# 位置情報を用いたウェブの事前キャッシュシステムと情報収集手法

佐々木威 †　　河口信夫 †

近年,”いつでも,どこでも”ネットワークに接続し,オンライン上の情報を検索・利用したいという要望が高まってきている.しかし,究極的に”いつでも,どこでも”ネットワークに接続できる環境を実現する事は難しい.よって,本稿では,ネットワークに接続していなくとも,位置情報を用いることによって，事前にユーザが使いそうな Web サイトやアプリケーションから情報を集め保存しておき,オフライン時に保存された情報を引き出すことによって,ネットワークへ接続できなくとも Web 上の情報･アプリケーションを利用できる仕組みと情報収集手法を提案する.

# Web PreCaching and Information Crawling Method based on Location Information

TAKESHI SASAKI†　NOBUO KAWAGUCHI†

Recently, to search information “anytime, anywhere”, demands for connecting to the Internet are increasing. However, it is difficult to realize “anytime, anywhere” Internet connection. In this paper, we propose a pre-cache system which enables us to extract information without the Internet connection by using pre-performed location and content based search results.

## 1. はじめに

近年，日本ではインターネット利用者数が増加の一途をたどっている．また，総務省の総務省通信白書[1]によると，携帯移動端末において，インターネットの利用者数が増加していることも分かる．このことから，ユーザが“いつでも，どこでも”インターネットに接続し利用するという傾向が高まってきていると言える.しかしながら，本当の意味で“いつでも，どこでも”インターネットに接続できる環境が整っているわけではない．これは，地下鉄やトンネル，時には航空機内など，インターネットへの接続が難しい様々な場面が存在するからである.

インターネットへの接続性が無い場合，様々な情報が容易に得られないため，有益な機会を失う事がある．例えば，出張に出かけた時に突然仕事の合間に時間が空いたとしても，行動できる範囲内にどんな店舗や施設があるかを調べることはできない．このため，近くに有名な店舗があっても，そこを訪れる機会を失ってしまう．用意周到であれば，あらかじめ出張先周辺にある店舗や施設の情報を調べることもあり得るが，その作業は煩雑であり，また網羅的に行えるとは限らない．煩雑さに耐え網羅的に出張先周辺の店舗や施設をあらかじめ調べたとしても，大量の情報は手軽に扱えない.

本稿では，移動先で携帯端末を使うことを前提として，インターネット接続が限られた状況でも快適に必要な情報をブラウズできるシステムの構築を目指す．インターネット接続が限られた状況で，快適なブラウズを実現するためには，端末内に周辺の店舗や電車の時刻表などの適切な情報が保持されていることが望ましい．任意の情報を保持することは困難であるため，本稿では位置情報を制約としてインターネット上の情報を収集する手法を採用する．情報収集は高速なインターネット接続を持つサーバで実行し，収集結果をコンパクトな形で携帯端末に送ることにより，短い接続時間で適切な情報の入手を実現する.

情報収集を行うサーバと情報ブラウズ用の端末をあわせて本稿では PreCache システムと呼ぶ．また，PreCache システムによってあらかじめ収集，保存される情報のことを PreCache と呼ぶ．PreCache システムを用いることにより，煩雑な事前調査を行うこと無く，位置情報に関連した情報を必要な時に手元で利用できる.

本稿では第 2 節において PreCache システムの概要について述べ，第 3 節で PreCache システムの情報集手法について示し，第 4 節では位置情報を用いた PreCache を作成するシステムの設計について記す．更に，第 5 節で提案した非定型データの収集手法について検討し，第 6 節で関連研究，関連技術について触れ，第 7 節において本稿をまとめる.

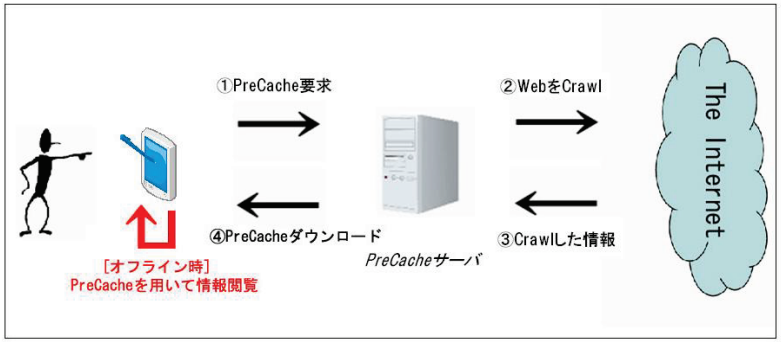
## 2. PreCache システムの概要

インターネット接続が限られた状況で PreCache システムは以下の課題に対応する必要がある．また，概要図を図 1 に示す.

