協調フィルタリング機能を持つ軽量 QRP の実装と評価

2001MT067 森 寿人　　2001MT111 吉田 廣志
指導教員　河野 浩之

1　はじめに

近年，P2P(Peer-to-Peer)[1] と呼ばれる技術が注目されている．P2P は，構築されたネットワーク上に存在するデータから処理能力に至るまで全てのリソースを共有，利用するための技術である．しかし，P2P はネットワークに非常に大きな負荷がかかるという問題点を抱えている．本研究では，その問題点を解決するために，JXTA[2] を用いて P2P システムを構築し，JXTA の機能である QRP を利用して協調フィルタリング [3] 機能を実装し，性能評価を行った．実験にはインターネットを用いることが望ましい．しかし，周囲への影響を考慮し NISTNET[4] によるインターネット環境のエミュレートで実験を行う．

2　P2P の問題点とフィルタリング

2.1　Peer-to-Peer

P2P は，ネットワーク上に資源を分散させ，一部に負荷が集中することを防ぎ，システム全体として対故障性を高めている．しかし，処理を分散させたことでシステムの管理が難しくなった．またネットワーク上のトラフィックが増加し，ネットワークにかかる負荷が問題となっている．後者の問題については，ネットワーク上を流れるメッセージの効果的なルーティングを行うことで軽減することが可能である．

2.2　フィルタリング

フィルタリングとは，何らかのアイテムとそれを渡すことのできる相手がいる場合に，そのアイテムまたはそれを必要としている相手を分析し，グループ化を行うことで，より効率良くアイテムを供給することを目的とする技術である．また，この技術は，無駄な情報の流れをを減らすためにも非常に有用である．

フィルタリング技術は，用いる情報とグループ化を行う対象によって 2 つに分類される．一方は，提供するアイテムをその内容によってグループ化する内容フィルタリング (Content-Base Filtering) である．他方は，第 3 者にアイテムを評価してもらい，その情報をもとに嗜好の近いユーザ毎にグループ化しグループ化を行う協調フィルタリング (Collaborative Filtering) である．

3　P2P への協調フィルタリングの適用

3.1　協調フィルタリング

協調フィルタリングを利用したシステムでは，アイテムの評価やユーザの振舞いによりユーザの嗜好や思考を分析し，似通った考え方を持つユーザをグループ化し，グループ内で人気のあるアイテムをグループメンバに推薦する．また質問や問い合わせをグループ内のみで回すことで効率良く，また効果的に情報の流れを制御することができる．協調フィルタリングの特徴として，次のものが挙げられる．

- 推薦対象の形態に関わらず適用可能
- 精度を上げるためにはより多くの対象・ユーザの情報が必要
- ユーザが評価値を誤入力するとグループ化・推薦内容に影響
- 新しい対象は評価情報が集まるまで推薦不可

3.2　ピアソン相関係数

情報推薦や，協調フィルタリングにおいて，ユーザの嗜好の類似度を求める尺度としてしばしばピアソン相関係数が用いられる．AnneDoucet らの研究 [5] でも，ピアソン相関係数が取り上げられている．相関係数は，対象を一定の基準で数値化する 2 種類の尺度があり，複数のアイテム (要素) について評価した場合に，二者間の基準の相関を求める手法である．

ピアソン相関係数は，式 (1) で与えられ，その値は式 (2) に従う．これは，ユーザ x，y について，x と y の共分散を x の標準偏差と y の標準偏差の積で除算したものである．式中で，x_i はユーザ x のアイテム i に対する評価値を表し，n はアイテムの総数を表す．そして，$\bar{x}$ はユーザ x の全ての評価の平均値を表す．係数は，1 に近付く程両者の考え方は似通っており，-1 に近付く程両者の考え方は正反対になる．係数が 0 のとき，両者の考え方の間に関連性は全く無い．

$$r_{xy} = \frac{\sum_{j=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{j=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

$$-1 \leq r_{xy} \leq 1 \quad (2)$$

3.3　協調フィルタリング機能の実装

本研究では，JXTA を用いて P2P に協調フィルタリングシステムを実装した．JXTA は, TCP/IP 上で P2P ネットワークを構成するためのプロトコル群を含む API である．JXTA は，オープンソースであり，Java プログラムであるため，機能の変更・追加が容易である．また，メッセージが XML 形式であるため，機能拡張がより簡単である．

