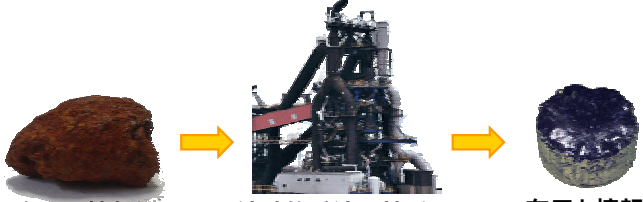
2012年10月25日　地域経済情報研究所研究会

データマイニングの手法を用いたカリキュラムに関する情報の提示効果の分析

経営学部　浮穴　学慈

データマイニングとは

- マイニング (mining) ：採鉱
- データの鉱山：DBに蓄積された大量のデータ
- 非自明で有用な情報を抽出 (統計的手法)
  - 採鉱より精錬のイメージの方が適切かも



有用な情報を含有するデータ　→　統計的手法に基づく大量のデータ処理　→　有用な情報

2

データマイニングの用途例

- 例1
  - 鉱山：小売店の販売データ (POS)
  - 情報：ビールを買う客は、紙オムツを買う (相関)
- 例2
  - 鉱山：通販サイトの閲覧履歴
  - 情報：購買につながる可能性が高い広告の提示
- 例3
  - 鉱山：クレジットカードの利用履歴
  - 情報：クレジットカードの不正利用の兆候の発見

3

[例]アソシエーション分析

- 次の買い物バスケットデータのなかで、共起性の高いアイテムの組み合わせは何か？

トランザクションID	アイテム集合
1	パン, 牛乳, ハム, 果物
2	パン, オムツ, ビール, ハム
3	ソーセージ, ビール, オムツ
4	弁当, ビール, オムツ, タバコ
5	弁当, ビール, ジュース, 果物

4

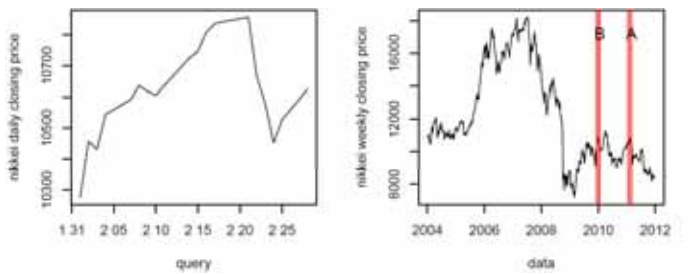
[例]テキストマイニング

- 次の文章のなかで、与えられたキーワードに関して、類似性の高い組み合わせはどれか？
  - A) まだあげ初めし前髪の　林檎のもとに見えしとき　前にさしたる花櫛の　花ある君と思ひけり
  - B) 赤い林檎に　唇よせて　だまって見ている　青い空　林檎はなんにも　いわないけれど　林檎の気持はよくわかる
  - C) バナナが一本ありました　青い南の空の下　子供が二人でとりやっこ　バナナはツルンと飛んでった
- キーワード＝林檎、空

5

[例]時系列分析

- 日経平均週次終値データから、ある特徴を持った部分を抽出する



DTWアルゴリズムSPRING（櫻井,2004）を用いて、浮穴が抽出

6

今回のテーマ

- 2011年度経営学科各コースモデル履修プログラム（履修ガイド掲載：資料）は、どの程度の情報提示効果があったのか？
  - 以下では、各コースモデル履修プログラム掲載科目を、次のように呼ぶ
    - 経営モデル科目、情報モデル科目、会計モデル科目
- このテーマの難しい点
  - 対照実験が出来ない
  - ドンピシャの先行研究が無い
  - アンケート調査の場合のバイアス除去

7

情報モデル科目

- 資料：2011年度履修ガイド
  - 経営情報コース (p.3)
  - 卒業要件 (pp.7-10)

8

A君の2セメ履修科目

科目コード	科目名	モデル科目
101310	くらしと経済	○
101320	計算機概論	○
101322	情報活用演習 II	○
101324	情報処理演習 II	○
101325	統計入門	○
101333	口語表現演習 II	○
103352	プラクティカル･イングリッシュ II	
103602	中国語 II	
104301	キャリア開発	○
104306	ビジネス実務概論 I	○
104308	ビジネス実務演習 I	○
104351	企業観察実習	○
104360	簿記演習 II	○
104361	企業論	○
105308	ビジネスの人間関係	○
136402	経営史	
136952	基礎演習 II	

