D-041

エージェントコミュニティを利用したP2P型情報検索の検索精度評価
Agent-Community-based Peer-to-Peer Information Retrieval and Its Retrieval Accuracy Evaluation

古後 陽大 † 峯 恒憲 ‡ 雨宮 真人 ‡
Akihiro Kogo Tsunenori Mine Makoto Amamiya

1. はじめに

現在のインターネットユーザは，日々増大する情報の中から自分にとって必要な情報だけを取り出す作業に追われている．そのような作業の手間を省くため，ユーザにとって必要な情報だけを残す「情報フィルタリング」(e.g. [1]) や他のユーザの評価情報を利用してユーザにとって興味ある情報を推薦する「協調フィルタリング」(e.g. [2]) などといった研究が盛んに行われている．しかしこのような研究で開発されるシステムの多くはサーバ・クライアント型のモデルに基づいており，情報の集中制御を行う際に生じるボトルネックに悩まされている．そのため情報の共有機能をピアツーピア (以下 P2P と略記) 型のモデルで実現する研究が現在盛んに行われている．

しかしそのようなモデルの多くは，各ノードで行われる処理内容は画一的であり単純な内容であることが多い．

このような背景から我々はエージェントコミュニティを利用した P2P 型情報検索手法（ACP2P 法）を提案してきた [9, 10, 11, 12]．ACP2P 法では，各エージェントが持つデータの内容に対して検索を行うほか，他のエージェントから受けた検索履歴を基に情報の在処の特定や同じトピックに関心を持つエージェント同士のグルーピングを実現する．これにより，必要な検索結果を得るために行う通信の量を徐々に削減していくことができる．

これまでのシミュレーション実験 [11, 12] によって，ACP2P 法が仮定していた「情報の通信量削減」効果と，クエリに多く答えられるエージェントほど自分の求める情報源へのパスを増やすことができ，その結果，検索効率が向上するという「give and take 」効果があることを示した．

そこで，本稿では，ACP2P 法の検索精度を測定するための実験とその結果について，評価および考察を行う．

2. エージェントコミュニティを利用したP2P型情報検索:ACP2P

ACP2P 法では，ユーザ毎にユーザインタフェースエージェント (**UIA**)，情報検索エージェント (**IRA**)，および履歴管理エージェント (**HMA**) の 3 種類のエージェントを 1 組として割り当てる．IRA は自分のユーザが所属するコミュニティ内の他の IRA との対話を中心に，自身のユーザの求める情報の探索を行う．もしそこで見つからない場合には，階層的に辿れる他のコミュニティ所属の IRA との対話を通して情報の探索を行う．

具体的には，IRA は自分のユーザが UIA に対して出したクエリを UIA から受け取ると，そのクエリを HMA に渡し，そのクエリの検索を依頼する他のユーザの IRA（検索対象 IRA と呼ぶ）を見つけさせる．その際 HMA は，**コンテンツファイル**と，**検索結果履歴 (Q/RDH)** と**クエリ受信ログ (Q/SAH)** と呼ばれる 2 つの検索履歴を利用して，検索対象 IRA の検出を行う．その検出方法については，2.2 節で述べる．

ここでコンテンツファイルとは，自身のユーザが作成したドキュメントファイルと，検索により獲得したドキュメントファイルのことである．Q/RDH は，IRA のユーザ自身が出したクエリと，その関連情報を返してきた IRA のアドレスの対からなる．また，Q/SAH はクエリと，そのクエリを送ってきた IRA のアドレスの対と，メッセージの受信形式からなる．

必要な数の検索対象 IRA が見つからなかった場合には，コミュニティ内の全 IRA のアドレスを管理しているポータルエージェント（**PA**）に対してコミュニティ内の全 IRA にそのクエリをマルチキャストするように依頼する．

クエリを受け取った IRA は，そのクエリと関連する情報があるか否かを 2.1 節の方法で調べ，その検索結果をクエリを送ってきた IRA（もしくは PA）に返す．

