2004－DBS－134　(I)　(27)
2004／7／13

P2P ネットワークにおけるコンテンツの多様性を考慮した検索手法の提案

羽多野　一磨[†]　大　島　裕　明[†]
是　津　耕　司[††,†]　田　中　克　己[†,††]

近年 P2P ネットワークが注目されて久しい．P2P ネットワークはサーバクライアントモデルに比べて，情報共有が容易なモデルであると考えられる．しかしながら，そこでの情報検索は検索の高速化という側面においては研究が盛んであるものの，情報検索という面においてはファイル名とクエリとのマッチングという原始的な手法が主であり，十分とはいえなかった．このため本論文では P2P ネットワークにおける検索についての問題点を発見し，従来より柔軟な検索手法について提案する．

Proposal of the Retrieval Technique on Peer-to-Peer Network considered with multi-attribute of contents

Kazuma HATANO,[†] Hiroaki OHSHIMA ,[†] Kouji ZETTSU[††,†]
and Katsumi TANAKA[†,††]

Recently, P2P network has been drawing attention. It is easier to share information on P2P network model than on server-client model. However, almost all researches study on improvement in the speed of retrieval. On the aspect of the information retrieval, matching file name with query is major, but it is very primitive and not enough to retrieve information efficiently. So, we find the problems of information retrieval on traditional P2P network, and we propose the system which provides more flexible information retrieval than one in the past.

1. はじめに

近年，P2P（Peer-to-Peer）ネットワークが注目されて久しい．Peer とは「対等な」・「同僚」といった意味を持つ語であり，ネットワーク上の一つのノードのことを指す．”P2P ネットワーク ”とは，複数のピアとそれらの間を結ぶいくつかのリンクで構成されたネットワークである．

現在，この様なネットワークを利用したアプリケーションとしては音楽ファイルや映画ファイルを主な対象とした Napster[5]，Gnutella[4] 等が上げられる．こういったネットワークではユーザがサーバに情報をアップロードする必要が無く，このような種類のファイルに限らずコンテンツを共有する際に有効な技術であると考えられる．

† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
†† 独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology

そのため P2P ネットワーク上に存在する，音楽や映画と言ったファイルに限らない汎用的なコンテンツを有効に扱うためには，ファイル名のみを対象とした従来手法とは異なる，より高度な検索技術が必要であると考えられる．本論文では，このような従来の P2P ネットワークにおける検索についての問題点を洗い出し，より柔軟な検索手法について提案する．

2. P2P 検索の現状と課題

Ｐ２Ｐネットワークを検索するに当たって，考慮すべき問題は以下の三点であると考えられる．

（1）どのように他のコンピュータを発見するか
（2）どのように検索するのか
（3）どのように結果を見せるのか

これらの点について Napster を例に，P2P ネットワークにおける検索技術の現状について考察する．

2.1 Napster

P2P ネットワークにおける情報検索の例として Napster の仕組みについて説明する．Napster は 1999 年

1月に発表されたインターネットを通じて個人間の音楽データの交換を行うためのアプリケーションである．

Napster では Napster 社が管理する中央サーバが存在する．このネットワークに参加するユーザはこの中央サーバに mp3 形式の音楽ファイルのリストを送信する．これを世界中のユーザが共有する事によって互いに他のユーザの所持している音楽ファイルを検索し，ダウンロードする事が出来る．

中央サーバはファイル検索データベースの提供とユーザの接続管理のみを行っており，ユーザは中央サーバにクエリを投げ，中央サーバがこれに対する検索結果を返すことになる．結果を見たユーザはそこから適当なファイルを選択する事によって，そのファイルを所持しているユーザと直接接続し，音楽データ自体のやり取りを行うことになる．

先の要素に当てはめると

(1) 中央サーバの把握している全てのピアから
(2) ファイル名にクエリキーワードを含むものを
(3) ヒットした順に一覧としてユーザに見せる

となる．

後発の Gnutella や Winny においては検索に中央サーバを一切用いないものになっているが，ファイル名やディレクトリ名がクエリのキーワードに含まれているもののインデックスをユーザに返す点においては同様である．図 1 は，Web 検索・Napster 型の P2P 検索・Gnutella 型の P2P 検索についてそれぞれの検索の流れを図式化している．図中の数字はいずれの方式においても 1 が検索クエリ，2 が検索結果，3 がコンテンツ要求で，4 がコンテンツの転送となっている．

