社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－DBS－139 (9)
2006－ＦＩ－ 83 (9)
2006／5／31

ライングラフと局所的類似度にもとづくマルチクラスタリングアルゴリズム

石田和成
†島根県立大学 総合政策学部
〒697-0016 島根県浜田市野原町 2433-2
E-mail: †k-ishida@u-shimane.ac.jp

Abstract: 潜在的ブログコミュニティ(LBC)を抽出するために、エッジの局所的な類似度と、半順序エッジ集合にもとづき、マルチクラスターを抽出できる、共有された関心(Shared Interest, SI) アルゴリズムを開発した。LBC とは、お互いに共通の関心を持つブロガーのクラスターのことである。ブロガーがお互いに知り合いではない場合、LBC は出会いのきっかけを提供できるため、インターネットにおける知識の自律的組織化に役立つ。関心の共通度が高い LBC を抽出するために、SI アルゴリズムにより密なクラスターが得られる条件を導出する。また、SI アルゴリズムが密なクラスターを抽出できることを示すために、疎および密なグラフのクラスター抽出例を示す。さらに、実際のブログ空間データへの適用例を示す。

Keyword: マルチクラスタリング、共有された関心、共参照、半順序集合、潜在的ブログコミュニティ、スケーラビリティ

A Multi Clustering Algorithm based on Line Graph and Local Similarity

Kazunari Ishida†
†Faculty of Policy Studies, the University of Shimane
2433-2 Nobara-cho, Hamada-shi, Shimane, 697-0016, JAPAN
E-mail: †k-ishida@u-shimane.ac.jp

Abstract: I developed the shared interests (SI) algorithm, a multi-clustering algorithm that extracts latent blog communities (LBCs) based on the local similarity of edge pairs and the node degree of partial ordered edge sets. As a means to promote the autonomous organization of knowledge on the Internet, LBCs can be used to create meeting spaces for bloggers who write about similar or closely related topics but do not know each other. To extract a cluster with closely related topics, the SI algorithm employs co-citation information as locally shared interests between bloggers and partially ordered edge sets in ascending order of node degree.

Keyword: multi-clustering, shared interest, co-citation, partial order set, latent blog community, scalability

1. はじめに

本研究では、ブロガー間で局所的に共有された関心(Shared Interest, SI)にもとづくマルチクラスタアルゴリズムを開発した。ブログ空間では、意味のあるもの、気まぐれなもの、自動生成されたもの、意図的に歪曲されたもの(スパムブログ[7])など、継続的に、様々な情報が生み出されている。この混沌とした空間から、ブロガーの関心のさまざまな側面を分類するために、SI アルゴリズムは、ブログから他のページへのリンクを、ブロガーの持つ何らかの関心として扱い、ブロガー間での局所的な関心の類似度にもとづき、マルチクラスターを抽出する。

この研究の背景には、ブログ空間の情報爆発がある。総務省の統計[8]によると、日本の社会にブログが浸透したことが分かる。コンテンツマネジメントシステム(content management systems, CMS)の普及とその利便性により、様々な人々が、個々の情報を開示できるようになり、消費者により生成されたメディア(consumer-generated media, CGM)を形成するに至っている。このメディアは、潜在的に価値のある情報を提供する可能性がある反面、気まぐれな書き込み、自動生成情報、あるいは、スプロガー[7]による意図的な歪曲により、無意味なもの、あるいは有害な情報が排出される易い状況になっている。

2. ブログ空間と情報の組織化

ブログ空間における自律的知識編集を促進するために、関連研究[2]において、潜在的ブログコミュニティ(Latent Blog Community, LBC)を提案した。これは、関心が近いにも関わらず、お互いの存在を知らないブロガーの集まりである。この背景にはブログ空間の爆発

的拡大がある。

ブログツールなどの CMS は、ブログ空間を含むインターネットにおける情報爆発を加速している。これらのツールは、人々をひきつけるために、ニュースヘッドライン、ブログを更新したブロガー、オンラインショップへのアフィリエイト、といったリンクを、自動的にブログに表示するための、様々な仕組みを提供する。そのため、ブログには、ブロガー自身の意図した情報以外に、意図せず自動的に生成された情報もたくさん含まれる。また、アフィリエイト収入のために、複数のブログを作成する人々(ペルソナ[2])も存在する。ブログ空間から意味のある情報を抽出するには、無意味な情報や、意図的に歪曲された情報などを、分離したクラスターとして抽出する必要がある。

これまでに、階層的クラスタリング、K-means、SDD[5]、相関ルール[1][8]、など様々な分類手法が開発されている。これらの手法は計算量の問題や抽出できるクラスターの制約により、莫大な情報を持つブログ空間から、多種多様な関心を抽出する課題には適していない。その課題に適した、アルゴリズムが満たすべき要件をあげると次のようになる。(1)ブロガーが複数のクラスターに所属することを許容(マルチクラスターを抽出)できること、(2)ブロガーの意図にもとづく情報と、その他の自動生成された情報とを分けてクラスター抽出できること、(3)一般的なグラフに適用できること、(4)データの増大に対して、計算量の爆発が生じないこと。次節では、この要件を満たすアルゴリズムを提案する。

