3P1-NFC-00a-5

# 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

## Comparison of diffusiveness of information based on clustering in the Great East Japan Earthquake

秦恭史[*1] 諏訪博彦[*1] 岸本康成[*2] 藤原靖宏[*2] 新井淳也[*2]
Hata Kyoji Suwa Hirohiko Kishimoto Yasunari Fujiwara Yasuhiro Arai Junya

飯田恭弘[*2] 岩村相哲[*2] 鳥海不二夫[*3] 安本慶一[*1]
Iida Yasuhiro Iwamura Soutetsu Toriumi Fujio Yasumoto Keiichi

[*1] 奈良先端科学技術大学院大学 Nara Institute of Science and Technology
[*2]NTT ソフトウェアイノベーションセンタ NTT Software Innovation Center
[*3] 東京大学 The University of Tokyo

Abstract : In the event of a disaster, it is important to spread information extensively to every corner, and as one of them, Twitter is useful as a tool for information distribution as evidenced by related research. In order to make more use of Twitter, it is necessary to find important accounts with high spreading ability, Ishihara et al. extract important accounts as one network as a whole network. However, according to Amac et al., the Twitter network is thought to have close connection among close people, and it is considered to be divided into multiple clusters. As a result, the overall important account is not always important to everyone. Therefore, we do clustering using high speed graph mining technology corresponding to large scale Twitter network and extract important accounts by using page rank.

## 1. はじめに

近年，日本の災害事情は深刻化してきており，このような災害時，いち早く情報を手に入れることは生死に関わる．そこで速効性という観点で注目されているのが Twitter 等の SNS である．実際に東日本大震災時，Twitter が災害情報の拡散において有用である事が総務省により証明されている．

このような背景において[石原 2016] は，Twitter をより活かすためには情報拡散能力の高いアカウントが重要となると考え，ネットワーク全体を一つのネットワークとして重要アカウントを抽出した．しかしながら，[AmaÇ 2013] によれば Twitter ネットワークは趣味嗜好の近い人同士でつながりを持つものと考えられる．すなわち，ひとつの均一なネットワークだけでなく，複数のクラスタに分かれたネットワークと考えられる．そのため，ネットワーク全体で重要とされるアカウントが，すべての人々にとって重要であるとは限らない．結果として，単にネットワーク全体から重要アカウントを抽出しても，効率的な情報流通には貢献できないと考えられる．

本研究では，クラスタリングに基づく重要アカウントの抽出を試みる．しかし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クラスタリングに膨大な時間とハイスペックなマシンを必要とするため，迅速な対応が迫られる災害時には不適切であった．この課題に対し我々は，大規模ネットワークに対しても高速にクラスタリング可能な Modularity ベースのクラスタリング[飯田 2015] を行うことで，ネットワークの分割を行う．

また，[石原 2016]は，次数中心性と媒介中心性に着目して重要なアカウントを抽出している．震災時のツイートネットワークは，リプライ，リツイートによるコミュニケーションネットワークとなっており，単にそのアカウントの次数や媒介性だけでなく，他のどのようなアカウントとつながっているかが重要となる．そのため，本研究では新たな重要アカウント抽出指標としてページランクに着目する．ページランクは，「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 Web ページ間のリンクから，Web ページのランク付けを行う指標である．この考えを採用し，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とるという考えに基づいて重要なアカウントの抽出を試みる．ページランクは，石原らが使用した次数中心性や媒介中心性よりも情報拡散時の重要アカウントをより適切に判断できると考える．大規模クラスタリングおよびページランクにより，災害時に情報を広範囲に隅々まで広げる情報拡散手法を実現する．加えて，提案手法を評価するために，情報の拡散を計る指標を定義する．