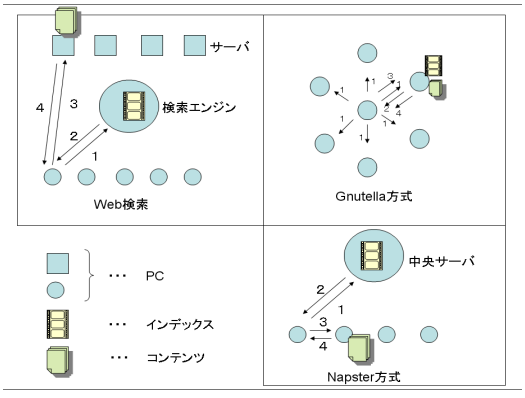

図 1 様々なコンテンツ検索

2.2 関連研究

CAN[7]，Chord[3] はファイル名の様なある一つの属性に対してハッシュ関数を適用することによって，高速な検索手法について提案している．これらの研究はハッシュ化を利用するために，ある特定の属性を持ったコンテンツを検索するためには適している．

RDFPeers[2] では RDF リポジトリを MAAN 上で作成することを目標としている．MAAN は Chord をコンテンツの多義性に応じて改良したものであり，情報が膨大になる RDF リポジトリを分散管理する方法について提案している．

REMINDIN[1] では，ソーシャルメタファーを用いたクエリのルーティング手法について提案している．ここでは対象データを RDF の三つ組として論じているが，実際にそのデータがどのように付加され，どのようなファイルを対象としているのかは述べられていない．この研究においてはクエリのルーティング手法に関する提案が主ではあるが，本研究ではこのような適当に形式化されたメタデータの付与は難しいと考えている．

湯川ら[6] は P2P ネットワークを通じた情報共有について提案している．ここでは小さなコミュニティ内での共有を想定しており，個人間の概念の相違を考慮した検索手法を提案している．

3. 問題点の考察

Napster や Gnutella ではファイル名・ディレクトリ名のみを対象として検索するという非常に単純な構造ではある．しかし，よく流通していた映画や音楽ファイルといったコンテンツの検索においてはこういった手法でも良好な結果が得られる．これは対象としているファイルが非常に明確であり，そこへのファイル名付けもタイトルやアーティスト名といったものが共通して付けられていたためである．しかし，こういった手法ではコンテンツ検索としては不十分である．例えばこういったものではなく，個人の撮影した写真などを対象として考えると，ファイル名だけで検索するのは難しく，また，メタデータの記述法を形式化するのも難しい．

本論文では P2P ネットワーク上に配置された多くのコンテンツの中から対象コンテンツを検索することを目標としている．先のような音楽・映画と言った，明確に一意に決定できるようなキーワードがファイル名に付けられるといった暗黙の了解のみではこういった検索は不可能であると本研究では考える．

そこで先にあげた，

(1) どのように他のコンピュータを発見するか
(2) どのように検索するのか

（3）どのように結果を見せるのか

について,（1）は 3.1,（2）については検索対象にどのようなメタデータが付けられているか（3.2），クエリに対してどのようなプロセスを行うか（3.3）,（3）については 3.4 で，それぞれ問題点を明らかにする．

3.1 関連ピアの発見

ピュア型の P2P ネットワークである Gnutella ではネットワークに参加する際に決定されるピアと静的に接続し，そのようなネットワーク上で全てのピアに対してクエリを投げ，それがフラッディングしていく．このようなシステムはピア数の増加と共にクエリが指数的に増大し，問題があった．

一方，Winny ではクラスタワードというユーザの興味を表す語を予め指定し，これの近いユーザを P2P ネットワーク上で近くするとともに，ネットワーク帯域によってツリー構造を構成する事によって効率化を図っていた．

このようにネットワーク上の全てのピアの情報を把握するのが難しいため，ユーザの興味に合わせてネットワークの構成やクエリの伝播方法を変更する事が必要である．ユーザが求める情報を発見するためには次のようなユーザにクエリを投げるのが効率的と考えられる．

- ユーザと全体的に興味が似ている
- ユーザが現在興味を持っている事に詳しい

これらの指標となるのは，所持しているコンテンツの内容や Web の閲覧履歴などが考えられる．このようなものが類似しているユーザ同士は求めるものも類似している可能性が高く，効率的な検索が可能になると考えられる．

3.2 メタデータ

コンテンツ検索において，機械的に検索を行うためのメタデータが必要である．メタデータとはデータにデータに対するデータであり，Napster・Gnutella ではファイル名・ディレクトリ名をメタデータとしていた．

