Webコンテンツの偏在性に着目したP2Pコンテンツ流通システム

高橋　昭裕† 　小林　亜樹†† 　山岡　克式† 　横田　治夫† 　曽根原　登†††

† 東京工業大学　〒 152–8552 東京都目黒区大岡山 2–12–1
†† (独) メディア教育開発センター　〒 261–0014 千葉県千葉市美浜区若葉 2–12
††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †akihiro@de.cs.titech.ac.jp, ††aki@nime.ac.jp, †yamaoka@ss.titech.ac.jp, †yokota@cs.titech.ac.jp, †††sonehara@nii.ac.jp

あらまし　P2P コンテンツ流通システムにおいて，コンテンツ，クエリの各ピアに投入される意味的なカテゴリのうち，最も優勢なカテゴリの比率を各ピアの優勢度とすると，優勢度は大きいほど効率がよく，また，同じ優勢カテゴリを持つピア同士を近くに配置すると効率が良いことを示す．一方，Web コミュニティの研究に見るように，Web サイトには優勢カテゴリが明確に存在し，同じ優勢カテゴリを持つサイト間はハイパーリンクで密に繋がっている．このことを利用して， Web リンク情報を用いてピアの優勢度を大きくし，同一優勢カテゴリを持つピア同士を近くに配置する手法を提案する．Web リンク情報を取得するため，協調型 Web アーキテクチャ上のシステムとして試作する．
キーワード　P2P, Web とインターネット, オーバレイネットワーク

A System of P2P Contents Distribution Taking Advantage of Uneven Web Contents Distribution

Akihiro TAKAHASHI†, Aki KOBAYASHI††, Katunori YAMAOKA†, Haruo YOKOTA†, and Noboru SONEHARA†††

† Tokyo Institute of Technology　Ookayama 2–12–1, Meguro-ku, Tokyo, 152–8552 Japan
†† National Institute of Multimedia Education　Wakaba 4–5–6, Mihama-ku, Chiba-shi, Chiba 261–0014 Japan
††† National Institute of Informatics　Hitotsubashi 2–1–2, Chiyoda-ku, Tokyo, 101–8430 Japan
E-mail: †akihiro@de.cs.titech.ac.jp, ††aki@nime.ac.jp, †yamaoka@ss.titech.ac.jp, †yokota@cs.titech.ac.jp, †††sonehara@nii.ac.jp

Abstract　In a peer-to-peer content distribution system, when the contents and queries classified by the semantic categories are submitted to a peer, the dominant rate is defined as the largest rate of the category distribution of each peer. We show that the higher the dominant rate, the more efficient distribution can be achieved in the system. It is also showed that the efficient distribution is achieved when the peers which have the same dominant category were neer one another. There is a dominant category in each web site. As many researches of web community, the web sites which have a same dominant category were strongly connected by hyperlink. Our proposed system takes advantage of such uneven distribution of web contents. We propose the peer-to-peer content distribution system with the way to enlarging dominant rate of each peer and to put the peers in the neighborhood that have a same dominant category. A prototype system is built on the cooperative Web architecture.
Key words　P2P, Web and the Internet, Overlay Network

1. はじめに

ピュア型 P2P コンテンツ配送システムにおける検索用ネットワークトポロジ構築手法は，構築方針によって大きく二つに分類することが出来る．

まず一つ目は，実トラヒック抑制を企図して，検索用ネットワークトポロジを物理トポロジに近づけるような検索用ネットワークトポロジ構築手法 [1] [2] [3] [4] [5] [6] [7] である．具体的には，ピア間の遅延や帯域幅を計測，または推定し，その値から実トポロジにおいて近くにあると考えられるピア同士を接続する．これらの手法では，検索効率を維持したまま，実トラヒックを抑制することが出来る．特に分散ハッシュテーブル

(DHT:DistributedHashTable) を用いた検索手法 [8] [9] [10] を利用するシステムにおいては，これらの手法が用いられる．

二つ目は，検索効率の向上を企図して，同種のコンテンツ，クエリを，検索用ネットワーク上において集中させるような検索用ネットワークトポロジ構築手法 [11] [12] [13] [14] [15] [16] [17] [18] である．これらの手法では，Flooding など，検索開始ピアの，検索用ネットワークにおける近くの情報を利用する検索アルゴリズムを用いたシステムにおいて，検索成功率の向上や 1 クエリあたりのホップ数の低減が可能とされる．

その前提として，ピアに投入されるコンテンツ，クエリの内容カテゴリに偏りがあることと，この偏在を利用したトポロジ構成によって効率が改善されることが挙げられるが，これらの点の検証は不十分である．

ところで，コンテンツの内容的なカテゴリが偏在することを示す事例として，Web コミュニティに関する研究がある．Web グラフにおいて密に結合された Web ページ群を抽出することで，互いに関連するページの集合を得る手法として，WWW の部分グラフから密に結合された Hub と Authority を抽出する提案 [19] がなされている．これを応用，発展させた類似ページを探す手法 [20] [21] のほか，Web グラフ中の完全二部グラフ構造に着目した手法 [22] [23] や，これらの手法を意味的なまとまりを持った Web ページ群である Web サイトに対して適用することで，Web コミュニティを抽出する研究 [24] [25] もなされており，互いに関連する Web サイトは Web グラフにおいて近くに存在することが示されている．これは，Web におけるグラフ構造を形成する鍵となるハイパーリンクが，制作者主導で利用者の誘導を促す仕組みであることから，グラフ空間上に自然に形成されるものと考えられる．

そこで本論文では，まず，コンテンツ，クエリが属する意味的なカテゴリの概念を導入し，各ピアに同カテゴリのコンテンツ，クエリの集中する度合，また，同カテゴリのピア同士の接続集中度が，検索効率に与える影響について検証する．次に，検索成功率の向上，1 クエリあたりの転送回数の期待値の低減を企図して，Web リンク情報を反映させることで，各ピアに特定カテゴリのコンテンツ，クエリが集中する利用方式および，同カテゴリのコンテンツ，クエリが集まるのピア同士を近くに配置する検索用ネットワーク構築手法を提案する．最後に，提案手法を実現する試作システムを Perl/CGI により実装を行う．