予備調査として情報の拡散に偏りがあることを明らかにする．まず，Twitter ネットワークが，均一なひとつのネットワークではなく，複数のクラスタをもつネットワークの集合であることを確認するために，大規模クラスタリングを行い各クラスタの特徴を把握する．その後，情報拡散の偏りを「関西電力の節電呼びかけチェーンメール」のデマ情報拡散の事例 に基づいて調査する．具体的には，このデマの情報が各ネットワーク（クラスタ）においてどの程度拡散されたのかを割合で算出する．クラスタリングの結果，1,149,490 のアカウントが 578 のクラスタに分かれ，一番規模の大きなクラスタのアカウント数が 151,435 アカウントとなった．また，規模の大きなクラスタには有名人のクラスタや Web サービスのクラスタ，海外アーティストのクラスタ等が存在した．また，デマのツイートをしたアカウントが各クラスタにどの程度存在しているのかを調べた結果，クラスタによって偏りがあったことから，ネットワーク全体において情報の拡散に偏りがあることが確かめられた．

提案手法の評価のために，東日本大震災前後である 2011 年 3 月 10 日と 12 日のデータを使用し検証を行う．分析は従来手法と提案手法を用いた場合の情報拡散度の比較により行う．検証の結果，小規模なクラスタに対してはクラスタリングに基づく重要アカウントの抽出手法が有効であることが確認できた．また，1 ホップの場合は次数中心性やページランクが有用なのに対し，2 ホップの場合は，媒介中心性が有用であることがわかった．

秦恭史 〒630-0192 奈良県生駒市高山町 8916-5 情報科学研究科 A 棟 4F Hata Kyoji hata.kyoji.gx0@is.naist.jp

## 2. 提案手法

本章では，情報拡散の偏りを起こさず，拡散度の高い重要アカウントの抽出手法について提案する．提案する手法は，ページランクに基づく手法と，クラスタリングに基づく手法である．

### 2.1 ページランクに基づく重要アカウント抽出手法

石原らの手法は，次数中心性と媒介中心性による重要アカウントの抽出を行っていた．次数中心性とは，ノード同士を繋ぐ関係の多寡によりノードの重要性を評価する指標で，情報拡散の基点となるアカウントの特定に使用していた．媒介中心性は，ノード間の連結関係上の重要性を評価する指標で，情報を仲介するアカウントの特定に使用していた．

我々は，新たな指標としてページランクに着目する．ページランクは，Google の検索エンジンのページ表示ランキングに使用されていたもので，web ページの重要度を計る指標の一つとして使われている．この指標は，被リンク数とその質により決定され，リンクがより集まっているページは重要であると定義されている．ここで我々は，twitter のコミュニケーションネットワークにおいて，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とるという仮説を立てた．

この仮説に基づいて，我々は，ページランクによる重要アカウントの抽出手法を提案する．具体的には，Twitter ネットワーク全体からページランクを算出し，ページランクが高い n アカウントを重要アカウントとして抽出する．

また，この提案手法の評価方法として，抽出された重要アカウントから 1 ホップ離れたアカウントが，各クラスタにどれくらいの割合で存在しているのか（情報拡散度）を算出する．拡散度は以下の式で表される．

$$Diffusivity(n) = \frac{AA(n)}{AiC}$$

$Diffusivity(n)$：対象のクラスタの拡散度
$AA(n)$：対象のクラスタ内における
　　1 ホップ離れたアカウントの数
$AiC$：対象のクラスタ内のアカウント総数

既存手法（次数中心性，媒介中心性を用いて重要アカウントを抽出）も同様に各クラスタにおける拡散度の算出を行い提案手法と比較する．これにより，抽出された重要アカウントが，情報拡散においてどの程度の拡散力を秘めているか検証する．

### 2.2 クラスタリングに基づく重要アカウント抽出手法

先行研究により，情報の拡散に偏りがあることが確認できた．そこで我々は，各クラスタから重要アカウントをそれぞれ抽出することによって，クラスタによる情報の偏りを軽減する手法を提案する．