1．　必要となりそうな情報の収集
2．　限られたインターネット接続への対応
3．　不足情報への対応
4．　携帯端末への適応

これらの課題に対応するため，PreCache システムは，ブラウズ用の端末と，インタ

†名古屋大学大学院，工学研究科
Graduate School of Engineering, Nagoya University..




*［PreCacheシステム概要図］*

図 1：PreCache システムの概要図

ーネット検索用の PreCache サーバから構成される．概要図を図 1 に示す．

インターネット接続が限られた状況で快適なブラウジングを実現するためには，端末内に適切な情報を保持していなければいけない．インターネット接続が限られていないブラウジングでは，なんらかの情報を欲した際に，検索ワードを打ち込み，検索を行う．

この過程を機械的に行う方法を検討し，検索ワードを事前情報として取得もしくは推測する手法を採用し，適切な情報を収集し PreCache を作成することとした．位置情報は，現在地と行動予定の目的地を入力として用い，移動先にある情報，例えば，バス時刻表やコンビニの位置，御土産屋，レストランなどの情報を収集する．また，目的地までの途中経路も利用する．これには，移動経路に含まれる公共交通機関において経路探索を行い，経路上の駅の情報も取得する．これにより，途中下車する際の情報収集にも利用できる．

位置情報を制約として Web から情報を収集するとしても，単純に駅名，街名などを使い検索を行って得られた情報を取得すればいいと言うわけではない．位置に関連した情報としてはバスの時刻表やレストラン情報などが想定できるが，Web の検索エンジンに位置情報を入力するだけで，これらの情報を効率的に収集できるとは限らない．我々は PreCache を作る為に，レストラン情報やバス時刻表を収集する為に各情報を収集するのに特化した機構を用意することが望ましいと考えた．

事前に Web から必要となりそうな情報を取得してあったとしても，必要な情報がすべて PreCache に保存されているとは限らない．不足した情報については，再度取得する為にサーバへ問合せをしなくてはならない．しかし，PreCache を使用する環境はインターネットへの接続性が限られていて，すぐにサーバへ問合せはできない可能性がある．これを解決するために，不足した情報を取得するためのクエリを端末に保持しておき，ホットスポットなどでインターネット接続可能になった際に改めて不足した PreCache を作成する手法をとることによって，不足した情報も取得可能にする．

インターネットの接続性が限られている中で，PreCache を作成するのは困難である．そこで，PreCache を作る役目を情報収集サーバに代行してもらい，高速で PreCache を作り，作り終えた PreCache をパッケージ化して一度に端末に送り返すことで，たとえ接続性が限られていても，PreCache の作成と受け渡しを可能とする．

PreCache は主に移動中もしくは移動先などインターネットに接続不可能な状況で利用する状況を想定している．移動中に快適に使える端末は PDA，携帯電話等の携帯端末に限られる． PreCache はこれらの携帯端末で利用可能にする必要がある．また，携帯端末で PreCache の快適なブラウズを可能にするためには，PreCache 専用ブラウザも必要となる．

## 3. PreCache システムの情報収集手法

本節では，PreCache システムの中核である情報収集手法について述べる．Web 上の情報は 2 種類に分類できる．この 2 種類の情報を定型データと非定型データと呼ぶ．我々が定義する定型データとは，WebAPI などを用いてデータを取得できるもの，特定のフォーマットで統一されたデータが取得できる情報を指す．位置情報を用いた Web の事前キャッシュ手法[2]では主に定型データの収集手法を述べた．また，非定型データは WebAPI を用いて情報を取得できないもの，つまり特定のフォーマットが定まっておらず，Web 上に散乱している情報の事を指す．また，定型データ，非定型データ収集にあたり，出発地と目的地情報をユーザに入力してもらい，経路探索を行う．さらに，経路上にある駅の位置や駅名などの情報を取得し，利用することで情報収集を行う．

定型データを収集するには，経路探索を行って得られた駅の位置情報や駅名などを使い，WebAPI を通じて各駅周辺の情報を収集する．定型データの収集は以上の方法で比較的容易に収集することが可能である．

次に，非定型データの収集手法について述べる．非定型データを収集している研究として，静的な検索ワード用いて検索を行い，検索結果の Web ページを収集する手法を取っている教官公募情報のダイジェスト自動作成[3]や，JPNIC が定めるドメイン名から地方公共団体の Web ページを特定し収集するワールドワイドウェブを知識源とした地域情報の自動編集[4]がある．また，オントロジーを用いて World Wide Web から情報を収集・分類する研究[5][6]もある．これらの方法は，検索ワードが適切でない可能性や，JPNIC が定めるドメインルールに従っていない地域情報ページの収集困難性があり，オントロジーを用いた情報収集手法は，ユーザが明示的に単語を与えるこ

