知能と複雑系　124−7
（２００１．５．１７）

# ピアツーピアネットワークにおけるトピック主導型検索手法の提案

中辻　真[†]　○川原　稔[††]　河野浩之[†]

ピアツーピアネットワーク環境下の情報交換システムとして，Gnutella 等のアプリケーションが活発に開発されつつある．しかしながら，ファイル名による分散型検索であるため，適切な検索結果を実用的な検索応答時間を保証しながら得ることは難しい．そこで，クライアント上のファイルの内容やネットワーク性能等を考慮した検索機能の提供を目標に，検索対象となる各ファイルの特性を示すインデックス方式を導入する．そして，各ホスト中の複数のインデックスファイルに基づいたヒューリスティックな関数を与えることによって，適切な検索ホスト選択を行うトピック主導型検索を提案する．なお，小規模なプロトタイプシステムを実装することで，提案したシステムの性能評価を行った．

# A Proposal of Topic-Driven Resource Discovery System over Peer to Peer Network

Makoto Nakatsuji,[†] Minoru Kawahara[††]
and Hiroyuki Kawano[†]

Many network applications over peer to peer network, such as Gnutella, has been developed drastically. However, by using distributed systems with simple search functions, it is difficult to discover suitable information resources and to guarantee practical response time. Therefore, in order to realize good network performance and advanced functions with content-based search, we introduce the index system showing the characteristic of each file. We propose the topic-oriented search mechanism, which selects optimal search hosts, based on heuristic functions with several indices in each host. We evaluate the performance of our proposed system on a small-scale prototype system.

## 1. はじめに

現在，クライアント・サーバ型の情報交換システムではなく，P2P(ピアツーピア) ネットワーク環境下において，個対個や，グループ対グループで情報資源の交換を実現するクライアント・クライアント型のシステムが注目されている [3]．そして，この種の情報交換システムは，パーソナルコンピュータ上の比較的小規模な蓄積データを対象とするだけではなく，広範囲に分散する膨大な数の Web サーバが保持する多様なデータをも統合的に扱う可能性を備えており，情報流通機構を実現する基盤となる広域分散型データベースシステムとして興味深い．

もっとも，現在の Gnutella 等の情報交換システムは，単純なファイル名検索機能を分散環境下で提供する程度であり，様々な問題を抱えている．事実，適切な検索結果を得ることは難しいし，ネットワーク環境下でばらつきが存在する検索応答時間の性能保証も課題となっている．

そこで，著者らが検索エンジン「問答」の実証実験 [9] などを通じて得た，データベース，情報検索，ネットワークなどに関わる技術 [10] を生かしながら，P2P ネットワークにおける情報流通に関わる技術課題の解決を試みる．特に，本稿では，P2P ネットワーク上の検索機構に焦点を当てて，ファイル名とパス名のみによる検索機能ではなく，情報資源に付随する複数の属性に基づいたトピック主導型検索手法を提案する．

本稿で提案する手法は，検索対象となる各ファイルの特性を抽出し，各ホストごとに蓄積されているデータに対するインデックスファイルを構築する．そして，ネットワーク上の多数のホストに格納されるインデックスファイルを効率良く利用しながら，検索者の要求に応じたトピック主導型検索を実行する．つまり，P2P ネットワーク環境下に広域分散する，異なる性能や性質を備えた多数のホストの全インデックスファイルを対象とした検索は，現実的に不可能である．そこで，トピック主導型検索により，問合せに応じた適切なイン

† 京都大学大学院情報学研究科
Department of Systems Science, Graduate School of Informatics, Kyoto University
†† 京都大学大型計算機センター
Data Processing Center, Kyoto University

デックスファイルをもつホスト群を絞り込む機能を提案する．なお，ホスト群の絞り込みには，検索式に含まれるキーワード群と各ホストのもつインデックスファイルの類似性などを利用する．また，Web クローリングアルゴリズムで提案されている分類子 (classifier) と抽出子 (distiller)[2] を用いる．

以下，2 章では，本研究の背景となる P2P ネットワークにおける典型的な検索モデルについて述べる．3 章では，P2P ネットワークにおけるトピック主導型検索機能の基本的構成を提案する．4 章では，トピック主導型検索を行う分類子の有効性を確かめる実験を行い，その結果と評価を述べる．そして，5 章において，抽出子を用いてインデックス交換を行う分散型インデックスシステムの簡単な議論を行い，6 章の結論と将来の課題で結ぶ．

