M-078

データマイニングによるデータハッシュテーブルの階層分類構造自動構築機能を利用した知的検索システムの提案

The Proposal of the Intelligent Search Engine Using the Automatic Construction Function of Classification Structure of the Data Hash Table by Data Mining

佐々木拓也† 澤本　潤† 瀬川典久† 杉野栄二† 加藤貴司† 和田雄次‡
Takuya Sasaki , Jun Sawamoto, Norihisa Segawa, Eiji Sugino, Takashi Katoh and Yuji Wada

1. はじめに

近年，ユビキタス環境が普及する中，モバイル機器を使って移動しながら活動する利用者に，場所（位置），時間などに応じてさまざまなサービスを提供することが求められている．例えば，ショッピングモール、国際会議場，スマートオフィス，駅構内，市街地等でのサービスが考えられる．こういったニーズに対して，無線 LAN，短距離無線通信といったアドホックな通信環境が移動を束縛しない自由なモバイルコンピューティング環境を実現しつつあり，各所に分散されたデータソースと計算処理ノードをつなぐことが可能となってきている[1].

一方，サービスを提供する仕組みとして，動的なサービスの発見や結合を可能とする技術が不可欠であるといわれ，セマンティック Web や Web サービス等の技術の適用が考えられる[2]．しかし、情報がいたる所に存在していても，その情報を効率よく組み合わせてニーズにあった情報を提供できていないのが現状である．これは情報を持っている個々のデータベースは独立して利用される事を想定しており，相互利用を想定していないので，管理手法が異なっている．そのためデータベース上の情報を組み合わせる事が難しい．また，ユーザの好みや空間的位置，時間を含むコンテキストの変化に応じて動的に適応していくことが不十分であるといった課題が挙げられる．

ショッピングモールにおける各小売店は，店舗毎に取り扱っている商品情報，在庫状況，キャンペーン情報などデータベースを利用することによって管理している．バス会社は市街地におけるバスの運行時間や運賃などの情報をデータベースと停留所間の時刻掲示板を利用し情報を表示する．Web 販売システムでは商品の情報だけでなく，Web 上で商品を購入する事が出来るなど，情報を必要とする人によって使い方も様々である．

そういった中，Peer-to-Peer(P2P)ネットワークを利用した情報の共有への注目が高まっている．クライアントが，接続したサーバから一方的にサービスの提供を受けるクライアントサーバモデルのシステムとは異なり，P2P 型システムでは，コンピュータ同士が相互に接続し平等な関係で直接情報やサービスをやり取りする．P2P ネットワークにおける検索手法の代表的なものとして，フラッディングによる方式と，分散ハッシュテーブルによる方式がある．分散ハッシュテーブルでは，非常に少ないメッセージ数で検索を行うことができるが，検索時に情報の識別子を指定しなければならないため，一度に複数のコンテンツを取得したい場合や，キーワードの部分一致などによる検索には不向きであるといわれている．また，アドホックなモバイル通信環境が P2P ネットワークに参加する場合，ノードの参加や離脱，再参加が頻繁に行われる状況にも対応が必要である．既に，分散ハッシュテーブルを用いた構造化された P2P ネットワークにおける，検索システムの研究と評価は種々行われている[3, 4, 5, 6]．分散ハッシュテーブルにおいてはコンテンツの柔軟な検索を行う事を目的とした階層分類化構造を導入した分散ハッシュテーブルの研究も事例として存在する[7].

本稿では，モバイル通信環境を利用しているユーザを対象とした P2P ネットワークでの分散ハッシュテーブルを利用した知的情報検索システム「GrowApp」を提案する．以下，2 章では，GrowApp の概要，構成要素，機能について説明する．3 章では，分散ハッシュテーブルにおける情報のグループ化機能について述べ，4 章では，そのシステム実装に関する構想について述べる．5 章では，実験構想について説明し，6 章で評価と応用に関する今後の課題について述べていく．

2. GrowApp について

2.1 概要

GrowApp はモバイル通信環境を利用するユーザを対象とした知的検索システムである．社会のあらゆる所に存在している情報を検索する事を目的とする．

GrowApp では情報を一つのデータベースでまとめて管理する事を行わない．一章で述べた通り，近年の情報技術の発達により個人，グループ，店舗などが独立してデータベースを利用した情報管理を行うケースが増えてきている．GrowApp はそれら既存のデータベースで扱われている情報を組み合わせて利用者に有益な情報を提供する事を目的とする．