2. 関 連 研 究

2.1 検索用ネットワークトポロジの構築手法

検索効率向上を企図して，同種のコンテンツ，クエリを，検索用ネットワーク上において近くに集中させるような手法である TellaGate [11] と Semantic Overlay Networks(SONs) [12] について述べる．

TellaGate は, コンテンツ発見率を維持した状態でのネットワークの通信負荷抑制を企図して，ピアに投入されるコンテンツ，クエリの内容からピアの嗜好を定義し，嗜好が類似したピア同士をピアネットワーク上で近くになるよう動的に再配置する手法である．TellaGate では，対象をテキストコンテンツとしているため，自然言語処理を用いて嗜好を決定しているが，映像，音声や複合コンテンツに対しては，個々に特徴量の抽出手法を開発しなければならないと考えられる．

SONs ではピアの保持しているコンテンツを考慮し，コンテンツのカテゴリ別にネットワークを形成する．具体的にはピアを，保持しているコンテンツの種類を基に，あらかじめ用意されている意味的なカテゴリに分類する．その上で，ピアが，カテゴリごとにネットワークを形成し，ユーザはコンテンツ，クエリを，同じカテゴリのピアへ投入する．

各カテゴリのピア数，コンテンツの種類数が少ない時は非常に効率が良い．そのため，流通するコンテンツが明確にカテゴリに分けられるときは有効な手段といえるが，カテゴリを事前に用意しておく必要があり，また，各カテゴリに対応する，ピアの数，コンテンツの数が多いと効果が期待できない．

2.2 検索アルゴリズム

コンテンツの偏在性と検索効率の実験で用いる検索アルゴリズムは Flooding [26],RandomWalkSearch [27],NISHA [29] [30] である．一般には，検索元のピアにおいて，ユーザなどから入力された検索クエリをピアが直接，接続しているピア (以下，隣接ピア) へ転送し，受け取ったピアがコンテンツを保持していれば検索成功，保持していなければさらに隣接ピアへ転送する．これを繰り返すことによって分散環境での検索を行っている．特に，クエリ転送時における，転送先ピアの選択方法が，各検索手法の特徴となる．

Flooding では，検索の際，ピアは隣接ピアすべてにクエリを転送する．クエリにはクエリループの防止およびネットワーク負荷の低減を目的として，Hops-To-Live（以下，HTL）と呼ばれる値が付与されており，クエリが 1 回転送されるごとに，ピアにおいて 1 減少させる．ピアは HTL が 0 になるまでクエリを転送する．

RandomWalkSearch（以下，RWS）では，検索元のピアはクエリをランダムに選ばれた k 個の隣接ピアへ転送する．クエリを受け取ったピアは，クエリ送信元以外のピアの中からランダムに一つのピアを選び，そのピアへクエリを転送する．Flooding と同様の理由でにクエリには HTL が付与されている．

NISHA は，各ピアがその時点でのコンテンツ配置に応じて自律的にコンテンツ間検索オーバレイネットワークを構築する，筆者らが提案している検索アルゴリズムである．NISHA では，各ピアが他ピアからコンテンツの位置情報を収集し，保持しているコンテンツにその情報を保持させることによって，検索用コンテンツ間グラフを分散構築，分散管理する．

3. コンテンツ偏在性と検索効率

意味的なカテゴリを考えたとき，コンテンツ，クエリはカテゴリに属するものとすると，ピアネットワークにおける，同じカテゴリのコンテンツ，クエリの集中度合いが検索効率に与える影響を，理論解析，およびシミュレーションを用いて検証する．実験では，検索手法として Flooding，RWS，NISHA を使用した．

本論文では，ピアに投入されるコンテンツ，クエリのカテゴリ分布をそれぞれコンテンツ分布，クエリ分布と呼ぶ．一つのピアに対応するコンテンツ分布，クエリ分布が同じである場合，

表 1　シミュレーション設定

優勢度 r_{prime}	20-100%
同カテゴリピア接続率 r_{conn}	20-100%
コンテンツ ID S_{con}	[0–999](1000 種類)
カテゴリ S_{cat}	[A–E](5 種類)
各ピアのコンテンツ数 S_{pc}	10
ピア平均次数 S_{pd}	4

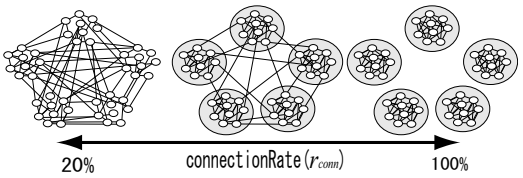

図 1　トポロジ

単にピアのカテゴリ分布と呼ぶ．それぞれのカテゴリ分布において最も大きな割合を占めるカテゴリを優勢カテゴリ，またその割合を優勢度とする．優勢度が大きい時，ピアに同カテゴリのコンテンツ，クエリが集中していることを表す．

3.1 実験内容

カテゴリの種類数を S_{cat} とした時，優勢度 $r_{prime}(1/S_{cat} < r_{prime} \leqq 1)$ と，同カテゴリのコンテンツ，クエリが集まるピア同士が接続されている確率である，同カテゴリピア接続率 $r_{conn}(1/S_{cat} < r_{conn} \leqq 1)$ を変数とし，検索成功率，および 1 クエリに対するホップ数の期待値を調べ，その値の変化状況などについて検証する．本論文では検索効率を効率性とし

$$(\text{効率性}) = \frac{(\text{検索成功率})}{(\text{1 クエリあたりのホップ数の期待値})} \tag{1}$$

とする．

Flooding,RWS は理論解析を用い，NISHA ではシミュレーションを行った．理論解析およびシミュレーション共通の設定は表に示す通りである．

トポロジの構成において，それぞれのピアは $S_{cat} = 5$ 種類のカテゴリ A–E のいずれかに属し，トポロジは各ピアが平均次数に相当するピアと接続されるようにする．このとき，同カテゴリピアとの接続数は同カテゴリピア接続率 r_{conn} によって決める．図 1 のように，同カテゴリピア接続率 r_{conn} が大きいと同カテゴリのピア同士が近くに集まるようになる．逆に同カテゴリピア接続率が小さい時はランダムなグラフとなる．