## 2. P2P 環境下の情報資源検索

P2P ネットワーク上のサービスとして，音声データに対してファイル名をサーバで検索し，ファイルの実体を個々のパソコン同士で交換させる Napster (http://www.napster.com)，音声データに加えて画像データの交換を対象とする Scour (http://www.scour.com) などがある．また，ファイル情報もサーバに集中せず，完全なクライアント・クライアント型システムとして構築された Gnutella，Freenet (http://freenet.sourceforge.net)，FreeHaven (http://www.freehaven.net) などがあり，情報流通の構造を変化させる枠組みとして注目されている [3),6),8)]．実際，Napster では，常時，約 8 千人程度のユーザが接続しながら，3.2TB 程度の約 80 万ファイルを交換している．また，Gnutella では，約 2 万人のユーザが，534TB 程度の 190 万ファイルを共有した状態になっており，P2P 形態の接続利用が急激に増加している．

もっとも，多くの P2P 上のサービスは，著作権の存在するファイルを大量に交換する状況を招いているため，Napster は RIAA (全米レコード工業会，http://www.riaa.com) 等から，Scour は MPAA (アメリカ映画協会，http://www.mpaa.org) から訴訟を起こされている．さらに，著作権問題以外にも，P2P システムでは，R 指定や X 指定といった公開が制限されるファイルのダウンロードが誰でもできることに対する危機感も高く，社会的問題は山積している．加えて，集中的なファイル管理システムが無い事によって，ウイルスが蔓延する可能性も懸念されている．

しかし，多くの社会的問題と別に，P2P ネットワークは，情報流通を促進する負荷分散を図る基盤技術として非常に重要である．したがって，既存の P2P システムにより明らかになりつつある技術的問題，例えば，クライアント・サーバ間の検索処理，クライアント・クライアント間の検索処理やネットワークトポロジーの最適構成，非常に頻繁な ping 発行によるネットワーク帯域の圧迫などを解決する必要がある．

なお，現在の P2P システムに共通する大きな問題として，ファイル名・パス名による検索機構しか備えないため，利用者の意図した検索と結果が全く異なる場合が多い．そこで，本稿では，これまでのクライアント・サーバ間の Web 検索支援システムにおける検索技術の成果を踏まえ，クライアント・クライアント間の検索処理について考察する．

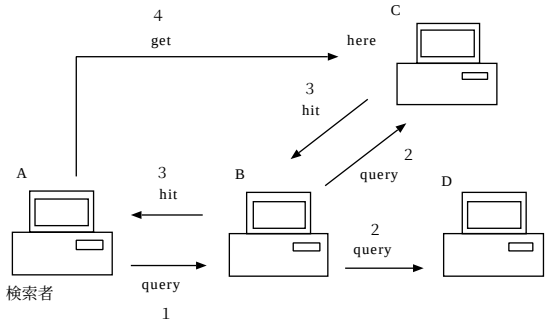
以後，クライアント・クライアント間の基本的動作を前提とした議論を進めるために，典型的 P2P システムである Gnutella の仕組みを簡単に説明する [7)]．

まず，Gnutella の基本的な動作は，図 1 に示すように，自分の知っている他のホストにコマンドを送り，その応答を待つという手順で動作する．そして，コマンドを受け取ったホストは，結果を返送し，更に他のホストへとコマンドを送信する．その繰り返しにより，ホストとファイルの情報を収集する．なお，コマンドを送信するパケットに中継回数を制限する TTL (Time To Live) を与え，中継されるたびに値を 1 ずつ下げ，ネットワークの巨大化を抑制する．また，一度受け付けたコマンドの重複受け付けを防止する識別子として GUID (Globally Unique IDentifier) を与える．

具体的な手順は，新規ホスト A がホスト B に対する TCP 接続を開始し，他ノードを検索するための GUID，TTL を含む PING コマンドを B へ送る．B は IP アドレスなどのホスト情報からなる PONG コマンドを A へ返答する．これにより，B がアクティブか否か，どのようなファイル公開を行っているかが分かる．そして，PONG コマンド送信後 TTL 値を 1 下げ，TTL 値が 0 でなければ PING コマンドを B の接続ホスト C,D へ送る．次に，ホスト C,D は，GUID により重複状態を判定し，続く接続ホストへは PING コマンドを，B に PONG コマンドを送る．ホスト B は他のホストから送られてきた PONG コマンドをホスト A へと転送し，その情報を元に A から C への接続が生じる．

## 3. P2P におけるトピック主導型検索システム

Gnutella で用いる GUID と TTL を用いて構築し




(1) ホスト A は直接接続しているホスト B へ，Query コマンドを送る．
(2) Query コマンドを受け取ったホスト B は，既に自分が受け取ったコマンド であれば破棄し，そうでなければ，自分の接続するホストに Query コマンドを送る．
(3) 条件に合致するファイルをもつホスト C は，Query HIT コマンドを，Query コマンドが送られてきた経路に沿って逆向きに送り返す．
(4) A は C に対して直接接続を行い，ファイルを要求する Get コマンドを送り，ファイル転送を行う．