とがシステム動作のトリガーとなる仕様である．何のルールもなく，インターネット上に散乱しているウェブページの中から，データのフォーマットが定まっていない情報を機械的に探し出す事は，既存の検索エンジンを用いても容易ではない．例えば，バス時刻表のような非定型データを収集するシステムを既存の検索エンジンを用いて実装するとしても，検索エンジンに「時刻表」という単語で検索を行うだけではユーザが欲する情報が得られないと容易に想像できる．一般に，バスの時刻表を検索するには，乗車駅や降車駅，路線名などの情報が必要である．また，既存の検索エンジンを使い，検索結果を出したとしても，その結果がバス時刻表ではなく不要なページであることもある．更に，検索結果のページからリンクを辿ることによって時刻表のページに到達することもある．以上の事を踏まえて非定型データの収集を行う手法を提案する．本稿では，人による検索キーワードや絞り込みの支援を考え，サンプルとしてバス時刻表を取り上げ，検索エンジンを用いて収集する．以下に，本手法で非定型データを収集する為の課題を以下に列挙する．

I.　最適検索キーワードの絞り込み
II.　検索結果表示された Web ページが収集対象か判定
III.　検索結果 Web ページ内のリンクを数ホップ先まで辿り，収集対象か判定

まず，I の課題を解決するために，バス時刻表検索エンジンを実装する．この検索エンジンは外部検索エンジンを用いて実装し，バス検索における人の検索エンジンの使い方を収集するために，検索が行われた際には検索ワードを保存する．この検索ワードを収集し統計を取ることで，よく使用される検索ワードを絞り込む．この結果を PreCache システムに反映し，バス時刻表を検索可能にする．II の課題は Web ページが収集対象か判定する方法である．判定方法は，ページ内に I の検索で使われた単語があるか，時刻表のページが持つ Web ページ中の数字含有率が高いなどの特徴を持つかで判定を行い，情報を収集する．III の課題はまず，検索結果の Web ページ内のリンクを辿り，リンク先の Web ページが時刻表ページかを判定する．同様に，その Web ページが含むリンクを辿ることを繰り返す．バスの時刻表かどうかの判定については II と同じ手法を取り，情報を収集する．以上を踏まえた上で PreCache システムを設計する．

## 4.　PreCache システムの設計

本節では，PreCache システムの設計について説明する．構成図を図 2 に示す．以下では，システムの各構成要素について述べる．

PreCache Request Generator

PreCache Request Generator はユーザから受け取った入力の組（現在地，目的地）を元に，経路探索を行い現在地から目的地までの駅名や街名をリストアップする．リストアップされた駅名･街名などの位置情報は Web に検索をかけるキーとして，PreCache Crawler へ送られる．

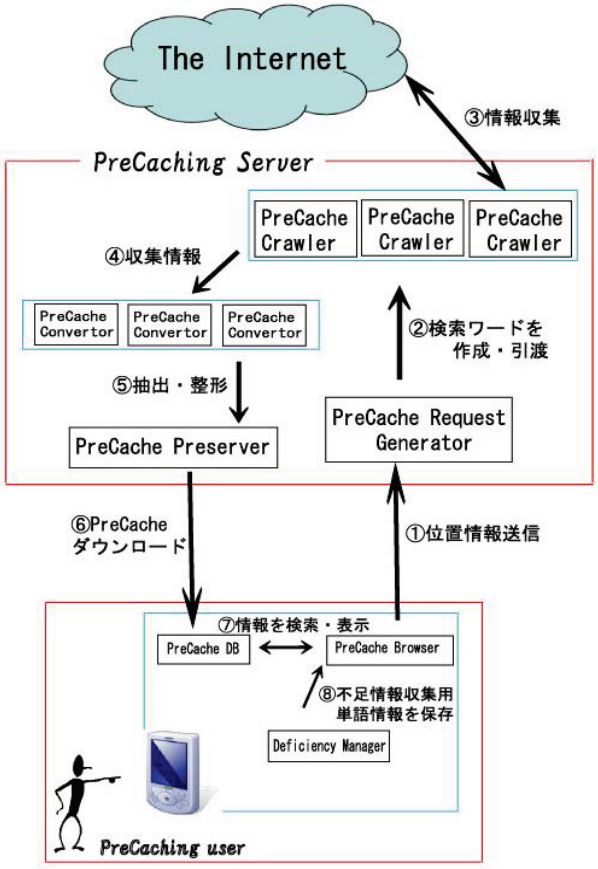



図 2：PreCache システム設計図

PreCache Crawler

PreCache Crawler は PreCache Request Generator から受け取った駅名・街名のリストを使い，Web から情報を収集する． PreCache Crawler は，収集する情報の種別毎に作成する．例として，Web 上のバス時刻表ドキュメントを集めるために，Google 検索で検索をし，条件に当てはまった Web ページの情報を取得してくるなどの機能が挙げら

