潜在的ウェブログコミュニティ抽出のための二部グラフ分割アルゴリズム

石田 和成

東京農業大学 国際食料情報学部 〒156-8502 東京都世田谷区桜丘 1-1-1

E-mail: k-ishida@nodai.ac.jp

概要 インターネットにおける自律的な情報組織化を促進するための概念として、潜在的ウェブログコミュニティ（Latent weBlog Community, LBC）を提案し、これを抽出するアルゴリズムとして、最弱対（the Weakest Pair, WP）アルゴリズムを開発した。LBC は類似の嗜好を持つ相互に知り合いではないブロガーたちの出会いの場であり、このブロガーたちのコミュニケーションにより、自律的な情報組織化を促進する。LBC は、PING サーバにより得ることができるウェブログの更新情報と、ブロガーたちの共参照情報にもとづく二部グラフ集合から抽出できる。この抽出を行う WP は、共参照、被共参照情報を用いることにより、完全二部グラフに近い状態でクラスタを分割するため、類似トピックの LBC が抽出できる。この評価のため、従来の最短路ビットウィーンネス（the Shortest Path Betweeness, SPB）を用いた分割手法との比較を行い、WP の有効性を示す。これを用いた 3 つの LBC の抽出例とともに、二次的効果、同一ブロガーによる複数ウェブログの抽出というペルソナ検出の例を示す。

キーワード 潜在的ウェブログコミュニティ、二部グラフ、共参照、分割アルゴリズム、ペルソナ検出

A Partitioning Algorithm for Bipartite Graphs to Extract Latent Weblog Communities

Kazunari Ishida

Tokyo University of Agriculture, Faculty of International Agriculture and Food Studies

1-1-1 Sakuragaoka, Setagayaku, Tokyo, 156-0054, JAPAN

E-mail: k-ishida@nodai.ac.jp

Abstract: I propose the concept of a latent weblog community (LBC), as a means to promote the autonomous organization of knowledge on the Internet. Such communities can be illustrated in terms of bipartite graphs based on weblog update information, and they can effectively function to create meeting spaces for bloggers who write about similar or closely related topics but do not know each other. To extract these communities from blogspace, I developed a partitioning algorithm known as the Weakest Pair (WP) algorithm, which separates the weakest pairs of bloggers and webpages, respectively, using co-citation information. As a result of numerical evaluation, the WP algorithm is more effective than the Shortest Path Betweenness (SPB) algorithm in terms of information loss and completeness of bipartite graphs. I will provide three examples of LBC extracted using the WP algorithm and report its secondary effects, i.e. personae detection, the detection of a set of weblogs owned by a single blogger.

Keyword: Latent weblog community, bipartite graph, co-citation, partitioning algorithm, personae detection

1. はじめに

インターネットを有効で使いやすい知識の共有地として維持するために、情報組織化の問題は重要である。情報を整理・組織化する仕組みとして、ロボット型、ディレクトリ型の検索エンジンがあるが、ウェブの規模や複雑さの増大により、その運営費用は増大の一途を辿っている。他方、情報発信者による自律的な情報組織化の現象として､ウェブコミュニティがある。これまでのウェブコミュニティに関する研究で、ウェブページ間の参照関係において、完全二部グラフを形成しているウェブページ群は、意味的にまとまったコミュニティを形成しているという観察結果が示されている（Kumar & et al., 1999)。

2005 年の現在、ウェブログなどのコンテンツ･マネジメント･システム（CMS）の普及により、多種多様な人々が簡単に情報発信できるようになっている。PING サーバの標準化（Winer, 2001）により、ブロガーは、自分のウェブログの更新をそのサーバ上で通知することができる。また、この更新情報は RSS フォーマット（Libby, 1999）により配信されている。ウェブログはもともと受動的な情報公開の仕組みであったが、更新情報を配信できる、PING サーバや RSS の標準化と普及により、能動的で即時的な情報公開がなされるようになった。これらの仕組みが利用できる現在、ウェブログ空間は、インターネットを介したボランタリな分散センサーが世界全体に配置されている状態とな

っている。この分散センサーにより、各地域での最新情報が公開されるため、これまでは現地以外では入手困難であった情報が、世界中どこでも入手できる可能性が高まっている。その反面、多種多様な情報が大量に更新されるため情報洪水を引き起こしている。このように情報量の爆発が加速している現在や今後のインターネットにおいて、有用な知識の共有地としてインターネットを維持発展させるためには、ウェブコミュニティを育てていく必要がある。特にウェブログは万人がインターネットで容易に発言できる仕組みであるため、ウェブログコミュニティを育てることは、今後のインターネットの発展に大きな影響を持つ。

本研究では、ウェブログコミュニティを育てるための概念として、潜在的ウェブログコミュニティ（Latent weBlog Community, LBC）を提案する。これは、持っている興味が近いにもかかわらず、お互いに知り合いでは無いブロガーたちの出会いの場である。この場を実現するために、PING サーバにより配信されるウェブログ更新情報と、ウェブログからの参照情報を利用し、二部グラフを構成する。PING サーバには、1 つのウェブログサイトにとどまらず、様々なブロガーが登録するので、この更新情報の中には、お互いに嗜好が似ているけれども、相互のウェブログの存在を知らないブロガーたちのペアがたくさんあるものと考えられる。そのため、PING サーバによって配信されるウェブログ更新情報にもとづき、それぞれのウェブログで参照しているリンク情報を抽出し二部グラフを構成する。2 人のブロガーの興味が近いかどうかは、どの程度、共通のホームページを共参照しているか、という情報で知ることができる。

しかしこの共参照情報には非常に多くのノイズが含まれている。例えば、ウェブログを観察すると、ブロガーの気まぐれなリンクや、ウェブログサイトにより自動的に埋め込まれるポータルサイトへのリンク、アフィリエイトによる収益増大を目論むブロガー、といった、トピックドリフトの原因となるリンクが非常にたくさん含まれていることが分かる。潜在的なウェブログコミュニティを見つけるためには、これらの雑多なリンク構造から、できるだけ意味のあるクラスタに分割されるようエッジを取り除く方法を考える必要がある。

2. 関連研究

これまでに様々なウェブコミュニティに関する研究が行われている。Kumar et al. (1999)は、ウェブコミュニティにはコアが含まれるという仮説を立て、ウェブ全体のスナップショットデータを調査した結果、10 万を超えるコミュニティのコアを発見した。コアとは参照ページ集合と被参照ページ集合からなる完全二部グラフである。これにより、完全二部グラフは、意味的にまとまったウェブコミュニティの特徴となっていることを示した。二部グラフの考えを利用した検索ランキングの手法として、Kleinberg (1998)による、オーソリティとハブの概念を用いた HITS(Hyperlink-Induced Topic Search)がある。Gibson et al. (1998)は、HITS で得ることができるコミュニティの性質について調査を行っている。

