社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－DBS－131 (54)
2003／7／17

地域ウェブ情報検索のための２次元領域質問処理法

李龍† 椎名宏徳‡ 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
‡京都大学工学部
††京都大学学術情報メディアセンター
Email:{ryong, shiina, yahiko}@db.soc.i.kyoto-u.ac.jp, takakura@rd.kudpc.kyoto-u.ac.jp

あらまし 近年、新しいウェブ検索手法の一つとして地理情報検索システムが注目され、さまざまな研究がなされている。特にユーザの質問作成とそれに対する検索の手法、ウェブページから的確な位置情報を獲得する手法が研究の焦点になっている。我々の提案では、ユーザの２次元領域質問の処理法として、従来空間索引に用いられていたMBR(Minimum Bounding Rectangle)を用いる。ウェブページを地図上のMBRとして管理することで、ユーザの領域質問との間で地理演算を行うことができる。しかし、各ウェブページに対応する領域を最適にしないと、正確な検索を行うことができない。本稿では、ウェブページに現れる地名の性質（ i) 地名の重要性、ii) 地名の階層構造、iii) 地名の多義性）を利用して、ページの対応する地理領域を的確に求める方法を提案し、その実験結果について述べる。

キーワード 地域ウェブ情報システム、地域ウェブページの索引、領域質問処理

Two-Dimensional Range Query Processing for Geographic Web Search

Ryong LEE†, Hironori SHIINA‡, Hiroki TAKAKURA††, and Yahiko KAMBAYASHI†
† Social Informatics, Kyoto University
‡ Faculty of Engineering, Kyoto University
†† Academic Center for Computing and Media Studies, Kyoto University
Email:{ryong, shiina, yahiko}@db.soc.i.kyoto-u.ac.jp, takakura@rd.kudpc.kyoto-u.ac.jp

Abstract Geographic web search engine is one of specialized web search engines to retrieve geographic information on the web. We have introduced a two-dimensional range query, which is specified by a rectangle region corresponding to a region. In order to handle such a query, we use location names appearing in each web page. For accurate computation, we have to find minimum regions corresponding to web pages. Methods to handle the following problems are developed;(1)Existence of non-important location names, (2)some redundancy of geographic hierarchy, and (3) identified location names used for different objects. In the major results of the paper is how to optimize the MBR (Minimum Bounding Rectangle) for a given web page. We have shown the effectiveness of our system by experiments.

Keyword Geographic Information System, Map-based Web Search, Range Query Processing

１．はじめに

近年、ウェブページの増加に伴い、ユーザが必要な情報を的確に入手するためにさまざまなウェブ検索技術が提案されている。特に、地図インターフェースを利用した地理情報検索システムは、日常生活に関係の深い検索システムと

	KyotoSEARCH[6]	ここのサーチ[9]	Digital City[10]	ローカル度[12]	Locality Ranking[5]
ユーザの質問方法	矩形領域	現在地を中心とした円	キーワード 目的別の分類	キーワード	キーワード
ウェブページと地理領域の対応方法	MBR	ポリゴン	点	MBR	点
地域性の判断方法	内的判断	内的判断	登録者が判断	内的・外的判断	外的判断
地域性判断の指標	タイトル・アンカーに含まれる住所、地理オブジェクト名（曖昧性があっても対処する）	文書に含まれる住所（曖昧性のないものに限定）	登録者が判断	地名とその出現回数や密度、日常用語、他ページとの類似度、ユーザの地理的分布などから計算したローカル度	対象地域に関連したリンクを利用して計算した人気度と地域指向性

図 1：関連研究との比較

して注目を集めている。従来のキーワード検索では、有名な場所を検索することはできるものの、ある地理領域内に関するウェブページを探すような検索は困難であった。地理情報検索システムにおける主な課題として、1．ユーザの質問方法とその質問処理方法をいかに行うかと、2．ウェブページに記述されている位置情報をどのように抽出するかの2点が挙げられる。

