劣モジュラ関数最大化による大規模クラスタリング

Large Scale Clustering by Maximizing Submodular Functions

入月卓也[*1] 川添高志[*1] 斉藤和巳[*1] 武藤伸明[*1] 池田哲夫[*1]
Takuya Iriduki Takashi Kawazoe Kazumi Saito Nobuaki Mutoh Tetsuo Ikeda

[*1]静岡県立大学
University of Shizuoka

In this paper, we consider formalizing the large-scale object clustering problem as a problem of maximizing the corresponding submodular function. As its solution method, we focus on the greedy method with lazy evaluation. Then we evaluate the performance of this method in comparison to a widely-used iterative improvement method for the clustering problem. In our experiments using a network constructed from Japanese Wikipedia and a document data set of Japanese newspaper, we demonstrate that the greedy method with lazy evaluation produces promising results within a reasonable amount of processing time.

1. はじめに

劣モジュラ構造を持つ離散最適化問題は，数理的枠組みの基盤が強固であり [室田 07]，潜在的に幅広い応用が展開可能なため，人工知能分野でも今後の重要な研究対象として注目できる．適用問題の例としては，多変量データから有用な変数集合を選択する問題，社会ネットワーク上の情報伝播で影響を最大にするノード集合を選択する問題などが知られている．また，ある種のコストや制約条件などを持つ複雑な問題に対しては，条件緩和や探索法の改良などにより，劣モジュラ最適化問題として定式化して解くための研究も最近活発に行われつつある．劣モジュラ最適化として定式化可能な新たな問題領域の探究も重要な研究課題と言える．

劣モジュラ最適化は，最小化と最大化の二つの問題に大別されるが，本研究では，後者の劣モジュラ最大化問題に焦点を当てる．この最大化問題は NP-完全クラスに属し，大規模問題では妥当な計算時間で厳密解を求めることが一般に困難となる．ただし，その望ましい性質として，いわゆる貪欲（greedy）法で効率良く求められる近似解により，ある程度妥当な精度で，最悪ケースの解品質を理論的に保証することができる．さらに，遅延評価（lazy evaluation）と呼ばれる手法の導入で貪欲法をさらに効率化することにより [Leskovec 07]，非常に大規模かつ複雑な問題でも妥当な計算時間で精度保証付き近似解を得ることが可能になる．

本論文では，まず，劣モジュラ最大化問題とその解法の遅延評価付き貪欲法について概説する．次に，オブジェクトのクラスタリング問題の一つである K-メディアン問題が劣モジュラ性を持つことを示す．一方，このようなクラスタリング問題の標準的解法と考えられる反復改善法について述べる．最後に，日本語 Wikipedia から構築した人名ネットワーク，および現実の文書データを用いた実験により，標準的な反復改善法と比較することで，遅延評価付き貪欲法の有効性を検証する．なお，後者の文書データの各オブジェクトは，ユークリッド空間の特徴ベクトルとして記述されるのに対し，前者の人名ネットワークでの類似度は，標準的なユークリッド距離では定義されない問題であり，両者は異なる特性を持つ問題と言える．

連絡先: 斉藤和巳，静岡県立大学経営情報学部，〒 422-8526 静岡市駿河区谷田 52 番 1 号，054-264-5436，k-saito@u-shizuoka-ken.ac.jp

2. 劣モジュラ最大化問題

劣モジュラ最大化について集合被覆問題を例に説明する．集合 $X = \{x_\alpha : 1 \leq \alpha \leq |X|\}$ と，集合 $Y = \{y_\beta : 1 \leq \beta \leq |Y|\}$，及び正の整数 $K < |Y|$ が与えられたとする．ただし，$|X|$ や $|Y|$ は集合 X や Y の要素数をそれぞれ表し，Y の各要素 y_β は X の部分集合 $y_\beta \subset X$ とする．いま，Y の任意の部分集合 $B \subset Y$ に対し，B の要素 y_β の和集合により被覆される X の部分集合 $A \subset X$ を $A = \{x_\alpha : x_\alpha \in y_\beta, y_\beta \in B\}$ で定義し，集合 B で定まる集合 A の要素数を $f(B) = |A|$ とし，劣モジュラ関数と呼ぶ．条件 $|B| = K$ で $f(B)$ を最大化する B を求めるのが集合被覆問題である．

集合被覆問題に代表される劣モジュラ最大化問題の標準解法，貪欲法の詳細は以下である．

1. 反復制御変数を $k = 0$ とし，結果を格納する集合を空 $B_0 = \emptyset$ に初期化;
2. 集合 B_k を固定し，Y から最良要素 $z = \arg\max_{y \in Y}\{f(B_k \cup \{y\}) - f(B_k)\}$ を計算;
3. 最良要素 z を追加 $B_{k+1} = B_k \cup \{z\}$ し，$k = k + 1$ に設定;
4. $k = K$ ならば終了，さらもなければステップ 2. へ戻る.