3. 共有関心(SI)アルゴリズム

ブログ空間からマルチクラスターを抽出するために、前節の最後で述べた要件を満たすアルゴリズムを開発した。この、共有関心(Shared Interest, SI)アルゴリズムは、ウェブグラフにおいて、ノードの代わりに、エッジをクラスタリングの基本単位として用いる。これは、ブロガーの関心が、ブログから張られているリンク(エッジ)によって示されていると考えられるためである。これにより、複数のクラスターに所属するブロガーを抽出できるとともに、リンクによって表現されるブロガーの関心の多様性を扱うことができる。共有された関心の尺度として、エッジペア間の局所的な類似度を定義する。密度の高いクラスターを抽出すると同時に、計算量を削減するために、エッジ両端のノードの次数にもとづき、エッジの半順序集合にもとづき、エッジ間を横断的に推移し、クラスターを抽出するアルゴリズムを開発した。

3.1. 半順序エッジ集合

エッジ間の関係を示すために、元のグラフのノードとエッジを入れ替えた、ライングラフを用いる。用語の混乱を避けるために、元のグラフにおける「ノード」「エッジ」を、ライングラフでは「エレメント」「パス」と呼ぶこととする(図 1)。「ノード」は「パス」に、「エッジ」は「エレメント」に対応する。

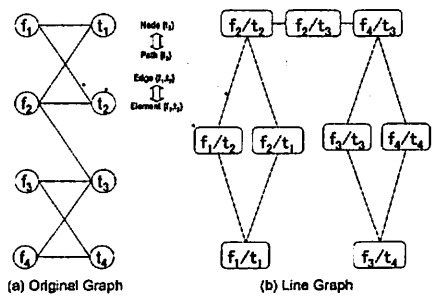


図 1: 元のグラフとライングラフ

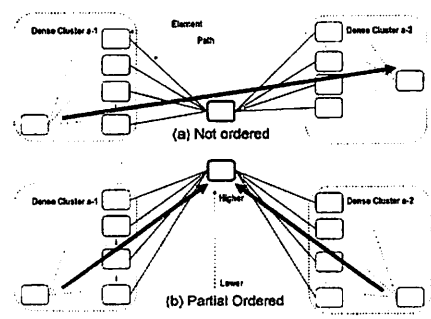


図 2: ライングラフとクラスター分離

類似度の観点からクラスターにエレメントを加えるとき、エレメントの順序を考慮しない場合、図 2(a)に示すように、2 つの異なる密なクラスターが、密度の低下した 1 つのクラスターに結合される場合がある。これは次数の高いエレメントが、次数の低いエレメントより先に追加されることが原因である。これに対して、図 2(b)のように、次数の小さい順にエレメントをクラスターへ追加していくと、密度低下をもたらすクラスターの結合を避けることができる。このようなエレメントの追加を行うために、半順序エッジ(エレメント)集合を定義する。この半順序集

合にもとづくエッジ横断の効果を、3.4 および 3.5.3 で示す。

半順序エッジ集合は、半順序ノード集合の直積集合において隣接関係の無いノードペアを取り除いたものとして定義する。ノードの次数にもとづき、昇順にノード ID を割り当てる。つまり、あるグラフにおけるノード n_i, n_j, n_k ($i<j<k$) の次数の関係は

$$d(n_i) \le d(n_j) \le d(n_k) \quad \ldots (1)$$

ここで、$d(n_i)$ はノード n_i の次数を表す。

エッジ間の順序は単純にノードの順序で定義できる。2 つのエッジ間に順序関係があるとき、それらは 1 つのノードを共有すると同時に、それぞれ 1 つずつ異なるノードを持つ。この異なるノードの順序に従い、2 つのエッジ順序が定義できる。この元のグラフにおけるエッジの順序は、ライングラフのエレメントの順序となる。

エレメントの順序にもとづき、ライングラフをハッセ図で表現する。例えば、順序関係の示されていない図 2(a) はライングラフ、順序関係が示されている(b)はハッセ図である。ハッセ図において、極小エレメントから極大エレメントへ横断的に推移することにより、マルチクラスターを抽出する。

