4N2-OS-01b-2

二部ネットワークからのコミュニティ検出に基づく新しい協調フィルタリング方法
A New Collaborative Filtering Method Based on Community Detection in Bipartite Networks

邱　シュウレ[*1]
Xule Qiu

岡本　洋[*1]
Hiroshi Okamoto

[*1] 富士ゼロックス(株)研究技術開発本部
Research & Technology Group, Fuji Xerox Co., Ltd.

近年，インターネットの爆発的な発展に伴い、ユーザに有用なものを抽出するための情報推薦技術が広く注目を集めている。購買履歴や映画鑑賞履歴などの行動履歴データは二部ネットワークで表現できる。我々は以前にランダムウォークの確率モデルに基づくコミュニティ抽出方法 MDMC を提案した。本研究は、MDMC を用いて、ネットワーク分析の視点からの新しい協調フィルタリング方法を提案した。さらに、映画評価履歴データ「FilmTrust」を用いた比較実験により、提案方法は従来の多くの推薦方法よりも高精度であることを示した。

1. はじめに

1.1 従来の推薦技術

近年インターネットの爆発的な発展により、情報過負荷(Information Overload)の問題が発生している。そこで、膨大な情報からユーザに有用なものを抽出するための情報推薦技術が広く注目を集めている。Amazon、楽天等のネット通販および Netflix、Hulu 等の動画共有サイトには、推薦技術が既に欠かせないものになっている。

推薦技術には大きく分けて二つある。一つはアイテム(商品、動画等)の内容によって推薦する方式(「内容ベース」)で、もう一つは過去の履歴の分析に基づいて推薦する方式(「協調フィルタリング」)である。内容ベースの方法と比較して、協調フィルタリングは、推薦内容の多様性が高く、ドメイン知識を必要としないため、広く用いられている。

協調フィルタリングは、更に二種類のアプローチに分かれる。近傍ベース(neighborhood based)の方法および潜在意味モデル(latent factor model)の方法である。

近傍ベースの方法は、嗜好の類似したユーザは互いに高く評価した映画を見る、あるいは一人のユーザが類似した映画を見る、との仮定に基づいて推薦を行う。前者「ユーザベース」、後者を「アイテムベース」[1]と呼ぶ。ユーザに新しいアイテムを推薦するには、前者の場合にはユーザ同士の類似度計算、後者の場合にはアイテム同士の類似度計算を介して推薦を行う。

一般に履歴データは非常にスパースであるため、近傍ベースの方法では同類推移問題および少カバー率問題が引き起こされる。同類推移問題とは、実際に類似したユーザ同士であっても、同じアイテムを評価したことない場合には類似していると判別されないことを指す。少カバー率問題とは、ユーザの評価の少ないアイテムについては、実際に類似したアイテムを提示することができなくなることを指す。

一方、潜在意味モデルの方法では、確率モデルなどを介して、過去の行動履歴(購買、映画鑑賞等)を解釈したユーザおよびアイテムの潜在特徴を推定する [2,3,4]。例えば、ユーザ A はアクション映画および監督 B の作品を好む場合、抽出した A の潜在特徴にはそれらの嗜好が反映される。また A の鑑賞した映画に関しても、「タイプがアクション」および/あるいは「監督が B」の属性を持つ潜在特徴が生成される。潜在特徴を用いればユーザ・アイテム間の相性スコアを計算でき、ユーザ同士あるいはアイテム同士の類似度計算を介せずに推薦できる。

潜在意味モデルの方法により、「同類推移」および「少カバー率」の問題を解決できる。しかしながら、従来の多くの潜在意味モデルの方法は、ユーザ・アイテム間の行列を Matrix Factorization(MF) [2,3]によって分解して、潜在特徴を求める。その際には、疎行列の扱いおよび大規模データにおける MF の計算コストが課題となる。

1.2 ネットワーク分析と情報推薦

本研究は、購買履歴や映画鑑賞履歴のデータを二部ネットワークで表現し、ネットワーク分析の手段を用いた情報推薦を試みる。

二部ネットワークとは、2種類のノードおよび異種のノードの間を結ぶリンクからなるネットワークである。特に購買履歴や映画鑑賞履歴は、二部ネットワークで表現できるデータの典型例である。購買履歴は、「ユーザ」と「商品」を二種類のノードとして、購買関係のある「ユーザ」ー「商品」ペアを結ぶことにより二部ネットワークで表現できる。同様に、映画鑑賞履歴は「ユーザ」と「映画」のノードからなる二部ネットワークで表現できる(図1)。