れる．また，PreCache Crawler は収集した情報を，各タイプの情報に解析するのに最適化された PreCache Convertor へ送る．

PreCache Convertor

PreCache Convertor は PreCache Crawler が集めてきた情報を解析し，PreCache として適切に保存・検索・利用できる形に整形する．例えば，PreCache Crawler がぐるなび[7]から検索条件に沿ったレストラン情報を xml 形式で取得し PreCache Convertor へ送るといった動作を行う．ぐるなびは株式会社ぐるなびが運営する飲食店の情報を集めたウェブサイトである．飲食店の情報を飲食店事業主から広告として募り，飲食店が各店舗の情報を登録・情報発信し利用者は無料で飲食店の情報を検索・閲覧できるサイトである．そして，PreCache Convertor は受け取った xml を解析し整形した上で，PreCache Preserver へ送る．

PreCache Preserver

PreCache Preserver は PreCache Convertor から受け取った整形された情報をリレーショナルデータベース（以降，情報を格納したデータベースを PreCacheDB と呼ぶ）へ保存し，PreCacheDB をユーザへ送る．PreCache Preserver は更に，一度格納された PreCacheDB を一定期間保持し，PreCache Server へ過去にあった要求があった場合に再び PreCache を作ることなく，ユーザに PreCacheDB を渡す．

PreCache Browser

PreCache Browser は携帯端末上でユーザの要求を受けて，要求に沿った情報を PreCacheDB から検索し，ユーザに提示するブラウザである．また，ユーザからの要求に沿った情報を PreCacheDB が含んでいない場合，不足した情報を検索するためのキーをユーザから入力してもらい，そのキーを Deficiency Manager へ送り一時的に保持させる．

Deficiency Manager

Deficiency Manager は PreCache Browser から不足した情報を検索するためのキーワードを取得・保持し，インターネット接続可能時に PreCache サーバへアクセス．不足した情報に対して再度 PreCache を作る要求を PreCache サーバへ送る．

## 5. 非定型データ収集手法の有効性

携帯移動端末の限られたメモリに，収集した情報を格納可能であるかを検討する為に，非定型データを収集する手法を用いて，バス時刻表を検索し，検索結果の Web ページから１リンク，2 リンク，３リンクで辿れる総数と，キーワードや時刻表ページの特徴でフィルタを行った場合の数を比較する．バス時刻表を検索する際の検索ワードは「(乗車駅名)，(降車駅名)，バス，時刻表」とし，検索エンジンを用いて検索する際に最初に表示される検索結果 10 件を対象とし，リンクのトレースをする．リンクで辿れる Web ページの総数を表１に．そのうち４つの検索ワードを含む Web ページ数を表２に，４つの検索ワードを含んだ時刻表ページ数を表３に示す．

表１：検索結果から辿れる Web ページの総数

| (乗車駅，降車駅) | 1 リンク先 | 2 リンク先 | 3 リンク先 |
|---|---|---|---|
| (名古屋大学 ，名古屋駅) | 10 | 450 | 12338 |
| (中村記念病院 ，いこいの里) | 10 | 1767 | 28501 |
| (池田港 ，福田港) | 10 | 163 | 1929 |
| (山口駅 ，センタービル前) | 10 | 307 | 4726 |
| (常滑 ，セントレア) | 10 | 390 | 14784 |

表２：キーワードが全て含まれるページ

| (乗車駅，降車駅) | 1 リンク先 | 2 リンク先 | 3 リンク先 |
|---|---|---|---|
| (名古屋大学 ，名古屋駅) | 10 | 13 | 36 |
| (中村記念病院 ，いこいの里) | 10 | 22 | 30 |
| (池田港 ，福田港) | 10 | 3 | 5 |
| (山口駅 ，センタービル前) | 10 | 15 | 29 |
| (常滑 ，セントレア) | 10 | 22 | 18 |

