ソーシャルネットワークのグラフマイニング

秦恭史[1] 諏訪博彦[1] 岸本康成[2] 藤原靖宏[2] 新井淳也[2]
飯田恭弘[2] 岩村相哲[2] 鳥海不二夫[3] 安本慶一[1]

奈良先端科学技術大学院大学[1]
NTTイノベーションセンタ[2]
東京大学[3]

研究室の目標

- NAISTユビキタスコンピューティングシステム研究室
  センサ・デバイス・ネットワークの連携により、環境から取得される実世界データを効率よく収集・分析・応用し、先進的なサービスの実現を目指す。

研究背景

- 近年，日本の災害事情は深刻化してきている

東日本大震災（2011）

熊本地震（2016）

災害時，いち早く情報を手に入れることは重要

研究背景

- Twitter上で拡散された災害情報

停電や電話回線が止まった場合，
インターネットを活用したSNSによる情報共有が効果を発揮

- 総務省の調べにより，東日本大震災時，
  Twitterは災害情報の拡散に貢献したことが報告されている

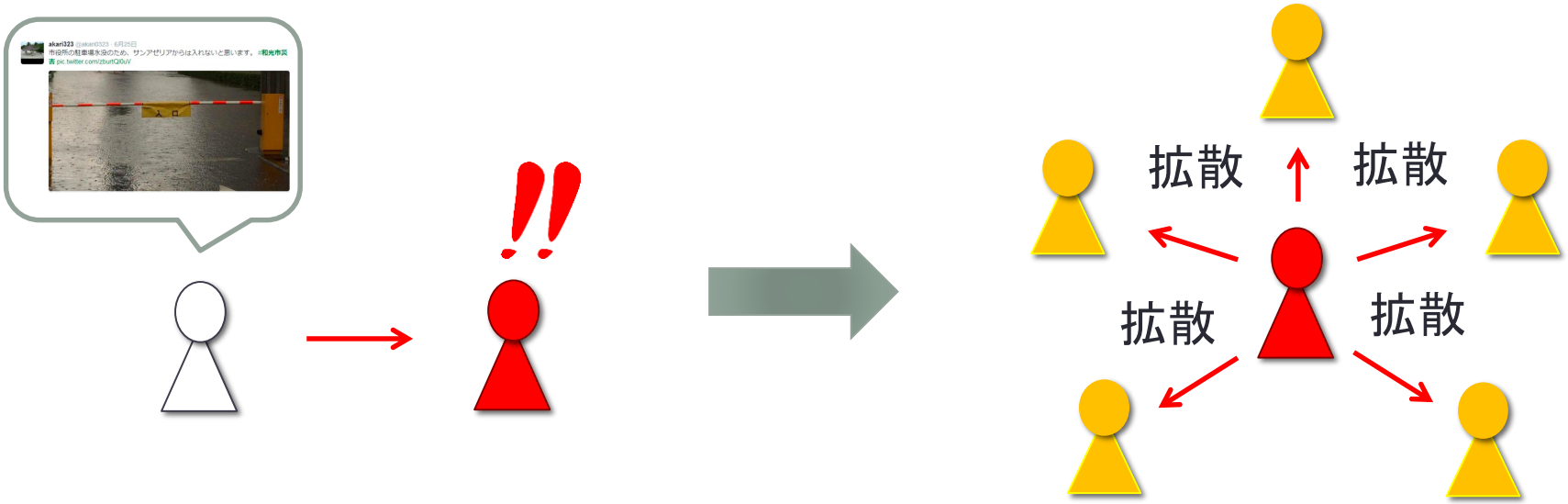
災害時，Twitterは情報流通を支える重要なツールである

研究背景

- 情報を流通させるためには重要アカウントが鍵

重要アカウント

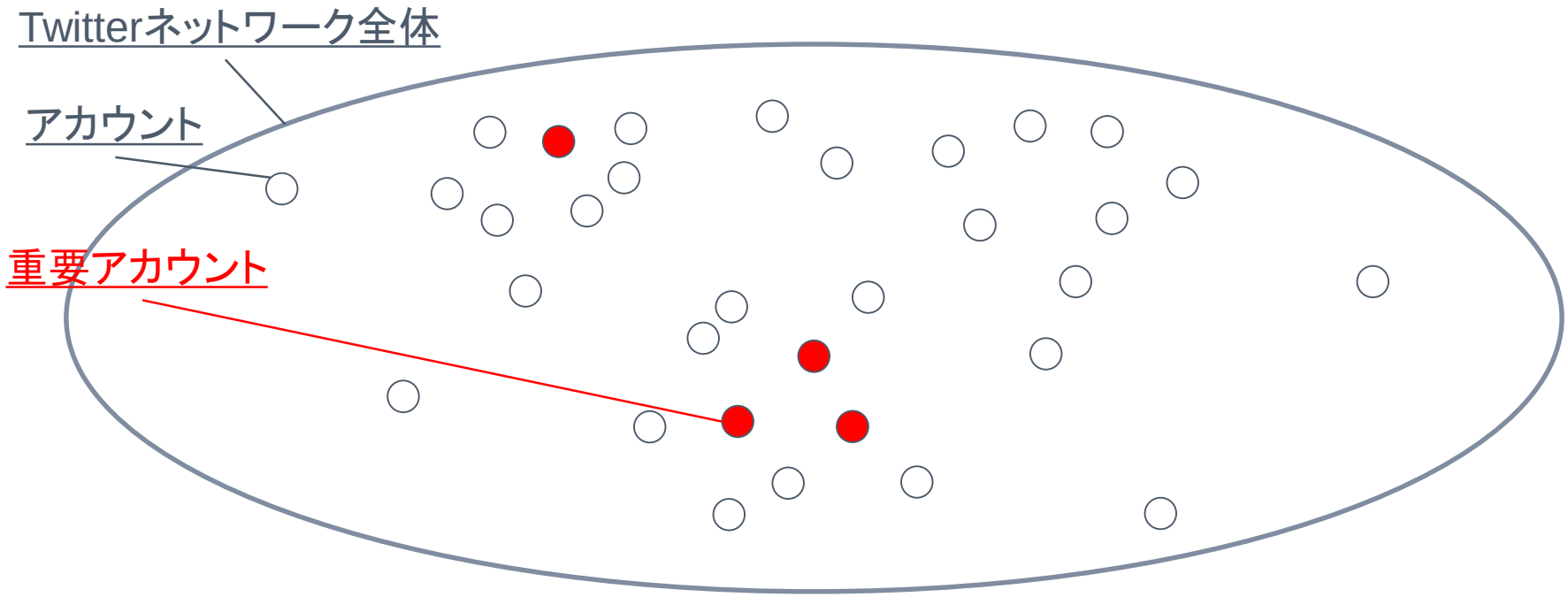
情報の拡散能力が高いアカウント



災害時に重要アカウントに対して
情報の拡散を依頼することで情報の流通をサポート

関連研究

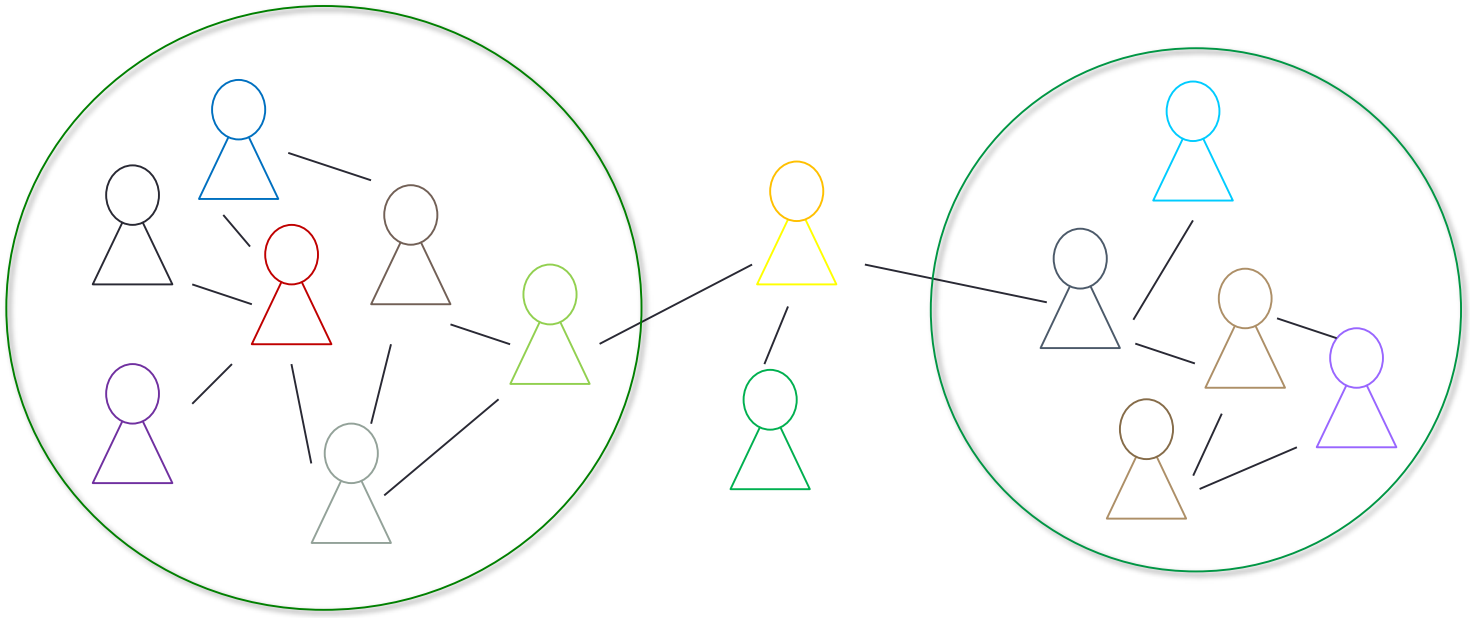
- 石原ら(2016)
  次数中心性と媒介中心性を用いてネットワーク全体としての重要アカウントを抽出



石原裕規ほか, "東日本大震災前後における重要アカウントの抽出とコミュニケーション形態の変容, "
電子情報通信学会論文誌D, Vol.99, No.5, pp501-513, 2016.

