第１６章
アソシエイション分析と意味処理

09t4020a　菊池裕紀

1、アソシエーション分析

アソシエーション分析とは

アソシエーション分析のルール

R言語による実装

2、評判分析

3、意味処理と辞書

1、アソシエーション分析

アソシエーション分析とは

アソシエーション分析のルール

R言語による実装

2、評判分析

3、意味処理と辞書

アソシエーション分析

Association

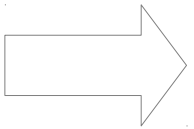
つながり　　関連性　　連関

連想　　相関

巨大なデータから
価値あるAssociation Ruleを抽出する分析手法

例　　POSデータ

購買データ



Association Rule

仮面ライダー変身ベルトを購入した人の９０％が関連武器玩具を購入している。

Association Rule

Aが起こると、Bが起こる

$$A \Rightarrow B$$

条件部(LHS)　　結論部(RHS)

例　　A：仮面ライダー変身ベルトを購入　　B：関連武器を購入

1、アソシエーション分析

アソシエーション分析とは

アソシエーション分析のルール

R言語による実装

2、評判分析

3、意味処理と辞書

Association Ruleを検出する際の評価指標
(良く用いられるもの)
Association Rule：X ⇒ Y

1,support(支持度)　条件Xと結論Yを含む事例が、全事例中に占める比率

アイテム集合Xを含む事例の数：Xの支持度数　→σ(X)

$$support(X \Rightarrow Y) = \frac{\sigma(X \vee Y)}{M}$$

※M：事例の総数

⇒Ruleが発生する確率

2,confidence(確信度) 条件X、結論Yを含む事例と条件Xを含む事例との比率

$$confidence(X \Rightarrow Y) = \frac{\sigma(X \vee Y)}{\sigma(X)} = \frac{support(X \Rightarrow Y)}{support(X)}$$

⇒条件Xが発生したとき、結論がYになる確率

3,lift(リフト)　条件X、結論Yの同時発生確率と条件X・結論Yの発生確率の積の比率

$$lift(X \Rightarrow Y) = \frac{cofidence(X \Rightarrow Y)}{support(Y)} = \frac{support(X \Rightarrow Y) / support(X)}{support(Y)} = \frac{support(X \Rightarrow Y)}{support(X) * support(Y)}$$