3.2. エッジペアの局所的類似度

2 つのエッジ間の類似度は、2 つの異なるノード n_i と n_j における共参照数により定義する(図 4)。

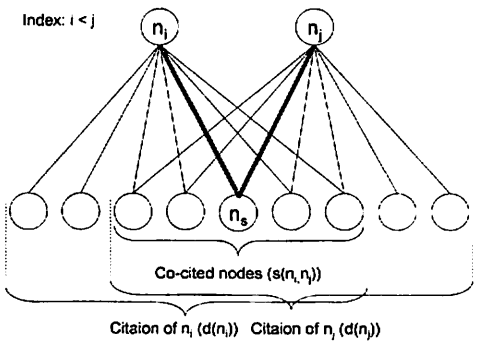


図 3: エッジペアの類似度

2つのエッジ$\{n_i,n_s\}$および$\{n_j,n_s\}$は共参照ノード集合において、共通のノード n_s を共有している。ここで、2つのエッジ間で異なる2つのノード、n_i と n_j の間で、Jaccard 係数を用いて、2つのエッジ間の局所的な類似度を定義する。

$$J(n_i,n_j) = \frac{s(n_i,n_j)}{d(n_i)+d(n_j)-s(n_i,n_j)} \quad \ldots (2)$$

ここで、$s(n_i,n_j)$ は共参照数である。

エッジ間の高い局所的類似度により構成されるクラスターは密になる。ハッセ図の横断的な推移における、密度や局所的類似度の詳細な分析については、3.4 で説明する。

3.3. 共参照の上限と下限

2つのエレメント(例えば、ライングラフにおける、$\{n_i,n_j\}$ と $\{n_j,n_k\}$)の間の、パスの特徴を示すために、元のグラフにおいて、2つのノード間の共参照数の上限と下限を定義する。上限は局所的類似度の定義より、ノードn_i 、n_jの間の最小次数である。つまり、$i<j$とすると

$$s^{\max}(n_i,n_j) = \min(d(n_i),d(n_j)) = d(n_i) \quad .. (3)$$

類似度の下限を定義するために、任意のノード間で、閾値 δ を以下のように定義する。

$$J(n_i,n_j) \ge \delta \quad \ldots (4)$$

この類似度の閾値 δ にもとづき、この不等式を共参照数についてまとめる。

$$\delta \le J(n_i,n_j) = \frac{s(n_i,n_j)}{d(n_i)+d(n_j)-s(n_i,n_j)}$$

$$s(n_i,n_j) \ge \frac{\delta}{\delta+1}\left[d(n_i)+d(n_j)\right] \quad .. (5)$$

よって、下限は

$$s^{\min}(n_i,n_j) = \frac{\delta}{\delta+1}\left[d(n_i)+d(n_j)\right] \quad \ldots (6)$$

この下限にもとづき、高い密度のクラスターを抽出するために、元のグラフのライングラフにおいて、閾値 δ 以下の、エレメント間のパスを取り除く。

パスと閾値 δ の関係を知るために、上限と下限の条件から、$d(n_i)$と$d(n_j)$ ($i<j$) の関係を導出する。

$$s^{\min}(n_i,n_j) \le s(n_i,n_j) \le s^{\max}(n_i,n_j)$$

$$\frac{\delta}{\delta+1}\left[d(n_i)+d(n_j)\right] \le d(n_i)$$

$$\delta d(n_j) \le d(n_i) \quad \ldots (7)$$

同様に、$d(n_j)$と $d(n_k)$(j<k)について、以下の関係が導出できる。

$$\delta d(n_k) \le d(n_j) \quad \ldots (8)$$

式(7)(8)より、以下の関係が成り立つ。

$$\delta d(n_j) + \delta d(n_k) \le d(n_i) + d(n_j)$$

$$\frac{d(n_j)+d(n_k)}{d(n_i)+d(n_j)} \le \frac{1}{\delta} \quad \ldots (9)$$

ノードの次数に関するこの条件は、類似度の下限の閾値 δ 以下の全てのエレメントが取り除かれたときに、残ったエレメント間で常に満たされる。

3.4. 局所的類似度と共参照

3.1の図1(a)で指摘したように、クラスターにエレメントを追加するとき、2つの異なる密なクラスターが密度の低下した1つのクラスターとなる可能性がある。このようなクラスターの密度低下を招く不要な結合を防ぐために、ハッセ図において、次数の低いエレメントから高いエレメントへ推移しながら、クラスターにエレメントを追加するとき、類似度の単調減少は良い指標となる。そのため、類似度の単調減少の条件を調べる。2つの隣接するエレメント$\{n_i,n_j\}$、$\{n_j,n_k\}$間の類似度増加割合を以下のように定義する。

$$\frac{J(n_j,n_k)}{J(n_i,n_j)} = \frac{\dfrac{s(n_j,n_k)}{d(n_j)+d(n_k)-s(n_j,n_k)}}{\dfrac{s(n_i,n_j)}{d(n_i)+d(n_j)-s(n_i,n_j)}}$$

$$\frac{J(n_j,n_k)}{J(n_i,n_j)} = \frac{s(n_j,n_k)}{s(n_i,n_j)} \frac{d(n_i)+d(n_j)-s(n_i,n_j)}{d(n_j)+d(n_k)-s(n_i,n_k)_{jk}} \quad (10)$$