しかし，これだけではメタデータとして十分とは言いがたい．一般的にファイル名は自分にとってどんな情報かを判別出来るためだけに付けられる情報であるためである．検索のための，より多くの情報を付加する必要があると考える．

メタデータを記述する方式としては RDF（Resource Descripiton Framework) が知られている．RDF はメタデータの表現方法についての枠組みである．RDF の利点は subject・predicate・object の三者関係によって，対象を記述する．

例えば Open Directory Project[8] は全ての Web ページについてカテゴライズを行っており，5年もの期間をかけて記述した，その RDF ファイルは 1 GB 以上に及んでいる．

これを効率的に利用するためには対象に対するどういった情報を示しているのかを共通項目として予め決定していなければならない．例えば Dublin Core[9] では幾つかの predicate が定義されている．これを音楽ファイルならば音楽ファイル用の RDF スキーマを，映画ファイルなら映画用スキーマを定義するといった形で拡張し，共通的に定義する事によって非常に柔軟に，また，機械が可読な形でメタデータを付加する事が出来る．

しかし，こういった属性を全てのコンテンツについて定義するのは難しい．属性と言うものは様々考えられ，追加・修正が頻繁に行われる必要があるためである．更に，こういった分類に従って多くの情報をユーザが入力するのはユーザにとって大きな負担となる．

一方でこういった属性分類を行わずにメタデータを記述する方式も考えられる．先ほどの RDF では，

『コンテンツは属性〜を持っており，その値は〜である』

といった記述であったが，

『コンテンツは値〜を持っている』

と言った場合である．つまりキーワードを羅列するだけの方法である．この方式であると，記述が容易になるが，各値の意味が分からない．

この様なメタデータの取得方法としては以下のようなものが考えられる．

- ユーザによる手動入力
- 機械的な入力

ユーザの手動入力は最も原始的だが，最も適当なメタデータを作成する事が出来ると考えられる．しかし，ユーザが全て行うのはあまりに手間が大きい．このためコンピュータによる自動的な入力方式が求められる．

これを実現する手法の一つとして，まずコンテンツ自身から抽出する方法が考えられる．例えば音楽ファイルについて考えると，CDDB を用いて音楽 CD にタイトルやアーティスト名を自動的に付加する事が出来る．更に，ユーザのディレクトリ内のファイル間にはある程度の関連がある，と言う仮定に基づいてディレクトリ内のファイル群から特徴を抽出する方法も考えられる．また，同様のコンテンツを所持しているピアのメタデータを参考に付加するような事も考えられる．

本節以降は，簡単のため単純にキーワードの羅列をメタデータとして付けているとする．

3.3 クエリの処理

ユーザがメタデータとして書くキーワードと言うものは様々である．このため単純なキーワードの一致だけで決めるのは不十分である．例えば，あるユーザが ”富士山 ”とインデックスに書いたファイルを持っているユーザが，クエリ ”河口湖 ”を受けた際，従来のキーワードの包含のみを見るとこれはヒットしない．しかし，”富士山 ”と ”河口湖 ”の移った写真においても，ユーザによってはどちらが主題かは変化し，メタデータを作成する際には違いが出てくる．このため ”富士山 ”と ”河口湖 ”が関連があり，クエリを投げたユーザの意図に ”一致しているかもしれない ”というような解釈を行う必要があると考える．

このため ”確信度 ”という概念を導入する．これは各ピアにとって，その答えがどの程度，クエリにヒットしている確率が高いと考えているかを表す尺度である．例えば ”日本一高い山 ”と ”富士山 ”が全く同じものを指していると思えば，クエリ ”日本一高い山 ”に対して ”富士山 ”というメタデータの付いたコンテンツは確信度１として返す事になり，一方，クエリ ”河口湖 ”に対してはもっと低い値になる．これに関しては 4.1 で詳しく説明する．

更に，河口湖がある程度関連すると言うことを知っていれば，これをクエリに付加して他のピアに送る事も考えられる．これによりクエリを受けたピアは ”河口湖 ”が ”富士山 ”と関連していると言うことを知らなくても，変換されたクエリを利用して返答する事も可能になる．場合によってはクエリからキーワードを削除したりする事も考えられる．

