データベースシステム 128－68
（２００２．７．１９）

地域的スコープと詳細度による WEB ページ分類とモバイルキャッシュへの応用

山田 直治† 李 龍† 高倉 弘喜‡ 上林 弥彦†

†京都大学情報学研究科社会情報学専攻 ‡京都大学学術情報メディアセンター
〒606-8501 京都市左京区吉田本町

E-mail: †{naoharu, ryong, yahiko}@db.soc.i.kyoto-u.ac.jp, ‡takakura@media.kyoto-u.ac.jp

あらまし WEB 上には多種多様な情報が大量に存在するため、利用者の要求に適した情報を収集することが困難になっている。ここでは携帯端末から WEB 上に存在する地域情報を収集するために、WEB 情報の地域性と記述形式に着目する。既存のキーワード検索では地名の位置情報を考慮していないため、任意の地域的範囲に関する情報を収集することが困難である。また要約情報なのか詳細情報なのかといった WEB ページの記述形式を考慮していないため、利用者の要求する情報の詳しさに対応することができない。記憶領域が限られ通信が不安定な携帯端末では利用者の要求する情報のみを提供する必要があるため、これらは大きな問題である。ここでは地名の位置情報を利用して WEB ページが着目する地域を特定する。また HTML タグや品詞の出現度数の特徴から WEB ページを目次型、要約型、詳述型の３つのタイプに分類する。最後に利用者の特定の地域に対する興味の深さに基づき、２つの尺度を用いた携帯端末へのキャッシュアルゴリズムについて述べる。

Classification of Web Pages with Geographic Scope and Level of Details for Mobile Computing

Naoharu YAMADA† Ryong LEE† Hiroki TAKAKURA‡ and Yahiko KAMBAYASHI†

† Department of Social Informatics, Kyoto University
‡ Academic Center for Computing and Media Studies, Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto-shi, 606-8501 Japan

E-mail: †{naoharu, ryong, yahiko}@db.soc.i.kyoto-u.ac.jp, ‡takakura@media.kyoto-u.ac.jp

Abstract Due to the rapid increase of the amount of web pages on the Internet, it is difficult to collect information that satisfies users' queries. This paper focuses on the geographic characteristics and description types of web resources. Keyword based search does not take account of the positional information of geographic names so that it cannot collect web resources on specific region. Furthermore, because a web page is treated without considering whether it contains detailed information or summarized one, the page may not satisfy users' requirements. In this paper, a method to determine the geographic scope and level of details of web pages is developed. Geographic scope is identified with the positional information of geographic names. Level of Details classified web pages into three types, *"table-of-contents type"*, *"summary type"*, and *"detailed description type"*, with HTML tags and frequency of parts of speech. The cache algorithm with these two measures for mobile computing based on users' interests is also described.

1. はじめに

携帯電話をはじめとする携帯端末の高機能化とその普及により、日本におけるモバイル通信利用者は２００５年までに全人口の７割に達するという調査結果が出ている[12]。それと同時に利用者の位置情報を利用したさまざまなサービスも登場し、地域情報に関する利用者のニーズは高いと言える。しかし携帯端末には以下の２つの大きな欠点がある。

格納できるデータ量の制限
最近ではフラッシュメモリの大容量化が進んでいるが、特定の地域に関するすべての情報を格納することは困難であり、利用者の要求を満たす地域情報のみを収集する必要がある。
低速な通信速度と通信の切断
最近では第３世代携帯電話で最大2Mbpsの伝送速度が実現しているが依然通信が不安定であり、それを補うために必要な地域情報を予め携帯端末上に格納しておく必要がある。

これらの背景からモバイルキャッシュに関する研究が行われている。従来のキャッシュ管理では、単体では意味を持たないデータに対してアクセス回数などを利用したキャッシュの置き換えアルゴリズムを提案してきた。一方モバイルキャッシュでは、位置情報を保有した意味をもつデータに対してその意味を考慮したキャッシュ手法が提案されている[1, 8]。しかしこれらの既存の研究では地域情報としてガソリンスタンドやコンビニエンスストアなどといった情報を扱い、WEB 上に存在する地域情報を扱っていない。WEB 上には地域ポータルサイト、特定の知識に関する解説説明ページ、旅行記など利用者にとって有益な地域情報が多く存在する。そこで我々は WEB 上の地域情報を旅行者に提供する手法について研究してきたが[6]、利用者の要求に応じた任意の地域に関する WEB 情報を提供することは困難であり、それは以下の二つに起因する。

