研究計画書
価値のある情報を発掘する情報流通基盤の構築

慶應義塾大学環境情報学部
自署：＿＿＿＿＿＿＿＿＿＿
学籍番号 79956832
平成 14 年 5 月 14 日

概要

コンピュータネットワークを利用して様々な活動が行われるようになり, そのうえで多様な情報が日常的に作り出されている. しかしそれらをネットワーク上で共有するには既存の技術では手間がかかり, 公開の意思があってもそのままコンピュータの中で無駄となっていることが多い. 本研究では Peer to Peer モデルを利用して低コストで情報公開ができ、高い検索性を持つ新しい情報流通基盤を構築する。

1　研究の背景と問題意識

私たちはコンピュータを利用して, 研究や講義に利用するための資料など様々な情報を日常的に作り出している. しかしこれらの情報は, 一度使われた後は多くの場合そのまま公開されることなく, コンピュータのディスクの中に埋もれてしまっている. これらの情報を他の人が資源として再利用可能にするには, 自分のためだけに作られた情報を他者向けに整理し公開するという作業が必要だが, たとえ公開したいという意思があったとしても情報を整理する時間をとれず, そのままになってしまうことが多い. その結果, 過去に同様の情報が生成されていることに気づかないまま, 再び時間を掛けて生み出されるという無駄が起きてしまう.

2　研究の目的

本研究では, コンピュータに埋もれている情報を再び流通させるために, 以下を実現する情報流通基盤を構築する. ただし, 本研究において情報とは, 授業のスライドや講義資料, 研究行為の過程で生成された各種資料などの一般にファイル単位で扱われるものを指している.

1. 情報を手間をかけずに公開できる

   これまでこういった情報が公開されなかったのは, 公開に大きな手間や時間がかかるからである. この手間をできる限り軽減することにより情報の公開を促す.

2. 自分の欲しい情報を的確に探し出せる

   ただ全文検索をするだけでは, 規模が大きくなった場合に検索結果が膨大な量となってしまうので, 単なる検索を行うだけでなく内容やその他の情報を利用して目的の情報に絞りこめることが必要となる. またその情報の存在を第三者によって確実に発見できなければならない.

3. 発見した情報を利用できる

   検索で見つかった情報を実際に利用できなければならない. たとえばその情報を公開していたノードがオフラインになった場合でもキャッシュから取得する手法などが考えられる.

3　既存の技術

現在情報の公開および共有に使われているモデルがいくつか存在する. この章ではそれらが上記の目的を満たすために不十分である点を分析する.

3.1　WWW(World Wide Web)

現在 WWW は情報の公開に広く利用されているが, 以下の理由から目的を満たさない.

- WWW 上で情報を公開するには, 原則として HTML 等の一般的な形式を用いて情報を記述し直す必要がある. その上で, それらの情報を WWW サーバに転送しなければならない. この手間は情報を公開するうえで大きな障壁となっている.
- WWW 自体は検索機能を持たないため, 情報を発見するには能動的に情報を収集して索引を作

成する検索エンジンに頼らなければならない. これらのサービスはインターネット全体を対象としているため, WWW サイトの増加に伴い目的の情報を探すことが難しくなっている.

3.2 ナレッジ・マネジメントツール

ナレッジ・マネジメントツールは, ある組織に属する者が得た知識をデータベースに蓄積し再利用することでその組織内での無駄を減らすアプリケーションである.

基本的にデータベースに蓄積された内容や既に公開された情報を対象としているためデータの質も高く検索性も高い.

しかし本研究で想定しているのはネットワーク上に分散したコンピュータが持つ情報からの検索であり, 特定の目的に関連する知識を対象とするナレッジ・マネジメントツールとは視点が異なる.

3.3 Peer to Peer

Peer to Peer モデルは, WWW のような Client-Server モデルと異なるもので, 全てのホスト (ノード) がその両方の機能を持ち, 直接相互に接続しデータを転送するモデルである. その性質上, データそのものをサーバに転送する手間が発生しない.

3.3.1 Peer to Peer モデルを利用したアプリケーション

Peer to Peer モデルを利用した主なアプリケーションとして以下の 3 つが挙げられる.

Napster[1][2] Napster 社により音楽データの共有を目的として設計されたアプリケーションである. 検索機能を中央の索引サーバによって行う "Hybrid P2P" モデルを利用している. 現在は著作権上の問題が有るデータの流通を制限している.

