人工知能学会研究会資料
SIG-KBS-B401-10

# Twitterのリンクベースでの会話コミュニティ抽出とコミュニティ内の単語使用特性分析

# Extraction of Link Communities on the Conversational Network in Twitter and Analysis on Word Usage among the Communities

丸井 淳己 [1*] 則 のぞみ [2] 榊 剛史 [1,3] 森 純一郎 [1]
Junki Marui[1] Nozomi Nori [2] Takeshi Sakaki [1,3] Junichiro Mori[1]

[1] 東京大学大学院工学系研究科
[1] School of Engineering, the University of Tokyo
[2] 京都大学大学院情報学研究科
[2] Kyoto University, Graduate School of Informatics
[3] 株式会社ホットリンク
[3]Hottolink, Inc.

**Abstract:** It is now common to have a conversation with others on social media. Many research have been taken to see the community structure on social media, but there are few studies that apply link-based community (link community) detection on a large social network. Link community detection allows users to belong to more than one community. We improve the method of existing link community detection of Ahn et al., which extracts many small communities. We evaluate existing and proposing methods by network indexes, and we characterize link communities from users' biographies. We found that link communities sharing users have similar characteristics from biographies.

## 1 はじめに

近年ソーシャルメディア上で多くの人々が情報の発信や日常的な会話を行うようになり，どのような人々が互いに情報のやりとりしているかについて大規模なデータを用いて分析することが可能になりつつある．特に Twitter ではそのような分析が多く行われ，ツイートと呼ばれる短文テキストを大規模に取得し，ユーザがツイート内に用いるハッシュタグや共有する動画の URL を用いて，感情分析を行ったり個人の属性や意見を推定したりする研究が行われている [Wang 11, Abisheva 14]．一方で分析の対象を個人からコミュニティへと拡大して分析を行う研究も多く，Twitter 上で購読する・される関係であるフォロー・フォロワー関係のネットワークからコミュニティを抽出し，コミュニティの性質をツイートに用いられる単語から推定する研究もなされている [Bryden 13].

また Twitter 上ではツイート内に「@アカウント名」と入れることで明示的に他のユーザを言及する (メンションと呼ばれる) ことができ，これを使って会話が行われている．このような会話のあるユーザの関係はフォロー・フォロワー関係よりも強い関係と考えることができ，特にソーシャルメディア上のターゲティング広告やバイラルマーケティングといったウェブ分野における応用につながると期待できる．著者らは Twitter の会話ネットワークからコミュニティを抽出し，その性質をプロフィール情報から特徴付けした [丸井 14]．この分析ではコミュニティ上位 38 個を取り出したが，ノードベースでコミュニティ抽出をしたため構成員は常に単一のコミュニティに属している．しかしこの分析では高校や大学といった属性ベースのコミュニティと，アニメや自転車ツーリングと言った趣味ベースのコミュニティに大別され，属性と趣味または複数の趣味のコミュニティに属するユーザもいると考えられる．またユーザを共有しているコミュニティ同士も情報が伝播しうる関係となり，コミュニティの連関をもより詳しく見ることが可能となる．

そこで本研究では，ユーザを介したコミュニティの関係を見るために，Twitter の会話ネットワークからリンクベースでコミュニティ抽出を行うことで，ユーザ

*連絡先：東京大学大学院工学系研究科技術経営戦略学専攻
〒 113-8656 東京都文京区本郷 7-3-1 工学部 3 号館 202
E-mail: marui@ipr-ctr.t.u-tokyo.ac.jp

が複数のコミュニティに含まれるような抽出を行う．リンクベースで行うとコミュニティがどのように抽出されるか見た上で，本研究の目的に沿う形で既存手法を改善する．その上で抽出されたコミュニティの性質を構成ユーザのプロフィール文から特徴付けを行い，コミュニティ間の関係を見る．本研究のようにリンクでのコミュニティ抽出を数百万ノードの社会ネットワークに適応した研究は少ないため，バイラルマーケティングといった応用から見ても，ユーザを介したコミュニティ間の関係についての有用な知見が得られると期待できる．

