

情報可視化ソフトウェア Hiddenによるデータ分析

コンピュータグラフィックス

Travel Reviews Data Set

1. データの概要

UCI Machine Learning Repositoryからデータを取得

【Travel Reviews Data Set】<https://archive.ics.uci.edu/ml/datasets/Travel+Reviews#>

東アジアの観光スポットを評価したデータをTripAdvisor(<https://www.tripadvisor.jp>)から入手したもの。

横軸

art galleries
dance clubs
juice bars
restaurants
museums
resorts
parks/picnic spots
beaches
theaters
religious institutions

計 10項目

縦軸

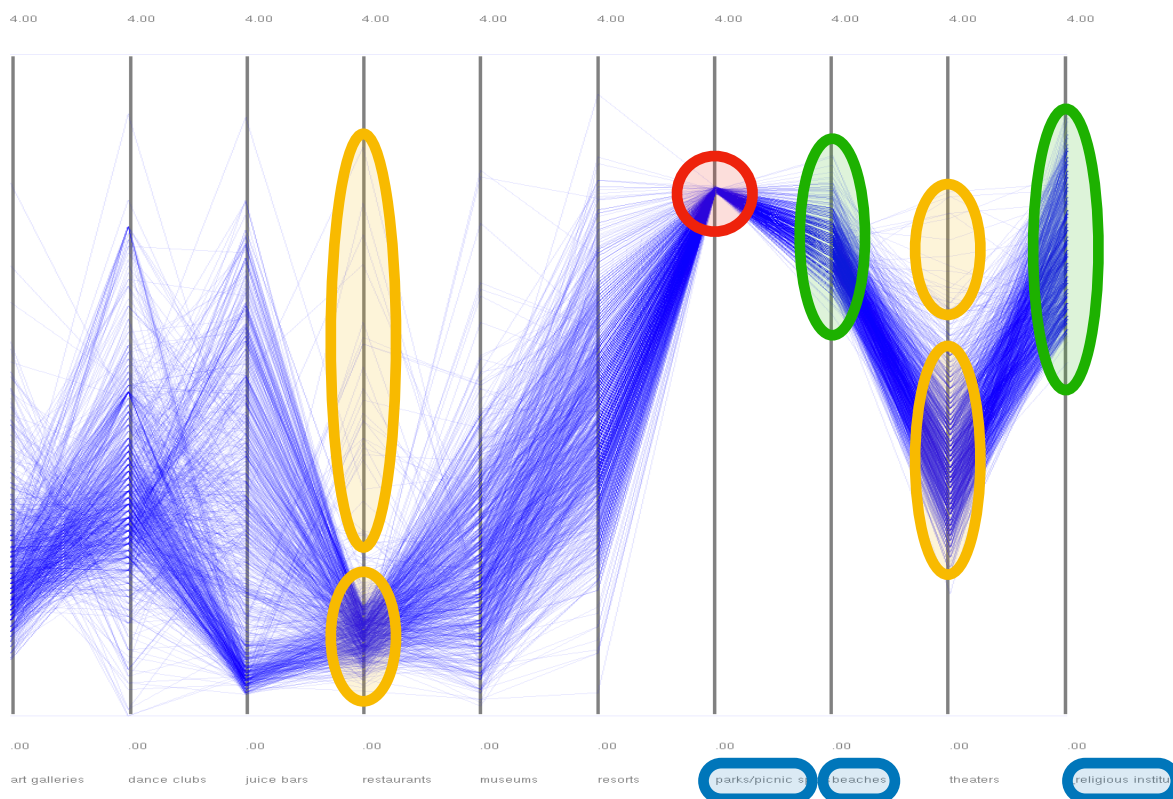
ユーザが各観光スポットに対して評価をつけ、複数あるものに関しては平均を取ったもの。