各コンテンツはコンテンツ ID を持っており，0–999 まで $S_{con} = 1000$ 種類のコンテンツ ID が存在するものとする．コンテンツ ID が等しいコンテンツは同一コンテンツである．また，各コンテンツはそれぞれいずれか 1 つのカテゴリに属しているものとする．

各カテゴリのコンテンツ数は均等に，各カテゴリ 200 種類とし，各ピアにコンテンツを 10 個配置する．コンテンツを配置する際，ピアのカテゴリと同カテゴリのコンテンツを配置する確率は優勢度 r_{prime} によって定める．カテゴリ数を 5 としたため，図 2 に示すように，優勢度 20% がコンテンツ，クエリとも，ランダムに投入される状態を表す．逆に，優勢度 100% はコンテンツ，クエリとも，完全にピアのカテゴリに則した状態である．

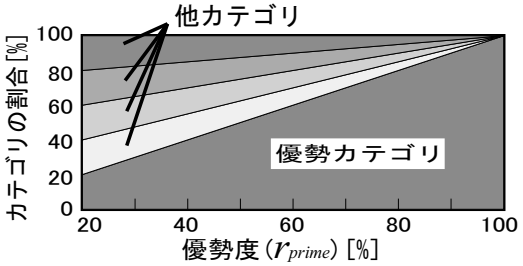

図 2　優勢度によるカテゴリ分布の変化

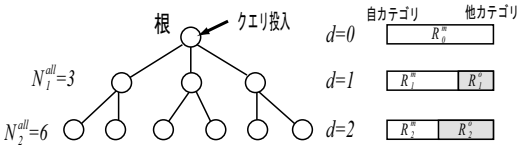

図 3　Flooding の探索範囲

3.1.1 Flooding の理論解析

検索成功確率 $R_f(r_{prime})$ を求める．Flooding の探索範囲はクエリのループを考えないとき，図 3 のように，投入されたピアを根とする木構造となる．

根のカテゴリを自カテゴリ，他のカテゴリをまとめて，他カテゴリとするとき，$R_f(r_{prime})$ は，自カテゴリに属するコンテンツの探索成功率 R_d^{fm} と，他カテゴリに属するコンテンツの探索成功率 R_d^{fo} とすると，

$$R_f(r_{prime}) = r_{prime}R_d^{fm} + (1 - r_{prime})R_d^{fo} \tag{2}$$

と書ける．それぞれの探索成功率は，各カテゴリに属するコンテンツが探索範囲内にいくつ入っているかを調べることができれば求める事ができる．すなわち，自カテゴリコンテンツが探索範囲内に n 個あるとすると，自カテゴリのコンテンツを発見できる確率 $R_d^{fm}(n)$ は

$$N_{am} = S_{con}/S_{cat} \tag{3}$$

$$R_d^{fm}(n) = 1 - \left\{\frac{N_{am} - 1}{N_{am}}\right\}^n \tag{4}$$

と表せる．N_{am} は自カテゴリに属するコンテンツの種類数である．

同様に，探索範囲内の他カテゴリコンテンツ数を n とすると

$$N_{ao} = (S_{con}/S_{cat})(S_{cat} - 1) \tag{5}$$

$$R_d^{fo}(n) = 1 - \left\{\frac{N_{ao} - 1}{N_{ao}}\right\}^n \tag{6}$$

となる．N_{ao} は他カテゴリに属するコンテンツの種類数である．

探索範囲内のコンテンツ数の期待値を求めるため，各ピアのコンテンツ数の期待値を求める．探索範囲内の自カテゴリコンテンツ数の期待値 N_d^{fm} および他カテゴリコンテンツ数の期待値 N_d^{fo} は，根からの距離 d のピア数を N_d^{all}，根からの距離 d のあるピアに存在する自カテゴリのコンテンツ数の期待値を N_d^{fmc}，他カテゴリのコンテンツ数の期待値を N_d^{foc} とすると，

$$\begin{cases} N_d^{fm} = \sum_{i=0}^{d} N_i^{all} N_i^{fmc} \\ N_d^{fo} = \sum_{i=0}^{d} N_i^{all} N_i^{foc} \end{cases} \tag{7}$$

と書く事ができる．

一方，根からの距離 d のピア数 N_d^{all} は

表 2　ピアにあるコンテンツ数の期待値

ピア＼コンテンツ	自カテゴリ	他カテゴリ
自カテゴリ	P^m_{mc}	P^m_{oc}
他カテゴリ	P^o_{mc}	P^o_{oc}

$$\begin{aligned} &N_0^{all} = 1 \\ &N_d^{all} = S_{pd}(S_{pd}-1)^{(d-1)} \quad (d \geqq 1) \end{aligned} \tag{8}$$

である．

N_d^{fmc}，N_d^{foc} は，ピアに存在するコンテンツ数の期待値を表 2 の様に表すとし，根からの距離 d のピアが自カテゴリである確率を R_d^m，および他カテゴリである確率を R_d^o とすると，

$$\begin{cases} N_d^{fmc} = P^m_{mc}R_d^m + P^o_{mc}R_d^o \\ N_d^{foc} = P^m_{oc}R_d^m + P^o_{oc}R_d^o \end{cases} \tag{9}$$

となる．P^m_{mc}，P^m_{oc}，P^o_{mc}，P^o_{oc} はそれぞれ

$$\begin{cases} P^m_{mc} = S_{pc}r_{prime} \\ P^m_{oc} = S_{pc}(1-r_{prime}) \\ P^o_{mc} = S_{pc}\frac{1-r_{prime}}{S_{cat}-1} \\ P^o_{oc} = S_{pc}\left\{r_{prime} + \frac{1-r_{prime}}{S_{cat}-1}(S_{cat}-2)\right\} \end{cases} \tag{10}$$

である．

最後に根からの距離 d のピアが自カテゴリである確率 R_d^m，他カテゴリである確率 R_d^o は，帰納的に，

$$\begin{cases} R_{mm} = r_{conn} \\ R_{mo} = 1 - r_{conn} \\ R_{om} = (1-r_{conn})/(S_{cat}-1) \\ R_{oo} = r_{conn} + \frac{(1-r_{conn})(S_{cat}-2)}{S_{cat}-1} \end{cases} \tag{11}$$

$$R_0^m = 1, R_0^o = 0 \tag{12}$$

$$\begin{cases} R_d^m = R_{d-1}^m R_{mm} + R_{d-1}^o R_{om} \\ R_d^o = R_{d-1}^m R_{mo} + R_{d-1}^o R_{oo} \end{cases} \quad (d \geqq 1) \tag{13}$$