関連研究

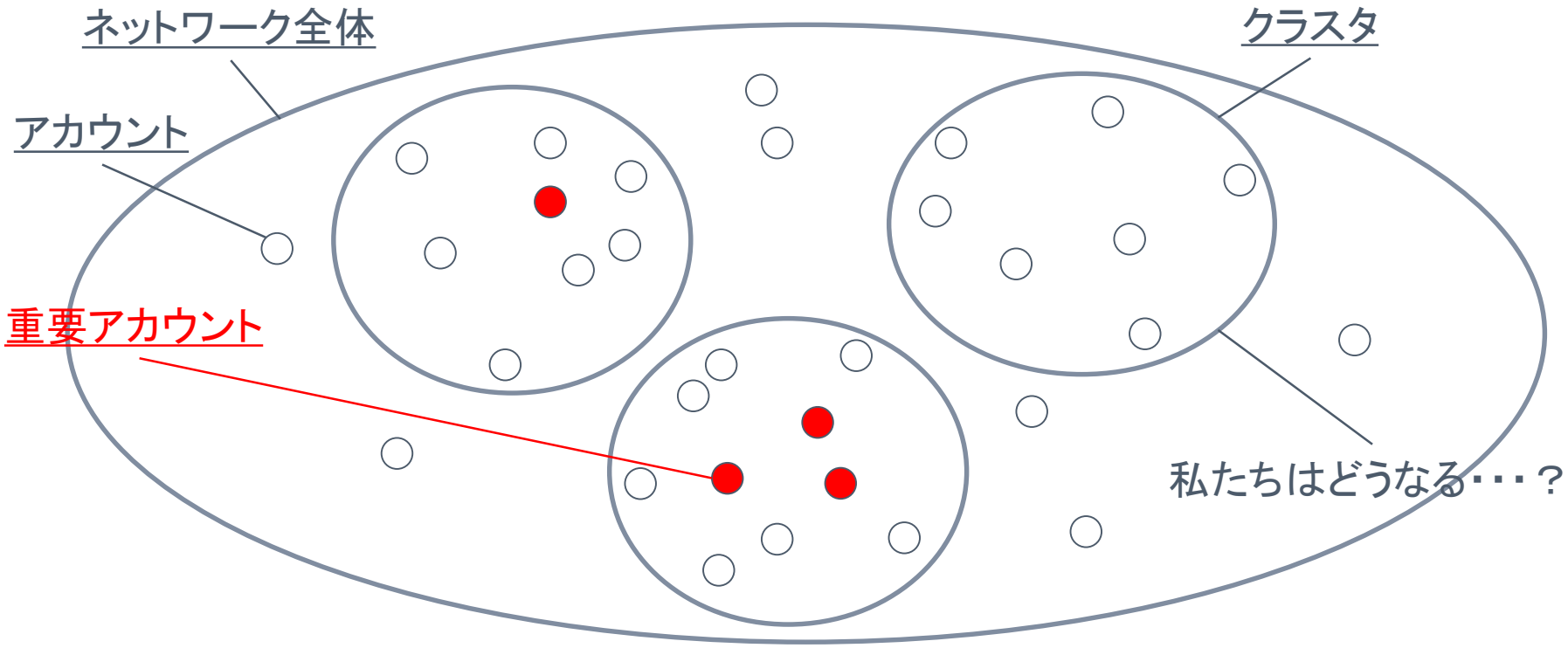
- Amacら（2013）
  Twitterがつくるネットワークは近い人同士でつながりを持つもので，**複数のクラスタに分かれているもの**



AmaC HerdaGdelen et al. An exploration of social identity: The geography and politics of news-sharing communities in twitter. Complexity, Vol. 19, No. 2, pp. 10–20, 2013.

問題点その1：情報が偏る可能性がある

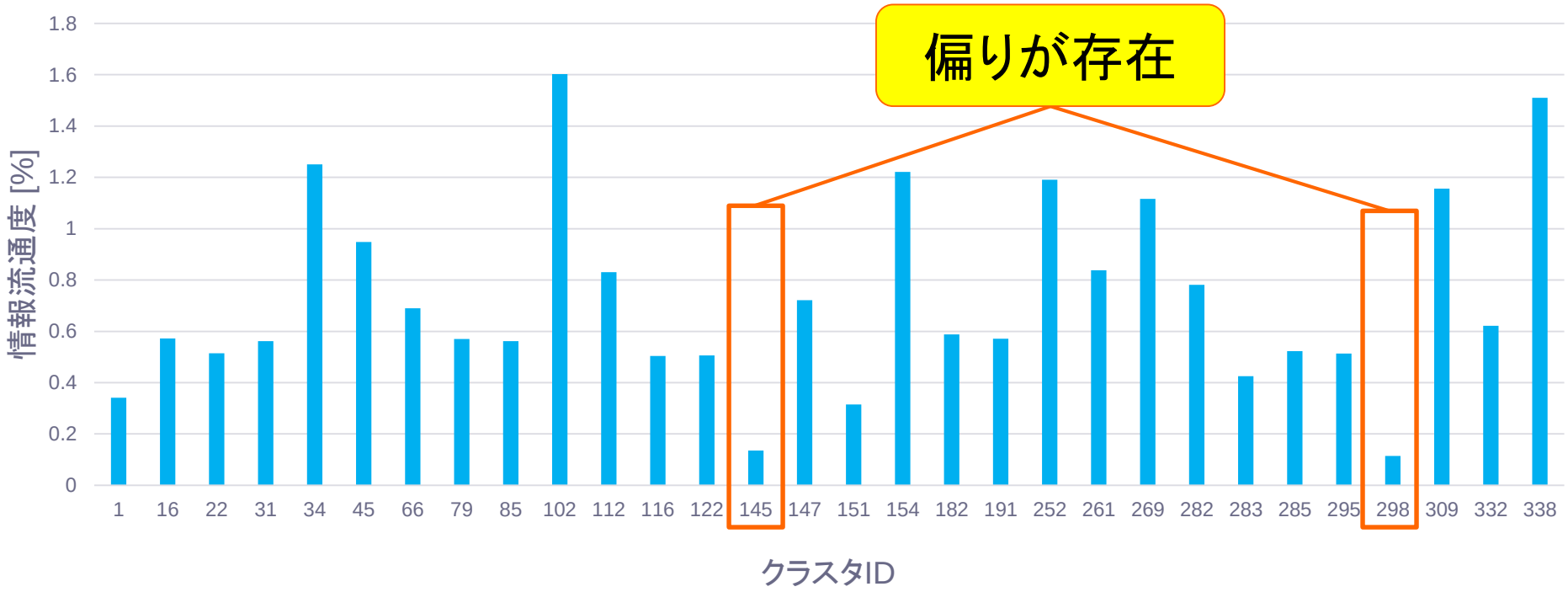
- 全体で重要なアカウントがすべてのクラスタにおいて重要とは限らない
  →もし，一部クラスタに重要アカウントが偏ると情報の拡散に
  偏りが出てしまう可能性がある



クラスタ毎に重要アカウントを抽出するべきでは？

先行研究

- 我々は先行研究として**情報の偏りを確認**した（2017.02）



- 情報の拡散に偏りが存在した
- 海外クラスタには情報があまり届かなかった

秦恭史，諏訪博彦，岸本康成，藤原靖宏，新井淳也，飯田恭弘，岩村相哲，鳥海不二夫，安本慶一
”東日本大震災におけるクラスタリングに基づく情報拡散度の比較，” 第23回社会情報システム学シンポジウム，2017.02.

