楽天データの解析について

白井康之(*1), 小山聡(*2), 櫻井祐子(*1), 川原純(*1), 鶴間浩二(*1)

(*1) JST ERATO 湊離散構造処理系プロジェクト
(*2) 北海道大学 大学院情報科学研究科

楽天データ公開

【2010年 楽天データ公開】

- 第3回楽天研究開発シンポジウムの一環として, 今年度より試行.
- 学術研究機関を対象にデータを公開
- 25件程度の申し込み
  (うち, 実際に分析を行っているのは10件程度か?)

【公開データ】
- 楽天市場の全商品データ
  (5000万商品)
- 楽天トラベルの施設データ
  (11,468施設)、
  レビューデータ(35万レビュー)
- 楽天GORA(ゴルフ)の施設データ
  (1,669施設)、
  レビューデータ(32万レビュー)

http://rit.rakuten.co.jp/rdr/index.html

楽天データ公開（商品データ）

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生地とデザイン
商品説明文	上質の素材を使用し、シルエットにも本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、デノのみジーンズにもコーディネイトに最適です。動きやすいだけでもなく、着心地にもこだわりました。
商品 URL	http://www.rakuten.co.jp/rakutenshop/12345/987654/
商品画像	http://thumbnail.image.rakuten.co.jp/@_0mall/rakutenshop/cabinet/tshirt_1234_56_78/img1234567890.jpg?_ex=64x64
商品価格	4,500
ジャンル ID	403862

http://rit.rakuten.co.jp/rdr/index.html

商品数 5000万，店舗数 7万

楽天データ公開（トラベルデータ）

Column	Sample
施設 ID	121212
ユーザー投稿本文	値段のわりには、きれいなホテルでした。以前泊まったときより、きれいに改装されたようで、入ったときにはびっくりしました。しかし、以前と同じくフロントでは丁寧な対応をして頂き、気分よく泊まることができました。接客には十分満足しましたが、駐車場が少し遠かったり、コンビニが近くになかったりしたところが少し残念です。朝食はバイキングでしたが、品数も多く、おいしかったですが、レストランスタッフの方を呼んでもなかなか出てこられなかったのは残念でした。
投稿番号	99887766
分類	感情・情報
プラン ID	242424
プランタイトル	新館、禁煙、朝食バイキング！平日お得プラン！
部屋種類	2
部屋名前	新館・禁煙・セミダブルルーム
施設回答本文	この度は、ご利用頂きまして誠にありがとうございます。また、貴重なご意見をお寄せ頂きありがとうございました。今年初めて改装しましたが、お褒め頂きましてありがとうございます。レストランスタッフの対応に関しましては、大変申し訳ございません。従業員の指導を徹底して参ります。またのご利用をお待ち申しあげております。

Column	Sample
投稿番号	99887766
ニックネーム ID	12345
目的	レジャー
同伴者	家族
評価1（立地）	4
評価2（部屋）	4
評価3（食事）	0
評価4（風呂）	4
評価5（サービス）	3
評価6（設備）	3
評価7（総合）	4

Column	Sample
施設 ID	121212
施設名	品川楽天タワーホテル

http://rit.rakuten.co.jp/rdr/index.html

施設数 1万，
レビュー数 35万

楽天データ公開（ゴルフ場データ）

楽天GORA:コースデータ

Column	Sample
コース ID	9876
コースナンバー	1
ゴルフ場 ID	1234
コース名	OUT
パー 1	4
パー 2	5
...	
全ホールのパー数の合計	35
ハンディーキャップ 1	7
ハンディーキャップ 2	15
...	
ホール説明 1	左にゆるやかにドッグレッグしたショートなホールです。左にある池に注意。
ホール説明 2	フェアウェイが複雑なアンジュレーション。右に傾斜しているので左狙い。
...	

http://rit.rakuten.co.jp/rdr/index.html

楽天GORA:ゴルフ場データ

Column	Sample
ゴルフ場 ID	1234
ゴルフ場名	マンハッタンゴルフコース
郵便番号 1	201
郵便番号 2	0035
住所	埼玉県東天市
開場年月日	1966/4/1
施設コメント	河川沿いのフラットなコースで、インターバルも短く、回りやすいコースになっています。河沿いの緩やかな形式のなかで、プレーを楽しめます。ビギナーにもオススメのコースです。初一人様や当日の受付も行っております。
ゴルフ場コメント	毎月 1 日は女性限定のビギナーズデータを開催しております。詳しくはHPをご覧ください。※毎月第 1 月曜日は定休日です。
シューズ指定	2
服装指定	Tシャツは不可です。
休場日	1/1、6/26,27 の 3 日間
ホール数	18
パー数	72
コースレート	70 9
ヤード	6379