となる．

次にクエリホップ数の期待値を考える．ルートからの距離 d の，あるピアで，ある自カテゴリコンテンツが発見できる確率を R_d^{fmc}，ある他カテゴリコンテンツが発見できる確率 R_d^{foc}，探索する深さ，すなわち HTL の値を D としたとき，根からの距離が d の，あるノードを根とした部分検索木の自カテゴリクエリのホップ数の期待値 N_d^{sm}，他カテゴリクエリのホップ数の期待値 N_d^{so} は，帰納的に，

$$N_D^{sm} = 0, N_D^{so} = 0 \tag{14}$$

$$\begin{cases} N_d^{sm} = (S_{pd}-1)(N_{d+1}^{sm}+1)(1-R_d^{fmc}) \\ N_d^{so} = (S_{pd}-1)(N_{d+1}^{so}+1)(1-R_d^{foc}) \end{cases} \quad (0<d<D) \tag{15}$$

$$\begin{cases} N_0^{sm} = S_{pd}(N_1^{sm}+1)(1-R_1^{fmc}) \\ N_0^{so} = S_{pd}(N_1^{so}+1)(1-R_1^{foc}) \end{cases} \tag{16}$$

と書ける．ただし，R_d^{fmc}，R_d^{foc} は

$$\begin{cases} R_d^{fmc} = 1 - \left\{\frac{N_{am}-1}{N_{am}}\right\}^{N_d^{fmc}} \\ R_d^{foc} = 1 - \left\{\frac{N_{ao}-1}{N_{ao}}\right\}^{N_d^{foc}} \end{cases} \tag{17}$$

である．

3.1.2　RWS の理論解析

RWS を行う根からの距離が d までの自カテゴリに属するあるコンテンツの探索成功率 R_d^{rm} と他カテゴリに属するあるコンテンツの探索成功率 R_d^{ro} は，

$$\begin{cases} N_d^{rm} = \sum_{i=0}^{i=d} N_i^{fmc} \\ N_d^{ro} = \sum_{i=0}^{i=d} N_i^{foc} \end{cases} \tag{18}$$

$$\begin{cases} R_d^{rm} = 1 - \left\{\frac{N_{am}-1}{N_{am}}\right\}^{N_d^{rm}} \\ R_d^{ro} = 1 - \left\{\frac{N_{ao}-1}{N_{ao}}\right\}^{N_d^{ro}} \end{cases} \tag{19}$$

となる．探索する深さを D^r とすると，Flooding とほぼ同様に，根からの距離が d の，あるノードを根とした部分検索木の自カテゴリクエリのホップ数の期待値 N_d^{rsm}，他カテゴリクエリのホップ数の期待値 N_d^{rso} は，帰納的に，

$$N_{D^r}^{sm} = 0, N_{D^r}^{so} = 0 \tag{20}$$

$$\begin{cases} N_d^{rsm} = (N_{d+1}^{sm}+1)(1-R_d^{fmc}) \\ N_d^{rso} = (N_{d+1}^{so}+1)(1-R_d^{foc}) \end{cases} \quad (0 \leq d < D^r) \tag{21}$$

となる．

3.1.3　NISHA のシミュレーション

NISHA は理論解析を行うことが難しいため，シミュレーションによる解析を行った．ピア数は 200，各カテゴリに属するピアの数は等分し，各 40 とした．NISHA における検索インデックス構築のための情報交換は隣接ピアとのみ行う．全ピアから全コンテンツを検索し，自カテゴリに属するコンテンツの探索成功率 R^{fm} と，他カテゴリに属するコンテンツの探索成功率 R^{fo} を求め，検索成功率 $R_n(r_{prime})$ を，

$$R_n(r_{prime}) = r_{prime}R^{fm} + (1-r_{prime})R^{fo} \tag{22}$$

として求めた．1 クエリあたりのホップ数の期待値も，同様に求めた．

3.2 結　　果

Flooding の実験結果を図 4–6 に示す．結果のグラフは各アルゴリズムにおいてほぼ同様の傾向を示した．次に，図 7–9 はそれぞれ，検索成功率，ステップ数，効率性のグラフを $r_{conn} = r_{prime}$ の平面で切ったときのグラフである．nishaRandom の直線は各メトリックで $(r_{prime}, r_{conn}) = (20\%, 20\%)$ のとき，すなわち，トポロジ，コンテンツ配置，クエリがすべてランダムの時の値である．すべての検索アルゴリズムにおいて，優勢度 r_{prime}，同カテゴリ接続率 r_{conn} がともに大きいとき，検索効率は良くなることがわかる．

次に各アルゴリズムについて個別に考察する．

3.2.1　Flooding の考察

検索成功率は優勢度，同カテゴリピア接続率を大きくすることによって，非常に良くなる．これは図 10 のように両パラメタが大きいときに同カテゴリが集中している場所と，Flooding による探索範囲が，ほぼ同一であるためである．

一方，クエリホップ数の期待値はよくならなかった．これは，ピアの近くでコンテンツを発見できても，他のルートを通っているクエリは探索を続けるためである．図 5 において，優勢度