問題点その2：使用する指標が適していない

- 石原ら
  - 次数中心性
    エッジの数で決まる．高いほど直接的なつながりが多い．
    →ネットワークの相互関係を無視している
  - 媒介中心性
    情報を伝達する際に経由されている回数で決まる．
    →次数中心性よりは良いがもっと良い指標がある
- 提案
  - ページランク（PR）
    「有名なページは有名なページへリンクを張る」という考えに基づいて，
    ページのランク付けを行う．
    「重要なアカウントは重要なアカウントとコミュニケーションを取る」
    という考えで抽出を行う．

研究目的

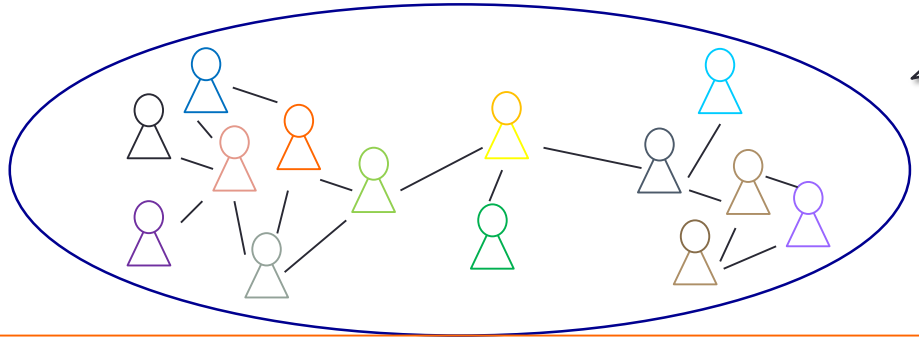
- 目的：偏りなく情報を拡散するための重要アカウントを抽出
  - 提案手法1：クラスタリングに基づく重要アカウント抽出手法
  - 提案手法2：ページランクに基づく重要アカウント抽出手法

研究目的

- 目的：偏りなく情報を拡散するための重要アカウントを抽出
  - 提案手法1：クラスタリングに基づく重要アカウント抽出手法
  - 提案手法2：ページランクに基づく重要アカウント抽出手法

クラスタリングに基づく重要アカウント抽出

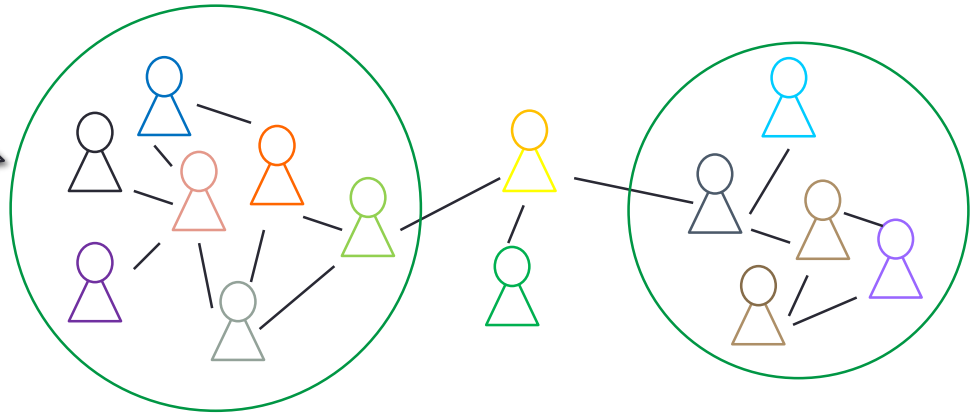
- 石原ら：ネットワーク全体から上位Nアカウントを抽出
  - 次数中心性，媒介中心性に基づいて上位300アカウントを抽出



1位 Cさん
2位 EFさん
3位 KJさん
⋮ ⋮
298位 STさん
299位 SZさん
300位 VVさん

- 提案：各クラスタから上位Nアカウント/対象クラスタ数を抽出
  - 上位30クラスタを対象に次数中心性，媒介中心性，ページランクに基づいて各10アカウントを抽出

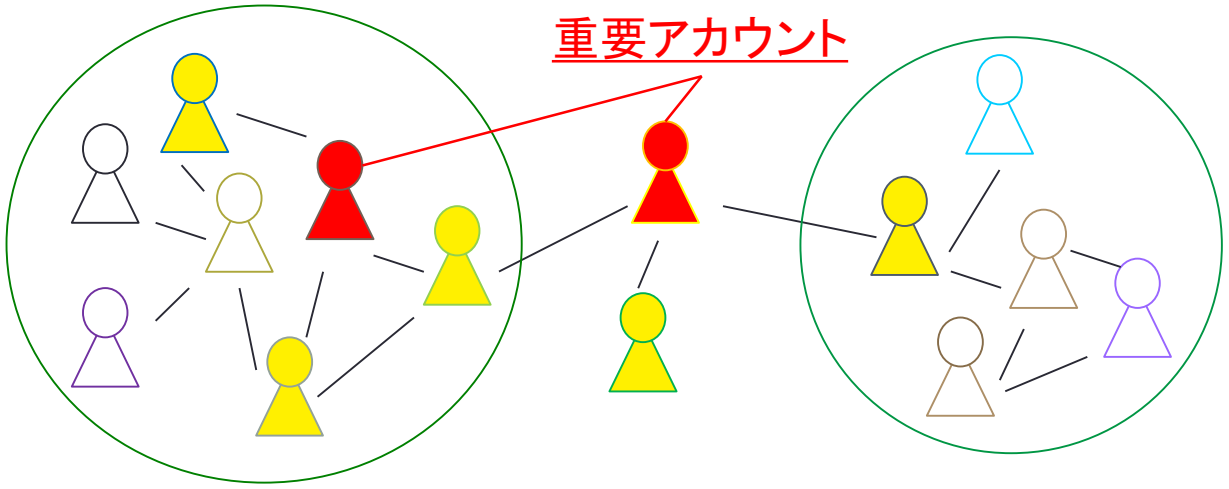
クラスタ(1)
1位 Kさん
2位 ESさん
3位 FGさん
⋮ ⋮
9位 HAさん
10位 HKさん