2.1. グラフ分割とコミュニティ抽出

Brandes (2001)、Newman (2001)は、グラフ内の全てのノードペア間の最短距離を計算量O(nm)で計算できるアルゴリズムを開発した。彼らは、隣接行列を用いたFreeman （1977）のアルゴリズムは、最短経路の発見にとって、必要以上の情報を処理している点を指摘した。これに対して、BrandesやNewmanのアルゴリズムは、漸進的な経路情報のみを記録・処理することにより、計算量削減に成功した。このアルゴリズムにもとづき、Girvan & Newman (2002)は、最短路ビットウィーンネス（Shortest Path Betweeness, SPB）を用いた、コミュニティ分割手法を開発した。この手法は、1 つのグラフを次々と 2 つのグラフに分割するトップダウンの分割方法である。これに対して、階層的クラスタリングは、ボトムアップな方法である。Girvan & Newman (2002)は、トップダウンな方法は、ボトムアップな方法と比べ、まとまったクラスタを得ることができることを指摘している。また、その他の分割手法として、グラフ理論における、最大フロー、最小カットを用いたもの（Flake, & et al., 2000)、スペクトル分解を用いた分割（Kannan, & et al., 2001）があるが、計算量の観点から大量データの分割にはには適さない。Newman (2003)は、スペクトル分解にもとづくアルゴリズムは最悪の場合、$O(n^4)$の計算量が必要であるため、SPBを用いた方法が優位であると主張している。

2.2. コミュニティ抽出における共参照情報の有効性

村田(2001)は、共参照情報を用いた Web コミュニティ発見手法を開発し、100hot.com に掲載されている URL 分類情報との一致度を用いて、その有効性を示した。原田、風間、佐藤(2001)は、参照共起情報を Web ディレクトリの自動拡張に利用する方法を開発し、Open Directory のデータを用いて評価を行った。原田、三木、石田(2004)は、Web サイトの引用関係と、書籍引用関係との比較から、それぞれの引用関係によって生じるクラスタ間の関係を議論した。豊田、吉田、喜連川(2005)は、共参照情報を利用したオーソリティ導出グラフにもとづき、ウェブコミュニティを抽出する方法を開発し、Yahoo Japan のウェブディレクトリとの比較を行った。Adar et al. (2004)は、共参照と公表時刻にもとづき、ウェブログスペースにおける情報拡散を推

測する手法と、その手法にもとづくランキングアルゴリズムを開発した。彼らはウェブログの場合、テキスト情報より、共参照情報の方が、有効な情報を得ることができると指摘している。

2.3. ウェブログ空間での問題の深刻化

Bharat & Henzinger (1998)は、二部グラフを用いたランキングアルゴリズムに関して、(1)ホスト間の相互強化関係(Mutually Reinforcing Relationships Between Hosts)、(2)自動生成されたリンク(Automatically Generated Links)、(3)適切でないノード(Non-relevant Nodes)、という問題点を指摘している。

ウェブログスペースにおいては、これら3つの問題が深刻化する。問題(1)については、同一人物が、複数のウェブログを持つことができるため、リンク情報を用いたアルゴリズムに影響を与えることができる。問題(2)については、ウェブログシステムが自動的に埋め込むポータルサイトへのリンクや、ニュース一覧、更新されたウェブログ一覧、アフィリエイトのバナー、といった、ユーザーが選択的に埋め込むことができるプラグインがある。問題(3)については、ウェブログが手軽に執筆することができるシステムであるため、その名のごとく日記として、テーマを定めず日々移り変わる時事についてコメントを書く傾向があり、意味のまとまりの無い、トピックドリフトを引き起こすリンクを張りやすい状態である。つまり、ウェブログスペースには、LBC を抽出するとき、不要なリンクやウェブページがたくさんある、というのが現状である。

2.4. ウェブログに対する従来の手法の問題点

LBC の抽出において、SPB を用いた手法はうまく働かない。SPB は、グラフをできるだけまとまりのある 2 つの部分グラフに分割する。これを繰り返すことにより、1 つのグラフを複数の部分グラフへと分割する。この手法は、実世界の社会的関係データに適用され、良好な分割結果が得られている。これは、例えば、研究論文の共著者関係などにおいて、研究者間のリンクは双方の了解がなければ張られることは無い（双方向リンク）。つまりグラフを構成するノードやエッジに無意味なものは無い。それに対して、ウェブでは片思いでリンクを張ることができる（一方向リンク）。そのため、グラフには無意味なエッジやノードが含まれる可能性がある。

SPB は、最短経路の情報のみを用いることにより、計算効率を高めることに成功した。しかし、それ以外のグラフの構造的特徴無視しているため、無関係なエッジやノードの存在を考慮したグラフの分割ができない。つまり、ウェブログのグラフ構造の分析には適していないため、ウェブログにおける潜在的なコミュニティを抽出するには、グラフの構造から得られる特徴を考慮した分割方法を新たに開発する必要がある。

3. 潜在的ウェブログコミュニティの抽出

潜在的ウェブログコミュニティ（Latent weBlog Community, LBC）を抽出するために、ウェブログと参照されるホームページにより構成される二部グラフにおいて、意味的まとまりがある部分二部グラフに分割する方法を提案する。

3.1. ウェブグラフの擬似二部グラフ化

LBC を抽出するために、PING サーバに更新情報を送るウェブログを *Fun* ノードとし、それに参照されるページを *Target* ノードとし、擬似的な二部グラフとして捉える。ウェブログの中には、他のウェブログから参照されているものもあるが、*Fun* ノードとなっているものは、*Target* ノードから除き、ノード集合全体を二部グラフ化する。このようにして作成した二部グラフ G、*Fun* ノード集合 *F*、*Target* ノード集合 *T*、*Fun* ノード *f*、*Target* ノード *t*、*F* から *T* へのリンクの集合 *E* の関係を以下に示す。

$$G = (F, T, E)$$

$$N = F \cup T,\ F \cap T = \phi,\ f \in F,\ t \in T$$

Fun ノードから *Target* ノードへの参照関係を示す、参照関係行列 R を以下に示す。

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n_t} \\ r_{21} & r_{22} & \cdots & r_{2n_t} \\ \cdots & \cdots & r_{ij} & \cdots \\ r_{n_f 1} & r_{n_f 2} & \cdots & r_{n_f n_t} \end{pmatrix},\ n_f = |F|,\ n_t = |T|$$

ここで、行数n_f、列数n_tはそれぞれ、*Fun*ノード集合のノード数、*Target*ノード集合のノード数である。*Fun*ノードiから*Target*ノードjへの参照間r_{ij}は、参照がある場合 1、無い場合は 0 である。