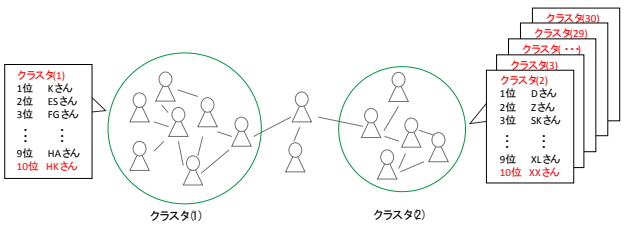


図 1 クラスタリングに基づく重要アカウントの抽出手法

提案手法は以下の通りである（図 1）．まず，クラスタリングを行い，クラスタ内の所属アカウント数が多い上位 k クラスタを抽出する．その抽出したクラスタ内で各指標に基づいた上位 n アカウント抽出し，2.1 節と同様に拡散度を算出する．なお，今回は，k=30，n=10 としている．これは，石原らが上位 300 アカウントを重要アカウントと議論しており，その数にあわせるためである．

クラスタリング手法としては，以下の 3 つの手法を検討した．

**Modularity ベースのクラスタリング：**Modularity とは与えられた分割に対して「グループ内のノード同士が繋がるリンクの割合」から「リンクがランダムに配置された場合の期待値」を引いた値として定義されるものである．この値が良いほど，より適切にグラフデータ内のクラスタを抽出できていることを示す．従来の Modularity ベースのクラスタリングではこのような大規模ネットワークにおけるクラスタリングは膨大な計算コストがかかるため難しかったが，本研究で使用した Grapon の Modularity ベースのクラスタリングは，計算対象となるノードとエッジの数を削減することにより，高速化を実現している [Shiokawa 2013].

**等粒度クラスタリング：**基本的には Modularity ベースのクラスタリングと同様に処理を行う．これに加えてグラフの分割数を k，クラスタの等粒度の度合いを決めるパラメータ a を指定可能とした手法である．この手法ではクラスタリングの進行中にクラスタ数が k×a を下回ると等粒度化に向けてクラスタをマージしていく．こうすることにより，各クラスタの大きさが同程度になるように柔軟にグラフを分割でき，並列処理に適したネットワークの分割が行える[藤森 2015].

**構造的クラスタリング：**ノード間の構造的類似度を計算し，閾値を超える類似度のノード群をクラスタに，複数のクラスタに接続しているノードをハブ（橋渡し役）に，それ以外のノードを外れ値として分類を行う手法である．この手法は主に橋渡し役となるノードを見つけることに特化している[Shiokawa 2015].

なお，今回の調査では Modularity ベースのクラスタリング以外の場合，一番大きいクラスタの人数が，他のクラスタに対し圧倒的に大き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．そのため，以降の分析は行わなかった．また，Modularity ベースのクラスタリングを行った結果，震災前はクラスタ数が 578 個に分類され，震災後では 346 個に分類された．

## 3. 重要アカウントの抽出と評価

本章では，2 章で提案した手法に基づいて重要アカウントの抽出と評価を行う．分析対象は，震災前である 2011 年 3 月 10 日と，震災後である 3 月 12 日に日本語で Twitter に投稿されたツイートである．この取得したツイートの一日毎のコミュニケーション(リプライ，リツイート)に基づいて無向ネットワークを作成する．

### 3.1 分析概要

提案手法を評価するために，従来手法と提案手法で抽出されるアカウントを比較する．また，それぞれの情報拡散度を比較する．重要アカウントを抽出する手法は，以下の 6 手法である．

1.次数中心性に基づいて抽出
2.媒介中心性に基づいて抽出
3.ページランクに基づいて抽出
4.クラスタ毎に次数中心性に基づいて抽出
5.クラスタ毎に媒介中心性に基づいて抽出
6.クラスタ毎にページランクに基づいて抽出

手法 1 から 3 までは上位 300 アカウントを，手法 4 から 6 はクラスタ毎に 10 アカウントを所属アカウント数上位 30 クラスタ分で計 300 アカウントを重要アカウントとして抽出する．大規模な Twitter ネットワークのページランク計算には F-Rank[藤原 2015]を用いた．

### 3.2 分析結果

抽出した重要アカウントの例を表 1，表 2 に示す．表 1 より，それぞれの指標順で並べた場合，近しい順位に同じアカウントがあることが多いことが確認できる．このことから，次数中心性，媒介中心性，ページランクはそれぞれと相関が高いことがわかる．この傾向は，クラスタリングした後でも同様に確認された．