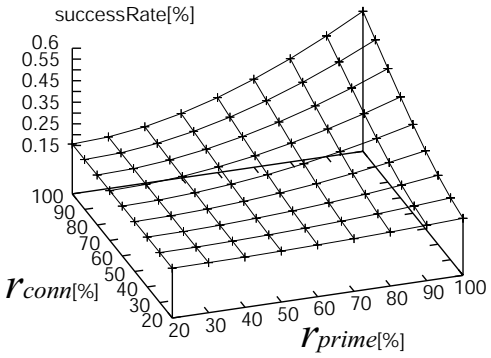

図 4　Flooding:検索成功率

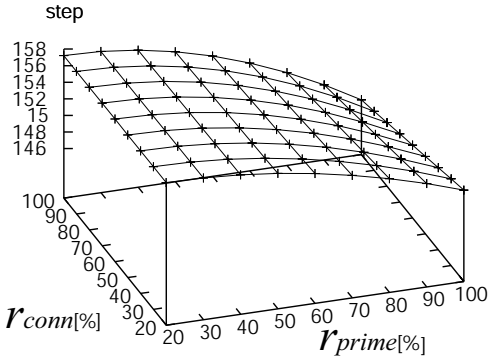

図 5　Flooding:1 クエリあたりのホップ数

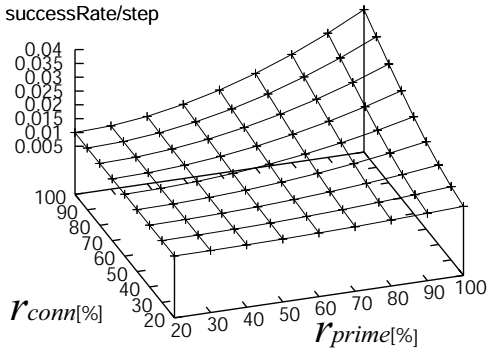

図 6　Flooding:効率性

r_{prime} が大きくなるにつれて，全体的に値が下がっているのは，クエリが投入されたピアで探索を終了する確率が高くなるためである．

3.2.2　RWS の考察

優勢度，同カテゴリピア接続率が極端に大きくないと Flooding に比べ検索成功率は低い．これは RWS による探索範囲が Flooding のように，自カテゴリピアが集中する近傍から探索するのではなく，図 10 に見るように，模式的に遠くまで調べるためである．同カテゴリピア接続率 r_{conn} が 100 になると値が高くなるのは，どこまで遠くに行っても自カテゴリピアにしか到達しないためである．

図 8 を見ると，クエリホップ数の期待値は Flooding に比べ，よくなっている．これは，RWS のクエリが 1 本のルートしか通らないため，ルートの途中で検索が成功すれば，以降クエリが転送がされなくなり，検索成功率が上がれば，クエリがルートの途中で止まる可能性が高くなるためである．

3.2.3　NISHA の考察

図 11 は，$r_{rate} = 20(\%)$ の時の NISHA の検索成功率である．このグラフにおいて，同カテゴリピア接続率 r_{conn} が高くなると検索成功率が悪くなっている．これに対する明確な理由はわかっていないが，隣接ピア同士が共通の隣接ピアを持つことが多くなるからではないかと考えられる．例えば，優勢度が 20% のとき，コンテンツのカテゴリ分布はすべてのピアにおいて同じであるため，r_{conn} が大きくなるにつれ検索成功率が悪くなる原因はトポロジにあると考えられる．ここで同カテゴリピア接続率が 100% の時を考えると，図 1 の最右図のように，それぞれカテゴリのピア同士のみで繋がる 5 つのネットワークに分割される．このとき，各ネットワークのピア数は 40 しかなく，ピアの平均次数が 4 であるので，例えば，図 12 のように隣接するピア同士が共通する隣接ピアを一つ以上持っている

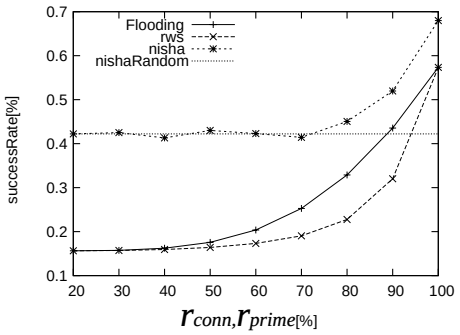

図 7　検索成功率:接続率＝優勢度

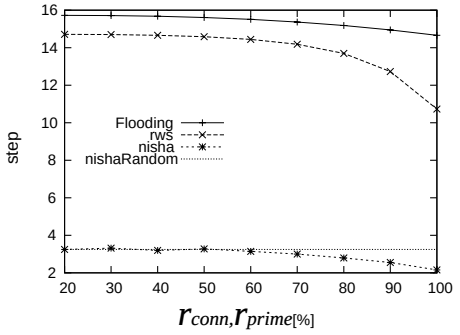

図 8　ホップ数:接続率＝優勢度

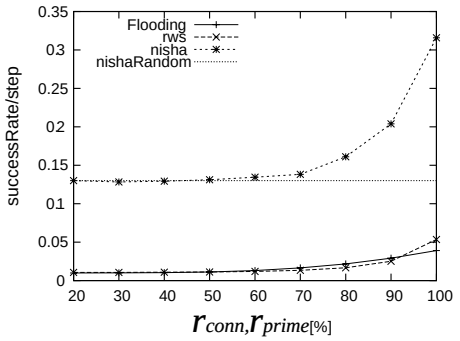

図 9　効率性:接続率＝優勢度

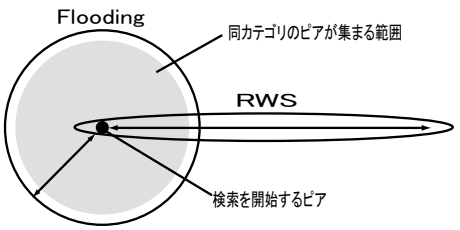

図 10　検索範囲の概念図

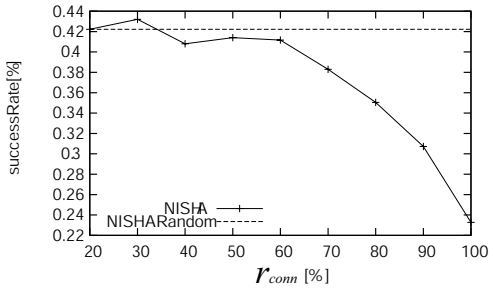

図 11　NISHA 検索成功率:優勢度=20%

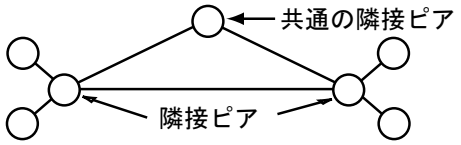

図 12　共通する隣接ピアを持っている状態

確率 $R(n)$ は，ネットワークに存在するピアの数を n とすると

$$R(n) = \frac{(n-2) \times {}_{n-3}C_2 \times {}_{n-3}C_2}{{}_{n-2}C_3 \times {}_{n-2}C_3} = \frac{9}{n-2} \qquad (23)$$