## 2　関連研究

コミュニティ抽出はソーシャルメディアでなく生物学の見地からも盛んに研究されるようになっている．ネットワークのグラフ構造からのコミュニティ抽出は応用先の多さから盛んに研究され，Newman は Modularity という指標を提案し，Modularity を最大化するようにリンクを切っていくことで良いコミュニティ分割がなされるとした [Newman 04]．Modularity を最大化する考えを用いてより高速なコミュニティ抽出を提案した研究も多くある [Clauset 04, Blondel 08]．

以上のような方法はリンクを切ってコミュニティを作っていくため，ノードは常に単一のコミュニティに属する．その一方でノードを複数のコミュニティに属するように抽出する方法も近年提案され，Ahn らはあるノードから伸びる 2 本のリンクに対して，リンク先ノードの隣人ノードが重なる割合から類似度を計算し，類似度からリンクを階層クラスタリングすることでリンクベースでのコミュニティ抽出を可能にした [Ahn 10]．従来よく研究されてきた *Zachary Karate Club* やタンパク質相互作用のネットワークでいくつかの指標を用いてリンクコミュニティの性質を調べている．しかし Ahn らの手法は数百万ノードといった大規模な社会ネットワークにおいてどのようにコミュニティ抽出されるかについて明らかにしていない．そこで本研究では Twitter の大規模な会話ネットワークに対して適用しその効果を見るとともに，より良いコミュニティ抽出を行うために手法を改善する．

## 3　リンクコミュニティ抽出

本章ではリンクベースでのコミュニティ抽出の基本的な考え方を述べた後，改善を行った点を中心に提案手法を述べる．ここで得られたコミュニティをリンクコミュニティと呼ぶ．

### 3.1　基本的な考え方

本節では前述の Ahn らによる手法を最初に述べる [Ahn 10]．まずノードを共有するリンクのペアを取り出す．図 1 の A-B, A-C のようなリンクを取り出した後，このリンクペアのスコアを計算する．リンクペアのスコアは，共有していないノード同士 (図では B,C) の共通隣接ノードのジャカード係数で計算される．この計算時にループバックするリンクを加えるため，ノードが直接つながっている場合には互いが隣接ノードとなる (図の B-C のリンク)．図では隣接ノードが A,B,C,F であり，また B,C の隣人の和集合は 8 であるため，ジャカード係数を取ると 1/2 ということになる．よって A-B,A-C のリンクペアのスコアは 1/2 となる．

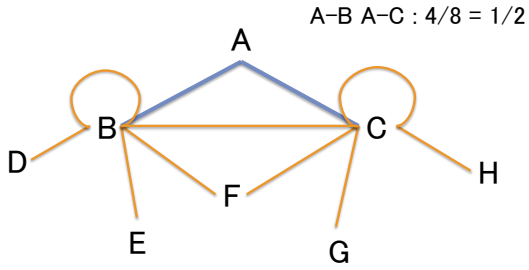


図 1: リンクペアのスコアの取り方

このような方法ですべてのリンクペアについて計算を行った後，リンクペアのスコアを使って最短距離での階層的クラスタリングを行う (図 2)．クラスタリングを以下の式を使って評価する．

$$D_c = \frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)} \quad (1)$$

$$D = \frac{1}{M} \sum_c m_c D_c \quad (2)$$

全体のリンク数が $M$, クラスタ $c$ のノード数が $n_c$, リンク数が $m_c$ としている．式 (2) は式 (1) のリンク数での重み付き平均になっていて，Ahn らはこの指標を *Partition Density* (図 2 の Score にあたる) と呼び，最大値となるところで階層的クラスタリングを区切れば良いとした．

リンク数が多い場合階層的クラスタリングを厳密に行うのは多大な計算時間を要する．しかし最短距離での階層的クラスタリングであることを考えると，閾値を設定した上で閾値を下回るリンクペアを削除し，その連結成分を同じラベルに振っても同じようにリンクコミュニティに分けることができる．彼らによる C++ 実装はこのような手法で実装されていて，以下閾値を設定しつつクラスタリングを行うこの手法を用いる．