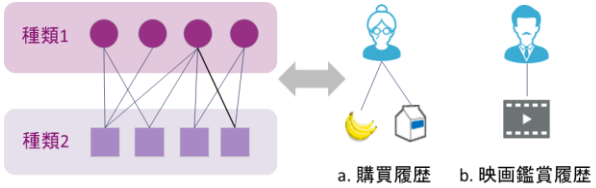

図 1：二部ネットワークとその例

ネットワーク科学では、何らかの共通性を持っているノード群を「コミュニティ抽出」を通じて発見することを試みる。コミュニティとは、ネットワークの中の密に繋がっているかたまりの部分を指す。共通性を持つノード同士は、他のノードとよりも強く関連しているため、ネットワーク上ではより密につながっていると考えられる。すなわち、コミュニティ抽出を通じて、ネットワークに潜在する共通性の情報を明らかにすることができる　。

二部ネットワークであっても、一般のネットワークと同様に、コミュニティ構造を持つ。特に、履歴データを表現する二部ネットワークでは、共通特性を持つアイテムとその特性を好むユーザに

連絡先：邱　シュウレ，富士ゼロックス(株)研究技術開発本部，
〒220-8668 神奈川県横浜市みなとみらい 6 丁目 1 番.
E-mail: qiu-xule@fujixerox.co.jp

より潜在コミュニティ構造が形成されると考える。したがって、履歴二部ネットワークからのコミュニティ抽出により、ユーザ・アイテムの潜在嗜好・特徴を抽出することができると考える。

コミュニティ抽出[7]を情報推薦に用いる試みは、まだほとんどない。それについてはいくつの理由が考えられる。二部ネットワークから効果的にコミュニティを抽出できる方法は[8,9,10,12]まだほとんど提案されていない。また、広く知られているコミュニティ抽出方法には、ノードを一個のコミュニティに割り当てるものが多い。しかしながら、適切な推薦効果を得るためには、対象の多様性(複数の特徴を用いること)を考慮することが重要である。

1.3 本研究の目的

本研究は、ネットワークにおけるコミュニティ抽出を利用した新しい潜在意味モデルの方法を提案する。

我々以前に、ランダムウォーク分解の確率モデルに基づくコミュニティ抽出方法(MDMC: Modular Decomposition of Markov Chain と呼ぶ)を提案した[11]。この方法は、二部ネットワークからも効果的にコミュニティを抽出できる[12]。一方、この方法はノードをただ一つのコミュニティに割り当てるのでなく、各コミュニティに属する割合を求める。さらに、階層的に(異なる粒度で)コミュニティ構造を抽出できる[13]。

本研究は、MDMC を行動履歴の二部ネットワークに適用し、抽出されたコミュニティ階層構造情報に基づいて、アイテムを推薦する方法を構築する。そのために、ユーザ・アイテムペアの相性を計算する手法「MDMC 推薦度」を提案する。この手法により、アイテムのポピュラー性およびターゲットユーザの個人趣味の両方を考慮した推薦を行うことができる。

さらに、映画評価履歴データ「FilmTrust」[14]を用いて、提案方法と従来の推薦方法との間の比較実験を行った。アイテム推薦の効果について、複数の指標を用いた評価の結果は、提案方法が従来の多くの推薦方法よりも高精度であることを示唆する。

2. 方法

2.1 MDMC によるコミュニティ抽出

まず、以前に提案した MDMC コミュニティ抽出方法について極簡単に振り返る。(詳細は文献[11]を参照)。

MDMC 法は、式(1)で定められた混合確率分布モデルを解くことによりコミュニティを抽出する。このモデルでは、ネットワーク全体におけるランダムウォーク(左辺)をそれぞれのコミュニティにおけるランダムウォーク(右辺)と分解することを試みる。

$$p^{stead}(n)=\sum_{k=1}^{K}\pi_k p(n|k) \qquad (1)$$

$p^{stead}(n)$ は、ランダムウォークの定常状態における確率分布である。$p(n|k)$ はコミュニティ k におけるノードの確率分布である。π_k はコミュニティ k の事前確率であり、$\sum_{k=1}^{K}\pi_k=1$ を満たす。

