メタサーチエンジンを用いたウェブ上の地域情報要約システム

井上　陽介† 　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　〒 606–8501 京都府京都市左京区吉田本町
†† Ubiquitous Computing Lab. Samsung Advanced Institute of Technology
P.O. Box 111, Suwon 440-600 Korea
††† 京都大学 学術情報メディアセンター ネットワーク研究部門 高機能ネットワーク研究分野
〒 606–8501 京都府京都市左京区吉田本町
E-mail: †yohsuke@db.soc.i.kyoto-u.ac.jp, ††ryong@sumsung.com, †††takakura@media.kyoto-u.ac.jp

あらまし　本論文では、地域情報検索を支援するメタサーチシステムを提案する。既存のキーワードベースのウェブ検索システムは急速に発展しその精度を高めてきた。しかしその一方で、地域情報を検索する場合、１）クエリとして地名が必要なため、対象地域に関する知識が要求される、２）キーワードだけでは他地域の同名の地名などと区別できない、などの問題点が存在する。よく似た内容のページが検索結果の上位を独占してしまうことも多い。地域情報検索の手法としてから広く用いられてきた地図をユーザインタフェースとして使用し、利用者の選択した地域から適切な地名を選択、さらに必要に応じて他の地名を追加・削除し、既存の検索システムへの橋渡しを行なうメタサーチシステムを提案する。また、得られた地域情報から類似トピックを要約して提示する手法についても議論する。

キーワード　GIS, WEB, メタサーチエンジン

Summarization of Geographical Information on the Web Using Meta Search Engine

Yohsuke INOUE†, Lee RYONG††, Hiroki TAKAKURA†††, and Yahiko KAMBAYASHI*†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto city, Kyoto, 606–8501 Japan
†† Ubiquitous Computing Lab. Samsung Advanced Institute of Technology
P.O. Box 111, Suwon 440-600 Korea
††† Academic Center for Computing and Media Studies, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto city, Kyoto, 606–8501 Japan
E-mail: †yohsuke@db.soc.i.kyoto-u.ac.jp, ††ryong@sumsung.com, †††takakura@media.kyoto-u.ac.jp

Abstract　In this paper, we propose a meta-search system, which summarizes geographical information on the web. Conventional search systems are improved dramatically, however they are mainly focused on keyword-base approarch. Hense, for the purpose of geographical retrieval, 1)Users' knowledge about the target geographical area is required, 2)Keyword-based retrieval system cannot distinguish one geographic area from another if they have identical names. In addtion, top ranked web pages often have the same topic, if they are ranked simply by their popularities. Our proposed system realizes efficient geographical information retrieval by extracting various information from the web and summarizing them, with the help of Geographical Information System(GIS) and conventional web retrieval systems. In this paper, we propose how to extract web pages, in which important topic of the area can be covered. Then, we will discuss how to summarize extracted topics.

Key words　GIS, WEB, Meta-Search-System

---

* 2004 年 2 月 6 日逝去

1. はじめに

インターネットの爆発的な普及により、世界中の情報に誰もがアクセスできるようになった。反面、膨大な情報はユーザに混乱を招き、目的の情報にたどり着くことが困難になっている。PDA や携帯電話などウェブにアクセスできる機器が急激に増加したことにより、ウェブはますます複雑になり、利用者はさらに混乱すると予想される。

このような現状に対応するために、さまざまなウェブ情報検索システムが提案され、実際に運用されている。これらの既存システムは、すでにある程度の成果を挙げ、高い評価を受けている [2]。しかし、その大半がキーワードを用いた一般的なウェブ情報検索に注力しているため、利用者はどのようなキーワードをクエリとして与えるか工夫しなければならない。これは、利用者に検索対象に対するある程度の知識を要求する。

また、システム固有のアルゴリズムにより各ページの重要度を計算してランキングを行うのが一般的だが、これは特定のトピックを扱うウェブページが検索結果の上位を独占してしまうことがある。そのため、ユーザは目的の情報を得るために試行錯誤を重ねなければならない。

