コロケーション・パーチェス2部グラフの対応分析階層クラスタリングによる顧客購買行動可視化システム

大槻明[1*], 川村雅義[2]
Akira OTSUKI[1*], Masayoshi KAWAMURA[2]

1 東京工業大学
Tokyo Institute of Technology
〒152-8550東京都目黒区大岡山2-12-1
E-mail: otsuki.a.aa@m.titech.ac.jp

2 MK future software
E-mail: kawamura.masa@nifty.com

*連絡先著者 Corresponding Author

本研究では，Twitterから位置（地域）情報及びツイート情報を取得し，ツイート情報から形態素解析及びルールベースによる係り受け解析によって「何処で何を買ったか（購買情報）」という情報を抽出する．そして，「地域情報」と「購買情報」を2部グラフとしてとらえ，Responsiveness Pair Clustering分析（対応分析を類似尺度としたクラスタリング）を行うことによって，同じような購買情報ノード群を参照する地域情報クラスタが生成される．最後に，この地域情報クラスタを地図上にマッピングすることにより類似地域を可視化する．Modularityでは，２部グラフのようなリンクに制限のあるネットワークを分析することが難しかったが，本研究では類似尺度にModularityではなくノード間の対応性を用いることで対応した．最後に，この可視化システムは顧客購買行動等の分析ツールとしての貢献が期待される．