3.2. ウェブログにおけるリンクの特徴

実際のウェブログを観察すると、テーマを絞った興味深いものがある反面、日々目に付いたニュースや事柄についてテーマを選ばず自分流の突っ込みを入れるものも多々見られる。また、オンラインショップへ誘導しアフィリエイトで収益を上げるものや、1 人のブロガーが複数のウェブログで情報公開している場合がある。さらに、ブロガーの意図にかかわらず、各ウェブログシステムが自動的に埋め込むリンクが多々あり、自社ポータルや関連ポータル、ニュース、他のウェブログ、Amazon などの商品ページへのアフィリエイトなど、情報抽出を困難にするリンクがたくさんある。これらの雑多なリンク構造から、できるだけ意味のあるクラスタに分割されるようエッジを取り除く必要が

ある。

Fun ノードにおいては、アフィリエイトによるショップへのリンク集は、通常のブログと比べ、非常に多くのリンクを持つ傾向がある。また、*Target* ノードにおいては、Yahoo、Excite、Goo などのポータルは、非常にたくさんの被リンクを持つ。これらのページは、トピックドリフトを引き起こすリンクを含みやすい、あるいはそのリンクを張られやすい。

この状況を形式的に表現するために、*Fun*ノードにおける共参照行列**F**、*Target*ノードにおける被共参照行列**T**を、参照行列**R**とその転置行列$\mathbf{R}^{\mathbf{t}}$による、2 種類の積和行列で定義する。

$$\mathbf{F}=\mathbf{R}\times\mathbf{R}^{\mathbf{t}}=\begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n_f} \\ f_{21} & f_{22} & \cdots & f_{2n_f} \\ \cdots & \cdots & \cdots & \cdots \\ f_{n_f1} & f_{n_f2} & \cdots & f_{n_fn_f} \end{pmatrix}$$

$$\mathbf{T}=\mathbf{R}^{\mathbf{t}}\times\mathbf{R}=\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n_t} \\ t_{21} & t_{22} & \cdots & t_{2n_t} \\ \cdots & \cdots & \cdots & \cdots \\ t_{n_t1} & t_{n_t2} & \cdots & t_{n_tn_t} \end{pmatrix}$$

これら行列において、共参照（被共参照）が多いノードペアが分かるため、ペア間のテーマの一致度を知ることができる。しかし、アフィリエイトリンク集や、ポータルは、他のページとの間で、非常に高い共参照数、被共参照数を持つため、それらページに対応するf_{ij}、t_{ij}の値が大きくなるため、他のページとのテーマの一致度が高くなってしまうという問題がある。

3.3. 最弱関係にもとづく分割アルゴリズム

本節では、前節で議論した、ウェブログのリンクの特徴を考慮した、二部グラフ分割アルゴリズムを示す。このアルゴリズムは、*Fun* ノード集合、*Target* ノード集合、それぞれの中での、ノード間の関係性にもとづき、一番弱い関係のノードのペアを分離するように、エッジを取り除く、というものである。この方法を最弱対（the Weakest Pair, WP）アルゴリズムと呼ぶ。WP は以下の 6 つのステップで構成される。

1. *Fun*、*Target* 集合それぞれの、正規化した共参照行列を計算する。
2. 各集合においてノードペアの間の関係性の強さを計算する。
3. 各集合において最弱ペアを見つける。
4. 最弱ペア間の最短経路を見つける。
5. 全ての最短経路を通過するときの各エッジの通過回数を数える。
6. 最も通過回数の多いエッジを取り除く。

ノードペア間の関係性は、前節で定義した、*Fun* ノードにおける共参照行列 **F**、*Target* ノードにおける被共参照行列 **T** にもとづき定義する。これらの行列の値をそのまま類似度とすると、たくさんリンクを張っているウェブログ、あるいはたくさんリンクが張られているページが、他のウェブログ、あるいはページと非常に類似度が高い、ということになり、リンクスパムに対して脆弱になってしまう。

この問題を解消するには、この性質を逆に利用し、ノードの持つリンク数が多くなればなるほど、共参照している、あるいは被参照されている、他のノードとの関係性は薄くなると考えるとよい。この考え方にもとづき、共参照行列**F**、被共参照行列**T**の各行の和f_i^{ref}、t_i^{refed}が 1 となるよう正規化した、共参照関係行列$\mathbf{F}^{\mathbf{rel}}$、被共参照関係行列$\mathbf{T}^{\mathbf{rel}}$を定義する。ここで、$f_{ij}$、$t_{ij}$はそれぞれ、共参照割合、被共参照割合となる。これらの行列は、第 1 ステップで求められる。

$$\mathbf{F}^{\mathbf{rel}}=\begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n_f} \\ f_{21} & f_{22} & \cdots & f_{2n_f} \\ \cdots & \cdots & \cdots & \cdots \\ f_{n_f1} & f_{n_f2} & \cdots & f_{n_fn_f} \end{pmatrix},\quad f_i^{ref}=\sum_{j=1}^{n_f} f_{ij}=1$$

$$\mathbf{T}^{\mathbf{rel}}=\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n_t} \\ t_{21} & t_{22} & \cdots & t_{2n_t} \\ \cdots & \cdots & \cdots & \cdots \\ t_{n_t1} & t_{n_t2} & \cdots & t_{n_tn_t} \end{pmatrix},\quad t_i^{refed}=\sum_{j=1}^{n_f} f_{ij}=1$$

これら共参照関係行列$\mathbf{F}^{\mathbf{rel}}$、被共参照関係行列$\mathbf{T}^{\mathbf{rel}}$の行列要素を用いて、2 つのノードペア(i,j)の関連度*FS(i,j)*、*TS(i,j)*を以下のようにそれぞれ定義する。

$$FS(i,j)=f_{ij}+f_{ji},\ TS(i,j)=t_{ij}+t_{ji}$$

これらは、リンクスパムのウェブログやポータルのようなホームページと、その他のウェブログやホームページとの関連性が低くなる定式化となっている。例えば、リンクスパムや、アフィリエイト収益を狙った多数のリンクを含むウェブログは、他のウェブログとの共参照割合が低下する。他方、ポータルサイトのようなたくさんの被リンクがあるホームページは、他のホームページとの被共参照割合が低下する。ウェブログペア、ホームページペアの関係度は、双方の共参照割合、被共参照割合の和でそれぞれ定義されているため、リンクスパムとその他のウェブログとの関連度*FS(i,j)*、ポータルとその他のホームページとの関連度

*TS(i,j)*はそれぞれ低い値にとどまる。第 2 ステップにおいてこの関連度が、*Fun* 集合、*Target* 集合それぞれにおける、ノードペアについて計算される。

第 3 ステップでは、この類似度が 0 以上で一番小さいノードペアを見つける。これは関連度 0 のノードペアを除いた、関連度での昇順ソートで見つけることができる。

第 4 ステップでは、*Fun* ノード集合、*Target* ノード集合それぞれにおける最弱関係ノードペアにおいて、ペアの間の最短路を見つける。第 5 ステップで、各エッジの通過回数を数え上げ、最終の第 6 ステップで、最も通過回数の多いエッジを取り除く。