Column	Sample
クチコミ ID	535770
ゴルフ場 ID	1212
投稿者ニックネーム	user0001
総合評価	3
コストパフォーマンス	3
スタッフ接客	4
コース/戦略性	4
食事が美味しい	5
設備が充実	2
...	
タイトル	楽しくプレーできました
コメント	きれいに手入れされたグリーンで、気持ち良かったのですが、初心者の私には少し難しかったです。練習してまたチャレンジしたいです。食事は、期待も多く
プレー日	2010/1/1

楽天GORA:クチコミデータ

楽天市場の現状のサービス

- 購入履歴の管理
- 閲覧履歴の管理
- 上記2つをもとにした推薦商品の提示
- 全文検索機能
  (and/or/not検索)
- ジャンル・価格絞り込み機能

楽天トラベルの現状のサービス

- 宿泊履歴の管理
- おすすめ宿泊施設
  （方法は不明．あまり精度は良くない？）
- 日時・場所指定検索
- キーワード検索機能

メンバー

白井康之	JST ERATO（札幌）	分析（1）
櫻井祐子	JST ERATO（札幌）	分析（2）
小山聡	北大情報科学研究科	分析（3）
川原純	JST ERATO（札幌）	分析（4）？
田部井靖生	JST ERATO（東京）	画像解析（分析1）
鶴間浩二	JST ERATO（札幌）	テキスト解析等

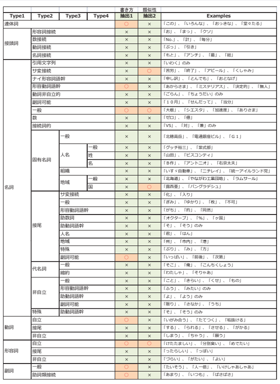
分析の概要

1. [楽天市場] 関連商品の提示・説明文ガイダンス機能（白井ほか）
   ◆商品説明から商品の特徴付け
   ➢ユーザ向け推薦，店舗向けガイダンス作成支援
2. [楽天トラベル] プランタイトル作成支援システムの構築（櫻井ほか）
   ◆評価に関連するキーワードの解析
   ➢店舗向けのプランタイトル作成支援
3. [楽天トラベル] 多面的数値・テキスト評価データに基づく推薦システム（小山ほか）
   ◆「立地」「設備」「食事」等の多面的な評価値を考慮
   ➢ユーザへの宿泊施設推薦精度の向上

1. 楽天市場データの分析　分析の対象

1. 楽天市場データの分析　（1/2）

- 商品データをもとに分析（購入履歴は非公開）
- 購入者の視点
  - 類似商品の提示（購入候補）
  - 本当は購入者の履歴を基にした推薦が良いが．
- 店舗の視点
  - 類似商品の提示（ライバル商品）
  - 商品説明文へのガイダンス
- ジャンルごとにユーザニーズ，店舗ニーズは異なる！
- とりあえずメンズファッション（紳士服）を対象に合計120万レコード（120万商品）を抽出
- 商品説明文から形態素解析により単語抽出
- 連続する名詞については 一語として抽出（名詞句n-gram連結）
- 変形TF-IDF（出現店舗数を考慮）により，4000単語程度を抽出．商品を特徴づけるベクトルとして利用
- 一商品当たり数個の特徴キーワードを持つ．
- 特徴キーワードをもとにして，Term-Document Matrix を作成



接続詞，助詞，助動詞，感動詞，記号，フィラー，その他（間投），未知語　は削除

1. 楽天市場データの分析　（2/2）

- VSOPでの読み込み時間 2～3分
- 想定されるアプリケーション
  ✓ 類似した商品の列挙（ユーザ）
  ✓ 関連するキーワードの抽出（店舗）
  ✓ 指定されたキーワードセットを含むもの（に含まれるもの）を列挙
  ✓ 商品検索　and/or（ユーザ）
  ✓ 特徴の差分の抽出（店舗）

【TO DO】
- データベースアプリケーションとの比較
- 潜在的意味インデクシングとの精度比較
- ただし…（わかっていること）：
  ✓ 商品の特徴をつかむには，現在の紹介記事の記述では明らかに不十分（分野によってはほとんど説明文なし）
  ✓ ソーシャルフィルタリングを行うにはデータがない（のでできない）

x12	あたたかい	x1107	スカイブルー
x80	おしゃれ	x1393	トラベル
x87	おもしろい	x2123	レザー
x100	お洒落	x2203	ヴィヴィッド
x119	かっこいい	x2346	出会い
x135	かわいい	x2460	可愛い
x154	きめ細か	x2485	味わい
x186	こだわり	x2536	大きい
x225	しなやか	x2629	引き締め
x254	ずっしり	x2657	徹底的
x286	たまらない	x2693	愛らしい
x358	どっしり	x2760	控え目
x560	アイデンティティ	x2847	極めつけ
x634	イタリアン	x2932	温かみ
x681	ウォーキング	x2989	男らしい
x711	エレガント	x2992	画期的
x725	オフィス	x3298	賑やか
x754	オールラウンド	x3329	透ける
x875	クッション	x3388	雰囲気抜群
x911	グラデーション	x3420	高密度
x940	コットン	x3421	高性能
x1080	ジャケット	x3424	魅せる

