Vol.2010-IOT-8 No.21
2010/3/1

分散したコンピュータによる公開コンテンツ観測システムの提案

永井俊行[†1]　坪川宏[†1]

Web による情報発信は，それがある時点に行われていたという事実が残らない．そのため，情報の初出が定められないという問題や，無責任な情報発信を容易にしている．情報発信事実を記録する既存のサービスは，その運用を絶対的に信頼しなければならない点で問題があると考えている．そこで，不特定多数の協力者によって情報発信事実の記録を行うことを提案し，その実現手法について述べる．

A Proposal of Observation System for Public Contents by Distributed Computer

Toshiyuki Nagai[†1] and Hiroshi Tsubokawa[†1]

The fact that published a document on the web doesn't remain. Therefore, there is a probrem whether the document is the original and the document is shown irresponsibly on the web. In existing archive service, the user must trust the web site which gives the service. In this paper, we propose the solution by the observation using a distributed computer and explain the realization approach.

1. はじめに

インターネットが一般の人々に普及したことにより，人々の情報との関わりが大きく変化している．インターネットが人々にもたらした最も大きな変化として，個人が簡単に情報発信を行うことを可能となった点が挙げられる．人々はインターネットを利用することにより，既存の大規模な流通網やメディアに頼ることなく，個人が世界規模での情報発信を行えるようになった．

特に Web がインターネット上のアプリケーションとして利用されてきた．電子掲示板といった Web アプリケーションを利用することによって，人々は簡単に情報発信を行うことができ，情報の提供や議論の場などになってきた．近年では，例えば Wiki による複数人でのコンテンツの共同編集など，インターネットはただ情報を発信する場ではなく，相互効果により情報を生み出す場としての効果を強めている．

インターネットを利用した情報発信は活発化しているが，Web での情報発信には既存のメディアにはないある特徴がある．それは，ある情報発信がある時には行われていたということを明確に示すことができないという特徴である．これは，Web が既存の出版といったメディアと異なり，ネットワークを通じて配布される電子的なメディアであることと，情報発信の主体は様々で，分散的に随時情報発信が行われるためであると考えられる．この特徴により，次のような状況において問題が発生する．

(1)　Web 上にある情報の引用における問題
(2)　Web で公開された情報の盗用や独占利用が起こる問題

まず（1）の問題について述べる．Web では Blog と呼ばれるスタイルが定着し，各々は他人の Blog やニュース記事などを引用し，それに対しての意見や感想を言及している．この際に通常は引用元の出典が併記される．しかし Web 上のコンテンツが出典の場合，ある時点ではそれが既に削除されたり，内容が更新されている可能性がある．その場合，引用文がその出典で確かに過去に発信されていたということがわからなくなる．そのため，その記事を書いた引用者が誤った引用を行っている，または捏造された引用を行っているといったことが考えられ，読者は記事を信用することができず，内容について評価を行うことができなくなる．

次に（2）の問題について述べる．Web 上に存在する情報は，それがいつから公開されている情報であるかわからない．つまり，ある情報の初出の日時と場所 (URL) を決定することができない．そのため，Web 上に公開した創作物が盗用される恐れがあり，オリジナルを創作し公開した者は，盗用されたことを示すことが難しい．また，共有を目的として情報発信された技術やアイデアなどが，第三者によって特許出願されることで，第三者に独占利用を許してしまう恐れがある．実際に特許庁では，インターネット上の情報について，公的機関などによるものではない Web ページを疑義のある情報とし，情報の管理者への照会を基に判断を行うものとしている[1)]．これは，インターネット上の情報について，正確な初出

†1 東京工科大学大学院バイオ・情報メディア研究科
Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology

日時を知る方法がないための措置である．そのため，一般的な Web サイトの情報は初出について確認できず，既にインターネット上で公開されていた新規性のない技術についての特許登録と独占利用が認められてしまう恐れがある．

視点を変えて（1）の問題を見ると，情報発信者は都合の悪くなった記事を随時変更，削除ができると考えられる．情報を発信する際のモラルを緩める原因ではないかと考えられる．実際に，Web を通じた安易な憶測記事や誹謗中傷，著作権侵害といった行為は跡を絶たない．また（2）の問題は，人々にインターネットを用いた情報発信に対して懸念を生み，さらなる情報の流通やネットワークを活用した創造活動を抑制する原因になるのではないかと考えられる．