WEB 情報の地域性
従来のキーワード検索では位置情報を持つデータとして地名を扱っていないため、特定の地域に関する情報を収集することが困難である。例えば既存のキーワード検索で「左京区」に関する WEB 情報を検索した場合、「左京区」というキーワードを含む WEB ページのみ収集し、京都大学や銀閣寺など「左京区」という地域に含まれる地域に関する WEB ページを収集することができない。
WEB 情報の記述形式の多様性
WEB ページは作者の意図によってさまざまな記述形式、例えばリストで情報を羅列したページ、情報を要約したページ、詳細に記述したページ、などが存在するが、キーワード検索ではそれらの相違を考慮していない。しかし特定の地域に対する利用者の興味の強さに応じて利用者の要求する記述形式は異なる。例えば銀閣寺に強い興味を持っている利用者は銀閣寺に関するより詳細な情報を要求するのに対し、銀閣寺に興味を持っていない利用者は銀閣寺に関するおおまかな理解ができる要約した情報を要求する。

本稿では旅行者の携帯端末に地域情報を効率的に提供するために、旅行者が興味を持つ地域に対して WEB の地域性や記述形式を考慮したキャッシュの置き換え手法を提案する。本研究では WEB ページが着目する地域を特定する尺度として地域的スコープを導入する。地域的スコープは MBR（Minimum Bounding Rectangle）によって測定される。ただしここでは WEB ページに出現するすべての地名を測定対象として用いるのではなく、不要な地名を除去することで地域的スコープの精度を高める。また記述形式を特定するための尺度として詳細度を導入する。トピックに基づく WEB ページの分類は既にいくつか行われているが[13]、記述形式による WEB ページの分類はまだ行われていない。ここで提案する詳細度によって WEB ページは目次型、要約型、詳述型の３つのタイプに分類される。これらは HTML のタグ、品詞情報、出現する地名数によって特徴づけられ、決定木によって３つのタイプに分類する。

以上２つの尺度を踏まえた上で、旅行者の興味の強さに基づく効率的なモバイルキャッシュ手法を提案する。そのために旅行者が興味をもつ地域と興味の強さを特定する。[8. 9]では携帯端末利用者は現在位置周辺に最も興味を持つと仮定している。しかし旅行者にとっては訪問予定の地域に対する興味が現在位置よりも強い場合が多く、訪問予定の地域から離れていてもその地域に関する情報を要求する可能性は高いと考えられる。ここでは旅行者が事前に訪問する地域を決定することを利用して興味のある地域と興味の深さを特定し、それに応じて二つの尺度を用いたキャッシュ手法を提案する。

以下２章では地域的スコープについて、３章では詳細度について詳しく説明する。４章では利用者が興味をもつ地域とその興味の強さについて分析し、２つの尺度を利用した WEB ページのキャッシュアルゴリズムについて述べる。

2. 地域的スコープ

地域的スコープとは WEB 情報が着目している地域を特定するための尺度である。この尺度を用いることで、WEB 上から特定の地域に関するページ集合のみを収集することが可能となる。一般に地域情報の地域的スコープを測定する手法として MBR（Minimum Bounding Rectangle）が提案されている。MBR とはひとつ以上の地域を空間的に包含する最小の矩形でかつその辺が緯度経度に平行な領域であり、Guttman によって提案された[2]。この手法により各 WEB ページの地域的スコープはページ内に出現するポリゴンで表現された地域集合に対してＸ、Ｙ座標の最大値と最小値のみを計算しすればよいので、計算量のコストを抑えることができる。また地域的スコープが４つの数値で表現できることからポリゴンの状態に比べてデータ量のコストも低く抑えることができる。さらに R-tree としてインデックス付けすることで特定の地域に関する情報を効率的に収集することが可能となる。MBR を用いて WEB 情報の地域的スコープを特定する手法は既に提案されているが[11]、WEB ページに出現する地名をすべて用いて MBR を計算すると WEB 情報が着目している地域より空間的に広い値になってしまう。これは WEB ページに出現する地名には以下の２つの種類が存在するためであると考えられる。