第 4 ステップ調べる最短路は常に距離 2 であるので、*Fun*側から*Target*側、あるいはその逆に、1 往復するだけで見つけることができる。これは、共参照関係行列 $\mathbf{F}^{rel}$、被共参照関係行列$\mathbf{T}^{rel}$を、Rの積和行列にもとづいて定義しているため､関連度があるノードのペアは、必ず共参照､あるいは被共参照されているからである。

3.4. 分割例

本節では、前節で示した WP の処理を具体的に示すために、ここでは、*Fun* ノード 6 つ、*Target* ノード 6 つで構成される、簡単な二部グラフの例を用いて各処理過程を説明する（図１）。このグラフは、完全二部グラフを構成する部分グラフ G1（ノード f1,f2,t1,t2)、G2（ノード f3,f4,t3,t4)、G3（ノード f5,f6,t5,t6）において、2 つのエッジ e1(f2-t3)、e2(f5-t4)を加え、1 つのグラフ G を構成したものと見ることができる。加えられたエッジは、ブロガーがこれまで参照しているページと若干テーマの異なるホームページにリンクを張ってみた、といった状況を表している。

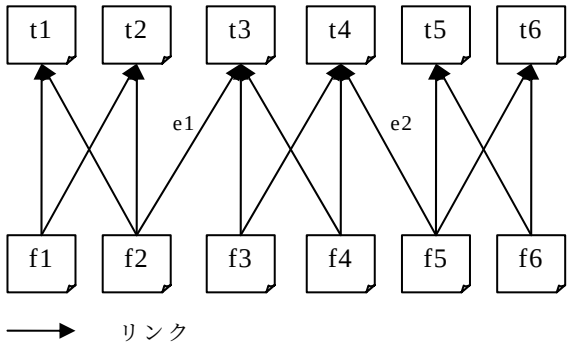

図 1　二部グラフの例

この二部グラフの参照関係行列、共参照行列、被共参照行列はそれぞれ、図 2、3、4 のようになる。これら行列要素の値にもとづき、*Fun* ノードペア、*Target* ノードペア、それぞれの関連度を計算した結果を、表 1、2 にそれぞれ示す。これら表によると、*Fun* ノード集合における､最弱対は､(f2,f3)､(f2,f4)､(f3,f5)､(f4,f5)､*Target* ノード集合における､最弱対は､(t1,t3)､(t2,t3)､(t4,t5)､(t4,t6)､である。

Fun ノード集合、*Target* ノード集合それぞれの最弱関係対の最短パスと、各エッジの通過回数を、図 5 に示す。また、エッジ通過回数を表 2 にまとめる。

表 2 によれば、最弱関係ペアを異なるサブグラフへと分けるためには、エッジ f2t3 と、f5t4 を取り除けばよいことが分かる。このように、本アルゴリズムは、ブロガーが気まぐれで張ったリンクを取り除き、分割されたグラフが、できるだけ完全二部グラフに近い形になるよう、グラフを分割する性質があることが分かる。

$$\mathbf{R}=\begin{pmatrix}1&1&0&0&0&0\\1&1&1&0&0&0\\0&0&1&1&0&0\\0&0&1&1&0&0\\0&0&0&1&1&1\\0&0&0&0&1&1\end{pmatrix}$$

図 2　参照関係行列

$$\mathbf{F}=\begin{pmatrix}2&2&0&0&0&0\\2&3&1&1&0&0\\0&1&2&2&1&0\\0&1&2&2&1&0\\0&0&1&1&3&2\\0&0&0&0&2&2\end{pmatrix},\ \mathbf{F}^{rel}=\begin{pmatrix}0.5&0.5&0&0&0&0\\0.29&0.43&0.14&0.14&0&0\\0&0.17&0.33&0.33&0.17&0\\0&0.17&0.33&0.33&0.17&0\\0&0&0.14&0.14&0.43&0.29\\0&0&0&0&0.5&0.5\end{pmatrix}$$

(a)共参照数　　**(b)共参照割合**

図 3　共参照行列

$$\mathbf{T}=\begin{pmatrix}2&2&1&0&0&0\\2&2&1&0&0&0\\1&1&3&2&0&0\\0&0&2&3&1&1\\0&0&0&1&2&2\\0&0&0&1&2&2\end{pmatrix},\ \mathbf{T}^{rel}=\begin{pmatrix}0.4&0.4&0.2&0&0&0\\0.4&0.4&0.2&0&0&0\\0.14&0.14&0.43&0.29&0&0\\0&0&0.29&0.43&0.14&0.14\\0&0&0&0.2&0.4&0.4\\0&0&0&0.2&0.4&0.4\end{pmatrix}$$

(a)被共参照数　　**(b)被共参照割合**

図 4　被共参照行列

表 1　ノードペアの関連度

(a)*Fun* ノードペア

i	j	FS
f2	f3	0.31
f2	f4	0.31
f3	f5	0.31
f4	f5	0.31
f3	f4	0.67
f1	f2	0.76
f5	f6	0.76

(b)*Target* ノードペア

i	j	TS
t1	t3	0.34
t2	t3	0.34
t4	t5	0.34
t4	t6	0.34
t3	t4	0.57
t1	t2	0.8
t5	t6	0.8

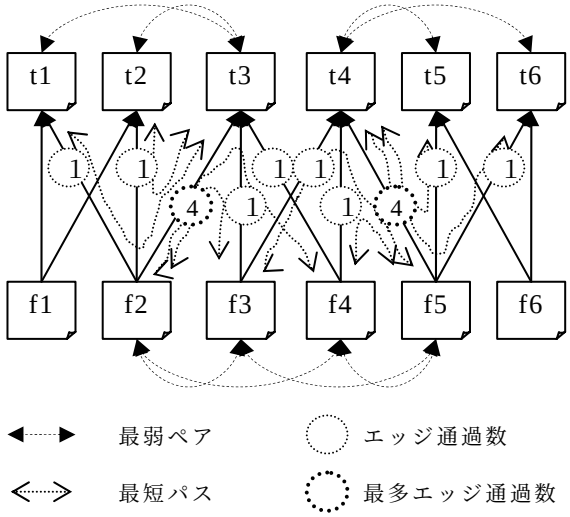

図 5　最弱ペアの最短パス

表 2　エッジ通過数

Edge	Freq
f2t3	4
f5t4	4
f2t1	1
f2t2	1
f3t3	1
f3t4	1
f4t3	1
f4t4	1
f5t5	1
f5t6	1

4. 評価

本研究で開発した WP と従来の SPB を比較する。この比較のため、コンピュータで二部グラフを実験用データとして生成し、これをそれぞれのアルゴリズムで分割する。2 つのアルゴリズムの評価のため、情報の損失、意味のまとまりの無さ（二部グラフの不完全性）という観点から、2 つの評価指標を定義し、これら評価指標にもとづき、評価を行う。