こうした Web における問題を改善するため，ある時間にある場所（URL）である情報が発信されていたということを，記録する手法が必要であると考えられる．現在いくつかの手法が提案され，Web サービスとして提供されている．しかしこれらのサービスには，

- コンテンツに対する視点が単一的である
- サービスが公平かつ正常に運営されていることを信用する必要がある

といった点で，問題があるのではないかと考えている．

そこで本論文では，インターネット上の分散して配置した多数のコンピュータの協力により，ある Web コンテンツが発信されている事実を記録する手法の提案を行う．インターネット上に，ボランティアなどによって分散してコンピュータを設置し，ネットワークを構成する．そして分散してコンテンツを観測することで，アクセス元が単一的にならない客観的な記録を可能とする．また，多数のコンピュータにより観測を行い比較するため，従来のような信頼できる第三者を必須としない．従来手法に対して，他システムが見た事実を記録し主張するのではなく，自身の記録に，客観的な事実を付与するという点に特徴がある．記録した事実は Web 上で引用に用いることができ，言及に活用できる．

以降では，2 章で既存の手法の問題点を挙げる．そして 3 章で提案手法の概要を述べる．4 章では提案手法を実現するための技術について述べる．5 章では，さらに詳細なシステムの構成と実際の流れについて述べる．

2. 関 連 手 法

ある過去の時刻にある情報が発信されていたという事実を記録するため，あらかじめ Web 上の情報を保存しておくことが考えられる．こうした情報発信の事実を記録するサービスとして，いくつかの Web サービスが存在する．

“ウェブ魚拓” は，Web サイトを簡単に記録して，ブログなどからリンクを張って言及に利用できることを目的として開発され，提供されている Web サービスである．

まずユーザは，記録する Web ページの URL をシステムに入力する．ウェブ魚拓のシステムは，入力された URL のコンテンツを取得し，そのコンテンツをサーバに蓄積する．蓄積されたコンテンツには” ウェブ魚拓” による新たな URL が付与され，記録を依頼したユーザ以外にも，一般に公開される．ユーザは新たな URL を用いて言及に利用することによって，元の Web ページのコンテンツが変更や削除されても，恒久的にある時点での Web ページについて言及することができる．

しかしこの場合，元のコンテンツの著作者以外が別のサーバにコンテンツの複製を置き公開することになるため，著作権の面で問題がある．コンテンツの著作者によって公開の停止を依頼された場合は従う必要がある．また，“ウェブ魚拓” はコンテンツ内や robots.txt によって，“ウェブ魚拓” によるキャッシュの作成の拒否を表明する手段を提供している．そのため，無責任な情報発信を抑制する効果がない．

そこで，“引用する” という機能が用意されている．この機能は，

(1) ウェブ魚拓は取得したコンテンツを画像化
(2) ユーザが矩形で引用範囲を指定
(3) 引用範囲以外に画像処理でぼかしを施す
(4) 画像として一定期間サーバ上で公開
(5) 一定期間終了後は期間中にアクセスのあった Web ページからのアクセスに対してのみ公開を継続

という手順で引用を実現し，ある Web ページに対する恒久的な言及を行うことができる．引用する者と引用される者から見て第三者であるサービスのサーバ上に複製を用意する仕組みのため，このような複雑な処理を必要としている．

“Web ページの存在証明サービス” は，Web ページがある時刻にある URL に存在したことを証明することを目的として開発された Web サービスである．

“ウェブ魚拓” と同様に，まずシステムはユーザが指定した URL のコンテンツを取得する．ここで，システムは取得したコンテンツからディジタルタイムスタンプを生成する．そして，取得したコンテンツとディジタルタイムスタンプをまとめてアーカイブしユーザに送付する．

後にユーザがその記録を証明するには，アーカイブ内のコンテンツとディジタルタイムスタンプをシステムに送信することによって，ディジタルタイムスタンプの検証結果を得るこ

とができる．そのため，コンテンツの内容が改竄されていないことを証明でき，過去のある時刻に自分が見た内容が，本当に存在していたことを第三者に証明できるとされる．

“ウェブ魚拓” は複製を作成することで引用や言及に利用する目的としているのに対して，“Web ページの存在証明サービス” は，取得したコンテンツが取得時点より変更されていないことを証明することを目的としているという違いがある．そのため，作成したアーカイブは Web 上での言及のために利用することができない．なぜなら，作成したアーカイブを Web に掲載した場合，ウェブ魚拓と同様に著作権上の問題が発生するためである．もしもコンテンツを引用のため必要範囲のみを取り出した場合，ディジタルタイムスタンプの検証の結果コンテンツが改竄されたと判断される．