キーワード :ビッグデータ俯瞰分析，顧客行動分析，クラスタリング，計量書誌学，データベース

発表内容

発表内容は，次ページ以降のスライドに示す．

コロケーション・パーチェス ビッグデータの対応分析 クラスタリングによる顧客 購買行動可視化システム

大槻明（東京工業大学）
川村雅義（MK future software）

1

はじめに

コロケーション情報

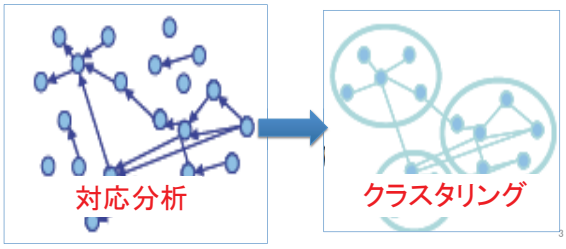
Twitterから位置（市区町村）情報と
何処で何を買ったか情報を取得

パーチェス情報

2

はじめに

前頁の情報を2部グラフとしてとらえ，対応性を類似尺度としたクラスタリングを行う．



3

はじめに

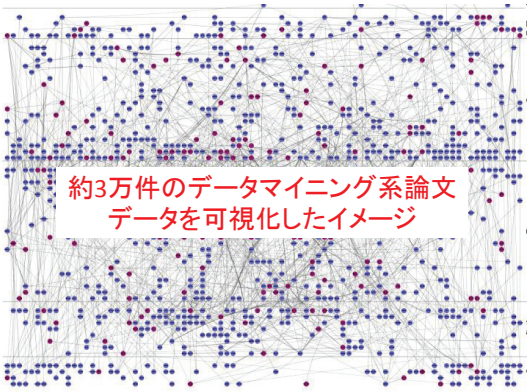
クラスタリングの結果生成される類似地域を可視化

4

関連・先行研究

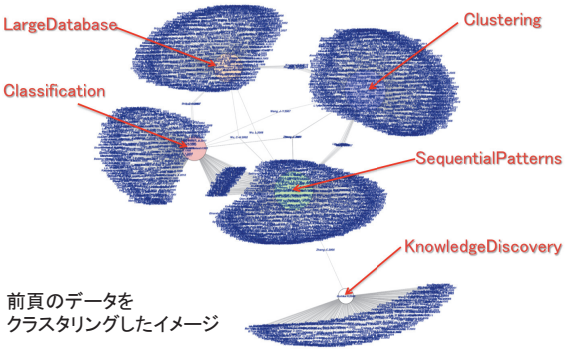
5

計量書誌学とModularity



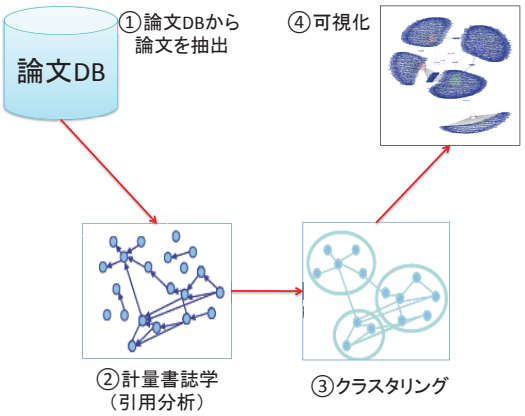
6

計量書誌学とModularity



前頁のデータを
クラスタリングしたイメージ

7



8

Modularity Qによる クラスタリング

9

モジュラリティQ

$$Q = \sum_i \left(e_{ii} - a_i^2 \right)$$

e_{ii}は，同じグループに所属するノード同士が繋がるリンク数の全リンク数に占める割合

a_i^2は，リンクの一方がグループ内のノードに繋がるリンク数の，全リンク数に占める割合の期待値，すなわちe_{ii}の期待値

10

モジュラリティQ

- コミュニティ分割時の評価関数
- モジュラリティQが高い時にグループ分割をすれば，
- グループ（モジュール）内での繋がりが密な状態で，グループ外との繋がりが疎な状態で分割できる．

11

しかし

12

Modularityは，リンクに制限の無いネットワーク
を処理することは出来るが

引用されている論文	被引用論文
大槻2010	著者A2013
大槻2010	著者B2011
大槻2010	著者C2012
東工大1995	著者D2000
東工大1995	大槻2010
東工大1995	著者E1998
東工大1995	著者B2011

13

リンクに制限が有る2部グラフネットワーク
を処理することは難しい

Buy_Place	City
家電_家電量販店	東京都新宿区
家電_家電量販店	埼玉県朝霞市
ケーキ_百貨店	大阪府大阪市
服_百貨店	千葉県市川市
ケーキ_スーパー	大分県大分市
ケーキ_百貨店	東京都足立区
生活用品_スーパー	秋田県北秋田市

14

ゆえに，
Modularity Qを2部グラフに対応させる研究が，
これまでに数多くなされてきた．

15

Modularity Qの2部グラフへの応用

先行研究

16

BarberのModularity Q^B

$$Q^B = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}(A_{ij} - P_{ij})\delta(c(v_i), c(v_j))$$

P_{ij}は，ランダム2部グラフにおける
ノードV_i, V_jにリンクが存在する確率

A_{ij}は，隣接行列の要素

17

村田のModularity Q^M

$$Q^M = \sum_i (e_{ij} - a_i a_j), j = \frac{argmax(e_{ik})}{k}$$

Q^Mは，最も対応するコミュニティ
とのリンク密度を評価

18

原田のModularity Q^H

$$Q^H = \sum_{i=1}^{|C_A|}\sum_{j=1}^{|C_B|} a_{ij}(e_{ij} - a_i a_j)$$

コミュニティ集合をC_A, C_Bとするとき，2部グラフのコミュニティ集合を$C=C_A \cup C_B$で表す(∪は論理積).

19

原田のModularity Q^H

$$Q^H = \sum_{i=1}^{|C_A|}\sum_{j=1}^{|C_B|} a_{ij}(e_{ij} - a_i a_j)$$

$(e_{ij}-a_ia_j)$は，Modularityと同じく，コミュニティ間のリンク密度とリンク密度の期待値との差を計算

20

先行研究の問題指摘

- Modularity Qは，同じような参照が増えてくればまとまっていくという特徴を持つため，
- 結果，より少ないコミュニティ数で分割がなされることになる.

21

先行研究の問題指摘

しかし，
SNSパーチェスデータにおいては，ノードの増加に比例してノード間の対応性は増加すると考えられる.