- **キーワードとしての地名**　そのページの主題となる地名であり、地名に関する情報は多い。
- **説明語としての地名**　キーワードとなる地名を説明するための地名であり、地名に関する情報は少ない。

例えば「金閣寺」というタイトルをもつ WEB ページで金閣寺に関する以下の記述があったとする。「金閣寺は銀閣寺と並んで日本で最も有名な寺のひとつだ」。この場合金閣寺はキーワードにあたり、銀閣寺は金閣寺を説明する地名である。

WEB ページの地域的スコープを測定する際には、キーワードとしての地名のみを用いて MBR を測定すれば、WEB ページが着目する地域を正確に求めることが可能となる。そこで地名のキーワード性を測定し、キーワード性が低い地名を除去した上で MBR を用いて地域的スコープを計算する。キーワードとしての地名の特徴として以下が挙げられる。

- HTML タグによって着目地名が強調されている
- 単一ページに複数回出現する
- 着目地名に関する情報を記述するために、着目地名の関連地名や関連名詞が多く出現する

３つの特徴のうち、上の２つは一般的なキーワードの特徴である。なおここでは HTML によるタグの強調度を表１のように定めた。３つ目の特徴に関して我々はデータマイニング技術を用いて地名と地名の関連性や地名と地名以外の名詞の関連性を抽出する研究を行っており[5, 14]、これを利用して WEB ページ内に出現する関連名詞を特定しその数を測定する。以上３つの値を元に決定木によってキーワードとしての地名と説明語としての地名を分類する適切な閾値を決定する。

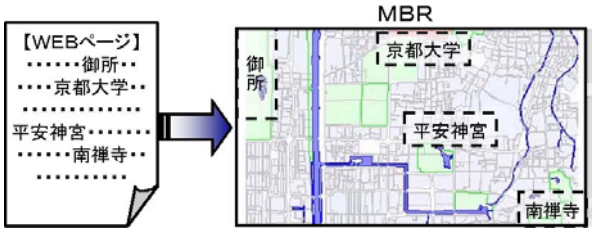

図 1　WEB ページの MBR

強調度	HTMLタグ
1	TITLE
2	SIZE=7 H1
3	SIZE=6 H2 B I EM STRONG
4	SIZE=5 H3
5	SIZE=4 H4
6	SIZE=3 H5
7	SIZE=2 H6
8	SIZE=1 H7

表 1　HTML タグによる地名の強調度

3. 詳細度

詳細度とは WEB ページに記述された情報の詳しさを測定する尺度である。WEB 上にはさまざまなトピックに関する情報が存在しているが、それと同様に情報の記述形式にもさまざまな種類が存在する。WEB ページは記述形式によって大きく以下の３つのタイプに分類される。

・ **目次型**

特定のトピックに従って地名をリストアップしたページであり、例として「京都の寺社一覧」が挙げられる。これは旅行者が計画段階で、興味のあるトピックを通して地域全体を把握するのに向いている。

・ **要約型**

特定の地域に関して数行程度で簡潔に記述されたページであり、特定の地域に関する情報サイトのトップページや簡単な施設紹介などが例として挙げられる。これは特定の地域についておおまかな情報を得るのに向いている。

・ **詳述型**

特定の地域に関して詳しく記述されたページであり、特定の地域に関するレビューや詳細な説明・解説が例として挙げられる。これは特定の地域についてより詳しい情報を得るのに向いている。