実装したシステムを Junius とし，コンテンツ検索と共有の機能を有する．

Junius の概要を図 1 を用いて説明する．図中のランデブーピアは, 他のピアからのリース要求を待っている．ランデブーピアは初期情報として，自身の持つ共有可能

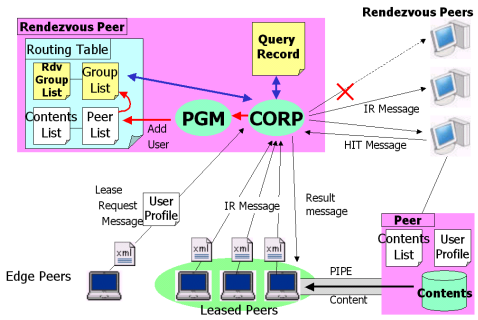

図 1　Junius のシステム概要

なコンテンツをリスト化した Peer List を持つ．また，図中の CORP は Collaborative Organize and Retrieve Protocol を略したもので，メッセージを受け取り，その内容を元にピアの登録，グループ化からメッセージのルーティングを行う．CORP には，QRP の機能が含まれている．

エッジピアは，初期情報として自身の持つ共有可能なコンテンツ (Contents) とそのリストである Peer List，そして既に動作しているランデブーピアの IP アドレス，ポート番号を最低 1 つ持つ．エッジピアは起動時にユーザから嗜好情報の入力を求める．方法は，各項目について興味の度合を数値で入力してもらう．ユーザからの入力を元にユーザプロファイルを作成する．その後，ランデブーの 1 つにリース要求，コンテンツリストとユーザプロファイルを含むメッセージを送信する．ランデブーピアは受け取ったメッセージを CORP から PGM(Peer Group Manager) に渡す．そこでピアの情報を Peer List と Contents List に追加し，ピアソン相関係数を用いて最も相関の高いグループに分類する．係数が一定の値 (閾値) に満たなければ新たなグループを生成し Group List に追加する．そしてリース提供メッセージをエッジピアに送信し，エッジピアがこれを受け取ることでエッジピアのネットワークへの登録が完了する．

登録が完了し，リースを提供されたエッジピア (Leased Peer) はランデブーピアに対して検索要求を送ることができる．検索要求を受け取ったランデブーピアは，JXTA でメッセージルーティングを行う QRP にあたる CORP によって検索メッセージを転送する対象を判断，転送する．その検索結果 (該当するものがあればそのリスト) がエッジピアに返され，ユーザの判断で受け取ったリストの中からコンテンツを提供してもらうピアを選択し，その相手との直接通信でコンテンツを受け取る．

ランデブーピア内の Query Record には検索要求の内容とその結果が格納される．Query Record は CORP での処理に利用される．また, Rdv(Rendezvous) Group List には Group List のグループの概要 (各グループの平均嗜好情報) が格納され，他のピアと接続した際にユーザプロファイルの代わりに送信される．これにより，ランデブーピアは他のランデブーピアの管理するグループを自身のグループと比較分類し，他のピアと同じようにルーティングの際に利用する．

4　仮想実世界ネットワーク上での検証

4.1　実験環境の構築

室内に PC8 台による LAN を構築し，これをコアネットワークとした．コアネットワークに属する PC には外部ネットワークへ接続するためのインターフェースが用意されており，ここから他のネットワークに接続することができる．本研究では，このコアネットワークに PC を接続することにより実験を行った．