明らかに，貪欲法で厳密解は得られないが，ある程度妥当な精度で最悪ケースを理論的に保証することができる．詳細には，厳密解を $B*$ とすると，貪欲法で求まる近似解 B の精度は，関係式 $f(B) \geq (1 - 1/e)f(B*)$ で抑えられることが証明されている．ここで，e は自然対数の底であり，貪欲解により，最悪でも厳密解の 63% 程度以上の性能が保証されることになる．

一方，$h < k$ の非負整数 h に対し，関係式

$$f(B_h \cup \{y\}) - f(B_h) \geq f(B_k \cup \{y\}) - f(B_k)$$

が成立し、劣モジュラ不等式と呼ばれ，集合 y を一つ加えることで新たに被覆可能となる X の要素数は，B_h の方が多いことを表している（$B_h \subset B_k$ の関係に注意）．この関係式を利用して，探索効率の改善を行うのが遅延評価である．具体的には，貪欲法ステップ 2. の第 k 反復目の最良要素選定において，各集合 y での改善上限 $f(B_h \cup \{y\}) - f(B_h)$ を用いて，

Y の各要素を降順にソートしてリストを作成し，先頭から順に探索する過程で，ある要素 $z \in Y$ での実際の改善値と比較して，探索リスト上で未探索の先頭要素での上限が小さくなれば，その時点で探索が終了し最良要素 z が求まる．

3. K-メディアン問題と反復改善法

K-メディアン問題について説明するとともに，この問題が劣モジュラ最大化問題となることを示す．オブジェクト集合 $X = \{x_i : 1 \leq i \leq |X|\}$ が与えられ，任意のオブジェクトのペア x_i と x_j に対して，適切な類似度関数 $\rho(x_i, x_j)$ が定義されているとする．ここで自然な設定として，類似度の値域は実数区間 $[0, 1]$ とし，自分自身との類似度は 1 とする．つまり，$\forall x_i, \forall x_j,\ 0 \leq \rho(x_i, x_j) \leq 1$ かつ $\forall x_i,\ \rho(x_i, x_i) = 1$ である．K-メディアン問題とは，与えられた非負整数 K に対して，下記の目的関数 f を最大にする K 個のオブジェクトの集合 $B_K \subset X, |B_K| = K$ を求める問題である．

$$f(B_K) \quad = \quad \sum_{x \in X} \max_{b \in B_K} \{\rho(x, b)\}. \tag{1}$$

ここで，$B_h \subset B_k$ とすれば，任意のオブジェクト $x, x_0 \in X$ に対して，以下の関係が成立する．

$$\begin{aligned} &\max_{b \in B_h \cup \{x\}} \{\rho(x_0, b)\} - \max_{b \in B_h} \{\rho(x_0, b)\} \\ &\geq \quad \max_{b \in B_k \cup \{x\}} \{\rho(x_0, b)\} - \max_{b \in B_k} \{\rho(x_0, b)\}. \end{aligned} \tag{2}$$

よって，目的関数 f が劣モジュラ関数となることは容易に確認できる．ゆえに，前節で述べた貪欲法により，精度保証付き近似解を求めることができる．

いま、$B_K = \{b_1, \cdots, b_K\} \subset X$ に対して，次のようなクラスタ分割を考えることができる．

$$C(b_k) \quad = \quad \{x \in X : b_k = \arg \max_{b \in B_K} \{\rho(x, b)\}\}. \tag{3}$$

ただし，オブジェクトに対して，類似度を最大にする B_K の要素が複数存在する場合，適当にタイブレークを行うものとする．つまり，オブジェクト集合 X と要素間の類似度 ρ に対して，式 (1) で定義した劣モジュラ最大化問題を解くことにより，式 (3) で規定されるクラスタ分割を得ることができる．

