アクセス履歴を利用した コンテンツメタデータベースによる情報流通支援

An Architecture for Supporting Information Distribution: Access Records as Metadata Database

亀田 尭宙[*1] 大向 一輝[*2*3] 武田 英明[*1*2]
KAMEDA Akihiro OHMUKAI Ikki TAKEDA Hideaki

[*1]東京大学 The University of Tokyo
[*2]国立情報学研究所 National Institute of Infomatics
[*3]総合研究大学院大学 The Graduate University for Advanced Studies

Issue of property rights and ambiguousness of reliability have blocked wider use of internet about commercial, legal or academic contents. To solve that problem, we designed and implemented a system to support the circulation of digital contents with providing access records as metadata database. Access records can certify existence of the content itself and the contexts: who, where, how and when it was accessed. The system uses cryptographic hash function to indicate content itself and OpenID to indicate a person who accessed it. Moreover, we propose to estimate reliability of each content using information of network structure among contents based on that database.

1. はじめに

近年は様々な分野において，今まで紙媒体でやり取りされていた情報を電子化することによって利便性向上が図られてきている．CiNii[*1]等の学術文献のデータベースでは論文やそのメタデータをウェブから提供し検索を可能にすることによって，より多くの文献へのアクセスを可能にすると共に，学術情報の間口を広げつつある．物質・材料研究機構の提供する物質・材料データベース[*2]等の科学データベースの公開は，情報の構造化によりその利便性を高めただけでなく，データマイニングによる既存の情報からの価値創造を容易にした．また，学術以外の分野においても，e-文書法[*3]によって企業内の財務・税務関係の法定保存文書を電子的に保存することが可能になり，紙文書の保管コストの削減や検索効率の向上などが期待されているといった動きもある．他にも，特許情報の電子化による公知及び情報取得の容易化，取扱説明書の電子化による環境負荷の低減と保管の必要性からの解放，YouTube[*4]等の動画共有サイトにおける映像を用いたコミュニケーションの登場など，電子化の潮流は多種多様なコンテンツに及んでいる．

その一方で，電子情報の特徴でもある複製や改変の容易性や，ウェブの特徴である全世界への発信の容易性は，不適切な情報の流通を生み結果的に適切な情報の流通を阻害する可能性がある．例えば，コンテンツの複製や改変はその著作権者の権利を侵害する可能性があり，法定保存文書の改竄による偽証の可能性は電子文書による保存を無効にすることになる．また，学術的文書や議論においては，論拠となるリソースへのリファレンスが重要であり，複数の人のアクセスしたリファレンスがウェブ上の URL の場合，前の人がアクセスしてから改変されていないかを確認する手段が必要になる．

また，人がコンテンツを発信利用することを表したネットワーク構造を用いて人やコンテンツの信頼性を推量する際にも基礎となる人やコンテンツのアイデンティティが確かであることが望まれる．そして，信頼のダイナミクスの表現や，先行権利の確認などのためには，時間情報のようなコンテクスト情報が重要になってくる．

よって,「誰が」「いつ」といったコンテクスト情報をコンテンツの確実な同定と共に記録することが求められている．

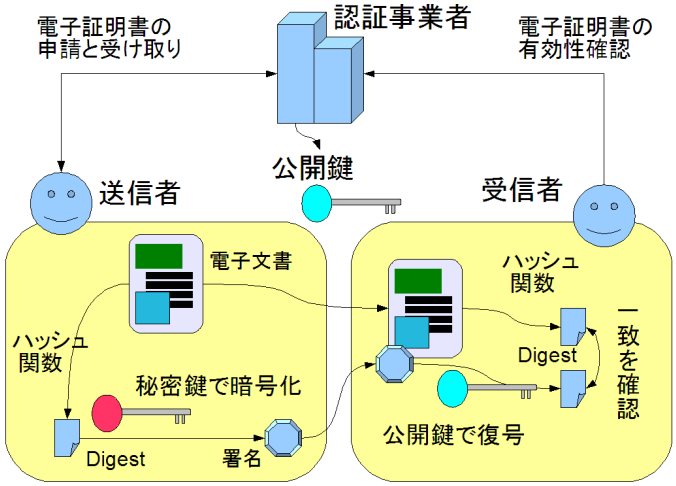

図 1: 公開鍵暗号に基づく電子署名

2. 関連システムと課題

前章のような背景から，電子署名のような技術や Web ページの存在証明サービス[*5]，ウェブ魚拓[*6]といったウェブのキャッシュを保存するサービスが注目を集めている．