この割合が 1 より小さいとき、エレメント間のパスを推移するとき、類似度は減少する。2つのエレメント$\{n_i,n_j\}$から$\{n_j,n_k\}$への局所的な類似度が常に減少する条件を見つけるために、以下の不等式をまとめる。

$$\frac{J(n_j,n_k)}{J(n_i,n_j)} = \frac{s(n_j,n_k)}{s(n_i,n_j)} \frac{d(n_i)+d(n_j)-s(n_i,n_j)}{d(n_j)+d(n_k)-s(n_i,n_k)_{jk}} \le 1$$

$$\frac{s(n_j,n_k)}{s(n_i,n_j)} \le \frac{d(n_j)+d(n_k)}{d(n_i)+d(n_j)} \quad \ldots (11)$$

ラングラフにおいて、式(11)が満たされるとき、エレメント$\{n_i,n_j\}$ から $\{n_j,n_k\}$へのパス$\{n_j\}$において、類似度は減少する。

密度の高いクラスター間の不必要な結合を避け、局所的類似度が単調減少するように、類似度の下限の閾値δまで、エレメントを加えることにより、密度の高いクラスターを抽出することができる。

3.5. SI アルゴリズムの動作例

SIアルゴリズムの動作を説明するために、2つの例を示す。SI アルゴリズムは一般のグラフに適用できるが、ブログの更新情報のグラフは二部グラフとなるため、動作例も二部グラフで説明する。

3.5.1. 疎なグラフへの適用

最初のサンプルグラフ G_1 は、2 つの完全二部グラフ G_a、G_bをエッジ$\{f_2,\ t_3\}$で結合したグラフである(図 4)。このサンプルグラフを以下に定義する。

$$G_a = (F_a, T_a, E_a),\ F_a = \{f_1, f_2\},\ T_a = \{t_1, t_2\}$$
$$E_a = \{\{f_1,t_1\},\{f_1,t_2\},\{f_2,t_1\},\{f_2,t_2\}\}$$
$$G_b = (F_b, T_b, E_b),\ F_b = \{f_3, f_4\},\ T_b = \{t_3, t_4\}$$
$$E_b = \{\{f_3,t_3\},\{f_3,t_4\},\{f_4,t_3\},\{f_4,t_4\}\}$$
$$G_1 = (F_1, T_1, E_1),\ F_1 = F_a \cup F_b,\ T_1 = T_a \cup T_b$$
$$E_1 = E_a \cup E_b \cup \{f_2, t_3\}$$

ここで、F_a はG_aにおけるブロガーの集合(Fun集合)、T_a はG_aにおいてブロガーに引用されたページの集合(Target集合)、E_aはG_aにおける、F_a(Fun集合)とT_a(Target集合)の間のリンク集合である。

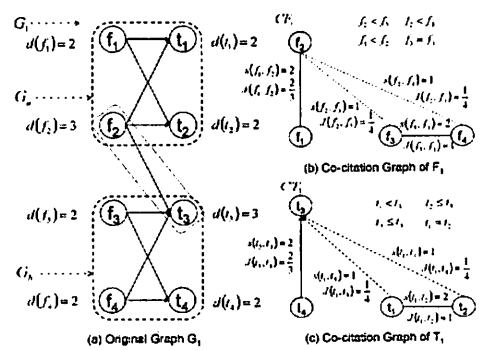


図 4: グラフ G_1

ノードの次数は

$$d(f_1)=2,\ d(f_2)=3,\ d(f_3)=2,\ d(f_4)=2$$
$$d(t_1)=2,\ d(t_2)=2,\ d(t_3)=3,\ d(t_4)=2$$

2 つのにおけるノード間の半順序関係は

$$f_3 = f_4,\ f_3 < f_2,\ f_4 < f_2,\ f_1 < f_2$$
$$t_1 = t_2,\ t_1 < t_3,\ t_2 < t_3,\ t_4 < t_3$$

Fun および Target 集合それぞれにおける共参照は

$$CF_1 = \{\{f_1, f_2\}, \{f_2, f_3\}, \{f_2, f_4\}, \{f_3, f_4\}\}$$
$$CT_1 = \{\{t_1, t_2\}, \{t_1, t_3\}, \{t_2, t_3\}, \{t_3, t_4\}\}$$

ここで、CF_1、CT_1は、それぞれFun、Target集合における共参照関係の集合である。

Fun 集合における各々のペアの共参照数と類似度は

$$s(f_1, f_2) = 2,\quad J(f_1, f_2) = 2/3$$
$$s(f_2, f_3) = 1,\quad J(f_2, f_3) = 1/4$$
$$s(f_2, f_4) = 1,\quad J(f_2, f_4) = 1/4$$
$$s(f_3, f_4) = 2,\quad J(f_3, f_4) = 1$$