⇒２事象の独立性の判定

1、アソシエーション分析

アソシエーション分析とは

アソシエーション分析のルール

R言語による実装

2、評判分析

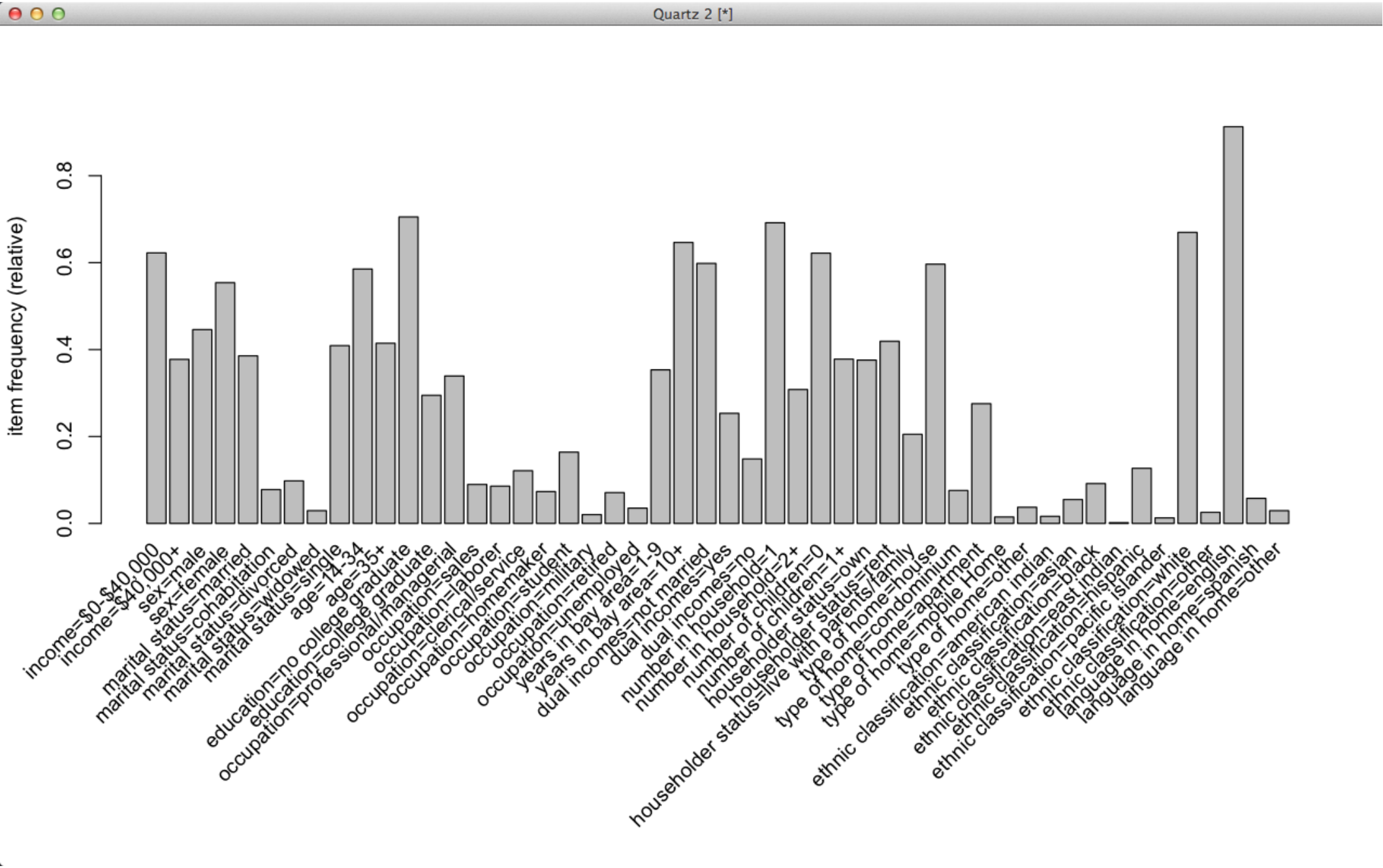
3、意味処理と辞書

サンフランシスコショッピングモールの顧客のアンケート結果(arulesパッケ)

14項目にYes：０、NO：１で答えたもの

```
> data(Income)
> Income
transactions in sparse format with
 6876 transactions (rows) and
 50 items (columns)
> incomematrix <- as(Income,"matrix")
> incomematrix[1,]
```

income=$0-$40,000	income=$40,000+	sex=male	sex=female	
0	1	1	0	
marital status=married	marital status=cohabitation	marital status=divorced	marital status=widowed	
1	0	0	0	
marital status=single	age=14-34	age=35+	education=no college graduate	
0	0	1	0	
education=college graduate	occupation=professional/managerial	occupation=sales	occupation=laborer	
1	0	0	0	
occupation=clerical/service	occupation=homemaker	occupation=student	occupation=military	
0	1	0	0	
occupation=retired	occupation=unemployed	years in bay area=1-9	years in bay area=10+	
0	0	0	1	
dual incomes=not married	dual incomes=yes	dual incomes=no	number in household=1	
0	0	1	0	
number in household=2+	number of children=0	number of children=1+	householder status=own	
1	0	1	1	
householder status=rent	householder status=live with parents/family	type of home=house	type of home=condominium	
0	0	1	0	
type of home=apartment	type of home=mobile Home	type of home=other	ethnic classification=american indian	
0	0	0	0	
ethnic classification=asian	ethnic classification=black	ethnic classification=east indian	ethnic classification=hispanic	
0	0	0	0	
ethnic classification=pacific islander	ethnic classification=white	ethnic classification=other	language in home=english	
0	1	0	1	
language in home=spanish	language in home=other			
0	0			


Quartz 2 [*]
item frequency (relative)
0.0
0.2
0.4
0.6
0.8
income=$0-$40,000
income=$40,000+
sex=male
sex=female
marital status=married
marital status=cohabitation
marital status=divorced
marital status=widowed
marital status=single
age=14-34
age=35+
education=no college graduate
education=college graduate
occupation=professional/managerial
occupation=sales
occupation=clerical/service
occupation=laborer
occupation=homemaker
occupation=student
occupation=military
occupation=retired
occupation=unemployed
years in bay area=1-9
years in bay area=10+
dual incomes=not married
dual incomes=yes
number in household=1
number in household=2+
number of children=0
number of children=1+
householder status=own
householder status=rent
householder status=live with parents/family
type of home=house
type of home=condominium
type of home=apartment
type of home=mobile Home
type of home=other
ethnic classification=american indian
ethnic classification=asian
ethnic classification=black
ethnic classification=east indian
ethnic classification=hispanic
ethnic classification=pacific islander
ethnic classification=white
ethnic classification=other
language in home=english
language in home=spanish
language in home=other