利用者がモバイル通信端末を利用し，情報を検索し取得するまでのシステムの動作は次の通りである．

(1) 利用者は GrowApp へアクセスする．
(2) 利用者はシステムへ検索ワードを入力する．
(3) 検索ワードを取得したシステムは利用者のコンテキストと分散ハッシュテーブルを持たせた検索エージェントを生成する．

†岩手県立大学　Iwate Prefectural University
‡東京電機大学　Tokyo Denki University

(4) 検索エージェントは利用者の検索ワードにハッシュをかけた値を分散ハッシュテーブルから検索する．
(5) 検索ワードのコンテンツを所持しているノードが見つかると検索エージェントはそのノードまで移動する．
(6) コンテンツを持っているノードの仮想化データベース管理エージェントへ検索ワードを投げる．
(7) 検索ワードを受け取った仮想化データベース管理エージェントは担当の仮想化データベースへクエリーを実行する．
(8) クエリーを実行して取得した値を検索エージェントへ戻す．
(9) 値を取得した検索エージェントは元の利用者の位置へ戻り利用者へ取得した情報を提示する．

2.2 構成

GrowApp は仮想化データベース，モバイルエージェント，P2P ネットワーク，分散ハッシュテーブルの大きく分けて４つの要素によって構成される．

2.2.1 仮想化データベース[8]

社会には様々な種類のデータベースが存在している．それら一つ一つは管理手法が異なっている為にデータベースによって同じ検索手法をとる事ができない．仮想化データベースとは異なったデータベースであってもアプリケーション側からデータベースの種類を問わず検索，更新を行えるように仮想データベーススキーマをユーザ，データベース間へ設置する．この仮想データベーススキーマがデータベースの種類を問わない検索クエリーへ変換する．図 1 は異種データベースに仮想データベーススキーマを用い，様々なデータベースに異種性を意識することなくアクセスしていることを示している．

GrowApp では様々なデータベースを検索対象としている．データベースの種類を問わない検索を容易に行うために仮想化データベースを利用することによって，データベース管理エージェントが仮想データベーススキーマへのクエリーを実行し，情報を取得するという手順になる．

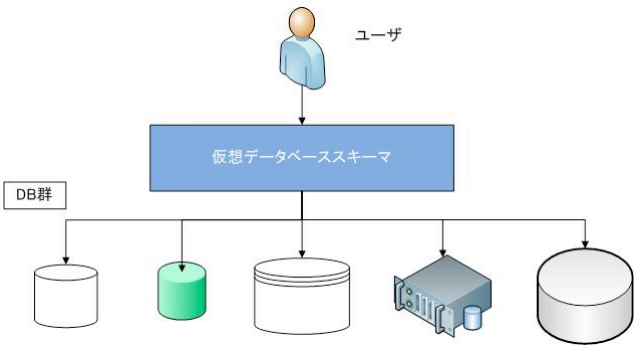

図 1　仮想化データベースの構成

2.2.2 モバイルエージェント

2.2.1 で述べた仮想化を，社会に存在している全てのデータベースに適用すると，一つの大きな仮想化データベースが出来る．仮想化データベースの最終的な理想は仮想化によって一つのデータベースのように動作させる事になる．しかし，現状では全てのデータベースを仮想化することはユーザ，仮想データベーススキーマ間がボトルネックになってしまう為に実現は難しいのが現状である．

本稿では，複数個のデータベースに仮想データベーススキーマを用い仮想化を行う．それらの仮想化を行ったそれぞれの仮想データベースにデータベース管理エージェントを配置させる．

さらに，ユーザがシステムを利用し検索ワードを投入した際に，システムはユーザのコンテキストを持った検索エージェントを生成する．ユーザのコンテキストとは，検索をかけた時点のユーザに関連する状況を表すものであるとする．空間的情報，時間的情報に加え，過去の検索履歴もコンテキストに加えることによって，過去の検索履歴も容易に現在の検索に反映することが出来るようにする．

図 2 に示すように，検索エージェントが検索を行う場合はデータベース管理エージェントへクエリーを投げる．クエリーを受け取ったデータベース管理エージェントは仮想データベーススキーマに対してクエリーを実行することによって，異種データベースに対して検索を行う．