Gnutella[3][4] "Pure P2P" モデルを利用した通信基盤である. キャッシュ機構が無く, 検索要求が無秩序に流れるため, ネットワークが大規模になるにつれて検索要求によってネットワークが高負荷になったり目的のデータを持つノードまで検索要求が届きにくくなるという問題がある.

freenet[5][6] 匿名の情報保存・検索ネットワークとして設計されたネットワーク基盤で, アップロードされたデータをネットワーク全体のノードに分散して保存してデータの提供者や取得者が分からないように設計されている. その代償として通信速度や確実性を犠牲にしている.

表 1: Peer to Peer モデルの分類

	検索方法	キャッシュ
Napster	Hybrid P2P	なし
Gnutella	Pure P2P	なし
freenet	Pure P2P	あり

3.3.2 Peer to Peer モデルの分類

これらのアプリケーションは以下に示す 2 点で分類することができる. その結果を表 1 に示す.

検索方法 ファイルの一覧とクライアントの位置の管理にあらかじめ指定されたロビーサーバを利用する方法を Hybrid P2P モデルと呼ぶ. ロビーサーバから目的のファイルを所有するクライアントの IP アドレスを取得し, ファイルの転送のみを直接相手に接続して行う. ロビーサーバに登録されている情報は確実に検索できるため効率は高いが, ロビーサーバやネットワークの障害で接続できないとネットワーク全体が分断される.

ロビーサーバに頼らない方法を Pure P2P モデルと呼ぶ. 特定のノードに依存しないため, どのノードがネットワークから切り離されてもネットワーク全体としては機能しつづける. 検索は条件を含む検索要求をネットワークに流す事によって行うことが多い. ネットワークの規模が拡大すればするほど目的のデータを持つノードまで検索や転送の要求がたどり着けない可能性が高まり, 検索の確実性が落ちるという問題点がある.

キャッシュの有無 各ノードがデータを別のノードから取得した場合に, そのデータをノードとしてキャッシュしてさらに別の第三者に再配布するかどうかでネットワーク全体の性質が変化する.

キャッシュを行わない場合には各ノードが持つデータはたとえそれが他のノードから取得したものであっても別個のものとして扱われる.

キャッシュを行う場合は, 最初の所有者がネットワークから切り離されてもそのデータを他にダウンロードしたノードから取得することができる. また, 人気のあるデータが存在した場合にキャッシュが無い場合はその所有者にアクセスが多発するが, そういったデータはキャッシュも増えるためボトルネックが発生しにくい.

3.3.3 既存のアプリケーションの問題点

Peer to Peer モデルはアップロード作業がいらないという点だけでなくサーバを通さず直接ノード間で転送を行い効率が良い等の点から注目されている.

しかし既存のアプリケーションではファイル名でしか検索を行うことができず, これは本研究の目的 2 を満たすことができない.

また,Napster は音楽データに特化することで音楽データにつけられたメタデータ [7] から取得した曲名やアーティスト名, あるいは符号化のパラメータで表される音質を条件とした検索ができる. しかし, メタデータの使い方や正確さはそれを設定した人の価値観によって大きく異なるため確実な検索性を提供しているとはいえない.

3.4 まとめ

以上の観点から既存の技術の特徴を表にまとめたものが表 2 に挙げる.

4 本研究のアプローチ

4.1 概要

本研究では公開に要する手間を最小限に押さえ, かつ確実に目的の情報を検索できるようにするためにインターネット上に分散した Hybrid P2P モデルを用いたアプリケーションを構築する.

さらに, ロビーサーバに接続できない場合は接続できるノードのどれかがロビーサーバとして機能することで, 閉鎖された環境等であってもネットワークを構築できるようにする.

4.2 要求事項

コンピュータの中に埋もれている情報を公開, 共有するためには以下の 5 点を満たさなければならない.

4.2.1 アクセス制御

必要な場合に相手を認証し, それによってデータの利用権限を制御できなければならない.

Peer to Peer モデルではサーバへのアップロード作業は必要なくなるが, 公開してもよいデータと公開したくないデータを区別して扱わなければならない. さらに, 不特定多数には公開したくないが特定のグループの参加者は取得できるようにしたい, というケースが考えられる.