となり，$n = 200$ すなわち，同カテゴリピア接続率 $r_{conn} = 20\%$ のとき $R(200) = 0.045$ なのに対し，$n = 40$ すなわち，同カテゴリピア接続率 $r_{conn} = 100\%$ のとき $R(40) = 0.238$ と非常に大きくなる．NISHA において検索インデックスを作成する際，各ピアは必要のないコンテンツの参照を他ピアに送る．このとき，隣接するピアが同じ情報を収集するのは効率が悪いと考えられる．

また，クエリホップ数の期待値は同カテゴリピア接続率が 60% を越えたあたりから良くなる．

4. 提 案 手 法

3. 章では優勢カテゴリが同じピアを近くに配置し，さらに，ピアの優勢度を大きくすることで，検索効率が向上する事を示した．すなわち，ある検索意図（クエリ）およびコンテンツへの適合度が高いコンテンツ群をクエリ，コンテンツのカテゴリとし，2 つのカテゴリが類似しているときそのクエリ，コンテンツは同種であるとすると，ネットワーク上において同種コンテンツ同士を近くに配置し，同種のコンテンツが集まる場所から検索を始めることで効率的な検索ができる．そこで本章では，3. 章の実験結果を基に，効率的な検索を行うことができるモデル，および実現手法を提案する．

4.1 提案モデル

従来の P2P コンテンツ配送システムでは図 13 のように 1 人のユーザに対して 1 つのピアを割り当て，ユーザは 1 つのピアに対してコンテンツ，クエリを投入するが，各ユーザが，コンテンツ，クエリの内容に応じて，投入するピアを選ぶことが出来れば，各ピアに投入されるコンテンツ，クエリの内容の分布は，多くの P2P システムに比べ，さらに大きな偏りを作ることが出来る．そこで，提案モデルではまず，利用者が検索を始める場所を指定することができるよう，図 14 のように利用者がどのピアでも使えるようにする．ここで，利用者がクエリおよびコンテンツと同種のコンテンツが集まる場所に投入できるよう，ピアにカテゴリを示す情報を付加する．この情報をピアのタグとする．また，タグの示すカテゴリが類似するピア同士を接続することにより，同種のコンテンツがネットワーク上において近くになるようにする．

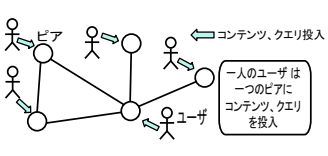

図 13　従来の P2P

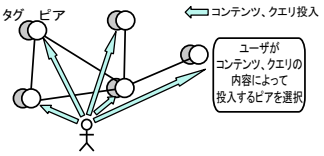

図 14　提案手法の P2P

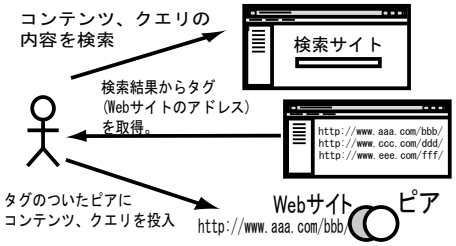

図 15　コンテンツ，クエリ投入方法の例

本提案モデルでは，利用者の手元にある端末をピアとするよりもむしろ，Web サーバのような常時稼働するサーバをピアとしても機能させることを志向している．このような形態では，ピアの安定性などでは優れるが，コンテンツの最終配送がボトルネックとなる可能性が指摘される．しかし，現在のコンテンツ配信の主流である Web においては，それぞれのユーザにとって十分なアクセス速度を提供できるサーバが相当数あることを前提としているため，これは問題とならない．

4.2 提 案 手 法

Web サイトは 1. 章で述べたとおり，ハイパーリンクによって同種の Web サイトとつながっており，Web コミュニティを形成している．したがって，Web サイトが示すカテゴリは Web サイト自体の内容と，そのサイトからハイパーリンクによって繋がっている他の Web サイトから特徴付けられると考えられる．

そこで，本論文ではまず，以下の二つの理由からピアと Web サイトを関連付け，ピアに関連付けられた Web サイトと同種のコンテンツ，クエリを集めることにする．

まず，一つ目の理由は，既存の Web サイトは多様であるためである．Web サイトの集合である WWW は分散型のデータベースとみなすことが出来，内容は非常に多種にわたる．また，本方式では目的としていないが，WWW は単一障害点を持たず，ピュア型 P2P の利点である可用性を損なわないという利点もある．

二つ目は，コンテンツ，クエリのカテゴリに類似するカテゴリを示す Web サイトを発見することは，容易であると考えられるためである．なぜなら，既存のサーチエンジンはコンテンツ，クエリのカテゴリと類似したカテゴリを示す Web サイトを発見するには十分な精度を持っているからである．

例えば，図 15 のようにコンテンツ，クエリの内容，もしくはコンテンツの名前などを用いて Web サーチエンジンで Web サイトを検索し，検索結果として表示されるサイトに対応付けられたピアにコンテンツ，クエリを投入すればよい．

次に，Web リンクの情報を利用し，同種のコンテンツを保持するピア同士を近くに配置するピアネットワーク構築手法を提

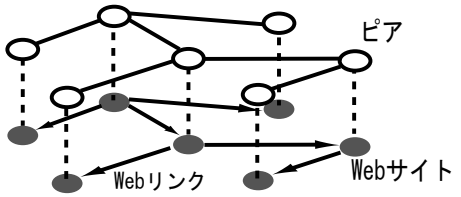

図 16　オーバレイネットワーク構築

案する．

各ピアは Web サイトのカテゴリと類似するカテゴリを持つコンテンツを保持しているものとすると，1. 章で述べたとおり，内容の類似する Web サイト同士はハイパーリンクによって密に繋がっているため，図 16 のように，Web グラフに倣って構築することで，同種のコンテンツを保持するピア同士を近くに配置することが出来る．

そこで本論文では，Web リンクからピアネットワークを構築する簡単な手法の一つとして，ピア間の Web リンクが多いピア同士を接続することでピアネットワークを構築する．ここで言うピアの Web リンクとはピアに関連付けた Web サイトの Web リンクのことである．