22

提案手法

コンセプト

23

コンセプト

先行研究

- コミュニティ形成時の類似尺度にリンクの構造に着目

本提案

- コミュニティ形成時の類似尺度にリンクの対応性に着目

24

コンセプト

Twitterから「位置情報」と「購買情報」を取得

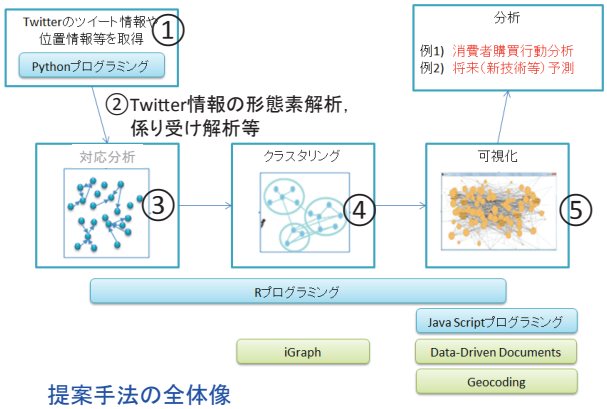
↓

同じような「何処で何を買ったのか」ノード群を参照する「地域」ノード群を抽出

↓

似た場所クラスタを抽出・可視化

25



提案手法の全体像

26

①Twitterからの情報取得

位置情報とツイート情報の取得

27

①Twitterからのデータ取得

○対象期間：クリスマス
2012年12月22日～25日

○取得件数：1，083

○使える情報： 731
→位置情報が無いものや
購買情報が無いものは削除

＊Pythonでスクリプトを作成

例：「買ったぜ！」

28

②Twitter情報の語彙解析

形態素解析と
ルールベースの係り受け解析

29

②Twitterから取得したデータのイメージ

①【日付】Tue, 22 Dec 2012,

②【ID】twitter_User_ID,

③【ツイート内容】○○百貨店でケーキを買った. (@ ○○百貨店w/7 others),

④【緯度経度】[35.628227, 139.738712]

30

②位置情報の取得

Geocoding

東京都港区

④【緯度経度】[35.628227, 139.738712]

緯度経度情報をGeocoodingに利用して，市区町村情報を取得

【参考】簡易ジオコーディングサービス
http://www.finds.jp/wsdocs/geocode/index.html#URL

31

②「何処で買ったか」情報

③【ツイート内容】〇〇百貨店でケーキを買った．(@ 〇〇百貨店w/7 others),

- 「何処で買ったか」は，(@ 〇〇百貨店～箇所を抽出、
- 残りは手動で抽出

32

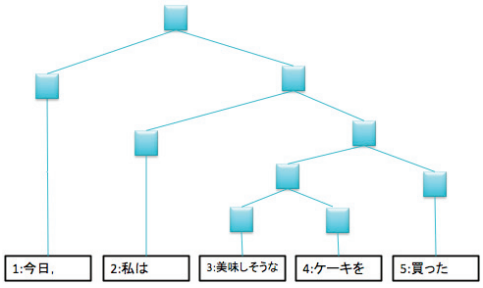
②「何を買ったか」情報

③【ツイート内容】〇〇百貨店でケーキを買った．(@ 〇〇百貨店w/7 others),

- 「何を買ったか（例：ケーキ）」は，ツイート内容を茶筅を使って形態素解析したうえで、ルールベースの係り受け解析にて抽出

33

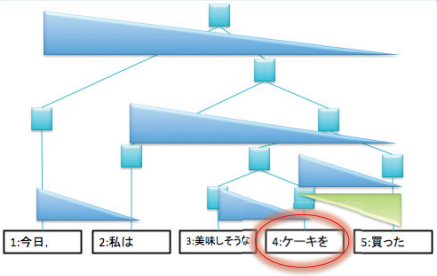
②係り受け解析



各単語（or文節）の係先を解析する手法

34

②ルールベースによる係り受け解析



- 動詞との組み合わせで検索することにより「目的語」を抽出したり，
- 助動詞及び動詞との組み合わせで検索することにより「名詞」や「固有名詞」を抽出

35

③対応分析

36

「何処で何を買ったか」が類似する地域を，
引用関係があると見なしてクラスタリング

Buy_Place	City
家電_家電量販店	東京都新宿区
家電_家電量販店	埼玉県朝霞市
服_百貨店	大阪府大阪市
ケーキ_百貨店	千葉県市川市
ケーキ_スーパー	大分県大分市
ケーキ_百貨店	東京都足立区
生活用品_スーパー	秋田県北秋田市

37

③対応分析

統計ソフトRで一般に用いられているVanables[*]の手法を参考に，多重対応分析（MCA）を応用する形で類似尺度を準備

[*] Vanables, W.N. and Ripley, B.D.: Modern Applied Statistics with S Fourth edition, Springer, 2002.