また，従来の手法ではメタデータ（ファイル名）に，検索キーワードが全て含まれているものをユーザの検索意図に適合したものとして，そのインデックスを返していた．しかし，キーワードが部分的に一致している情報も，完全に不必要なわけではなく，ユーザにとって有用な情報を部分的にしろ含んでいる可能性がある．

3.4 結果の提示

従来の手法では結果はファイル名や回線速度などがリストとして単に羅列されているに過ぎなかった．例えば図 2 は Napster の検索画面である．Napster では対象とするファイルが音楽ファイルであるため，欲しいアーティスト名やタイトルを入れる事によって不要が紛れ込んでくると言う事は少なかった．このため，回線速度等を見てユーザがダウンロードするファイルを選択するような形でも問題は無かった．しかし，これでは自分が欲しい情報を選択するのに効率的とは言えない．

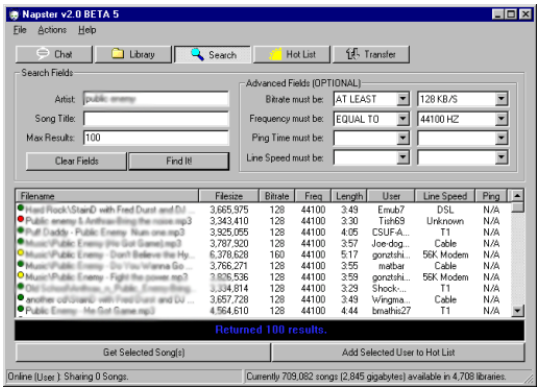

図 2 napster の検索結果表示画面

このためユーザの検索結果をユーザの意図にマッチしていると思われる順にランキングを行ったり，また，メタデータに含まれる内容によって適当にクラスタリング，視覚化する必要性があると思われる．

4. クエリ処理手法の提案

前節のような問題点のうち，本論文では特にクエリに対する処理という点について，提案する．

ピアがクエリを受け取った際に考えられる対応は以下のような要素が挙げられる．

- クエリの解釈（関連キーワードの拡張など）
- クエリの適合評価
- 返答
- クエリの伝播

まず，クエリが一体どういうものかを判断しなければならない．これは例えば ”富士山 ”というクエリを受け取った場合に,『 ”富士山 ”といえば ”日本一高い山 ”だ』，というような言い換えである．このようにピアがまず自分の持っている情報を持ってクエリに処理を加える．

このような変換の方法としてユーザのローカルコンテンツにおけるキーワードの共起度を利用する．ユーザが所持しているローカルコンテンツは，ユーザの意図を持って収集されたものであり，これらにはユーザの意図が反映されていると考えられるからである．

クエリを受け取った際，クエリに含まれる各キーワードに対して共起語を算出し，共起度が閾値より高いものを関連語と見なし，クエリ変換の候補とする．また，キーワードの共起度に応じて，後に述べる確信

度の値を変更する．検索クエリを投げたユーザに返答する場合，コンテンツのインデックスにこの確信度を付加して送り，ユーザに提示する際のランキングに反映する．

このように算出して作成されたクエリキーワードを用いてコンテンツを評価するに当たっても単なるキーワードの包含だけで意図したコンテンツが検索できるわけではない．キーワード間の関連も考慮し，かつ，部分一致も含めた検索が必要である．また，伝播させるのにも単に伝播するだけでは不十分である．クエリをそのまま送るのか，あるいは変換したものを送るのかなど様々考えられる．これに関しては後の 4.2 節で述べる．

各ピアの処理の流れとしては図 3 のような形になる．

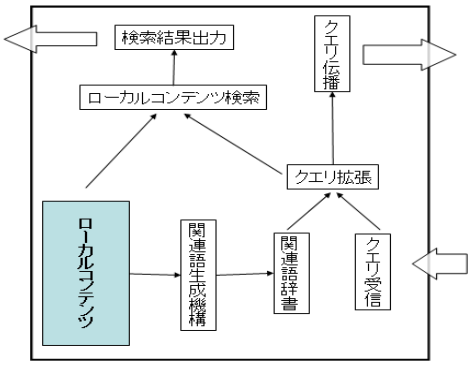

図 3　各ピアでの処理の流れ

4.1　確信度の提案

ユーザが求める情報は，クエリに表されるキーワード群によって全て表されるわけではない．こういった事に関して Web 検索においてもキーワード拡張について研究されてきた．また，本論文の提案するシステムもこういった理由と共に，自らの保持しているコンテンツに対して付けたメタデータと，他人が同じコンテンツに対して付いているであろうと推測するキーワードは異なる事が考えられる．

