JSAI 2011 全国大会 in 盛岡
2G3 データマイニング「頻出パターン」
06月02日(Thu) 14:45～17:00 G会場

モジュラリティの差異に基づくコントラスト法
(Patterns with Emerging Modularity and their Detection)

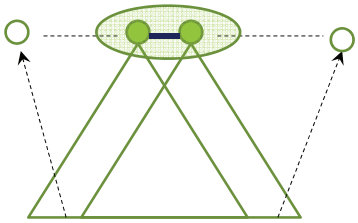
北海道大学 大学院情報科学研究科
鶴田 哲章 (現在 NTTコムウエア) 、原口 誠

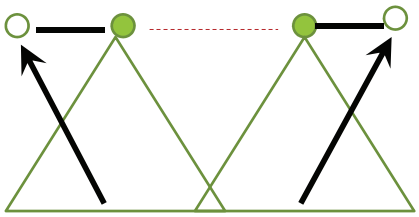
背景: マイニング分野における変化の抽出

- 2つ(以上)のDBを比較:
  パターンの頻度(アイテムの共起)の違いを検出
  - Emerging Pattern(G.Dong et al '99)
  - Contrast Set(S.D.Bay et al '01, P.K.Novak et al '09)
  - lift ($p(A|B)/p(A)$) の差異(比率) (Taniguchi et al '06)
- 相関マイニング (Brin '98): 独立な場合との差異 (統計量、情報量)
  相関の差異を検出 (Li et al '11)
- 今回: **必然性(in Newman clustering) の差異**
  単なる共起頻度・相関ではなく、稀なものの結合を重要視
  抽出目標: base と target を比較し、必然性が増加するもの

共起の必然性の差が大きくなる場合

必然性大 ー必然性小





低次数の項の共起
共起しにくい項の共起
比較的に小規模のクラスタ抽出

高次数の項の低い共起
共起しやすい項が共起していない
or 例え共起していても有り難がらない

ボランティア代表者の被災者の生活の調査
話題が特定化され、次数が低くなる

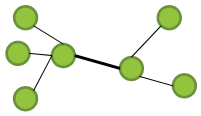
家,代表,生活,住宅

?

様々な話題で使用され、次数が高い

モデュラリティ行列

m 回、「ランダムに」辺を生成。m：辺の総数
ただし、ランダムグラフ s.t.
　頂点周りの辺期待値の総和 ＝ 現実の次数
（次数保存性＋独立性）k_i k_j /2m が導出



$$B = A - P = (A_{ij} - P_{ij})_{ij}$$
$$A = D^t D, \quad P_{ij} = k_i k_j / 2m$$

D：文書ー項（ブール値）、A：共起頻度

低次数の共起：　ランダム性から遠くなり必然。regard
高次数の共起：　当たり前。予想できるから disregard

Newmanモジュラリティの利用に向けて

- Newman(クラスタリング)の場合
  - 目的は全ネットワークのグラフ分割
  - 分割を評価する関数Qを定め、モヂュラリティ行列 B のスペクトル分解に基づく、近似的最適クラスタリングの計算
- 本研究
  - 目的はクラスタリングではなく、モジュラリティが増加するパターンをマイニングする
    - モジュラリティの差異を計算
    - 単一の準最適なパターンではなく、可能性のあるパターンを枚挙

Newmanによる分割の評価(‘06)

$$\mathbf{S} = (\mathbf{s}_1 \quad \cdots \quad \mathbf{s}_c)$$

各グループの評価　　グラフ分割の評価

$$\mathbf{s}^T\mathbf{B}\mathbf{s} = \sum_{i,j} B_{ij}s_i s_j = \sum_{s_i=s_j=1} B_{ij} \qquad Q = \sum_{k=1}^{c} \mathbf{s}_k^T\mathbf{B}\mathbf{s}_k = \mathrm{Tr}\left(\mathbf{S}^T\mathbf{B}\mathbf{S}\right)$$

$$x^T\mathbf{B}x = \lambda_1 z_1^{\,2} + .. + \lambda_p z_p^{\,2} - \beta_1 z_{p+1}^{\,2} - .. - \beta_{n-p} z_n^{\,2}$$

$$\mathbf{s}^T\mathbf{B}\mathbf{s} = \lambda_1(u_{11}+u_{21})^2 + .. - \beta_1(u_{13}+u_{23})^2 - ...$$

$$s = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

	λ_1	λ_2	$-\beta_1$	$-\beta_2$
v_1	u_{11}	u_{12}	u_{13}	u_{14}
v_2	u_{21}	u_{22}	u_{23}	u_{24}
v_3	u_{31}	u_{32}	u_{33}	u_{34}