各変数 $\{\pi_k\}$ 及び $\{p(n|k)\}$ は、機械学習の標準手法 EM (Expectation Maximization)アルゴリズムに従って求める。

ノード n の各コミュニティ k への帰属度 $g(k|n)$ はベイズ定理により計算し、$\sum_k g(k|n)=1$ を満たす。

$$\gamma(k|n)=\frac{\pi_k\cdot p_t(n|k)}{\sum_{k=1}^{K}\pi_k\cdot p_t(n|k)} \qquad (2)$$

また、このモデルには唯一のパラメタ α が存在する。α はコミュニティの分解粒度を制御する。α の値が大きい(/小さい)場合、ネットワークが粗く(/細かく)分解される。前の研究では、α の準静的な増加により起きる相転移現象を利用して、コミュニティ階層構造を自動的に抽出する方法を提案した[13]。

2.2 MDMC 法を用いたユーザ・アイテム間の相性計算

本研究では、MDMC で得られたコミュニティ情報を用いて、新しいユーザ・アイテム間の相性を計算する手法を提案する。

まず以下に用いる符号を定義する。

u :ユーザ;

i :アイテム;

l :階層/分解粒度(その層におけるコミュニティ数で標記する);

k :コミュニティ;

$\gamma(k|u)$:ユーザ u がコミュニティ k に属する割合;

$p(i|k)$:コミュニティ k におけるアイテム i の重要度。

$\gamma(k|u)$ と $p(i|k)$ はコミュニティ分解の結果により得られる。

抽出されたコミュニティは個々の潜在属性を表すと考えて、階層 l におけるユーザ・アイテム間の相性スコア $C^l(u,i)$ の計算手法を式(3)で定める。

$$C^l(u,i)=\sum_k \gamma(k|u)\cdot p(i|k) \qquad (3)$$

$g(k|u)$ はユーザ u が属性 k を持つカテゴリに興味持つ度合いを表す。$p(i|k)$ はそのカテゴリにおけるアイテム i のランキング(ポピュラー性)を表す。$g(k|u)\times p(i|k)$ は、ユーザ u の興味の高い領域における人気の高いアイテムを高く評価する。$C^l(u,i)$ は、$g(k|u)\times p(i|k)$ をすべてのコミュニティ k について合計することにより、ユーザの多様性(複数のカテゴリに興味持つ)とアイテムの多様性(複数の属性を持つ)と考慮した総合スコアを表す。

2.3 推薦に用いるコミュニティ分解粒度/階層

行動履歴二部ネットワークを含む多くの現実世界のネットワークは、階層的なコミュニティ構造を持つと考える。例えば、ユーザ・商品を{「食べ物」、「服」、「電子用品」、…}という粒度で分けてもよく、あるいは「食べ物」をさらに{「果物」、「肉類」、「お菓子」、…}という粒度で分けてもよい。どの粒度の分け方が正解であるかには意味なく、用途に応じて粒度を使い分けるべきであろう。

階層/分解粒度の果たす役割を、「FilmTrust」データを用いた実験により確認した。上記の相性計算式 $C^l(u,i)$ にどの層のコミュニティ結果を用いるかにより、異なる推薦効果を得られることを、以下の通りに見出した(図2)。