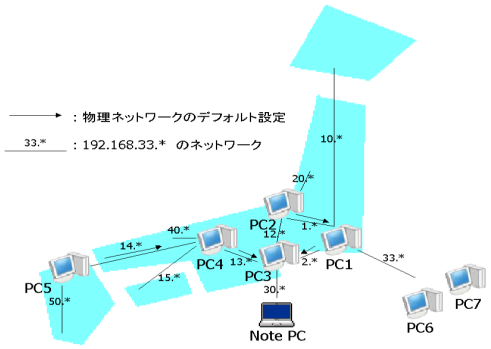

図 2　実験環境のネットワークトポロジー

図 2 が本研究で利用したネットワークの概要図である．PC1〜PC5 にはネットワークカードが複数枚搭載されており，NISTNET というアプリケーションがインストールされている．NISTNET は，稼働させることにより帯域幅やパケットロス，伝送遅延といった回線環境を制御可能なルータとして動作する．また，伝送中のパケットを検出する機能もあるため，本研究のトラフィック検出にも NISTNET を利用している．

実験用ネットワークは日本列島をモデルとしており，特徴として，PC1，PC2，PC3 によるループがあること，物理的に PC1 にトラフィックが集中することがある．これは実際のインターネットの環境に近づけるためであり，NISTNET の機能で東京ー名古屋間の応答時間を近似する，というように利用する．

ピアはネットワーク全体で 20 動作しており，1 つのピアが 10 個のコンテンツを所有し，公開している．全てのピアはランデブーとして起動させてあるためルーティングにおける制限は無い．リレーとプロキシーの設定は行っていない．

4.1.1　実験 1：協調検索機能の効果検証

Junius によるネットワークを構築し，JXTA プロジェクトの一つである CMS(Contents Management System) の API を用いた検索を 10 秒に 1 回の割合で 10 分

間行い，発生したトラフィックを PC1〜PC4 で計測した．検索文字列はランダムに生成される．

その後グループを構築し，協調フィルタリングの機能を実装した検索を行った．ピア数が 20 という制限があるため，今回の実験ではグループを構築できるように嗜好情報を設定した．その結果 3 つのグループが生成され，本研究ではグループ数を 3 として実験を行った．

4.1.2 実験 2：グループ数によるトラフィックへの影響の検証

Junius によるネットワークを構築し，グループ数によるトラフィック変化を計測する実験を行った．

また，本ネットワークは小規模のため，キャッシュを利用した検索を行うとネットワーク上のコンテンツをほとんどローカルキャッシュに集めてしまい，トラフィックを計測する実験に支障が出る可能性が有る．そのためコンテンツに関するアドバタイズメントにはキャッシュを用いない設定で実験を行った．

トラフィックの計測方法は以下のとおりである．

- 乱数により生成された文字列を検索文字列としたクエリーを，図 2 の NotePC から 5 秒に 1 回送信する．
- 30 秒間に PC1〜PC4 で通過したトラフィックを記録する．
- 上記を 30 分間繰り返す．

グループ数を 3 とした状況でトラフィックを計測した後，グループ数が 1 になるように嗜好情報を変更し，ピアアドバタイズメントを公開した．各ピアでローカルキャッシュに保存されたピアアドバタイズメントの更新作業を行った後，上記の方法でトラフィックを計測した．

4.2 実験結果

4.2.1 協調フィルタリングによる検索効率の向上

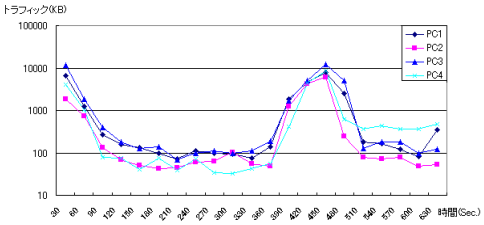

図 3　CMS 検索時のトラフィックの変化

図 3 はフィルタリングを適用しない通常の検索である．全てのピアが最上位グループである NetPeerGroup に属しているため，ピア全てに検索クエリーを送信していると考えられ，そのことは図中に示した各 PC での計測結果に裏付けられている．全ての計測点において同じタイミングでほぼ同量のトラフィックが流れていると考えられる．