図 1 ファイル検索とダウンロード手順

た P2P ネットワーク上で，ファイル内容を効果的に検索するトピック主導型検索機構を提案する．

### 3.1 データ抽出とインデックス構築

トピック主導型検索の実現のために，ネットワーク上に存在する各ホストが蓄積しているファイルの諸特性を記述したインデックスファイルの構築を考える．例えば，ファイル名や拡張子，ファイルサイズや日付，ファイル作成方式などの数多くの属性を，インデックス化できる．もちろん，半構造データ [1] である XML, HTML ファイルなどを対象とするならば，参照ファイルのアンカー情報や，ハイパーテキストへの参照ファイル数などを利用することもできる．以下，蓄積されたファイルの種類に応じて，どのような属性をインデックス化できるかを簡単に示す．

【テキストファイル】

Web 検索システムと同様に自然言語処理技術を用いた単語抽出を行う．そして，検索式に含まれるキーワードとして単語集合を含む時，単語集合の単語がテキストファイルに表れる頻度，ファイルサイズ，文書構造などの総合的評価によって，あるホスト $h_i$ 内のテキストファイル $f$ に対するスコア $S$ を与える．なお，この種のスコアとして，例えば，TF-IDF 法 [4),5)] や，テキストマイニング技術 [12)] で研究されている様々な手法を用いることができる．TF-IDF 法を用いるならば，問合せに対するスコアは次式のように与えられる．

表 1 MP3 ファイルのタグ情報

| ファイル形式 | タグ情報 |
|---|---|
| MP3 | トラック番号，タイトル，アーチスト，アルバム名，西暦，ジャンル，コメント，作曲，Orig. アーチスト，著作権，URL，エンコーダ |

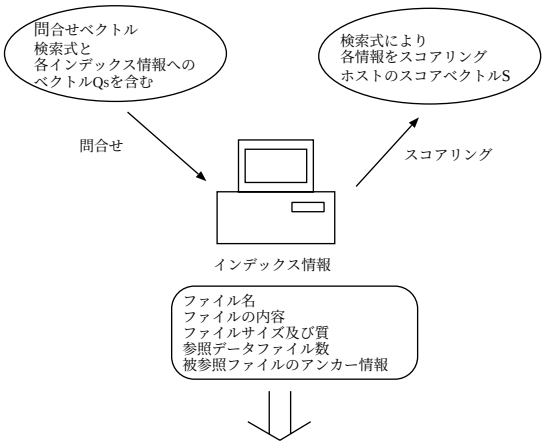



図 2 問合せとホストに対するスコア

$$S = n(d,t) \cdot \log \frac{|\cup_{f \in F(h_i)}|}{|\cup_{f \in F(h_i)}\{t \in f\}|}$$

$n(f,t)$ : キーワード $t$ がその対象ファイル $f$ 中に含まれる出現回数

$F(h_i)$ : ホスト $h_i$ に含まれる全ドキュメント数

【オーディオファイル】

マルチメディアファイルは，将来的には MPEG7 で標準化されるメタ情報などを利用することが適切である．ただし，現時点では，表 1 に示す MP3 ファイルのタグ情報などを用いて，オーディオファイルに属性を付与する．

### 3.2 インデックスに対するスコアリング

利用者の入力した検索式を元に効率的なファイルの絞込み検索を実現するために，それぞれのホストのもつインデックスファイルに対してスコアリングを行う．

例えば，利用者による問合せが，図 2 に示すように入力される．この時，利用者は各ホストのもつインデックスファイル（ここでは,「ファイル名 $(f_i)$，記述内容 $(c_i)$，サイズ $(q_i)$，参照データファイル数 $(d_i)$，被参照ファイルのアンカー情報 $(r_i)$」の属性をもつとする）に対する重みベクトル $\mathbf{Q_S} = (f, c, q, d, r)$ も同時に与える．よって，問合せ $\mathbf{Q}$ は，式 (1) のベクトルとなる．

$$\mathbf{Q} = (t, \mathbf{Q_S}) \quad (1)$$

また，各ホスト $h_i$ に対するインデックスファイルに対するスコアベクトル $\mathbf{S_i}$ を式 (2) で表す．

$$\mathbf{S_i} = (T(f_i), T(c_i), T(q_i), T(d_i), T(r_i)) \quad (2)$$

ここで，各ホストのもつインデックスファイルに対する重みを考慮する関数 $T$ として，平均 $\overline{x}$，分散 ${\sigma_x}^2$ をもつ式 (3) を用いる．

$$T(x_i) = \alpha \cdot \frac{x_i - \overline{x}}{\sigma_x} + \beta \quad (3)$$

次に，利用者の与える各インデックスファイルへの重みと，各ホストのもつインデックスファイルによって，ホストに対するスコア $S$ を与える．ここでは，各ホスト $h_i$ のインデックスを用いたスコア $S$ は，重みベクトル $\mathbf{Q_S}$ とホスト $h_i$ のスコアベクトル $\mathbf{S_i}$ の内積により，式 (4) で与る．

$$S = \mathbf{Q_S} \cdot \mathbf{S_i} \quad (4)$$

そして，3.1 節の議論に従って，各ホストの持つファイルのインデックスファイルを用いてスコアを決定する．一例としてファイル名と記述内容に対するスコアリング手法を述べる．