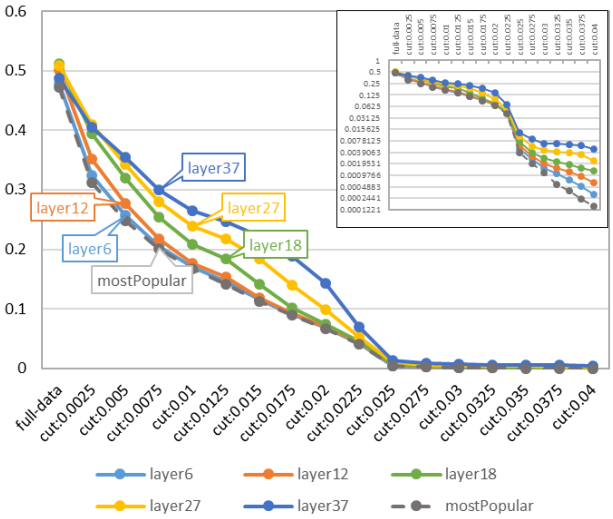

図 2：映画の人気順位でトップ t の割合のものに関わる評価履歴をテストデータから除いて、残り下位 1−t のアイテムについ

ての推薦精度を5−CV 実験で求める。t を[0.0~0.0325]の範囲で少しずつ増やし、それぞれの層を用いた $C^l(u,i)$（l = 6,12,18,27,37)によるトップ 60 の推薦結果、および、ポピュラーアイテムを全ユーザに推薦する「MostPopular」法によるトップ 60 の推薦結果を、MAP (Mean Average Precision)指標を用いて比較する。(ただし、各層を用いた推薦結果は、その層におけるコミュニティ数で標記する。例えば、層 $l=6$ における結果を layer6 で表す。)右上図は縦軸を対数目盛で表したものである。

図2により、ポピュラーでないアイテムの推薦について、より深い層(細かいコミュニティ分解)を用いた場合には推薦予測の精度がより高くなる;一方、より浅い層(粗いコミュニティ分解)を用いた場合には、より MostPopular 法のパーフォマンスに近づいていくことがわかった。すなわち、$C^l(u,i)$ に用いるコミュニティ分解の粒度を変えることにより、よりポピュラーなアイテムを推薦するか、よりユーザの個人趣味を反映したアイテムを推薦するかを調整できることが示唆された。

2.4 階層を考慮した MDMC 推薦度指標

多くの研究はパーソナライズされた推薦を目指しているが、実際にはユーザがポピュラーなアイテムを欲している場合も多い。ゆえに、推薦を行う際には、ターゲットユーザの個人趣味およびアイテムのポピュラー性の両方を考慮すべきであろう。

そこで、我々はコミュニティ階層構造を利用し、式3の相性計算式 $C^l(u,i)$ を階層で平均した指標を提案する。これを「MDMC 推薦度」と呼ぶ。

$$C(u,i)=\frac{1}{L}\cdot\sum_l C^l(u,i) \tag{4}$$

ただし、L は階層の数である。

MDMC 推薦度は全ての階層のコミュニティ結果を用いることにより、ポピュラー性とユーザの個人趣味との相性について総合的にスコア高いアイテムを推薦する。

「FilmTrust」での比較実験を通じて、一層のみの情報を利用した $C^l(u,i)$ よりも、全階層の豊かな情報を利用した MDMC 推薦度の推薦結果が優位であることを確認した。

3. 結果

前節で提案したユーザ・アイテム相性計算手法 MDMC 推薦度を用いて、映画評価履歴データ「FilmTrust」(「映画」数 1,400、「映画」数 1,987、評価レコード数 35,497)でのトップ N アイテム推薦の実験を行った。

ユーザ毎に履歴データをランダムに5分割し、5-CV で評価実験を行った。

トレーニング用データから「ユーザ」-「映画」の二部ネットワーク(リンクをユーザが映画に付けた点数で重み付けした)を構築し、それから MDMC 法を用いてコミュニティの階層構造(l=2~35)を抽出する。提案 MDMC 推薦度を用いて、ユーザ・アイテムペアの相性スコアを計算する。目的ユーザにスコアの最も高いアイテムを N 個推薦する。

提案推薦方法がどのぐらいユーザの趣味を捉えているか、どのぐらいユーザの行動を予測できているかを、推薦アイテムとテストデータに含むアイテムとがどれくらい一致するかで評価する。

一人のユーザ u に対して、テストデータに登場したアイテム集合を R_u 、推薦されたアイテム集合を R'_u と記述する。用いた評価指標は以下の通りである。

(a). CR(Conversion Rate):

全てのユーザのうち、少なくとも一個のアイテムが正しく推薦された人数の割合を示す。一人のユーザに対しての計算は以下である。

$$CR=\begin{cases}0 & \text{if } |R\cap R'|>0\\ 1 & \text{otherwise}\end{cases} \tag{5}$$

(b). MAP (Mean Average Precision)