このような問題は、地域情報検索の場合は特に大きな問題となる。ウェブ上から地域情報検索を行うユーザは、検索対象となる地域に詳しいとは限らない。これに対し、既存システムでは、検索時にどのようにして地域を限定するような検索条件を加えるか考慮しなければならないため、このようなユーザには使いにくいシステムになっている。また、検索結果の上位が特定のトピックで埋まってしまうために利用者はその地域のさまざまな情報を認知することすらできないという課題もある。

我々はウェブ上の情報と、地理情報システムを統合して利用することで、効率的に特定地域に関する情報検索を支援するシステム KyotoSEARCH [7] を開発し、その中で特定地域の情報を扱うウェブページをランキングする手法 [6] を提案してきた。本論文では、これをさらに進め、既存のウェブ情報検索システムと地理情報システムを活用し、地域情報検索を効率的かつ効果的に行うメタサーチシステムを提案する。

我々の提案するシステム (図 1) では、クライアントサイドのユーザインターフェースに地域情報検索の手法として広く用いられてきた地図を用い、利用者は地図インターフェースを用いて検索対象地域を領域として選択する。システムは、利用者の選択した地域から適切な地名をクエリとして生成し、既存のウェブ検索システムを最大限に活用した地域情報検索を実現する。さらに、特に検索エンジンから対象地域の情報を網羅的に取得し、類似情報をクラスタリングして提示することにより、利用者に地域情報の要約を提示することも提案する。

以下、本論文は以下のような構成をとる。２章で我々の提案する地域情報検索のためのメタサーチシステムの概要について説明する。３章では外部のウェブ情報検索システムを最大限に活用するためのクエリの生成手法について議論する。４章では、外部のシステムから検索結果として得られたウェブページを本システムの目的に沿って評価し、必要な情報を抽出する手法について述べる。５章では我々のシステムの実現・実装について述べる。６章が結論である。

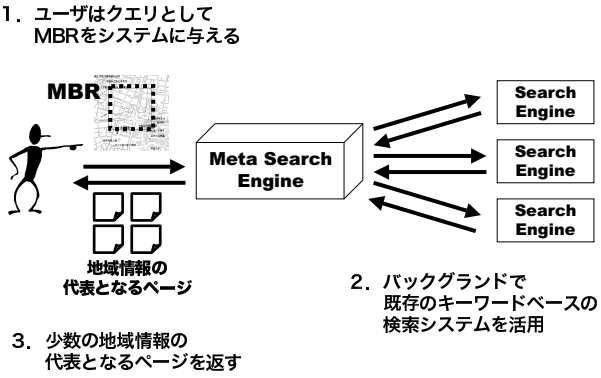

図１　メタサーチエンジンを用いた地理情報要約システム

2. システム概要

本章では、提案するメタサーチシステムの概要について説明する。提案するシステムは、「外部システムに送るクエリの生成」「外部システムの返したウェブページの評価」「ウェブページのクラスタリングを用いた地域情報の要約」の３つのモジュールから構成される。以下、それぞれのモジュールの動作・役割について説明する。

2.1 クエリの生成

我々の提案するメタサーチシステムはバックグラウンドで、既存のウェブ情報検索エンジンを用いているため、これらの検索エンジンに与えるクエリを生成しなければならない。ここでは、ユーザがクライアントプログラム上で指定した検索対象地域を、地理情報システムの知識を活用して地名集合に変換する役割を担う（図 2)。