抽出キーワードの例

S = 0
S = S + x940 x985 x1435 x1642 x1652 x1711 x1875 x1887 x2053 x2224 x3040 x3126
S = S + x985 x1435 x1642 x1652 x1711 x1875 x2053 x2224 x3040 x3126
S = S + x940 x985 x1435 x1642 x1652 x1711 x1875 x1887 x2224 x3040
S = S + x940 x985 x1435 x1875 x1887 x2224 x3040
S = S + x985 x2224 x3040
S = S + x761 x985 x1735 x2224 x3040
S = S + x761 x985 x1642 x1735 x1846 x1875 x2224 x3040
S = S + x761 x794 x985 x1642 x1735 x1846 x1875 x2224 x3040 x3080
S = S + x985 x1831 x1846 x2224 x3079 x3362
…………

Term-Document Matrix

2. プランタイトル作成支援システムの構築　分析の対象

2. プランタイトル作成支援システムの構築（1/2）

- プランタイトル作成支援の必要性
  - ホテルは50文字以内でプランタイトルを作成
    例)【東京タワーの夜景に乾杯ー】～タワービュークリスマス～
    IN17:00/OUT10:00(朝食付)
  - アピールしたい単語を入れることが一般的
  - ユーザはプランタイトルからホテルを選定する場合も多い
- 楽天公開データ(分析対象)
  - プランタイトル，ホテル名，ユーザ評価(7項目：総合評価，サービス，立地，部屋，設備・アメニティ，風呂，食事)など
  - タイトル総数：84,000 件余
- プランタイトル作成支援システム
  - プランタイトルに含まれる単語とユーザ評価(項目別)の相関関係を考慮し，ホテルが重視する評価項目に応じたプランタイトルの作成支援を行う．

2. プランタイトル作成支援システムの構築（2/2）

単語	プラン数	投稿数	施設数	出現回数
プラン	56599	187853	8613	58906
【	47726	177538	6502	52782
】	47729	177513	6504	52788
～	31828	155692	6390	45473
の	22759	143848	5350	44013
～	26317	108117	6141	34007
の	23782	84588	6164	29025
♪	21763	83908	4780	26884
☆	15240	86855	3876	25500
～	15689	76867	4410	23633
★	12671	75336	3641	22941
付	20596	65085	5350	22087
朝食	21106	82440	4900	21782
限定	19455	72580	5332	20817
と	15995	55804	4996	19899
お	16412	56332	5262	17907
ト	11812	51516	4208	16299
で	13994	43297	4663	15114
ポイント	9928	40933	3459	12257
～	8208	38419	2700	11985
に	10150	33670	4053	11149
◆	5668	37804	1463	10391
得	9776	36640	3552	10157
付き	9403	28769	3679	9731
個	9037	34111	3343	9720

- 現状
  - 形態素解析を行い，単語毎に出現回数や項目別評価点(平均，分散)を分析
  - 単語数：17.000 語余
- 今後の方針
  - 得点に影響する単語の組合せ(positive /negative synergy)の抽出手法の検討
    - とりあえず二単語間
  - 50文字制限の下で，しきい値以上の評点を持つ単語の組合せの検出方法の検討

3. 多面的数値・テキスト評価データに基づく推薦システム　分析の対象

3. 多面的数値・テキスト評価データに基づく推薦システム　(1/2)

■ 従来の協調フィルタリングによる推薦システム

ユーザのアイテムに対する評価値

	Item 1	Item 2	Item 3	Item 4
User 1	5	1	3	1
User 2	?	3	4	3
User 3	4	2	4	?
User 4	3	4	?	3

◆既知の評価値を基に未知の評価値を予測
◆協調フィルタリング：「似ているユーザの評価値は似ている」と仮定して未知の評価値を予測

■ 従来の方法の問題点
- ◆既知の評価値の相関をユーザ間の類似度として用いる
- ◆評価値が似ているからといって、ユーザの嗜好が似ているとは限らない
- ◆同じホテルに付けた評価値「5」は、あるユーザは「立地」を、別のユーザは「設備」を重視して評価したかもしれない

3. 多面的数値・テキスト評価データに基づく推薦システム　(2/2)

■ 楽天トラベルデータ(ホテル評価)の特徴
- ◆ 総合評価だけでなく「立地」「設備」「食事」等の多面的な評価値を付与
- ◆ 数値評価に加えて感想や苦情がレビューテキストとして与えられている

■ 目標：複数側面での(多面的な)評価値とテキストの内容を考慮して推薦精度を向上

■ 方針：
- ◆ 各側面での評価値に加え、レビューでの語の使用回数も「側面」として行列に保持
- ◆ これらをまとめた「テンソル補完」の問題として定式化
- ◆ 語の組合せやフレーズを考慮する場合、ZDD等の技術が必要となる