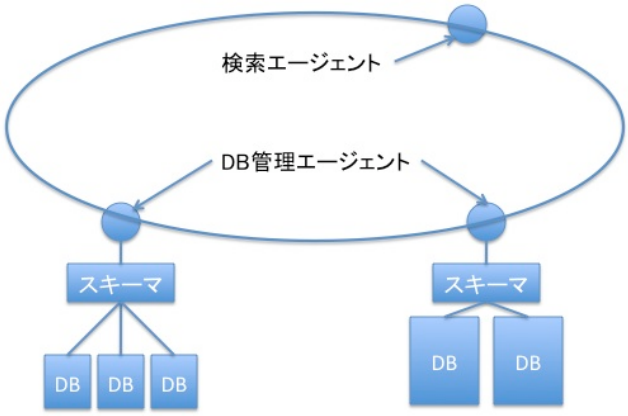

図 2　検索エージェントと DB 管理エージェント

2.2.3 P2P ネットワーク

データベースを一ヶ所にまとめた場合，ユーザとデータベース間のボトルネックが生じてしまう．本稿では，この問題を解決する為に．ユーザが検索を行う場合はエージェントを生成し，エージェント間の P2P 通信によって情報を取得することを考える．

2.2.4 分散ハッシュテーブル

本システムでは，ノードのアドホックなネットワークへの参加が予想できる．そこで頻繁なノードの参加，離脱が行われる事を想定し P2P ネットワークを構築する際に Kademlia アルゴリズムを用いた分散ハッシュテーブルを持たせている．Kademlia アルゴリズムは，ノードが頻繁に参加，離脱を行う事に対して特別な更新作業が不要であることから，実装が比較的容易なアルゴリズムとして評価されている．

2.3 機能

GrowApp は，多くの異種データベースを対象とするモバイルコンピューティング環境を考慮した知的検索システムとして以下の機能を備えている．

2.3.1 現在位置取得

ユーザによる，モバイルコンピューティング環境でのアドホックなシステムへの参加が予想される．そこで，ユーザの空間的情報を取得し，その空間的情報をユーザのコンテキストとして扱う．空間的情報によって現在ユーザがおかれている空間的状況を把握し，現在のユーザに対して最も適している情報への重み付けとして利用する．

今日の携帯電話には GPS 機能が搭載されているため，ユーザがシステムへアクセスした際にユーザの GPS 情報を取得することが可能である．この情報はユーザがシステムにアクセスする際に取得し，ユーザ毎に生成される検索エージェントが保持する．

2.3.2 分散ハッシュテーブルにおけるコンテンツのグループ化

本システムは P2P ネットワークを構築する．ノードがコンテンツを検索する際に分散ハッシュテーブルを利用する．ユーザの実行したクエリーから分散ハッシュテーブルを参照し，コンテンツを保持しているノードをたどることによって，目的とするコンテンツをダウンロードする．しかし，分散ハッシュテーブルでは検索ワードとの完全一致が要求される．分散ハッシュテーブルは，部分一致検索やグループ検索を行う事が出来ないという欠点を持っている．つまり柔軟な検索を行う事が難しい．この問題は検索システムを構想するにあたって解決しなければならない問題となる．そこで，本稿では分散ハッシュテーブルの階層化を行いコンテンツのグループ化を行う．コンテンツのグループ化が導入することで，検索ワードに対して，同時に検索される可能性の高い情報を同時に表示することで，ユーザの検索に柔軟さを導入することが出来る．分散ハッシュテーブルのコンテンツグループ化を行うことによって分散ハッシュテーブルの弱点である柔軟な検索機能に欠ける点を補うことが可能であると考える．

2.3.3 データマイニングによるカテゴリの計量

社会に存在しているデータベースの多くを検索対象とした場合，ユーザが検索を行った際に，大量の情報がユーザに提示されることになる．その中からユーザにとって有益な情報を見つけ出す事は非常に難しい[9].

そこで，本稿ではデータマイニングによるカテゴリの計量を行う．検索エージェントには，ユーザの位置的情報のほかにユーザの過去の検索履歴を保持させる．ユーザが本システムを利用し検索を行った際に，ユーザの検索履歴と返ってきた情報の嗜好について計量を行う．計量方法に関しては，ユーザの過去のコンテンツのダウンロードしたものから優先的に表示させる．Google 検索エンジンのように，多数の検索結果のページの中から，計量した結果を反映させ重み付けをする事で有効と考えられる情報から順に表示することが可能となる．