Target 集合における各々のペアの共参照数と類似度は

$$s(t_1, t_2) = 2,\quad J(t_1, t_2) = 1$$
$$s(t_1, t_3) = 1,\quad J(t_1, t_3) = 1/4$$
$$s(t_2, t_3) = 1,\quad J(t_2, t_3) = 1/4$$
$$s(t_3, t_4) = 2,\quad J(t_3, t_4) = 2/3$$

類似度の下限を 1/2 (δ=1/2)とすると、図 4 の(b)(c)の点線で示される、CF_1 における共参照関係$\{f_2, f_4\}$および$\{f_2, f_1\}$、CT_1 における共参照関係$\{t_1, t_2\}$および$\{t_2, t_3\}$は無視される。

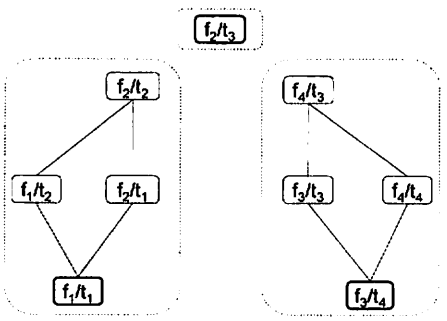


図 5: G_1 のハッセ図

元のグラフ CF_1 におけるエッジ集合 E_1 は、F_1 と T_1 の直積集合から、リンクの無いノードペアを除いたものである(図 5 において取り除かれたペアは、$\{f_1, t_3\}$, $\{f_1, t_4\}$, $\{f_2, t_4\}$, $\{f_3, t_1\}$, $\{f_3, t_2\}$, $\{f_4, t_1\}$, $\{f_4, t_2\}$)。この図は、半順序のエレメントによるライングラフ、つまり、ハッセ図となる。図 6 は、半順序エレメント(エッジ)集合の順序に従う、エレメントの類似度行列である。

最小エレメントは、$\{f_1, t_1\}$、$\{f_3, t_4\}$、そして$\{f_2, t_3\}$である($\{f_1, t_2\}$と$\{f_4, t_4\}$は、$\{f_1, t_1\}$と$\{f_3, t_4\}$の代わりに、最小エレメントとして用いることができる)。

半順序集合における、極小エレメントから極大エレメントへの推移において、類似度の下限より小さいパスは無視される(図 5 のハッセ図で取り除かれたパスは、$\{f_3, t_3\}$-$\{f_2, t_3\}$、$\{f_4, t_3\}$-$\{f_2, t_3\}$、$\{f_2, t_1\}$-$\{f_2, t_3\}$、そして$\{f_2, t_2\}$-$\{f_2, t_3\}$)。図 5,6 における推移により、図 5 のハッセ図において、点線で囲まれた 3 つのクラスターが抽出される。

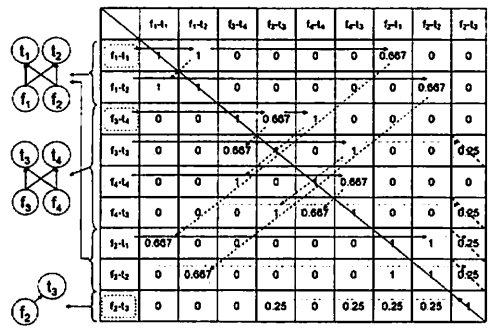


図 6: G_1 エッジ (エレメント) 類似度行列

3.5.2. 密なグラフへの適用

図 7 は、1 つ目のグラフ G_1 に、2 つのエッジ$\{f_1, t_3\}$、$\{f_2, t_4\}$を加えたグラフ G_2 を示している。

$$G_2 = (F_2, T_2, E_2),\ F_2 = F_1,\ T_2 = T_1$$
$$E_2 = E_1 \cup \{f_1, t_3\} \cup \{f_2, t_4\}$$

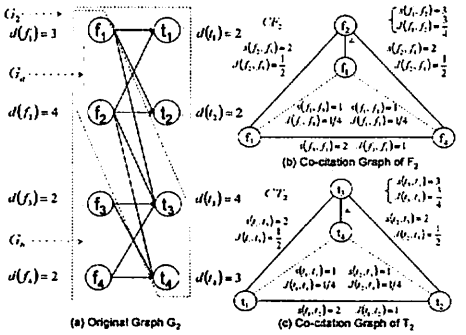


図 7: グラフ G_2

グラフ G_1 と G_2 の間で異なるノードの次数は

$$d(f_1) = 3,\ d(f_2) = 4,\ d(t_3) = 4,\ d(t_4) = 3$$

Fun 集合 F_2、Target 集合 T_2 における半順序関係は、F_1、T_1 それぞれに対して以下を追加したものとなる。

$$f_3 < f_1,\ f_4 < f_1 \quad \text{および} \quad t_1 < t_4,\ t_2 < f_4$$

Fun、Target 集合における共参照は

$$CF_2 = CF_1 \cup \{f_1, f_3\} \cup \{f_1, f_4\}$$
$$CT_2 = CT_1 \cup \{t_1, t_4\} \cup \{t_2, t_4\}$$

