L-12　プロキシキャッシュに基づく共有 Visitedlink 機構

A Shared VisitedLink Function Based on Proxy Cache

張　剣鋒†　　成　凱†　　上林　弥彦†

Jianfeng Zhang　Kai Cheng　Yahiko Kambayashi

1．はじめに

現在、プロキシサーバを経由してインターネットにつながる利用方式はますます重要となっている。複数の利用者が共有するプロキシサーバにキャッシュを配置し大量のウェブデータを一時格納して、再利用できるようにしている。プロキシサーバに記録している利用履歴とキャッシュされたデータには貴重な情報が含まれているが、これまでキャッシュコンテンツは利用者に透明的にしか利用できないし、ウェブ利用履歴も個人レベルのパーソナライゼーションにしか用いられない状況であった。ウェブデータがキャッシュに入っているうちに、興味のありそうな利用者に知らせることによって、キャッシュデータの利用率を高めることができる。さらに、同じサーバを経由してネット上で検索するユーザ同士は、共通の特徴を多く持っていると考える。また、個々の利用者の知識、検索方式、能力などに格差があると仮定できる。ユーザの訪問したページを共有すると、ユーザ同士の融合も促進できる。本研究は、多数の利用者が共にウェブを利用する時得られた知識、経験ないしは結果を共有するため、ウェブブラウザにある VisitedLink 機能を拡張し、自分だけでなくほかの利用者の訪問したページも VisitedLink 機能を通じて興味のある利用者に知らせることができるような「共有 VisitedLink」機構を提案し、その実装を行った。

2．共有 VisitedLink 機構

これまでのウェブクライアントソフトウェアにはほとんど Visitedlink という機能を実装している[1]。利用者が最近アクセスしたページはクライアントのキャッシュに一時格納し、後でこのページへリンクしているドキュメントを閲覧する時、Visitedlink として特別な色で表示される。Visitedlink 機能によって利用者が既に辿ったリンクであることと、リンク先のページは自分のコンピュータに入っていることが分かる。しかしこのような Visitedlink 機能はあくまでも個人レベルであり、ほかの利用者が同じページをアクセスしても、他人がこのページに興味があることを示す Vistedlink は表示されず、共有できなくなってしまう。この問題を解決するため、VisitedLink を共有することが重要である。

ユーザの要求するページにあるリンクがよくアクセスされた場合、そのリンクをハイライトする。ユーザの普段のナビゲーション活動に影響を及ばさず、ユーザ同士のアクセス履歴情報を加えることができる。ユーザがナビゲーションをする際に、ハイライトされたリンクをユーザに提示することにより、そのリンクが興味のあるリンクである可能性を示す。

2.1．ページの利用状況管理

プロキシのアクセスログデータに、ユーザのアクセスした URL を記録する、これらの URL を抽出して、各 URL のアクセス頻度を集計する。URL を抽出する時に、いくつかの前処理が必要である、

1. URL の接頭辞（ftp://,http://,mailto,etc）があれば、削除する。
2. ウェブストップワード(next,back,home,etc)があれば、削除する。

前処理以後、プロキシサーバにあるアクセスログデータに含んでいる urlL[I]　(I:＜＝N；N：url 数）を全部抽出して、ユーザが各ページ UnqL[I]（I:＜＝M；M＜＝N：url 数）をアクセスした回数 FreqUnqL[I]を統計して、記憶しておく。日・週・月の時間単位で統計して、そのアクセス頻度を統計して、アクセス頻度表を作る、以下の形式で動的に管理維持している。

ページ ID	本日 利用回数	今週平均 利用回数	今月平均 利用回数
ページ 1	7	10	15
ページ 2	0	5	7
ページ 3	12	5	4
…	…	…	…

表 1：アクセス履歴から抽出したページ利用状況

2.2．ページの書き換えの仕組み

ユーザのアクセス履歴から抽出したページの利用状況にもとづいて、ページの重要度を判断する[2]。重要度の判断はその当日、今週と今月のアクセス頻度（Vd, Vw, Vm）と重み（Wd,Ww,Wm）により構成された関数 f（Vd, Vw, Vm）で計算する。重みはウェブ内容によって決める。たとえば、音楽ではクラシック、ポピュラーなど分類がある。クラシックは、長期重視型であり、ポピュラーは、短期重視型であると分類する。

型/重み	Wd	Ww	Wm
短期重視型	0.8	0.1	0.1
長期重視型	0.1	0.1	0.8
どちらか	0.33	0.33	0.33

表 2：ウェブ内容種類により重みを決める

f（Vd, Vw, Vm）＝Vd*Wd＋Vw*Ww＋Vm*Wm

計算した値により、ページの重要度を判断する、その重要度に基づく、アクセス頻度表と色の対応関係を設定して、ページの書き換えルールを考えている。重要度によって、リンクの色を書き換える。もしもともとの色と同じであれば、その属性を点滅するなどの表示を加えても良い。

型/重要度 f	f＞10	f＞1	f＝0
短期重視型	red	purple	fuchsia
長期重視型	green	lime	olive
どちらか	navy	blue	aqua

†京都大学大学院情報学研究科

表3：アクセス頻度と書き換え色との対応関係

2.3. 共有 VisitedLink 生成のアルゴリズム

具体的なアルゴリズムは以下に説明する。

1. ユーザ usr から第 j 回目にリクェストするページ Req[usr][j]を受ける。ユーザにデータを送るために、ユーザの IP アドレスを記録する IP[usr].
2. 要求されたページのリンクを抽出して、アクセス頻度表を調べる。