それに対して協調検索の場合を示したグラフである図 4 では，変化するタイミングは同じであるが，各計

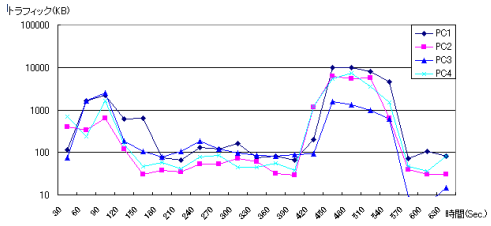

図 4　協調検索時のトラフィックの変化

測点で異なる量のトラフィックを検出している．計測開始から 8 分後あたりにある大きな変動では，PC3 のトラフィックは他の検出点の 10 分の 1 まで低減されている．これはフィルタリングによりメッセージの送信先に偏りが生じたためであり，発見に対する応答が無いためより大きな差として現れている．また，グラフ全体を見ると，必ず 1 つ以上の計測点ののトラフィックが CMS 検索や他の計測点と比べて目に見えて小さな値を計測している．これはフィルタリングの効果が一時的なものではなく，常にトラフィックの低減に効果があることを示している．

実験後に検索によって得られたコンテンツを調べたところ，どちらの検索方法もネットワークのほとんどのコンテンツを入手していた．これは検索文字列が短かったこと，ピアの所持するコンテンツ名が全てアルファベットの文字列であったこと，ネットワークが小規模なためほとんどのピアのコンテンツアドバタイズメントが共有されてしまったことが原因として考えられる．コンテンツの入手量としてはグループ全体にメッセージを送信する CMS 検索の方が多いかもしれないが，重複したコンテンツを考えた場合，目的のコンテンツを入手する確率としては同等であると言える．

以上により，帯域消費が少なく，かつコンテンツの入手率で同等の結果を示した協調検索法は検索の効率が高いと言える．

4.2.2 グループ化によるトラフィックの低減

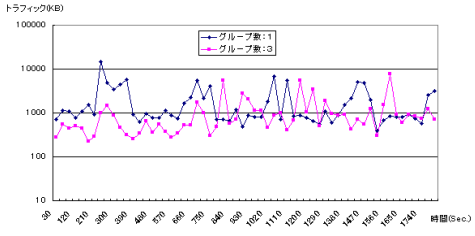

図 5　グループ数によるトラフィックの変化

図 5 は PC1〜PC4 で 30 秒ごとに計測し，30 秒間に発生したトラフィックの総和をグラフにしたものである．840 秒まではグループを 3 つに分けた方が一方的にトラフィックが少ない．これは，クエリーを送信した

ピアの嗜好値に近い嗜好値を示すグループにのみ，メッセージが転送されているためであると考えることができる．クエリーの条件にあうコンテンツのアドバタイズメントを発見すればメッセージの転送は終了するため，少ないホップ数で確実にコンテンツアドバタイズメントを発見することができた結果と言える．

しかし，それ以後にも単一グループの場合よりもトラフィックを抑えることができているが，以前よりもはっきりとした差は見られず，時に大きなトラフィックを発生させている．これは目的のコンテンツアドバタイズメントを発見することができず，最終的に全てのピアにメッセージをを転送する結果になったためであると考えられる．しかしながら，最もトラフィックを抑えることができている下方部に注目すると，低い値を出しているのはいずれもグループ数が 3 つの場合であり，その頻度も単一グループと比較して多くなっている．

5 協調フィルタリングの適用に対する考察

フィルタリング技術をグループ化に適用するシステムには，グループ外に存在するコンテンツを取得する方法が必要である．本システムでは検索結果によって，次に嗜好値が近いグループへメッセージを転送しているが，これだけでは近い嗜好情報を持つグループを 1 つにまとめたにすぎない．メッセージは，複数ではあるが特定のグループ間を常に転送されることになり，グループ外の有用なコンテンツに対してのアプローチとしては依然不十分である．