一方，別の解法として，K-平均法に類似した以下のような反復改善法を考えることができる．

1. $t = 0$ とし，X からランダムに K 個のオブジェクトを選定し $B_K^{(0)} = \{b_1^{(0)}, \cdots, b_K^{(0)}\}$ を初期化;
2. 式 (3) でクラスタ $C(b_k^{(t)})$ を計算（$1 \leq k \leq K$）；
3. 次式で $B_K^{(t+1)}$ の要素 $b_k^{(t+1)}$ を計算（$1 \leq k \leq K$）；

$$b_k^{(t+1)} = \arg \max_{x \in C(b_k^{(t)})} \{ \sum_{y \in C(b_k^{(t)})} \rho(x, y)\} \tag{4}$$

4. $B_K^{(t)} = B_K^{(t+1)}$ ならば終了，さらもなければ $t = t + 1$ としステップ 2. へ戻る．

このアルゴリズムでは，ステップ 2. でオブジェクトの分割を行い，ステップ 3. では，各 k 毎に分割集合 $C(b_k^{(t)})$ の範囲で最良オブジェクトを探索する．特に，X のオブジェクトがユークリッド空間のベクトルとして与えられ，類似度関数 ρ が通常のユークリッド距離の自乗と実質等価で，さらに，$B_K \subset X$ の制約をなくして任意のユークリッド空間のベクトルまで許容すれば，標準的な K-平均法に帰着される．すなわち，ステップ 3. では，クラスタ $C(b_k^{(t)})$ に属すベクトルの重心を計算することになる．ここで，$B_K \subset X$ の制約を採用すれば，いわゆる K-メディアン問題になることが知られている．

一般に，K-メディアンの枠組みは，外れ値などにロバストなことが知られており，ユークリッド空間以外のオブジェクト集合にも適用可能な汎用性を有する．以下の実験では，2 章で述べた貪欲（greedy）法と反復改善（iteration）法で得られる解の精度比較を行う．ここでは，解集合 B_K に含まれるオブジェクトをピボットと呼び，K を特にピボット数と呼ぶ．

4. 評価実験

4.1 人名ネットワークへの適用

評価データとして，Wikipedia の人名ネットワークデータを用いた [Kimura 07]．具体的には，人名一覧に登場する人物において，ウィキペディア内の記事中に 6 回以上共起した 2 人の人物をリンクすることから得られる無向グラフの最大連結成分を抽出した．ここに，ノード数は 9,481 であり，各ノードの次数の総和は 245,044 であった．一方，ノード x_i と x_j 間のネットワーク上での最短パス長を $d(x_i, x_j)$ とするとき，類似度関数を $\rho(x_i, x_j) = 1/(1 + d(x_i, x_j))$ で定義した．なお，このような最短パス長の逆数の利用は，非連結グラフへも適用可能となり，より汎用性の高い手段であることが指摘されている [Newman 03]．

図 1 では，ピボット数を $K = 1$ から 10 まで変化させたときの，人名ネットワークデータにおける貪欲（greedy）法と反復改善（iteration）法で得られる解の精度を比較する．ここで解の精度とは (1) 式で定義した目的関数のことである．反復改善法では，初期値に依存して解の精度が変わるため，初期値を変え 5 回の実験を行った．図より，$K = 1$ から 3 程度では，2 つの手法は同等程度の結果も得られるものの，K が大きくなるにしたがい，貪欲法が優位になっていることが分かる．すなわち，反復改善法での 5 回全ての解は，貪欲法の解に劣っている．この結果は，クラスタリング手法としての貪欲法の有望性を示唆していると考える．