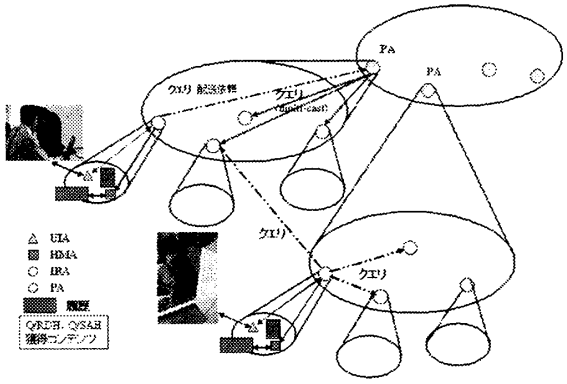


図 1: エージェントコミュニティの構造

図 1 は，ACP2P 法が仮定しているエージェントコミュニティ構造の例を示している．PA は，上位のコミュニティのメンバエージェントでもあり，コミュニティが 1 エージェントとして扱われることにより，コミュニティの階層構造を実現する．本研究では ACP2P 法を Kodama[8] を利用して実装した．Kodama の PA は，コミュニティ内のエージェントのアドレスのみを管理しているにすぎず，コミュニティ内のエージェントのコンテンツの管理等は行わない．

†九州大学大学院システム情報科学府, Graduate School of Information Science and Electrical Engineering, Kyushu University
‡九州大学大学院システム情報科学研究院

2.1 クエリとコンテンツとの類似度の計算方法

エージェントが他のエージェントからクエリ Q を受けた際，Q に関連するドキュメント D を求めるため，Q と D の類似度計算を行うが，それは情報検索において実績のある BM25[7] を修正した式 (1) において，$dl/avdl$ を 1 と近似した式によって算出する．

$$Sim_d(Q,D) = \sum_{T \in Q} w^{(1)} \frac{2tf}{\frac{dl}{avdl} + tf} \quad (1)$$

ここで，T は Q に含まれる単語である．
tf は D に含まれる T の数である．
dl は D のドキュメント長（D に含まれる単語の数）である．
$avdl$ は平均ドキュメント長である．
$w^{(1)}$ は以下の式で表される T の重みである．

$$w^{(1)} = \log \frac{N - n + 0.5}{n + 0.5} \quad (2)$$

ここで，N は各 IRA がコンテンツとして保持する全ドキュメント数である．
n は，N 個のドキュメントにおいて T を含むドキュメント数である．
$Sim_d(Q,D)$ の値が 0 より大きい値となる D を Q と関連のあるドキュメントと判断する．

2.2 クエリと検索対象 IRA との間の類似度計算

クエリ Q と，IRA $agent_j$ $(j = 1 \cdots M)$ との類似度計算式 $Score(Q, agent_j)$ を，式 (3) に定義する．ただし M は，履歴中に登録されている IRA 数である．

$$\begin{aligned} Score(Q, agent_j) = & \sum_{i=1}^{k} \cos(\vec{Q}, \vec{qh_{d_i}}) \\ + & \sum_{i=1}^{m} (\cos(\vec{Q}, \vec{qh_{sa_i}}) + \varphi(i)) \\ + & \max_{1 \le i \le n} Sim_d(Q, doc_i) \end{aligned} \quad (3)$$

$$\varphi(i) = \begin{cases} \delta & qh_{sa_i} \text{ が他の IRA から直接送られた場合} \\ 0 & \text{それ以外（PA から送られてきた場合）} \end{cases}$$

ここで，第 1 項は，Q と，$agent_j$ に出した k 個のクエリ qh_d とのスコア値であり，第 2 項は Q と，$agent_j$ が送ってきた m 個のクエリ qh_{sa} とのスコア値である．また，$\varphi(i)$ は，qh_{sa_i} が PA を経由せず，他の IRA から直接送られた場合の重みであり，本実験では $\delta = 0.1$ とする．そして第 3 項は，$agent_j$ がコンテンツの $original$ フィールドに登録されている n 個のドキュメントとのスコア値である．$Sim_d(Q, doc)$ は，クエリ Q と検索対象ファイル doc との類似度値を計算する式であり，2.1 節で述べた BM25 の簡易式を使用する．

検索対象 IRA は，$Score(Q, agent_j)$ の上位から N_R 個を取り出した $agent_j$ とする．

3. 実験

文献 [11, 12] では，ACP2P 法が仮定していた「通信量の削減」と「検索効率の向上」の効果があることを示した．また，Q/SAH の効果から，IRA 間の「give and take」効果があること，および，仮想的なエージェントコミュニティの創出についても効果があることを示した．そこで本実験では，2 つの履歴ファイルを使用することで，ACP2P 法において検索精度の向上の効果があることを確かめる．