そこで，ユーザがクエリに対して適当な解釈を行ってコンテンツを評価を行う必要があると考える．しかし，こういった変換が正しく行われるとは限らない．クエリに表されたキーワードというものはユーザの意図を完全とは言えなくとも，およそ適当に表していると考えられるが，それに何らかの解釈を加えた際にはクエリが表す範囲は変化する．クエリにユーザの意図が多く反映されていることを考えると，解釈後に生成されたキーワードと言うものは元のクエリに比較してユーザの意図からずれている可能性が高い．図 4 はクエリの変換候補にクエリを変換した場合の適合率のおよその様子である．

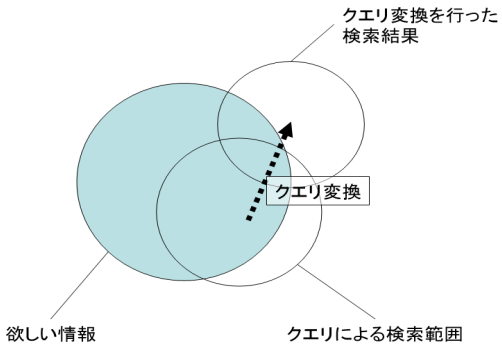

図 4　クエリ解釈による確信度の変化

そこで，検索ユーザの意図に沿っているかどうか，についてどの程度確信を持っているかを表す尺度を導入する必要があると考える．各ユーザが独自の判断も加えてユーザに返答するため，どの程度確実にクエリにマッチしていると思われるかを表す指標が必要であるからである．

ユーザが変換前のキーワードと変換後のキーワードがほぼ同値として捉えている場合は確信度は 1 に近い．一方で幾らか関連している，キーワードを概念的な上位語に変換する，と言った場合にはこれは 0 に近づく事となる．

4.2　クエリ処理プロセス

クエリに対する対応を考える場合，ローカルコンテンツの検索結果に応じて様々考えられる．検索結果の質が非常に良い場合は，それ以上検索クエリを伝播させる必要はないと考えられ，一方であまりよくなかった場合は変換したキーワードも含め伝播させる必要がある．更に全く自分の知らないキーワードについてクエリを投げられた場合は何も処理する事が出来ない．

さて，検索結果の評価について述べる．検索結果の質を述べるにあたって次の二つの要素が重要であると考える．

- ヒット数
- 確信度平均

ヒット数はどの程度の結果を出せたかという指標であり，確信度はどの程度ユーザの意図とマッチしているものを返す事ができたかと言う指標になる．これらの値によって図 5 の様に四つに場合わけを行う

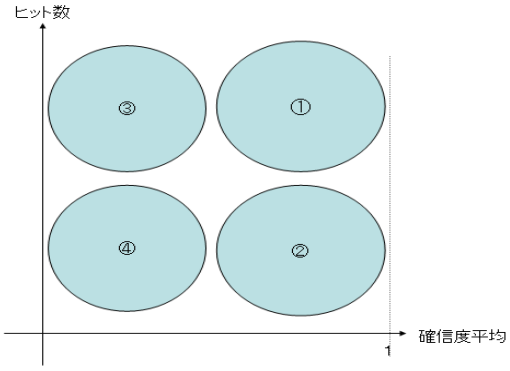

図 5　検索結果の場合分け

4.2.1　　：共に高い

確信度・ヒット数共に高い場合，ユーザのクエリに対して十分な返答が出来ていると考えられる．このためこれ以上の伝播の必要はないと判断し，返答を返すのみに留まる．

4.2.2　　：ヒット数のみ低い

ヒット数が低い場合は更なるクエリの伝播を必要とする．これはクエリに対する答えが十分にローカル上に存在しなかった際に起こりえる事であり，ユーザによって返答できたクエリに関しては not で取り除き，解釈語のキーワードも含め，これを他のピアに伝播する．

4.2.3　　：確信度のみ低い

確信度が低い場合は，クエリ元ユーザの意図に沿った返答を出来ていないと考えられる．このためクエリを確信度の高いもののみに限定して他のピアに伝播する．

4.2.4　　：共に低い

共に低い場合は，クエリ元ユーザの興味のあるコンテンツをあまり所持していないと言う事が考えられる．このため特に処理を加えることなく，単にそのままクエリを伝播させる．