クラスタ(30)
クラスタ(29)
クラスタ(・・・)
クラスタ(3)
クラスタ(2)
1位 Dさん
2位 Zさん
3位 SKさん
⋮ ⋮
9位 XLさん
10位 XXさん

評価方法

- 「重要アカウントから1ホップ離れているアカウント」が各クラスタ内で何割存在しているか（拡散度）を検証



クラスタAの場合

$$\frac{3人}{7人} \times 100 = 43\%$$

クラスタBの場合

$$\frac{1人}{5人} \times 100 = 20\%$$

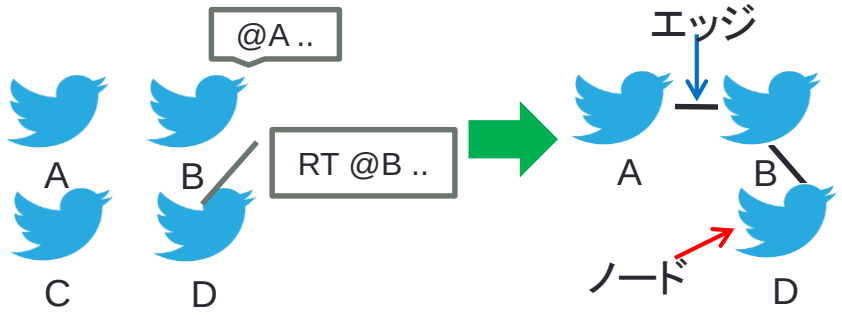
クラスタリング手法の選択

- 今回のようなビッグデータにおいて従来のクラスタリング手法では膨大な時間とハイスペックマシンが必要
- 3つの手法を検討
  1. Modularityベースのクラスタリング（Shiokawaら2013）
     従来のModularityベースのクラスタリングのノード数，エッジ数を削減することにより高速化を実現
  2. 等粒度クラスタリング（藤森ら2015）
     Modularityベースのクラスタリングに加え，分割数k，等粒度の度合いaを指定可能とする手法
  3. 構造的クラスタリング（Shiokawaら2015）
     ノード間の構造的類似度を計算し，クラスタ，ハブ（橋渡し役），外れ値の3属性に分類を行う手法

2.と3.の手法では，一番大きなクラスタのアカウント数が，他のクラスタに対し圧倒的に多くなる等のことから有意なクラスタリングはできなかったため，Modularityベースのクラスタリングを使用した

[Shiokawa 2013] Hiroaki Shiokawa, Yasuhiro Fujiwara, Makoto Onizuka: Fast Algorithm for Modularity-based Graph Clustering, In Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI 2013), Bellevue, Washington, USA, July 2013.
[藤森 2015] 藤森 俊匡，塩川 浩昭，鬼塚 真：分散グラフ処理におけるグラフ分割の最適化，第7回データ工学と情報マネジメントに関するフォーラム(DEIM2015)，E5-2，2015.
[Shiokawa 2015] Hiroaki Shiokawa, Yasuhiro Fujiwara, Makoto Onizuka: SCAN++: Efficient Algorithm for Finding Clusters, Hubs and Outliers on Large-scale Graphs, The Proceedings of the VLDB Endowment (PVLDB), Vol. 8, No. 11, pp. 1178–1189, 2015.

調査対象



- データ：ツイート
  - 収集条件
    - 日本語で，ある時点においてツイート数200以上のアカウント
  - 収集期間
    - 2011年3月5日から3月24日(20日間)
  - 収集ツイート数
    - 約3億件

- ネットワークの生成
  - 一日毎のコミュニケーションネットワークの構築
    - RT，Replyに基づき構築
    - 無向グラフ
    - 約115万ノード

クラスタの分析結果

2011年3月10日
震災前のクラスタリング

- ノード数：1,149,490
- エッジ数：6,334,207
- クラスタ数：578

No	クラスタID	人数	No	クラスタID	人数	No	クラスタID	人数
1	274	151435	11	424	30092	21	184	15966
2	226	133305	12	513	27953	22	16	15062
3	443	77657	13	530	26692	23	287	13831
4	124	69729	14	40	24769	24	114	10897
5	573	68435	15	43	24764	25	271	10302
6	420	53999	16	476	22647	26	365	9325
7	244	53647	17	209	22548	27	540	9232
8	353	46376	18	146	20176	28	149	8662
9	170	45315	19	447	20176	29	12	7451
10	92	44611	20	222	18377	30	375	7357

クラスタの分析結果

2011年3月10日
震災前のクラスタリング

- ノード数：1,149,490
- エッジ数：6,334,207
- クラスタ数：578

No	クラスタID	人数	No	クラスタID	人数	No	クラスタID	人数
1	274	151435	11	424	30092	21	184	15966
2	226	133305	12	513	27953	22	16	15062
3	443	77657	13	530	26692	23	287	13831
4	124	69729	14	40	24769	24	114	10897
5	573	68435	15	43	24764	25	271	10302
6	420	53999	16	476	22647	26	365	9325
7	244	53647	17	209	22548	27	540	9232
8	353	46376	18	146	20176	28	149	8662
9	170	45315	19	447	20176	29	12	7451
10	92	44611	20	222	18377	30	375	7357