しかし，特異的なアカウントがいくつか確認された．表 1 において他指標と比べて明らかにページランクが高かったアカウントは，MentionKE や papatah，justinbieber や eseMendiola など，海外のアカウントであった．これらのアカウントは，予備調査で明らかになった情報があまり届かなかったクラスタに所属するものであり，情報の偏りを是正する効果が期待できる．

表 1 ネットワーク全体の各指標上位 30 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 |
| 11 | karashichan | 47news | Yomiuri_Online |
| 12 | sazae_f | SoalCINTA | sazae_f |
| 13 | mainichijpnews | Yomiuri_Online | mainichijpnews |
| 14 | kenichiromogi | rinrin_kit | 47news |
| 15 | issonson_8374 | MentionKe | kenichiromogi |
| 16 | rinrin_kit | Mujina30 | pepatah |
| 17 | Mujina30 | viratter | AddThis |
| 18 | MentionKe | kenichiromogi | gizmodojapan |
| 19 | gizmodojapan | shuzo_matsuoka | issonson_8374 |
| 20 | lgm_ | karashichan | rinrin_kit |
| 21 | shuzo_matsuoka | kakusan_RT | DamnItsTrue |
| 22 | now_fes | issonson_8374 | Mujina30 |
| 23 | justinbieber | lgm_ | BangMir |
| 24 | d_v_osorezan | winwin88 | lgm_ |
| 25 | AddThis | mainichijpnews | mariko_dayo |
| 26 | mariko_dayo | twinavi | masason |
| 27 | ogiri_tweet | BangMir | shuzo_matsuoka |
| 28 | masason | gizmodojapan | karashichan |
| 29 | OttikiCharlie | masason | cnet_japan |
| 30 | batounohito60 | DamnItsTrue | twinavi |

表 2 上位 3 クラスタのページランク上位 10 アカウント

| ページランク順位 | クラスタ ID:274 | クラスタ ID:226 | クラスタ ID:443 |
|---|---|---|---|
| 1 | swedenhills | setsulla | justinbieber |
| 2 | Yomiuri_Online | karashichan | eseMendiola |
| 3 | mainichijpnews | shuzo_matsuoka | JavaJoeMyspace |
| 4 | 47news | OttikiCharlie | yarrjerrica |
| 5 | kenichiromogi | htmk73 | professor_adail |
| 6 | gizmodojapan | Le_potiron | Sexstrology |
| 7 | rinrin_kit | scarletrain193 | JiNxBeatz |
| 8 | Mujina30 | souha00 | Mazaroddi |
| 9 | lgm_ | zenra_bot | iUsX |
| 10 | masason | now_fes | OfficialJaden |