この手法で実験したところ，多くが単一リンクの最小コミュニティとなったため，以下の提案手法でこの

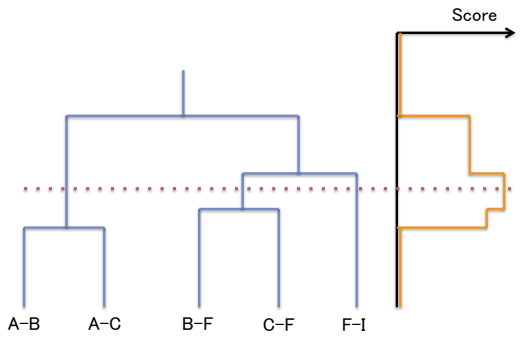


図 2: リンクペアの階層的クラスタリング

点を改善する．主に改善できる箇所はリンクペアのスコアの取り方である．今回のネットワークには重みを定義することができ，重みを考慮することによって同一スコアとなるリンクペアを減るので，より良いクラスタリングが行われると期待できる．また，いくつかのリンクペアは単一ノードを共有しているだけであるのに同じコミュニティに振られている．リンクペアが単一ノードを共有するのは要請から明らかであるので，共有された単一ノードを計算から外すことで改善が図られると考えられる．

### 3.2 提案手法

ネットワークの重みがある場合について Ahn らは言及しているものの，その効果は示されていない．また図 1 の例で A-B,A-C の隣接ノードを取る際に A が含まれているため，B,C の次数が低い場合は A を介してつながっているというだけで A-B,A-C のリンクペアのスコアが上がってしまう．この 2 点について以下の改善を行った．

まず式 (1) に重みを導入する場合には，いくつかの方法が考えられ，Ahn らは Jaccard 係数の拡張である Tanimoto 係数 [Tanimoto 57] を用いることができるとしている．今回は [Ioffe 10] 等いくつかの論文で Weighted Jaccard と言及される以下の指標を用いる．

$$J(\mathbf{S},\mathbf{T}) = \frac{\sum_k \min(\mathbf{S}_k, \mathbf{T}_k)}{\sum_k \max(\mathbf{S}_k, \mathbf{T}_k)} \tag{3}$$

この重み付きジャカード係数は非負成分のベクトル $\mathbf{S} = (S_k)$，$\mathbf{T} = (T_k)$ に対して定義される．この定義を用いると，ループバックの重みをいくつに設定するべきかが明確になる．図 1 の例では B-C が直接つながっているので，B のループバックと C のループバックの重みは B-C の重みと同一として定義すれば良い．式にまとめると以下のようになる．

$$w_{BB} = w_{CC} = \begin{cases} w_{BC} & \text{BC 間にリンクがある場合} \\ 0 & \text{BC 間にリンクがない場合} \end{cases} \tag{4a}$$

また，A-B,A-C のリンクペアのスコアを計算する場合には B と C の共通隣接ノードとして A をカウントしないようにする．このようにすることで，従来手法ではすべてのリンクペアは 0 より大きい値を持っていたが，提案手法では 0 となるリンクペアが生まれ，低次数の影響も受けにくくなる．最短距離で階層的クラスタリングを行うと大きなクラスタが生まれやすいので，スコアが 0 となるリンクペアが生まれるのはそのような現象を緩和することが期待される．

リンクペアに対して以上のようにスコアを計算した後に閾値を設定してクラスタリングを行い，式 (2) を使ってクラスタリングを評価する部分に関しては前節と同じである．

## 4 実験

本章では前章の従来手法と提案手法を Twitter の会話ネットワークに適用した結果を述べる．

### 4.1 データセット

2012 年 1 月 1 日から 12 月 31 日に渡って，日本語でツイートをしていると TwitterAPI で判定されたユーザを対象に，プロフィールとタイムスタンプ付きのツイートを集めた．取得された 49 億ツイートの中から，会話のつながりを取り出すためにメンション付きのツイートを抽出した．この期間に 1.24 億ペアが相互にメンションを行い，ペアが相互に行ったメンションの回数 (入次数と出次数の平均) をネットワークの重みとした．メンションを行ったユーザは 740 万であった．

### 4.2 手法の適用とその効果

従来手法については Ahn らが配布しているプログラムを[1]，提案手法についてはそのプログラムを改変した上で OpenMP で高速化を施し実行させた．従来手法の実装ではリンクペアのスコアの計算は Intel Xeon E5-2650 v2 2.6GHz のデュアルプロセッサ環境で 31 時間，提案手法の実装では 72 分であった．閾値を設定したクラスタリングは従来手法・提案手法ともほぼ同一