クラスタリングに基づく重要アカウント抽出

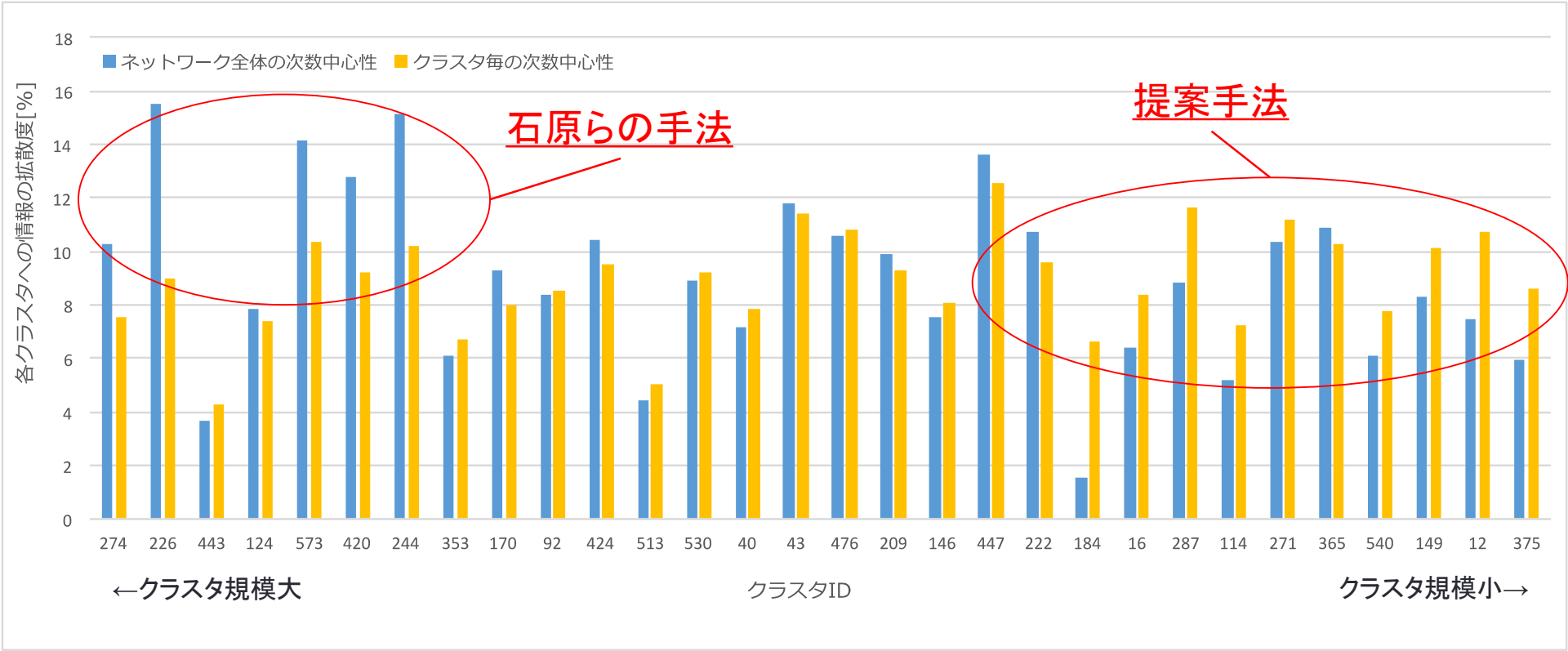
- 次数中心性の場合（赤字がクラスタリングによって新たに抽出されたアカウント）

次数順位	最大クラスタ	2番目のクラスタ	・・・	29番目のクラスタ	30番目のクラスタ
1	Swedenhills	setsulla		gerilon_bot	2nomatomeR
2	Yomiuri_Online	karashichan		tv1986	sheonite
3	47news	shuzo_matsuoka		santacloseeyes	filemente
4	mainichijpnews	now_fes		teru_0531	nyanyanchi
5	kenichiromogi	OttikiCharlie		tensyontakada	tonto_kaimo
6	rinrin_kit	zenra_bot		ryuuw0505	oO_Kizna310_Oo
7	Mujina30	Le_potiron		ka_10_yo	fukurow
8	gizmodojapan	htmk73		hirokisas	holineko
9	lgm_	scarletrain193		kx2880	msz3i
10	masason	ultrasoul_bot		tanjun_nayatu	poppoyatakasago

各クラスタ10アカウント × 30クラスタ = 300アカウント

クラスタリングに基づく手法の評価

- 「ネットワーク全体で重要とされているアカウント」と
  「クラスタ内で重要とされているアカウント」の比較（次数中心性）



→ 規模の大きなクラスタにはネットワーク全体から抽出する手法が，
小さなクラスタにはクラスタ内から抽出する提案手法が有用であった

研究目的

- 目的：偏りなく情報を拡散するための重要アカウントを抽出
  - 提案手法1：クラスタリングに基づく重要アカウント抽出手法
  - 提案手法2：ページランクに基づく重要アカウント抽出手法

ページランクに基づく重要アカウント抽出

- ページランクに基づく重要アカウント抽出手法の評価
  - 次数中心性，媒介中心性，ページランクに基づいて上位300アカウントを抽出し比較
- ページランクの算出方法（藤原ら2015）
  以下の式を再帰的に収束するまで繰り返し計算