38

③出発行列

$$Z_{m\times n} = (z_{ij})\ (f_{ij} \geq 0, i \in I, j \in J)$$

$I=\{1,2,\cdots,m\},\ J=\{1,2,\cdots,n\}$

（m×n）の2元クロス表を考える．
IとJはそれぞれ行と列の項目の選択肢の集合を表す．

39

③対応分析の成分スコア

$$z_{ik} = (i \in I, k = 1,2,\cdots K)$$

行，いわゆる「何処で何を買ったかノードi」に対する第k成分のスコア

$$z_{jk} = (j \in J, k = 1,2,\cdots K)$$

列，いわゆる「地域ノードj」に対する第k成分のスコア

40

③対応分析の成分スコア

成分スコア（z_{ik}，z_{jk}）を
クラスタリングの行列座標として使用する．

41

④クラスタリング

今回は階層クラスタリングで説明

42

④階層クラスタリング

2変数の場合，
対象iとj間のユークリッド距離を求め，この距離が近いものを結合してクラスタを生成していく，

43

④階層クラスタリング

次に，クラスタ間の距離を測定し，
最も距離の近い2つのクラスタを逐次的に結合する．

44

④-3階層クラスタリング

この結合を，全ての対象が1つのクラスタに併合されるまで繰り返すことで階層構造を獲得

45

④-1ユークリッド距離の算出

2変数の場合 $d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$

多変量にした場合 $d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$

46

④-1ユークリッド距離の算出

多変量にした場合 $d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$

↓ 対応分析の成分スコア，z_{ik} とz_{jk}を適応

$$d_{ij} = \sqrt{\sum_{s=1}^{n} (z_{ik_s} - z_{jk_s})^2}$$

47

④-1ユークリッド距離の算出

このようにすることで，
最も対応関係のあるノード同士からクラスタ化されていくようになる．

48

④-2クラスタ間の距離の測定

最短距離法 $d(C_1, C_2) = \min_{x_1 \in C_1, x_2 \in C_2} d(x_1, x_2)$

最長距離法 $d(C_1, C_2) = \max_{x_1 \in C_1, x_2 \in C_2} d(x_1, x_2)$

群平均法 $d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x_1 \in C_1} \sum_{x_2 \in C_2} d(x_1, x_2)$

ウォード法 $d(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2)$

49

④-2クラスタ間の距離の測定

ウォード法以外の手法は，重心を使うことで，クラスタ結合後に距離が短くなってしまう場合もあり，距離の単調性が保証されるわけではない．

↓

本研究ではウォード法を採用

50

④-2クラスタ間の距離の測定

$$d(C_n^1, C_m^2) = E(C_n^1 \cup C_m^2) - E(C_n^1) - E(C_m^1)$$

- ユークリッド距離が近いiとj結合して生成するクラスタのうち，隣接するものをC^1_1, C^1_2, …C^1_n及びC^2_1, C^2_2, …C^2_nとして表現
- これをウォード法のC_1, C_2箇所に適応

51

以上の対応分析及びクラスタリングをRで試行

52

元データの例

city	place	buy
愛知県名古屋市	スーパー	衣服
愛知県名古屋市	スーパー	ワイン
茨城県つくば市	スーパー	お菓子
茨城県つくば市	モール	衣服
埼玉県鴻巣市	モール	衣服
広島県広島市	衣料量販店	衣服
広島県広島市	衣料量販店	衣服
愛知県長久手市	洋菓子店	ケーキ
東京都板橋区	コンビニ	ケーキ
東京都渋谷区	モール	衣服
東京都渋谷区	モール	アクセサリー