グラフ G_1 と G_2 との間で異なる、Fun 集合における共参照数と、類似度は

$$s(f_1, f_2) = 3, \quad J(f_1, f_2) = 3/4$$
$$s(f_2, f_3) = 2, \quad J(f_2, f_3) = 1/2$$
$$s(f_2, f_4) = 2, \quad J(f_2, f_4) = 1/2$$

グラフ G_1 と G_2 との間で異なる、Target 集合における共参照数と、類似度は

$$s(t_1, t_3) = 2, \quad J(t_1, t_3) = 1/2$$
$$s(t_2, t_3) = 2, \quad J(t_2, t_3) = 1/2$$
$$s(t_3, t_4) = 3, \quad J(t_3, t_4) = 3/4$$

類似度の下限を 1/2 (δ=1/2)とすると、図 7 の(b)(c)の点線で示される、CF_2 における共参照関係$\{f_1, f_3\}$および$\{f_1, f_4\}$、CT_2 における共参照関係$\{t_1, t_4\}$および$\{t_2, t_4\}$は無視される。元のグラフ CF_2 におけるエッジ集合 E_2 は、F_2 と T_2 の直積集合から、リンクの無いノードペアを除いたものである(図 8 において取り除かれたペアは、$\{f_1, t_4\}, \{f_3, t_1\}, \{f_3, t_2\}, \{f_4, t_1\}, \{f_4, t_2\}$)。

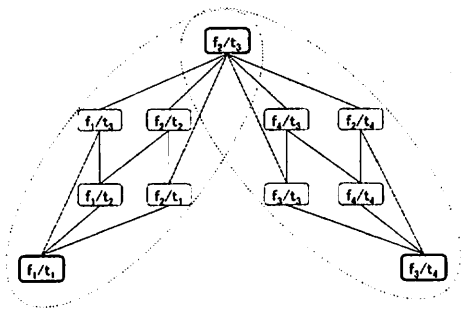


図 8: G_2 のハッセ図

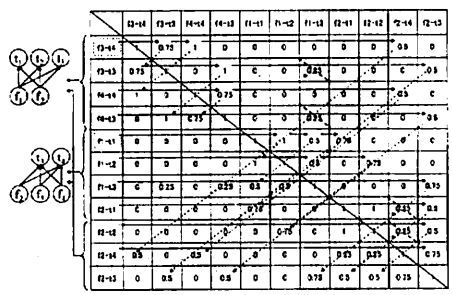


図 9: G_2 のエッジ（エレメント）類似度行列

図 9 は、半順序エレメント集合における、エレメント類似度行列である。極小エレメントは$\{f_1, t_1\}$と$\{f_3, t_4\}$である。グラフ G_1 に示したように、$\{f_1, t_1\}$と$\{f_3, t_4\}$の代わりとして、$\{f_1, t_2\}$と$\{f_4, t_4\}$を極小エレメントとして用いることができる。

半順序集合における極小エレメントから極大エレメントへの推移において、類似度の下限以下のパスは無視される(例えば、$\{f_3, t_3\}$-$\{f_1, t_3\}$, $\{f_4, t_3\}$-$\{f_1, t_3\}$, $\{f_2, t_1\}$-$\{f_2, t_4\}$, そして$\{f_2, t_2\}$-$\{f_2, t_4\}$は、図 8 におけるハッセ図から取り除かれる)。図 8,9 に示されるように、横断的な推移により、2 つのクラスターが抽出される。

3.5.3. 考察

エッジ集合における順序の影響を調べるために、図 10 に示す 3 つ目の例を分析する。この図に示すハッセ図の構造は、Target 集合 T_2 において t_3 と t_4 の順序を入れ替えた以外は、図 8 と同じである。ここで、図 7(c)に示した Target 集合に注目する。ノードペア$\{t_1, t_2\}$と$\{t_2, t_3\}$の間のパスは不等式(11)を満たす。

$$s(t_2, t_3)/s(t_1, t_2) = 2/2 = 1$$
$$(d_{12} + d_{13})/(d_{11} + d_{12}) = (2+4)/(2+2) = 3/2$$

しかし、ノードペア$\{t_2, t_3\}$と$\{t_3, t_4\}$間のパス$\{t_3\}$は不等式(11)を満たさない。

$$s(t_3, t_4)/s(t_2, t_3) = 3/2$$
$$(d_{13} + d_{14})/(d_{12} + d_{13}) = (4+3)/(2+4) = 7/6$$