【ファイル名に対するスコアリング法】

入力された検索式 $t$ に含まれるキーワードが，ホスト $h_i$ に格納されているファイルのファイル名 $f$ の文字列中に含まれる状態を $\{t \sim f\}$ で表す．また，ホスト $h_i$ における全ファイルを $F(h_i)$ とする．この時，ホスト $h_i$ において検索式に対するファイル名の表れる頻度 $f_i$ を式 (5) で与える．

$$f_i = |\cup_{f \in F(h_i)} \{t \sim f\}| \quad (5)$$

ネットワーク内の全ホストに関して式 (5) で得られた $f_i$，その平均と分散を計算し，ホスト $h_i$ のファイル名に対するスコアを，式 (3) を用いて得られる $T(f_i)$ とする．

【記述内容に対するスコアリング法】

ホスト $h_i$ 内のあるファイルの記述内容に，検索式 $t$ の構成要素が出現するという情報を $t \in f$ で表す．また，ファイル $f$ の記述内容に対する検索式 $t$ の構成要素の出現頻度を $n(f,t)$ とする．そして，各ファイルの記述内容に対する検索式 $t$ の構成要素の出現頻度を計算し，ホスト $h_i$ 内の全ファイルに対して和をとったものが，式 (6) の $c_i$ で与えられる．

$$c_i = |\cup_{f \in F(h_i)} \{n(f,t) \cdot log\frac{|\cup_{f \in F(h_i)}|}{|\cup_{f \in F(h_i)} \{t \in f\}|}\}| \quad (6)$$

式 (6) により得られる $c_i$ の平均と分散をネットワーク内の全ホストに関して計算し，式 (3) に代入して得られる $T(c_i)$ をホスト $h_i$ の記述内容に対するスコアとして与える．

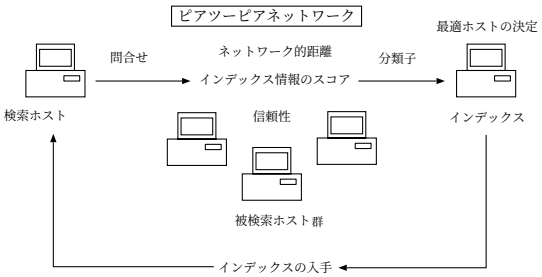


図 3　分類子による最適ホスト選択法

以上，各ホストのもつインデックスファイルに対するスコアベクトルと各属性に対する重みベクトルを与えることで，ホストのもつインデックスファイルのスコアが決定される [11]．

3.3　分類子による最適ホスト選択

本節では，リソース発見機構を備えた Web サーバに対するクローリング手法の研究 [2] で用いられている分類子の考え方を発展させ，最適なインデックスファイルをもつホスト選択への応用を提案する．

最適なファイルを持つホストであるか，また，ファイルを入手するのに最も適したホストであるかを分類する分類子を，以下に述べる条件などを考慮しながら設計する (図 3)．

【インデックスに対するスコア】

利用者が検索したいコンテンツと共通領域の多いファイルを多く蓄積しているホストが，隣接ホストとして望ましい．ここで，問合せに応答したホスト数を $k$ とする．そして，ネットワーク内の全ホストに対して，ホスト $h_i$ に対するスコア $S_i$ を計算し，そのスコアの分布を用いて，各ホストのインデックスファイルの適切さを表す関数 $T(S_i)$ を求める．

$$T(S_i) = \frac{S_i - \overline{S}}{\sigma_S} \quad (7)$$

【ネットワーク距離の利用】

P2P ネットワークにおけるファイル交換を実行する際，ネットワーク的により近いホスト間でコネクションを張る方が望ましい．ここで，ネットワーク距離 $R$ は，ホストのファイル処理能力や転送速度，さらに，ホスト間の回線速度に依存する値と考えて良い．したがって，ネットワーク距離の値は，ホストからファイルをダウンロードするファイル転送速度 $B_i$（KB/s）を用いることで評価できる．そこで，検索元ホスト $h_0$ とホスト $h_i (0 \leq i \leq k)$ とのホスト間のリモート距離