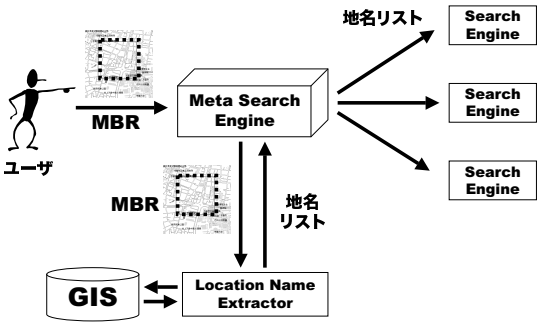

図２　クエリの生成

ユーザからはウェブブラウザ上で動作するアプレット上で動作する地図インターフェースを通じ、メタサーチエンジンに検索対象地域がクエリとして渡される。地図インターフェース上で地域を指定する方法としては、地図上の任意の地理オブジェクトを選択したり、地図上の任意の領域を選択したりといった方法が考えられるが、ここでは最小矩形領域 (Minimum Bounding Rectangle, MBR) をクエリとして用いる方法を選択する。

図 2 にあるように、ユーザから渡された MBR はサーバ側で

地名集合に変換され、順に検索エンジンにクエリとしておくられるが、地理情報システムの持つ地名集合は、ランドマーク・住所・施設名称などさまざまなものがあり、そのすべてをあわせると小さな領域であっても膨大な地名の数になる。また、同名の別の地理オブジェクトが多く存在するようなあいまい性の高い地名、検索結果の期待できない地名などが多い。そのため、効果的なクエリ生成が必要となる。これについては、3. で述べる。

また、提案するシステムでは最終的に類似ページをクラスタリングして、クライアントサイドで出力することを前提としている。そのため、多様なトピックを網羅的に含むウェブページ群を抽出できるようなクエリの生成が期待される。

2.2 ウェブページの内容分析

各検索エンジンが返したウェブページを、そのままの形でユーザに返すことはできない。まず、これらのウェブページは、キーワードに対する検索結果であり、同名の他の地名の情報が混ざっている場合があり、対象地域の情報だけを含むようにフィルタリングする必要がある。また、集めたウェブページをトピックごとにまとめるために、各ページの特徴となるキーワードベクトルが必要となる。

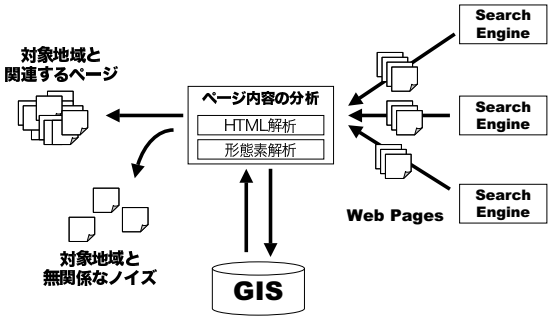

図 3　ウェブページの解析

そこで、各ウェブ検索エンジンの返した検索結果のウェブページを実際にアクセス・取得し、内容解析したうえで、評価する必要がある。この内容解析は、主に以下の二つの処理から構成される。

- **対象地域と無関係なページのフィルタリング**

検索対象地域と無関係な情報を含むページをフィルタリングによって排除する処理。ウェブページの HTML および、日本語の形態素解析により、ページ内に含まれるすべての地名について、タグによる強調、出現頻度などが取得される。これらの地名の出現の仕方から、対象となるウェブページが検索地域と関連のある情報かどうかを判断する。

- **各ページの特徴を表現するキーワードの抽出**

TF/IDF 法により、このウェブページを特徴付けるキーワードを抽出する。この特徴となるキーワードは、類似トピックを扱うウェブページを分類する際に用いられるが、それ以外にもの検索時に利用される。

2.3 ウェブページの分類・選択

検索結果は整理された状態でユーザに返されなければならない。提案するシステムでは、ウェブページはトピックが類似したものをまとめて、各トピックごとに代表となるページを利用者に返す。

このステップは、図 4 に示すように、前処理で抽出されたウェブページを、内容に応じてトピックごとに分類する役割を担う。各ページからは前処理で特徴となるキーワードが抽出されているので、このキーワードを用いて各ページの類似度を定義することができる。類似度を計算し、それを用いてクラスタリングし、各クラスタの代表となるページを利用者に返す。