これらのサービスに共通する問題として，あくまでコンテンツの取得を行うのはユーザではなくサービスであるということが挙げられる．そのため，記録される Web ページがユーザが見たページとは大きく異なってしまう可能性がある．例えばこうしたサービスからのアクセスに対して，一般に公開しているものとは違ったコンテンツを返す Web ページがある場合，サービスは客観的でない事実に対して記録を行い主張することとなる．また，本来はユーザが得ることができない，アクセス制限されたコンテンツをシステムが配信してしまうという問題もある．

また，サービスが公平な第三者として振る舞い，正常にシステムが運用されていることを期待し，信用する必要がある．もし，システムを運用する者に悪意があった場合，記録されるコンテンツを改竄することが可能である．また悪意がなくとも，システム内部の不具合や外部からの攻撃により，誤った記録が行われる可能性がある．

そのため，視点が単一的である従来の手法では，客観的な情報発信事実の記録を行っているとは言えないと考えられる．

3. 提　　案

3.1 概　　要

従来手法は情報発信の記録を単一の主体が行うため，その記録の客観性に問題があることを述べた．そこで，単一の第三者に依存しない，分散されたコンピュータを用いた情報発信事実の記録手法を提案する．

提案手法は，学術機関やボランティアなど独立した組織によりコンピュータを設置し，それらのコンピュータを用いて多角的な観測を行う点を特徴とする．既存の手法が一人の審判による判定と例えることができるのに対して，提案手法は，複数のコンピュータの観測情報を集め，それらを比較することで，情報発信事実の記録についての客観性を高める．提案手法の概要を**図 1** に示す．

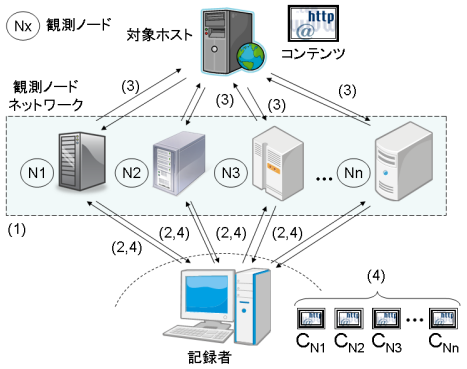

図 1　提案手法の概要

ここで，提案手法によって実現されるシステムは，次のように構成される．

観測ノードネットワーク　システムに賛同する学術機関やボランティアなどにより，分散してネッ トワーク上に設置されるコンピュータを，観測ノードと呼ぶ．観測ノードはお互いの存在を知っており，ネットワークを構成するものとする．観測ノードは随時システムに参加することができ，その際は既存の姦賊ノードと情報を交換し，システムへの参加処理を行う．

情報発信事実の記録　観測ノードネットワークを利用して情報発信事実の記録を行うユーザのことを，記録者と呼ぶことにする．記録者は，まず既知の観測ノードに問い合わせを行い，観測ノードリストを得る．次に観測ノードの中から N 台を選択し，選択した N 台に対して，URL などを指定し，情報発信事実記録のための観測を依頼する．

コンテンツの取得と観測情報作成　依頼を受けた観測ノードは，依頼された URL のコンテンツを取得し，時間の情報などと共に，記録者に配信する．これを観測情報と呼ぶ．

観測情報の収集　記録者は集めた観測情報がそれぞれ一致することを示し，観測情報が客観的な情報であることを示す．

以下では，提案手法を実現するためにの核となる部分について述べる．

3.2 観測ノード認証

記録者は様々な観測ノードより観測情報を集め，それぞれが一致することを示すことで，

客観的な情報発信事実の記録を主張する．しかし，確かにそれぞれの観測情報を作成したのが，それぞれの観測ノードであることを示せなければ，観測情報はただの出処不明のデータに過ぎない．そこで，観測ノードはそれぞれが秘密鍵を持ち，観測情報を作成する際は，観測情報を作成した責任の所在を明確にするため観測ノードによるディジタル署名を付ける．