$|\mathbf{X}_k|^2 - |\mathbf{Y}_k|^2$　固有値の平方根で補正したベクトルの2乗ノルムの和．正と負のベクトル

2つのDBにおける「モジュラリティの差異」

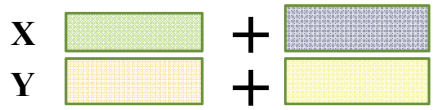
モジュラリティの差

$$B_{ij} = \left(A'_{ij} - \frac{k'_i k'_j}{2m'} \right) - \left(A_{ij} - \frac{k_i k_j}{2m} \right)$$

$$s = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

	λ_1	λ_2	β_1	β_2
v_1	$\sqrt{\lambda_1}u_{11}$	$\sqrt{\lambda_2}u_{12}$	$\sqrt{-\beta_1}u_{13}$	$\sqrt{-\beta_2}u_{14}$
v_2	$\sqrt{\lambda_1}u_{21}$	$\sqrt{\lambda_2}u_{22}$	$\sqrt{-\beta_1}u_{23}$	$\sqrt{-\beta_2}u_{24}$
v_3	$\sqrt{\lambda_1}u_{31}$	$\sqrt{\lambda_2}u_{32}$	$\sqrt{-\beta_1}u_{33}$	$\sqrt{-\beta_2}u_{34}$

s に対し、



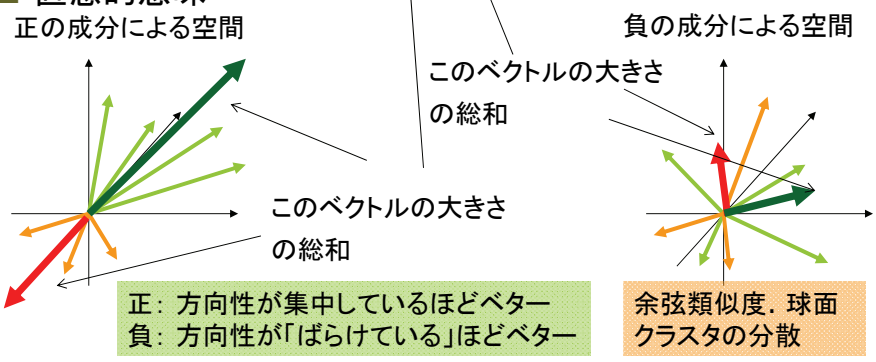
パターン評価のためにグループ評価法を借用

グループの評価 $= |\mathbf{X}_k|^2 - |\mathbf{Y}_k|^2$

正の成分での大きさ

負の成分での大きさ

- 直感的意味



正：方向性が集中しているほどベター
負：方向性が「ばらけている」ほどベター

余弦類似度．球面クラスタの分散

予備実験用ストラテジー

- Pattern 生成と評価:
  縦型、set enumeration tree
  |正のベクトル|^2 − |負のベクトル|^2
  アイテムの追加に対し、非単調

  制約: 負の成分に対し、
  中心ベクトルとの余弦類似度の総和が一定の値以下

  アイテムの追加に対し**安全でない制約**: 要改良

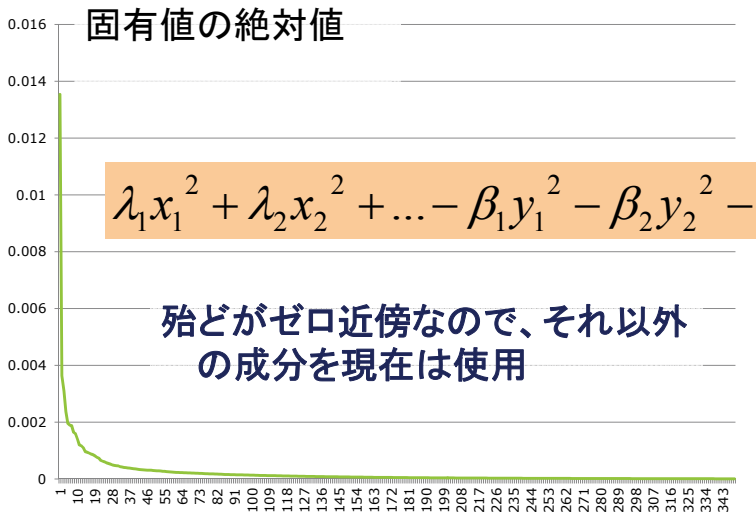
  幅限定探索: 制約を満たす候補アイテムで、正の 成分評価がトップN個に絞り込む