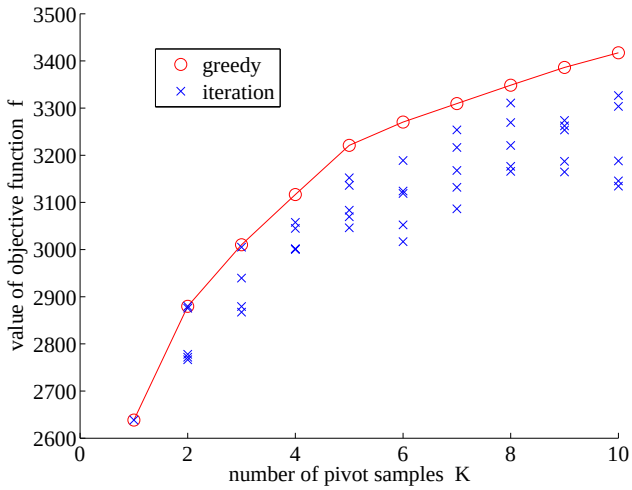

図 1: 人名ネットワークでの解品質の比較

図 2 では，上記実験設定での遅延評価付き貪欲（greedy with lazy evaluation）法，反復改善（iteration）法，および，遅延評価なし貪欲（greedy without lazy evaluation）法の計算効率を比較する．遅延評価なし貪欲法の処理時間は，K が大きくなるにしたがい線形に増加し，反復改善法では，試行毎に差があるものの，$K \geq 2$ で K によらず殆ど一定の処理時間だったと言える．一方，遅延評価付き貪欲法では，$K \geq 2$ での処理時間は微増で，遅延評価が有効に機能していることが分かる．

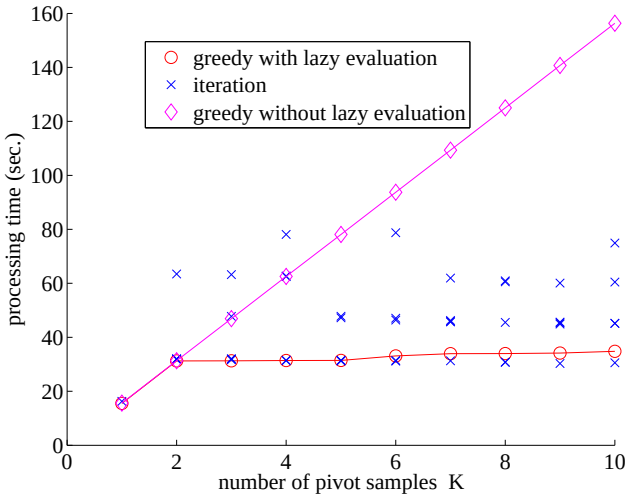

図 2: 人名ネットワークでの計算効率の比較

表 1 には，貪欲法と 5 回の反復改善法の解集合 $B_K(K = 10)$ のピボットとして実際に選択された人名を示す．ここで，貪欲法では選定された順に上から人名を並べているが．反復改善の順番は任意である．また，手法名の下の括弧内の数値は，結果の解精度を示している．表より，どちらの手法でも多様な分野の代表的な人名が選択されていることが分かる．

4.2 文書データへの適用

もう一つの評価データとして，1994 年 1 月から 12 月までの毎日新聞国際面記事を用いた [斉藤 05]．総記事数は 5,464 であり，文書データの前処理には，Chasen による形態素解析を施し，助詞などを削除したところ，出現した異なる単語総数は 23,100 となった．各文書は，出現した単語の頻度ベクトルを tf-idf 変換した特徴ベクトルを用いた．すなわち，23,100-次元のベクトルで各文書を表現した．なお，文書の平均単語数は 121.7 であった．一方，文書間の類似度としては，特徴ベクトルのコサイン類似度を用いた．

図 3 に，ピボット数を $K = 1$ から 10 まで変化させたときの，毎日新聞データにおける貪欲法と反復改善法で得られる解の精度比較を示す．ここでも反復改善法では，初期値を変え 5 回の実験を行った．図より，人名ネットワークを用いた実験結果と同様に，全体として貪欲法が優っていることが分かる．

図 4 には，上記実験設定での遅延評価付き貪欲法，反復改善法，および，遅延評価なし貪欲法の計算効率比較を示す．図より，ここでも人名ネットワークを用いた実験結果と同様に，遅延評価なし貪欲法の処理時間は，K が大きくなるにしたがい線形に増加し，反復改善法では，試行毎に差があるものの，$K \geq 2$ で K によらず殆ど一定の処理時間だったと言える．一方，遅延評価付き貪欲法では，$K \geq 2$ での処理時間は微増で，遅延評価が有効に機能していることが分かる．ここで，反復改善法とし比較して，遅延評価付き貪欲法の処理時間が大きく