4.1. 評価指標

情報の損失、意味のまとまりの観点から、ここでは、分割によって生じたクラスタの数を基準とし、(1)孤立ノードの数、および、(2)二部グラフの不完全度、を用いる。1 つ目の評価指標について、孤立ノードとは、分割により参照が無くなった *Fun* ノード、または、被参照が無くなった *Target* ノードである。分割によって生じる孤立ノードの数が少なければ少ないほど、情報の損失が少ないと考えられるため、この値が小さければ小さいほど良いアルゴリズムであると考える。また、2 つ目の評価指標は、完全二部グラフは、意味的まとまりとなっているという観察結果にもとづいている。つまり不完全度が低く、できるだけ完全二部グラフの状態に近ければ近いほど、意味的にまとまっていると考える。不完全度を定義するために、*Fun* ノードの参照ベクトル fi、fj を考える。

$$f_i = \left(r_{i1}, r_{i2}, ..., r_{in_t}\right),\ f_j = \left(r_{j1}, r_{j2}, ..., r_{jn_t}\right)$$

各*Fun*ノードペアにおいて、ハミング距離D_{ij}^{Hf}を計算し、それを 0 から 1 の間で正規化したものを、1 つの二部グラフkにおいて計算したものを、二部グラフkの不完全度$IBPR_k$と定義する（範囲は[0,1]）。

$$D_{ij}^{Hf} = \sum_{k=1}^{n_t} \left| r_{ik} - r_{jk} \right|,\ \ IBPR_k = \frac{1}{\left(n_t\right)^2} \sum_{i,j \in F}^{n_t} D_{ij}^{Hf}$$

この各二部グラフの不完全度の平均を、全データに対する不完全度 IBPR と定義する（K は全二部グラフ数、IBPR の範囲は[0,1]）。

$$IBPR = \frac{1}{K} \sum_{k=1}^{K} IBPR_k$$

この不完全度 IBPR が低ければ低いほど、元の全データが、意味的にまとまったクラスタに分割されていると考える。この指標は *Target* 集合においても同様に定義する。

4.2. 実験条件

2 つのアルゴリズムを比較するための試験用データを生成する。この試験用データは、グラフにおけるノード数、エッジ密度を変化させ、作成する。ノード数は、*Fun* ノード、*Target* ノードの合計ノード数である。この実験では分割アルゴリズムの基本的性質を調べるため、双方のノード数が等しい場合を調べる。エッジ濃度は、完全二部グラフのエッジ数を 100%として定義する。用いるノード数は、20、40、60、エッジ密度は、10%刻みで、10%～90%である。リンク構造は、様々な組み合わせがあるため、同じノード数、エッジ密度の二部グラフをランダムに複数生成し、これらの処理結果の平均を取る。この実験ではこのグラフ生成数を 10 とする。実験条件を表 3 にまとめる。

表 3　実験条件

パラメータ	値
ノード数	20,40,60
エッジ密度	10%～90%（10%刻み）
同一条件でのサンプル数	10

4.3. 実験結果

実験結果を図 6～17 に示す。水平軸は分割アルゴリズムを繰り返し適用した後の二部グラフの数を示している。繰り返しにおいては、二部グラフができるだけ完全形に近づくよう、$IBPR_k$が一番大きい二部グラフを選び、分割を適用した。図 6,7,10,11,15 の垂直軸は、平均孤立ノード数である。図 8,9,12,13,16,17 の垂直軸は、平均IBPRである。これらのグラフにおいて、値の軌跡が低い方が、他方よりも有効性が高いということ

を示している。これらの図において、WPはブルーの軌跡、SPBはピンクの軌跡で示されている。

ノード数が 20、エッジ密度が 40%の場合、WP は SPB より有効性が低い(図 6,8)。それに対して、ノード数が 20、エッジ密度が 50%の場合、WP が SPB より有効となる(図 7,9)。ノード数が 40、エッジ密度が 30%の場合、平均 IBPR の観点から、WP は SPB より有効性が低い(図 12)。それに対して、ノード数が 40、エッジ密度が 40%の場合、WP が SPB より有効となる(図 11,13)。ノード数が 60、エッジ密度が 20%の場合、WP は SPB より有効性が低い(図 14,16)。それに対して、ノード数が 60、エッジ密度が 30%の場合、WP が SPB より有効となる(図 15,17)。