つまり、類似度は、パスの経路において、類似度は単調減少していない(Fun 集合においても同様)。

$$J(t_1, t_2) \geq J(t_2, t_3) \leq J(t_3, t_4)$$

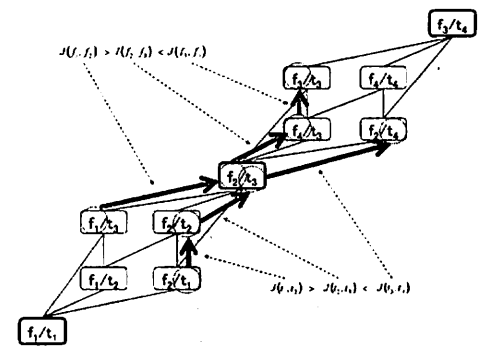


図 10: G_2 の t_3 と t_4 の順序を入れ替えたハッセ図

図 10 のハッセ図では、極小エレメント$\{f_1, t_1\}$から極大エレメント$\{f_3, t_4\}$への推移により、1 つのクラスターが抽出される。エレメント$\{f_2, t_1\}$から$\{f_2, t_2\}$へのパスにおいて、最初

の 2 つのエレメント間で、類似度が減少するが、続く 2 つのエレメント間のパスでは、類似度が増加する。 結果として、エレメント$[t_2,t_3]$の通過は、2 つの密度の高いクラスター間の、不要な結合をもたらす。つまり、次数が高いエレメント$[t_2,t_3]$は、不要な結合を避けるために、極大エレメントとなる必要がある。

3.6. アルゴリズムの実装

SI アルゴリズムの実装には、GNU c++ (g++)コンパイラと STL(standard template library)を用いた。実装においては、計算量削減のため、3.5.1 および 3.5.2 において説明した 2 つの例とは逆に、極大エレメントから、極小エレメントへの推移を行うアルゴリズムを用いた。極小エレメントからの推移では、極大エレメントに近づくにつれ、重複して通過するパスが多くなる。これは、閾値 δ が小さいとき、ハッセ図における極大エレメント数が少なくなり、より重複度が高くなる。これに対して、極大から極小エレメントへの推移の場合、ここのエレメント(エッジ)を 1 度だけ調べると、極小エレメントかどうか判断できる。そのため計算量は、O(e)となる(e はエッジ数)。

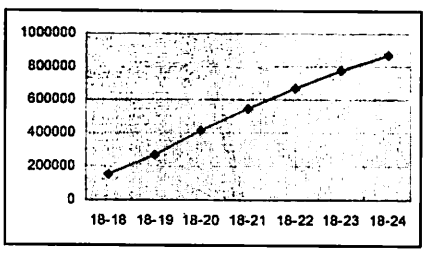


図 11: エッジの累積数

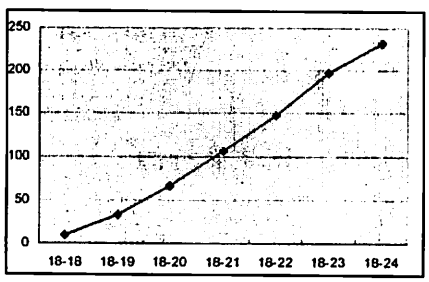


図 12: データ量と処理時間

これを実験的に調べるために、日本のブログサイトと PING サーバの 1 週間(2005 年 12 月 19 日～24 日)の更新情報に、SI アルゴリズムを適用した。図 11 はブログから張られるリンク数と期間毎との関係を示しており、期間の長さに比例して、リンク数が増大することがわかる。

図 12 は、全ての最小エレメントの抽出にかかる処理時間と、データ期間との関係を示している。最小エレメント数は、抽出されるクラスタ数に対応する。この図は、SI アルゴリズムが、計算量 O(e)(e はリンク数、あるいはエッジ数)となることが実験的正しいことを示している。

4. ブログ空間における LBC 抽出

実際のブログ空間データに対して、SI アルゴリズムを適用した例を示す。ブロガーの関心の対象として、NHK と首相官邸、および、ブログをマーケティングツールとして活用している 1 つのオンラインショップ [5][6] を取り上げ、これらのページへのリンクがあるブロガーを含むクラスターを紹介する。

4.1. NHK (Japan Broadcasting Corporation)

図 13 は、NHK の URL http://www.nhk.or.jp/を含む 2 つのクラスターを示している。1 つ目のクラスターのブロガーは、TV 番組やその出演者の視聴を楽しむ人たちである。2 つ目のクラスタは、石川県能登半島周辺の農業や食品について共通の関心を持っている人たちである。このクラスターにおいて、1 人のブロガーが NHK 大河ドラマへのリンクを張っていた。図 14 は、アニメーション番組について共通の関心を持っている人々のクラスターであり、NHK のアニメーション番組へのリンクが張られていた。これらの LBC にもとづき、NHK は TV プログラムの改善を、ブロガーは、共通の関心を持つブロガーとの交流範囲の拡大を、それぞれ行うことができる。