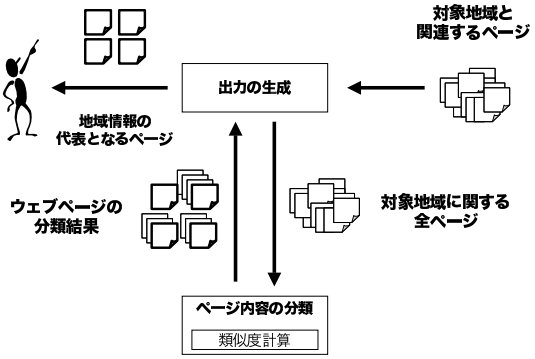

図 4　ウェブページの分類・選択

3. クエリの生成

本章では、提案するシステムが外部のウェブ情報検索システムに渡すクエリの生成手法について議論する。

提案するシステムは、ユーザから地図上のある領域をクエリとして受け取ることを前提としている。ユーザの指定した地域の情報を既存のウェブ情報検索システムを用いて検索するためには、地図上の領域をキーワードに変換しなければならない。本システムでは地理情報システムから取得した領域内の地名を用いることでこの問題を解決する。

また、集めたページをトピックごとに分類するという本システムの特性上、外部の検索システムから返されるウェブページは、検索対象地域に関する多様な情報を網羅的に含むことが期待される。

以下、本章ではユーザから与えられた検索対象領域に存在する地名を用いて、外部の検索システムから多様な情報を網羅的に取得するクエリの生成について述べる。

3.1 地名の種別とウェブ検索結果

地名には、郵便番号辞書に含まれるような行政区分けを基にした地名、その地域に存在する施設名、またそれらの通称・別称など、さまざまなものがある。本論文ではこれらのすべてを含めたものを「広義の地名」、そのうち地理情報システムが扱うことのできる「地名」として扱う。

外部の検索システムに与えるクエリには、地理情報システムの持つ地名を用いる。地理情報システムに含まれる地名は、一般的に **1) 住所型地名**、**2) ランドマーク**、**3) その他の施設名称**に分類される。しかし、地名の個数は膨大であり、またあいまい性を含むものも多く、そのすべてを外部の検索システムでク

エリとして用いることは適切ではない。そこで地名の選択が要求される。

以下、本章では各カテゴリの地名をクエリとして与えた場合に、外部の検索システムがどのような振る舞いを示すか考察する。

3.1.1 住所型地名を用いた検索

地理情報システム上の地名でもっとも扱いやすいのが住所型の地名である。指し示す領域が行政区界として明確に定義されており、また郵便番号辞書などの形で入手もしやすい。しかし、ウェブ情報検索においてキーワードとして用いた場合は、以下に示すような問題があり不適切である。

図 5、図 6、図 7 はある郵便番号辞書に含まれる「田中門前町」という住所型の地名を一般のキーワード型のウェブ情報検索エンジン [11] で検索した場合に上位に現れる検索結果の一部である。「田中門前町」は京都市左京区にある地域で、百万遍の名で知られる知恩寺があり、フリーマーケットが定期的に開催され、学生街として学生向けのマンションや飲食店が立ち並ぶなどのさまざまな特徴があり、このような情報を提供するウェブページが見つかることが期待される。

図 5、図 6 は、それぞれ郵便番号辞書サイトに含まれるページのひとつ、天気予報サイトに含まれるページのひとつである。これらは「田中門前町」以外にも近隣のすべての住所型地名に対して同様のページが存在すると思われる情報である。そのため、この地域を代表する情報としては不適切である。

図 7 は京都の古書店をリストアップしたウェブページである。田中門前町に古書店が存在することがわかる点で、前述の図 5・図 6 に示すページよりも地域情報を含むと判断できるが、ページ内に多量に含まれる住所型地名のなかにたまたま「田中門前町」が含まれていただけで、このページ内に含まれる「田中門前町」の情報はわずか数行にすぎない。