表３：キーワードが全て含まれた時刻表ページ数

| (乗車駅，降車駅) | 1 リンク先 | 2 リンク先 | 3 リンク先 |
|---|---|---|---|
| (名古屋大学 ，名古屋駅) | 2 | 0 | 1 |
| (中村記念病院 ，いこいの里) | 0 | 1 | 0 |
| (池田港 ，福田港) | 2 | 1 | 2 |
| (山口駅 ，センタービル前) | 3 | 2 | 0 |
| (常滑 ，セントレア) | 1 | 1 | 1 |

表１の結果から，Web ページ中のリンクを辿るとその情報量は爆発的に増えることが分る．しかし，キーワード全てを含む Web ページのみを抽出すると表２のような結果となる．このことから，PreCache システムを作る際の情報量を大幅に減らすことができることがわかる．また絞り込んだ Web ページ中に十分なバス時刻表ページが含まれることが分る，以上より，この手法は有効であると考えられる．

## 6. 関連研究・関連技術

**Google Gears[8]**

Google Gears は本来，オンラインでしか動かない Web アプリケーションをオフライン状態でも使えるようにするという発想で開発された．Google Gears の基本構成はローカル･サーバ，データベース，ワーカープールの３つである．まず，Web アプリケーションをローカルサーバに保存しローカルホスト内でも動くようにする．次に，アプリケーションがデータの入出力を必要とする場合は，データをデータベースに格納する．そして，最後のワーカープールは，オンライン上のコンテンツをキャッシュし，コントロールする．これらの働きによって，オフラインでもユーザは Web アプリケーションを使えるようになる．ただし，Google Gears は Google Gears に対応した Web アプリケーションしかローカルホストでしか動かないという仕様である．

**Supporting Cooperative Caching in Ad Hoc Networks[9]**

この研究はアドホックインターネットを用いて，ユーザ達が持っているキャッシュを共有するというものである．アドホックインターネットを用いてキャッシュを共有できれば，ユーザが直接インターネットに接続できなくても周囲のユーザが持っているキャッシュから情報を検索できる可能性がある．また，同様の研究として，モバイルアドホックインターネットにおける効率的な情報共有[10]がある．

**Web Cache**

Web Cache は Web Proxy や Web ブラウザが，処理を高速化するために一度ユーザが閲覧した Web の情報をキャッシュとして一定期間保存しておくものである．次に同じ Web を閲覧する要求が来た場合は，キャッシュを使うことによって高速な応答が実現される．Web の情報は更新頻度が高いものもあるため，必要に応じてキャッシュも更新される．この Web Cache を使えばユーザが過去に見た Web ページの情報を利用できるが，必要な情報をキャッシュの中から適切に検索する手法は確立されていない

## 7. まとめと今後の展望

本稿では，位置情報を用いたウェブの事前キャッシュシステムを提案し，更に Web を定型データと非定型データに分け，情報を収集する手法も提案した．PreCache システム全体の設計も行った．また，非定型データの収集手法について，バス時刻表をサンプルとして本手法が有効であるかを確かめた．今後は本稿で提案した手法を実装し，Web アプリケーションとして実運用を行い，評価を行う．

## 8. 参考文献


[1]総務省通信白書平成 20 年版， 総務省情報通信統計データベース

[2]佐々木 威，河口 信夫，”位置情報を用いた Web の事前キャッシュ手法”， 情報学ワークショップ WiNF,pp117-122,2008.

[3]見舘 潔，佐藤 理史，“教官公募情報のダイジェスト自動生成”，第 58 回全国大会講演論文集，pp.”3-87”-“3-88”,1999.

[4]大槻 洋輔，佐藤 理史，“ワールドワイドウェブを知識源とした地域情報の自動編集”，第 119 回情報処理学会「知能と複雑系」研究会(ICS-119)，pp165-172，2000.

[5]野田 武史，大島 裕明，小山 聡，田島 敬史，田中 克己，“主題語からの話題語自動抽出とこれに基づく Web 情報検索”，信学技報，vol.106，no.149，DE2006-90，pp. 239-244, 2006.

[6]松平 正樹，上田 俊夫，大沼 宏行，森田 幸伯，“Web コンテンツの分析に基づくオントロジー構築および情報整理の試み”，人工知能学会 第 4 回セマンティックウェブとオントロジー研究会，pp0801-0808, 2003.

[7] ぐるなび， http://www．gnavi．co．jp/

[8]Google Gears, Google, http://gears.google.com/

[9] L. Yin and G. Cao, "Supporting Cooperative Caching in Ad Hoc Networks", Proceedings of IEEE INFOCOM, pp.2537-2547, March 2004.

[10]榎本 真, “モバイルアドホックインターネットにおける効率的な情報共有 “, ユニシス技報, Vol.27, No.4, 通巻 95 号,pp.37-54, 2008.