3. 分散ハッシュテーブル自動グループ化機能

3.1 概要

一般的に分散ハッシュテーブルは更新に非常にコストがかかってしまう．本稿では，アドホックなノードの参加を想定しているので，頻繁にノードの入れ替わりが起きてしまう．さらに，ユーザの検索状況に応じたコンテンツのグループ分けを頻繁に行うようになれば，なおさら，分散ハッシュテーブルを更新する回数は増えてくる．そこで，本稿では分散ハッシュテーブルの自動グループ化機能を提案する．グループ化を自動で行い分散ハッシュテーブルの更新を全て自動で行えるようになることで，リアルタイムに近いコンテンツのグループ化を提供することが可能になる．

本稿では，分散ハッシュテーブルに Kademlia アルゴリズムを使用している．Kademlia アルゴリズムはノードの参加，離脱に強いアルゴリズムとして評価されているが，本稿では分散ハッシュテーブルの階層化を行っている為に，ノードの参加，離脱に加えて新規カテゴリが作成された時や，古いカテゴリが削除された場合においてもテーブルの更新が行われる．分散ハッシュテーブルに変更が入った際に更新を行わないと，コンテンツ検索の精度が減少してしまう．その為，変更が入る度に更新を行うのが理想的だが，更新を行う事は人手を要する作業になってしまう．更新は増えれば増えるほど理想的だが，その反面何度も人手を介す必要がある．

そこで，本稿では新規コンテンツグループの生成，古いコンテンツグループの削除をユーザの検索状況に応じて自動処理する手法について提案する．

あるユーザがシステムを利用し検索を行った場合，一つのアイテム検索だけではなく，さらに連続して検索を行う事でそれら検索対象を一つのまとまりとして考えることができる．あるユーザが利用したコンテンツのグループは，他のユーザにとっても有用なコンテンツグループと利用される可能性を持っている．そのユーザだけでなく，他ユーザも同様な検索を行っていた場合それらは新たなカテゴリとして分散ハッシュテーブルに登録される．しかし時間の流れや新しい種類のアイテムが現れた際に，検索対象として外れていくアイテムを永久的にカテゴリとして残すのではなく，検索回数などに応じてカテゴリの更新を行う．そうする事で常にユーザの求めている嗜好に合った検索を効率よく行うことが可能となる．

3.2 グループ生成の例（PC 機器の例）

ユーザが実際に検索，コンテンツのダウンロードを行う際に起きる新規グループの自動生成，使用されなくなったグループの削除の具体的な流れについて以下に記述する．

(1) 分散ハッシュテーブルに登録されたコンテンツは初期状態ではどこのグループにも属していない状態である．図 3 でコンテンツについて示している．それぞれ「CD-552GA」は CD ドライブ，「Core i7 920」は CPU. 「P6T」はマザーボードのコンテンツの名称である．
(2) これら 3 つのコンテンツが多数のユーザによってコンテンツの検索，ダウンロードが頻繁に行われるようになる．

(3) これら３つのコンテンツはよく同時に検索されるグループとして分散ハッシュテーブルに登録される．図４ではグループ化されたコンテンツを示している．
(4) DVD ドライブの出現．ここでは DVD ドライブは「DVR-216DBK」というコンテンツ名であり．図５にその状態を記している．
(5) DVD ドライブの出現によりユーザは CD ドライブを含んだグループに代わって DVD ドライブ，CPU．マザーボードの組み合わせの情報検索を行うようになる．
(6) DVD ドライブ，CPU，マザーボードの組み合わせの検索がユーザによって頻繁に行われるようになった為，新たにグループを生成し，分散ハッシュテーブルへ登録される．図６は新たにグループが生成された時点での状態を示している．
(7) DVD ドライブを含むグループが CD ドライブを含んだグループにかわって多数のユーザにコンテンツの検索，ダウンロードが行われるようになる．一方 DVD ドライブ出現後の CD ドライブは検索，ダウンロード回数は減少してしまった．
(8) CD ドライブを含んだグループはユーザにとって有益な情報とは言えない状態になる．そこで CD ドライブ，CPU，マザーボードの組み合わせのグループを分散ハッシュテーブルから削除する．図７では CD ドライブを含んだグループを削除した状態である．しかしグループが削除されただけであり，コンテンツとして CD ドライブは検索，ダウンロードを行う事は可能である．

