符号付き２部グラフを用いた同義表現抽出の現状

-格フレームの極性付き概念化を用いたセンチメント分析-

北野 道春　同志社大学大学院文化情報学研究科
谷岡 健資　同志社大学大学院文化情報学研究科
宿久 洋　同志社大学文化情報学部

ERATO合宿 in 札幌

dil0017@mail.doshisha.ac.jp

2011/11/1

発表構成

- はじめに
- 符号付きグラフ
- 重複クラスタリング
  - 先行研究:Greedy Clique Expansion (GCE)
  - 提案手法:Greedy Signed Clique Expansion (GSCE)
  - 極性付き概念抽出
- まとめ

dil0017@mail.doshisha.ac.jp

はじめに

- テキストマイニング
  - テキストから情報を抽出

テキスト

将来の暮らしについての
インターネット調査
年金制度に不安が広がる.

・・・

単語，フレーズ

はじめに

- テキストマイニング
  - テキストから情報を抽出

テキスト

将来の暮らしについての
インターネット調査
年金制度に不安が広がる.

・・・

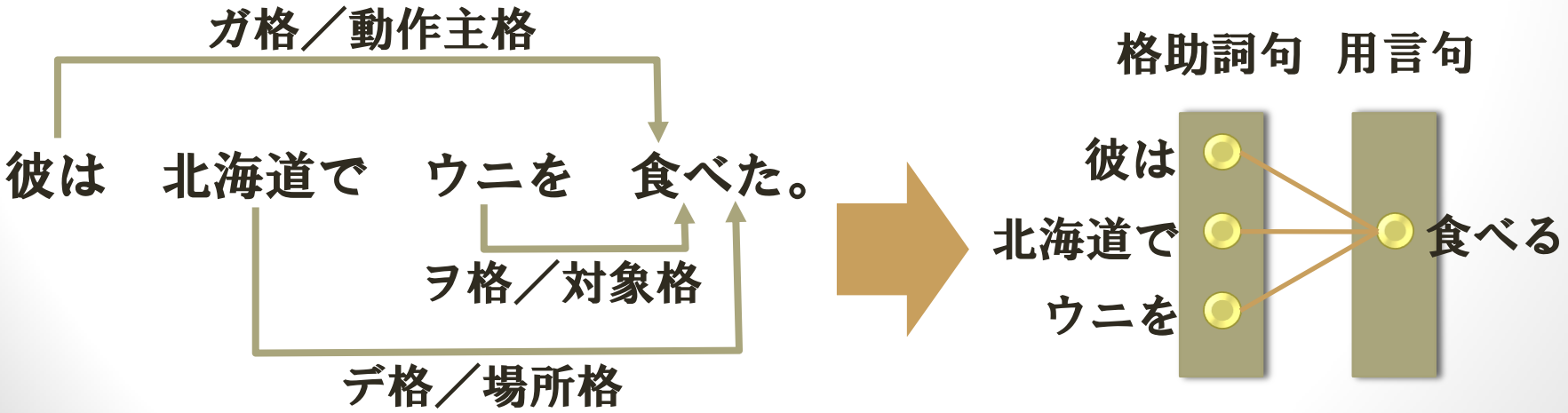
単語，フレーズ

	将来	不安	広がる	…
テキストA	1	0	0	…
テキストB	0	1	1	…
テキストC	0	0	0	…
⋮	⋮	⋮	⋮	⋮

dll0017@mail.doshisha.ac.jp

はじめに

- テキストマイニング
  - 係り受け解析
    - 文の意味構造を「動詞–深層格–名詞」という関係の集合で捉える

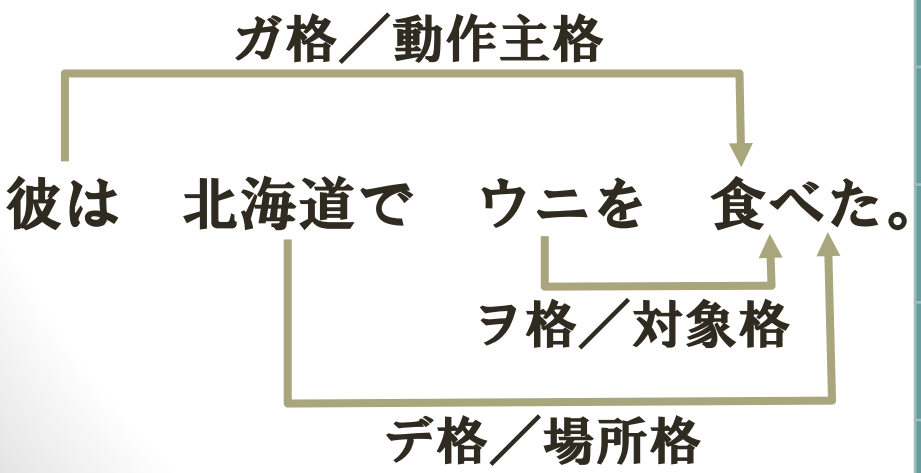


※ 小室，岡田先生のスライドを参考

はじめに

- テキストマイニング
  - 係り受け解析
    - 文の意味構造を「動詞–深層格–名詞」という関係の集合で捉える

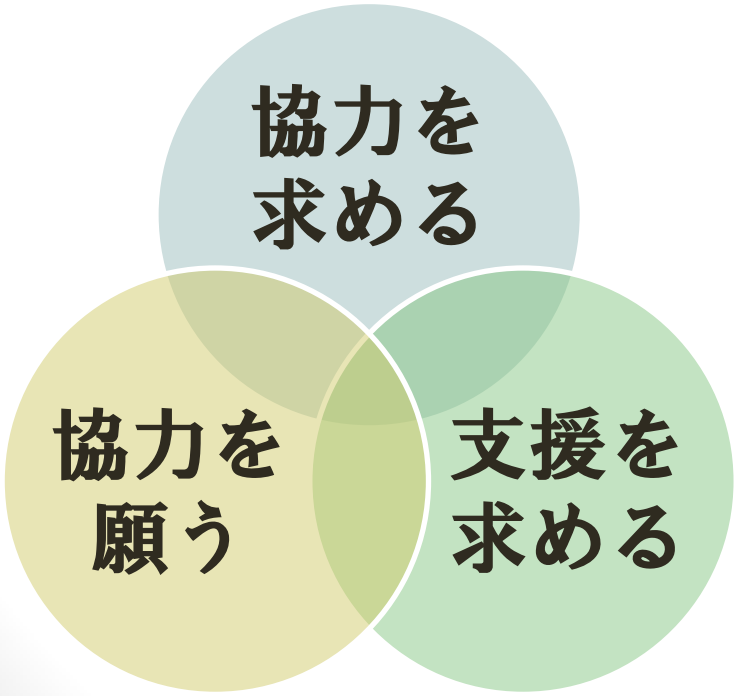
格フレーム



	彼は 食べる	北海道で 食べる	ウニを 食べる	…
テキスト A	1	1	1	…
テキスト B	0	0	1	…
テキスト C	0	1	0	…
⋮	⋮	⋮	⋮	⋮

はじめに

- テキストマイニング
  - 類義語・同義語の問題

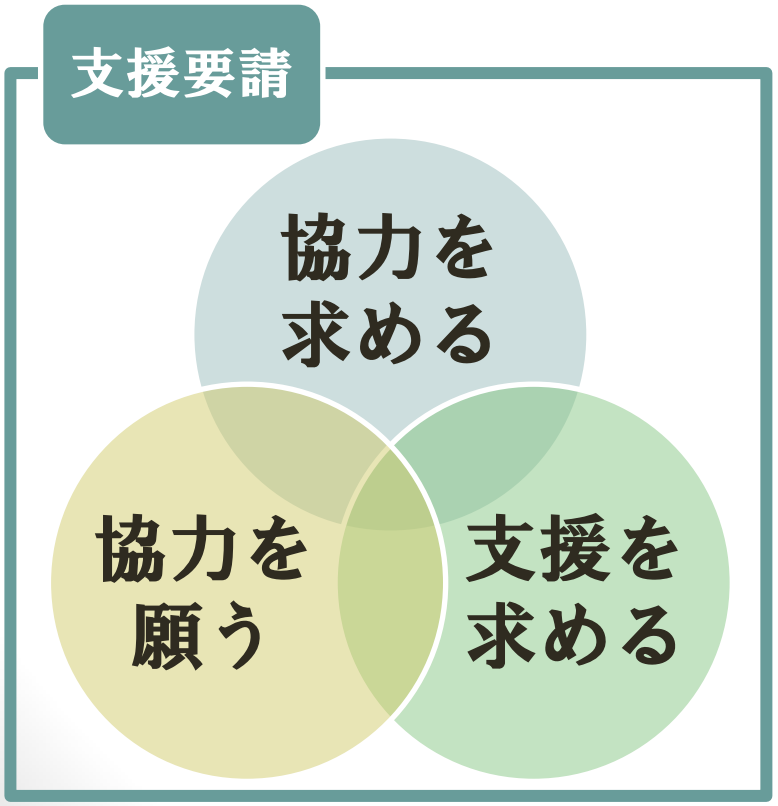


1. 変数の増加
2. スパース
3. モデルへの適合

dll0017@mail.doshisha.ac.jp

はじめに

- テキストマイニング
  - 類義語・同義語の問題



主成分分析

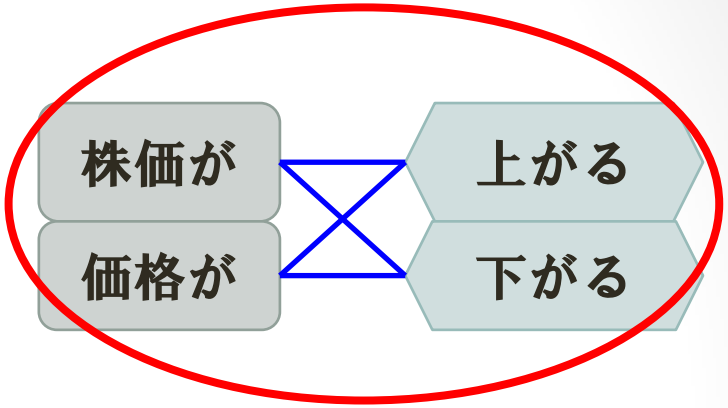
潜在クラスモデル

グラフ理論を用いた重複クラスタリング

dil0017@mail.doshisha.ac.jp

はじめに

グラフ理論を
用いた重複
クラスタリング



上野，他 (2004)
相澤，中渡瀬 (2007)
羽室，岡田先生によるERATO合宿での発表 (2010)

