確率的セマンティック P2P ルーティングの性能評価

M2006MM020 森下 広史
指導教員　河野 浩之

1　はじめに

P2P システム [4] は，資源を共有する目的でネットワークトポロジの自己組織化が可能なノードの相互連結で成り立つ分散システムである．また，ピュア型の P2P システムは管理構造の欠如のために検索効率の低さや不安定な構造などの問題点を抱えている [4].

本研究では，ピュア型の P2P システムがクエリーをブロードキャストすることによるトラフィックの増大を抑制することを目的として，意味的な手法に基づいて効率の良いピアの検索や P2P オーバレイネットワークの組織化を達成する．また，本研究は，過去の研究 [3] を拡張したものである．

2 章では，P2P システムにおける構造化手法について述べる．また 3 章では，セマンティック指向のルーティングアルゴリズムを提案する．そして 4 章では，提案した手法についてシミュレーションによる実験とその評価を述べる．5 章でネットワークの組織化について述べ，6 章でまとめる．

2　P2P システムの構造化手法

2.1 節はインデックス指向で構造化した手法を 2.2 節ではセマンティック指向で構造化した手法を述べ，これらの P2P システムの制御手法を表 1 にまとめる [4] .

構造化手法			例
インデックス指向 P2P システム			
	チェーンモード伝播		Freenet
	DHT	関連フィンガーテーブル	Chord
		多次元の座標空間	CAN
		plaxton メッシュ	Tapestry
			Pastry
			Kademlia
セマンティック指向 P2P システム			
	トピックベース	トピック階層	SONs
		ローカルインデックス	RIs
	JXTA	RDF ベース	Edutella
		SWAP	REMINDIN’
			INGA

表 1　P2P システムの分類

2.1　インデックス指向 P2P システム

インデックス指向による構造化手法は，効率，回復力，柔軟性などが設計の関心事となる．また，システム内のスーパーピアと呼ばれる一部の特別なピアがサーバとしての側面を持ち，制御のための特別な動作を行う．ただし，このスーパーピアは，ハイブリッド型システムにおけるサーバのような問題を抱える．

2.2　セマンティック指向 P2P システム

セマンティックなアプローチによる構造化は，ユーザの興味や嗜好によるコンテンツの偏りに基づいている．しかし，ピアのメタ情報の収集やインデックスの構築などを必要とする．このような複雑なメカニズムは，トラフィックの増大を引き起こす．

INGA(Interest-based Node Grouping Algorithms) [2] は存在するルーティングアルゴリズムと異なり，応答と対照でインデックス化情報を確立するピアを経由するクエリーを使用する．また，転送する目的で，インデックス情報を評価するための意味的な情報を使用する．加えて，インデックス更新のために意味的な情報を使用する．

2.3　構造化に関する課題

P2P システムの制御手法について，セマンティックな関係を考慮する手法が考案されているが，検索効率の向上に対してトラフィックを低減させる効果は十分ではない．そこで，本稿ではトラフィックの低減について検討する．また，知識のローカルな集合を利用する場合に，オーバレイネットワークの変化やユーザの興味が変化する状況に対応する必要がある．そのため，P2P ネットワークにおける柔軟で適応的でスケーラブルなネットワークの構築及び制御手法が要求される．

3　P2P ネットワーク組織化アルゴリズム

確率的ルーティングは，ピアの接続状態に応じた乱数を用いてパケット送出を確率的に行うことで，トラフィックを低減する．しかしながら，コンテンツや検索式の偏りを考慮していないため，全体的に検索効率を低下させる問題がある．

本研究では，ユーザの興味や嗜好に基づいた意味的なルーティングの制御を行う．ここでは，確率的ルーティングをユーザの嗜好に関する情報に基づいて拡張する．そして，クエリーの応答が得られる可能性が高いピアと積極的に通信できる動的なネットワークの組織化を図るメカニズムを提案する．提案するセマンティックな確率的ルーティングでは，キーワードに対する類似度に応じて分配した重みから閾値を決定して，その閾値とランダム値を比較してクエリーの送信や制限を行う．

