D_039

協調型Webアーキテクチャのためのリンク情報同期方式の検討

Comparison of synchronization methods for Cooperative Web Architecture

小林 亜樹 † 山岡 克式 ‡ 酒井 善則 ‡ 曽根原 登*
Aki Kobayashi Katsunori Yamaoka Yoshinori Sakai Noboru Sonehara

1 はじめに

PageRank の成功に見る Web 特有の価値であるリンク構造を利用した，Web 構造マイニングに代表されるリンク情報を用いる技術の多くは研究段階 [1]–[2] にあり，商用サービスは Google など少数に留まる．これは，Web のコンテンツ配送を支える HTTP が（HTTP/1.1 での拡張を考えても）基本的に単一コンテンツを取得するプロトコル（図 1(a)）であるためクローリングを要し，研究として成立 [3]–[5] しても，使い勝手などでの障壁があるものと思われる．集中型検索サービスが，文脈依存検索にも利用 [6] されるなど，リンク構造をも考慮した情報探索のエージェントとして利用されている．

リンク構造の有用性が明らかにされつつある中，その情報取得を情報源の外部の巨大設備に頼る方法では，単にスケーラビリティ*1だけでなく，深層 Web への対応性や，分散型のインターネットのアーキテクチャにそぐわないなどの問題が指摘される．そこで筆者らは，Web サーバが最低限の協調性を備え分散的な手法（超分散的手法）で，部分的リンク情報（部分 Web グラフ）を提供（図 1(b)）できる，協調型 Web アーキテクチャ（Cooperative Web Architecture; CWA）を提案 [7] している．

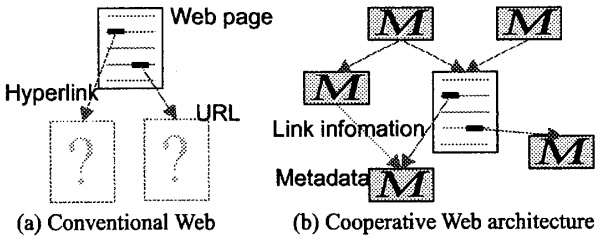

図 1 コンテンツ取得アーキテクチャ

2 協調型 Web アーキテクチャ

協調型 Web アーキテクチャは，リンク構造を活かした情報探索やマイニング，ナビゲーション支援などを目的に，目的の Web コンテンツを中心とする，部分的な周辺リンク構造やメタデータを一括して提供できるよう Web サーバの機能を拡張する．同様の機能性を外部サーバによって実現しようとした研究 [8] もあるが，深層 Web の広がりを考えるとサーバ自体の機能拡張が将来性の面で有利であると考える．

本機能性のうち本質的な問題は，周辺リンク構造をいかにしてサーバが「知り」「維持管理していく」か，にある．このリンク情報をサーバ間で共有し同期し続けることを，部分 Web グラフの同期と呼ぶ．一見，この問題は IP における経路情報の共有と同じであるが，Web グラフの大きさから全てを共有することは不可能であり，また，階層的に分割することは，自由にリンクを張ることのできる Web コンテンツの特性を犠牲にする．そこで，コンテンツ毎に一定の範囲内の部分 Web グラフ情報を保持し，維持管理することを，サーバをエージェントとして実現する本アーキテクチャを提案する．

一方，blog に見られるトラックバックの仕組みは，利用者が「誰から」参照されたかを欲している端的な事例であり*2，部分 Web グラフの応用を考えても逆リンクを辿れることが重要である．逆リンクを他のコンテンツが（サーバをエージェントとして）知るためには，リンク元コンテンツがリンクの存在を知らせればよく，更新通知による同期方式を提案 [9, 10] している．