3.1 準備

本実験でも，文献 [11, 12] と同様に検索に利用するデータとして，Yahoo! JAPAN [13] に登録されている Web サイトのコンテンツを利用した．Yahoo! JAPAN では，登録された Web サイトがカテゴリ分けされているが，ここで使用したカテゴリは，大きく分けて「動物」，「スポーツ」，「コンピュータ」，「医療」，「金融」の 5 つに分類されるものである．実験では各カテゴリから登録数の多い順に 20 個（計 100 個）のサブカテゴリを選んで利用した．各々のサブカテゴリを仮想的なユーザと見なし，それぞれに 1 つの IRA を割り当てた．またそのカテゴリに登録されているサイトから収集した Web ページを，IRA のコンテンツ（ユーザの保持する情報）として利用した．次に IRA が利用するクエリとして，長さ 1 のもの $(QL = 1)$ と長さ 2 $(QL = 2)$ の 2 つのセットを各 IRA 毎に用意した．各セットのクエリを作成するために，IRA に割り当てられたコンテンツから出現頻度の高い名詞 N 個 ($QL = 1$ の場合 10 個，$QL = 2$ の場合 5 個) を自動抽出した．$QL = 1$ の場合，抽出された各名詞をクエリとし，$QL = 2$ の場合は，抽出された 5 個の名詞から 2 個を選択し，その組み合わせにより計 10 個のクエリを作成した．実験では，100baseT の LAN で接続された Linux の動く 4 台の PC を利用し，各サブカテゴリに割り当てた計 100 個の IRA 全てを 1 つのコミュニティ上に配置した．クエリの送信は，各 IRA とも出現頻度の高いものから順に送信する．

次の三つの手法で，検索精度の比較を行った．

(1) 検索対象 IRA を見つけるために Q/RDH と Q/SAH の両方の履歴を利用し，N_R 個の検索対象 IRA を検出し検索を依頼する手法．これを**両履歴利用法** と呼び，**wQ/SAH** と略記する．

(2) 検索対象 IRA を見つけるために Q/RDH のみを利用し，N_R 個の検索対象 IRA を検出し検索を依頼する手法．これを**検索結果履歴利用法**と呼び，**woQ/SAH** と略記する．

(3) 両方の履歴を利用せず，常に PA にマルチキャストを依頼する手法．**マルチキャスト法**と呼び，**MulCST** と略記する．

マルチキャストを PA に依頼する場合，PA はある一定時間内に早く返されたものから順に N_R 選ぶ．(1)，(2) で N_R 個の IRA が見つからない場合には，PA にマルチキャストを依頼する．

3.2 検索精度の測定方法

P2P ネットワーク上で，各 Peer からすべてのコンテンツを集めて，それらをインデックス化するのは極めて困難である．そこで理想となる，すべてのドキュメントをインデックス化して行う従来の検索手法（以下 CIR：Conventional IR と呼ぶ）と，ACP2P 法とで検索結果を比較しその類似度を，ACP2P 法の近似的な検索精度の指標とした．CIR 法として 2.1 節の式（1）を各エージェントの保持するすべてのコンテンツに適用する確率型の検索モデルを使用した．

ACP2P 法と CIR 法の検索精度を比較する指標として，以下の計算式を使用した．

$$\sum_{i=1}^{N_R} \frac{1}{r(i)} / \sum_{i=1}^{N_R} \frac{1}{i} \tag{4}$$

ここで $r(i)$ は，ACP2P 法で i 番目にクエリとの関連性が高いと判断されたドキュメントの CIR 法における順位を表す．例えば，あるクエリとの類似度が，ACP2P 法において 3 番目であり，CIR 法においては 5 番目だった場合，$r(3) = 5$ ということになる．式（4）によって求められる値を **RRS(Reciprocal Rank Similarity)** と呼ぶことにし，両手法のドキュメントのランキング間の類似度を計算する．RRS 値は，N_R 個の中にランク上位のドキュメントが含まれているほど大きくなる．例として，$N_R = 3$ の下で，ACP2P 法でのランキング上位 3 個のドキュメントが，CIR 法でそれぞれ 3，5，1 というランキングだったとすると，RRS 値は $\frac{1/3+1/5+1/1}{1/1+1/2+1/3} = 0.84$ と計算される．またランク 1 位が含まれず，CIR 法で 3，5，2 が返されたとすると，$RRS = 0.67$ と小さくなる．