- 単語の係り受け関係に関する情報を使う．
- 単語の意味(極性)は考慮していない．

dil0017@mail.doshisha.ac.jp

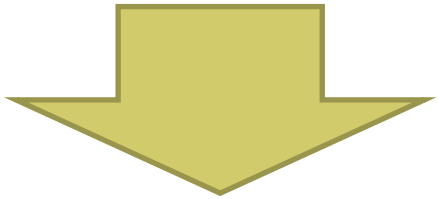
はじめに

- 極性
  - その単語（格フレーム）がポジティブかあるいは，ネガティブかを表す指標.
  - 羽室，岡田先生によって，周辺文脈法(那須川・金山2004)を利用した極性付き格フレームを得る方法が提案されている.

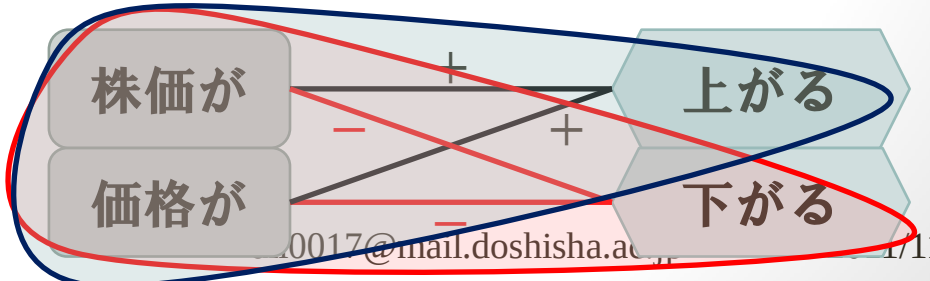
dil0017@mail.doshisha.ac.jp

はじめに

- 本発表の目的
  - 極性付き格フレームが与えられたとき，その**極性を考慮した**同義表現抽出を行う．



  - 極性付き格フレームを符号付き２部グラフを用いて表し，符号を考慮した重複クラスタリング手法を提案する．



0017@mail.doshisha.ac

符号付きグラフ

dil0017@mail.doshisha.ac.jp

符号付きグラフ

- 符号付きグラフ
  - グラフ内の全ての辺に，positive (＋) もしくは negative (－) という符号が付与されたグラフ

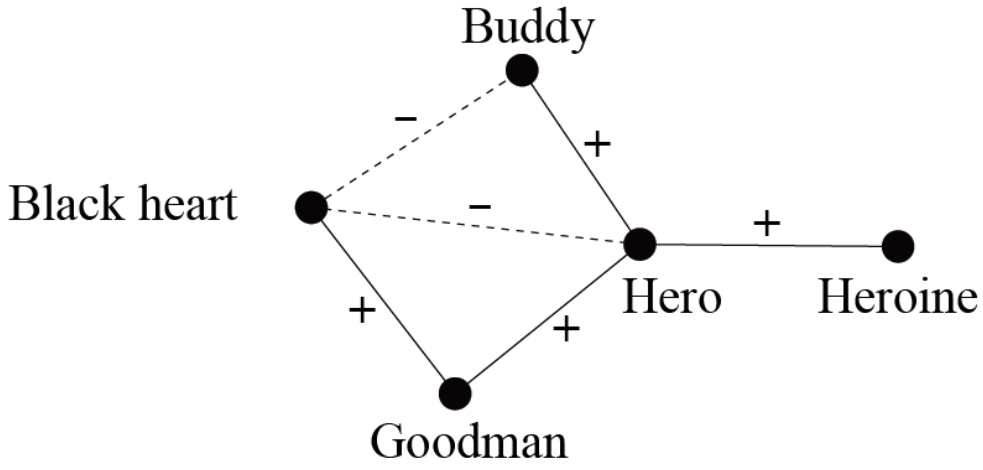


図１：符号付きグラフの例

dl0017@mail.doshisha.ac.jp

符号付きグラフ

- 符号付きグラフ
  - グラフ内の全ての辺に，positive (＋) もしくは negative (－) という符号が付与されたグラフ

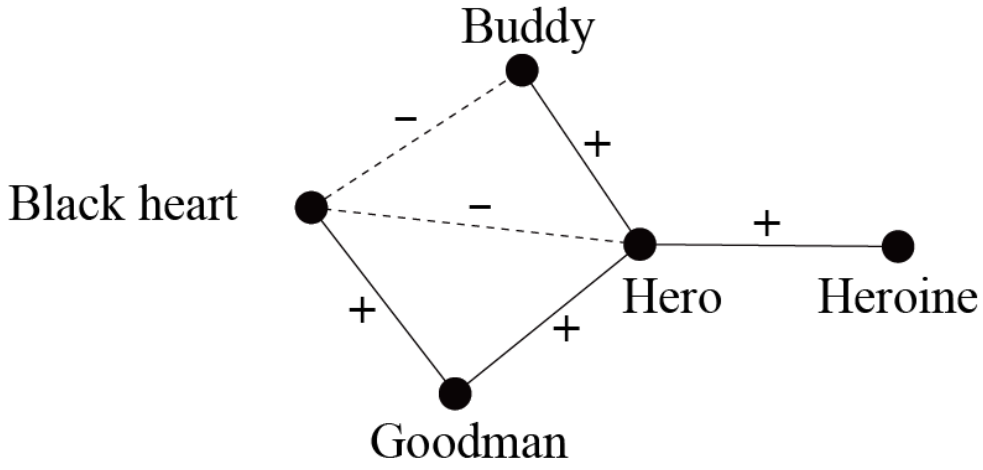


図１：符号付きグラフの例

符号付きグラフ

- **符号付きグラフ**
  - **グラフ内の全ての辺に，**positive (＋)**もしくは** negative (－) **という符号が付与されたグラフ**

$$\boldsymbol{A} = \left(a_{ij}\right)_{(n\times n)}$$

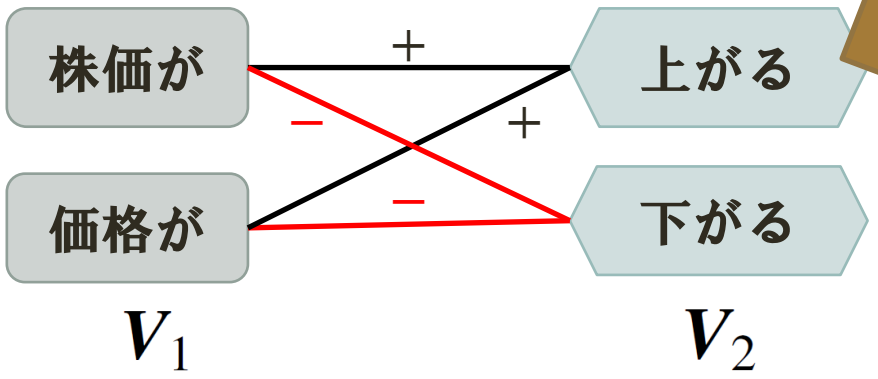
$$a_{ij} = \begin{cases} 1 & \text{頂点 } i \text{ と頂点 } j \text{ の間に positive な辺がある} \\ -1 & \text{頂点 } i \text{ 頂点 } j \text{ の間に negative な辺がある} \\ 0 & \text{頂点 } i \text{ と頂点 } j \text{ の間に辺がない} \end{cases}$$

$$\left(\text{ただし，} i \neq j\right)$$

dil0017@mail.doshisha.ac.jp

符号付きグラフ

- 符号付き2部グラフ
  - 集合間にしか辺がないような2つの集合に頂点を分割できる符号付きグラフ



極性付き
格フレームを
表現できる.