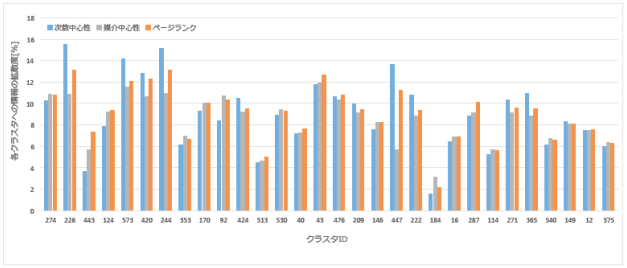


図 2 各中心性指標における拡散度の比較

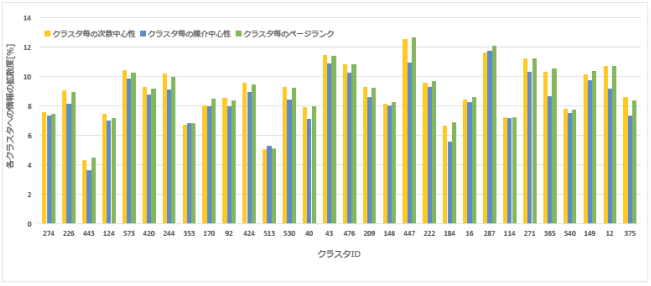


図 3 クラスタリングにおける拡散度の比較

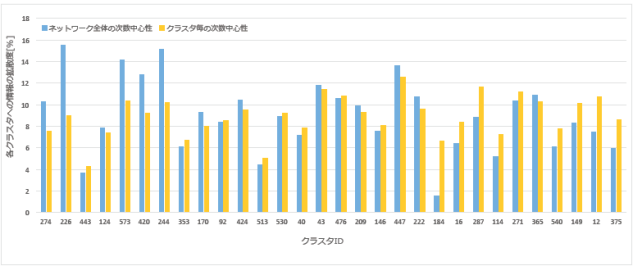


図 4 クラスタリングの有無による拡散度の比較

図 2，図 3，図 4 は，次数中心性，媒介中心性，ページランクの 3 つの指標において抽出した重要アカウントについて，そこから 1 ホップ離れているアカウントが，所属アカウント数上位 30 クラスタの中にどれだけ含まれているかの割合を表したグラフである．横軸のクラスタ ID については，図中で左に示したものほど，サイズの大きいクラスタとしている．すなわち，左端のクラスタ ID274 が，図の横軸に示した上位 30 クラスタのうち，最も大きいクラスタであり，右端のクラスタ ID375 が最も小さなクラスタである．

図 2 は次数中心性，媒介中心性，ページランクから重要アカウントを抽出した結果である．クラスタ毎に違いはあるものの，媒介中心性よりもページランクの方が高い拡散度であることが確認できる．また，この傾向が特に見られたのはクラスタ ID:124 や 443 などの 3.4.節で確認した情報があまり届かなかった海外音楽クラスタや海外 web サービスのクラスタであった．

図 3 はクラスタ毎の次数中心性，媒介中心性，ページランクから重要アカウントを抽出した結果である．図 2 に比べて，各クラスタ内の情報拡散度が増加しているのがわかる．クラスタ毎に見ていくと，媒介中心性が他の指標に比べて若干低いことが確認できる．また，この図からも次数中心性，媒介中心性，ページランクの相関性が高いことがわかる．

図 4 は，次数中心性に基づいてネットワーク全体から重要アカウントを抽出する手法（従来手法）と，次数中心性に基づいてクラスタ毎から重要アカウントを抽出する手法（提案手法）の情

報拡散度を比較した結果である．横軸は，クラスタに所属する人数が多い順にクラスタを並べていることに注意すると，グラフの左に位置する所属アカウント数が多いクラスタは従来手法が，アカウント数が少ないクラスタは提案手法が，高い情報拡散度を示すことがわかる．また媒介中心性，ページランクに関しても同様の比較を行った場合，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重要アカウントを抽出する石原らの手法が，規模の小さなクラスタに対しては，クラスタ毎に重要アカウントを抽出する提案手法が効果的であることが同様に確認できた．

しかし，1 ホップでの評価は次数中心性に有利な評価であるため 2 ホップ先のアカウントまで同一クラスタに所属しているかどうかの評価を実施した．その結果を図 5，図 6，図 7 に示す．

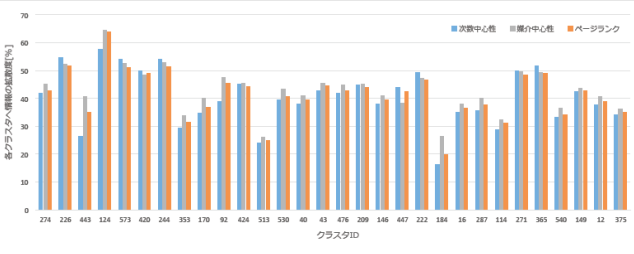


図 5　各中心性指標における拡散度の比較(2 ホップ)

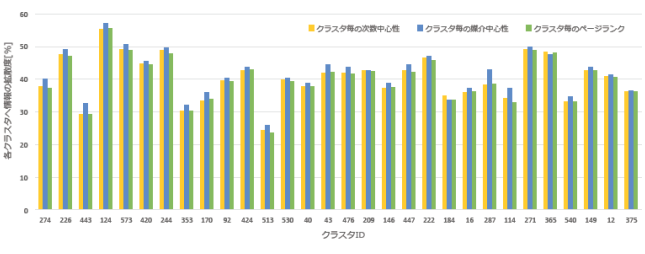


図 6　クラスタリングにおける拡散度の比較(2 ホップ)