$R_i$ を式 (8) で与える．

$$R_i = \frac{1}{B_i}(0 \leq i \leq k) \tag{8}$$

次に，問合せに合致するホスト $h_i(0 \leq i \leq k)$ に対する $R_i$ を計算し，スコア分布を考慮して，各ホストのリモート距離の適切さを，つぎのように評価する．

まず，問合せに合致した全ホスト $h_i(0 \leq i \leq k)$ に対する $B_i$ の平均 $\overline{B}$ と分散 ${\sigma_B}^2$ を求める．

これらの値を用いて，検索ホスト $h_0$ から対象ホスト $h_i$ のリモート距離情報の適切さを評価する $T(B_i)$ が，式 (9) で求まる．

$$T(B_i) = \frac{B_i - \overline{B}}{\sigma_B} \tag{9}$$

ところで，分類子を構成するにあたって，検索者がインデックスファイルの適切さを重視するのか，それとも，ネットワーク距離の適切さを重視するのかという評価関数 $f(X)$ が必要である．よって，分類子を与える式 $A_i$ を，ホスト間のファイル転送率 $B_i$ と，対象ホストの蓄積しているファイルの適切さ $S_i$ と，利用者による重み関数 $f(T(X))$ を用いて，式 (10) のように表す．

$$A_i = T(S_i) \cdot f(T(B_i)) \tag{10}$$

すなわち，式 (10) により計算される値が大きいホストほど，検索者から見たホストの適切さが高いということであり，ホストの適切さを用いた分類が実行できる．

【検索履歴の利用】

分類子を用いるごとに，式 (10) により求まるホストに対するスコアを，各ホストが記憶する．$k$ 回前の検索において，式 (10) を用いた結果求まったホスト $h_i$ の適切さのスコアを ${P_i}^{(k)}$ で表す．更に，過去の検索時のホストのスコアよりも，より新しい問合せに対して決まるスコアを重視するためのパラメータ $\theta(0 \leq \theta \leq 1)$ を与える．ここで，過去 $n$ 回の検索履歴を残しているとして，ホストの適切さを表す過去の検索結果の情報も含む分類子を表す式 $P_i$ を次式により定義する．

$$P_i = \frac{1}{n} \cdot \sum_{j=0}^{n} \theta^j P_i^{(j)} \tag{11}$$

なお，検索者は分類子を用いながら，最適なファイルを選択し，そのファイルをもつホストのインデックスファイルを要求する．また，他のホストのもつインデックスファイルの転送要求も出しながら，検索の実行を進める．

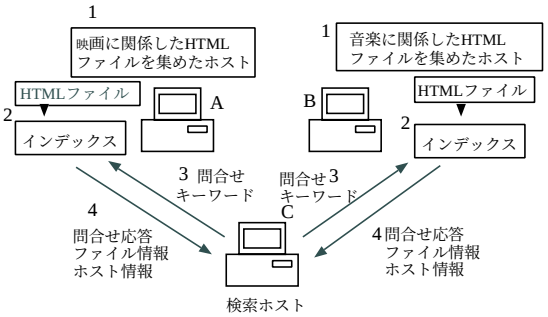



図 4　トピック主導型検索システムにおける検索手順

## 4. トピック主導型検索手法の実装と性能評価

本章では、提案手法の有効性を簡単に評価するために，3 ホストからなる P2P ネットワークを構築し，HTML ファイルを対象とするトピック主導型検索を行うプロトタイプシステムを実装する．そして，検索式に合致するホスト内のファイル数とホスト間のネットワーク距離を用いた分類子を設計し，ホストの絞り込み機能の実装可能性についての評価と検討を行う．

### 4.1 トピック主導型プロトタイプシステム

本稿で提案した検索モデルの性能を確かめるために，図 4 に示すように，3 ホストのうち，2 ホストを被検索ホスト (A, B) とし，1 ホストを検索元ホスト C とする．そして，被検索ホストに対する特徴として，ホスト A は映画情報に関する HTML ファイルを中心に格納し，ホスト B は音楽情報に関する HTML ファイルを中心に格納する．なお，それぞれのホストに格納した HTML ファイルは，表 2 に示した URL などからリンクされる下位 10 階層のページであり，Web ページ収集プログラム Gethtml (http://hp.vector.co.jp/authors/VA014425/gethtml.html) を用いて収集した．

実験は，利用者の入力した検索式を他のホストに転送し，検索式を受取ったホストのもつインデックスファイルを利用して検索を実行し，ファイル情報とホスト情報からなる応答を検索元ホストへ返さねばならない．そこで，各ホストに格納した全ファイルから簡単な記述内容を抽出するために，全文検索システム Namazu[13] (http://www.namazu.org/) を用いてインデックスファイルを生成する．さらに，P2P ネットワークにおけるリソース発見ソフトである gnut (http://www.umr.edu/ jjp/) に，本稿で提案したトピック主導型検索を実現する分類子を実装する．