図 3 単体コンテンツの状態

図 4　コンテンツグループ生成

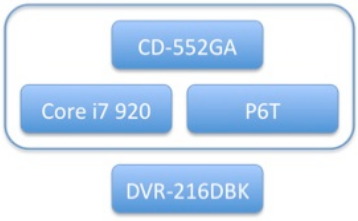

図 5　新たなコンテンツの出現

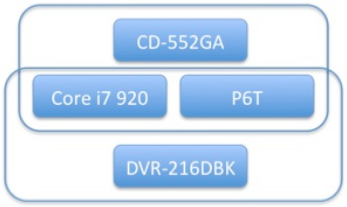

図 6　新たなコンテンツグループの生成

図 7　コンテンツグループの削除

4. システム実装

本システムはユーザとアプリケーションの接点のユーザインタフェース部分と，モバイルエージェント部分，P2P ネットワーク部分，仮想化データベース部分の４つを構築し実装を行う．本研究では，プログラミング言語 Python を使用し，分散ハッシュテーブルのコンテンツのグループ化を P2P シミュレータ上で動作の確認，仮想化データベーススキーマ，検索，データベース管理エージェントの作成，GrowApp のインタフェースの作成と順を追って構築していく．本稿では P2P シミュレータを用いた分散ハッシュテーブルのコンテンツのグループ化の有用性について実験を行い評価を行う．

ユーザインタフェースは，唯一ユーザが本システムに接触する部分である．検索ワードを入力するフォームと，コンテンツの情報を表示する機能を持つ．

本システムでは二種類のモバイルエージェントを利用している．一つはユーザの検索ワードをコンテンツ情報から検索する検索エージェント．そして仮想化データベース内のコンテンツ情報と検索エージェントからのクエリーを受信する機能を持っているデータベース管理エージェントである．

GrowApp の一機能である分散ハッシュテーブルにおけるコンテンツの自動グループ化が正確に行われ，P2P ネットワークが検索システムとして優位的な結果を残せるかを確認する為に P2P シミュレータを利用し実験を行う．今回は，P2P シミュレータは Kademlia アルゴリズムを実装した P2P ネットワークを構築する．

仮想化データベースに関しては，[8]で試作された仮想化データベーススキーマを利用し，複数個のデータベースを仮想化する．

5. 実験

本稿では，分散ハッシュテーブルにおけるコンテンツの自動グループ化が有用的に行われているか，システムとしてユーザへかかる負担度の計測をする為のトラフィックの取得を行い，同時に検索精度の定量的評価を行う．

5.1 実験概要

本稿では以下の実験内容を考える．以下の実験内容をノード数を初期状態 200 ノードから 1 台 PC（200 ノード）ずつ増加させていき計量を行う．

- 情報量のトラフィックの計量
- 検索ワードに対するコンテンツの調査
- 検索をかけてから表示されるまでのレスポンス時間

実験内容は，PC1 台当たりに 200 個程度のノードを生成し，P2P ネットワークを構築する．そして今回の実験対象である，トラフィックの計量と分散ハッシュテーブルによる検索精度の検証を行う．

5.2 実験構成図

実験時の構成図を図８に示す。本稿の実験はトラフィックの比較，自動グループ化された分散ハッシュテーブルの精度を評価するので，模擬的に 1 台の PC で 200 ノード程度生成し P2P ネットワークを擬似的に構築する．　それらのノードが所持している分散ハッシュテーブルによる検索を行う事によって分散ハッシュテーブル内のコンテンツがユーザの検索されたワードのコンテンツグループとして正確に追加されるのか，また，サーバ・クライアント方式に比べトラフィックがどのように変化していくのかを検証する．

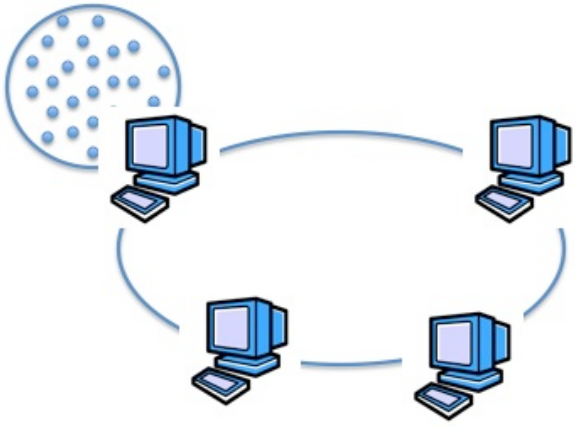

図 8　実験構成図

5.3 ソフトウェア構成図

ソフトウェア構成について図 9 に示す．各端末の OS 上で P2P シミュレータを起動する．OS に関しては，様々なモバイル通信端末での利用を想定し Windows，Linux を実験に使用する．P2P シミュレータはユーザインタフェースから入力された検索ワードにハッシュをかけ分散ハッシュテーブルを参照する．ノードの位置を指定されたシミュレータはノード位置へ移動し，モバイルエージェントがデータベース管理エージェントと通信を行い，ユーザの求めているコンテンツ情報を要求し取得する．最後にユーザへコンテンツ情報を表示する．