関数　　apriori

Association Ruleの抽出

```
> library(arules)
> data(Income)
> result <- apriori(Income)

parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.8    0.1    1 none FALSE            TRUE     0.1      1     10  rules FALSE

algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[50 item(s), 6876 transaction(s)] done [0.01s].
sorting and recoding items ... [30 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.06s].
writing ... [8664 rule(s)] done [0.00s].
creating S4 object  ... done [0.01s].
```

関数　　inspect

抽出ルールの呼び出し

```
> inspect(head(SORT(result, by="support"), n=20))
   lhs                                  rhs                                 support confidence      lift
1  {}                                => {language in home=english}      0.9128854  0.9128854 1.0000000
2  {ethnic classification=white}     => {language in home=english}      0.6595404  0.9847991 1.0787763
3  {number in household=1}           => {language in home=english}      0.6495055  0.9388270 1.0284171
4  {education=no college graduate}   => {language in home=english}      0.6343805  0.8995669 0.9854106
5  {years in bay area=10+}           => {language in home=english}      0.6013671  0.9300495 1.0188020
6  {number of children=0}            => {language in home=english}      0.5801338  0.9328812 1.0219040
7  {income=$0-$40,000}               => {language in home=english}      0.5578825  0.8962617 0.9817899
8  {number of children=0}            => {number in household=1}         0.5532286  0.8896165 1.2858951
9  {type of home=house}              => {language in home=english}      0.5446481  0.9129693 1.0000919
10 {dual incomes=not married}        => {language in home=english}      0.5426120  0.9069033 0.9934470
11 {age=14-34}                       => {language in home=english}      0.5248691  0.8966460 0.9822109
12 {number in household=1,
    number of children=0}            => {language in home=english}      0.5213787  0.9424290 1.0323629
13 {number of children=0,
    language in home=english}        => {number in household=1}         0.5213787  0.8987215 1.2990559
14 {number in household=1,
    language in home=english}        => {number of children=0}          0.5213787  0.8027318 1.2908287
15 {sex=female}                      => {language in home=english}      0.5122164  0.9246521 1.0128896
16 {income=$0-$40,000}               => {education=no college graduate} 0.5018906  0.8063084 1.1433649
17 {number in household=1,
    ethnic classification=white}     => {language in home=english}      0.4941827  0.9880779 1.0823680
18 {number of children=0,
    ethnic classification=white}     => {language in home=english}      0.4474985  0.9868505 1.0810235
19 {income=$0-$40,000,
    education=no college graduate}   => {language in home=english}      0.4454625  0.8875688 0.9722675
20 {education=no college graduate,
    ethnic classification=white}     => {language in home=english}      0.4384817  0.9827249 1.0765041
```

関数　　subset

結果の部分抽出

```
> sub <- subset(result, subset = rhs %in% "income=$40,000+" & lift > 2)
> inspect(SORT(sub)[1:9])
  lhs                                          rhs                    support confidence     lift
1 {occupation=professional/managerial,
   householder status=own}                  => {income=$40,000+} 0.1384526  0.8074640 2.138722
2 {occupation=professional/managerial,
   householder status=own,
   language in home=english}                => {income=$40,000+} 0.1336533  0.8075571 2.138969
3 {dual incomes=yes,
   householder status=own}                  => {income=$40,000+} 0.1260908  0.8156162 2.160315
4 {marital status=married,
   occupation=professional/managerial,
   language in home=english}                => {income=$40,000+} 0.1233275  0.8030303 2.126979
5 {marital status=married,
   dual incomes=yes,
   householder status=own}                  => {income=$40,000+} 0.1227458  0.8194175 2.170383
6 {education=college graduate,
   householder status=own,
   language in home=english}                => {income=$40,000+} 0.1212914  0.8011527 2.122005
7 {dual incomes=yes,
   householder status=own,
   language in home=english}                => {income=$40,000+} 0.1207097  0.8258706 2.187476
8 {occupation=professional/managerial,
   householder status=own,
   type of home=house}                      => {income=$40,000+} 0.1192554  0.8175474 2.165430
9 {marital status=married,
   dual incomes=yes,
   householder status=own,
   language in home=english}                => {income=$40,000+} 0.1175102  0.8304214 2.199529
```