この問題を解決するには定期的にブロードキャストでメッセージを送信すればよく，発見したコンテンツ情報を保持しておくことにより，グループ外のコンテンツ情報をグループ内でも共有することができる．また，被参照量が多く，ピアの嗜好情報に近いコンテンツを発見した場合，そのコンテンツをダウンロードして自分の保持するコンテンツとして公開することにより，ピアが所属するピアグループにとって有用なコンテンツをミラーリングすることができる．ミラーリングによる検索効率向上からトラフィックのさらなる低減が見込まれ，嗜好情報に近いコンテンツを共有することになるためピアの嗜好情報の信頼性が向上する．

本研究でのグループ化はピアの発見順に行ったが，グループ化の手順自体にも改善が必要である．発見したピアの順序が異なる場合，それらのピアの嗜好情報が同一であったとしても全く同じピアグループを構成するとは限らず，常に最適なグループを構成しているという保証は無い．また，構成されたグループ数が多すぎる，あるいは少なすぎる場合，フィルタリングの効果は小さくなることがわかっている．グループ数が最適化されていない場合でも総トラフィックはブロードキャスト以下に抑えることが可能であるが，レスポンスメッセージを受信してから他のグループへメッセージを送信する処理が発生するため，検索結果を得るのにより多くの時間を必要とする．このためグループを最適なメンバで構成するための最適化処理が必要であり，システム稼働中にはグループを定期的に最適化しなければならない．

また，グループ化に用いる嗜好情報にも注意をはらう必要がある．初期値はユーザの入力を用いるため，未入力や入力間違いが最初に所属するグループに直接影響し，その影響はグループに所属する他のピアにも及んでしまう．グループの最適化のためには検索ログ等からピアの本当の嗜好を発見し，嗜好情報を最適化する必要がある．未入力の嗜好情報を持つピアに対しては最初の検索単語からグループを決定する方法も考えられる．

さらに，本研究で行った実験は小規模であることを考えておかなければならない．例えばインターネットで 10 万のピアが稼働した場合，この実験結果をそのまま延長した結果が得られるとは限らない．しかしグループを再構成し，今回検証したサイズまでグループの規模を縮小することにより，結果を改善することができると考える．ネットワーク規模に対し最適なグループのサイズと数を発見することも今後の課題である．

6 おわりに

本研究では，P2P のネットワークトラフィックの問題に着目し，その問題の解決のためにピアソン相関係数を用いた協調フィルタリングを P2P システムに組み込み，検証を行った．そして，トラフィックを低減しつつ検索力を維持することに成功し，その実効性を示した．

しかしながら，グループ化の処理においてグループの最適なサイズと個数を検討する必要があり，グループ数を最適にするための相関係数の閾値や，嗜好情報に関するパラメータをいかに設定するかという問題が今後の課題として残っている．

参考文献

[1] Andy Oram, “PEER-TO-PEER Harnessing the Power of Disruptive Technologies”, O’REILLY(2001).

[2] Sun Microsystems, “ProjectJXTA”, available from <http://www. jxta. org/>, (accessed 2004-9-10).

[3] Peng Han, Fan Yang, Ruimin Shen, “A Novel Distributed Collaborative Filtering Algorithm and Its Implementation on P2P Overlay Network”, PAKDD 2004, Sydney, Australia, pp. 106-115.

[4] NIST(National Institute of Standards and Technology), “NISTNET”, available from <http://snad.ncsl.nist.gov/itg/nistnet/>, (accessed 2004-9-10).

[5] Anne Doucet, Nikolas Lumineau, “A Collaborative Approach for Query Propagation in Peer-to-Peer Systems”, SWDB 2003, Berlin, Germany, pp. 251-257.