評価4点 Excellent
評価3点 Very Good
評価2点 Average
評価1点 Poor
評価0点 Terrible

計 982ユーザ

2. 全体を見て

縦軸の数値の幅が全て同じなので、全て0に評価したユーザと全て4に評価したユーザをデータに加えた。



parks/picnic spots 全てのユーザが3.2程度の評価をしている。

beaches , religious institutions 各ユーザが同じような評価をしている。

restaurants , theaters

同じような評価をしているユーザが多いが、一部それよりも高い評価をしているユーザもいる。

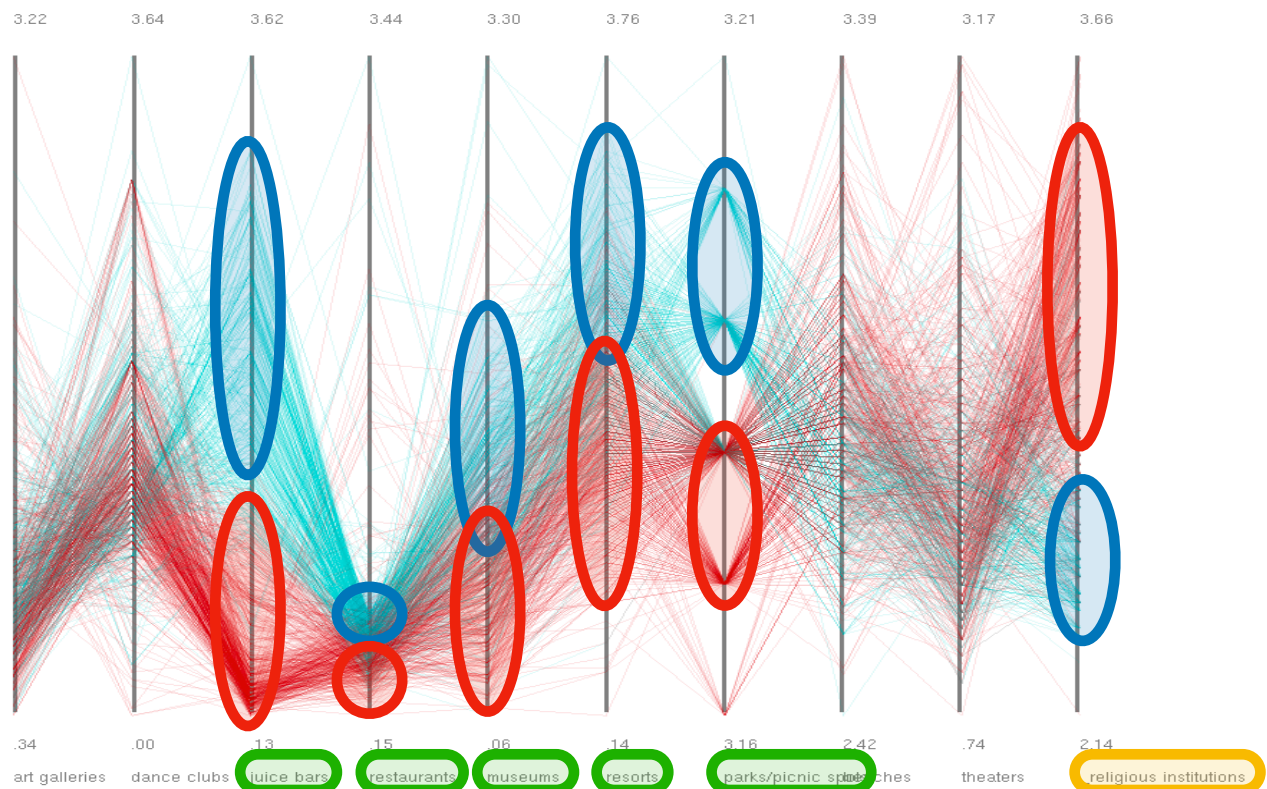
parks/picnic spots , beaches , religious institutions

全てのユーザから高い評価を得ている。

考察

- parks/picnic spotsについて、「この項目については3.2前後で評価してください」と書いてあるのか？という程に評価がまとまったので驚いた。データとして不自然な気がする。
- beachesは、「よっぽど海が汚い」「クラゲに刺された」「怪我をした」という特別に悪いことがない限り、「海が綺麗だな」といったような感想しか出てこないのでは、評価のばらつきがないと考える。
- restaurantsにおいては、町の露店や屋台などの評価が低く、高級レストランに行った人だけが高評価をつけていると考えられる。
- 比較的高評価のスポットは評価のばらつきがなく、低評価のスポットは評価にばらつきがある。