これら３つのタイプは利用者の状況によってその重要度が異なる。例えば「足利家ゆかりの地」などの特定のトピックに関連する地域を調べたい場合には目次型が適しており利用者によっての重要度は高いが、「金閣寺」など特定の地域に関する感想などを知りたい場合には詳述型が適しており利用者にとっての重要度は高い。しかし既存のキーワード検索では利用者がこれらのタイプを直接指定することはできない。そこでWEB ページをこの３つのタイプに分類するためにここでは各品詞の出現度数、ページの表記方法、出現する地名数に着目する。それぞれについて３つのタイプの特徴を示す。

・ **目次型**

特定のトピックについて地名を羅列する形で記述することが多いため、名詞の品詞比率が高く、箇条書きで記述されることが多い。地名を羅列するため、その出現数は多い。

・ **要約型**

特定の地域に関して簡単な情報のみを記述するため、箇条書きもしくは文章で記述され、記述量は少ない。そのため動詞の数が少なくなる。また特定の地域に関してのみ記述するため、地名の出現数は少ない。

・ **詳述型**

特定の地域に関して詳細な情報を記述するため、文章で記述され、記述量も多い。そのため動詞の数は多く、動詞の比率も高い。他の地名を引用しながら記述することが多く、地名の出現数は多い。

以上の特徴を表２にまとめる。これを元にページ毎に名詞の出現度数、動詞の出現度数、修飾語の出現度数、リスト表記やテーブル表記を行うためのタグである LI、TD、P、BR の出現度数、地名の出現度数を測定した。P や BR は本来リスト表記を行うためのタグではないが、実際にはこれらのタグを利用してリストタグ表記している WEB ページが多く見つかったため、リスト表記のタグとして加えた。WEB ページ内に出現する名詞や動詞の数を得るためにここでは形態素解析システムの茶筅[8]を利用する。そしてこの値を元に名詞と動詞の出現比率を計算する。以上の要素を用いて決定木によって WEB ページを３つのタイプに分類する。

	目次型	要約型	詳述型
品詞の出現度数	名詞比率 高	動詞数 少	動詞数 多 動詞比率 高
ページ表記	リスト表記	リスト・文章	文章表記
地名数	多	少	多

表 2　ページタイプの特徴

4. 利用者の興味の強さに基づくモバイルキャッシュ

旅行者が WEB 上の地域情報を効率的に参照できるように、利用者が要求する情報を予測しその地域に関する情報をプリフェッチする必要がある。またメモリの領域に制限があるため、不要な情報はメモリ上から除去する必要がある。本稿ではまず旅行者が興味を持っている地域を特定し、その地域に対する旅行者の興味の強さについて考察する。そして旅行者が興味のある地域に関する WEB 情報を適切に提供するために地域的スコープを利用する。また旅行者の興味の度合いに応じて適切な WEB 情報を提供するために詳細度を利用する。その結果これらふたつの尺度によって効率的なモバイルキャッシュを実現する。

4.1. 地域に対する旅行者の興味

特定の地域に対する旅行者の興味の強さは地域情報の要求に強く依存しており、キャッシ

ュアルゴリズムを考える上で、旅行者がどの地域に対してどの程度強い興味を持っているかを分析する必要がある。

興味の強さ

旅行者は計画時に訪問予定地域をいくつか決定するが、それらの地域は旅行者にとって興味深い地域であると考えられる。また訪問予定地域が複数存在する場合、次に訪れる訪問予定地域へ移動する必要があるが、その際通過する地域に対して興味を持つ可能性は高いと考えられる。例えば金閣寺から銀閣寺へ移動中する際に、道中に存在する御所に興味を持つ可能性は高い。しかし計画段階で訪問予定地域に含めていないことから、訪問予定地域に比べて興味度は低いと言える。一方既に訪問した地域に関する興味はほとんどなくなると考えられる。