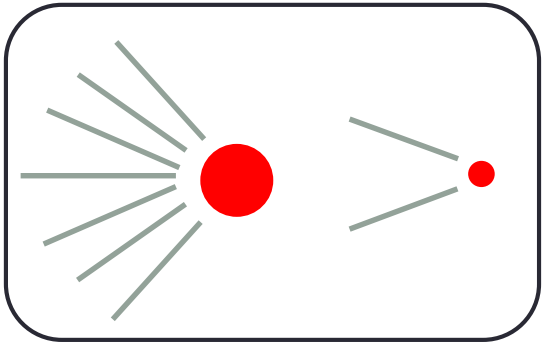
$$p_i = sWp_{i-1} + (1-s)e$$

1位	Cさん
2位	EFさん
3位	KJさん
⋮	⋮
298位	STさん
299位	SZさん
300位	VVさん

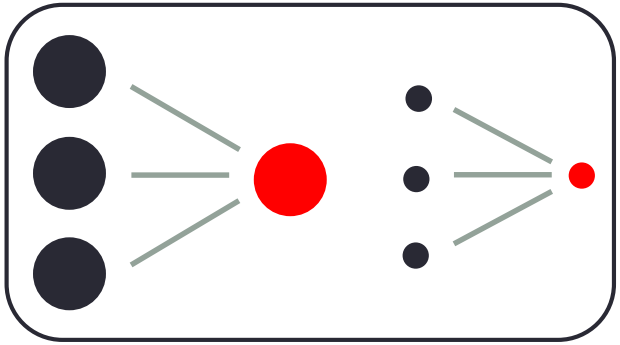
p：第u成分p[u]がノードuのページランクをスコアに対応するような列ベクトル
s：ランダムウォーク確率(0<s<1)

W：W[u,v]をノードvからノードuへ移動する確率とした時に列成分が正規化されたグラフの隣接行列
e：Nをグラフのノード数とした時に全ての成分の値が1/Nである列ベクトル

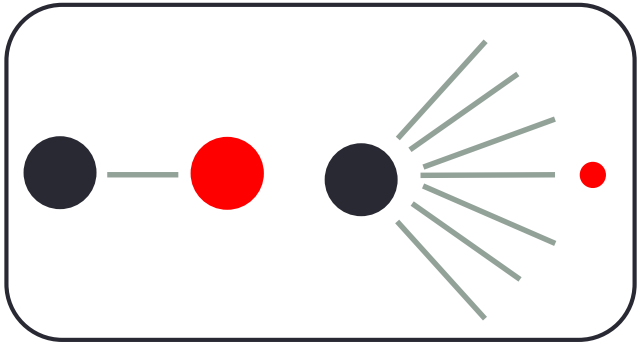
たくさん受け取る時



重要なノードから受け取る時



貴重なエッジを受け取る時



藤原靖宏, et al. "PageRank のための高速な Top-k 検索." 人工知能学会論文誌 30.2 (2015): 473-478.

ページランクに基づく重要アカウント抽出

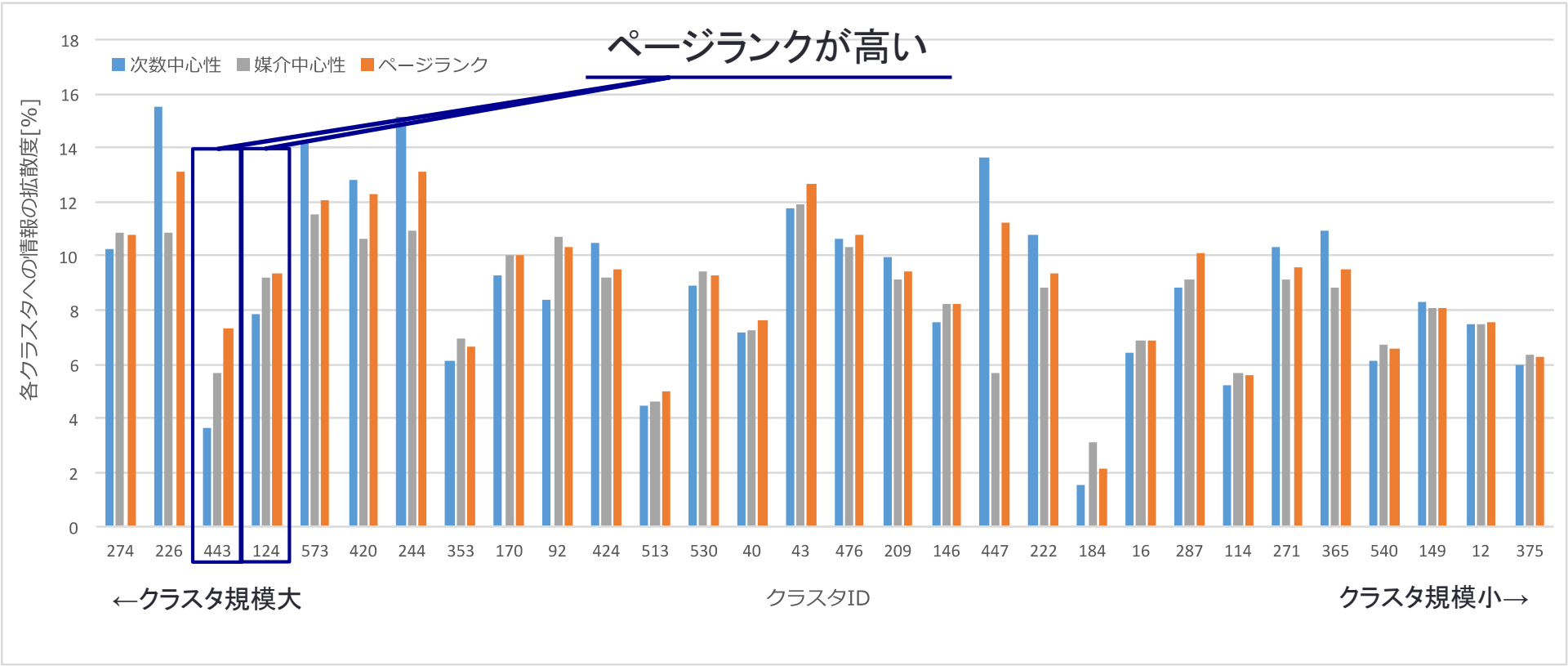
- ネットワーク全体から抽出した各指標毎の重要とされているアカウント

順位	次数中心性	媒介中心性	ページランク
1	youtube	youtube	youtube
2	shuumai	foursquare	shuumai
3	setsulla	shuumai	foursquare
4	natalie_mu	AddThis	SoalCINTA
5	wwwwww_bot	wwwwww_bot	natalie_mu
6	foursquare	swedenhills	MentionKe
7	swedenhills	natalie_mu	setsulla
8	Yomiuri_Online	setsulla	swedenhills
9	SoalCINTA	justinbieber	wwwwww_bot
10	47news	sazae_f	justinbieber
…	…	…	…