9

情報の類似性を測るには

- 距離
  - ユークリッド距離
  - マンハッタン距離
  - ミンコフスキー距離
  - マハラノビス距離
- 情報量
  - 相互情報量
  - 条件付きエントロピー
- 類似度
  - コサイン類似度
  - ピアソンの相関係数
  - 偏差パターン類似度
  - ジャッカード係数
  - ダイス係数
  - シンプソン係数

10

コサイン類似度

- ベクトル空間内の、2つのベクトル間の角度
  - ベクトル内積や余弦定理に関係あり

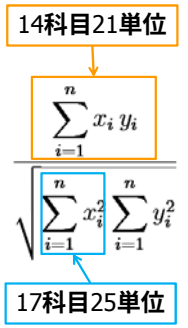
n 次元ベクトル空間における，ベクトル $\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i$, $\mathbf{y} = \sum_{i=1}^{n} y_i \mathbf{e}_i$ の間の類似度は，次のように定義される．

$$\text{コサイン類似度}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

11

コサイン類似度の計算

科目コード	科目名	モデル科目
101310	くらしと経済	○
101320	計算機概論	○
101322	情報活用演習 II	○
101324	情報処理演習 II	○
101325	統計入門	○
101333	口語表現演習 II	○
103352	プラクティカル･イングリッシュ II	
103602	中国語 II	
104301	キャリア開発	○
104306	ビジネス実務概論 I	○
104308	ビジネス実務演習 I	○
104351	企業観察実習	○
104360	簿記演習 II	○
104361	企業論	○
105308	ビジネスの人間関係	○
136402	経営史	
136952	基礎演習 II	○



12

計算に何を使う？

- 科目数 or 単位数
- モデル科目全体 or 2セメ限定

14科目21単位

$$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

17科目25単位

y	科目数･単位数
情報モデル科目　全体	89科目139単位
情報モデル科目　2セメ限定	16科目24単位
カリキュラム　全体	184科目297単位
カリキュラム　2セメ限定	44科目70単位
1年生平均	15.9科目24.3単位
1年生（2年次「情報」選択）平均	16.1科目24.9単位

- 科目数＆情報モデル科目全体　を採用
  - 科目数と単位数の相関係数 (0.92)
  - 解析的な計算が簡単
  - 追跡がし易い

13

計算に何を使う？

- 科目数
- ベクトル空間の次元=89

14科目21単位

$$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

17科目25単位

y	科目数･単位数
情報モデル科目　全体	89科目139単位
1年生平均	15.9科目24.3単位
1年生（2年次「情報」選択）平均	16.1科目24.9単位

A君のコサイン類似度

=14/√17√89

=0.35

14

たまたま履修する可能性①

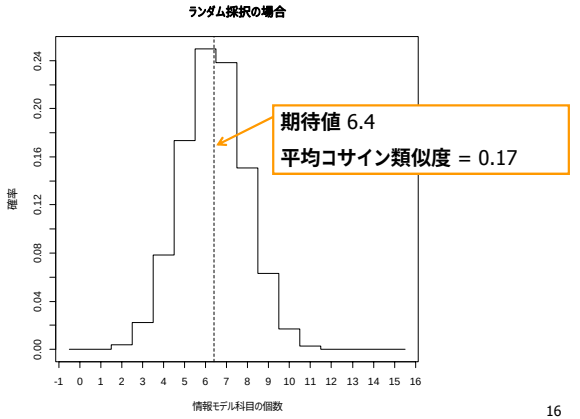
- ランダムに履修するとどうなるか

2 セメスタ配当の科目が N 科目 $(N = 44)$ 存在し，そのうち情報コースのモデル履修プログラム掲載の科目が n 科目 $(n = 16)$（以下，情報モデル科目），M 科目 $(M = 16)$ を履修したなかに，情報モデル科目が m 科目含まれる確率は，

$$Pr(m) = \frac{\binom{n}{m}\binom{N-n}{M-m}}{\binom{N}{M}}$$

15

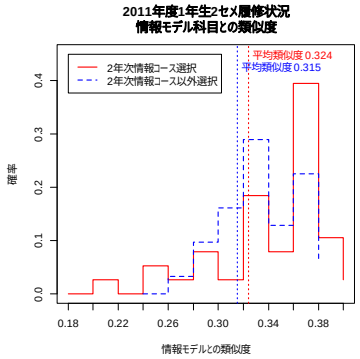
たまたま履修する可能性②



16

傾向がハッキリ

- 情報コース志向学生は、情報モデル科目を履修



17

今後の課題

- 各コース先修科目図（科目系統図）の情報提示効果の測定
  - 実はこちらが本命

18