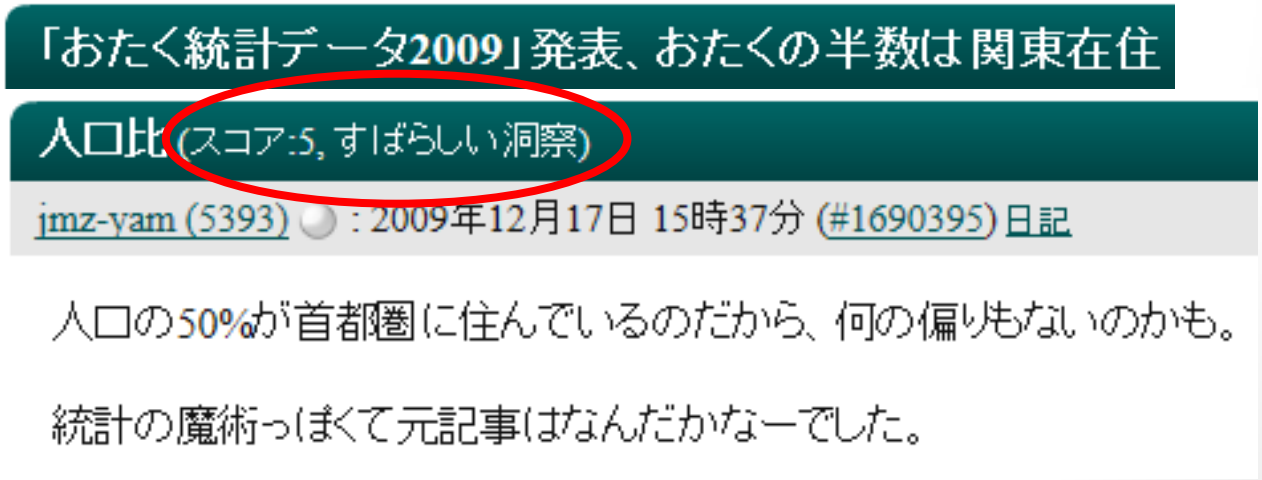
図１：符号付き2部グラフの例

dil0017@mail.doshisha.ac.jp

符号付きグラフ

- **近年，ソーシャルメディアの普及とその機能の多様化によって，その有用性は増えつつある．**

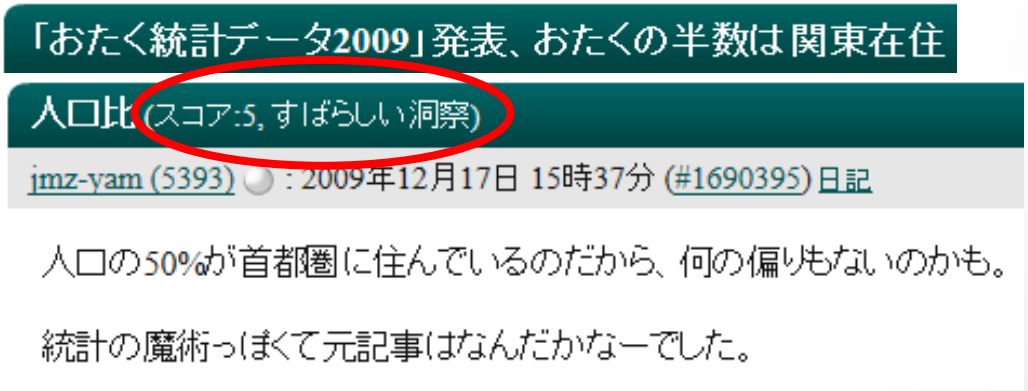
（Leskovec, Huttenlocher and Kleinberg, 2010）



Slashdot **参照**

符号付きグラフ

テキストマイニングに応用された例はない.



Slashdot 参照

符号付きグラフ

- balance
  - **集合内は全てpositiveの，集合間は全てnegativeの辺が張られるような，2つの集合に頂点を分割できる状態**

（Harary, 1956）

dil0017@mail.doshisha.ac.jp

集合内にnegativeな辺が含まれなくなるだけでなく，極性反転も考慮した分割

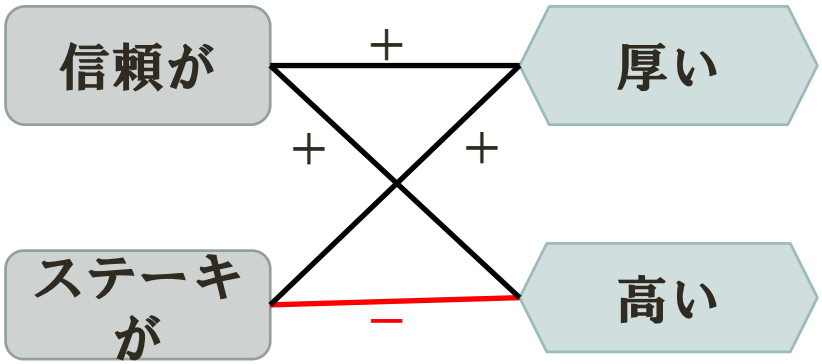


図2：balanceでないグラフ

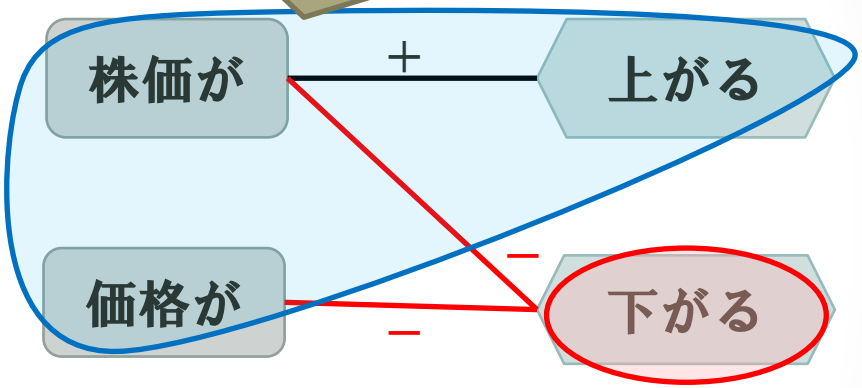


図4：balanceなグラフ

dil0017@mail.doshisha.ac.jp

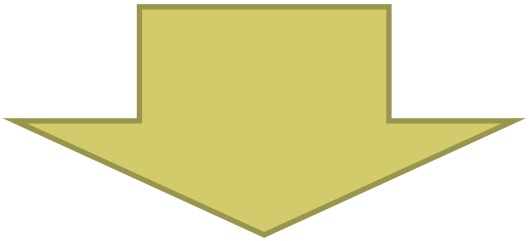
符号付きグラフ

- **符号付きグラフを，できるだけbalanceな分割に近い状態に分割する手法がいくつか提案されている．**
  - （Doreian and Mrvar, 1996）
    - **2値近似法**
  - （J. Kunegis, 2010）
    - **スペクトル分析**

dil0017@mail.doshisha.ac.jp

符号付きグラフ

符号付きグラフに対する重複クラスタリングは提案されていない

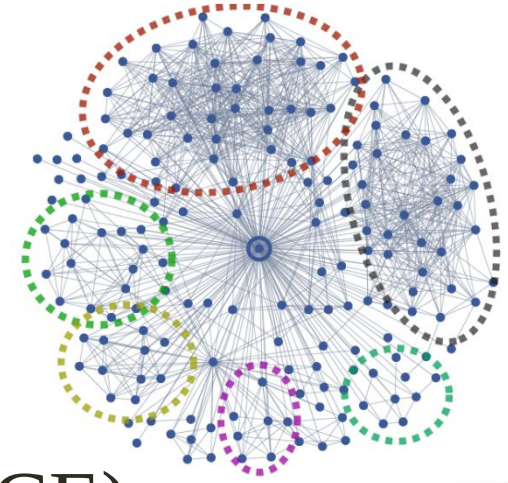


既存の手法を符号付きグラフに適用する

dil0017@mail.doshisha.ac.jp

重複クラスタリング

重複クラスタリング



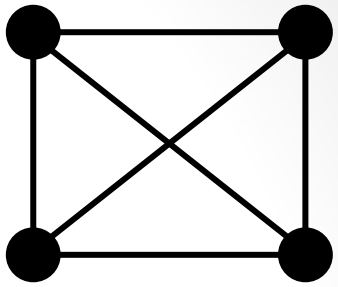
- **先行研究**
  - Greedy Clique Expansion (GCE)

    Lee, et al. (2010)

  - **極大クリークとSeedとして，評価関数を局所最大化するようなクラスターを抽出.**
  - **各クラスターが独立に抽出されるので，重複部分が生まれる.**

dil0017@mail.doshisha.ac.jp

重複クラスタリング



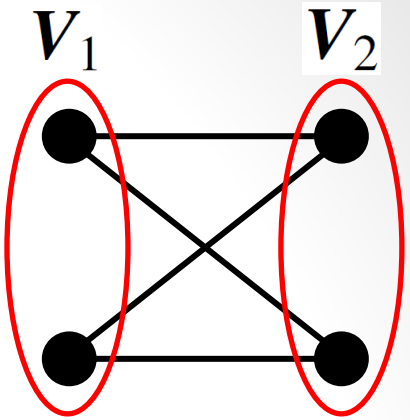
クリーク

- **先行研究：GCE**
  - **クリーク**
    - **全ての頂点間に辺がある部分グラフ**
  - **極大クリーク**
    - **他のどのクリークにも含まれないクリーク**