本実験では，RRS を全エージェントで平均したもの，すなわち $\frac{1}{N}\sum_i^N RRS(i)$ を比較指標として使用する．ここで，N は全エージェント数，$RRS(i)$ は i 番目のエージェントの RRS 値を表す．

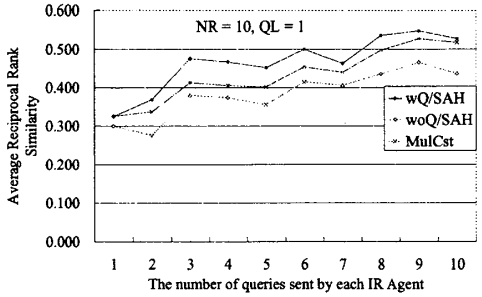
3.3 実験結果

三つの手法の RRS 値の比較を行った実験のうち，$N_R = 10$ の結果を図 2 に示す．

クエリ送信回数が増えるにつれて MulCST の RRS 値が大きくなるのは，検索を重ねることでコンテンツがコミュニティ内のエージェント間に分散し，クエリとの類似度の高いドキュメントが検索結果として返される確率が上昇するためだと考えられる．

$QL = 1$ の場合，wQ/SAH の RRS 値が他の 2 手法よりも高い数値を示しているものの，その値がそれほど高くない．この原因は，初期の検索の段階で蓄積されるコンテンツファイルおよび検索履歴は，PA がマルチキャストを行い，早く返された順に N_R 個選んだことにより得られるものであり，検索回数を重ねても，CIR 法との類似度があまり上昇しなくなるためだと考えられる．しかし $QL = 2$ の場合では，3 手法共に高い RRS 値を示している．

また，QL の値に関わらず，wQ/SAH は他の手法よりも高い RRS 値を記録していることから，Q/SAH は検索精度の向上にも効果があると言える．



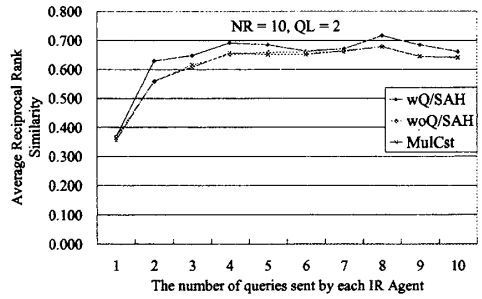


図 2: 各方式における平均 RRS 値の比較（$N_R = 10$，上から $QL = 1$，$QL = 2$）

同様の比較を，Yahoo!JAPAN に登録されている別のカテゴリ（「芸術と人文」，「生活と文化」，「教育」，「メディアとニュース」，「政治」）に対して同様の方法で行った．その結果は，図 2 とほぼ同様の結果になった．実験で利用した 2 種類のコンテンツに対して，IRA の保持するクエリが他の IRA からもどれだけ使われているかを調べると，同じクエリを保持する IRA 数とクエリ数の両対数グラフ（図 3）が類似していることが分かる．さらに，このグラフはほぼべき乗則に従っていることから，利用した両コンテンツが一般的なものであり，IRA 間で利用されるクエリがコンテンツの内容と関連のある場合，ACP2P 法の効果は一般的なコンテンツでも期待できると考えられる．

4. まとめと今後の課題

本稿では，エージェントコミュニティを利用した P2P 型情報検索手法（ACP2P 法）について，その検索精度についての評価実験の結果について報告した．実験の結果，検索履歴を利用することで，マルチキャストで単純に返答の早い順に検索結果を選ぶ場合よりも高い検索精度を得られることが分かった．CIR 手法と比較した ACP2P 法の RRS 値があまり高くなかったのは，PA によって行われるマルチキャストによって蓄積されたコンテンツファイルおよび検索履歴が影響している為である．PA は検索依頼に対して YES と返答してきた順にエージェントをアドレスリストに追加するので，良い精度は期待できない．これを解決するための方法としては，PA が全