AP を全てのユーザで平均したものである。AP は推薦結果のゴミの少なさとモレの少なさを評価する指標である。

$$AP_u=\frac{1}{|R_u|}\cdot\sum_{i=1}^{|R'_u|}rel(i)\cdot\frac{count(i)}{i} \tag{6}$$

$$MAP=\frac{1}{|U|}\cdot\sum_u AP_u \tag{7}$$

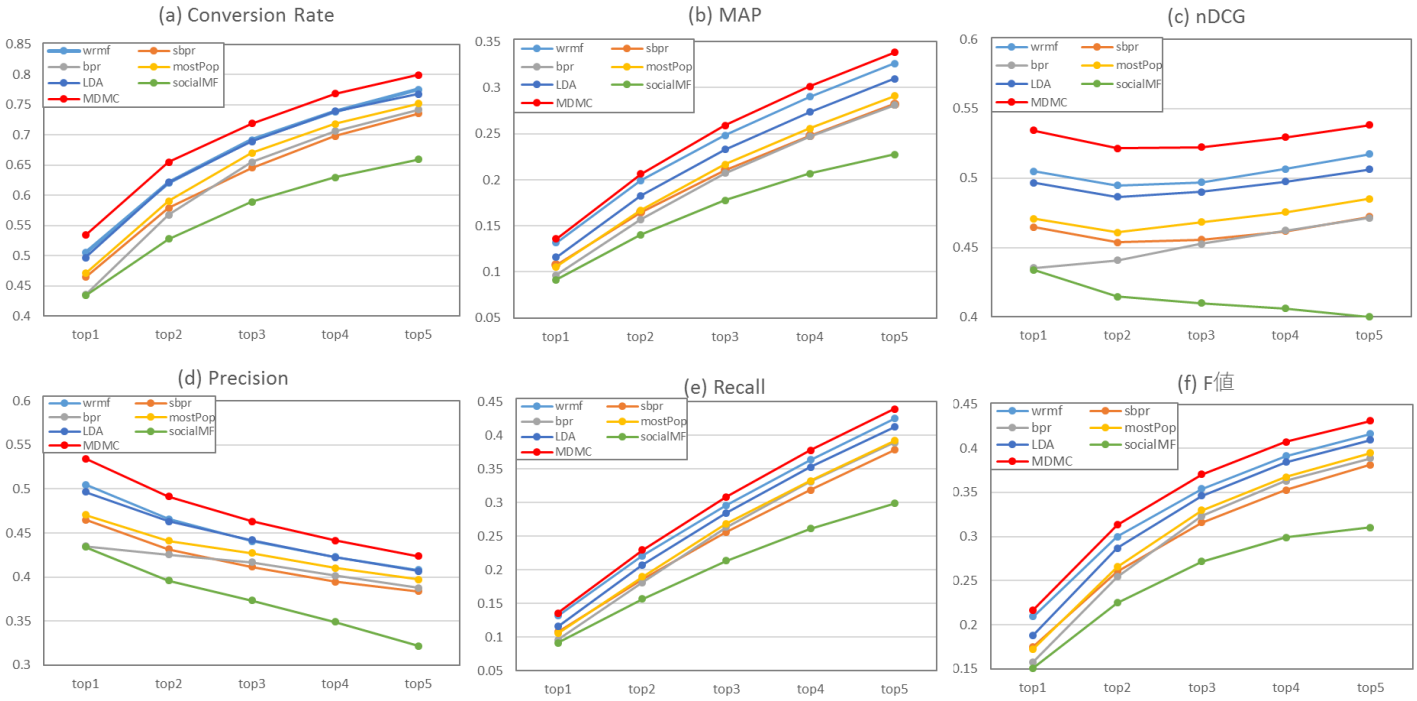

図 3： FilmTrust における Top5アイテム推薦の比較結果

ただし、iは推薦リストにおける順位(1から$|R'_u|$まで)、$rel(i)$はi位の推薦アイテムが正解か否か(0 or 1)、$count(i)$は上位iまでの推薦アイテムの正解数、$|R'_u|$、$|R_u|$はそれぞれR'_u、R_uに含むアイテムの数を示す。

(c). nDCG (normalized Discounted Cumulative Gain)
順位付けの正しさを示す指標である。

$$nDCG_k = \frac{DCG_k}{idealDCG_k} = \frac{1}{idealDCG_k} \cdot \sum_{i=1}^{k} \frac{2^{rel(i)} - 1}{\log_2(i+1)} \quad (8)$$

ただし、kは順位の最大数(トップN個のアイテムを推薦する場合は$k = N$)、$rel(i)$はi位の推薦アイテムが正解か否か(0 or 1)、$idealDCG_k$は理想の最大DCG_k値を示す。