4.2.2 操作性

一般的にソフトウェアが高機能になるにつれ操作体系は複雑化するが, それは本研究の目的 1「容易に公開できる」という条件に反する. したがって, 設定や操作を直感的に行えるユーザインターフェースが必要不可欠である.

特に公開すべき情報の指定では, 不特定多数に公開, あるグループの参加者のみに公開, 公開しないといった指定方法が考えられるが, これら複雑な設定を簡便に指定できなければならない.

4.2.3 検索性

欲しい情報をより細かい粒度で確実に見つけだすためには, 設定した人によって使い方の異なるファイル名やメタデータの単純検索だけでは不足で, データの内容そのものに踏み込んで検索できなければならない. また, 検索の精度を向上する工夫を積極的に取り入れるべきである.

4.2.4 位置透過性

求める情報がネットワーク上のどこにあっても, 検索して取得できなければならない.

また, データを取得する際に所有者の端末がネットワークに接続されていない場合, そのデータをキャッシュしている端末が他にあれば取得できなければならない. また, 携帯端末を考慮してネットワークに接続されている間にデータをキャッシュする端末があればデータ取得の確実性が向上する.

4.2.5 規模性

Peer to Peer モデルではノードやネットワーク上に流れる情報量が大規模化する傾向にある。本研究で実現するアプリケーションでは、このように大規模化したネットワークにおいても検索や情報の取得が実用的でなければならない.

5 設計

上記の条件を満たすため, 新しいアプリケーションを設計・実装する.

5.1 ノード

このアプリケーションは以下の 4 種類の論理的なノードから成り立つ. これを図示したものが図 1 である. 実際には, 各ノードがこの論理的ノードを含むことになるが, たとえば検索機能のみを提供するノードの運用等も考えられる.

- 公開ノード

表 2: 既存の技術

	公開の手間	検索の精度	情報の取得
WWW	△	○	○
ナレッジ・マネジメントツール	△	◎	○
Peer to Peer(Pure P2P)	○	△	△
Peer to Peer(Hybrid P2P)	○	○	○

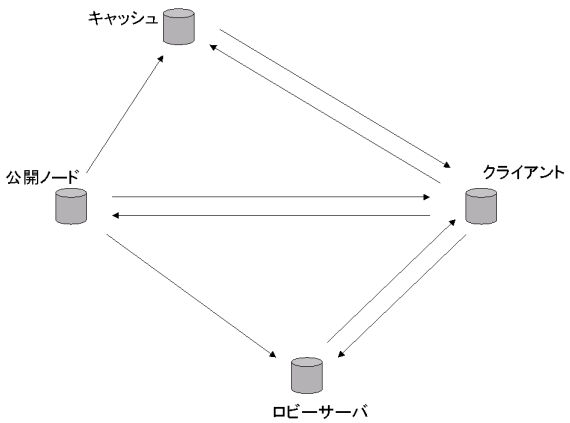

図 1: 構成図

- ロビーサーバ
- クライアントノード
- キャッシュノード

公開ノードがデータを公開し, クライアントはロビーサーバに検索を要求する. その結果からクライアントは目的のデータを持つ公開ノードに接続し, そのデータを取得するという基本的な Hybrid P2P モデルの構成を踏襲している.

本アプリケーションでは一度自分が取得したデータをキャッシュとして提供するほか, 積極的に特定の公開ノードの複製を行いキャッシュを提供することもできる. これは移動端末の複製を自宅のコンピュータで行うといった用途を想定している.

また各ノードは基本的にロビーサーバとしての機能も持ち, ロビーサーバに接続できない場合は自らがロビーサーバとなり近隣のノードに検索機能を提供する.

5.2 高度な検索性の提供

4 章で示したように, ファイル名やデータに付随するメタデータだけでは人によるばらつきが加味されるために検索条件として不十分である. そこで, 人の手によらない以下のような手法が考えられる.

リンク情報の利用 複数のファイル間の関係を示すリンク情報を共有しこのリンク情報を利用して検索精度の向上を図れる. 例えば引用が行われた場合に引用元のデータに対するリンクを作成することで, そのデータの閲覧者が引用元となった 1 次情報のデータを検索できるようになるといったことが考えられる.