3.1　キーワードの類似性を考慮した重みの導入

ここでは，クエリーの応答から得られるキーワードの関係性に注目して重みを分配することを提案する．s_1 と s_2 のトピック間の類似度関数 [1] は，$Sim(s_1, s_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}$ で定義される．ただし，$s_1 = s_2$ のとき，$Sim(s_1, s_2) = 1$ である．そして，$\alpha = 0.2, \beta = 0.6$ として，l はトピック s_1 と s_2 の間の最短経路，また，h は語彙階層のトップへの包摂からレベルをカウントする

ことで得られる包摂の深さである．

これより，送信したクエリーに含まれるキーワード (k) に対して，クエリーを受信したピア (P_i) との類似度を求める (式 (1))．

$$Sim(k, P_i) = max_{l \in J_i}\{Sim(k, s_i^l)\} \tag{1}$$

s_i^l は P_i が保持するキーワードであり，J_i は P_i が保持するキーワード集合である．この式 (1) の値を類似度とする．

3.2 確率的セマンティック P2P ルーティング (Probabilistic Semantic P2P Routing，PSPR) アルゴリズム

本節では，セマンティックな確率的ルーティング (PSPR) を行う以下のアルゴリズムを提案する．

確率的セマンティック P2P ルーティング (PSPR)

[Step 1]
ピアが接続しているピア数である接続数 N とトラフィックの総量に基づく仮想的なピアの接続数である接続設定数 N_S を設定する．$N > N_S$ なら **[Step 2]** へ，$N \leq N_S$ なら **[Step 7]** へ進む．

[Step 2]
N_S から $\sum_N W$ と $\sum_N R$ を分配する (式 (2))．N_S について $\sum_N W$ と $\sum_N R$ の関係は，$0 \leq a < 1$ とすると，$N_S \cdot a = \sum_N W$ と $N_S \cdot (1-a) = \sum_N R$ である．

初期閾値 W について式 (3) で与える．

また，送信するクエリーに含まれるキーワード $Keyword$ とすると，クエリー送信の対象となるピアに対する類似度 S は，3.1 節の類似度関数 [1](式 (1)) から式 (4) となる．ここで，正規化した類似度を S_n とすると，R は各ピアの S_n に基づいて分配する類似度重みとなる．この R を式 (5) で示す．

$$N_S = \sum_N W + \sum_N R \tag{2}$$

$$W = \frac{N_S}{N} \cdot a = \frac{(N_S \cdot a)}{N} \tag{3}$$

$$S = Sim(k, P_i) \tag{4}$$

$$\begin{aligned} R &= \{N_S \cdot (1-a)\} \cdot \left(\frac{Sim(k, P_i)}{\sum_N Sim(k, P_i)} \right) \\ &= \{N_S \cdot (1-a)\} \cdot \left(\frac{SV}{\sum_N SV} \right) \\ &= N_S \cdot (1-a) \cdot SV_n \end{aligned} \tag{5}$$

[Step 3]
上限値 L は，確率的にルーティングの制限を扱うために，$L = 1$ とする．T が L に等しくなる基準の値を基準値 D とする (式 (6))．また，S_n の最大値を S_{max} とする．S_{max} が D より大きな値である場合は，T が $L(= 1)$ より大きな値となるので，重みの分配を修正する必要がある．$S_{max} \leq D$ なら **[Step 4]** へ，$S_{max} > D$ なら **[Step 5]** へ進む．

$$\begin{aligned} D &= L \cdot \frac{1}{N_S \cdot (1-a)} - \frac{N_S}{N} \cdot a \cdot \frac{1}{N_S \cdot (1-a)} \\ &= \frac{L}{N_S \cdot (1-a)} - \frac{a}{N \cdot (1-a)} \end{aligned} \tag{6}$$