   もし、抽出したリンクに対応する URL がアクセス頻度表になければ、アクセス頻度表に追加して、リクェストされたページを、何も処理しないで、そのまま送り返す。

   もし、抽出したリンクが頻度表に存在すれば、以下手順に従って処理する。ページ Req[usr][j]にあるリンク ReqL[i]をすべて抽出して、そのリンクに対する、保存しているリンクとそのアクセス頻度表を調べて、以下の操作を行う。

```
for(I=0;I<=M;I++){
if   ReqL[i]= UnqL[I]      then
   if  FreqL[I]>=10   then   linkflg=1;
   else if FreqL[I]>=1   then  linkflg=0;
else
   linkflg=-1;}
FreqL[I]=FreqL[I]+1;
```

   要求されたページのソースファイルにある、リンク ReqL[i]が現れたところ、HTML ファイルを書き換え方法でリンク文字の色 linkcolor を変え、NewReq[usr][j]を生成する。ここでも、日・週・月の時間単位で統計して、色などの属性はアクセス頻度との対応表を作る。検査後と、属性タグをつけて、ページを書き換える。（ここで、CSS も使ってみる）
3. 生成した新しいページ NewReq[usr][j]をユーザ IP[usr]宛に送り返す。

以上の説明のような URL のアクセス頻度表と URL アクセス頻度と色の対応表がある、その二つの表を動的に作成、維持、後者は前者の変化に従って、書き換える。それら表の内容調べに基づいて、常に、ユーザ同士の最新のアクセス履歴情報を調べて、ユーザに提示する。

3. 共有 VisitedLink 機構のトップページ

ユーザの検索活動は、個人の特徴を持っているが、もしあるページの存在を意識すれば、また、ユーザ同士はそのページにあるリンクを良くアクセスする場合に、そのページに載せるリンクの重要性を認識できる。しかし、そのページの存在を知らない場合は、当然、そのページにあるリンクの訪れることが不可能である。そのために、共有 VisitedLink 機構のトップページが必要となる。当日、今週と今月の三つの各期間内に、ランキングされた人気度の高いリンクをまとめて、トップページに載せて、ユーザはそのトップページから、情報を得て、人気ページを訪れて、検索していく。

トップページはユーザの興味に基づいて生成されるが、ユーザの興味は各ユーザ個人の基本情報を記録するプロファイルに管理されている。ユーザの基本情報はユーザの興味があるカテゴリを含んでいる。ユーザ入力、あるいはユーザナビゲーション活動を監視して生成している。ユーザのプロファイルによりグループに分かれ、従って、ユーザグループを動的に生成し、ユーザプロファイルも動的に変化する。各ユーザグループに基づいて、人気度ページを分類して、ユーザに提示して、検索知識を共有する VisitedLink 機構を実行している[3]。

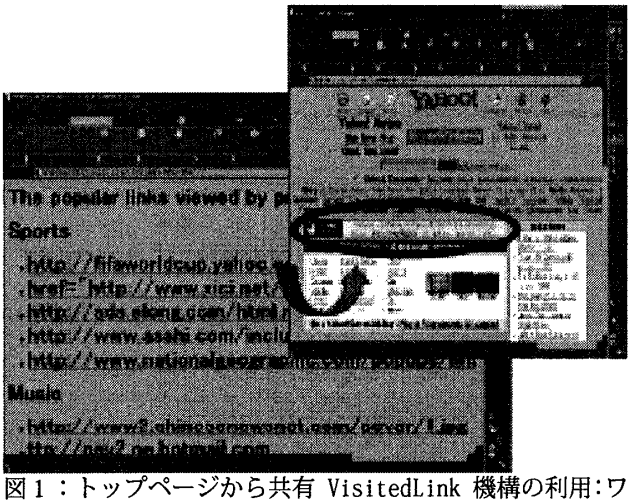

図1：トップページから共有 VisitedLink 機構の利用:ワールドカップのリンクをハイライトする

共有 VisitedLink 機構のトップページは、ユーザ同士の検索知識をさらに共有するため共有スペースを作る。ユーザ同士は最近よくアクセスされたページを、その共有スペースに載せれば、ユーザ自身の検索に役立つと考えられる。

4. 今後の課題

ユーザ同士の検索能力、知識などは差があるので、ここで導入した方式は、ユーザ同士間の検索能力差に基づいて設定した。利用者によって優先順位を付けることを考えると、かなり知識を持っている、あるいは、よくインターネットを利用しているユーザに、より良いアクセス環境を与えて、より高い優先順位を与える、知識が少ない、あるいはあまりインターネットを利用しないユーザに、より低い優先順位を与えることが考えられる。共有 VisitedLink 機構を導入すれば、経験者が見つけたページを初心者に提示することができ、プロキシサーバからデータを取得することが多くなるので、ローカルネットワーのク出口のボトルネック問題にも対処できる。

参考文献

[1]K. Cheng, Y. Kambayashi, A Semantic Model for Hypertext Data Caching, 21st International Conference on Conceptual Modeling (ER2002), October 7-11, 2002, Tampere Finland. (to appear)

[2]K. Cheng and Y. Kambayashi, Enhanced Proxy Caching with Content Management, Knowledge and Information Systems (KAIS), An International Journal. Springer-Verlag London Ltd, UK. ISSN: 0219-1377. April 2002, 4(2): 202-218

[3] Y. Hara and K. Hirata, "Augmented Hypermedia: System Integration and Usability". In K. Tanaka, S. Ghandeharizadeh and Y. Kambayashi ed. Information organization databases Foundations of Data Organization. Kluwer Academic Publishers, pp330-343 (2000).