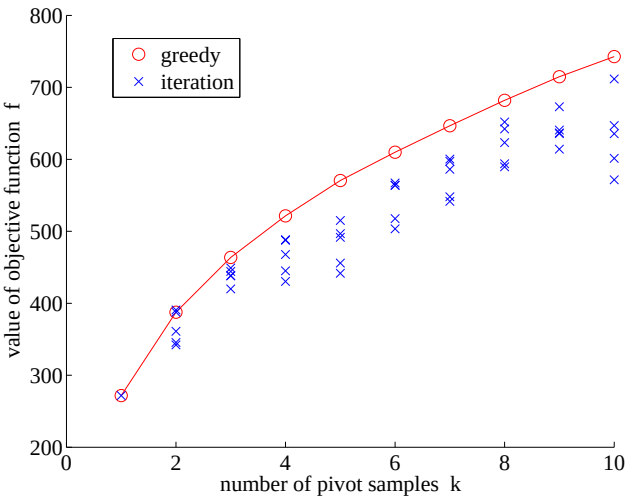

図 3: 毎日新聞データでの解品質の比較

なっている．ただし，各文書はユークリッド空間の特徴ベクトルであるため，式 (3) で定義したクラスタ分割内での重心を求めることが可能であり，コサイン類似度の場合には，重心との類似度が最も高いベクトルを最良ピボットとして選択できる．この性質に基づく反復改善法プログラムを利用したため，人名ネットワークでの実験結果と比較して，反復改善法の計算効率が向上したと考えられる．すなわち，遅延評価付き貪欲法には，比較的不利なタイプの問題と言えるが，図 4 に示すように，K が増大しても 2 から 3 倍程度の処理時間で結果が得られることは，遅延評価付き貪欲法の大規模データへの有望な適用性を示唆していると考える．

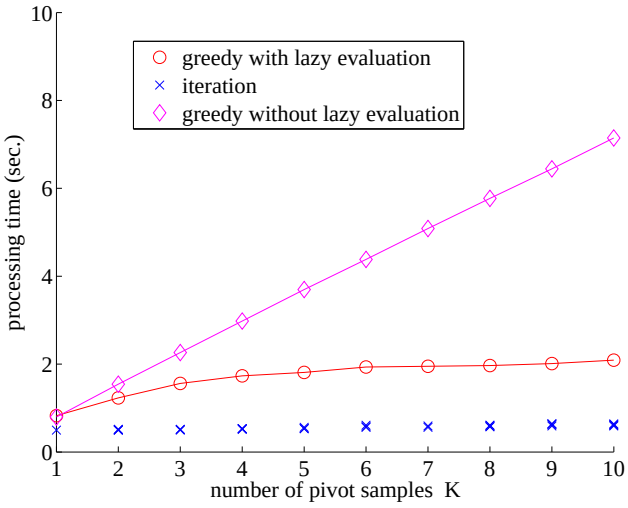

図 4: 毎日新聞データでの計算効率の比較

表 2 には，表 1 と同様に，貪欲法と反復改善法の解集合 $B_K(K = 10)$ のピボットとして実際に選択された記事名を示す．ここで，反復改善法の結果は随意に選んだものである．表より，貪欲法でも反復改善法でも多様なトピックの記事が選択されていることが分かる．詳細な検討は今後の課題であるものの，貪欲法のクラスタリング手法としての有望性を示唆する結果と考える．

表 1: 貪欲法と反復改善法で抽出したピボット（人名）の比較（$K = 10$）

貪欲法 (3,417.38)	反復改善法 1 (3,326.64)	反復改善法 2 (3,187.88)	反復改善法 3 (3,145.68)	反復改善法 4 (3,303.54)	反復改善法 5 (3,134.52)
徳川家康	長嶋茂雄	長嶋茂雄	ビートたけし	徳川家康	昭和天皇
ビートたけし	クビライ	ビートたけし	徳川家康	ベートーヴェン	ビートたけし
ナポレオン	徳川家康	徳川家康	小泉純一郎	長嶋茂雄	谷亮子
長嶋茂雄	明石家さんま	黒澤明	ベートーヴェン	ルイ 14 世	サンプラス
置鮎龍太郎	ナポレオン	小野伸二	神山繁	源頼朝	エディ・ゲレロ
手塚治虫	アントニオ猪木	ベートーヴェン	西郷隆盛	香取慎吾	松浦亜弥
小泉純一郎	ピカソ	足利尊氏	クビライ	置鮎龍太郎	置鮎龍太郎
和田アキ子	置鮎龍太郎	梨田昌孝	桑田佳祐	ナポレオン	源頼朝
ベートーヴェン	始皇帝	後醍醐天皇	渡辺いっけい	アレキサンダー	石ノ森章太郎
ヒトラー	小泉純一郎	星野仙一	日蓮	松田聖子	清和天皇