ウェブページと地理領域との対応に関する研究はさまざまな形で行われている。ウェブページの地理領域を表すために、地図上で多角形に対応させたり、点に対応させたりするものが提案されている。ウェブページの地域情報を評価する手段としては、ページ内に現れる情報を利用するものだけでも、住所、地名の分布や日常用語まで利用した研究まである。その他に、リンクなどを利用してページの地域性を評価することもできる。

現在、我々が開発しているKyotoSEARCH[6]もこの地理情報検索システムの一つである。このシステムの機能のひとつとして、ユーザが質問領域を地図インターフェース上に指定することで、その質問領域に関するウェブページのリストを得ることができる。検索には、従来の空間索引に利用されていたMBR (Minimum Bounding Rectangle)を利用している。MBRはページ中の地名に対応する場所をすべて含む最小の矩形領域である。このMBRをウェブページごとに計算し、ユーザの質問領域との間で包含(Contain)などの地理演算を行うことでページを検索する。しかし、ただ単にページ中に出現する地名をすべて用いてMBRを生成すると、重要でない地名の存在などさまざまな問題からMBRは不正確なものとなり、サイズが大きくなってしまう。その結果、ユーザの質問領域に包含されるMBRが少なくなり、検索がうまくいかなくなってしまう。このような問題に対処するためには、ウェブページがどこについて述べているのかを的確に判断しなくてはならない。本稿では、ページに現れる地名の性質を利用してそのページに対応する最適な地理領域を求める手法について提案する。

以下、２章では、まず地域情報検索システムに関する研究を簡単に紹介する。３章では、これらの関連研究をふまえて、MBR を利用した領域質問処理を提案する。４章ではその処理を効率化するために、地名の性質を利用する方法を述べ、５章で、実験による評価を行う。６章では本稿のまとめと今後の計画について述べる。

２．関連研究

これまでも地域情報検索に関してさまざまな研究が行われてきた。ここでは、いくつかの研究を簡単に紹介する。図にそれぞれの特徴を表の形式で示してある。

２．１ ユーザの質問処理について

まずは、ユーザの質問形式とそれに対する検索方法についての研究について触れる。

- 横路氏らの研究[9]（図２のここのサーチ）では、ウェブページを地図上で、住所に対応したポリゴンとして管理している。一方ユーザの質問は、現在地を中心とした円であり、各ポリゴンとの重なりを調べることでページの検索を行っている。
- Digital City の GeoLink[10]☆（図２の Digital City）は、地図上にあらかじめ登録されたウェブページへのリンクが置かれており、ユーザはこの地図上の点をクリックすることで、ページの閲覧ができる。また、キーワード検索を備えているとともに、各ページは観光、買物などの目的別に分類されているので、ユーザは目的に応じてページを絞り込むことができる。現在ウェブ上で使われている地図サービスにはこのような形式が多い。

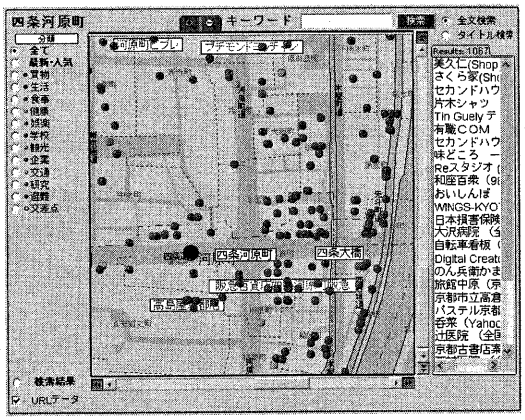

図 2：GeoLink of Digital City Kyoto Project

２．２ ウェブページの地域性について

続いて、ウェブページの地域性をどのように判断するかについて触れる。ページの地域性を求める主な手法として、**内的判断**と**外的判断**の２種類がある。内的判断は対象ページの内部に含まれている情報のみを用いてページの対応する地域を求める手法である[2][7][8]。外的手法は、対象ページが、どのようなページからリンクされているか、どのような地域で利用されているかといった情報を用いてページの地域性を判断する手法である[3][4]。

