2U-7

# P2P 上で情報共有をおこなうための情報管理機構の構築

大塚　雅史 †
東京工科大学大学院工学研究科 †

上原　貴夫 ‡
東京工科大学コンピュータサイエンス学部 ‡

## 1　はじめに

近年の飛躍的なコンピュータ技術の発展と，ネットワーク環境の整備により Peer To Peer(以下 P2P) という通信形態に対する可能性に期待が高まっている．P2P は MP3 形式の音楽ファイル交換用アプリケーションソフト Napster の登場により注目を集めた．P2P は負荷の分散，設備コストの削減など多くの利点があり，様々な分野においてその可能性が注目されている．

本研究では P2P ネットワークをもちいた情報共有をおこなうためにシステムを提案する．P2P ネットワーク上で情報を共有した場合，端末が自由に情報の送受信をおこなうために，情報漏洩といったセキュリティ上の危険がある．さらに，単に情報を P2P ネットワーク上で共有するだけでは効率の良い情報共有とはならない．そこでそれらの問題点を解決し P2P ネットワーク上で効率よく情報を共有するために，本研究ではサーバによる情報管理機構を構築することで，情報の管理とユーザの管理を実現し効率の良い情報共有と，セキュリティの向上を図る．

## 2　情報共有支援システム

本研究で提案する情報共有支援システム [1] は，Hybrid 型 P2P ネットワーク上において情報を共有する．P2P では共有する情報は各ピアの記憶領域において保持し，サーバでは共有情報の管理をおこなう．共有情報はカテゴリごとに分類し管理し，ユーザの検索効率の向上を図る．

各ピアが P2P ネットワーク上で共有される情報を取得する際には，必ずサーバに接続し各ユーザが取得できる情報を示したカタログを入手する．それざれのユーザがサーバに対して必ず接続することで，サーバにおけるユーザの管理をおこなう．カタログには入手することのできる情報のみが記されているため，情報への不正なアクセスや第三者による情報の取得を防ぐ．本研究で提案する情報共有支援システムの概要図を図 1 に示す．

## 3　情報管理機構

P2P ネットワーク上で情報を共有するため，情報は各ピアの記憶領域において保持される．そのため共有されている情報について一元管理することは難しい．しかし，情報共有という観点から考えると P2P ネットワーク上でどのような情報が共有されているのかを管理者が管理できる機能が必要である．よって，提案するシステムではサーバ上に情報管理機構を構築し情報を管理する．情報の管理はピアにおいて生成されるメタデータをもちいる．

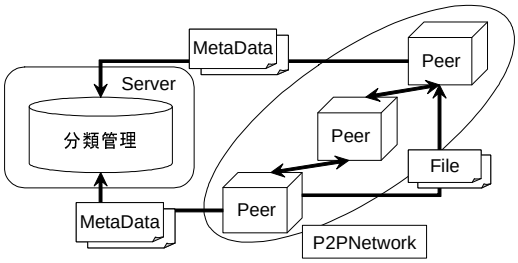


図 1　システム概要概要図

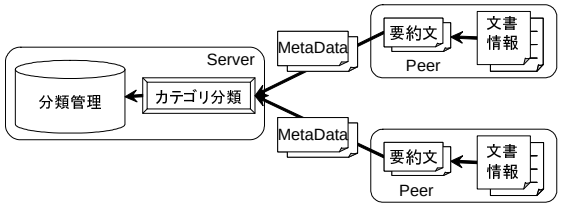


図 2　共有情報の蓄積・管理部の概要図

メタデータにはファイル名，情報の作成者，更新日時，要約文，要約文生成日時，発信者 ID，アクセス制御コードなどを属性として付加する．

情報の管理は，アクセス制御コードを利用してユーザごとのファイルへのアクセス制御，要約文を利用してカテゴリ分類による情報の整理との二点についておこなう．本研究で提案する情報の蓄積・管理部の概要図を図 2 に示す．

### 3.1　アクセス制御

本研究では共有する情報に対してアクセス制御をおこなう．アクセス制御をおこなうためにメタデータにアクセス制御コードを属性として付加し，ユーザからファイルへのアクセスを制御する．

ユーザが情報を取得するためサーバにアクセスした際に，サーバにおいてそのユーザが取得可能な情報についてのメタデータを抽出し，生成したカタログをユーザに対して提供する．ユーザはカタログに記された情報から取得したい情報を選択する．P2P ネットワーク上から情報を取得するのに必要な情報は，ユーザが情報を選択した際にサーバから取得する．

### 3.2　文書分類器

本研究ではサーバにおいて情報をカテゴリごとに分類し情報を管理する．情報を分類するために各ピアにおいて生成されたメタデータに含まれる要約文を利用する．カテゴリごとに情報を分類することで，ユーザの検索効率の向上が見込める．また，情報がカテゴリごとに分類さ

---

Structure of information management mechanism where information is shared by P2P

† Masafumi Ootsuka, Graduate School of Engineering
‡ Takao Uehara, School of Computer Science, Tokyo University of Technology