3. 2つにクラスタリング



juice bars , restaurants , museums , resorts, parks/picnic spots

赤：評価が低いユーザ，青：評価が高いユーザにわかれた。

religious institutions

赤：評価が高いユーザ，青：評価が低いユーザにわかれた。

考察

- 赤：全体的に評価を低くつけがちなユーザ,青：全体的に評価を高くつけがちなユーザにわかれた。
分類の考え方として以下の2つが挙げられる。
 - ①評価を厳しく付ける人と易しく付ける人の分類
例えば私の場合、基本3に付けて良かったら4,ちょっと気になる点があれば2に付けるが、評価2点 Averageになっているので、基本2に付ける人も多いと思う。
 - ②東アジア旅行が楽しかったユーザと楽しなかったユーザ
- 分類について上のように考えたが、religious institutionsだけ逆になったことが不思議である。
religious institutionsを高く評価したユーザは宗教信仰心が強く、religious institutionsだけを観に行く目的で、一般的な観光地を目的として旅行していない…？

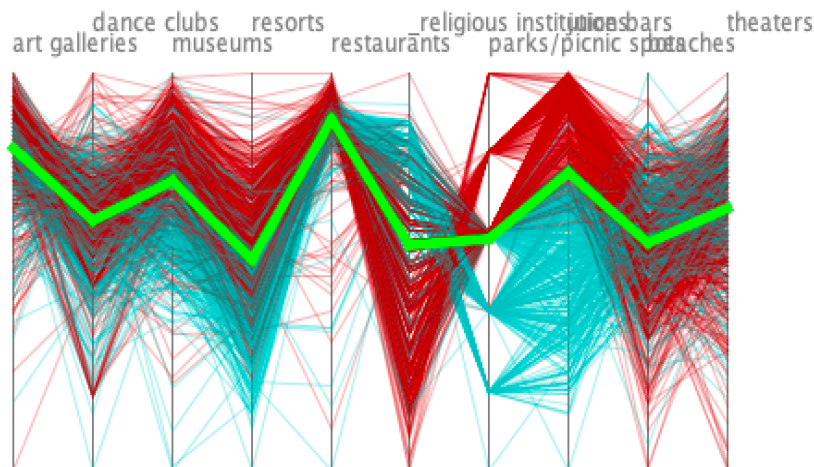
余談

parks/picnic spotsを見てみると、評価が5つにわかれている。

他のデータは評価がこのように綺麗にわかれることがないので2. 全体を見て述べたように不自然なデータになっており、評価に手が加わっていると考えられる。

AWT版のプログラムで平均線を描画 縦軸の上下が逆になっている

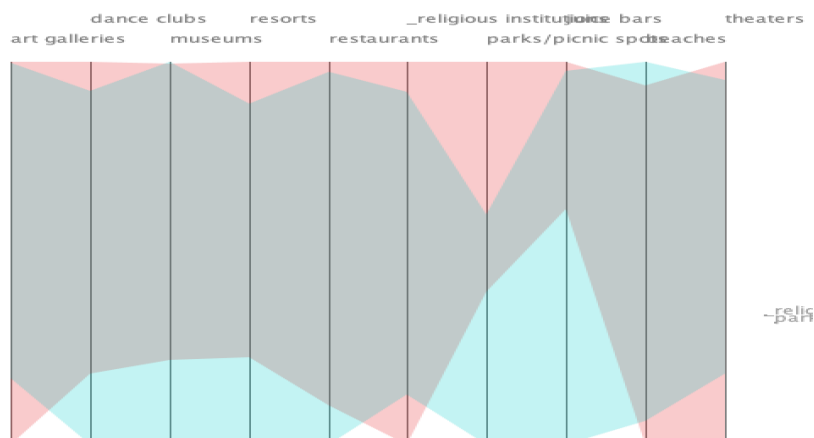
綺麗にクラスタリングされた項目については平均線はおおよそ、赤と青の境目となっていた。