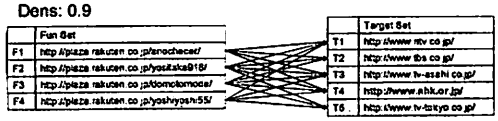
4.2. 首相官邸

図 15 は、首相官邸の URL http://www.kantei.go.jp/ を含むクラスターである。このブロガーたちは、拉致問題の解決について議論している。1 人のブロガーが首相への意見投稿ページへのリンクを張っていた。

4.3. オンラインショップ

ブログをマーケティングに活用しているオンラインショップ[5][6]の名前の文字列「kenko」を URL に含むページへのリンクを持つブロガーのクラスターを抽出したところ、全てのブロガーから、全て共通のページ群にリンクが張られている、完全二部グラフのクラスターとなっていた。このクラスターは、ペルソナ[4]の可能性が高い。

CID: 59660 Type: M2M
Fnum: 4 Tnum: 5
Dens: 0.9



CID: 30204 Type: M2M
Fnum: 4 Tnum: 6
Dens: 0.5

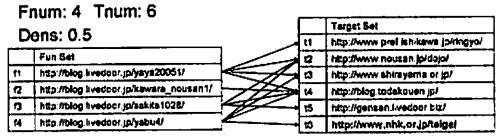


図 13: NHK を含んだクラスター(1)

CID: 64804
Type: M2M
Fnum: 5
Fnum: 21
Dens: 0.752381

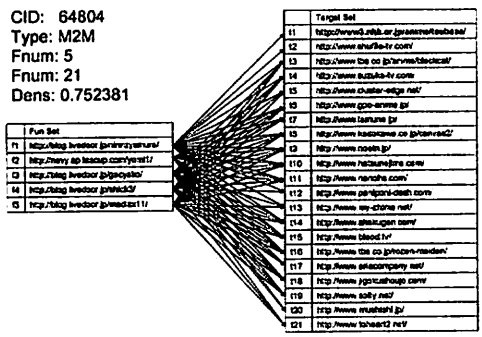


図 14: NHK を含んだクラスター(2)

5. まとめ

本研究では、ブログ空間からマルチクラスターを抽出するための、SI アルゴリズムを開発し、その特徴と、実データへの適用例を示した。今後、実データへの適用を継続し、社会現象の分析を行うとともに、アルゴリズムや実装のさらなる開発を行う。

参考文献

1. Agrawal, R. and R. Srikant, "Fast Algorithms for Mining Association Rules," in *Proc. 20th Int. Conf. Very Large Data Bases (VLDB'94)*, 1994, pp. 487–499.
2. Ishida, K., "Extracting Latent Weblog Communities: A Partitioning Algorithm for Bipartite Graphs," Proceedings of the 2nd Annual Workshop on the Weblogging Ecosystem - Aggregation, Analysis and Dynamics in the 14th International World Wide Web Conference (WWW2005), May 10 - 14, 2005.
3. IT Media, "7 分で分かる 4 月のブログ界", http://www.itmedia.co.jp/enterprise/articles/0505/05/news011.html.
4. IT Media, "7 分で分かる 8 月のブログ界", http://www.itmedia.co.jp/enterprise/articles/0512/09/news108_2.html.
5. Koyutürk, M., A. Grama, and N. Ramakrishnan, "Compression, Clustering and Pattern Discovery in Very High Dimensional Discrete-Attribute Datasets," IEEE Transaction on Knowledge and Data Engineering, April 2005, pp. 447 – 461.
6. 総務省, "ブログ・SNS(ソーシャルネットワーキングサイト)の現状分析及び将来予測", 報道資料, http://www.soumu.go.jp/s-news/2005/050517_3.html, 2005 年 5 月.
7. Wikipedia, http://en.wikipedia.org/wiki/Splog
8. Zaki, M. J., "Scalable Algorithms for Association Mining," IEEE Transaction on Knowledge and Data Engineering, Vol. 12, No. 3, May/June 2000, pp. 372 – 390.

CID: 36999 Type: M2M
Fnum: 4 Tnum: 34
Dens: 0.338235

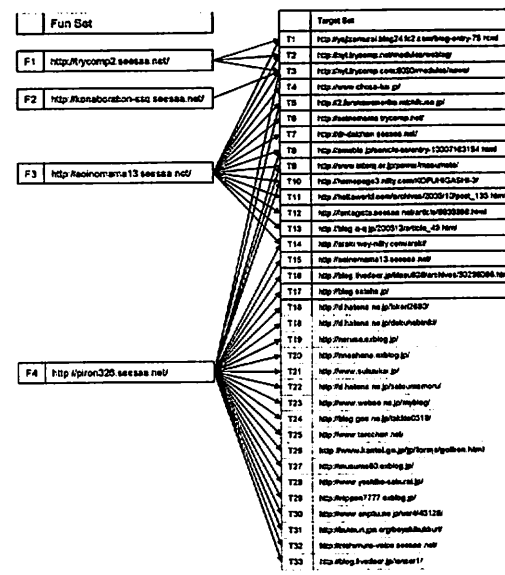


図 15: 拉致問題