また，Gnutella 等の従来の手法のように一定のホップ数は伝播させる事により，全体的な検索結果数の底上げ等が可能であると考えられる．

4.3　クエリ変換とメタデータの改善

今まで述べた手法だけでは，全体のメタデータの質は向上しない．ローカルコンテンツ内で共起度の低いキーワード間の関連性はいつまで経っても抽出できないことになる．これはある種，ユーザの意図を反映していると言えるが，クエリに対してよい良好な結果を返すためには，概念体系のすり合わせが必要であると考える．このため，他ピアのクエリ変換情報等をメタデータに反映する機構を提案する．この機構の概観を表したのが図 6 である．

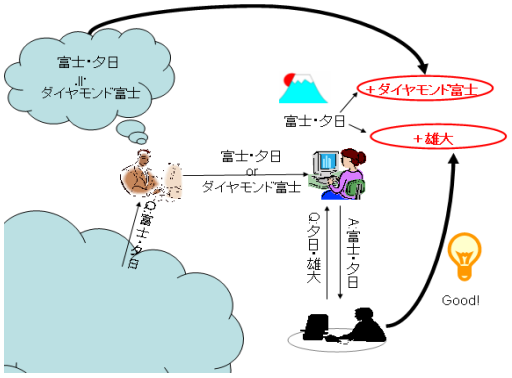

図 6　メタデータの改善

非常に簡単な手法としては前節でのクエリ変換を利用したメタデータの追加が考えられる．クエリ変換では他のピアがキーワード間が関連しているというと判断していることであるため，これをメタデータに追加することは有用である．また，クエリ変換の確からしさも加味し，自らの意図と重み付けを行ったうえで加える事によって，グローバルな意図を反映した上でのメタデータが作成出来る．

もう一つの手法としては検索結果を見たユーザからのフィードバックが考えられる．クエリ変換により得られた結果がユーザの意図していたものに合致しているような場合，このクエリ変換は適当であったと考えられる．ユーザの意図と合致しているかどうかは最も簡単な手法としてユーザが実際に判断し，その評価を下すというものが考えられる．意図していたものと少し違っている，とか，全く違うというような情報をユーザから入力してもらうことにより，適当なフィードバックが可能であると考えられる．このような場合はユーザの評価を含めて，メタデータに追加するような形になる．

5.　まとめと今後の課題

本論文では従来の P2P ネットワークにおける検索手法の持つ問題点について提起し，より柔軟な検索を実現するための手法について述べた．従来のファイル交換ソフトで用いられるような，ファイル名との一致に限られたファイル検索とは異なり，より汎用的なコンテンツ検索手法を提案している．．また，コンテンツに対する記述の多様性に対応するための機構として

クエリ変換とメタデータの改善機構について述べた．本論文ではユーザのクエリに対しての処理を中心に述べた．

本論文の提案する手法を用いると従来手法に比べてユーザの求める情報について発見しやすくなると考えられる．P2P ネットワークでの検索において，ネットワークトラフィックを鑑みると，クエリの伝播先を制限し，選択する必要がある．また，メタデータの付け方や，本論文で提案した確信度などを考慮に入れた上で，ユーザはそれらをアグリゲーションする必要がある．このような点も含め，今後考えていく必要がある．

謝　　辞

本研究の一部は，平成１６年度科研費特定領域研究（２）「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：16016247，代表：田中克己），および，２１世紀ＣＯＥプログラム「知識社会基盤構築のための情報学拠点形成」による．ここに記して謝意を表します．

参 考 文 献

1) Tempich, Christoph and Staab, Steffen and Wranik, Adrian REMINDIN ’:Semantic Query Routing in Peer-to-Peer Networks Based on Social Metaphors, In Proceedings International WWW Conference 2004, New York, USA.
2) Min Cai, M. Frank, RDFPeers: A Scable Distributed RDF Repository based on A Structured Peer-to-Peer Network, In Proceedings WWW Conference 2004, New York, USA.
3) I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan, Chord: A scalable peer-to-peer lookup service for internet applications., In ACM SIGCOMM, 2001.
4) gnutella，http://www.gnutella.com/
5) Napster, http://www.napster.com/
6) 湯川高志，吉田仙，桑原和宏，パーソナル・リポジトリに対するピア・ツー・ピア型協調検索機構の提案，信学技報 AI2001-48(2001).
7) S. Ratnasamy, et al., A Scalable Content-AddressableNetwork, Proc. ACM SIGCOMM 2001
8) Open Directory Project, http://dmoz.org/
9) Dublin Core, http://dublincore.org/