- **GCEはクリークをSeedとして評価関数を最大化するように広げていく．**

dil0017@mail.doshisha.ac.jp

重複クラスタリング



クリーク
(2部グラフ)

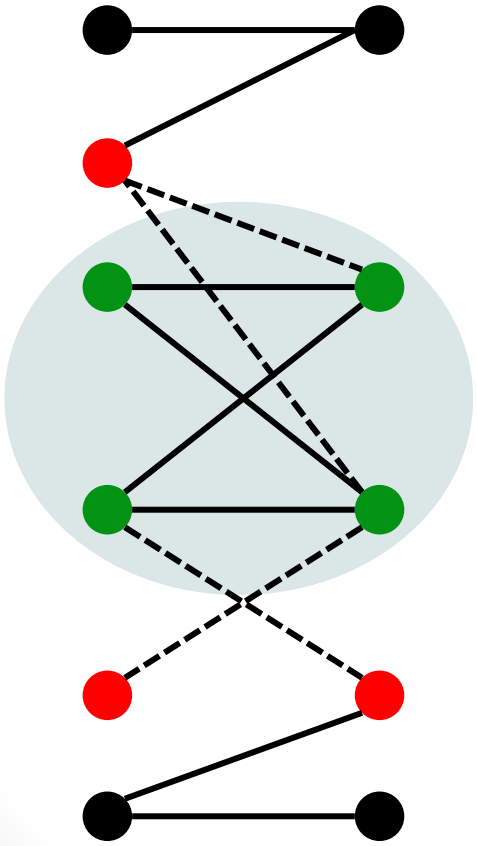
- 先行研究：GCE
  - クリーク
    - 全ての頂点間に辺がある部分グラフ
  - 極大クリーク
    - 他のどのクリークにも含まれないクリーク

- GCEはクリークをSeedとして評価関数を最大化するように広げていく．

dil0017@mail.doshisha.ac.jp

重複クラスタリング

- **先行研究：GCE**



Seed
S

Seedの近傍にある頂点の中から，評価関数 F_s を最大にする頂点を算出する．

$$F_s = \frac{k_{in}^s}{(k_{in}^s + k_{out}^s)^\alpha}$$

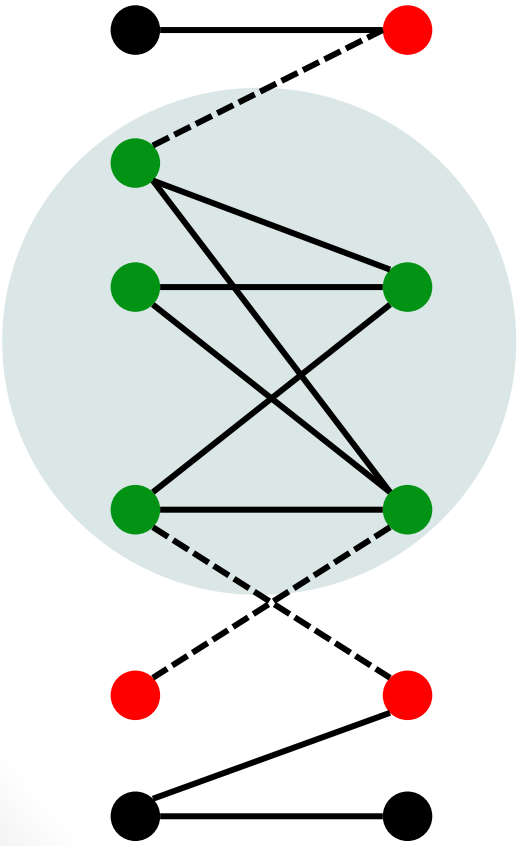
$$k_{in}^s = \sum_{i,\, j \in S} a_{ij}, \quad k_{out}^s = \sum_{i \notin S,\, j \in S} a_{ij}$$

α：パラメータ

dil0017@mail.doshisha.ac.jp

重複クラスタリング

- **先行研究：GCE**



Cluster S

Seedの近傍にある頂点の中から，評価関数 F_s を最大にする頂点を算出する．

$$F_s = \frac{k_{in}^s}{(k_{in}^s + k_{out}^s)^\alpha}$$

$$k_{in}^s = \sum_{i,\, j \in S} a_{ij}, \quad k_{out}^s = \sum_{i \notin S,\, j \in S} a_{ij}$$

α：パラメータ

重複クラスタリング

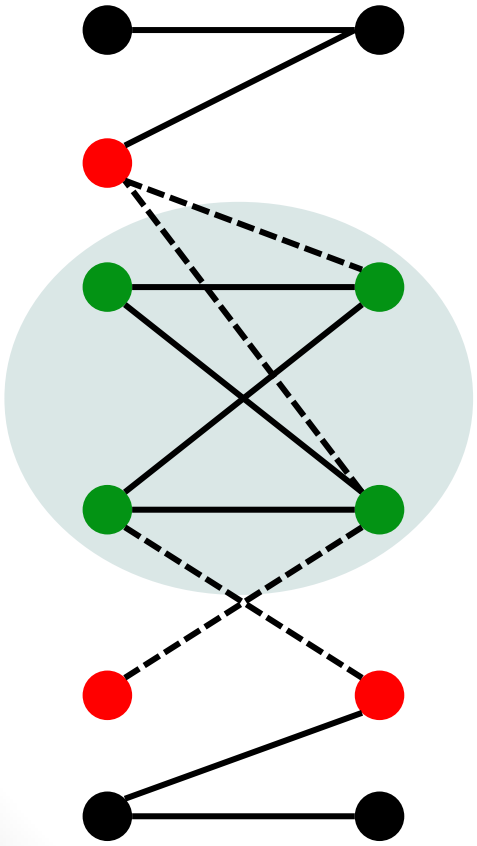
- **先行研究：**GCE
  - **クラスターの類似性**
    - **同じ**Seed**から，類似するクラスターが生成される可能性がある．**
    - **クラスター間の非類似性** $\delta_E(S, S')$ **を定義し，酷似するクラスターは捨てる．**