表 2: 貪欲法と反復改善法で抽出したピボット（記事名）の比較（$K = 10$）

貪欲法 (742.72)	反復改善法 1 (635.57)
［検証・核の疑惑］識者座談会／上　「査察」問題の核心を探る	「見切り発車」に不安も――南北首脳会談・開催合意
［探眼複眼］出口見えぬボスニア紛争　冷戦終結を引き金に	［探眼複眼］あすから欧州議会選挙　関心は失業、不況問題に
イスラエルのラビン首相がモスクワ入り	セルビア人勢力、重火器すべて返還　再空爆は回避の見通し
［探眼複眼］中台統一の構図に異変――台湾、「独立派」の勢力台頭	［フロント・ライン］一党支配６５年、緊迫のメキシコ大統領選
［探眼複眼］あすから欧州議会選挙　関心は失業、不況問題に	韓国大統領、来月２６日から訪中
［探眼複眼］ハイチ駐留、長期化の恐れも	［探眼複眼］東ティモール問題国際会議の２国間摩擦
［探眼複眼］東ティモール問題国際会議の２国間摩擦	［探眼複眼］北朝鮮とＩＡＥＡ、袋小路の核査察協議
［探眼複眼］北朝鮮主席「空席」６カ月　金正日氏の就任遅れ	イラク軍集結…国民の不満解消狙う？　示威なら逆効果
［探眼複眼］仏軍介入で混迷――ルワンダ内戦	連立暫定内閣を発表――ルワンダ
シアヌーク国王、円卓会議を再提案　政府軍とポル・ポト派の停戦求め	ルツコイ氏ら保守派結集、「ロシアのための合意」を組織

4.3 議論と考察

本実験では，遅延評価つき貪欲法, 遅延評価なしの貪欲法をクラスタリング問題で標準的に利用される反復改善法と比較することにより，貪欲法の有効性を評価した．貪欲法の利点として，図 1 と 3 の結果からわかるように，反復改善法よりも安定して精度の高い解が求まる点があげられる．すなわち，貪欲法のクラスタリング手法としての有望性を示唆している．一方，遅延評価なしの貪欲法の問題点は，図 2 と 4 の結果からわかるように，反復改善法よりも K に比例して処理時間が必要となる点である．しかしながら，この問題点は貪欲法に遅延評価を導入することで，図 2 と 4 の結果からも分かるように，K に比例した処理時間の増大を大幅に抑制できることが分かる．すなわち，貪欲法の大規模データへの望ましい適用性を示唆している．ただし，図 4 の新聞の記事データにおける場合のように，ユークリッド空間の特徴ベクトルとして表現されるオブジェクトでは，遅延評価付き貪欲法でも反復改善法より常に高速な処理手法であるとは一概に言えない．しかしながら，反復改善法よりも安定して優れた精度の解が探索可能であり，遅延評価を導入することで反復改善法との探索時間の差を縮小できることを考慮すると，一般に反復改善法よりも優れた解法であることが期待できる．

5. おわりに

本論文では，劣モジュラ最大化問題とその解法の遅延評価付き貪欲法について概説するとともに，オブジェクトのクラスタリング問題の一つである K-メディアン問題が劣モジュラ性を持つことを示した．また，このようなクラスタリング問題の標準的解法と考えられる反復改善法について述べるとともに，日本語 Wikipedia から構築した人名ネットワーク，および現実の文書データを用いた実験により，標準的な反復改善法と比較することで，遅延評価付き貪欲法の有効性を検証した．今後の課題として，貪欲法のさらなる探索効率の改善，探索によって求めた解の詳細な分析などを進める予定である．

参考文献

[Kimura 07] M. Kimura, K. Saito, and R. Nakano: Extracting influential nodes for information diffusion on a social network. Proceedings of the 22nd AAAI Conference on Artificial Intelligence, pp. 1371–1376 (2007).

[Leskovec 07] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance: Cost-effective outbreak detection in networks. Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining, pp. 420–429 (2007).

[室田 07] 室田一雄: 離散凸解析の考えかた 最適化における離散と連続の数理. 共立出版 (2007).

[Newman 03] M. E. J. Newman: The structure and function of complex networks. SIAM Review, vol. 45, no. 2, pp. 165–256 (2003).

[斉藤 05] 斉藤和巳, 木村昌弘, 上田修功: 文書トピックに関する認知科学的実験. 知識ベースシステム研究会 69, pp. 51–56 (2005).