↓ クロス表で表現すると

	コンビニ	スーパー	モール	衣料量販店	洋菓子店	アクセサリー	お菓子	ケーキ	ワイン	衣服
愛知県長久手市	0	0	0	0	1	0	0	1	0	0
愛知県名古屋市	0	2	0	0	0	0	0	0	1	1
茨城県つくば市	0	1	1	0	0	0	1	0	0	1
広島県広島市	0	0	0	2	0	0	0	0	0	2
埼玉県鴻巣市	0	0	1	0	0	0	0	0	0	1
東京都渋谷区	0	0	2	0	0	1	0	0	0	1
東京都板橋区	1	0	0	0	0	0	0	1	0	0

↓ グラフで表現すると

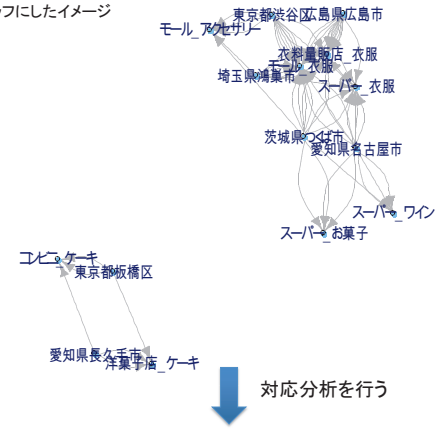
【補足】2部モジュラリティー拡張エッジ

元データ

city	place	buy
愛知県名古屋市	スーパー	衣服
愛知県名古屋市	スーパー	ワイン
茨城県つくば市	スーパー	お菓子
茨城県つくば市	モール	衣服
埼玉県鴻巣市	モール	衣服
広島県広島市	衣料量販店	衣服
広島県広島市	衣料量販店	衣服
愛知県長久手市	洋菓子店	ケーキ
東京都板橋区	コンビニ	ケーキ
東京都渋谷区	モール	衣服
東京都渋谷区	モール	アクセサリー

- 「愛知県名古屋市_スーパー_衣服」と「茨城県つくば市_スーパー_ワイン」が"スーパー"で繋がった時に，"ワイン"繋がりの，「愛知県名古屋市　スーパー　ワイン」も同じコミュニティに追加する。
- つまり，他のbuy,placeエントリのどちらかに一致した場合、そのエントリの組にcityからのリンクを追加する。

グラフにしたイメージ



対応分析

57

クロス表から対応分析を行った例

```
> #  対応分析.
> #  列の得点
> corr$cscore
                  [,1]         [,2]
コンビニ       2.1213203 -5.970967e-16
スーパー      -0.4714045  1.677030e+00
モール        -0.4714045 -3.512214e-01
衣料量販店    -0.4714045 -1.813102e+00
洋菓子店       2.1213203  1.777265e-16
アクセサリー -0.4714045 -7.594381e-01
お菓子        -0.4714045  7.024427e-01
ケーキ         2.1213203  1.303394e-17
ワイン        -0.4714045  2.164324e+00
衣服          -0.4714045 -3.512214e-01
```

```
> #  行の得点
> corr$rscore
                  [,1]         [,2]
愛知県長久手市  2.1213203  2.958228e-31
愛知県名古屋市 -0.4714045  1.672080e+00
茨城県つくば市 -0.4714045  5.426826e-01
広島県広島市   -0.4714045 -1.400739e+00
埼玉県鴻巣市   -0.4714045 -4.546173e-01
東京都渋谷区   -0.4714045 -5.867152e-01
東京都板橋区    2.1213203  0.000000e+00
>
> #  正準相関係数
> corr$cor
[1] 1.0000000 0.7725649
>
> # 寄与率
> round(100* eig/sum(eig),2)
[1] 62.62 37.38
```

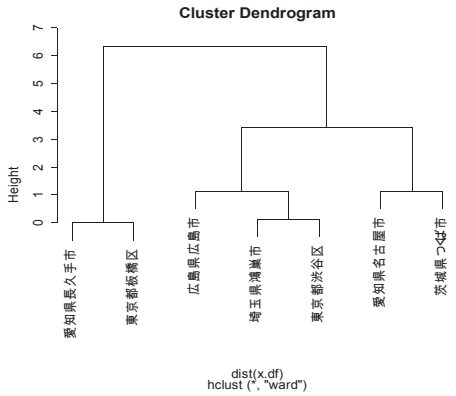
階層クラスタリング

60

- 行のマージ結果に対して、クラスタ数または高さ(upperTail法)を指定して分割
- 下記は，前項によりcityのクラスタ分割を行った結果

```
> (stepj <- StepByUpperTail(x.hc,k=1))
[1] 4.461264
[1] 6
> corr.members <- cutree(x.hc,h=x.hc$height[stepj])
> corr.members
愛知県長久手市 愛知県名古屋市 茨城県つくば市　広島県広島市　埼玉県鴻巣市　東京都渋谷区　東京都板橋区
             1              1              1              1              1              1              1
```