[1] http://barabasilab.neu.edu/projects/linkcommunities/ にて入手

のプログラムで，それぞれ一回あたり 50 分程度の所要時間であった．

それぞれの Partition Density は図 3,4 に示したとおりである．それぞれの最大値は従来手法で 0.0426357(閾値 0.223)，提案手法が 0.0513722(閾値 0.199) となり提案手法が上回る結果となった．

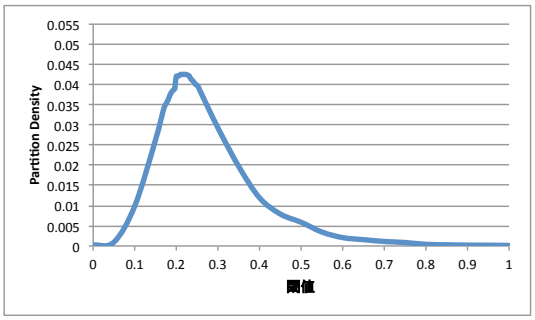


図 3: 従来手法の Partition Density 推移

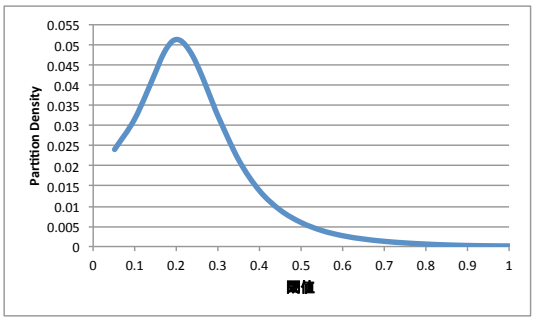


図 4: 提案手法の Partition Density 推移

Partition Density が最大値となるときのコミュニティをそれぞれ取り出し，以下議論する．Ahn らはコミュニティ分割の評価として *Community Quality*, *Overlap Quality*, *Community Coverage*, *Overlap Coverage* を提案している．そのうち今回のデータセットでネットワークから計算のできる Community Coverage と Overlap Coverage から評価する．

Community Coverage とは非自明なリンクコミュニティが全体のノードに占める割合であり，Overlap Coverage は非自明なリンクコミュニティにノードが平均していくつ所属しているかを示す指標と Ahn らは定義している．非自明なコミュニティとして所属ノード数が 3 以上のもの，100 以上のものという 2 つの定義でそれぞれの値を算出した．Community Coverage で所属ノード数 $n$ 以上のリンクコミュニティのみカウントしたものを $CC_n$，Overlap Coverage も同様に $OC_n$ として表 1 に示した．

表 1: Community Coverage と Overlap Coverage による手法の比較

| 手法 | $CC_3$ | $CC_{100}$ | $OC_3$ | $OC_{100}$ |
|---|---|---|---|---|
| Ahn ら | 68.9% | 11.7% | 2.725 | 0.141 |
| 提案手法 | 58.8% | 12.5% | 2.788 | 0.147 |

### 4.3 抽出されたリンクコミュニティ

Ahn らの方法は小さいリンクコミュニティがノードをカバーする割合は高いものの，大きいリンクコミュニティがカバーする割合は提案手法の方が大きくなる．ノードが所属するコミュニティの数はどちらの定義でも提案手法が上回った．

提案手法を用いてどのようなリンクコミュニティが抽出されたか調べるために，コミュニティに属しているユーザのプロフィール文からコミュニティの特徴付けを行った．ユーザ数が 500 以上の 51 コミュニティについてプロフィール文を収集し，コミュニティごとにプロフィール文をつなぎ合わせる．1 コミュニティ 1 文書として TF-IDF を計算し，上位の単語がそのコミュニティを表す単語として，人手によるラベリングを行った．その結果の一部は表 2 に示したとおりである．

最も大きいコミュニティは声優ファンのもので 6091 人，その後は格闘ゲーム・音楽ゲームのプレーヤー，自転車ツーリング愛好者と続いている．

それぞれのコミュニティのユーザがもし重複していれば，それらのコミュニティの関連は高いと考えられる．そこで，コミュニティメンバーの重複を重み付きのエッジとして表現して，コミュニティの関連を調べた (図 5)．最もユーザの共有が大きかったものは声優ファンとアイドル系アニメ・ゲームファンのコミュニティ間で 41 人，次いで AKB 系ファンとももクロファンの 38 人，J-POP ファンと三浦大知ファンの 29 人と続いている．このように共有するユーザが大きいコミュニティ同士はラベルを見ても関連しているものが多いことがわかる．

