```
vsop> P = a b c + a b d + b c d
vsop> N = b f + a b f
vsop> print (P-N).Restrict(a b) > 0
 a b c + a b d
```

% アイテム a, b をともに含み，ポジデータに含まれる組み合わせ

```
vsop> print (P-N).Restrict(a+b) > 0
 a b c + a b d + b c d
```

% アイテム a または b を含み，ポジデータ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

情報推薦の実行例

実行例

- { 味わい, 米, ストレート } を含むネガデータ.
  全部を含むポジデータの頻出パタンはなし.
- { 味わい, 米 }$+\alpha$ でポジデータになるものを探す.
  （例）$\alpha =$ { 香り }, $\alpha =$ { 杜氏 }, ...
- { 味わい, ストレート }, { 米, ストレート } を含むポジデータはないので，上記以外では $cost(\sigma) = 2$ の推薦は不能.
- より感覚に訴えるキーワードを追加，あるいは製造過程に言及するような記述を追加，等.

情報推薦機能の課題

- アイテムの性質や関連を考慮しない推薦は，ポジデータとネガデータの差分評価により実現可能．
- 一方，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．
  - たとえば，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？
    ex. 「プレゼント」や「お祝い」
  - また，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．
    ex. 「米麹」
- ただし，100%完全な推薦である必要はない．
- 否定をどのように取り込むか．本質的かつ意味のあるリテラルのみ否定を扱いたい．

まとめ・今後の課題

- ショッピングサイトにおける商品推薦文に対する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を示した．
- 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
- また，顕在パタンに基づく推薦技術の可能性を示した．
- 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

参考文献


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


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


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


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


`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**)


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


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.


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


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.