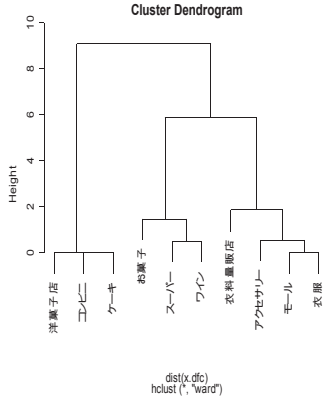
行の得点から距離行列－＞階層的クラスタリング



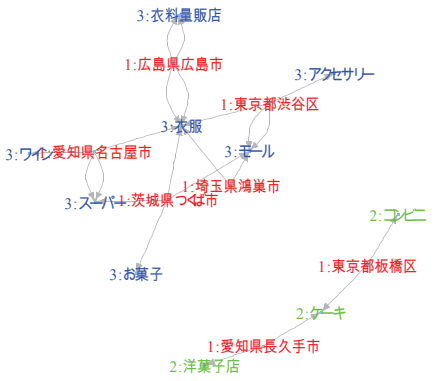
- 列のマージ結果に対して、クラスタ数または高さ(upperTail法)を指定して分割
- 下記は，前項によりplaceとbuyのクラスタ分割を行った結果

```
> (stepj2 <- StepByUpperTail(x.hcc,k=1))
[1] 5.338775
[1] 8
> corr.col.members <- cutree(x.hcc, h=x.hcc$height[stepj2])
> corr.col.members
コンビニ スーパー モール 衣料量販店 洋菓子店 アクセサリー お菓子 ケーキ ワイン 衣服
      1      2      2      2      1      2      2      1      2      2
```

列の得点から距離行列－＞階層的クラスタリング



対応分析－＞階層クラスタリング& upperTail による分割



しかし，

upperTail法ではkというパラメータ予め指定しなければならない．

例）
高さ平均 ＋ k ＊ 高さ標準偏差

ゆえに，

- 対応分析までは手法1と同じだが、
- 行の得点と列の得点に対し、upperTail法ではなくxmeans法によりクラスタリングを行う。

x-means法

k-means法を使って2分割を繰り返すが、分割の基準としてBIC（ベイズ情報量基準）を使って最適な分割数を求める。

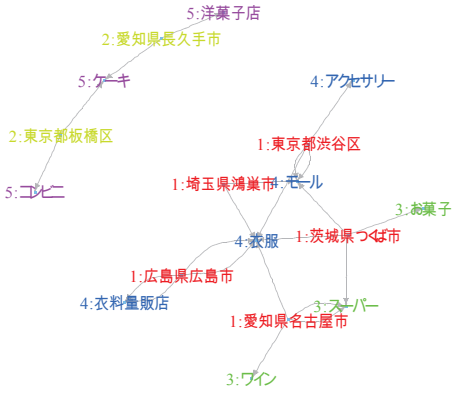
予めパラメータを指定する必要がない

【Rの実装例】

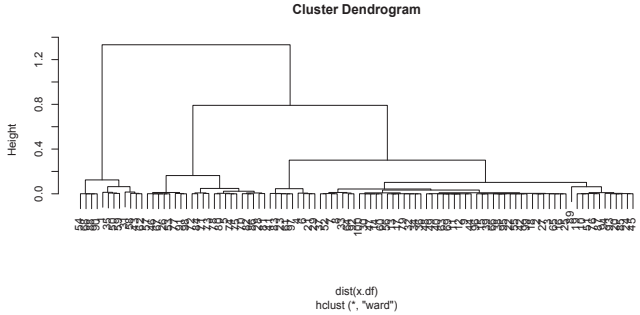
```
x.xmr <- xmeans(as.matrix(dist(x.df)),merge.cls=T)
  #列についてmerge.cls=Tにするとエラーがでるので、重要性はないのでFにしておく

x.xmc <- xmeans(as.matrix(dist(x.dfc)),merge.cls=F)
```