これらの情報は、確かに「田中門前町」の情報ととらえることも可能ではあるが、期待される情報とは大きく異なる。これはウェブ検索エンジンの多くが、「田中門前町」というクエリに対して、「田中門前町」が強調されたウェブページを返そうとしていることに起因する。入力されたキーワードを強調したウェブページを上位に配置する傾向があり、「田中門前町」が強調表示された上記の図 5・図 6 などが上位で返されやすい。外部のウェブ検索システムには、このような住所型地名ではなく次に説明する「ランドマーク」を用いるべきである。

3.1.2 ランドマークを用いた検索

ランドマークは地理情報システムのもつ大きなアドバンテージである。一般に、日常会話である地域について話題になる場合は、わかりやすい目標物に対して、「〜の近く」「〜のそば」などの表現で地域を指し示すことが多い。ウェブ上の地理情報についても、位置情報としてこのランドマークが用いられていることが多分に期待される。

しかし、ランドマークを用いた検索には次で議論するような傾向があり、そのままランドマークをクエリとして用いることには問題がある。

まず、ランドマーク自身が重要な情報を持つ場合に、周辺情報をかき消してしまうという問題がある。図 8、図 9 は京都の代表的な観光名所である「銀閣寺」を一般のキーワード型のウェブ情報検索エンジン [11] で検索して得られた結果である。

図 8、図 9 ともに、観光客に向けた観光名所としての銀閣寺を解説したページである。銀閣寺そのものについて歴史・見所などに触れられてはいるが、周辺の地理情報については、銀閣寺への交通などを除き、触れられていない。このようにランドマーク自身が強すぎるために周辺情報をかき消してしまう。

また、ランドマークの名称が一意ではないという問題もある。地理情報システム内に含まれるランドマークには、その地域の代表的な施設名が多く含まれているが、その中には名称にあい

図 5　「田中門前町」の検索結果 1
京都府 京都市左京区 田中門前町の天気予報

図 6　「田中門前町」の検索結果 2
郵便番号辞書

図 7　「田中門前町」の検索結果 3
京都の古書店一覧

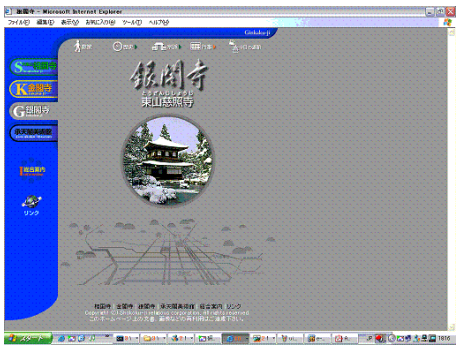

図 8　「銀閣寺」の検索結果 1
公式のウェブサイト

図 10　「熊野神社」の検索結果 1
秋田 三皇熊野神社

図 9　「銀閣寺」の検索結果 2
観光客向けのガイド

図 11　「熊野神社」の検索結果 2
自由が丘　熊野神社

まい性を含むものが存在する。

「銀閣寺」と同様に、京都市左京区内にある「熊野神社」というランドマークをウェブ情報検索に用いた結果が、図 10、図 11 である。図 10、図 11 はどちらも京都の情報ではない。「熊野神社」という同じ名称をもったまったく違う施設の情報である。「熊野神社」というランドマークは日本全国に散在しており、結果として、京都の熊野神社の情報が埋もれてしまう。

いずれの問題も、ランドマークの名称を単体で使った場合のものであり、ランドマーク名に別のキーワードを加えることである程度防ぐことが可能である。この問題点を解決する手法については 3.2 章で議論する。