電子署名は電子文書に付与する署名情報で，紙文書における印やサインに相当する役割を果たす．一般的には，公開鍵暗号方式に基づくデジタル署名が用いられる（図 1）．山地らは電子署名とタイムスタンプを用いて学術コンテンツの内容証明システムを開発している [Yamaji 08]．また，ウェブのキャッ

連絡先: 亀田 尭宙，東京大学大学院新領域創成科学研究科，千葉県柏市柏の葉 5-1-5, 04-7136-4275, kameda@race.u-tokyo.ac.jp

∗1 http://ci.nii.ac.jp/
∗2 http://mits.nims.go.jp/
∗3 正式には「民間事業者等が行う書面の保存等における情報通信の技術の利用に関する法律」「同法施行に伴う関係法律の整備等に関する法律」を指す
∗4 http://www.youtube.com/
∗5 http://www.existingproof.jp/
∗6 http://megalodon.jp/

シュを保存するサービスは比較的削除の早いニュースサイトなどを引用して議論する際などに用いられてきた．

しかし，これらの技術にはそれぞれ次のような問題点がある。署名の場合は，

- 認証事業者への登録が有料であるので敷居が高い．
- 署名をつけることのできるソフトウェアが限られており，その結果，署名をつけることのできるファイルも限られる．
- 情報の発信者にしか証明を用意する手段が無く，「見たこと」の証明をすることができない．

という問題があり，一方でウェブのキャッシュを保存するサービスでは，

- 証明のための複製自体が著作者の権利の侵害になりうるので，アクセスを遮断しているサイトも増えている．
- 証明できるのが，ウェブ上のコンテンツに限られる．
- 大きなファイルを大量にキャッシュすることはできない．

といった問題がある．

我々はこれらの点を解決するべく，暗号技術的ハッシュ関数を用いたコンテンツの同定と OpenID によるユーザの同定を日時や URL と共に記録することで，情報の提供者および取得者が内容についての証明を得られるようなメタデータベースを構築した．また，このメタデータ同士の関連を記述することでコンテンツの流通をネットワークとして把握し，それを分析することで，コンテンツの信頼性評価に用いることを提言する．

3. コンテンツメタデータベース

先に述べた「誰が」「いつ」以外にも，アクセシビリティ提供のための場所の情報（「どこに」），そしてコンテンツを得たのかアップロードしたのかといった付帯情報（「どのように」）をコンテンツの同定のための暗号技術的ハッシュ関数（「何を」）と共に記録することで，アクセス履歴がコンテンツのメタデータベースの役割を果たすようにする．「いつ」については世界標準時を用い，「どこに」についてはウェブにおける URL を用い，「どのように」については自由書式としたが，「誰が」「何を」について以下に詳しく述べていく．また，このデータベースが扱えるコンテンツの範囲について言及する．

3.1 暗号技術的ハッシュ関数とコンテンツの同定

コンテンツのハッシュ値を得るためのハッシュ関数としては，具体的には Secure Hash Algorithm（SHA）[*7]の中の SHA256 と呼ばれるアルゴリズムを用いた．SHA はインターネット上で情報を暗号化して送受信するプロトコルである Secure Sockets Layer（SSL）などにも使われている関数で，暗号技術的ハッシュ関数（Cryptographic hash function）と呼ばれる種類のもので，次のような条件が強く求められている．

- ハッシュ値の計算が容易である．
- ハッシュ値から元の情報を得ることが困難である．
- 元の情報をハッシュ値を変えずに改変することが困難である．
- 同じハッシュ値を持つ 2 つの異なる情報を見つけ出すことが困難である．

このような強力な関数を用いて一対一対応を保障することで，非可逆性によって著作権やプライバシーなどの権利に抵触せずに，コンテンツを同定可能にすることができる．そして，対象のデータは HTML や XML のようなテキストコンテンツだけではなく PDF や JPEG などのバイナリデータでも扱え，計算の容易性からサーバでもクライアントでも計算できるため非常に汎用性がある．また，ハッシュ値は固定長であるためスケーラビリティの面でも優れており，大規模な検索システムなども実現することができる．