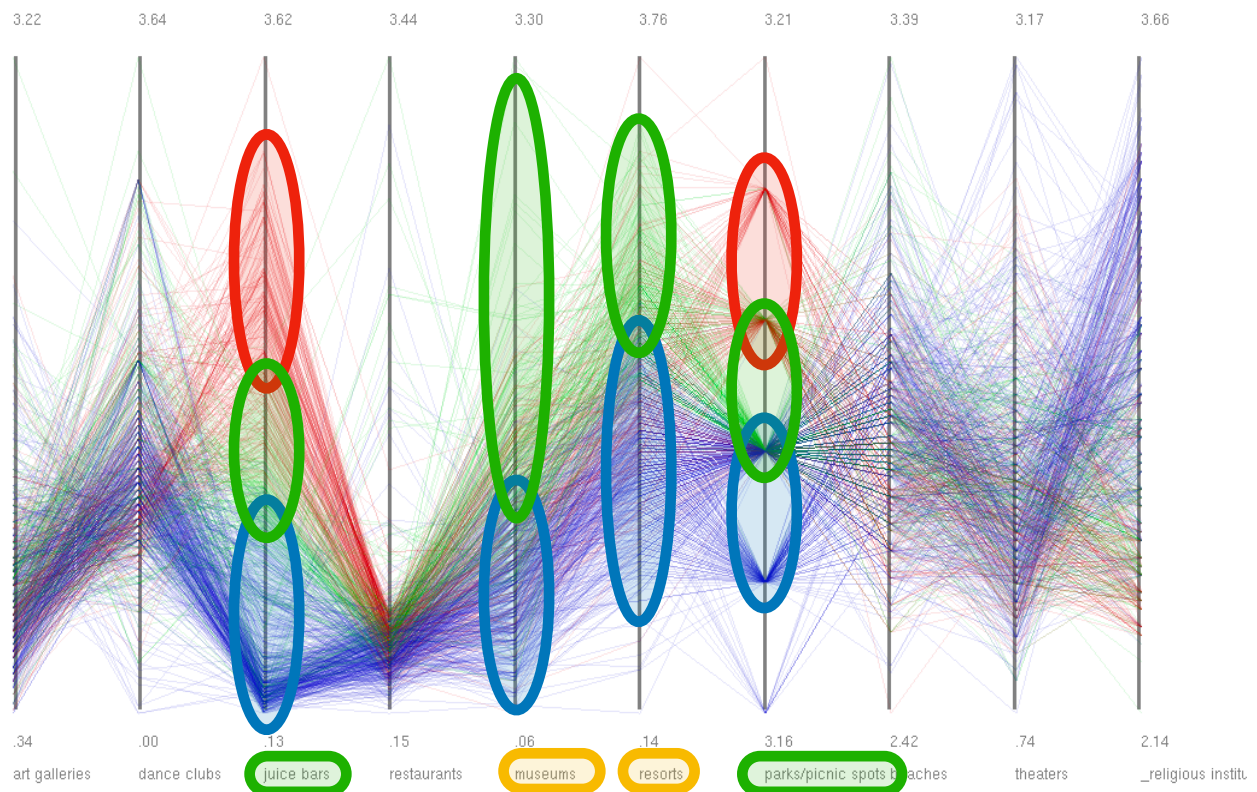
AWT 版のプログラムで図形を描画 縦軸の上下が逆になっている

重なっている部分は広いものの、考察の1つめで述べたように、

赤：全体的に評価を低くつけがちなユーザ，青：全体的に評価を高くつけがちなユーザ
ということがパッと見でなんとなくわかる。



4. 3つにクラスタリング



juice bars , parks/picnic spots

2つにクラスタリングした際の赤青と同じで(色は逆になっている)、間の値が緑になった。

museums , resorts

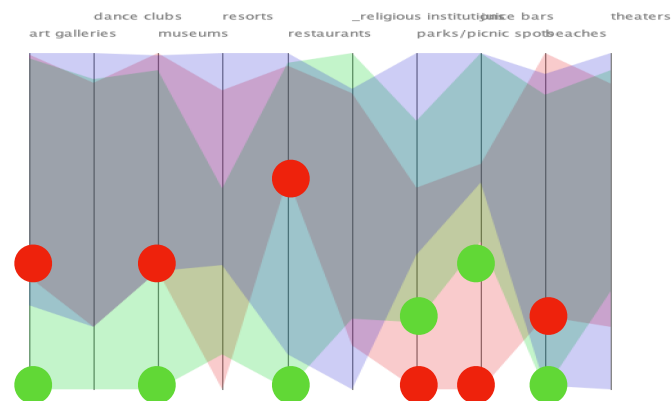
緑：評価が高いユーザ，青：評価が低いユーザにわかれた。

考察

- juice bars , parks/picnic spotsで高い評価をした人はmuseums , resortsでも高い評価をしている訳ではない。
- 青にクラスタリングされているユーザーは全体的に評価を低くつけがちである。(religious institutionsを除く)

AWT 版のプログラムで図形を描画 縦軸の上下が逆になっている

(juice bars , parks/picnic spots , museums , resortsについては上記のようにクラスタリングされているとわかったが、正直全体的にどのようにわかれているかわからなかったのです)