3 部分 Web グラフ同期問題

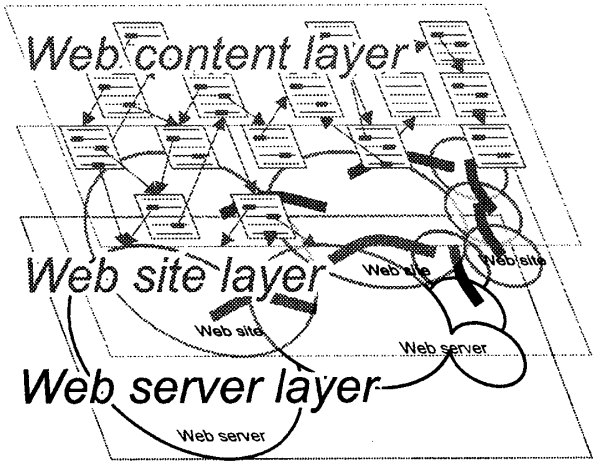

図 2 CWA のレイヤーモデル

部分 Web グラフ同期は，CWA 実現の核となる技術である．サーバを媒介としてコンテンツ間における更新通知を用いることは既に述べた．このとき，情報更新の確実性，すなわち同期の整合性は，サーバを考えずコンテンツ間での通知という形で議論できる．また，コンテンツ層でのリンク空間の変化は，リンクの追加と削除に還元される．更新通知のコンテンツ管理への応用や通信量を議論する際には，図 2 のような Web サイトや Web サーバの層へと写像して，方式に修正を加える．

† (独) メディア教育開発センター, National Institute of Multimedia Education
‡ 東京工業大学大学院理工学研究科, Tokyo Institute of Technology
* 国立情報学研究, National Institute of Infomatics

*1 Google では，過負荷時に「ただいまキーワード検索には多大な要求が寄せられ，結果を表示できない場合があります．ご迷惑をおかけしますがご了承ください．」と表示される

*2 同時に，表層的な HTML においてリンクの方向性を失わせる「強制的な双方向リンク」である点で，machine readable な Web から遠ざかるものである．

各コンテンツにおいて部分 Web グラフ情報を提供できるようにするためには，提供するグラフ情報の範囲を定める必要があり，コンテンツ間の距離が別に定める閾値以下であるようなコンテンツと定義する．ここでコンテンツ間の距離の定義には，直観的なリンクによる最短パスや，コンテンツ間のフロー，あるいはコンテンツ間の意味的類似性を用いるものなどが考えられる．スケーラビリティの確保のためリンク構造としての近さを反映し，制御の単純さを目的に，リンクの追加や削除の発生時に当該リンク元，リンク先コンテンツがそれぞれ管理している Web サブグラフの範囲を越えて，その影響が及ばないことが望ましい．本稿では，最短パスを用い，便宜上各リンク距離は同等とする．

4 部分 Web グラフ同期方式

4.1 バケツリレー

ネットワークプロトコルなどで使われるバケツリレー方式は，リンク hop 数が閾値以下のコンテンツまでのフラッディングとして知られる通知方式となる．

通知内容を更新のあった当該リンク情報のみに限定することが原理的には可能である．この場合，リンク link($c \to d$) の更新時にコンテンツ c, d の管理する Web サブグラフ範囲 $\langle c\rangle \cup \langle d\rangle$ に含まれるリンク本数 $|\langle c\rangle \cup \langle d\rangle|_l$ 程度の通知が必要である．通知自体に $\langle c\rangle$, $\langle d\rangle$ の各コンテンツへの通知状態を持たせても，後述の直接更新通知並のコンテンツ数 $|\langle c\rangle \cup \langle d\rangle|_c$ 程度まで抑えることはできない．

4.2 直接更新通知

最短パスをコンテンツ間の距離として用いる場合，リンク link($c \to d$) の更新時に当該コンテンツ c, d の管理範囲 $\langle c\rangle \cup \langle d\rangle$ を越えて，部分 Web グラフ情報を更新しなければならないコンテンツはないことが保証されるため，管理範囲内のコンテンツへ直接更新情報を通知する直接更新通知方式を提案した [9]．通知内容は，必要に応じて管理範囲内の部分 Web グラフ全体を含むこともあるが，通知数は原則として $|\langle c\rangle \cup \langle d\rangle|_c$ 程度に収まる．

しかし，あるコンテンツ c において，$\langle c\rangle$ が増加するような更新が c 以外で発生したにも関わらず，当該通知到着以前に c において他の更新が発生した場合に，通知が行き渡らない「通知範囲問題」が存在する．また，間接的な更新情報も通知されることから，通知受信時の順序によって，破棄すべき情報による上書きや，その逆の破棄が発生する，「受信順序問題」がある．

4.3 順序制御更新通知

順序制御更新通知方式 [10] がこれを解消する．通知範囲問題を，更新発生後一定時間に渡り未通知のコンテンツが，管理範囲内に入った場合に当該更新通知を発し，受信順序問題は，一定時間内に連続的に受信した通知が互いに矛盾する場合，当該リンクを直接保有するコンテンツからの情報を優先することで解消を図る．