同種のファイルのグルーピング 検索を支援するために, 似た内容の情報で異なる内容のものをグルーピングして扱うことも考えられる. テキストやグラフィックデータではデータの部分的な変更があり得るが, 既存のシステムでは少しでも変更されたデータは全く別の存在として扱われている. また, そうでなくても他の人が持つデータが同じかどうかは, md5 などのハッシュを利用するか, 人間が判断するしかない. また, 全く同じ内容であっても記録方式によって同じ内容に対しバイナリとして異なるデータが無数に存在する. これらを同一視ないしグループ化して検索できれば, 求めるデータの発見がしやすくなりネットワーク全体の効率を向上させられる.

6 評価方針

本研究で作成するアプリケーションが要求事項を満たしていることを確認するため, 実際に運用を行い, 以下の点から評価を行う.

- 情報を手間をかけずに公開できる
- 自分の欲しい情報を的確に探し出せる
- 発見した情報を利用できる

7 予想される成果と将来への展望

7.1 予想される成果

以上のような情報流通基盤を構築・運用することにより, これまで使い捨てられていた情報を容易に他の人が資源として使うことができるようになる. その結果, 社会全体がより価値の高い情報の生産に専念でき, 技術や文化の発展に貢献する.

7.2 将来への展望

7.2.1 ファイル以外への対応

対象とするデータをファイルに限らず, 例えばマルチキャストで流れている映像ストリームを同じように扱うことで, ユーザは検索して自分のみたい映像を探し, そのマルチキャストグループに参加するという使い方が考えられる.

7.2.2 課金システムとの協調

本システムではあらかじめ設定されたアクセス制御リストしか考慮していないが, 何らかの電子キャッシュ系と組み合わせることで, 動的にアクセス制御リストを更新し, 情報の受信に課金するという使い方ができるようになる.

8 これまでの活動

徳田・村井・楠本・中村・南合同研究会 環境情報学部入学当初より徳田・村井・楠本・中村・南合同研究会に所属し, 高度なネットワークインフラの構築及び運用に携わっている.

pie[8],sprng[9] 大学 1 年春よりインターネットの普及に関する研究グループに所属し, インターネットインフラが普及する上での問題点や, 普及が社会に与える影響について研究を行っている.

neco[10] 大学 3 年春よりネットワーク上のコミュニケーションに関する研究グループに所属しネットワーク上で行われる新しい形態のコミュニケーションの研究に携わっている.

9 志望理由

本研究では, 新しい情報流通ネットワークモデルを提案し, 構築する. これを実現するためには, ネットワーク技術のみならず, 効果的な索引の作成や検索にかかわるマルチメディアデータベースや, アプリケーションとしてより多くの人に使ってもらうためのユーザインターフェースに関する議論が必要不可欠である. このように多角的な面から研究を進めるにあたり, これらの分野の研究が盛んに行われている政策・メディア研究科を志望したい.

10 共同研究者・関連団体

本研究では, 慶應義塾大学大学院政策・メディア研究科 “モービル広域ネットワーク (MAUI)” プロジェクトにおいて, 村井純教授, 楠本博之助教授, 中村修助教授の指導のもとに行う. また, 本研究の基礎技術として必要不可欠なデータベースおよびユーザインターフェースの教員の指導を受けたいと考えている.

また, 以下の団体と協力しながら研究を進める.

- 慶應義塾大学村井研究会内ワーキンググループ neco
- WIDE Project

参考文献

[1] napster messages
http://www.clip2.com/gnutellaprotocol04.pdf.

[2] Napster Inc. Napster
http://www.napster.com.

[3] Clip2. The gnutella protocol specification v0.4.
http://www.clip2.com/gnutellaprotocol04.pdf.

[4] Jnutella.org
http://www.jnutella.org/.

[5] Ian Clarke, Oskar Sandberg, Brandon Wiley, and Theodore W. Hong. Freenet: A distributed anonymous information storage and retrieval system. In *Workshop on Design Issues in Anonymity and Unobservability*, pages 46–66, 2000.

[6] Freenetproject.org
http://www.freenetproject.org.

[7] Id3v2
http://www.id3.org.

[8] インターネットの普及に関する研究グループ
http://www.sfc.wide.ad.jp/kg/pie/.

[9] 通信基盤としてのインターネット整備戦略、およびその上での政策・社会環境の整備に関する研究グループ
http://www.sfc.wide.ad.jp/kg/sprng/.

[10] ネットワーク上のコミュニケーションに関する研究グループ
http://www.sfc.wide.ad.jp/kg/neco/.