以上をまとめると旅行者の特定の地域に対する興味度は

既に訪問した地域<<通過予想地域<訪問予定地域

である。

興味がある地域の特定

旅行者が興味をもつ地域は上述の考察より、訪問予定地域と移動の際に通過する地域である。訪問予定地域の位置情報は容易に得ることができる。移動の際に通過する地域に関して、ここでは[9]のように旅行者の移動手段（車、バス、徒歩）を限定していないため、訪問地間の移動経路を特定することができない。そこでここでは訪問予定地域集合に対する MBR を通過領域とする。これにより利用者が移動経路を自由に選択しても通過予想領域から逸脱することは少ないため、通過領域を更新する必要がない。また通過地域に関する情報は R-tree によってインデックス付けされた WEB ページ集合から効率的に収集することができる。R-tree に関してはいくつか改良手法も提案されており[3, 7]、ここでは既存の手法を利用する。

興味がある地域と興味の強さの動的な変更

旅行者が興味を持っている地域やそれらの地域に対する興味の強さは以下の２つによって動的に変更する。

- 旅行者による変更・・・旅行者は旅行中に訪問予定地域を追加・キャンセルする。
- 移動に伴う変更・・・訪問を終えた地域に対する旅行者の興味は弱くなる。

後者を例に挙げると、図２において旅行者はあらかじめ「御所」、「南禅寺」、「八坂神社」を訪問予定地域として登録しており、八坂神社を現在訪問しているとする（①）。その後、旅行者は南禅寺へ向かう（②）。このとき八坂神社は訪問予定地域ではなくなるため、新たな通過領域は南禅寺と御所の MBR となり「円山公園」は旅行者がまだ通過していないにも関わらず通過領域からはずされてしまう（図３）。そこでここでは通過地域と興味の強さの更新は利用者が訪問地の見学を終えた時点、もしくは訪問予定地を追加・キャンセルした時点で行う。そして通過予想領域は訪問地集合に旅行者の位置座標を含めた MBR で測定する。これにより旅行者は更新後には常に訪問予想地域内に存在することになり、上記の２つの変更に伴う問題を回避することができる。

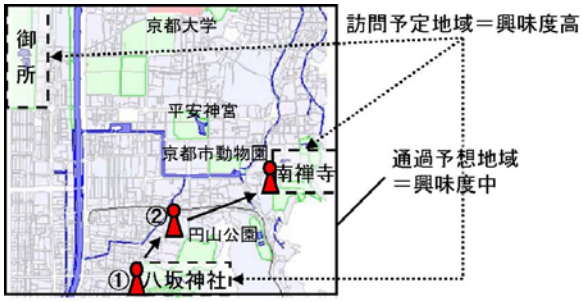

図 2　訪問地域から分かる利用者の興味

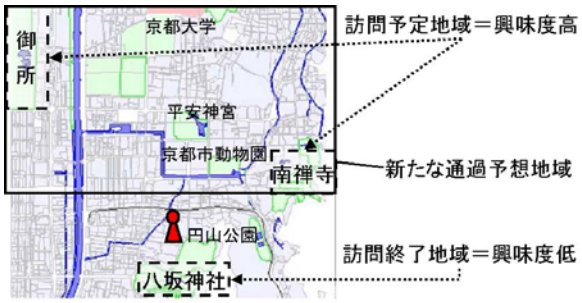

図 3　通過予想領域の問題点

4.2. 2つの尺度を利用した効率的な WEB ページのキャッシュ

前節で特定した旅行者が興味のある地域に対して。旅行者の興味の強さに応じた適切なタイプの地域情報をキャッシュする必要がある。旅行者は強い興味を持っている地域に関して詳しい情報を要求し、逆に興味のない地域に関しては情報を要求しない。このような興味と情報

要求の関連性を利用して、キャッシュのプリフェッチとインバリデート（無効化）を以下のように行う。

- 訪問予定地域・・・興味が強い地域であるため詳述型ページをプリフェッチする。
- 通過予想地域・・・訪問予定地域に比べて興味は弱いが、興味をもつ可能性は高いため、要約型ページをプリフェッチする。
- 既に訪問した地域・・・興味度は低くなっているため、詳述型ページをインバリデートする。