図 9　ソフトウェア構成図

6. 評価と応用

実験内容に対応した評価の基準については以下の通りである．評価の基準対象として，本システムと同じ情報量を一つのデータベースに格納した，サーバクライアントモデルを利用する．

- 本システムとサーバクライアントモデルで同じ検索ワードをかけた際の，情報のトラフィックの比較
- 本システムで，ワード検索を行い表示されるコンテンツの精度評価
- 本システムで，コンテンツの同時検索を行い分散ハッシュテーブル内のコンテンツの自動グループ化が行われているか
- 本システムと，サーバクライアントモデルで，コンテンツの検索を行った際のレスポンス時間の比較
- 参加ノード数が増加した場合にも適したシステムであるか

実験結果を踏まえた上で，分散ハッシュテーブルにおけるコンテンツのグループ化が有用な手段であるかを評価する．評価の基準は，実験概要で述べたように，分散ハッシュテーブルの構築、検索，更新などに要するトラフィック量の，クライアントサーバモデルを用いた場合との比較値，およびユーザからの検索要求に対する検索精度を用いる。本実験の結果内容を踏まえて，今後分散ハッシュテーブルのコンテンツのグループ化機能の改良を行っていく．

さらに，本提案方式を評価した上で、今後は本稿で提案した GrowApp の主要機能として実装していく事を検討していく．

7. おわりに

本稿では，ユビキタスネットワーク社会における既存のデータベースを利用した知的情報検索システム GrowApp について提案を行った．GrowApp の主要機能である分散ハッシュテーブルにおけるコンテンツのグループ化について，実装を行い実験評価を行っていく．その評価を踏まえ，コンテンツのグループ化機能に改良を加えながら，仮想化データベーススキーマとの連携，モバイルエージェント，インタフェース部分の構想を行い，最終的に GrowApp 全体のシステム設計を行っていく予定である．さらに，提案した GrowApp の具体的な応用システムを考え，実際に適用

実験を行っていくことによりその有用性の検証を進めていく予定である．

謝辞

本研究は科研費（20500095）の助成を受けたものである．

参考文献

[1] D. Chakraborty et al., “Toward Distributed Service Discovery in Pervasive Computing Environments,” IEEE Trans. Mobile Computing, vol. 5, no. 2, 2006, pp. 97–112.
[2] 小倉弘敬, 村上佐枝子, 佐藤宏之, 小島富彦, 清水昇, 細見　格, “セマンティック Web の応用システム,” 情処会誌 43 巻 7 号, pp.743-750, 2004.
[3] Ben Y. Zhao, John Kubiatowicz, and Anthony Joseph. Tapestry: an infrastructure for fault-tolerant wide-area location and routing. Technical Report UCB/CSD-01-1141, U.C. Berkeley, April 2001.
[4] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrish-nan. Chord: A scalable peer-to-peer lookup service for internet applications. In Proceedings of the ACM SIGCOMM ’01 Conference, San Diego, California, August 2001.
[5] Petar Maymounkov, David Mazi`eres, Kademlia: A Peer-to-peer Information System Based on the XOR Metric, In Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS), March 2002.
[6] 葛野弘樹，中井優志，渡邊集，川原卓也，加藤貴司，ビスタベッドバハドゥール，高田豊雄，モバイルエージェントを用いた分散型インターネット観測システムの提案，情報処理学会論文誌 47(5)　pp.1393-1405　2006.
[7] Yi WAN, Takuya ASAKA and Tatsuro TAKAHASHI, A Hybrid P2P Overlay Network for Non-strictly Hierarchically Categorized Content, IEICE Transactions on Communications 2008 E91-B(11):3608-3616; doi:10.1093/ietcom/e91-b.11.3608
[8] 渡辺裕太，菖蒲佳右，三井田浩，和田雄次，澤本潤，加藤貴司，異種データベースの仮想化技術，FIT2009.
[9] 森薫，倉林修一，石橋直樹，清木康, “モバイルコンピューティング環境におけるユーザ情報の動的計量による能動型情報配信方式,” 電子情報通信学会第 15 回データ工学ワークショップ (DEWS2004)論文集, March, 2004.