また，ホストの適切さを決定するために，各ホストのもつインデックスファイルの適切さ，および，検索ホストからのネットワーク距離の 2 つの要素を用いた

表 2　比較対照ホストの蓄積データの収集元の URL (一部)

| ホスト | URL |
|---|---|
| ホスト A | http://www.geocities.com/eventmovies/godzilla.htm |
| | http://www.geocities.com/dragonlady255/princessmononoke.html |
| | http://www.geocities.com/coner_m2000/index.html |
| ホスト B | http://www.gbv.com/ |
| | http://www.superchunk.com/ |
| | http://www.sebadoh.com |
| | http://www.bernardbutler.com/ |

分類子を次のように設計した．

簡単のため，各ホスト $h_i$ のインデックスファイルのスコア $S_i$ を，問合せに応答したホストのファイル数 $H_i$ を用いて評価する．したがって，ホスト $h_i$ のインデックスファイルに対するスコア $T(H_i)$ は，式 (7) を用いて次式のように与えることができる．

$$T(H_i) = \frac{H_i - \overline{H}}{\sigma_H} \tag{12}$$

次に，ネットワーク距離は，検索に応答したホスト $h_i$ から，直接ファイルをダウンロードする際の転送速度 $B_i$(KB/s) の値により表す．そこで，検索ホスト $h_0$ から，対象ホスト $h_i$ のリモート距離を表す値 $T(B_i)$ として次式で与える．

$$T(B_i) = \frac{B_i - \overline{B}}{\sigma_B} \tag{13}$$

また，ホストの適切さを評価する分類子は，式 (14) で表される．そして，$P_i$ の最大値を取るホストが最適ホストとなる．

$$P_i = T(H_i) \cdot T(B_i)^{\theta} (0 \leq \theta) \tag{14}$$

なお，ここで用いた $\theta$ は，ホスト選択の適切さに対するネットワーク距離の重みを変化させるパラメータであり，ネットワーク距離をどの程度考慮するかを調整する値である．

4.2　実験結果と性能評価

検索式に含むキーワードとして，映画により密接な関係のある単語を採用した場合，式 (14) により，どちらのホストが適切なホストとして選択されるかという実験を行った．

まず，第一の問合せとして “scream”（映画のタイトル）を与えた場合に，$\theta$ の変化とホスト A，B に対する評価値を図 5 に示す．

図 5 は，ホスト A のヒット数が 14，転送率が 9.94(KB/s)，B のヒット数が 8 で，転送率が 11.77(KB/s) という結果に基づいている．そして，$\theta$ が 0 に近いときは，式 (14) によりホスト A が最適ホストとして選ばれる．実験結果は入力した検索式に対する予想と合致する．また，$\theta$ が 1，つまり，ヒット数で表されるインデックスの適切さと，ネットワーク距離の重みが等しい時グラフが交差する．これは，この検索において，ホストのヒット数によるインデックスの適切さの分布と，ネットワーク距離の望ましさの分布がホストに関して対称であるためである．また，$\theta$ が 1 より大きいときはホスト B が式 (14) により最適ホストとして選ばれることを，グラフから読み取ることができる．

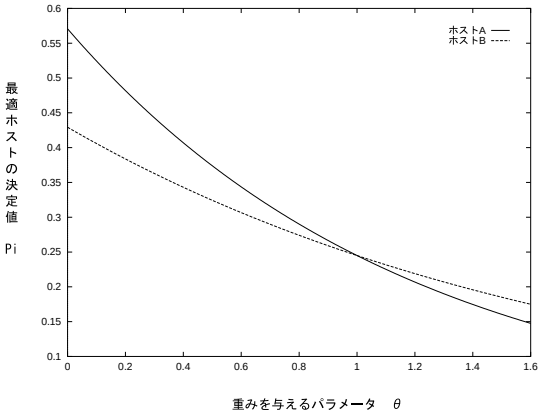


図 5　$\theta$ の変化とホスト A,B に対する決定値の変化 (キーワード “scream”)

次に，第二の問合せとして “pearl”(バンド名の一部) を与えた場合の $\theta$ の変動に伴うホスト A，B の適切さの評価値の変化を図 6 に示す．実験結果として，ホスト A のヒット数が 5，転送率が 10.67(KB/s)，B のヒット数が 26 で，転送率が (10.79KB/s) という結果を得た．この結果に基づくと，$\theta$ の変化によらず，常にホスト B が適切なホストとして選択されることになる．