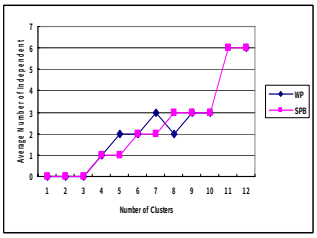

図 6：平均孤立ノード数
（20 ノード、エッジ密度 40%）

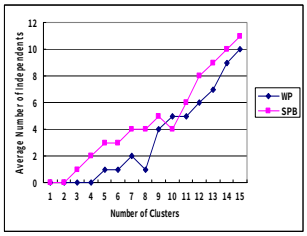

図 7：平均孤立ノード数
（20 ノード、エッジ密度 50%）

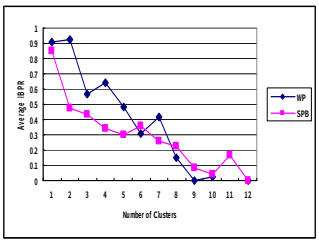

図 8：平均 IBPR
（20 ノード、エッジ密度 40%）

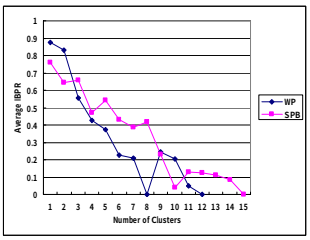

図 9：平均 IBPR
（20 ノード、エッジ密度 50%）

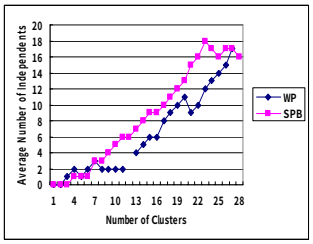

図 10：平均孤立ノード数
（40 ノード、エッジ密度 30%）

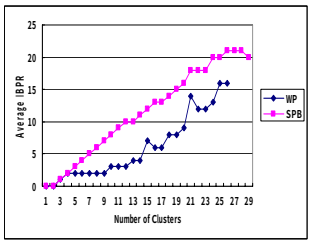

図 11：平均孤立ノード数
（40 ノード、エッジ密度 40%）

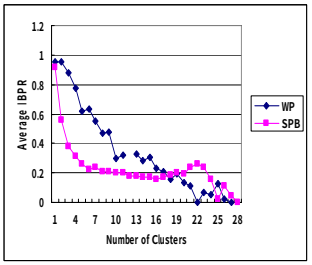

図 12：平均 IBPR
（40 ノード、エッジ密度 30%）

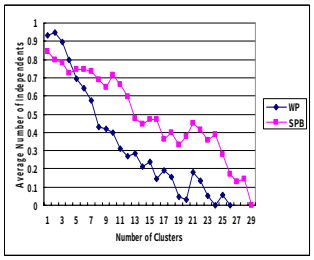

図 13：平均 IBPR
（40 ノード、エッジ密度 40%）

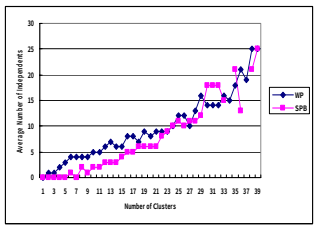

図 14：平均孤立ノード数
（60 ノード、エッジ密度 20%）

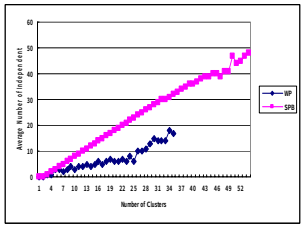

図 15：平均孤立ノード数
（60 ノード、エッジ密度 30%）

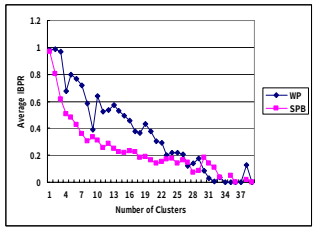

図 16：平均 IBPR
（60 ノード、エッジ密度 20%）

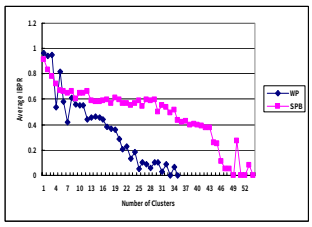

図 17：平均 IBPR
（60 ノード、エッジ密度 30%）

この結果をまとめると、エッジ密度が低い場合は SPB、高い場合には WP が、それぞれ良好な分割となる。また、ノード数が多くなればなるほど、2 つのアルゴリズムの有効性が入れ替わる境界条件のパーセンテージが下がる。つまり、グラフのノード数が多くなればなるほど、WP アルゴリズムが有効な範囲が拡大する。3 種類のノード数それぞれにおける WP と SPB との大まかな境界条件を、図 18 に示す。

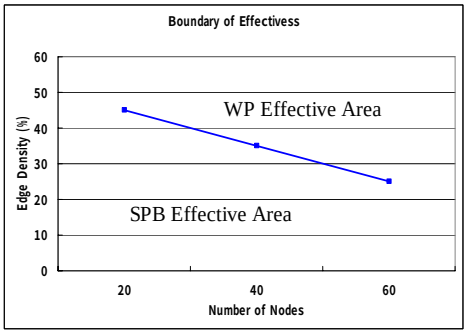

図 18：WP と SPB の間の境界条件

この図によれば、ノード数が増大すればするほど、WP が有効となる領域が増大すると考えられる。つまり、膨大なノード数を含んでいるウェブログ空間においては、WP による分割が、LBC の抽出に有効であると考えられる。

4.4. 考察

実験の結果、ノード数が少なく、エッジ密度が低い場合を除いて、SPB と比べ、WP が有効であることが分かった。ここで、その条件で、WP が不利になる理由を考察する。エッジ密度が低く、1 ノードあたりのエッジ密度が低い場合、二部グラフは疎になり、グラフにおけるノード間の最短路は長くなり、代替的な経路数は少なくなる（図 19)。このような場合、トピッ

クドリフトの問題が発生し、図 19 における f1 と f6 の扱うトピックは非常に異なるものと想像できる。WP は両端以外のエッジを均等に切るので、細長いグラフを細切れに粉砕し、孤立ノードが生じやすい。図 19 の例では、5 つのエッジ(f3t2,f3t3,f4t3,f4t4,f5t4)が取り除かれ、4 つのノード（f3, f4, t3, t4）が孤立ノードとなり、両端に 2 つの部分二部グラフ G1(f1, f2, t1, t2)および G2(f5, f6, t5, t6)が残る。他方、SPB はグラフの真ん中のエッジを切るため、細長いグラフの分割において、孤立ノードが生じにくい。このため、ノード数が少なく、エッジ密度が低い場合、SPB が有利になると考えられる。

実際のウェブログでは、ブロガーの嗜好があるため、人工的に生成したランダムな二部グラフと比べ、グラフ内で密な部分と疎な部分が生じていると考えられる。このような場合、3.4 の図 5 の例で示したように、気まぐれで張られた疎な部分のリンクが取り除かれる。そのため、実際のデータに対して、人工的データを用いた場合と比べ、SPB に対する WP の優位性が高まるものと考えられる。

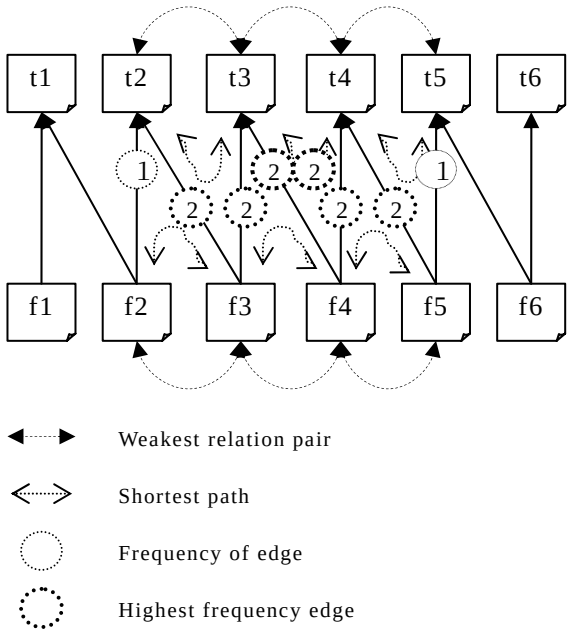

図 19：直径の長い二部グラフ

計算量については、WPは共参照行列を得るために、行列積の計算が必要であり、この計算量がアルゴリズムの効率に影響を与える。SPBの計算量O(nm)（nはノード数、mはエッジ数）であるが、グラフの全てのノードが直接連結している場合は、エッジ数がn(n-1)本となるため、$O(n^3)$である。二部グラフの場合は、$O((f+t)ft)$、つまり$O(f^2t)$あるいは$O(ft^2)$である（fは参照側ノード数、tは被参照側ノード数、n=f+t)。WPにおける行列積の計算量は、基本的に$O(n^3)$、二部グラフに限定すると$O(ft^2)$あるいは$O(f^2t)$であるが、参照行列のデータ構造を、スパース行列の表現に用いられるHB（Harwell Boeing）形式(Duff, & et al., 1992)で実装すれば、リンクの無い部分の処理を省くことができるので、エッジの密度が低い場合はO(fte)(eは参照ノード１つあたりの平均参照数、あるいは被参照ノード１つあたりの平均被参照数）となり、計算量は軽減でき、SPBに大きく劣ることはないと考えられる。