リンク情報に付随する情報は最小限で済み，自律的に整合性を確保できるが，通知の遅延や順序の入れ換えが発生する最大時間を見積もっておく必要がある．実際の通知が Web サーバ層において行われることを考えると，Web サーバのネットワーク上の距離や処理能力，負荷状況に依存するこのような時間の見積もりは難しい．

4.4 順序付与更新通知

コンテンツにおいて自身のリンクが更新される度に順序番号（sequence number）を付与し，通知にはその時点における自身の順序番号を併せて通知する方式である．受信した情報にも各コンテンツの順序番号を保持しておくことで，到着した情報との新旧を判別できる．

図 3 は，c, d において相前後して link の更新が発生したとき，d が c に対して $\langle c\rangle$ が増加したことを通知する様子である．ここで，コンテンツ名左肩の数字が順序番号であり，当該コンテンツでのリンク更新毎に 1 増加される．notify:$d \to c;\langle d\rangle$ の通知によって，c は $\langle {}^{i}c\rangle_{k+1}$ へと管理範囲が増加したことを知るが，この通知が届く前に c 自身のリンク情報の更新が発生していると，その通知は $\langle {}^{i}c\rangle_k$ に留まり，この差が不整合となる．

そこで，順序付与更新通知方式では，d からの通知に，その時点で d の知る c の状態，すなわち ${}^{i}c$ を付与し（notify:$d \to c;{}^{i}c|\langle d\rangle$），${}^{i+1}c$ のように c がさらに更新された状態であればこの差分情報を $\langle c\rangle_k \cup \langle d\rangle$ 全体に通知することで，この問題を解決する．

また，原理的に古い情報での上書きは発生しない．しかし，受信順序問題のうち，管理範囲外の部分 Web グラフを受信した際に，本当に破棄しても良いかの判断には，やはり一定時間の待機が必要である．

本方式は，各コンテンツに新しさを表す順序番号を付与して管理し，通知時にもその情報を追加する必要があるが，これは比較的小さな情報であり，見積もりの難しい待機時間の設定が不要な点が利点である．

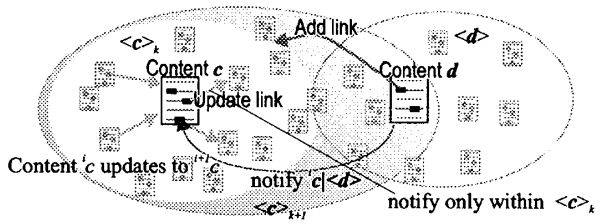

図 3 通知範囲問題

5 おわりに

協調側 Web アーキテクチャを実現する上で核となる部分 Web グラフ同期方式について，新たに順序付与方式を導入し，他方式との理論的な比較を行った．実装に当たって，通知量や通知数，また，DB の管理コストなどがサーバ負荷となるため，引き続き検討したい．

参考文献

[1] 村田 他, "Web 構造マイニングと Web コミュニティ発見", 情処研報 2006-DBS-139, pp.85-90, 2006
[2] 宇野 他, "孤立クリークを用いたウェブ構造マイニングとリンクファームの検出", IEICE WI2-2006-26, pp.83-88, 2006
[3] 内藤 他, "超分散サーチエンジンを用いた効率的情報探索について", 信学技報, Vol.99, No.674, IN99-163, pp.123-128, 2000
[4] 小林 他, "自律的情報収集による超分散 Web 検索システム PIRCS の設計と試作", 信学技報 DE/情処研報 DBS, pp.95-102, 2001
[5] 藤本 他, "ウェブ検索 API とトピック主導型クローリングに基づくロボット型住所関連情報検索システム", IEICE WI2-2006-66, pp.125-130, 2006
[6] 石谷 他, "連鎖検索と近傍検索に基づく Web コンテンツへの効率的なアクセス方法", IEICE WI2-2006-43, pp.31-36, 2006
[7] A.Kobayashi, et al., "Cooperative Web Architecture for Search and Navigation Assistance", Proc. of ICWI2002, CD-ROM, 2002
[8] K.Randall, et al., "The LINK database: Fast access to graphs of the Web", Research Report 175, Compaq Systems R.C., 2001
[9] Y.Takasago, et al., "Synchronization Method of Web-Subgraph of Contents", Proc. of ICWI2005, pp.65-69, 2005
[10] 高砂 他, "Web サーバ間での部分 Web グラフ同期方式の提案", DEWS2006, 7C-o2, 2006