$$\delta_E(S, S') = 1 - \frac{|S \cap S'|}{min(|S|, |S'|)}$$

dil0017@mail.doshisha.ac.jp

重複クラスタリング

- **先行研究：GCE**　パラメータ：k, ε, α



1．最低 k 個の頂点をもつ極大クリークを抽出する．

2．最も大きいSeedから，評価関数が収束するまでクラスターを広げる．

3．すでに得られていたクラスターとの非類似性が ε より大きい場合，そのクラスターを受け入れる．

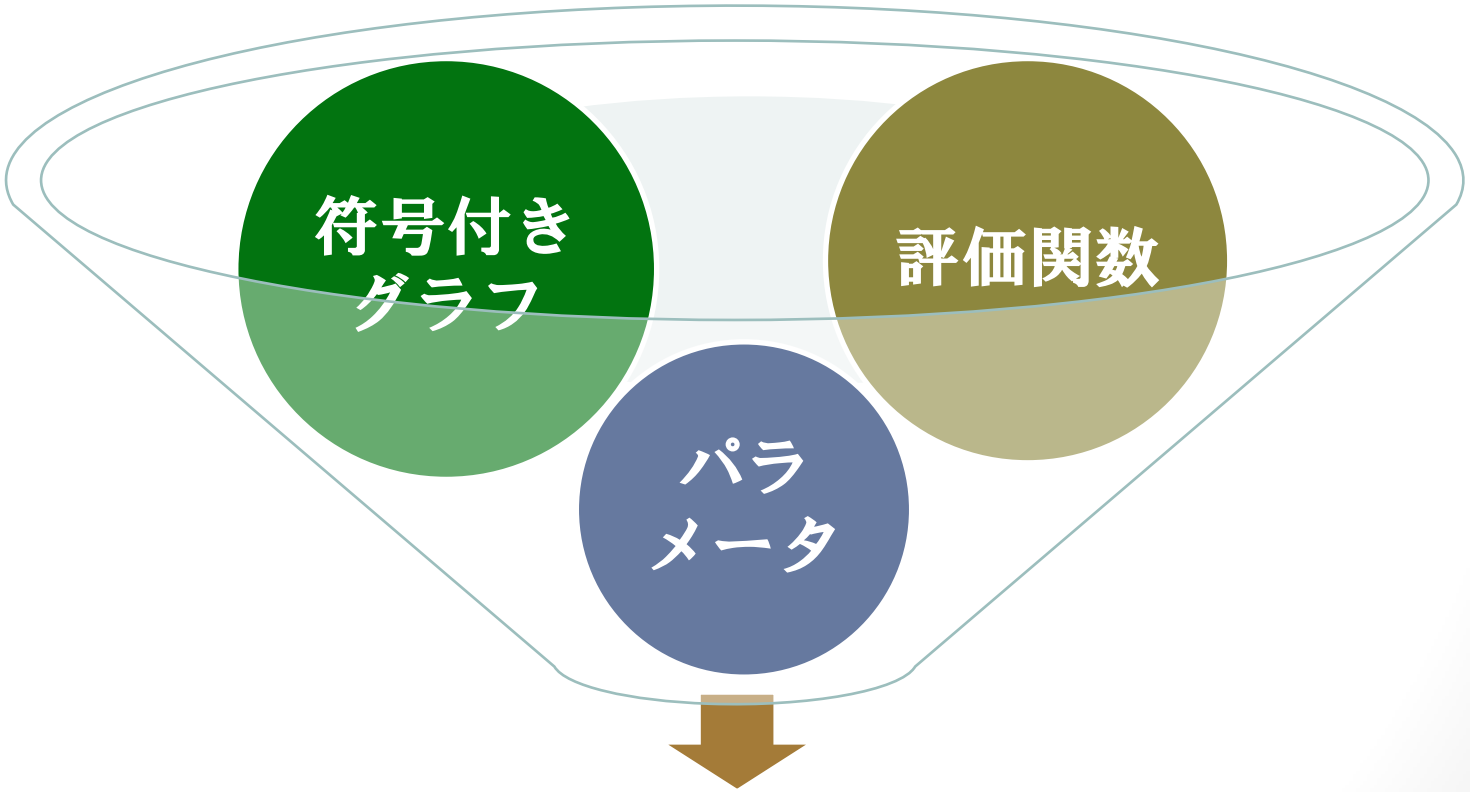
重複クラスタリング

- **提案手法**
  - Greedy Signed Clique Expansion (GSCE)
  - **新しい評価関数を導入する.**

dil0017@mail.doshisha.ac.jp

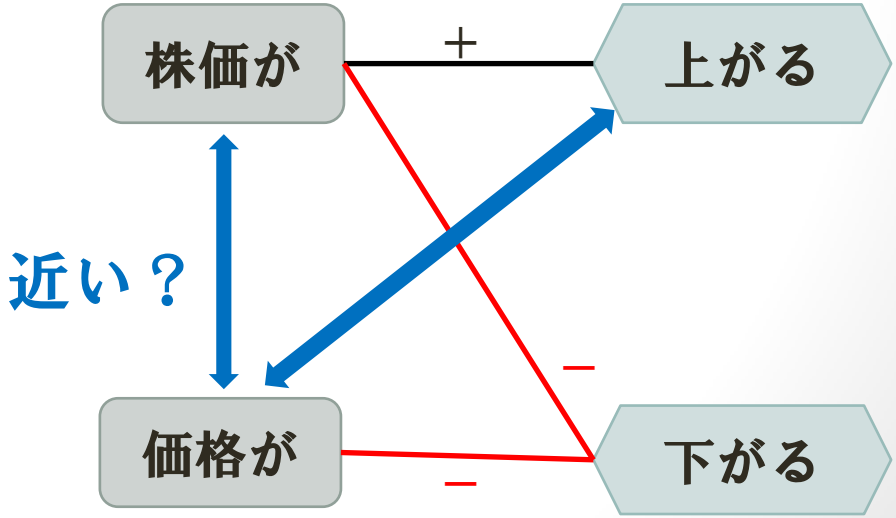
重複クラスタリング

- **提案手法：GSCE**



重複クラスタリング

- 提案手法：GSCE
  - 評価関数
    - 単なる隣接情報だけでは極性反転を考慮できない.

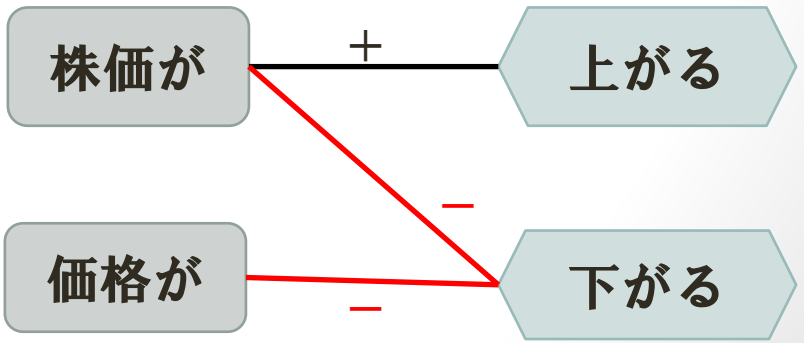


重複クラスタリング

- **提案手法：GSCE**
  - **評価関数**

$$p_{ij}^{(d)} = \begin{cases} \text{negative な辺を偶数個含む最短経路の数} & (i \text{ と } j \text{ の最短距離が } d \text{ のとき}) \\ 0 & (\text{others}) \end{cases}$$

$$n_{ij}^{(d)} = \begin{cases} \text{negative な辺を奇数個含む最短経路の数} & (i \text{ と } j \text{ の最短距離が } d \text{ のとき}) \\ 0 & (\text{others}) \end{cases}$$



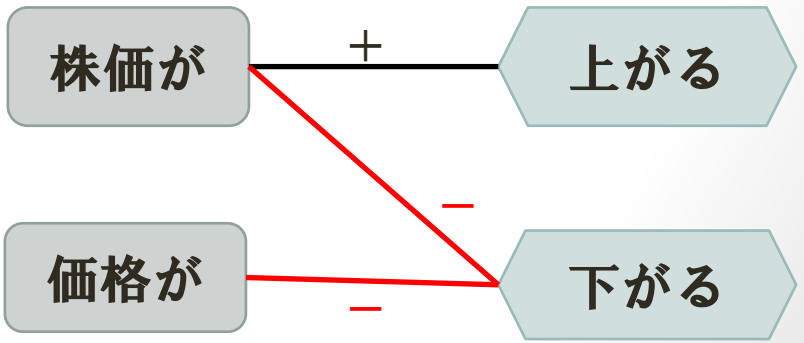
重複クラスタリング

- **提案手法：GSCE**
  - **評価関数**

$$p_{ij}^{(d)} = \begin{cases} \text{negative な辺を偶数個含む最短経路の数} & (i \text{ と } j \text{ の最短距離が } d \text{ のとき}) \\ 0 & (\text{others}) \end{cases}$$

$$n_{ij}^{(d)} = \begin{cases} \text{negative な辺を奇数個含む最短経路の数} & (i \text{ と } j \text{ の最短距離が } d \text{ のとき}) \\ 0 & (\text{others}) \end{cases}$$

$d = 1$	株価が	価格が	上がる	下がる
株価が	0	0	1	1
価格が	0	0	0	1
上がる	1	0	0	0
下がる	1	1	0	0



dil0017@mail.doshisha.ac.jp

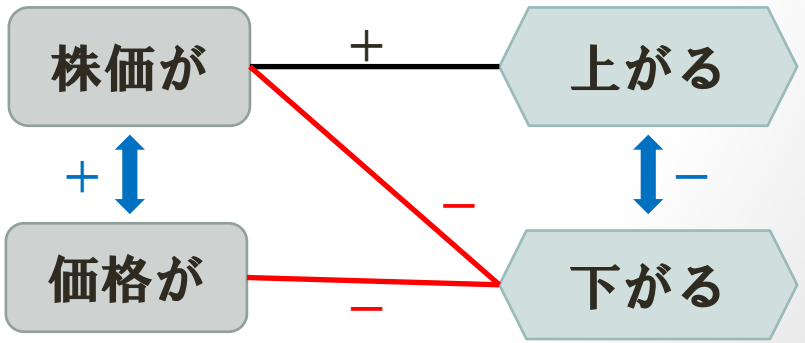
重複クラスタリング

- **提案手法：GSCE**
  - **評価関数**

$$p_{ij}^{(d)} = \begin{cases} \text{negative な辺を偶数個含む最短経路の数} & (i \text{ と } j \text{ の最短距離が } d \text{ のとき}) \\ 0 & (\text{others}) \end{cases}$$

$$n_{ij}^{(d)} = \begin{cases} \text{negative な辺を奇数個含む最短経路の数} & (i \text{ と } j \text{ の最短距離が } d \text{ のとき}) \\ 0 & (\text{others}) \end{cases}$$

$d = 2$	株価が	価格が	上がる	下がる
株価が	0	1	0	0
価格が	1	0	0	0
上がる	0	0	0	1
下がる	0	0	1	0



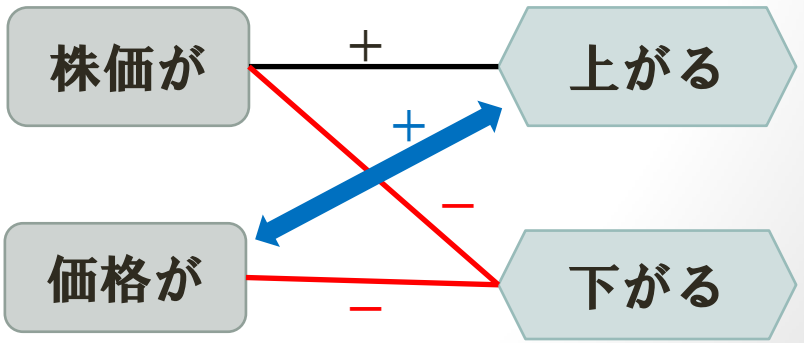
重複クラスタリング

- **提案手法：GSCE**
  - **評価関数**

$$p_{ij}^{(d)} = \begin{cases} \text{negative な辺を偶数個含む最短経路の数} & (i \text{ と } j \text{ の最短距離が } d \text{ のとき}) \\ 0 & (\text{others}) \end{cases}$$

$$n_{ij}^{(d)} = \begin{cases} \text{negative な辺を奇数個含む最短経路の数} & (i \text{ と } j \text{ の最短距離が } d \text{ のとき}) \\ 0 & (\text{others}) \end{cases}$$