5. 潜在的ウェブログコミュニティの抽出

本論文では、潜在的ウェブログコミュニティ(Latent weBlog Community, LBC)を抽出するデータとして、更新されたウェブログと、そのウェブログが参照するホームページとのリンク関係を用いる。更新されたウェブログは、PING サーバやウェブログサイトの更新情報 RSS により知ることができる。これにもとづき、更新されたウェブログの URL と、それが参照するページの URL を抽出する。

このウェブログと参照されるホームページとの間のリンク数は膨大であるため、以下の条件で抽出に用いるデータを選別した。1 つ目は、データを 1 週間毎に区切り、日曜日から土曜日までを１つのデータ群としデータを分割した。2 つ目は、参照されるホームページを、企業、行政に制限した。3 つ目は、更新されたウェブログは、他の更新されたウェブログに参照されていても、被参照ページに含めないこととした。

利用データには複数の二部グラフが含まれる。分割する二部グラフの選択と、分割を終了する条件が必要である。ここでは、次の分割戦略を用いた。二部グラフを構成するページ数が 100 を超えるものがある場合は、その内の一番大きなものを分割する。また、ページ数が 100 を超える二部グラフが無くなった場合には、不完全性が 0.9 を超えるものの中で、一番大きなものを分割する。不完全性が 0.9 を超える二部グラフが無くなったとき、分割を終了する。不完全性の閾値 0.9 は高い値ではあるが、1 週間単位でまとめたデータはエッジの密度が低く、完全性の高い、LBC が少ないと考え、この値とした。この分割戦略にもとづき利用データを分割したクラスタのいくつかを例示する。

表 4: 3 週間データの統計

データ番号	1	2	3
期間	1 月 9 日～15 日	1 月 16 日～22 日	1 月 23 日～29 日
ウェブログの数	4725	5426	5415
ウェブページの数	3926	4635	4832
リンク数	8665	9932	11127
初期の二部グラフ数	1791	2043	2156

5.1. 例1： 鉄道

関西から北陸までの鉄道会社に関するLBCである[1]。ウェブログf1～f4 は全体的にコメントやトラックバッ

[1] http://webcom.dr-k.info/show_sec_grp.php?sid=103&gid=1471

クがある。f1 は温泉、お寺参りをテーマにした旅行に興味がある。f3 は北陸の鉄道をテーマとしており、金沢駅ｔ５を参照している。t1～t8 は関西から北陸までの鉄道関連のページであるため、鉄道好きのブロガーが、各自のホームタウン以外へ旅行する場合に役立つと考えられる。

表 5　鉄道関連 LBC の URL

f1	http://www.doblog.com/weblog/myblog/2974/
f2	http://plaza.rakuten.co.jp/dongurimama/
f3	http://kissy21.seesaa.net/
f4	http://blog.livedoor.jp/yuikugayama/
t1	http://www.kintetsu.co.jp/
t2	http://www.jr-central.co.jp/
t3	http://www.hankyu.co.jp/rail/index.shtml
t4	http://www.jreast.co.jp/top.html
t5	http://www.westjr.co.jp/branch/kanazawa/
t6	http://www.keihan.co.jp/
t7	http://www.central.co.jp/club/mizonokuchi.html
t8	http://www.westjr.co.jp/index.html

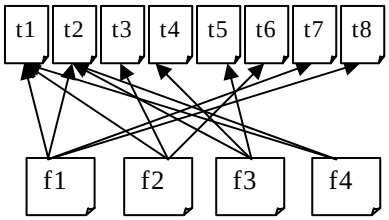

図 20: 鉄道関連 LBC

5.2. 例 2：美術ギャラリー

美術ギャラリーに関するLBCである[2]。t2 はDIYショップのページで、f1、f4 により共参照されている。f1 はDIYショップでの芸術的な木片の購入を記述していた。それに対し、t2 のみ参照するf4 は女性のバイクライダーのウェブログで、14 得ナイフの購入を記述しており、意味的につながりの解釈が困難なトピックドリフトが残っている。

表 6：　美術ギャラリー関連 LBC の URL

f1	http://d.hatena.ne.jp/simplife/
f2	http://nohibi.nonnnonn.net/
f3	http://blog.livedoor.jp/itakeaway/
f4	http://blog.livedoor.jp/kirakira_vtr/
f5	http://tout.jugem.cc/
t1	http://www.dnp.co.jp/gallery/ggg/index.html
t2	http://www.toho.co.jp/diy/diy-top.html
t3	http://www.shiseido.co.jp/house-of-shiseido/html/exhibition.htm
t4	http://www.dnp.co.jp/gallery/ggg/gki/g224/g224ki.html
t5	http://www.matsuya.com/ginza/design/0124e_suetomi/index.html

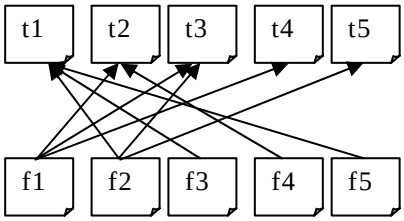

図 21：美術ギャラリー関連 LBC

[2] http://webcom.dr-k.info/show_sec_grp.php?sid=102&gid=936

5.3. 例 3：食品への不純物混入

食品への不純物混入に関するLBCである[3]。ウェブログf1～f4 は全てこのトピックに関する情報提供を行うものである。このクラスタはこのトピックに関する情報を得るのに役立つと考えられるが、情報提供を目的としているため、他のブロガーの参加無しでは、コミュニケーションの活発な、ウェブログコミュニティに変化するのは難しいかもしれない。

表 7：食品への不純物混入関連 LBC の URL

f1	http://blog.gozonji.com/
f2	http://blog.goo.ne.jp/yujikatu/
f3	http://blog.livedoor.jp/sogo_security/
f4	http://blog.livedoor.jp/kishinamin/
t1	http://www.olc.co.jp/news/20050119_01.html
t2	http://www.mhlw.go.jp/houdou/2002/07/h0719-3.html
t3	http://www.kagome.co.jp/news/2004/050118-info.html
t4	http://www.maff.go.jp/www/press/cont2/20050118press_6.htm
t5	http://www.mhlw.go.jp/kinkyu/diet.html
t6	http://www.mhlw.go.jp/kinkyu/diet/index.html

