コンテンツの整合性維持と情報検索を実現する P2P 共有システム

[1]太田 晋　[2]美馬 秀樹　[1]苫米地 英人

[1]コグニティブリサーチラボ株式会社　[2]東京大学大学院工学系研究科

概要

多くの P2P コンテンツ共有システムでは，コンテンツの更新が考慮されておらず，単純な検索機能しか提供されていないため，テキストを含むコンテンツや頻繁に更新されるコンテンツの共有に適していない．こうした問題を解決するため，我々はコンテンツの整合性維持および全文検索を実現する P2P コンテンツ共有システムを開発している．本システムでは，コンテンツの最新版のダウンロード先を中央サーバで管理することにより整合性を維持する．ダウンロード情報の改ざんを防止するため非対称暗号鍵を用いたコンテンツ転送プロトコルを用いる．また，中央サーバに転置インデックスを配置し，クライアントで自然言語処理および検索結果のキャッシュを行うことにより効率的な全文検索を実現する．本論文では，これらの手法と本システムの実装について報告する．

1. はじめに

P2P コンテンツ交換システムは, 少ない資源で大量のコンテンツを流通させられるという優れた特徴を持っている．しかし，実際に運用されている P2P コンテンツ交換システムでは，著作権を侵害した動画や音楽などの交換が行われるケースも目につく．

我々は, 技術面からこうした問題の解決を支援するため，P2P 環境において，コンテンツの改ざんを防ぎ，実際にユーザが行ったコンテンツの交換とシステムの運用者が把握するコンテンツの交換との整合性を維持(以下ではこれを単にコンテンツ交換の整合性の維持という)する手法および情報検索の手法を開発し，成果を標準化することを目標としている[1].

本システムは Napster[2]に類似した中央サーバ型の P2P システムであるが，以下の特徴を備えている．

- **コンテンツ整合性の維持**
  コンテンツの版ごとに電子署名を保持し，非対称暗号鍵を用いたダウンロードプロトコルを用いることにより，コンテンツの改ざんを防止し，サービスの提供者がコンテンツの交換を追跡することを可能にする．
- **全文検索**
  転置インデックスの作成に必要な形態素解析等の自然言語処理は各クライアントで行う．また，検索結果をクライアントにキャッシュすることにより，中央サーバにかかる負荷を軽減する．

本稿では，我々が開発中の P2P コンテンツ交換システムの概要と，そのシステムで使用しているコンテンツ交換の整合性維持手法および全文検索手法を紹介する．

[1]Susumu Ota, Hideto Tomabechi, Cognitive Research Lab. [2]Hideki Mima, School of Engineering, University of Tokyo

2. システムの構成

本研究で開発中のシステムの構成を図 1 に示す. 本システムでは，コンテンツの検索機能を提供するために中央サーバを用いた方式を採用する．

中央サーバでは, サービスを利用する各クライアントの公開鍵が保持される．クライアントは，中央サーバに要求することにより，サービスを利用している他のクライアントの公開鍵を取得することができる．また，中央サーバ自身も秘密鍵を保持しており，公開鍵を各クライアントに公開している．

近年はスケーラビリティなどの観点から中央サーバを用いない方式が数多く研究されているが[3], 既存のコンテンツ交換システムで用いられているような部分一致検索や AND 検索の実現が用意であること，サービスの運用者がユーザ間のコンテンツの交換を把握することが可能であることなどを考慮にいれると，現時点では中央サーバを用いる方式の方がより実現性が高いと考えられる．

この他に，中央サーバは，コンテンツ交換の整合性維持のための機能と全文検索機能を提供する．これらについては以降の節で説明する．

3. コンテンツ交換の整合性維持

3.1. コンテンツ管理のデータ構造

本システムでは，コンテンツの公開者は，自分の秘密鍵を用いてコンテンツに電子署名を行い，コンテンツのユニークな ID とともに中央サーバに登録する．中央サーバは，コンテンツの ID と版に対応した電子署名とを管理する．クライアントは，コンテンツの ID と版を元に中央サーバに問い合わせることにより，コンテンツの版に対応した電子署名を得ることができる．これにより，クライアントは，コンテンツが公開された状態から改ざ

んされていないことを確認することができる.

また, 中央サーバでは, 最新の版を常にダウンロードできるように, 最新の版をダウンロードしたクライアントのリストを管理し, ダウンロード元をリスト内で分散させる手法を用いている.

3.2. コンテンツ交換の記録

本システムでは, 悪意のあるクライアントがコンテンツのダウンロード状況を偽って中央サーバに伝えることを防止するため, 中央サーバで管理されている公開鍵を利用してコンテンツのダウンロードを行うプロトコルを採用している.