[Step 4]
初期閾値 W と類似度重み R から閾値 T を設定する (式 (7))．

$$T = W + R \tag{7}$$

[Step 5]
$S_{max} > D$ のとき，W の値を増加させることで，T を L 以下の値とする．式 (8) で L と S_{max} の関係を与え，式 (9) で修正する初期閾値 W_r を示す．式 (2) より，W が増加すると R は減少する．これより，T は L を越えない値に修正される．閾値 T は，$T = W_r + R_r (R_r = N_S - W_r)$ とする．

$$(N_S - N \cdot W_r) \cdot S_{max} + W_r = L(= 1) \tag{8}$$

$$W_r = \frac{L - N_S \cdot S_{max}}{1 - N \cdot h_{max}} \tag{9}$$

[Step 6]
$V \leq T$ を満たすピアにクエリーを送信する．また，$V > T$ なら **[Step 7]** へ進む．

[Step 7]
終了 □

4 シミュレーションによる性能評価

シミュレーションによる実験を行った結果を示す．

4.1 シミュレーションのセッティング

INGA[2] では，グリーディに類似度の上位 k ピアを選択する．この選択を類似度の上位 n ピアから確率的に k ピアを選択するように PSPR を適用する．クエリーを送信する場合，ピアが保持するインデックスからクエリー送信の対象となるピアを決定してルーティングテーブルを生成する．ピアの情報をストアするインデックスは，クエリー送受信の際に更新される．

INGA[2] に基づいて，ピア数 (P_{num}) を 1,000，クエリー数 (Q_{num}) を 30,000 回まで，最大ホップ数 (H_{num}) を 6 とする．興味の変化をクエリー 15,000 回で導入する ($IC_{time} = 15,000$)．

オープンディレクトリプロジェクトの DMOZ [*1] より，カテゴリの一部分からキーワード数 (K_{type}) を 282 とする．また，キーワードの選択は，Zipf 分布を考慮し，興味の変化の導入で再度選択する．

最大接続数 (C_{num}) は 5 とする．各接続数に対して，確率的に 2 ピアを選択するような仮想的接続設定数

[*1] http://www.dmoz.org/

(V_{num}) とする．INGA[2] では，ランダムな選択を 20% としていることから，初期閾値 (T_{base}) を 0.2 とする．また，インデックス数は 40 として，インデックス更新に LRU 戦略を用いている．

4.2 PSPR によるトラフィック量の低減と応答数

4.1 節で設定したパラメータによってシミュレーションによる実験を行った．転送数，応答数によってシミュレーションによる実験の結果を評価する．まず，トラフィック量の低減を確認するために転送数を評価する．そして，トラフィック量の低減に対して十分な応答が得られているのかを確認するために応答数を評価する．

4.2.1 クエリー制御によるトラフィック量の低減

図 1 は，V_{num} に対するトラフィック量となる転送数を示している．V_{num} の低減に対して，転送数が減少し

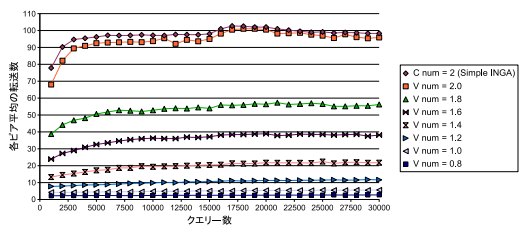


図 1　転送数

ていることが分かる．“ $IC_{time} = 15,000$ ”の直後に，転送数が一時的に増加していることが確認できる．ピアは同一のクエリーを受信した場合の処理を行わないので，ネットワークの再構築のため受信したクエリーが増えて転送数が一時的に増加したと考えられる．

4.2.2 クエリー応答受信が可能なクエリー送信の制限

図 1 でトラフィックの低減量を示したが，クエリーを送信した際に応答が得られることを考慮する必要がある．応答数が 1 以上となれば，送信したクエリーに対して回答が得られるので，応答数が 1 となるまでのトラフィックの低減が可能である．

図 2 は，トラフィック低減に対する応答数の変化を示している．V_{num} を低減させた場合には，応答数は減少

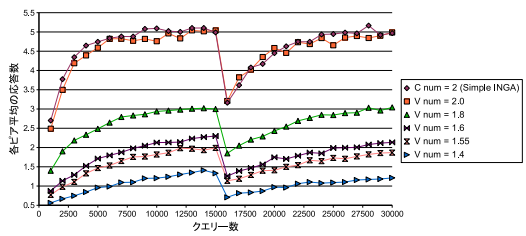


図 2　トラフィック低減に対する応答数

する．“ $IC_{time} = 15,000$ ”において興味の変化を導入したときに応答数が減少しているが，応答数 1 以上を維持できる値は，“ $V_{num} = 1.55$ ”となる場合である．このとき，図 1 で示されるトラフィック量は，67% の低減であった．