$d = 3$	株価が	価格が	上がる	下がる
株価が	0	0	0	0
価格が	0	0	1	0
上がる	0	1	0	0
下がる	0	0	0	0



dil0017@mail.doshisha.ac.jp

重複クラスタリング

- **提案手法：GSCE**
  - **幅優先探索**（Breadth First Search）

(Erwin Kreayszig, 1999)

で最短経路と同時に求める．

$$p_{ij}^{(d)} = \begin{cases} \text{negative な辺を偶数個含む最短経路の数} & (i \text{ と } j \text{ の最短距離が } d \text{ のとき}) \\ 0 & (\text{others}) \end{cases}$$

$$n_{ij}^{(d)} = \begin{cases} \text{negative な辺を奇数個含む最短経路の数} & (i \text{ と } j \text{ の最短距離が } d \text{ のとき}) \\ 0 & (\text{others}) \end{cases}$$

dil0017@mail.doshisha.ac.jp

重複クラスタリング

- **提案手法：GSCE**
  - **評価関数**

$$F(S,\ g) = \frac{\sum_{d=1}^{g} \sum_{i \in V_1} \sum_{j \in V_2} (p_{ij}^{(d)} - n_{ij}^{(d)}) f(d)}{|V_1||V_2|}$$

グラフが全てpositive**な辺からなり，**$f(1) = 1$**，**$g = 1$**ならば二部グラフにおける密度と同値.**

dil0017@mail.doshisha.ac.jp

重複クラスタリング

- **提案手法：GSCE**
  - **評価関数**

$f(d)$：**単調非増加関数**

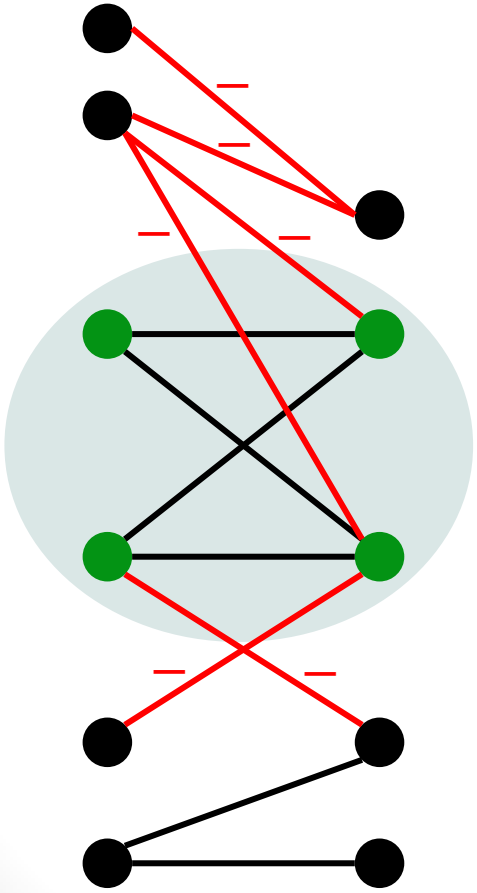
$$F(S,\ g) = \frac{\sum_{d=1}^{g} \sum_{i \in V_1} \sum_{j \in V_2} (p_{ij}^{(d)} - n_{ij}^{(d)}) f(d)}{|V_1||V_2|}$$

グラフが全てpositiveな辺からなり， $f(1) = 1$，$g = 1$
ならば二部グラフにおける密度と同値.

dil0017@mail.doshisha.ac.jp

重複クラスタリング

- **提案手法：GSCE**



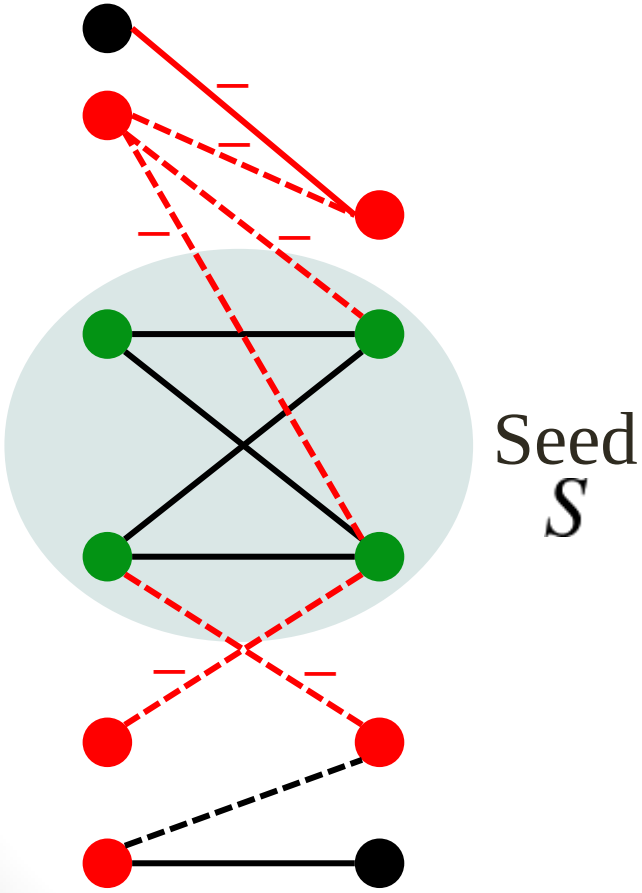
Seed
S

positiveな辺のみからなる極大クリークを抽出する．
(negativeなクリークは符号を入れ替えて抽出する．)

dil0017@mail.doshisha.ac.jp

重複クラスタリング

- **提案手法：GSCE**



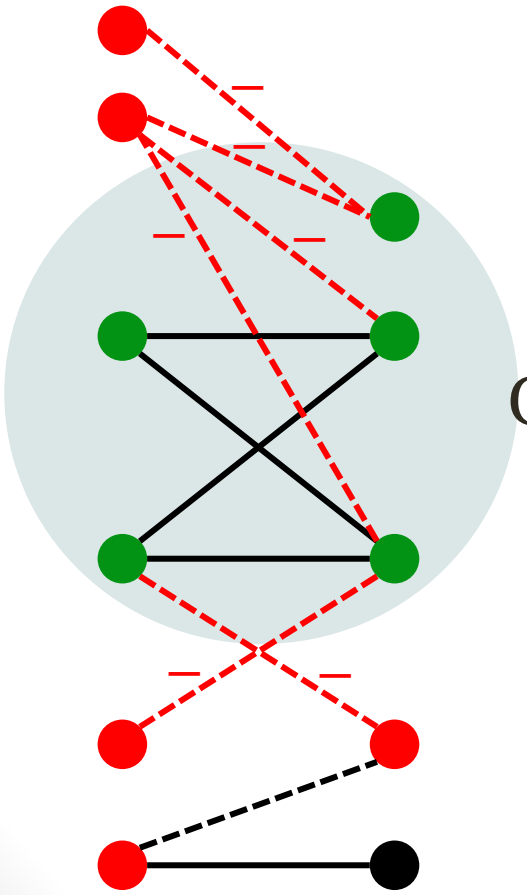
Seed
S

Seedのg-近傍にある頂点の中から，評価関数$F(S, g)$を最大にする頂点を算出.

dil0017@mail.doshisha.ac.jp

重複クラスタリング

- **提案手法：GSCE**



Cluster
S

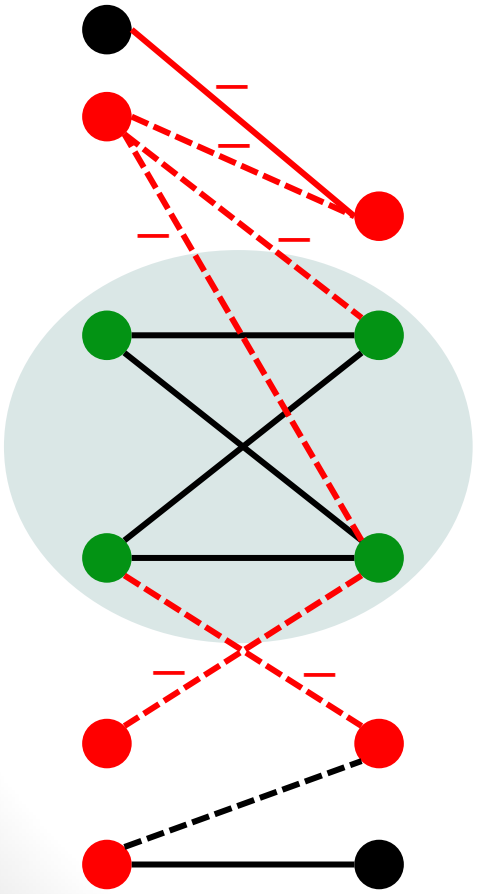
Seedのg-近傍にある頂点の中から，評価関数 $F(S,g)$ を最大にする頂点を算出.