- 「ここのサーチ」は、ウェブページに含まれる住所を抽出して、地域性を求めている（内的判断）。厳密な住所マッチや、丁目表記のばらつき（全角数字、半角数字、漢数字）を正規化して抽出する点が特徴的である。
- 馬氏らの研究[12]（図２のローカル度）では、さまざまな情報からローカル度という値を計算している。内的判断としては、地名の出現頻度、詳細度、MBR を利用した分布状況に加え、日常用語の出現回数なども用いている。外的判断としては、他のページとの類似度や、IP を利用したユーザの地域分布などを用いている。

☆ http://www.digitalcity.gr.jp/openlab/kyoto/geolink/2dmap.html

- 我々の研究の一つ[5]（図２の Locality Ranking）に外的判断を用いたものがある。この研究では、ウェブページのリンク構造を利用して、ウェブページがどれだけその地域から評価されているかという、地域における人気度と、どれだけその地域を意識した内容を持っているかという、地域に対する指向性からウェブページの地域性を評価した。

次章ではこれらの研究と比較しながら、我々の考案した地域ウェブ検索手法（図２の KyotoSEARCH）を紹介する。

３．MBR を利用した領域質問処理

地図インターフェースの利点として、ユーザが直感的に検索領域を指定できる点がある。本研究では、この利点を生かすために、ユーザが地図上に自由に矩形の領域を描くことで、検索範囲を指定するという方法をとった。このような領域質問に対しては、ウェブページも２次元領域に対応させて、地理演算で検索を行うことが望ましい。ウェブページの領域表現については、簡単さと計算量の面で MBR（Minimum Bounding Rectangle）による提案がなされている[7][8]。その他の領域表現として関連研究「ここのサーチ」では、ポリゴンを用いている。これらの２次元検索には R-Tree[1]を用いた空間索引が利用されている。本研究では、管理と計算の簡単さを重視して、MBR を用いることとする。ポリゴンが複雑な形を正確に表せるのに対してノイズの可能性が大きいが、２つの点の緯度経度座標だけで表せるので記憶、計算の面でサーバの負荷を軽減できる。また、MBR の分解と合成によって、複雑な形に対応することも可能である。

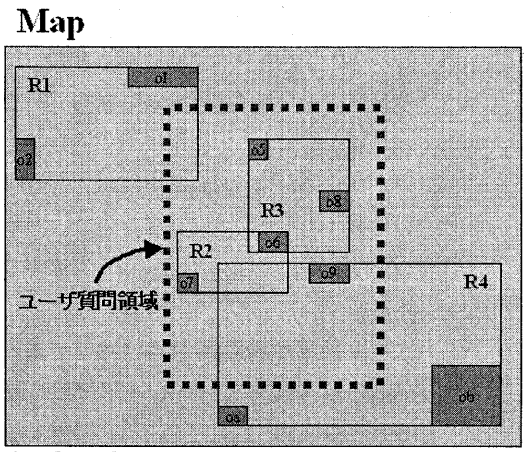

図 3　MBR を用いた領域質問処理

図３に MBR の例を示す。ページ P1 が地名 o1,o2 に関連する場合、P1 は領域 R1 で表される。地理演算には包含(Contain)と共通部分(Intersection)がある。この場合、ユーザの質問領域と各 MBR に対して Contain 演算を行うと R2,R3 が検出される。さらに Intersection 演算で R4 も見つかるが、質問領域と関係のない地名しか含まない R1 まで見つかってしまう。ユーザの要求に合ったページを見つけるためには、Contain 演算のみで検索を行うほうが望ましい。そのためには、質問領域に包含されやすいように本当に関連のある地名だけを使って、できるだけ MBR のサイズを小さくする必要がある。

そこで、どのようにウェブページの対応地域を求めるかが問題となる。２章で述べたとおり、この問題には内的判断と外的判断の２種類の方法が存在する。外的判断については、膨大な母集団の収集／分析の必要性やリンクを利用した地理領域決定方法の妥当性から、的確な実現が困難であるため、本研究では、内的判断の手法を用いてページの対応する地理領域を求める。