3.2 OpenID とユーザの同定

OpenID[*8]と呼ばれる仕組みで人の同定を行っている．OpenID はあるサービス内におけるユーザのアカウントに対応した URL をユーザの ID とし他のサービスでも使えるように認証を代行する仕組みであり，ソーシャルメディアが増えてきている現在においては，「人」を Web の中に位置づけ，複数のサービスを透過的に利用する手段として普及してきている [Ohmukai 06]．これは，電子署名とは異なり必ずしも現実の人間との対応を保障しないが，ウェブ上の行動の履歴によって人の信頼性が推量されるようになると，同じ OpenID を用いて信頼ある行動をすることが動機付けられると考えられる．

3.3 ウェブだけではない情報資源

また，オンラインデータベースの中の情報や P2P ネットワーク，個人のオフラインリソースにも大量の情報があり，それらの扱いが可能になる点も，ハッシュ値をキーにして情報を纏めることのメリットである．例えば，オンラインデータベースシステムは一般的に，ユーザが欲しい情報をクエリとよばれる情報で指定することで提示するようになっている．そのため，ディレクトリ型の検索エンジン[*9]にクエリを一つ一つ報告し登録を行うのはその量から非現実的であるし，ロボット型の検索エンジン[*10]はリンクされないこれらの情報をインデックスすることができない．このように検索エンジンに登録されないウェブの情報は Deep Web と呼ばれており，その大量な情報をどうやって生かすかということが課題となっている [Bergman 01]．ハッシュでコンテンツを同定し，アクセシビリティをそれぞれのネットワークの基盤システムが担保することにより，このようなコンテンツも変わりなく扱えるので，ウェブ以外のネットワークへの広がりを実現することができる．

4. コンテンツメタデータベースの利用

コンテンツメタデータベースは他のアプリケーションを介してユーザから利用されることを想定している．そのため，API を備えたり，コンテンツの指示のための URI を提供したりしてすることで，連携アプリケーションの開発を可能にしている．また，ユーザの利用のためのサンプルアプリケーションも作成した．全体像は，図 2 のようになっている．

4.1 記録する

外部のアプリケーションから投稿されたコンテンツと人のアイデンティティに関する情報を記録する．JSON による RESTful

*7 National Institute of Standards and Technology(NIST): Announcing the secure hash standard. Federal Information Processing Satndards Publication 180-2, 2002.

*8 http://openid.net/

*9 人手により分類し，登録を行う形式の検索エンジン．Yahoo! カテゴリ http://dir.yahoo.com/ など．

*10 クローラと呼ばれるプログラムが自動的にリンクを辿ることで情報をインデックスする形式の検索エンジン．Google http://www.google.com/ など

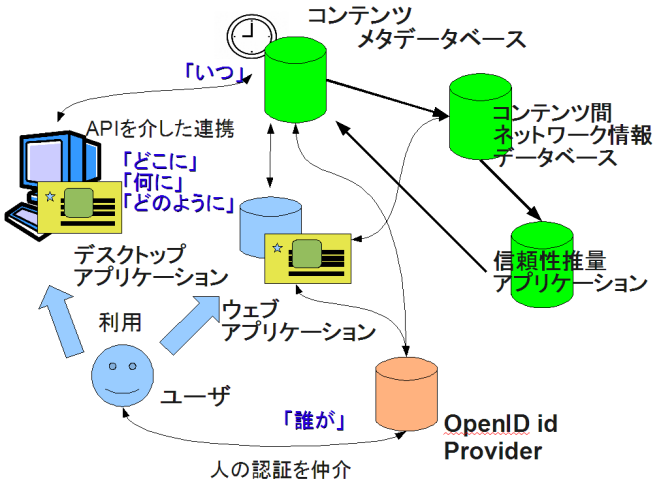

図 2: システムの全体像

な API を備えており，サービス提供者が新たな連携アプリケーションを作成することも可能になっている．例えば，コンテンツメタデータベースが持つウェブ上のインタフェース，外部のデータベースシステムなどが協力的に発信するもの，ユーザがデスクトップアプリケーションとして利用するものなどがある．