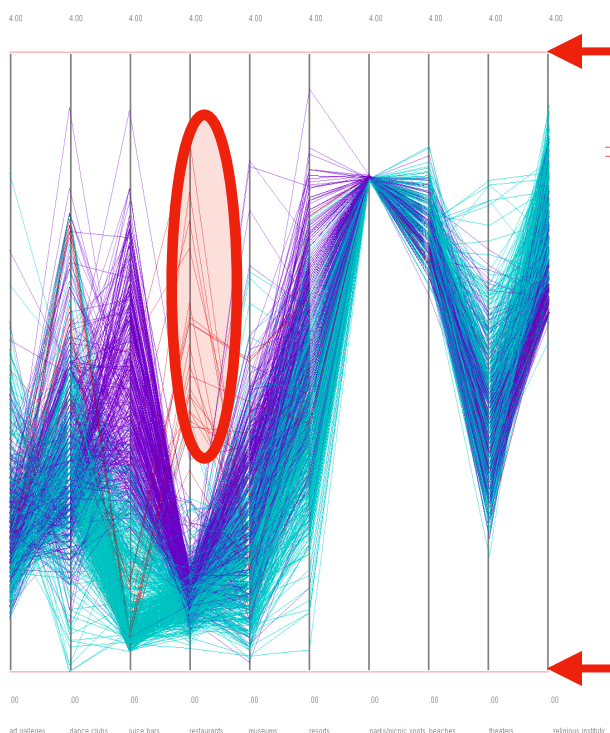


青：全体的に評価が低いユーザ

赤のユーザが高評価した項目は緑で評価が高くないことが多い。逆も言える。

赤と緑は価値観が違うユーザ同士であるのかもしれない??