対応分析 ー > xmeansクラスタリングによる分割

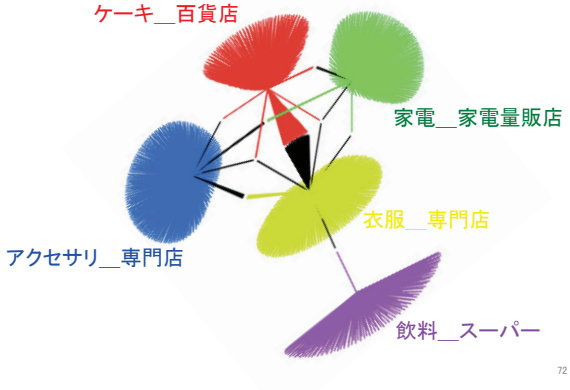


購買行動の似た場所クラスタイメージ



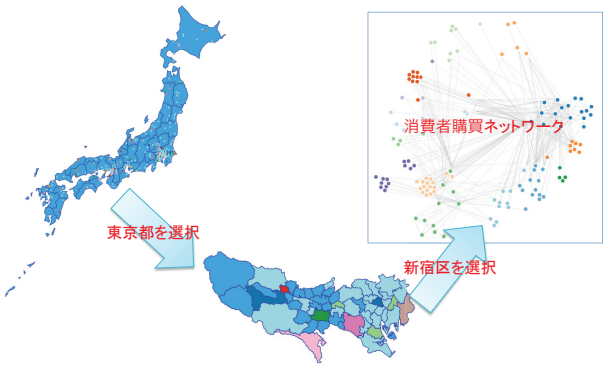
71

⑤購買行動の似た場所クラスタイメージ



72

⑤購買行動の似た場所クラスタの可視化



73

評価実験

階層クラスタリング版

74

評価実験1

75

評価実験の概要

① 原田の手法と比較

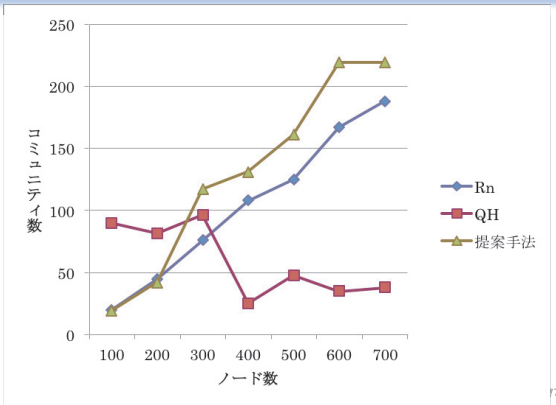
② 実際のSNS2部グラフデータを使用

③ 予め正しいコミュニティ(クラスタ)分割数をR_nを定義

④ ノードを増やしながらR_nの変遷と比較検証

76

評価実験1(コミュニティ数推移)



77

評価実験1の結果の考察

- Q^HはModularity Qに基づいている
- Modularity Qは，同じような参照が増えてくればまとまっていくという特徴を持つため
- 結果，より少ないコミュニティ数で分割がなされることになる．

78

評価実験2

79

評価実験2（t検定）

帰無仮説を，

「提案手法と原田の手法間に差がない」

と仮定

80

評価実験2（t検定）の結果

- それ以下ならば帰無仮説が棄却される有意水準5%において，
- t検定の結果は0.046<0.05であり，統計学的な有意差が認められた

81

まとめ

- 先行研究：コミュニティ形成時の類似尺度としてリンクの構造に着目して2部グラフをクラスタリング
- 本提案：コミュニティ形成時の類似尺度としてリンクの対応性に着目して2部グラフをクラスタリング
- 評価実験：評価実験の結果，先行研究からの有用性を確認

82

本発表の関連技術を著書として出版

http://www.amazon.com/Big-Data-Analysis-Akira-Otsuki/dp/3659383066