具体的に，次に示すアルゴリズムによる自律分散なピアネットワーク構築手法を提案する．ピアが他ピアとの接続する際，ピアは接続要求を行い，相手のピアが接続を許可すると，ピア同士が接続した状態になる．まず，接続要求を行おうとしているピアを自ピアとすると，Web リンクにおける隣接ピアを自ピアとの Web リンク数が多い順に並べる．次にピアとしての接続要求をその順番で行い，自ピアが接続要求を出した先のピアとの Web リンク数の累計を N_{sum}，自ピアの総 Web リンク数を N_{all}，パラメタを $p(0 \leqq p \leqq 1)$ とした時，

$$N_{sum}/N_{all} > p \tag{24}$$

となるまで接続要求を行う．また，接続要求を受けたピアは，ピアネットワークに参加できないピアが出ないよう，必ず接続を許可する．

5. 実装システム

ピアと Web サイトを結びつけるための簡単な手法として，ピアの機能を CGI/Perl で実装し，Web サイトに付加する形で利用できるようにする．提案方式はオーバレイネットワーク構築時に Web リンク情報を取得する必要がある．筆者らは，周辺リンク情報を容易に取得できるようにする協調型 Web アーキテクチャ [31]（Cooperative Web Architecture; 以下 CWA）を提案している．

そこで，試作システムは図 17 のように，CWA を実現する vAWS モジュールと P2P 検索，配送モジュールから構成する．CWA は P2P 系とは独立に既存の Web に追加される形で機能し，周辺リンク情報を他の Web サーバと協調して維持，管理する機能を持つ．

vAWS により Web グラフを保持し，この情報から，P2P コンテンツ配送システムが関連付けられた Web サイトと密に接続されている Web サイトを見つけることによって，内容の類似する Web サイトを発見する．具体的な手法は 4.2 節で述べた手法を用いる．

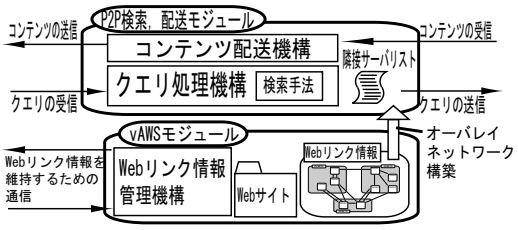

図 17　ブロック図

5.1 vAWS モジュール

CWA では，Web サーバが自身の保持するコンテンツの更新を確実に検知できる特性を利用して，Web サーバの付加機能としてサーバ内 HTML コンテンツの Web グラフ構築，ならびに管理を行う．同時に各 Web サーバが近隣の最新の Web グラフ（以下，部分 Web グラフ）を保持することが出来るよう，近傍コンテンツの更新時に当該コンテンツを保持するサーバからの通知，およびリンク情報の提供を受ける仕組みを導入する．このような付加機能を導入した Web サーバのことを AWS (AdvancedWebServer) と呼んでおり，アクセスした利用者に対してコンテンツだけでなく，部分 Web グラフ情報を提供することが出来る．

CWA の理念としては，付加機能は完全に Web サーバと一体で動作すべきだが，ここでは導入の容易さなどから CGI による実装とし，これを virtualAWS（以下，vAWS) と呼ぶ．vAWS モジュールはサーバの持つ HTML コンテンツから周囲 n ホップの Web リンク情報を維持，管理するモジュールであり，ディレクトリ単位で Web リンク情報を管理できるようになっている．また，筆者らは Web リンク情報の同期方式 [32] を提案しているが，本システムでは基礎的な方式を用いた．

5.2 P2P 検索，配送モジュール

P2P 検索，配送モジュールは，ユーザからのクエリを受け，他ピアと協調動作することで，コンテンツを検索，配送するモジュールであり，次の三つの機能を有する．

- vAWS が保持している Web リンク情報から他ピアを発見し，接続要求を出す
- ユーザもしくは他ピアからのクエリを受信した時にコンテンツの返送，もしくは検索アルゴリズムに応じて適切なピアへクエリを転送する
- コンテンツを検索ルートに沿って配送する

5.3 考　　察

本実装システムは，DEWS2006 デモセッションにおいて用意した数台の PC を接続した動作デモを行い，数 10query/s 程度の処理が可能であることなどを示した．本システムの現実の Web 空間における効率性や負荷の検証は今後の課題である．しかしながら，3. 章の結果と Web コミュニティの研究成果を考慮すると，ある程度のカテゴリに関する偏在性は実現可能であり，かつまた，その場合に相応の検索効率の向上が期待できる．

6. おわりに

本論文では，まず，P2P コンテンツ配送システムにおいて，優勢度および，同カテゴリピア接続率の変化による検索効率に

ついて分析を行い，優勢度および，同カテゴリピア接続率がともに大きい時に検索効率が良くなることを明らかにした．

次に，優勢度および，同カテゴリピア接続率が大きくなるよう，ピアと Web サイトを関連付け，ユーザがコンテンツ，クエリの内容に対応する Web サイトに関連付けられたピアへ投入する手法と，意味的に同じカテゴリに属する Web サイト同士が Web リンクによって密に繋がっている事を利用し，Web リンク情報に倣ってピアネットワークを構築することで，類似するピア同士をネットワークにおいて近くに配置する手法を提案した．また，実システムによる性能評価は行っていないが，3. 章の実験より，明確なカテゴリ分類がされていない提案手法においても検索効率の向上が期待できる．

最後に，Perl/CGI を用いて Web リンク情報に倣ってピアネットワークを構築する P2P コンテンツ配送システムの実装を行った．

今後の課題として，トポロジ構築アルゴリズムの改善や，本論文では対象外とした，コンテンツ配置法などが挙げられる．

本提案は，Web におけるコンテンツ流通モデルを考慮したとき，ピアの機能をサーバ側に持たせ，利用者の嗜好毎に接続するピアを検索エンジンなどを用いて選択させることが，3. 章での検証によって合理的であるとの発想に基づく．また，Web コミュニティの研究を踏まえた上で，Web リンク情報を利用したトポロジ構成法を提案し，容易に導入可能な実装システムとして実現することで，実システムへの導入を今後進めていきたい．

文　献

[1] Miguel Castro, Peter Drushel, Y.C. Hu, and Antony Rowstron: “Exploiting Network Proximity in Peer-to-peer Networks. Technical Report,” MSR-TR-2002-82, Microsoft Research, 2002.