全て0に評価したユーザと全て4に評価したユーザをデータに追加



赤：restaurantsで高く評価したユーザ
全て0に評価したユーザ
全て4に評価したユーザ

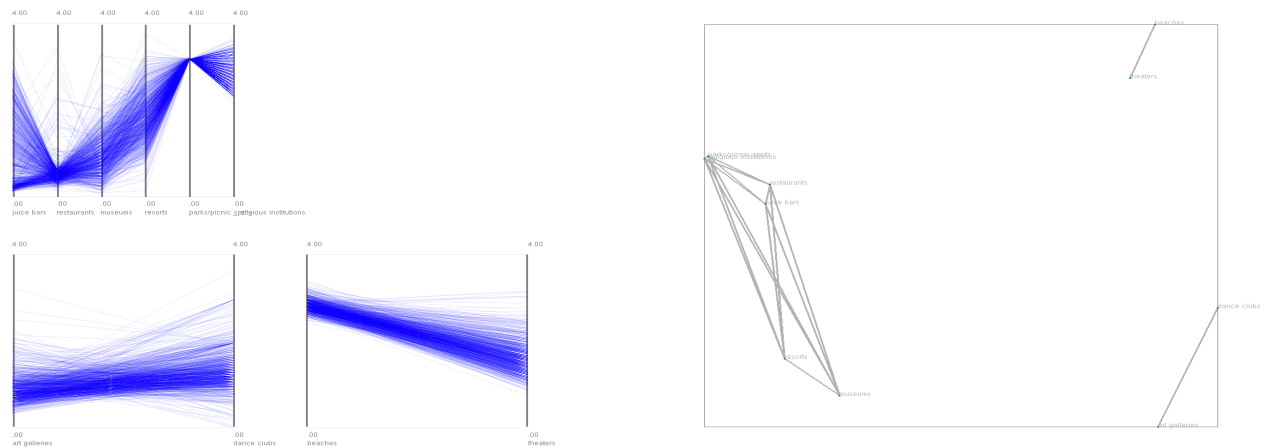
水色：3. 2つにクラスタリングの赤と同じ。

紫：3. 2つにクラスタリングの青と同じ。

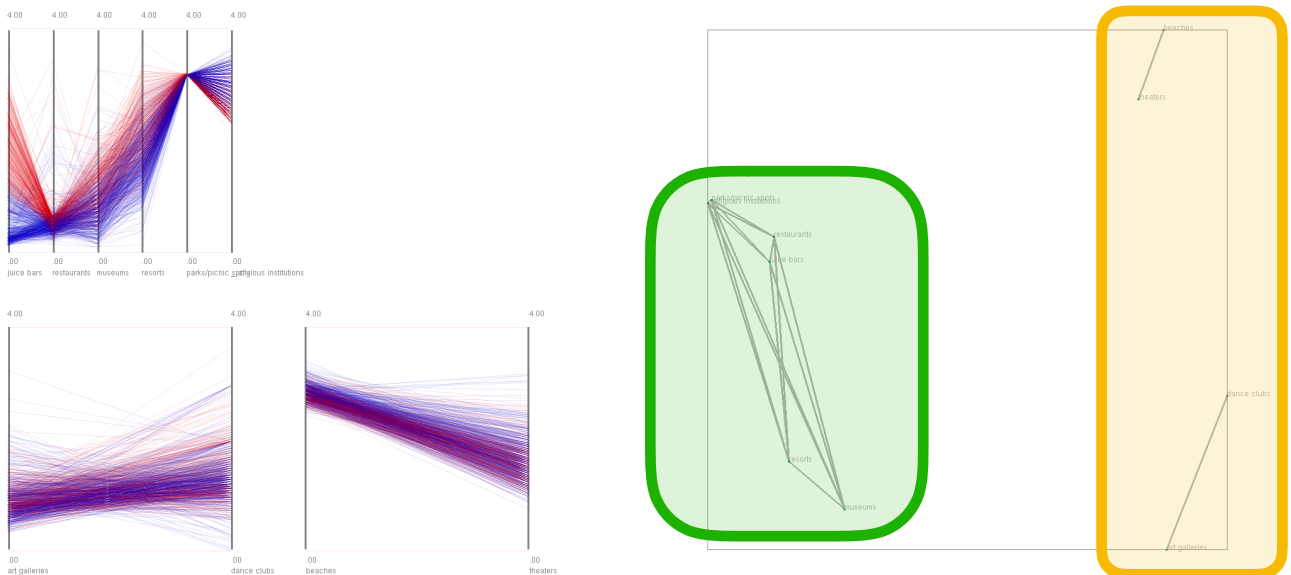
考察

- ・ 全て0に評価したユーザ・全て4に評価したユーザは自分で追加した特別な線であることからわかるように、赤に分類されたユーザは例外として扱われたと考える。

5. 分類分け



このように大きくわけて3つのブロックにわかれた。確かに3つとも強い相関関係が現れているが、なぜこのようにわかれたのかわからなかったので2つにクラスタリングしてみた。

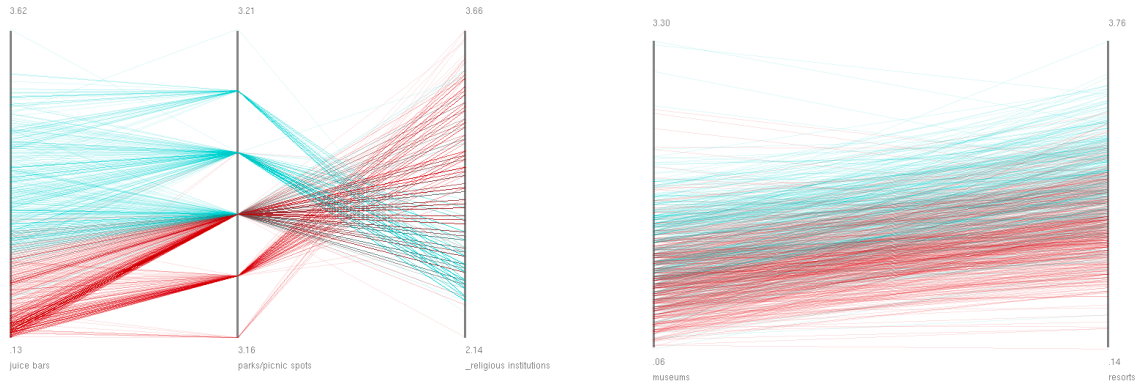


考察

- 緑で囲んだ部分に関しては綺麗にクラスタリングされた。
- 3. 2つにクラスタリングという項目で juice bars , restaurants , museums , resorts, parks/ picnic spots , religious institutionsについて綺麗にクラスタリングされていると述べたが、この散布図を見れば一目瞭然だ。正しく考察できた。