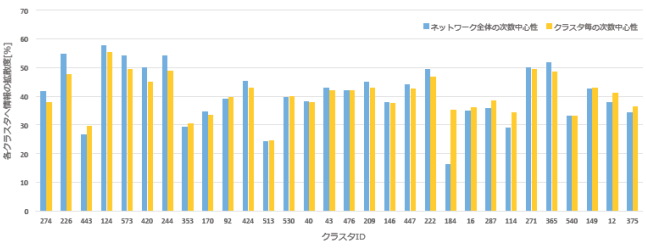


図 7　クラスタリングの有無による拡散度の比較(2 ホップ)

図 2，図 3，図 4 のそれぞれと，図 5，図 6，図 7 のそれぞれとを比較すると，重要アカウントの 2 ホップ先までが同一クラスタに含まれているかについて調べた後者のほうが，情報拡散度が大きくなっていることが確認される．また，図 3 と図 6 を比較すると，図 6 においては，ほぼ全てのクラスタにおいて媒介中心性による拡散度が良い結果を示している．これは 2 ホップ先まで見たことにより，1ホップの時に見ることができなかった媒介中心性の利点が活かされた結果であると考えられる．ただ今回は，2 ホップ先の拡散度を算出する際に全ての1ホップ先のアカウントが情報を拡散させたという仮定のもと拡散度を算出しているため，一概に媒介中心性が有用であるとは言えない．よって今後は 1 ホップ先のアカウント数を絞って 2 ホップ先の拡散度を算出することが求められる．

また，クラスタサイズが小さいほど，ネットワーク全体の中心性指標で情報拡散度を評価するよりも，クラスタ毎に中心性指標で情報拡散度を評価するほうが，高い情報拡散度を示すという点は，情報の拡散力を 1 ホップ先までのアカウントで評価するか 2 ホップ先までを評価するかに関わらず，同じ傾向が見られた．

## 4. 結論

本研究では，石原らが行った次数中心性，媒介中心性による重要アカウントの抽出に対し，ページランクに基づく重要アカウントの抽出手法の提案と，クラスタリングに基づいた重要アカウントの抽出手法の提案を行った．東日本大震災時における twitter のコミュニケーションネットワークに基づいて，重要アカウントの抽出と情報拡散度の評価を行った．その結果，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重要アカウントを，規模の小さなクラスタに関しては各クラスタから重要アカウントを抽出することが必要であることがわかった．加えて，1 ホップの場合は次数中心性やページランクが有用なのに対し，2 ホップの場合は，媒介中心性が有用であることがわかった．

また，拡散させる情報によって，情報の拡散具合に影響が出ることが示唆された．よって，拡げたい情報に合わせて重要アカウントを抽出することが必要になると考えられる．

## 参考文献

[石原 2016] 石原裕規，諏訪博彦，鳥海不二夫，太田敏澄：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.J99-D，No.5，pp.501-513，2016.

[AmaÇ 2013] HerdaĞdelen AmaÇ, Zuo Wenyun, Gard-Murray Alexander, Bar-Yam Yaneer: An exploration of social identity: The geography and politics of news-sharing communities in twitter, COMPLEXITY, Volume 19, Issue 2, Pages 10-20, November/December 2013.

[飯田 2015] 飯田恭弘，岸本康成，藤原靖宏，塩川浩昭，新井淳也，岩村相哲：大規模グラフ向けの先進的な処理・分析技術，NTT 技術ジャーナル，Vol.27，No.12，pp.24-28，2015 年 12 月．

[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.

[藤原 2015] 藤原靖宏，中辻真，塩川浩昭，三島健，鬼塚真，PageRank のための高速な Top-k 検索，人工知能学会論文誌 30(2)，pp.473-4778，February 2015.