これらは旅行者が興味のある地域とその強さを更新した後に更新する。

5. おわりに

本稿では WEB 上に存在する地域情報から旅行者の興味のある地域に関する情報のみを収集するために地域的スコープを提案し、利用者の興味の強さに適合するタイプの WEB 情報を収集するために詳細度を提案した。さらに利用者が興味を持っている地域とその強さを予測することで、二つの尺度を用いたモバイルキャッシュアルゴリズムを提案した。今までは携帯端末へ配信するための地域情報として WEB の情報を利用されることはなかったが、WEB 情報の増大により特定の地域に関する感想など他の情報源では得ることが困難な情報も多く存在し、地域情報源としての価値は大きいと考えられる。

本稿で提案したキャッシュアルゴリズムについて、計画時に訪問地域を多く選択するとそれだけ多くの情報をキャッシュに格納しなければならないという問題があるが、WEB ページ間の類似度を測定することで、特定の地域に関する全く異なる情報のみを携帯端末にキャッシュすることができる。また地域毎、タイプ毎に人気度を定義することで WEB ページのランキングを定めることができ、これを元にランキング上位のみキャッシュに格納するといった改良も可能である。

我々は現在京都に関するＷＥＢページを収集しており、今後はここで提案した地域的スコープと詳細度を実際に測定し、手法の評価実験を行っていく予定である。

謝辞

本研究は科学技術振興事業団 (JST)・戦略的基礎研究推進事業 (CREST)における「デジタルシティのユニバーサルデザイン」プロジェクトの支援によって行われた。

文　献

[1] B Zheng, and D. Lee, Semantic “Caching in Location-Dependent Query Processing”. SSTD, pp.97-116, 2001.

[2] G. Antonin, R-TREE A Dynamic Index Structure For Spatial Searching, In proceedings of ACM SIGMOD, pages 47-57, 1984

[3] N. Beckmann, H. Kriegel, and R. Seeger, R*-tree: An Efficient and Robust Access Method for Points and Rectangles, In proceedings of ACM SIGMOD pages 322-331, 1990

[4] N. Davies, K. Cheversi, K. Michell, and A. Friday, “Caches in the Air: Dissemination Tourist Information in the Guide System”, Proc. of the 2nd IEEE Workshop on Mobile Computing Systems and Applications (WMCSA '99), pp.25-26, Feb. 1999.

[5] R. Lee, H. Takakura and Y. Kambayashi, “Visual Query Processing for GIS with WEB Contents”, Proc of the 6th IFIP Working Conference on Visual Database Systems, pp.171-186. May 2002.

[6] R. Lee, K. Goshima, Y. Kambayashi and H. Takakura, “Caching Schema for Mobile Web information Retrieval”, 2nd International workshop on Web Dynamics, 2002.

[7] Timos K. Sellis, Nick Roussopoulos, Christos Faloutsos, The R+-Tree: A Dynamic Index for Multi-Dimensional Objects, VLDB 1987, pp.507-518, 1987.

[8] 坂田一拓, 倉島顕尚, 市村重博, “通知型の位置関連情報提供サービスの提案と、その実現方式の検討”, モバイルコンピューティングとワイヤレス通信研究報告 vol.15-9, pp.73-80, Dec. 2000.

[9] 佐藤健哉, 最所圭三, 福田晃, “移動計算機における位置依存情報のキャッシュ方式に関する考察”, モバイルコンピューティング研究報告 vol.7-5, pp.33-38, Dec. 1998.

[10] 茶筅, http://chasen.aist-nara.ac.jp/index.html.ja

[11] 松本知弥子, 馬強, 田中克己, “WEB ページの地理情報と話題の日常性を考慮したローカル度検出とフィルタリング機構”, DBWEB2001, pp.193-200, Dec.2001.

[12] モバイルコンピューティング推進コンソーシアム, http://www.mcpc-jp.org/

[13] 山田洋志, 福島俊一, 松田勝志, “Web ページからのタイプ別情報抽出・分類方式”, 情報処理学会研究会報告 FI-57-19, pp.143-150, 2000.

[14] 李龍, 高倉弘喜, 上林弥彦, “地域ウェブ情報を利用した地域情報検索と地域分析”, 空間 IT ワークショップ, pp.8-16, Dec.2001