4.2 確認・検索する

記録したものを検索し，データとして返す．これによって外部のアプリケーションは，コンテンツに対しての関連する利用法を提示したり，同じコンテンツを有する人を検索したり，ある URL におけるコンテンツの変遷を確認したりすることができる．

4.3 指示・アクセスする

ハッシュ値から URL に結びつけてアクセスを可能にするサービスを URL を HTTP リダイレクトを用いて提供している．これは，コンテンツメタデータベース自体はアイデンティティの提供のためにアクセシビリティから切り離されて設計されているので，外部のアプリケーションやデータがコンテンツメタデータベースによるアイデンティティをできる限りアクセスのために用いられるようにするための機能である．また，URL の形式でアイデンティティを簡便に統一的に記述することができるというメリットもある．

5. コンテンツの関連付け

さらに，コンテンツメタデータベースに基づいて自らがアクセスしたコンテンツ同士の関連を記述できるようなウェブアプリケーションも提供している．これは各人の知識の体系化をインセンティブとした活動で，体系化された知識は，本人の思い出しや知識の利用を促進すると共に，他人の知識を参照する場合もその体系を通して理解しやすくするのに役立つ．また，外部のデータベースなどからデータを引き受けて処理を行うアプリケーションは，元のデータと自らが生成したデータの間の関連を知っているので，自動的にその関係をメタデータ編纂システムの方へ送ることができる．さらに，自然言語処理などの知能的処理によってエージェントがコンテンツを関連付けることもできる．そして，これらの活動を通じて生成されたデータがコンテンツの関連付け情報として蓄積され，全体としてコンテンツのネットワークを表現することができる．この関連付けは，次章の信頼性推定のためのデータになると共に，既に述べた Deep Web の問題に対し，今あるコンテンツをキーにして他のコンテンツを探す考え方である Query By Example という方法でコンテンツへのアクセスを提供することを可能にしている．

6. コンテンツの信頼性評価

前章のネットワーク構造のデータを元に，PageRank アルゴリズムによる個々のコンテンツの信頼性の推量を実装した．PageRank は Google[*11] のランキングで用いられているアルゴリズムで，ネットワークの重み付き隣接行列の固有値を計算することで，有向のグラフにおける入次数中心性の相対的な強さを，「より信頼できるものから参照された資源は，信頼のない資源から参照されたものより，信頼に値すると考えられる」という推移的な関係を反映して推量している [Page 98]．隣接行列の重みはリンク元の出次数の逆数になっている．

現在はコンテンツの信頼性の和を，それを発信した人の信頼性としているが，今後人を含めたネットワークの中から相互の人とコンテンツの相互のかかわりを考慮に入れた信頼性の推量アルゴリズムに取り組みたい．

7. おわりに

本研究では，コンテンツの流通を支援するために，情報の提供者および取得者がコンテンツに関わる証明をするための手段の提供，それに基づくコンテンツ同士のネットワーク構造の記録，さらにそのネットワークに基づく信頼性評価を行うシステムを提案及び実装した．コンテンツの証明手段の提供は，権利や正確さが問題になってくるようなコンテンツに対してもインターネットが社会における情報流通に貢献するために重要であるだけでなく，上位層のアプリケーションがコンテンツのネットワークを把握する上で重要な情報のアイデンティティをその基盤として提供することになる．ネットワーク構造の記録については，他のアプリケーションやアルゴリズムを使用した人手及び人工知能によるコンテンツ同士の結びつけによって多様で豊かな情報が共有されることが期待される．

参考文献

[Yamaji 08] 山地一禎, 片岡俊幸, 行木孝夫, 曽根原登: プレプリントへの長期署名付与および検証システムの構築, Journal of Japan Society of Information and Knowledge, Vol.18, No.3, pp.240-248, 2008.

[Ohmukai 06] 大向一輝: SNS の現在と展望 -コミュニケーションツールから情報流通の基盤へ-, 情報処理, Vol.47, No.9, pp.993-1000, 2006.

[Bergman 01] Bergman, M.:The Deep Web: Surfacing Hidden Value, Journal of Electronic Publishing, Vol.7, No.1, 2001.

[Page 98] L. Page, S. Brin, R. Motowani and T. Winograd: The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Progect, 1998.

*11 http://www.google.com