１、アソシエーション分析

アソシエーション分析とは

アソシエーション分析のルール

R言語による実装

２、評判分析

３、意味処理と辞書

Chrome ファイル 編集 表示 履歴 ブックマーク ウインドウ ヘルプ (100%) 日 5:25 菊池 裕紀

coco - 映画レビューサイト

coco.to

ブックマーク ログイン Active! mail 自然言語処理 新納研究室 ネットワーク認証 染谷翔の自転車世界一

Twitter連動型映画レビューサイト 「ロード・オブ・ザ・リング」は、この冒険から始まった

coco http://coco.to × ホビット 思いがけない冒険

ログイン

映画名 Twitter ID

最新映画ニュース

▶ 他のニュースも見る

- 及川正通が久しぶりに映画作品を書き下ろし！ びあ映画特別号発売
- 『ホビット』ピーター・ジャクソン監督がホビットにほれ込む理由とは？
- 女の本音と男の本性が浮き彫りに 『つやのよる』ダイジェスト映像が到着！
- 「DanceDanceRevolution」YUNIの号令でお仕事終了...

おしらせ

- 『アウトロー』ジャパンプレミア5組10名様
- 『ホビット 思いがけない冒険』特製グッズ20..
- マイベスト10機能をリリースしました

coco 映画特集

▶ 特集一覧

ホビット 思いがけない冒険 new

映画ファン感嘆の声！
12/1特別試写会レポート

LOOPER/ルーパー new

今度の未来は2044年
31年待たずとも今目撃せよ！
タイムトラベル映画の新機軸

砂漠でサーモン・フィッシング

監督x脚本家xキャストが贈る
信じる心を取り戻す旅

coco 映画特集

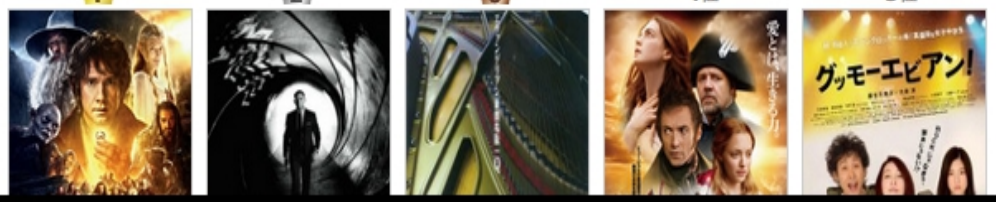
注目度 | 評価順 | 公開中 | 近日公開 | 全ての映画

Twitterのつぶやきが多い新作映画ランキング

1週間 | 1日

ポスター表示 タイトル表示

1位 2位 3位 4位 5位



映画マニア達の注目作品ランキング

1 ホビット 思いがけない冒険 307pt.
2 007 スカイフォール 227pt.
3 砂漠でサーモン・フィッシン.. 114pt.
4位 レ・ミゼラブル 63pt.
4位 ルビー・スパークス 63pt.
6位 フランケンウィニー 47pt.
6位 ヱヴァンゲリヲン新劇場版：.. 47pt.
8位 人生の特等席 39pt.
9位 ゼロ・ダーク・サーティ 35pt.
10位 ライフ・オブ・パイ トラと.. 30pt.

もっと見る

CGM(Consumer Generated Media)

出現

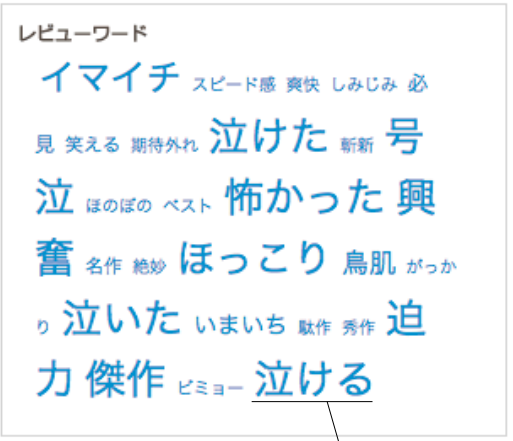
↓

Amazonに投稿されたレビューやブログなど
一般人が作成、発信するコンテンツであるのがポイント

NLP的に見ると...

CGMの出現で、新しい解析技術が必要に！
　–テキスト中の評判や意見などの把握

従来、テキスト中の客観的事実が対象
　–レビューなどはきわめて主観的である。



語や語句のカテゴリ化には意味情報を用いる。

–分類語彙表のような辞書が必要である！

？
ポジティブ
or
ネガティブ
？

１、アソシエーション分析

アソシエーション分析とは

アソシエーション分析のルール

R言語による実装

２、評判分析

３、意味処理と辞書

現段階のテキスト解析...

形態素解析、構文解析を用いてテキストに現れている要素やそれらの共起関係などを集計して分析

↓

意味情報の分析には限界がある。

- シーソラス辞書
- 概念辞書
- 文脈情報

さまざまな辞書を取り入れた研究開発が必要！