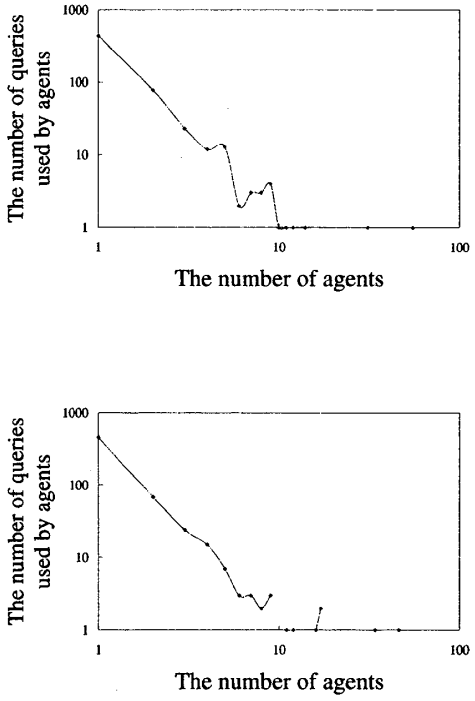


図 3: クエリを保持する IRA 数とクエリ数の両対数関係.（上：従来のコンテンツ，下：比較実験で利用した新しいコンテンツ）

IRA に対して YES，NO という答え以外にクエリとのスコアも要求し，リスト中の上位 N_R 個のスコアのエージェントを検索依頼してきた IRA に渡す方法が考えられるが，この場合は，現手法に比べて時間がかかるというデメリットがある．そこで，検索履歴にはないが関連性の強い新しいエージェントを見つけることができるような工夫を行えば，検索精度の上昇に寄与されると予想されるが，これについては今後の課題とする．

その他の課題として，階層化されたコミュニティを使用することによって，ブロードキャストの範囲を制限するというコミュニティ利用の利点を示すことや，ユーザの検索結果に対する評価情報を次回の検索に利用する手法の検討などが挙げられる．

謝辞 本研究の一部は，科学研究費 基盤研究 (C)(2)（課題番号 16500082）の支援を受けて行われた．

参考文献

[1] Ken Lang, NewsWeeder: learning to filter netnews, Proceedings of the 12th International Conference on Machine Learning, Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 331–339, 1995, citeseer.nj.nec.com/lang95newsweeder.html

[2] Badrul Sarwar and George Karypis and Joseph Konstan and John Riedl, Item-Based Collaborative Filtering Recommendation Algorithms, WWW10, 285–295, 2001

[3] Chord: A Scalable Peer-To-Peer Lookup Service for Internet Applications, Ion Stoica and Robert Morris and David Karger and M. Frans Kaashoek and Hari Balakrishnan, Proceedings of the 2001 conference on applications, technologies, architectures, and protocols for computer communications, 149–160, 2001

[4] Ian Clarke and Oskar Sandberg and Brandon Wiley and Theodore W. Hong, Freenet: A Distributed Anonymous Information Storage and Retrieval System, Designing Privacy Enhancing Technologies: International Workshop on Design Issues in Anonymity and Unobservability, http://www.doc.ic.ac.uk/~twh1/academic/, 2001

[5] Gnutella, http://gnutella.wego.com/, 2000

[6] Napster, http://www.napster.com/, 2000

[7] S. E. Robertson, S.Walker, S. Jones, M. M.Hancock-Beaulieu, and M. Gatford, Okapi/keenbow at trec-8, NIST Special Publication 500-246: The Eighth Text REtrievalConference (TREC-8), 151-162, 1999

[8] Guoqiang Zhong, Satoshi Amamiya, Ken'ichi Takahashi, Tsunenori Mine, Makoto Amamiya, The Design and Implementation of KODAMA System, IEICE Transactions INF.& SYST., Vol. E85-D, No. 04, pp. 637-646, 2002

[9] 峯 恒憲, 松野大輔, 雨宮真人, エージェントコミュニティを利用した P2P 型情報検索, 人工知能学会論文誌 J-STAGE, http://tjsai.jstage.jst.go.jp/ja/, vol. 19, no. 5, pp. 421–428, 2004

[10] Tsunenori Mine, Daisuke Matsuno, Koichiro Takaki, and Makoto Amamiya, Agent Community based Peer-to-Peer Information Retrieval, AAMAS2004, Poster, 1484-1485, 23 July 2004

[11] Tsunenori Mine, Daisuke Matsuno, Akihiro Kogo, Makoto Amamiya, Design and Implementation of Agent Community based Peer-to-Peer Information Retrieval Method, CIA 2004, LNAI 3191, 31-46, September 27 2004

[12] 峯 恒憲, 古後 陽大, 雨宮 真人, エージェントコミュニティを利用した P2P 型情報検索とその評価, 電子情報通信学会論文誌：ソフトウェアエージェントとその応用特集号, D-I, Vol. J88-D-I, No.9, pp.1-10, 2005

[13] Yahoo, http://www.yahoo.co.jp/, 2003