このプロトコルでは, ダウンロード先のクライアントの要求に応じて中央サーバが共通暗号鍵を生成し, ダウンロード元のクライアントはその暗号鍵を使ってコンテンツを暗号化する. また, ダウンロード元のクライアントは暗号化したコンテンツに自分の署名を行う.

このプロトコルにより, コンテンツのダウンロード元がプロトコル通りに動作していれば, ダウンロードの記録は必ず行われる. また, 目的のコンテンツがダウンロードできていないにもかかわらず中央サーバでダウンロードの記録が行われることも防ぐことも可能である.

4. 全文検索

本方式では, 全文検索の実現のために, 各クライアントにおいてテキストの形態素解析を行い, 転置インデックスの作成に必要な情報を生成し, コンテンツの更新時に中央サーバに登録する. また, 中央サーバにかかる負荷を軽減するため. 検索結果をクライアントにキャッシュする手法を採用している. 中央サーバは, 長時間接続しているクライアントを一定個数選び, ハッシュ値の値域を等分した区間に割り当てる. クライアントはネットワークに接続する際にこのリストを取得する. クライアントが検索を行うときには, 検索キーワードのハッシュ値を求め, そのハッシュ値の区間に割り当てられているクライアントに対して要求を出す.

[1]における実験では, この手法により, 検索要求が比較的均等にクライアントに分散され, 中央サーバへの要求を減らすことができることが確認できた.

5. システムの実装

本システムのプロトコルは, SOAP を用いて規定している. また, コンテンツの転送を効率よく行うため, HTTP を用いた転送手法も規定している. これらのプロトコルは IETF のインターネットドラフトとして公開する予定である.

本システムは, サーバ, クライアントともに Java を用いて実装しており, [4]の URL より入手可能である.

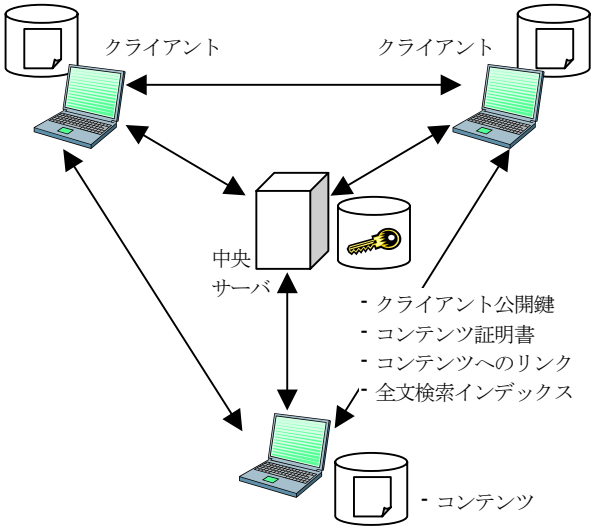

図 1. 本システムの構成

6. まとめ

本稿では, 現在普及している P2P コンテンツ交換システムの問題点を指摘し, 技術面からそれらを解決するための手法を提案した. また, 現在我々が開発中の P2P コンテンツ交換システムの概要について説明した.

本システムでは, 中央に管理サーバを設け, コンテンツ公開者の電子署名を管理することによりコンテンツの改ざんを防止する. また, 非対称暗号鍵を用いたプロトコルにより, ダウンロード状況の偽装を防ぐ.

また, 中央のサーバへの検索負荷の集中を防ぐため, 各クライアントにおいてテキストの形態素解析を行い, ハッシュを用いた手法でクライアントに検索結果を分散させてキャッシュする手法を用いている.

これらの手法により, P2P コンテンツ交換システムにおいて, 高度なコンテンツ交換の整合性維持と効率的な全文検索の実現が可能となることが期待できる.

謝辞 本研究は, 総務省戦略的情報通信研究開発推進制度により, 国際技術獲得方研究開発プロジェクトとして総務省の支援を受けています.

参考文献

[1] 竹辺靖昭, 美馬秀樹, 苫米地英人. 次世代 P2P コンテンツ交換システム –コンテンツの整合性維持と高度な情報検索の実現–. 第 11 回 DPS ワークショップ, 2003

[2] Napster. http://www.napster.com/, http://opennap.sourceforge.net/.

[3] Patrick Reynolds, Amin Vahdat: Efficient Peer-to-Peer Keyword Searching. Middleware 2003.

[4] Cogny. http://cogny.sourceforge.net/