[2] Kirsten Hildrum, John D. Kubiatowicz, Satish Rao, and Ben Y. Zhao: “Distributed Object Location in a Dynamic Environment,” In Proc. of the ACM SPAA, 2002.

[3] Sushant Jain, Ratul Mahajan, and David Wetherall: “A Study of Performance Potential of DHT-based Overlays,” In Proc. of the 4th USITS, 2003.

[4] David R. Karger and Matthias Ruhl: “Finding Nearest Neighbours in Growth-restricted Metrics,” In Proc. of the ACM STOC, 2002.

[5] Sylvia Ratnasamy, Mark Handley, Richard Karp, and Scott Shenker: “Topologically-Aware Overlay Construction and Server Selection,” In Proc. of the INFOCOMM, 2002.

[6] Ben Y. Zhao, Anthony Joseph, and John D.Kubiatowicz: “Locality Aware Mechanisms for Large-scale Networks,” In Proc. of the FuDiCo 02, 2002.

[7] 阿武孝文, 朝香卓也, 高橋達郎: “Unstructured 型 P2P ネットワークにおける検索メッセージ転送方式,” 電子情報通信学会論文誌，vol.J88-B, no.2, pp.372-382, 2005.

[8] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker: “A Scalable Content-Addressable Network,” In Proc. of the ACM SIGCOMM 2001, 2001.

[9] Ion Stoica, Robert Morris, David Karger, Frans Kaashoek, and Hari Balakrishnan: “Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications,” In Proc. of the ACM SIGCOMM 2001, 2001.

[10] A.Rowstron and P.Druschel: “Pastry : Scalable, decentralized object location and routing for large-scale peer-to-peer systems,” Middleware, 2001.

[11] K.Kojima: “Self-Organizable P2P Search Engine for Knowledge Management,” the first International Workshop on Peer-to-Peer Knowledge Management, 2004.

[12] A. Crespo and H. Garcia-Molina: “Semantic overlay networks for p2p systems,” Technical report, Computer Science Department, Stanford University, 2002.

[13] H. Zhang, A. Goel, and R. Govindan: “Using the Small-World Model to Improve Freenet Performance,” In Proc. of IEEE Infocom, 2002.

[14] V. Kalogeraki, D. Gunopulos, and D. Zeinalipour-Yazti: “A Local Search Mechanism for Peer-to-Peer Networks,” In Proc. of the Eleventh International Conference on Information and Knowledge Management, pp. 300-307, 2002.

[15] S. Joseph: “NeuroGrid: Semantically Routing Queries in Peer-to-Peer Networks,” In Proc. of the International Workshop on Peerto-Peer Computing, 2002.

[16] P. Haase, R. Siebes, and F. van Harmelen: “Peer Selection in Peer-to-Peer Networks with Semantic Topologies,” In Proc. of the International Conference on Semantics in Networked World (ICNSW ’04), volume 3226 of LNCS, pp. 108-125, 2004.

[17] Ka Cheung Sia. P2P information retrieval: “A self-organizing paradigm,” Technical report, 2002.

[18] Tsunenori Mine, Daisuke Matsuno, Akihiro Kogo, Makoto Amamiya: “Design and Implementation of Agent Community Based Peer-to-Peer Information Retrieval Method,” CIA 2004, pp.31-46, 2004.

[19] Jon M. Kleinberg: “Authoritative sources in a hyperlinked environment,” In Proc. of the ninth annual ACM-SIAM symposium on Discrete algorithms, p.668-677, 1998.

[20] Jeffrey Dean and Monika R. Henzinger: “Finding related pages in the World Wide Web,” In Proc. of the 8th WWW Conference, 1999.

[21] 豊田正史: “WWW における関連コミュニティ群の発見,” 情処研報 DBS122-40, IPSJ, 2000.

[22] R. Kumar, P. Raghavan, S. Rajagopalan and A.Tomkins, “Trawling the web for emerging cybercommunities,” In Proc. of the 8th WWW Conference, 1999.

[23] 村田剛志: “参照の共起性に基づく Web コミュニティの発見,” 人工知能学会誌、Vol.16 No.3, pp.316-323, 2001.

[24] 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,” IPSJ 研究報告 NL-142-007, 2001.

[25] 浅野泰仁、吉田雄介、西関隆夫、豊田正史、喜連川優：“サイト間グラフの最小カットを用いたウェブ上のコミュニティ発見法,” アルゴリズム研究会 情処研報, AL-095-8, pp.51-58, 2004.

[26] Gnutella. The Gnutella Protocol Specification: www9.limewire.com/developer/gnutella protocol 0.4.pdf, 2004.

[27] Qin Lv et al: “Search and Replication in Unstructured Peer-to-Peer Networks,” In Proc. of the 16th ACM ICS, 2002.

[28] N. Bisnik and A. Abouzeid: “Modeling and Analysis of Random Walk Search Algorithms in P2P Networks,” to appear in HOT-P2P 2005, 2005.

[29] スレスタサンブ, 小林亜樹, 山岡克式, 酒井善則 : “CDN における近傍ノードのコンテンツによる検索木の分散構成に基づく効率的コンテンツ検索手法,” 情処研報 DBS-137-28, pp.207-214, 2005.

[30] Shambhu Shrestha, Aki Kobayashi, Yoshinori Sakai, Katsunori Yamaoka, Noboru Sonehara: “Efficient Content Location Algorithm for Content Destribution Networks based on Distributed Construction of Search Tree from Contentts of Proximal Nodes,” In Proc. of IASTED DBA2006, pp.101-108, 2006.

[31] Kobayashi Aki, Yamaoka Katsunori, Sakai Yoshinori : “Cooperative Web Architecture for Search and Navigation Assistance,” In Proc. of ICWI2002, IADIS, 2002.

[32] 高砂幸代、小林亜樹、山岡克式、酒井善則: “Web サーバ間での部分 Web グラフ同期方式の提案,” DEWS2006 7C-o2, 2006.