4.2.3 キーワードの数を変化させた場合の応答数の変化

次に，K_{type} を変化させた場合の応答数の変化を確認した．図 3 では，階層構造の深さを保って K_{type} を 141

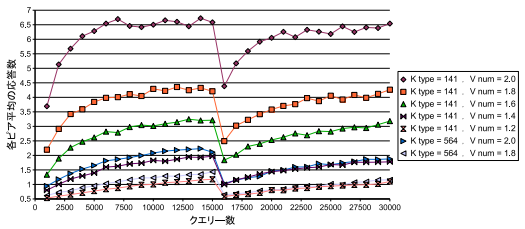


図 3　キーワード数を変化させた場合

と 564 と変化させた場合を示す．“ $K_{type} = 141$ ”とした場合では，応答数 1 となる V_{num} は 1.4 であった．このとき，トラフィック量は，77% の低減であった．また，“ $K_{type} = 564$ ”とした場合では，応答数 1 となる V_{num} は 2.0 となり，トラフィックの低減を図ることができなかった．

4.2.4 ピア数を変化させた場合の応答数の変化

そして，ピア数を変化させた場合の応答数の変化を確認した．図 4 は，“ $V_{num} = 1.55$ ”として P_{num} を変化

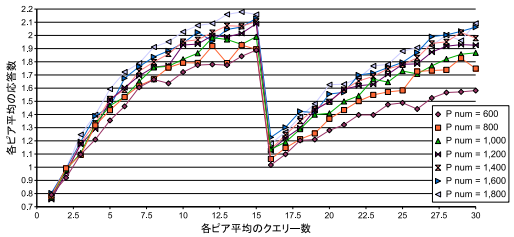


図 4　ピア数を変化させた場合（$V_{num} = 1.55$）

させた場合を示す．ここでは，P_{num} が大きいほど応答数が多いことが分かる．実験終了時に，ピアの増加に対する平均の応答数は対数関数に近い増加の様子が確認できる．

4.2.5 PSPR と通常の確率的ルーティングの比較

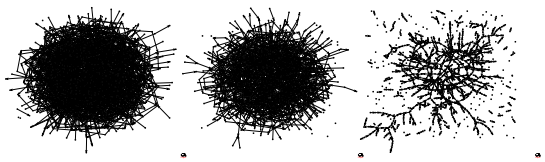
PSPR と既存の確率的ルーティングを比較した．システムは，平均クエリー数 4 回以降で差分が収束して安定する．システムが安定した状態となる“ $Q_{num} = 90,000$ ”，“ $IC_{time} = 15,000$ ”の場合と，システムが不安定な状態となる“ $IC_{time} = 3,000$ ”の場合に比較したが，PSPR の既存手法に対する変化は見られなかった．

これより，興味の変化の導入により性能が低下する可能性が考えられる．興味の変化を導入しなかった場合，PSPR は既存手法に対して平均応答数が約 0.1 高い結果となった．システムが安定した状態で PSPR が効果的であることが分かった．しかし，多くのキーワードを持つピアを早期に発見できるかにより結果に影響する可能性が考えられる．

5　ネットワークトポロジの視覚化

視覚化ツール Pajek[*2]と Kamada-Kawai のアルゴリズムによる配置によってネットワークを視覚化した．

図 5(a) は“ $V_{num} = 2.0$ ”，図 5(b) は“ $V_{num} = 1.55$ ”，図 5(c) は“ $V_{num} = 0.8$ ”の場合である．中心部分に接



(a)　(b)　(c)