3.1.3　施設名称を用いた検索

この分類には地域内に含まれる飲食店、マンション名などランドマークに含まれなかったすべての施設が含まれる。ここに含まれる地名は、**1)** あいまい性が高く名称から地域が一意に定まらない、**2)** 地理情報システム上に記載された名称と実際の名称が必ずしも一致しない、**3)** 個数が膨大であり、クエリとして用いる地名を選択する手法が困難、などの問題があり検索に用いるクエリとしては不適切である。

3.2　ランドマークを用いた再帰的な検索

3.1 章で示されたように、地名をウェブ検索用のクエリとして用いた場合にはさまざまな問題が発生する。問題は「単純に地名を単体でクエリとしてもちいること」に起因する。本章では、検索条件を足し合わせることでこの問題を解決する手法と

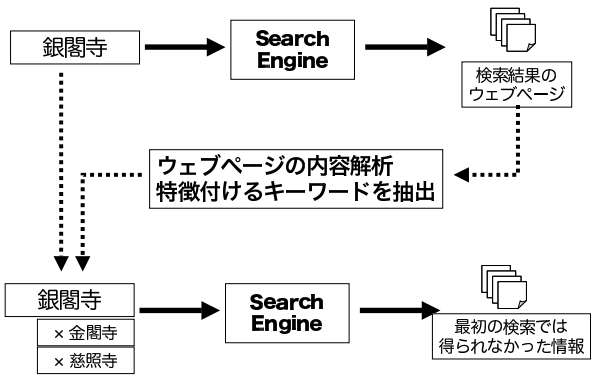

図 12　再帰的なクエリの生成

して、再帰的なウェブ情報検索を提案する。

さきほどの「銀閣寺」の例のようにランドマーク自身が強すぎるためにランドマークが周辺の情報をかき消してしまう場合、地域の情報を網羅的に取得するために、ランドマーク自身を除外するような検索条件が要求される。また、特定のトピックが検索結果の上位を独占してしまう場合なども同様である。

検索結果の上位に特定のキーワードが多く含まれていた場合には、そのキーワードを除外して再検索する。観光地としての「銀閣寺」を紹介するウェブページでは、同じ京都の観光名所である「金閣寺」と併記されるケースや、「慈照寺」という銀閣寺の別名を含むページが多く、実際「銀閣寺」での検索の上

位のページのほとんどにこの２つのキーワードが現れている。そこで「銀閣寺」を含み、「金閣寺」「慈照寺」を含まないウェブページを再検索する（図 12)。

図 13　「銀閣寺 -金閣寺 -慈照寺」の検索結果 1
Yahoo! グルメ 銀閣寺 貝

図 14　「銀閣寺 -金閣寺 -慈照寺」の検索結果 2
銀閣寺 あじろ

図 13、図 14 は「銀閣寺」を含み、「金閣寺」および「慈照寺」を含まないウェブページを検索したときに得られた結果の一部である。どちらも「銀閣寺」周辺にある飲食店の情報であり、新しいキーワードによって、周辺の新しい情報が得られたといえる。対象地域に関する情報を網羅的に取得するという「クエリ生成」における目標をより確実に達成できる。

逆に、「熊野神社」のように対象地域と無関係な情報が多く含まれてしまう場合についても考察する。「熊野神社」の場合、同名のさまざまな神社の情報が検索結果に混ざっているために目的の情報が十分に得られないという問題を持つ。ノイズとなる情報に共通したキーワードがないため、「京都の熊野神社」であることを明確にするために何らかの地理情報を付加することで対応する。

例えば、「熊野神社」を含み、「京都」または「聖護院」を含むページでウェブ情報検索をした場合の検索結果の一部が図 15、図 16 である。どちらも、京都の熊野神社自身の情報、または周辺の施設の情報である。

このように、検索条件を付加することで、ノイズを減らし、より網羅的に地域情報検索ができる。このためには、「外部の検索エンジンから検索結果として得られたウェブページが対象地域に関する情報であるかを判断することができること」、また「外部の検索エンジンから検索結果として得られたウェブページの特徴を示すキーワードを抽出できること」が要求される。この点については、4. 章で議論する。