6. 相関関係

一番強い相関関係について取り上げる。



考察

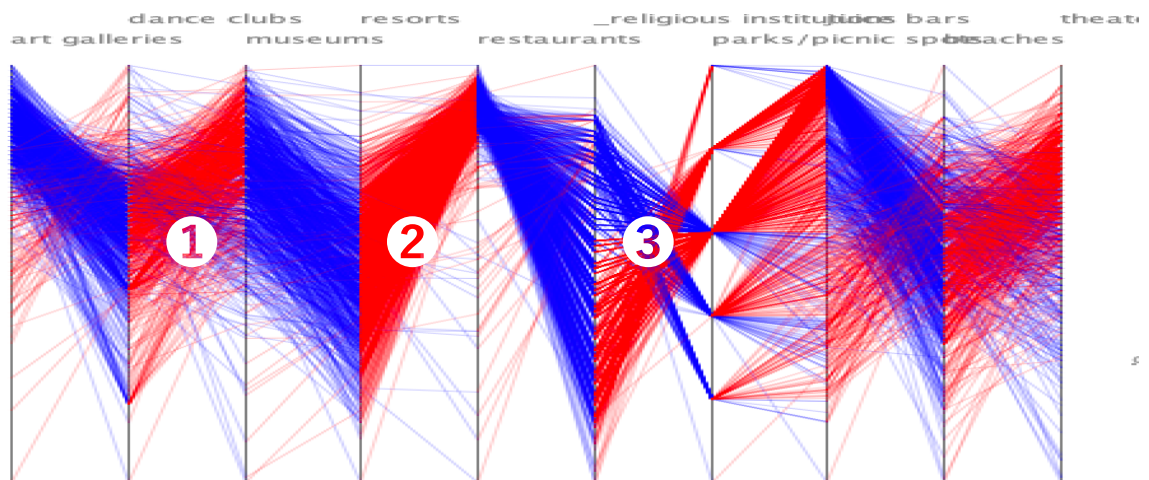
右図より、museumsとresortsは**正の相関**がある。

左図より、juice barsとparks/picnic spotsは**正の相関**があり、parks/picnic spotsとreligious institutionsに**負の相関**がある。つまり、juice barsとreligious institutionsには**負の相関**がある。

7. おまけ

線が右下がりの場合青、右上がりの場合赤の線で描画

AWT 版のプログラムで描画したので、縦軸の上下が逆になっている



考察

- ①は7割程度のユーザが同じような評価している。
(museumsよりdance clubの方に高い評価をした 青の線が例外のユーザ)
- ②はほとんどのユーザが同じように評価している。
(restaurantsよりresortsの方に高い評価をした 青の線が例外のユーザ)
- ③はユーザの意見が二つにわかれている。
- 以上のように、この表現手法によりユーザの意見の一致度が測れる。

8. まとめ

- ・ クラスタリングをすることによってユーザの分類が出来、ユーザの分類ごとに観光スポットの提案ができる。

〇〇を高く評価した人は××を高く評価している

↓

〇〇のページを見ている人に××をお勧め

- 4. 3つにクラスタリングでjuice barsを高く評価している人はparks/picnic spotsも高く評価していることがわかったので

juice barsを高く評価した人はparks/picnic spotsを高く評価している

↓

juice barsのページを見ている人にparks/picnic spotsをお勧め

ということが可能になる。

- ・ 3. 2つにクラスタリング 余談で述べたように、いかにもおかしいデータもPCPで表すことによって明確になる。(食べログの3.8→3.6問題もPCPで表すと不正が見つかるのでは……?)
- ・ 4. 3つにクラスタリング 全て0に評価したユーザと全て4に評価したユーザを自分でデータ加えたで述べたように、例外のユーザが見つけれられる。

例えば下記のようにしてデータ分析の精度向上に貢献できると考える。

- 4. 3つにクラスタリングではrestaurantsに例外が見られたので

- ① 例外を切り落として、より平均的な結果を表示

restaurantsは評価がとっても低い！東アジアのご飯は美味しくない！

- ② 逆に例外に注目して原因を追求し、新たな意見として取り上げる

restaurantsで高く評価している人は高級レストランに行った人だ！

(この場合の原因は予想だが、実際のデータを見れば原因が明らかになる)

restaurantsの項目を予算を元に2つにわけよう！

- ・ 数値からは項目一つ一つの評価しかわからないが、PCPで表すことによって項目間の関係がわかった。
- ・ 今回はデータの特性上、クラスタリングをすることによりわかることが多かった。