ウェブページ内の情報の中でもページ中に出現する地名を利用している。ここでは地名として、区名、町名といった住所のほかに、建物などの地理オブジェクト名を用いた。しかし、ただ単にページ中に出現する地名をそのまま使ってMBR を生成してもよい結果は出ない。ページ中の地名から対応する地理領域を求めるには以下のような問題がある。

問題となる地名の性質

i) **地名の重要性**：ウェブページ内の地名にはページの主題となる重要な地名と説明のために用いられるあまり重要でない地名がある。例えば、「銀閣寺」について述べたページ中に『「京都駅」からバスで○○分』というアクセス方法の説明があったとする。ここで、「銀閣寺」「京都駅」という地名が出現するからといって両方を用いて MBR を生成すると無関係な領域を多く含む MBR になってしまう。

ii) **地名の階層構造**：地名には市、区、町字、地理オブジェクトなどさまざまな階層がある。例えば、「銀閣寺」について述べたウェブページに所在地として「京都市左京区」と書いてあったとする。この時、「銀閣寺」「京都市左京区」両方を用いてMBR を生成すると「京都市左京区」全体を含む大きなMBR が生成されてしまう。

iii) **地名の曖昧性**：同じ地名を持つ場所が複数存在することがある。例えば、「薬屋町」という町字がページ中に出現したとする。「薬屋町」という町字は「京都市上京区」「京都市中京区」の両方に存在する。ここで、両方の「薬屋町」を含む MBR を生成すると不正確かつ大きなMBR が生成されてしまう。

4．地名の性質を利用した地域情報抽出

これらの問題への解決策を以下に述べる。まず、地名のマッチを行うために、たとえば町字を認識するために以下のような住所リストを作成した。

(1) 町字名　ex.)薬屋町
(2) 区／町字名　ex.)上京区薬屋町
(3) 市／区／町字名 ex.)京都市上京区薬屋町
(4) 府／市／区／町字名 ex.)京都府京都市上京区薬屋町

このようなリストと最長マッチを行うことで地名の曖昧性と階層の問題に対処することができる。(2)～(4)の表記に限定してマッチを行えば、町字を唯一に特定することができる。関連研究「ここのサーチ」はこのような方法で住所の特定を行っている。しかし、本研究では、(1)の表記のものについても後の述べる手法で対処できるため、このような曖昧性を持った表記も抽出する。なお、京都市内には 3,658 個の町字が存在し、それに対応する MBR は 30,063 個あり、平均で約 8.21 の曖昧性があった。

他には、区名と建物などの地理オブジェクトについての住所リストを作成した。ただし、店名が普通名詞や人名になっているお店などは、無関係なページを検出する可能性が高いので除いてある。

京都市内について、以上の住所リストからその対応する場所の MBR（複数の候補がある場合はすべての候補を用意）を求めるデータベースを作成した。MBR は©ZENRIN☆の地理データを用いて計算し、緯度／経度の座標を用いて表した。　続いて、i)～iii)の問題への解決策を

☆ ©Zenrin の住宅地図電子データ(TOWNII)

ウェブページ内の情報の中でもページ中に出現する地名を利用している。ここでは地名として、区名、町名といった住所のほかに、建物などの地理オブジェクト名を用いた。しかし、ただ単にページ中に出現する地名をそのまま使ってMBR を生成してもよい結果は出ない。ページ中の地名から対応する地理領域を求めるには以下のような問題がある。

問題となる地名の性質

i) **地名の重要性**：ウェブページ内の地名にはページの主題となる重要な地名と説明のために用いられるあまり重要でない地名がある。例えば、「銀閣寺」について述べたページ中に『「京都駅」からバスで○○分』というアクセス方法の説明があったとする。ここで、「銀閣寺」「京都駅」という地名が出現するからといって両方を用いて MBR を生成すると無関係な領域を多く含む MBR になってしまう。

ii) **地名の階層構造**：地名には市、区、町字、地理オブジェクトなどさまざまな階層がある。例えば、「銀閣寺」について述べたウェブページに所在地として「京都市左京区」と書いてあったとする。この時、「銀閣寺」「京都市左京区」両方を用いてMBR を生成すると「京都市左京区」全体を含む大きなMBR が生成されてしまう。

iii) **地名の曖昧性**：同じ地名を持つ場所が複数存在することがある。例えば、「薬屋町」という町字がページ中に出現したとする。「薬屋町」という町字は「京都市上京区」「京都市中京区」の両方に存在する。ここで、両方の「薬屋町」を含む MBR を生成すると不正確かつ大きなMBR が生成されてしまう。