評価関数 $F(S,g)$ が閾値 ε を超えない範囲で広げる.

dil0017@mail.doshisha.ac.jp

重複クラスタリング

- **提案手法：GSCE**　　**パラメータ：g, k, ε**



1．**最低 k 個の頂点をもつ極大クリークを抽出する．**

2．**評価関数が ε を超えない範囲でクラスターを広げる．**

3．**類似しているクラスターをまとめる．**

dil0017@mail.doshisha.ac.jp

極性付き概念化

- GSCEを用いて，極性を考慮した同義表現抽出を行う．

	noun	case1	verb	case2	polarity
2	悪化	ニ格	つながる	用言	-1
3	改善	ニ格	つながる	用言	1
4	株価上昇	ニ格	つながる	用言	1
5	赤字	ニ格	修正	用言	-1
6	赤字	ヘ格	修正	用言	-1
7	赤字	ヲ格	上回る	用言	1
8	赤字	カラ格	上方修正	用言	1
9	5%	修飾格	上回る	用言	1
10	5%	修飾格	上昇	用言	1
11	9%	修飾格	上回る	用言	1

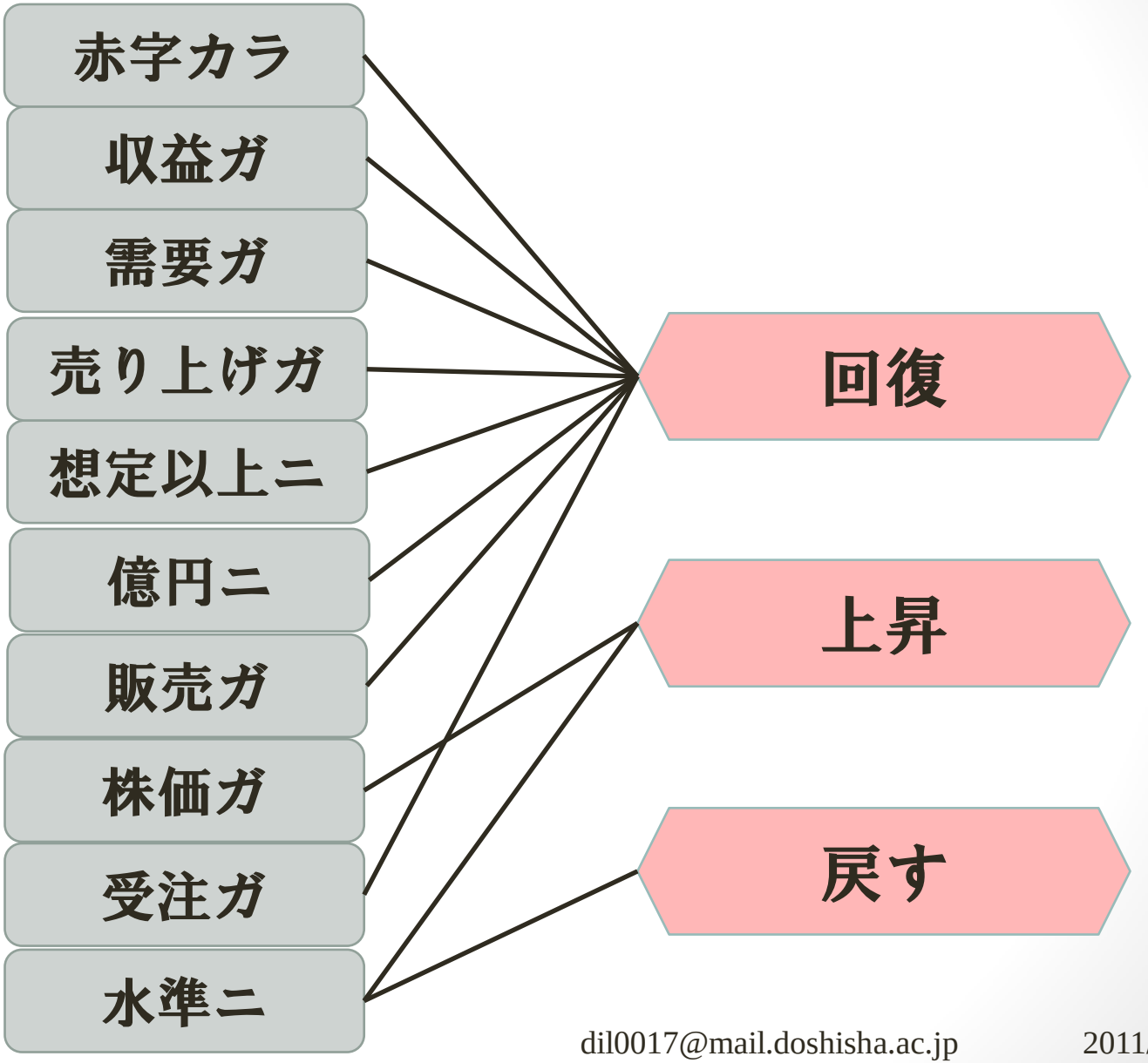
dil0017@mail.doshisha.ac.jp

極性付き概念化

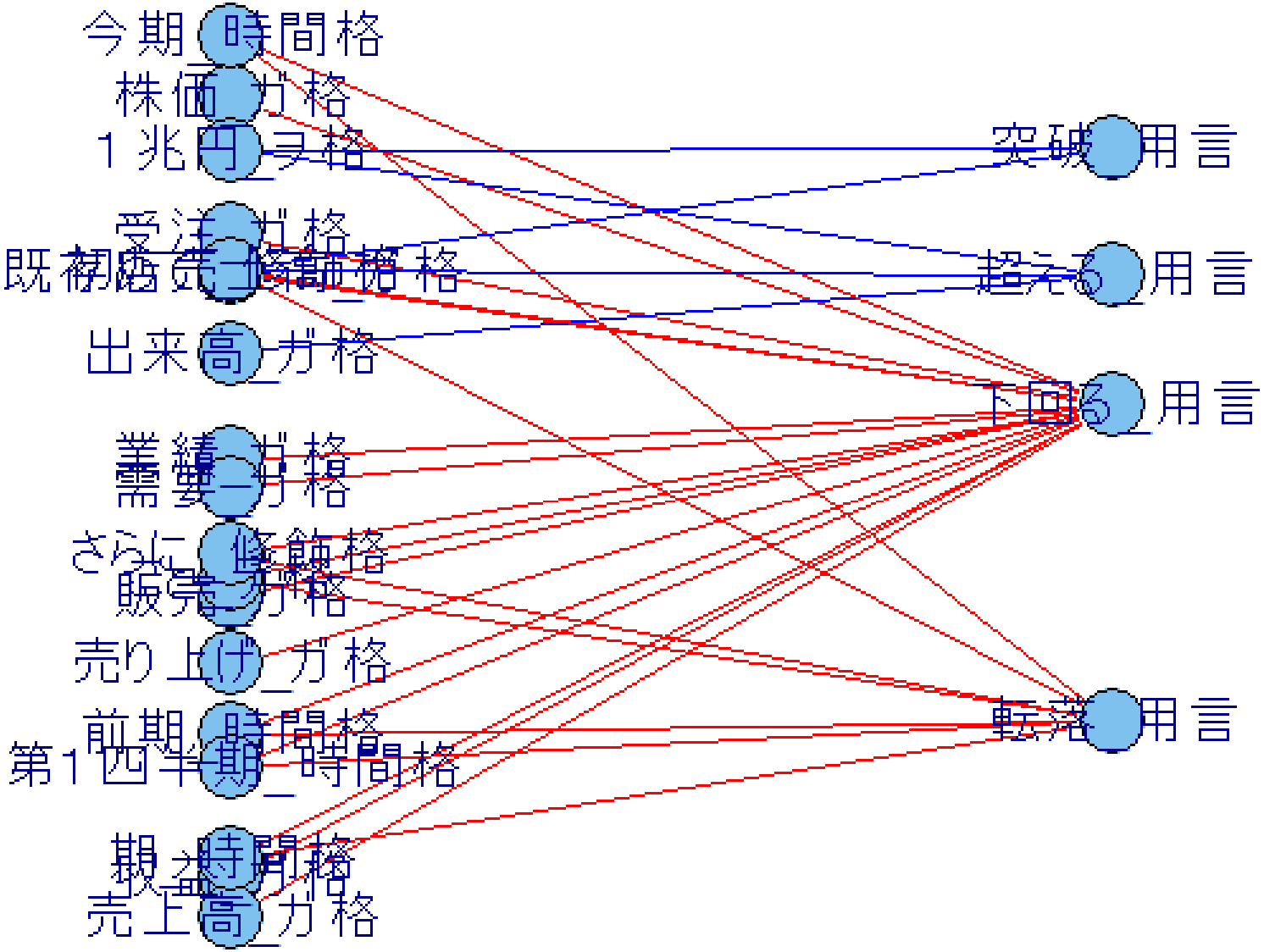
- が，うまくいかなかった．(昨日)
- $p_{ij}^{(d)}$ のばらつきが大きいことが原因なのでは？
- そこで，
- $p_{ij}^{(d)} f(d)$ が１を超えないように $p_{ij}^{(d)}$ の値を調整
- 頂点数も20に制限する．