一般的に，ディジタル署名の検証はディジタル証明を用いてオフラインで行われる．あるディジタル署名について，ディジタル証明書で示される公開鍵を用いて検証することで，その署名がディジタル証明書で示される組織や個人により生成されたかどうかを確認できる．しかし，ディジタル証明書の発行には信頼できる第三者が必要である．これをシステムで用意するとなると，認証局 (CA) や登録局 (RA) の運用が必要となり管理者やリソースを必要とする．また既存手法と同様に，局を運用するシステムの提供者が，記録の信頼性について責任を負うことになる．

これらは設計方針に反するため，証言ノードは信頼できる第三者によるディジタル証明書を求めないことにする．つまり，自己認証局と自己署名証明書でもシステムに参加可能とする．もちろん，それだけでは確認者から見て信頼できる第三者へ認証パスを辿れない (検証できない) ため，証明書自体の意味は無くなってしまう．

そのため，確認者は証明書で示されるノードに対してオンラインで実在の確認を行うことが有効であると考えてある．ある観測情報に付与された署名について，付随するディジタル証明書を基に，観測情報を生成した観測ノードに問い合わせを行う．観測ノードは確認者に対して，ディジタル証明書に内包される公開鍵に対応する秘密鍵を有していることを示す．

このプロセスにより，確認者は署名を行った者がシステムに参加していることが確認できる．そして，IP アドレスといった情報により，記録者と異なる複数の組織から観測情報が得られ示されていることを知ることができる．

3.3 ブロックハッシュ処理

システムにおいて，記録者は指定した URL のコンテンツを，多数の観測ノードを介して取得することができる．この時，アクセス制限されているために記録者が本来取得できないはずのコンテンツを，記録者が取得することができてしまう可能性がある．例えば，観測ノードが設置されたイントラネット内のコンテンツを取得できてしまう恐れがある．

こうした問題に対し，各観測ノードがブラックリストやホワイトリストといった手法を用いて制限することも考えられる．しかし，リストの保守作業が発生する，人為的ミスが発生する，といった問題がある．そのため，観測ノードは取得したコンテンツを，記録者によって理解できない形式に変換してから渡すことが望ましい．

しかし提案手法では，記録者は多くの観測情報を集めそれらを比較することによって，客観的な観測結果の一致を示す．そのため，得られた観測情報は比較が可能な形式でなければならない．すなわち次の二つを満たす処理が必要とされる．

- 記録者に対して元のコンテンツを漏洩しない
- コンテンツ同士の比較は可能である

そこで，コンテンツからハッシュ値を算出し，記録者には観測情報としてコンテンツ自身ではなくハッシュ値を送付することが考えられる．しかし Web コンテンツに特徴的な点として，同じ URL で同じ時間に取得したとしても，まったく同一のコンテンツが取得できるとは限らないという点がある．例えば，コンテンツ内に埋め込まれた動的生成日時，アクセスカウンタ，広告などである．そのため，各観測情報として得られるハッシュ値が異なり，比較ができなくなってしまう．また，コンテンツ全体に対してハッシュ値が算出されるため，記録者は観測情報を引用に用いることができない．

この問題を解決するために，コンテンツの分割処理を行い，それぞれについてハッシュ値の計算を行う．

(1) 証言ノードはコンテンツを取得し，これを文書 M とする．
(2) 文書 M を，コンテンツの形式に適した方法で，部分文書 $M_0 \dots M_n$ に分割する．
(3) $M_0 \dots M_n$ からそれぞれハッシュ値を算出し，$H_0 \dots H_n$ を得る．

以上の操作をブロックハッシュ化と呼び，こうして得られた $H_0 \dots H_n$ を観測情報のコンテンツ部分とする．観測情報はハッシュ値の列であるため，ユーザは元のコンテンツが復元できない．記録者は，記録者自身で対象のホストより取得したコンテンツを同様にブロックハッシュ化することによって，観測情報と自身のコンテンツと比較することが可能である．また，複数の観測ノードから返された観測情報同士を比較が可能である．

また，ユーザは証言情報を用いて引用を提示する場合，引用部分以外をハッシュ値での置き換え，または引用部分とそのインデクスのみ残し他を削除することによって，引用部分のみを公開し参照するといったことが可能となる．

3.4 ホストへの負荷抑止

システムでは記録者の依頼により，対象の URL のコンテンツを複数の観測ノードによって取得する．そのため，対象の URL で示されるホストに短期間に複数のアクセスが発生し過負荷を与えるおそれがある．