4．地名の性質を利用した地域情報抽出

これらの問題への解決策を以下に述べる。まず、地名のマッチを行うために、たとえば町字を認識するために以下のような住所リストを作成した。

(1) 町字名　ex.)薬屋町
(2) 区／町字名　ex.)上京区薬屋町
(3) 市／区／町字名 ex.)京都市上京区薬屋町
(4) 府／市／区／町字名 ex.)京都府京都市上京区薬屋町

このようなリストと最長マッチを行うことで地名の曖昧性と階層の問題に対処することができる。(2)～(4)の表記に限定してマッチを行えば、町字を唯一に特定することができる。関連研究「ここのサーチ」はこのような方法で住所の特定を行っている。しかし、本研究では、(1)の表記のものについても後の述べる手法で対処できるため、このような曖昧性を持った表記も抽出する。なお、京都市内には 3,658 個の町字が存在し、それに対応する MBR は 30,063 個あり、平均で約 8.21 の曖昧性があった。

他には、区名と建物などの地理オブジェクトについての住所リストを作成した。ただし、店名が普通名詞や人名になっているお店などは、無関係なページを検出する可能性が高いので除いてある。

京都市内について、以上の住所リストからその対応する場所の MBR（複数の候補がある場合はすべての候補を用意）を求めるデータベースを作成した。MBR は©ZENRIN☆の地理データを用いて計算し、緯度／経度の座標を用いて表した。　続いて、i)～iii)の問題への解決策を

☆ ©Zenrin の住宅地図電子データ(TOWNII)

まず、各ページに対してページ中に現れる地名をすべて含み、同名の場所があればそのすべてを含むMBR（PureMBRと呼ぶ）を作成し、その面積を求めた（累積ヒストグラムを図4にPureMBRとして示す）。この結果、1km²以下に収まったページは40%程度に過ぎず、50%程度が8km²以上になってしまった。この状況でユーザがContain演算だけで的確にページ検索を行うことは困難である。

続いて、4章に挙げたMBRの最適化案(A)Title/Anchorに現れる地名の使用（図4のTitle +Anchor）、(C)上位階層の考慮（図4の区名削除）、(B)同名候補の解決（図4の曖昧性解決）をそれぞれ適用した。さらに3つの手法を(A)(B)(C)の順に一度に適用した場合（図4のLastMBR）、70%以上のページを1km²以内に収めることができた。この改善結果から、ユーザの質問領域に対してContain演算のみでも検索を行うことができる。

6.今後の課題

本稿では、MBRを用いた領域質問処理と、地名の性質を利用して的確にウェブページの位置情報を摘出する手法を考案した。これによって、ユーザは地図インターフェース上で、自由な領域質問を行い、その領域内に含まれるウェブページを検索することができる。一方、サーバも効率よく、ウェブページの管理及び検索を行うことができる。しかし、実用に耐えうる地理情報検索システムを構築するためには今後も改良が必要である。

本研究ではウェブページの地名をタイトル・アンカーからのみ抽出したが、これは実験の結果でも大きな効果が出ている。特にタイトルに現れる地名は数も少なく、ページ中の最も重要な地名をさしている可能性が高い。しかし、観光地のような地理オブジェクト名はタイトルやアンカーに現れることが多いが、住所が現れることは少ない。京都市は観光地が多いため、地理オブジェクト名で多くのページが検索できるが、他の地域もそうとは限らない。今回、お店の名前が普通名詞であるようなものを省いたが、このようなお店に関するページを見つけるためには、ページ内の住所を見つけることが重要である。したがって、タイトルとアンカーに限定してしまうのではなく、ページ内に現れる住所なども考慮する必要がある。