れているため，ユーザが求める情報だけでなくカテゴリ内の他の情報も同時に閲覧することができる．これは求める情報だけでなく，カテゴリ内の情報も付加情報として利用することができる．これにより共有する情報が過去に利用され情報提供者にとっては価値の薄れた情報であったとしても，他者に再利用されることで情報としての価値が高まる可能性が生まれる．カテゴリ分類ではクラスタリングの手法をもちいる．クラスタリングはデータ解析手法の 1 つで，データの集合をクラスタに切り分け，それぞれのクラスタに含まれるデータが共通の特徴を持つようにする．

## 4 要約文生成手法

本研究では共有する情報の対象を文書情報としているため，サーバでの情報管理をおこなうのに，共有する情報から要約文を生成しもちいる．要約文はユーザがネットワークに対して情報の共有を許可した際に生成される．生成した要約文はメタデータに付加しサーバに登録する．

重要箇所の抽出は主に単語の重みを指標としてもちいる．本研究では単語に対する重み付けには TF-IDF 法をもちいる．TF-IDF 法は単語の重み付けをおこなう場合には一般的にもちいられる手法である．しかし，出現頻度をもちいて計算しているので文書の長さによって精度に差が出る．そのため，出現頻度を正規化し文書長の長さによる影響を少なくする必要がある．また，TF-IDF 法は高い精度で計算するためには大量の文書が必要になる．これらの問題点をふまえると TF-IDF 法の結果のみを利用して要訳文を生成するには，精度の点において問題がある．よって本研究では TF-IDF 法に加えて情報利得比 [2] の考え方を導入する．情報利得は単語が指定されたカテゴリにとってどれだけ特徴的であるかを示す．

### 4.1 情報利得比

情報利得比 [2] はクラスタの分割構造に対してえられる値でクラスタの分割の際に毎回算出される．情報利得比とは本来，決定木学習システム C4.5 において属性選択をおこなうために導入され，C4.5 ではある属性の決定木の分岐におけるテストとしたときに，その属性がどれくらい適切にクラスの出現を予測できるかを表す尺度としてもちいられている．本研究では属性ではなく，クラスに対応する単語の評価値として情報利得比をもちいる．$C_i$ を C の部分クラスタとすると，クラスタ C における単語 w の情報利得比 $\mathrm{gain_r(w, C)}$ は式 (1) により求める．

$$gain_r(w, C) = \frac{gain(w, C)}{split_{info}(w, C)} \quad (1)$$

情報利得比を語の重要度にもちいた場合には，出現頻度分布についてクラスタの下位分岐構造との整合性が高い語ほど重要度が高くなる．つまり，クラスタを特定できるような語の情報利得比は高くなり，クラスタを特定するには難しい語の情報利得比は低くなる．

### 4.2 要約生成結果

表 1 は TF-IDF 法による計算結果，表 2 は情報利得比を組み込んだ場合の計算結果から上位 10 単語を示す．計算結果を比較すると TF-IDF 法のみで重要語を抽出した場合は,「集ま」,「増やし」など文書を要約する上であまり重要でない考えられる語を含んでいる．しかし，情報利得比を組み込み計算した場合には，上記で示したノイズが除去されている．このことから TF-IDF 法のみをもちいた場合よりも情報利得比を組み込んだ場合のほうが，重要語の抽出においては精度を向上させることができた．

表 1 TF-IDF による計算結果

| 順位 | 抽出語 |
|---|---|
| 1 | 急増 |
| 2 | 安心 |
| 3 | 事件 |
| 4 | 公民館 |
| 5 | 合同 |
| 6 | 施設 |
| 7 | 守る |
| 8 | 集ま |
| 9 | 増やし |
| 10 | 進めて |

表 2 情報利得比による計算結果

| 順位 | 抽出語 |
|---|---|
| 1 | 急増 |
| 2 | 安心 |
| 3 | 事件 |
| 4 | 公民館 |
| 5 | 不安 |
| 6 | 施設 |
| 7 | 守る |
| 8 | 容疑者 |
| 9 | 捜査 |
| 10 | 子供 |

要約文生成においてはこれらの単語を含む文書を抽出し連結することで生成した．生成された要約文についてアンケートを実施した結果，要約内容に関してはおおむね良いという意見が多かったが，文書長に関しては長いという意見が多かった．また，文書中から重要文を抽出し連結しているため，文書としてのつながりを欠いてしまう部分があった．

上記の文書長と文の整合性の問題については，要約内容を加味しつつ，重要語により文書中から重要文を抽出し要約文を生成する過程のアルゴリズムを検討する必要である．

## 5 おわりに

本稿では P2P ネットワーク上での情報共有における，サーバでの情報管理機構の概要を述べ，実装した要約生成手法について述べた．要約生成では文書中からの重要箇所の選定については，TF-IDF 法と情報利得比を組み合わせることで，ノイズが減少しおおむね満足いく結果がえられた．しかし要約文を生成した場合，生成された要約文の文書長が長くなるという傾向が見られた．文書長については要約内容を加味しながら，重要語からの要約生成法を検討する必要がある．

## 参考文献

[1] 大塚雅史, 上原貴夫著「P2P をもちいた情報共有支援における情報管理」FIT2005,2005

[2] 佐々木拓郎，野澤正憲，森辰則著「情報利得比に基づく語の重要度と MMR の統合による複数文書要約」自然言語処理研究会報告，情報処理学会 2002