このような，場合によって対象のホストをサービス拒否に陥らせてしまうことは，システムとして許されるものではないと考えている．同時に複数の記録者が利用する可能性があ

り，各記録者が自由なタイミングで依頼を行うことができることを考えると，各観測ノードが自律してホストへの過負荷を抑止する必要がある．

しかし，それぞれの観測ノードがホストへの現在の負荷を計測する場合，その観測自体が負荷を与える原因となる．例えば，Ping により応答速度を定期的に計測する場合，その計測行為がホストへ負荷を与える．また，アプリケーション層での負荷は計測できない．アプリケーションレベルの負荷を計測しようとすると，ますます負荷を与えることとなる．

また，負荷の計測を行わずに各観測ノードが，それぞれ単一ホストへの同時アクセス数やアクセス間隔を制限することを考えてみる．この場合でも，システムに参加する観測ノードが多くなる程，ホストへの負荷が増える可能性があるという問題がある．

そこで，システムで一律に対象ホストへの同時アクセス数を制限する必要があると言える．この場合，あるホストへのシステムからの同時アクセス数をカウントし，管理と判断を行うコンピュータが必要となる．しかし，そのような特別なコンピュータを必要とすることは設計方針に反する．

これを解決するため，同時アクセス数を管理する役割を，観測測ノード全体に動的に分散することができれば，各観測ノードの公平な負担によりアクセス数の制限を行うことが考えられる．そのために，Consistent Hashing の手法を用い，対象ホストごとに同時アクセス数の制限を管理する観測ノードを決定することを考える．観測ノードに割り当てられたその役割を，便宜的に制御ノードと呼ぶことにする．観測ノードは記録者の依頼に応じて対象のホストにアクセスを行う前に，制御ノードより対象のホストへのアクセスする許可を得る必要があるとする．

(1) 観測ノードの ID の決定
まず各観測ノードに重複のない ID を割り当てる．例えば，H をハッシュ関数とし，$H(IP\ address)$ や，$H(Digital\ Certficate)$ として決定する．これを NID とする．各観測ノードは，例のように一貫して求めるか通信により伝達することにより，お互いの ID を理解しているものとする．

(2) 対象ホスト固有の ID の決定
対象のホストから固有の ID を算出する．例えば，対象ホストの $H(IP\ address)$ として決定する．これを CID とする．NID と CID は同じ範囲をとる整数値であるとする．

(3) 制御ノードの決定
ある CID を担当する観測ノード NID(制御ノード) を一貫した手法で決定する．例えば，ID のとる最大値の次が ID のとる最小値となるリングと考え，$NID > CID$ となる最小の NID を制御ノードとする．

(4) アクセスの許可要求
観測ノードは，前手順で決定した制御ノード NID に，アクセスの許可を要求する．

(5) アクセス許可の判断
制御ノードは自身のデータベースから，対象ホストへのアクセス数を参照する．アクセス数の制限値に達していなければ許可を返答し，データベースを更新する．達していなければ不可を返答する．

(6) 対象ホストへのアクセス
返答内容が許可であれば，対象ホストへのアクセスを行う．不可であれば，一定時間待機した後 4 から再度実行するか，記録者に混雑を通知し依頼の実行を中止する．

(7) 完了の通知
対象ホストへのアクセスが完了すると，制御ノード NID へその旨を通知する．制御ノードは，データベースを更新する．

Consistent Hashing はもともと，複数台でウェブのキャッシュサーバを構成する状況を想定して考え出された．本提案において，制御ノードを決定するための ID として資意的な値ではなく乱数やハッシュ値を用い，キー (提案手法では対象ホスト) が偏りなくやってくるならば，各観測ノードの制御ノードとしての担当数に，高い確率でひどい偏りは生じないため，公平な分散であると考えられる．また，このノードの ID とキーの ID で同じ ID 空間を有し距離で結びつけるこの方法は，観測ノードの増減に対して，ある対象ホストへの制御ノードの担当が代わるのは，全体の $1/N$ 程度で済むという利点がある．

4. ま と め

本論文では，分散したコンピュータによる観測システムについて提案し，概要を述べた．そして，提案手法に関して，必要な機能や実現方法について述べた．提案した手法については，今後も議論を進めていく必要があると考えている．

参 考 文 献

1) 特許庁:インターネット上の情報の引用文献としての取り扱い運用指針, 特許庁ホームページ, 特許庁（オンライン）, 入手先 ⟨http://www.jpo.go.jp/tetuzuki/t_tokkyo/shinsa/unsisin.htm⟩(参照 2010-01-28).