dil0017@mail.doshisha.ac.jp

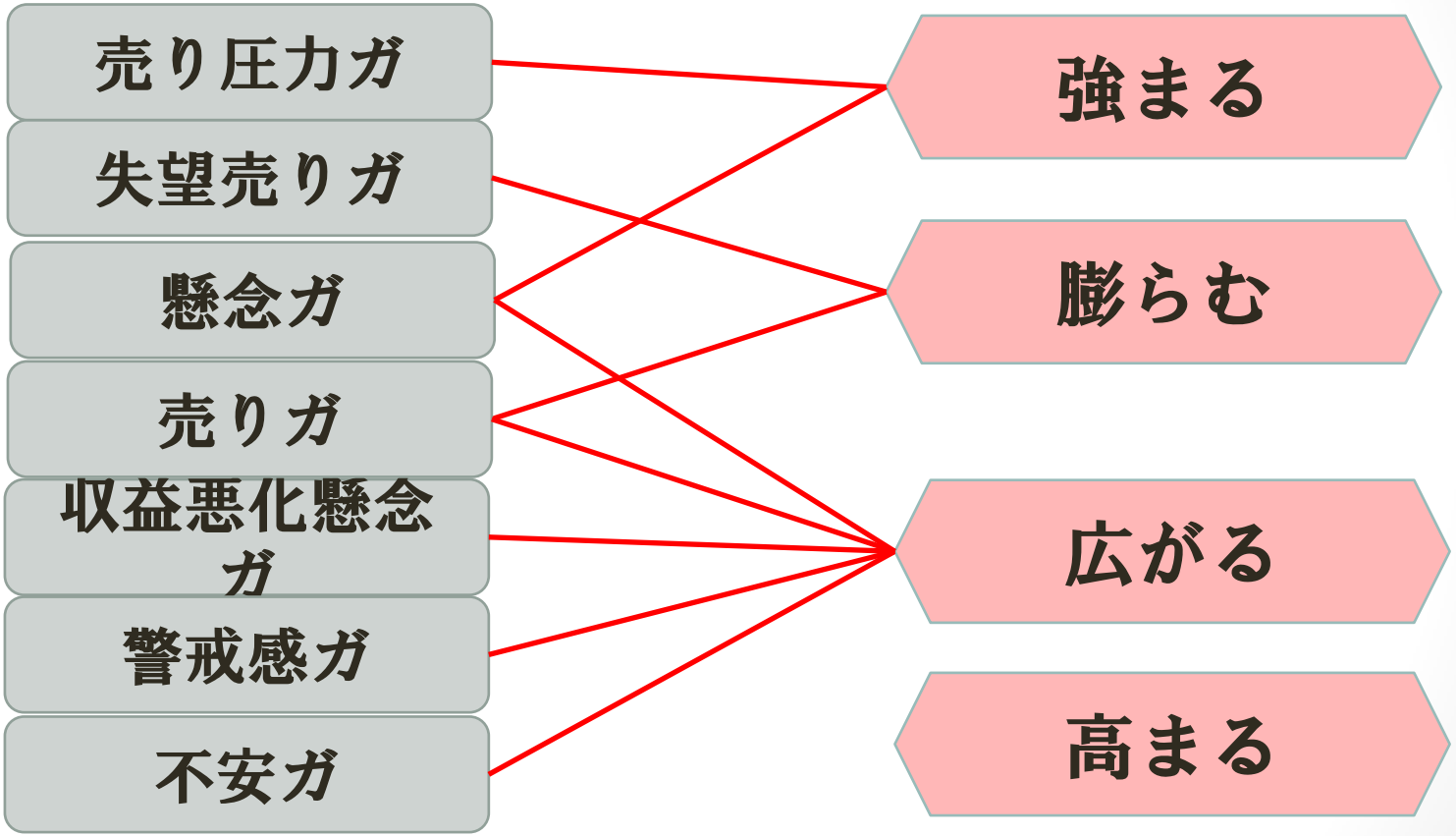
得られたクラスターの例（正の極性）



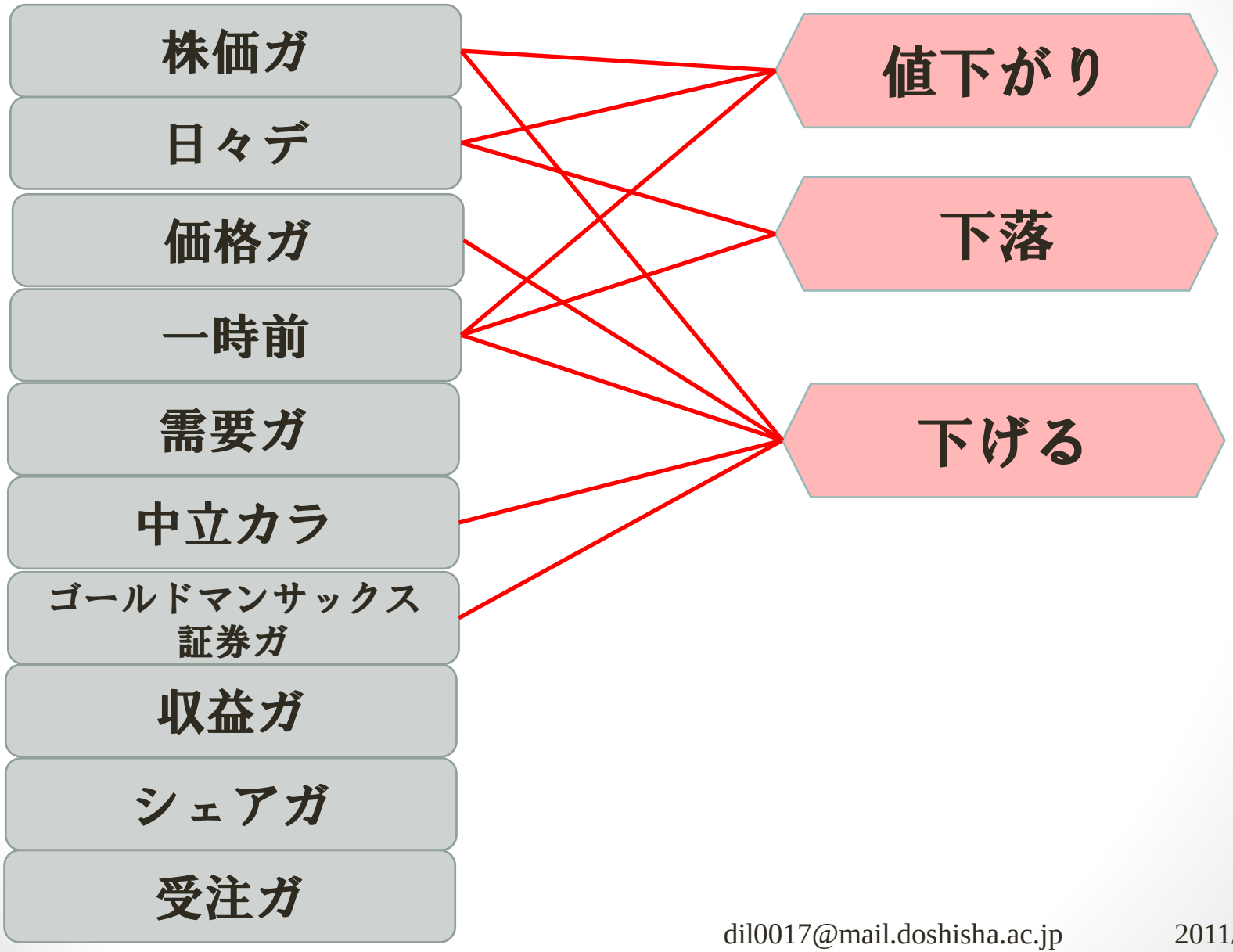
得られたクラスターの例（正の極性）



得られたクラスターの例（負の極性）



得られたクラスターの例（負の極性）



まとめ

- 符号付きグラフに対する重複クラスタリング手法である，GSCEを提案した．
- それを用いて極性付き概念抽出を行った．

dil0017@mail.doshisha.ac.jp

問題点と今後の課題

- クラスターの精度と解釈の容易さを上げるため，アルゴリズムと評価関数についてより検討.
- 得られたクラスターが株価予測に有用に働くか確認.
- 符号をより一般化させる.
  （ニュートラルの導入）

dil0017@mail.doshisha.ac.jp

ご清聴ありがとうございました

dil0017@mail.doshisha.ac.jp

- **上野友司・森　辰則・木戸冬子・中川裕志**（2004）**係り受けの2部グラフと共起関係を利用した同義表現抽出，情報処理学会研究報告，**2004-NL-159: 169-176.
- **相澤彰子, 中渡瀬秀一: 係り受け関係を利用した類語・例文辞書構築法と大規模コーパスへの適用**,*Proceedings of the 20th annual conference of the Japanese Society for Aritificial Intelligence*(2007). 2E1-5.
- Lee, C., Reid, F., McDaid, A. & Hurley, N. (2010). Detecting highly overlapping community structure by greedy clique expansion. In Proc. 4th Workshop on Social Network Mining and Analysis.

dil0017@mail.doshisha.ac.jp