図 5　ネットワークの視覚化

続が多いピアと外周に接続の少ないピアが確認できる．

図 6 は，V_{num} と P_{num} を変化させて，接続先を選択しなかったピアの数を示している．V_{num} が小さくなる

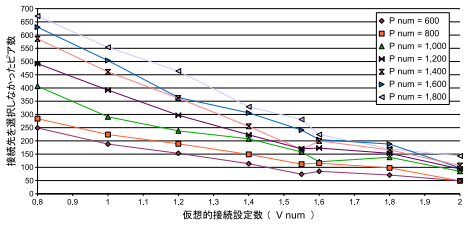


図 6　接続先を選択しなかったピア数

と接続先の無いピアの数は大きく増加している．

そして，平均ホップ数を図 7 で示す．全ての場合で，

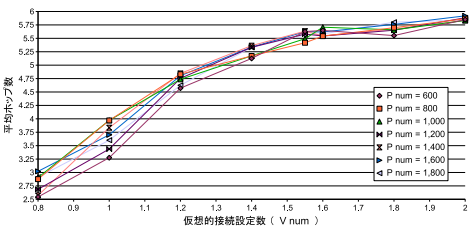


図 7　平均ホップ数

最大のホップ数は上限である 6 であった．V_{num} が小さくなると平均ホップ数は大きく減少している．

H_{num} を変化させると，“ $\sum_{n=1}^{H_{num}} Vnum^n \leq P_{num}$ ”が成り立つとき，平均ホップ数は H_{num} に収束する．そうでなければ，平均ホップ数は“ $\sum_{n=1}^{H_{num}} Vnum^n = P_{num}$ ”を満たす H_{num} に収束する．V_{num} を変化させると，$\sum_{n=1}^{H_{num}} {V_{num}}^n \leq P_{num}$ ”が成り立つとき，平均ホップ数は H_{num} に接近する．そうでなければ，最終的に経路上に存在する全ピアと接続するので，平均ホップ数は 1 に収束する．

[*2] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

6　まとめ

本研究では，類似度関数 [1] を利用して確率的ルーティングを拡張した確率的セマンティック P2P ルーティングを提案した．先行研究のピア選択アルゴリズムに適用してシミュレーションによる性能評価を行った．

ピア数が 1,000，282 種類のキーワードで，クエリー数を各ピア平均 30 回，興味の変化をクエリー数が半分となる周期に導入するとした．PSPR は，接続先を 1.55 ピアまで確率的に低減しても応答を得ることができ，トラフィック量を 67% 低減できた．また，キーワードの種類を半分にすると，接続先を 1.4 ピアまで確率的に低減しても応答を得ることができ，トラフィック量を 77% 低減できた．そして，ピア数を変化させると，ピア数が大きいほど応答数が多いことが分かった．実験終了時に，ピア数の増加に対する平均の応答数は対数関数に近い増加となった．さらに，既存の確率的ルーティングと比較すると，システムが安定した状態で PSPR が効果的であるが，興味の変化の導入により性能が低下する可能性と多くのキーワードを持つピアを早期に発見できるかにより結果に影響する可能性が考えられる．最後に，ネットワークトポロジを視覚化して考察した．これらの実験より，検索効率を維持したままトラフィック量の低減が効果的であることを示した．

PSPR をピア選択アルゴリズムに適用したが，単純な INGA[2] として一部分に対して評価を行ったので，そのシステム全体に対する評価が未確認となっている．また，PSPR の適用先の環境を良く考える必要がある．

参考文献

[1] Li, Y., Bandar, Z. A., McLean, D., “ An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources, ” IEEE Transactions on knowledge and data engineering, Vol. 15, No. 4, pp. 871-882, 2003.

[2] Loser, A., Staab, S., Tempich, C., “ Semantic Social Overlay Networks, ” IEEE Journal on Selected Areas in Communications, Vol. 25, No. 1, pp. 5-14, 2007.

[3] 森下 広史, 河野 浩之, “ セマンティックな確率的 P2P ルーティングの提案, ” 第 21 回人工知能学会全国大会 (JSAI2007), 1G1-5, CD-ROM, ISSN 1347-9881, http://www.ai-gakkai.or.jp/jsai/conf/2007/data/pdf/100025.pdf (accessed 2008.2), 2007.

[4] Risson, J., Moors, T., “ Survey of Research towards Robust Peer-to-Peer Networks: Search Methods, ” Internet informational RFC 4981, http://www.ietf.org/rfc/rfc4981.txt (accessed 2008.2), 2007.