関連研究「Digital City」のような形式はウェブページを地図に登録しなくてはならないために、検索できるページ数は少ないが、確実にその地域に関連したページが検出できる。また、ページを分類できるなど、ユーザの目的に合った検索ができ、現在最も実用的である。ページの検索に加えてこのような分類まで行うシステムが理想的である。ユーザの目的に合わせた検索を行うためには、地理検索の検索結果にキーワード検索を行うといった手法が考えられる。

今回の研究や、関連研究「ここのサーチ」では地名を単なる位置情報としか見ていないが、関連研究「ローカル度」は、地名の出現回数や、日常用語の検出など言葉の面からも地理情報検索を行っている点が特徴である。KyotoSEARCHはウェブ情報、地理情報、キーワード情報の3つを扱うことを目的としているので、このような言葉の面での分析もおおいに興味の対象となる。今後は、地名をキーワードとしての観点から分析するなど、地理情報とキーワード情報の関連について研究を行う予定である。

謝辞：本研究は、21世紀COEプログラム「知識社会基盤構築のための情報学拠点形成」および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（代表石田亨）の支援によって行われた。また、韓国科学技術部・韓国科学財団指定「韓国航空大学インターネット情報検索センター」と京都大学情報学研究科上林研究室との共同研究成果の一部でもある。

文　献

[1] G. Antonin, R-TREE: A Dynamic Index Structure for Spatial Searching, In proceeding of ACM SIGMOD, pages 47-57, 1984.

[2] M. Arikawa, K. Okamura, "Spatial Media Fusion Project," In Proc. of Kyoto International Conference on Digital Libraries: Research and Practice, pp.75-82, Nov. 2000.

[3] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar, "Exploiting geographical location information of web pages," In Proc. of the ACM SIGMOD Workshop on the Web and Databases, WebDB, 1999

[4] J. Ding, L. Gravano and N. Shivakumar, "Computing Geographical Scopes of Web Resources," VLDB2000, pp.545-556, 2000.

[5] Y. Inoue, R. Lee, H. Takakura, and Y. Kambayashi, "Web Locality Based Ranking Utilizing Location Names and Link Structure," The 2nd Int. Workshop on Web Geographical Information Systems, IEEE CS Press, Singapore, Dec. 2002. BEST PAPER.

[6] R. Lee, H. Takakura and Y. Kambayashi, "Visual Query Processing for GIS with Web Contents," The 6th IF Working Conference on Visual Database Systems, pp.171-185, May 29-31, 2002.

[7] C. Matsumoto, Q. Ma, and K. Tanaka. Web Information Retrieval Based on the Localness Degree. In Proc. of the 13th int'l Conf. on Database and Expert System Applications 2002 (DEXA '02), pages 172-181, 2002.

[8] N. Yamada, R. Lee, H. Takakura, and Y. Kambayashi, "Classification of Web Pages with Geographic Scope and Level of Details for Mobile Cache Management," The 2nd Int. Workshop on Web Geographical Information Systems, IEEE CS Press, Singapore, Dec. 2002.

[9] 横路誠司，高橋克巳，三浦信幸．島健一，“位置指向の情報の収集,構造化および検索手法”，情報処理学会論文誌，Vol.41，No.7，pp.1987-1998，2000

[10] K.Hiramatsu and T.Ishida, “An Augmented Web Space for Digital Cities”, IEEE/IPSJ Symposium on Applications and the Internet 2001, pp.105-112, IEEE CS Press, 2001.

[11] 相楽毅，有川正俊，坂内正夫，“分散位置参照サービス”，情報処理学会論文誌，Vol.42，No.12，pp.2928-2940，2001.

[12] 馬強，松本知弥子，田中克巳，“ページ内容と位置情報に基づく Web コンテンツのローカル度検出とその応用”，情報処理学会研究報告，Vol.2002，No.67,2002-DBS-128-69，pp.515-522，2000.