## 5 考察・議論

リンクコミュニティを抽出する方法について，Ahn らの手法とそれを改善した手法を通じて以上議論してきた．Ahn らの手法は Twitter の会話ネットワークに適用すると細かいコミュニティに分かれすぎてしまい，大きいコミュニティの量も少ないが，提案手法によって大きいコミュニティがより多く抽出できるようになり，コミュニティ分割の指標でも上回ることがわかった．最短距離での階層的クラスタリングは大きなクラスタが発生しやすく，閾値を上げてしまうとネットワークが

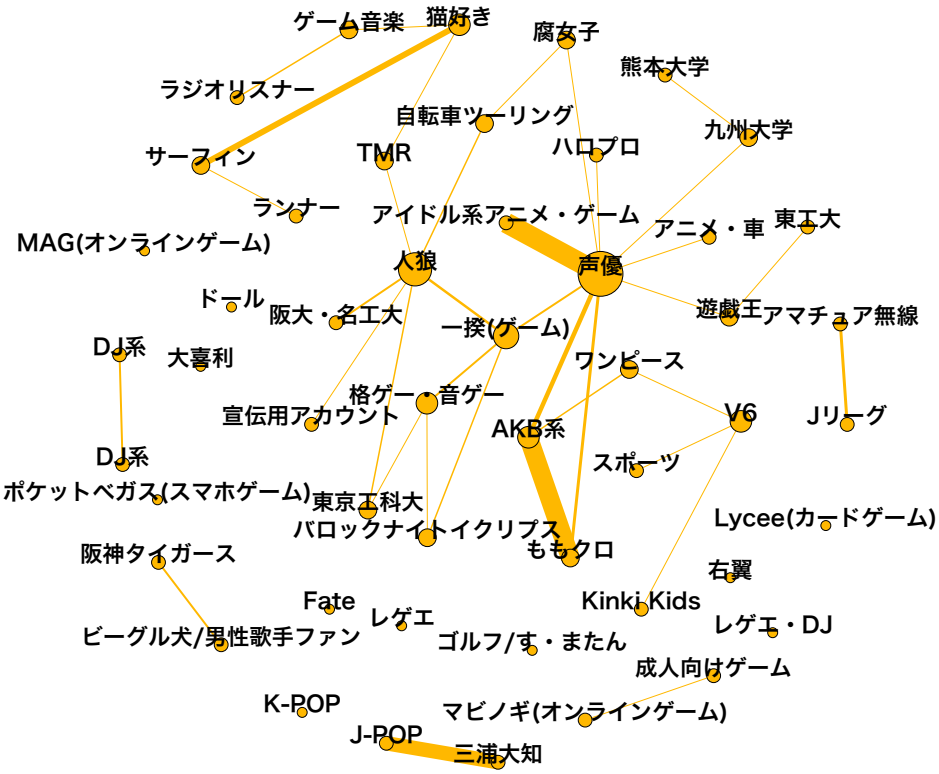


図 5: コミュニティ同士のユーザの重なり具合

表 2: リンクコミュニティのキーワードとラベルの一部

| | | | | |
|---|---|---|---|---|
| 人数 | 6091 | 2816 | 2064 | 1779 |
| ラベル | 声優 | 格ゲー・音ゲー | 自転車ツーリング | AKB 系 |
| キーワード 1 | スフィア | 勢 | ブルベ | 乃木坂 46 |
| キーワード 2 | 茅原実里 | KOF | JCRC | NMB |
| キーワード 3 | 奈々 | 格ゲー | 自転車 | まゆゆ |
| 人数 | 1711 | 1702 | 1511 | 1227 |
| ラベル | ももクロ | 遊戯王 | 人狼 | ポケットベガス |
| キーワード 1 | ももクロ | YP | 人狼 | ポケベガ |
| キーワード 2 | エビ中 | 遊戯王 | やる夫 | ベガ |
| キーワード 3 | あーりん | CS | 東方 | レイズ |
| 人数 | 1161 | 1088 | 1078 | 1040 |
| ラベル | アイドル系アニメ・ゲーム | レゲエ・DJ | 東京工科大 | バロックナイトイクリプス |
| キーワード 1 | ラブライブ | REGGAE | 虹裏 | PBW |
| キーワード 2 | みもり | SOUND | ボダ | バロックナイトイクリプス |
| キーワード 3 | アイマス | 毎月 | TUT | BNE |