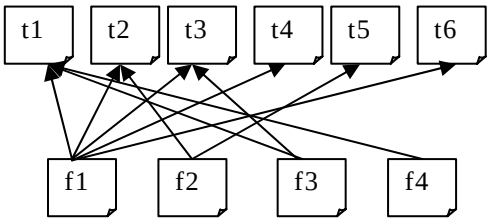

図 22:食品への不純物混入関連 LBC

5.4. ペルソナ検出

ペルソナ検出とは、同一人物に所有されている複数のウェブログを検出することである。例えば、健康食品のオンラインショップ[4]の商品への共参照を含むウェブログのクラスタが抽出された[5]。これらのウェブログはアクセス数を増やすために、オンラインショップの関係者が記述しているウェブログの可能性がある。このクラスタをさらに分割すると、完全二部グラフを構成する 2 つのウェブログが抽出できる。

また、オンラインモールへのアフィリエイトページ[6]へのリンク集ウェブログのクラスタが抽出された[7]。このクラスタは、分割の結果、完全二部グラフとなっていた。これらのウェブログは、アフィリエイトによる収益を狙ったページである[8]。これらのクラスタは、検索エンジンの検索結果に大きな影響を及ぼすかもしれない。ペルソナ検出は、その影響を排除するための仕組みとして用いることができる。

5.5. 考察

いくつかの分割されたクラスタを紹介した。全クラ

[3] http://webcom.dr-k.info/show_sec_grp.php?sid=102&gid=695
[4] http://www.kenko.com/
[5] http://webcom.dr-k.info/show_sec_grp.php?sid=102&gid=387
[6] http://pt.afl.rakuten.co.jp
[7] http://webcom.dr-k.info/show_sec_grp.php?sid=102&gid=86
[8] 例えば、http://s01.agentsonic.jp/camera/

スタを観察した結果、意味のまとまりが高いクラスタが多く観察されたが、いくつかは複数のトピックが分割されていないトピックドリフトを含むクラスタもあった。これは分割戦略の問題であり、さらに分割を繰り返すと分けることができる。

さほど遠くないトピックドリフトは話題の広がりとして考えることができる。例えば、「電車はお寺や温泉めぐりのための交通機関」、「ギャラリーに興味があるブロガーは、DIY ショップでは芸術的な木片を購入する」といった話題の広がりである。

ただ、ギャラリーのクラスタでは、DIY ショップのページにより、テーマに合わない、女性バイクライダーのウェブログがクラスタに含まれていた。このクラスタをもう一度分割すると、DIY ショップと女性バイクライダーのサブグラフが LBC から分離される。

コメントやトラックバックの多いウェブログとそうでないものとの間で差がある。コメントやトラックバックの多いものは、テーマを絞っている傾向あり、少ないものは、どんなテーマでも取り上げている。同じテーマで同じ LBC に分類された、コメントやトラックバックの少ないブロガーは、コメントやトラックバックの多いブロガーのテーマの選び方や書き方を参考にする、また、そのブログにトラックバックするなどして、自分のウェブログの質や注目度を改善できるかもしれない。

6. まとめ

インターネットにおける自律的な情報組織化を促進するための概念として、潜在的ウェブログコミュニティ（Latent weBlog Community, LBC）を提案した。また、これを抽出するアルゴリズムとして、共参照、被共参照情報にもとづき、最弱対（the Weakest Pair, WP）アルゴリズムを開発した。

この評価のため、従来の最短路ビットウィーンネス（the Shortest Path Betweeness, SPB）を用いた分割手法との比較を行い、WP の有効性を示した。これを用いた 3 つの LBC の抽出例と、二次的効果、同一ブロガーによる複数ウェブログの抽出というペルソナ検出の例を示した。

LBC は類似の嗜好を持つ相互に知り合いではないブロガーたちの出会いの場である。今後、この出会いの場を実装し、ブロガーたちのコミュニケーションによる、自律的な情報組織化の促進を計画している。

謝辞

有益な助言と計算資源をお貸しいただいた太田敏澄先生、山本仁志先生、Ms. Bonnie Huie に感謝いたします。

参考文献

[1] Adar, E., Zhang, L., Adamic, L. A. and Lukose, R. M.: Implicit Structure and the Dynamics of Blogspace, in Glance, N., Adar, E., Hurst, M. and Adamic, L. eds., WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2004).

[2] Bharat, K. and Henzinger, M. R., “Improved Algorithms for Topic Distillation in a Hyperlinked Environment,” the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.

[3] Brandes, U., “A Faster Algorithm for Betweeness Centrality,” Journal of Mathematical Sociology, Vol. 25, No. 2, pp. 163 – 177, 2001.

[4] Dhillon, I. S., “Co-clustering documents and words using Bipartite Spectral Graph Partitioning,” ACM SIGKDD 2001, San Francisco, California, USA., 2001.

[5] Duff, I. S., Grimes, R. G., and Lewis, J. G. Users’ guide for the Harwell-Boeing sparse matrix collection (Release I). Technical Report tr/pa/92/86, CERFACS, 1992.

[6] Flake, G. W., Lawrence, S., and Giles, C. L., “Efficient Identification of Web Communities,” in Proc. KDD 2000, 2000.

[7] Girvan, M. and Newman, M. E. J., 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826.

[8] Gibson, D., Kleinberg, J., and Raghaven, P., “Inferring Web Communities from Link Topology,” in Proc. HyperText98, 1998.

[9] 原田早恵子、三木武、石田亨、「コミュニティマイニングにおける Web 引用解析と文献引用解析との比較」、電子情報通信学会誌 D-I, Vol. J87-D-I, No. 3, pp. 382 - 389, 2004 年 3 月

[10] 原田昌紀、風間一洋、佐藤進也、「参照共起分析の Web ディレクトリへの適用」, 研究報告 「情報学基礎」 アブストラクト No.061 - 007, 2001.

[11] Kannan, R., Vempala, S., and Vetta, A., “On Clusterings: Good, Bad and Spectral,”July, 2001.

[12] Kleinberg, J. K., “Authoritative Sources in a Hyperlinked Environment,” in Proc. ACM-SIAM Symposium on Discrete Algorithm, 1998.

[13] Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A., “Trawling the web for emergin cyber-communities,” Proceedings of the 8th WWW conference, 1999.

[14] 村田剛志、「参照共起にもとづく Web コミュニティの発見」、人工知能学会誌、Vol. 16, No. 3B, 2001 年.

[15] Newman, M. E. J., 2001. Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. Phys. Rev. E 64, 016132.

[16] Newman, M. E. J., “Fast algorithm for detecting community structure in networks,” 2003.

[17] Newman, M. E. J. and Girvan, M., “Finding and evaluating community structure in networks,” 2003.

[18] Seary, A. J. and William, D. R., “Partitioning Networks by Eigenvectors,” Proceedings of the International Conference on Social Networks, Vol. 1: methodology, pp. 47 – 58.

[19] 豊田政史、吉田聡、喜連川優、「ウェブコミュニティチャート：膨大なウェブページを関連する話題を通して閲覧可能にするツール」、電子情報通信学会誌、Vol. J87-D1, No. 2, Feb. 2005.