  制約を満たすものの中でトップN個を探索する、
  分枝限定トップN枚挙とは異なる。
  本実験では 動的順序付けのもとで、最適トップN解のみを枚挙。

実験

- 用いたデータについて
  - 毎日新聞の「神戸」を含む1994年の文書集合(DB)と1995年の文書集合(DB')を用いて実験
  - DBの文書数:2337
  - DB'の文書数:9324
  - DB・DB'に共通して出現する名詞(アイテム)

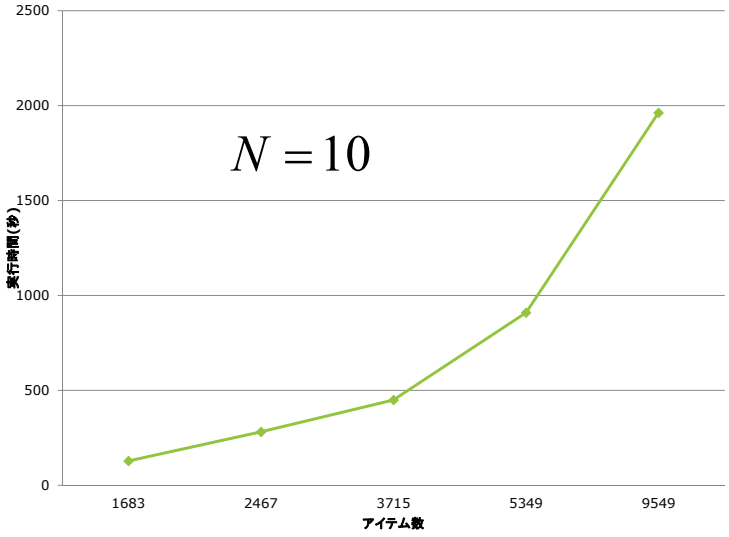
成分の選択について



$$\lambda_1 x_1^{\,2} + \lambda_2 x_2^{\,2} + \ldots - \beta_1 y_1^{\,2} - \beta_2 y_2^{\,2} - \ldots$$

殆どがゼロ近傍なので、それ以外の成分を現在は使用

計算時間



品質

得られたパターン	モジュラリティの差	支持度の差	反映していた(DB'での)文書の概要
JR,次,地震	0.0002465	0.0068673	地震による被害を受けた、各種交通機関の復旧
伊丹,南,明石,西宮,須磨,中央	0.0002072	0.0002178	
伊丹,南,明石,宝塚,尼崎	0.0001887	0.001397556	
決定,発生,被害	0.0000590	0.0006479	地震による被害と、その地域の再開発計画の決定
家,代表,生活,住宅	0.0000535	0.0009686	ボランティア代表者の被災者の生活の調査
滋賀,明石,宝塚	0.0000448	0.0037571	各種交通機関の復旧
システム,全国,被害	0.0000439	0.0030063	地震による被害によって表面化した、行政システムの問題
南,明石,宝塚	0.0000436	0.0022567	各種交通機関の復旧
伊丹,南,明石	0.0000430	0.001507	各種交通機関の復旧
センター,会場,中央	0.0000422	0.0037571	チャリティーコンサートの会場案内

- 必然性(modularity)の差が大きいもの
- 差は小さすぎて emerging pattern としては実質的に抽出不能

今後の課題

- 探索の完全性の観点から
  - 負の成分ベクトルに対する枝刈りの安全性は NG
    アイテム追加により分散は単調に増加するとは限らない。
  - 他の安全性を保証する制約に替える as 分散が一定以内の部分を含まない、等
  - 正の成分ベクトル：論文では枝の本数を限定（高々N個）
    動的順序で、分枝限定枝刈りOK（concept vector を基準にした、）
- 成分選択指針： ゼロ近傍が多数でないとき
- 今回：１部グラフモデュラリティ。２部グラフ版がより正確.