(d). Precision 精度

$$precision = \frac{|R \cap R'|}{|R'|} \quad (9)$$

(e). Recall 再現率

$$recall = \frac{|R \cap R'|}{|R|} \quad (10)$$

(f). F値

$$F値 = \frac{2 * precision * recall}{precision + recall} \quad (11)$$

提案方法による結果を、既存の推薦方法による結果と、図3において比較する。ただし、競争方法として、mostPop、bpr (Bayesian Personalized Ranking)[2] 、wrmf (weighted regularized matrix factorization)[3] 、LDA[4]、sbpr (social bayesian personalized ranking)[5]、socialMF (social matrix factorization)[6]を用いた。

どの指標でも、提案方法は競争方法よりも高い推薦精度を示した(図3)。

4. 議論

本研究は、履歴データを二部ネットワークで表現し、ネットワーク分析の新しい視点からアイテム推薦を試みた。我々が以前に提案したランダムウォークの確率モデルに基づくコミュニティ抽出方法 MDMC を用いて、新しい協調フィルタリング方法を提案した。

コミュニティ分解により、二部ネットワークで表現された履歴データに潜在する情報、すなわち、共通特性を持つアイテムとその特性を好むユーザを発見できる。本稿は、そのコミュニティの情報を用いて、新しいユーザ・アイテムの間の相性計算手法―「MDMC 推薦度」―を提案した。

また、推薦効果におけるコミュニティ階層が果たす役割を調べ、浅い層からはポピュラーなアイテムが推薦され、深い層からはユーザの個人趣味を反映したアイテムが推薦される傾向がわかった。「MDMC 推薦度」は、階層的な(マルチスケール)コミュニティ構造を用いて、アイテムのポピュラー性とユーザの個人趣味の両方を考慮した総合スコアを与える。

MDMC は潜在意味抽出モデルでもあり、MDMC を用いた提案方法は以下の特徴を有する。コミュニティ抽出では近傍から遠くまで(コミュニティ範囲)の情報を見ているため、近傍ベース方法で発生する「同類推移」と「少カバー率」の問題を回避できる。さらに、コミュニティという局在化された構造のみを見ているため、同様に潜在意味モデルである行列分解方法の大規模データ計算コストの問題を緩和できる。

現在、ユーザ同士のソーシャル関係を考慮したソーシャルネットワークベースの方法も研究されている[5,6]。ユーザ同士の嗜好や行動がソーシャルネットワークを介して相互に影響し合っていることが考えられている。ソーシャルネットワークの情報を履歴二部ネットワークと組み合わせてより適切な推薦を行うことは、今後の研究課題であると考える。

参考文献

1. Deshpande M, Karypis G. Item-based top-n recommendation algorithms. ACM Transactions on Information Systems. Springer-Verlag, 22/1, 2004.
2. Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback. Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. AUAI Press, 2009.
3. Hu Y, Koren Y, Volinsky C. Collaborative filtering for implicit feedback datasets. In IEEE International Conference on Data Mining (ICDM 2008), 2008.
4. Griffiths T. Gibbs sampling in the generative model of latent dirichlet allocation. 2002.
5. Zhao T, McAuley J, King I. Leveraging social connections to improve personalized ranking for collaborative filtering. Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. ACM, 2014.
6. Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks. Proceedings of the fourth ACM conference on Recommender systems. ACM, 2010.
7. Fortunato S, Hric D. Community detection in networks: A user guide. Physics Reports, 2016.
8. Larremore D B, Clauset A, Jacobs A Z. Efficiently inferring community structure in bipartite networks. Physical Review E, 2014.
9. Guimerà R, Sales-Pardo M, Amaral L A N. Module identification in bipartite and directed networks. Physical Review E, 2007.
10. Barber M J. Modularity and community detection in bipartite networks. Physical Review E, 2007.
11. 岡本 洋. マルコフ連鎖のモジュール分解:ネットワークからの重なりと階層構造を持つコミュニティの検出.JWEIN2014.
12. 邱シュウレ, 稲木誓哉, 貫井駿, 等. 二部ネットワークからのコミュニティ検出およびその実課題への適用. JWEIN2016.
13. 邱シュウレ, 岡本 洋. ネットワークからのコミュニティ階層構造の効果的かつ安定な検出. JWEIN2015.
14. FilmTrust: http://www.librec.net/datasets/filmtrust.zip.