最後に，ヒット数が全く同じで，転送率が，ホスト A が 9.94(KB/s)，ホスト B が 11.77(KB/s) と仮定して $\theta$ を変動させた場合，ホスト A，B の適切さの評価値が，どのように変化するかを示すグラフを図 7 に描いた．これによると，$\theta$ が 0 の時，つまり，転送率をホストの選択の考慮に入れない場合，どちらのホストも同じスコアをもつが，$\theta$ を大きくすると，転送率の早いホスト B が選択されることがわかる．

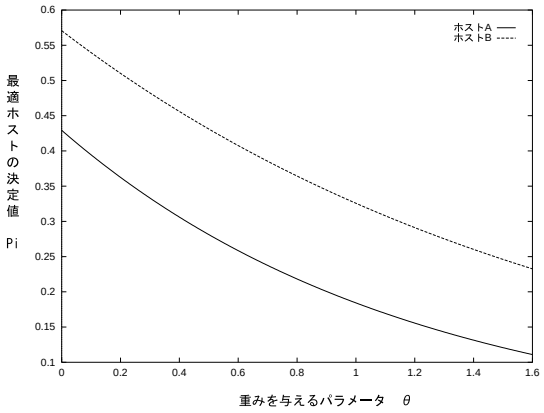



図 6 $\theta$ の変化とホスト A,B に対する決定値の変化 (キーワード “pearl”)

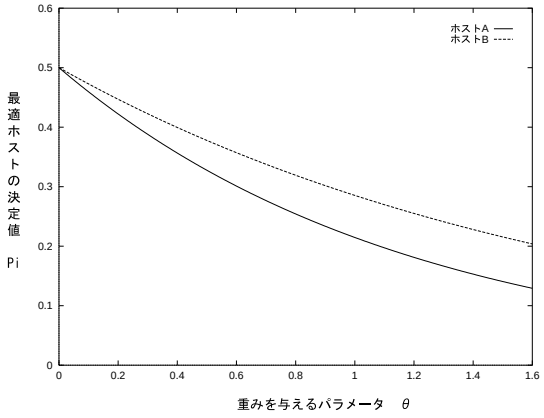



図 7 $\theta$ の変化とホスト A,B に対する決定値の変化 (ヒット数が等しい場合)

以上の実験結果から，我々が提案する式 (14) は，ホストの適切な選択に有効であることを示せた．ただし，今回の実験のような状況では，ファイルサイズの小さい場合のスコアは，ホスト間のネットワーク距離よりもヒット数の方が重要と考えられる．また，今回は，$\theta$ をユーザが調整する値としたが，実際のシステムでは，対象とするネットワークに対応して動的に $\theta$ の値を変化させるモデルを考える必要がある．

## 5. 分散型インデックスシステム

前章までに述べた分類子を用いると，利用者にとって最適なインデックスをもつホスト選択が実現できる．そして，さらに，そのホストのインデックスの全体，もしくは，該当部分をダウンロードしてくれば，利用者のホストに全てのファイルを格納することなく，他のホストのインデックスを利用できる．また，様々な検索を繰り返すことによって分類子を用いながら，多くのインデックスファイルを利用者が収集した状況を考えると，各ホストごとに蓄積される複数のインデックスファイルは，高い関心を示す領域を多く含む．

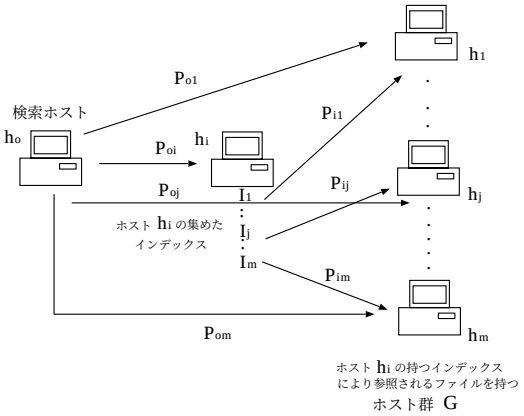



図 8 抽出子を用いた最適ハブホストの選択

そして，このようにしてホスト内に集積されたインデックスが，検索者の関心のある領域と重なり合う場合は，そのホストのもつ他のインデックスも検索者にとって興味深い内容が含まれる可能性が高いと推測できる．つまり，検索者は，より適切なインデックスを多数蓄積したハブとしての役割を果たすホストを発見することで，より効率的な検索が可能となる．そこで，以下に，ハブホストを発見するための抽出子について簡単に述べる．

あるホストが，適切なインデックスファイルをどれだけもつかを示すパラメータをハブスコアと呼ぶ．まず，図 8 のように，ホスト $h_i$ が，ホスト群 $G$ で示した他ホストのもつファイルへのポインタを含むインデックスファイルを獲得しているものとする．但し，適切な幾つかのインデックスファイルと，不適切な多数のインデックスファイルをもつ場合もあることを考慮して，適切なホストを抽出する必要がある．