切れすぎてしまう問題が発生してしまう．提案手法はこの問題を，(1) 重みを用いて類似度が分散されるようにし，(2) 関係性の薄いリンクペアのスコアを 0 とすることの 2 点で緩和している．従来手法では全く同じスコアをもつ複数のリンクペアがあった場合に，ある閾値で一気に結合してしまう問題があったが，(1) によってこの問題が緩和されている．また (2) によって関係性の薄いリンクペアが低い閾値で切れるのでよりまとまりのあるリンクコミュニティが取れるようになるとかんがえられる．

実際にリンクコミュニティを取り出すと，似た興味・属性の人が互いに会話していることが分かり，比較的

容易にラベル付けができることが分かった．また，声優ファンやアイドル系アニメ・ゲームファン，AKB系ファンとももクロファンなど，ユーザの重なりがあるリンクコミュニティ同士は，似た性質を持つことが多いことも分かった．ここから会話ネットワークから抽出されたリンクコミュニティは，強い興味対象でのまとまりがあることが推察される．

## 6 むすび

提案手法によってある程度の改善は見られたものの，最短距離での階層的クラスタリングをしている限り，クラスタが大きくつながりすぎてしまう問題は解決せず，その結果としてリンクコミュニティが小さいものしか取れない．将来的にこのクラスタリングの部分での改善が図られれば，より大きなコミュニティがいくつか取ることができ，より多用なコミュニティが観察されるだろう．

また今回はプロフィールの単語使用特性のみに着目したが，会話文を取り出してコミュニティ内でどのような会話が行われるか分析することで，Twitterのコミュニティ構造だけでなく情報伝播の観点からも有用な知見が得られるだろう．

## 謝辞

本研究は、独立行政法人新エネルギー・産業技術総合開発機構（NEDO）「学術・産業技術俯瞰システム開発プロジェクトの支援により行われた。

## 参考文献

[Abisheva 14] Abisheva, A., Garimella, V. R. K., Garcia, D., and Weber, I.: Who Watches (and Shares) What on Youtube? And when?: Using Twitter to Understand Youtube Viewership, in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pp. 593–602, New York, NY, USA (2014), ACM

[Ahn 10] Ahn, Y.-Y., Bagrow, J. P., and Lehmann, S.: Link communities reveal multiscale complexity in networks, *Nature*, Vol. 466, No. 7307, pp. 761–764 (2010)

[Blondel 08] Blondel, V., Guillaume, J., Lambiotte, R., and Mech, E.: Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, pp. 10008–10019 (2008)

[Bryden 13] Bryden, J., Funk, S., and Jansen, V. A. A.: Word usage mirrors community structure in the online social network Twitter, *EPJ Data Science*, Vol. 2, No. 1 (2013)

[Clauset 04] Clauset, A., Newman, M. E., and Moore, C.: Finding community structure in very large networks, *Physical review E*, Vol. 70, No. 6, p. 066111 (2004)

[Ioffe 10] Ioffe, S.: Improved consistent sampling, weighted minhash and l1 sketching, in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 246–255IEEE (2010)

[Newman 04] Newman, M. E.: Fast algorithm for detecting community structure in networks, *Physical review E*, Vol. 69, No. 6, p. 066133 (2004)

[Tanimoto 57] Tanimoto, T.: An Elementary Mathematical theory of Classification and Prediction, *Internal IBM Technical Report* (1957)

[Wang 11] Wang, X., Wei, F., Liu, X., Zhou, M., and Zhang, M.: Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach, in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pp. 1031–1040, New York, NY, USA (2011), ACM

[丸井 14] 丸井 淳己, 則 のぞみ, 榊 剛史, 森 純一郎：分散表現を用いたコミュニティにおける単語使用傾向の分析, 人工知能学会全国大会 (JSAI 2014) (2014)