各指標300アカウント抽出

ページランクに基づく手法の評価

- 次数中心性，媒介中心性，ページランクの拡散度

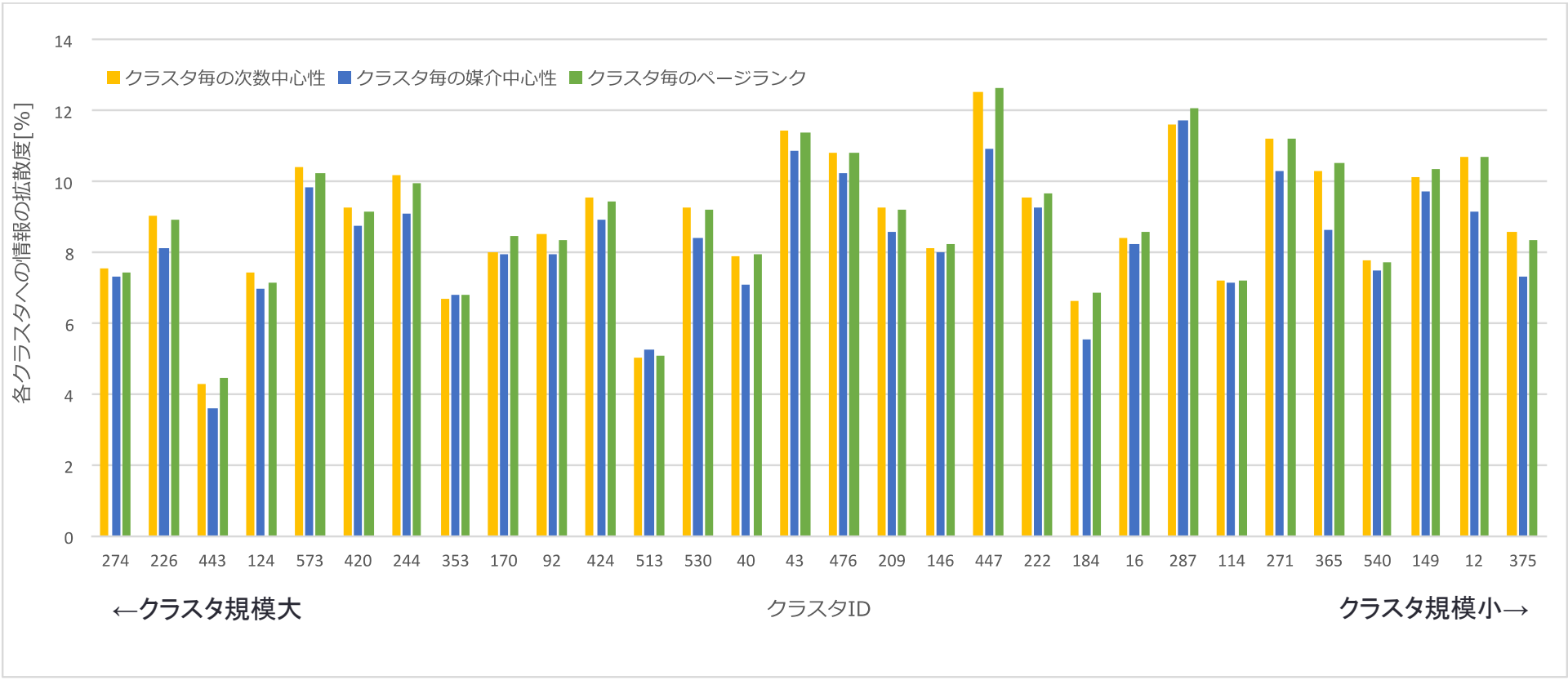


→ 各指標それぞれ相関性が高い

→ 海外クラスタ等の情報が届きにくいクラスタには
ページランクが有用であった

クラスタリング+ページランクに基づく手法の評価

- クラスタ毎の次数中心性，媒介中心性，ページランクから抽出した重要アカウントに基づいて求めた拡散度



→ 提案手法をミックスした手法ではページランク等の指標による違いは無し

考察

- 規模の大きなクラスタに対しては，ネットワーク全体から抽出する手法が，
  規模の小さなクラスタに対しては，クラスタ内から抽出する手法が効果的

- 次数中心性，媒介中心性，PRの各指標はそれぞれ相関性が高い

- 海外クラスタ等の情報が届きにくいクラスタに対しては，
  ページランクが効果的

- 今回は1ホップのみでみたため，次数中心性に有利な評価手法であった．
  今後は2ホップの調査が必要．

まとめ

- 災害時，情報流通のツールとしてTwitterは非常に有用である
  それをより活かすためには拡散能力の高い重要アカウントの発見が必要である

- 石原らはネットワーク全体から重要アカウントを抽出したが
  それでは情報に偏りが出てしまうため，
  クラスタリングに基づく重要アカウントの抽出が必要である

- 重要アカウントの抽出手法の選択は以下の表を基に行うのが
  効果的であることがわかった

	次数中心性	媒介中心性	ページランク
クラスタリング有り	規模の小さなクラスタ（提案手法）	ー	規模の小さなクラスタ（提案手法）
クラスタリング無し	規模の大きなクラスタ（既存手法）	ー	海外クラスタ（提案手法）

おわりに

- 大規模データを処理する場合、汎用的なツールでは不足
  - 本研究ではNTT様のGraponを使用してコミュニティ抽出
  - Graponは大規模・高速に解析が行えるため非常に助かった
- 今後もこういったツールが開発、発展されることを強く望む