そこで，ホスト群 $G$ に含まれるホストのうち，ホスト $h_i$ から見て，適切さの高いホスト $m$ 個を対象とする．すなわち，ホスト $h_i$ のハブスコアは，ホスト群 $G$ に含まれる検索ホスト $h_o$ からみた適切さのベクトル $\mathbf{S_O}$ と，$h_i$ からホスト群 $G$ に含まれるホストへの適切さのベクトル $\mathbf{S_I}$ を用いて，式 (15) を求める．なお，ホストの適切さ $P$ の決定は，先に述べた式 (11) を用いる．このようなハブスコアを用いて，適切なインデックスファイル群をもつ可能性の高いホストの抽出を行う．

$$\frac{\mathbf{S_O} \cdot \mathbf{S_I}}{m} \tag{15}$$

$$\mathbf{S_O} = (P_{o1}, P_{o2}, \ldots, P_{oj}, \ldots, P_{om})$$

$$\mathbf{S_I} = (P_{i1}, P_{i2}, \ldots, P_{ij}, \ldots, P_{im})$$

## 6. 結論と今後の課題

本研究では，P2P ネットワーク環境下で，各ホストが蓄積しているファイルから属性データを抽出したインデックスファイルを構築し，利用者の検索式に含まれるキーワードに対するインデックスのスコアや，ネットワーク距離に基づき最適なインデックスファイルをもつホストを評価しながら，トピック主導型の検索絞り込みを行う手法を提案した．更に，小規模なネットワーク上のプロトタイプシステムを用いた実験により，HTML ファイル検索に対する手法の有効性を示した．さらに，複数のホストから適切なインデックスファイルを抽出する手法についても簡単な議論を行った．

今後の課題として，3.1 節で述べたインデックスファイル構築などがある．さらに，P2P ネットワーク上のセキュリティやプライバシーの観点から，ホストの信頼度情報等を考慮して，実システムで利用可能な分類子の構築も必要である．さらに，抽出子の有効性を検証すると共に，トピック主導型検索システムにおける P2P 上の分散インデックス更新などの研究を進める必要もある．

謝　　辞

本稿の一部は，文部省科学研究費 (11130211, 12780278, 12792015) の研究成果による．

## 参 考 文 献

1) Abiteboul, S., Buneman, P. and Suciu, D.: *Data on the Web*, Morgan Kaufmann Publishers Inc. (2000).
2) Chakrabarti, S., van den Berg, M. and Dom, B.: Distributed Hypertext Resource Discovery through Examples, *Proc. of the 25th International Conference on Very Large Data Base*, pp. 375–386 (1999).
3) Clark, D.: Face-to-Face with Peer-to-Peer Networking, *Computer, IEEE*, pp. 18–20 (2001).
4) Ellis, D.: *New Horizons in Information Retrieval*, The Library Association, London, UK (1990). [エリス: 情報検索論：認知的アプローチへの展望, 細野 公男 監訳, 丸善株式会社, (1994)].
5) Schäuble, P.: *Multimedia Information Retrieval, Content-Based Information Retrieval from Large Text and Audio Databases*, Kluwer Academic Publishers (1997).
6) OpenP2P: *O'Reilly P2P Directory*, O'Reilly and Associates, Inc., http://www.oreillynet.com/pub/q/p2p_category (2000-2001).
7) 井口圭介: ネットワーク管理者のための Gnutella 入門, デジタルアドヴァンテージ, http://www.atmarkit.co.jp/fwin2k/experiments/Gnutella_for_admin/Gnutella_for_admin_0.html (2000).
8) 上村圭介: ファイル交換ソフトウェアの行方, *Glocom Review*, Vol. 5, No. 10, 国際大学グローバル・コミュニケーション・センター (2000).
9) 河野浩之, 川原稔: Web 検索におけるテキストマイニング, 人工知能学会誌, Vol. 16, No. 2, pp. 212–218 (2001).
10) 西尾章治郎, 田中克巳, 上原邦明, 有木康雄, 加藤俊一, 河野浩之: 情報の構造化と検索 (岩波講座マルチメディア情報学 8), 岩波書店 (2000).
11) 中辻真: ピアツーピアネットワークにおけるトピック主導型検索手法の 提案と評価, 京都大学工学部情報学科数理工学コース卒業論文 (2001).
12) 人工知能学会: 特集：テキストマイニング, 人工知能学会誌, Vol. 16, No. 2, pp. 191–238 (2001).
13) 馬場肇: 日本語全文検索システムの構築と活用, ソフトバンク (1998).