図 15　「熊野神社 (京都 or 聖護院)」の検索結果 1
京都十六社 朱印めぐり

図 16　「熊野神社 (京都 or 聖護院)」の検索結果 2
京都市左京区聖護院 旅館 丸家

ユーザが MBR で指定した検索対象領域が極端に狭い場合、検索に用いるランドマークが十分に得られない場合が考えられる。この場合は地理情報システム上の「その他の施設」に分類される施設名をランドマークの代用に用い、「熊野神社」のケースと同様に「京都」などの住所型の地名を付加することであいまい性を軽減して用いることで、検索が可能となる。

4.　ウェブページの解析

各検索エンジンから返された検索結果は解析され、吟味されなければならない。この段階で各ウェブページが扱う地域を判定し、そのトピックについて評価される。この解析の目的は次の通りである。

（1）　検索結果として返された各ウェブページの扱う地域を評価し、与えられた検索対象地域の情報を含むかどうか評価し、地域と無関係なページをフィルタリングする。

（2）　検索結果として返されるウェブページ集合の上位のページについて、共通するトピックを抜き出し、再検索に利用するキーワードを追加する。

（3）　各ページに含まれる単語に重みを与え、特に重要なものを選択して特徴ベクトルを生成する。この特徴ベクトルにより各ページの類似度を定義し、ページのクラスタリングを行う。

4.1　ウェブページの地域判定

本章では、ウェブページ内の地名を評価して、ウェブページ

の内容が検索対象地域の地域情報として適切かどうかを評価する手法について述べる。

一般にランドマーク名を使って検索して得たウェブページには検索語として用いたランドマーク名称以外にも複数の地名が含まれる。ここでは、これらの地名を用いて、ランドマーク名が本当に目的となるランドマークを指し示すかどうかを判断しなければならない。

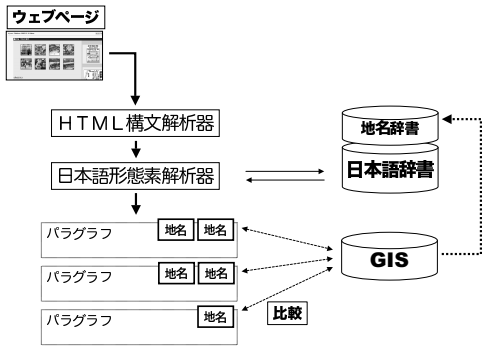

図 17　地域の判定

ここでは、各ウェブページを図 17 のようにして評価する。

（1） HTML 解析によりウェブページを HTML ブロックタグで区切られたパラグラフに分割する。

（2） 地理情報システムの地名知識を利用しながら、各パラグラフの日本語形態素解析を行うことで、地名・キーワードを抜き出す。

（3） HTML タグの強さに応じて、各地名の有効範囲を定義し、それぞれのパラグラフに地名を付加する。地名の有効範囲は基本的にその地名が出現したパラグラフだが、<Hn>タグや<title>タグなどのように出現箇所以降のパラグラフに意味的な影響を与える場合がある。本稿では、地名出現箇所より HTML 的に強調度の高いパラグラフがあらわれるまでを地名の有効範囲と定義する。

（4） 各パラグラフの地名を地理情報システムを用いて、検索対象地域に関連する地名であるか判断する。

（5） 地名情報の付加されたパラグラフのうち対象地域の地名を含むものの割合が一定の閾値 α を超えたページを地域情報を含むページと判断する。

山田ら [9] は、ウェブページが指し示す地域を MBR で表現するために、地名を **(1) キーワードとしての地名**と **(2) 説明語としての地名**に分類した。HTML タグでの強調度、地名の出現頻度、地名の関連後の出現頻度で、キーワードとしての地名を抜き出し、キーワードとしての地名を含む MBR をウェブページの地域的スコープと定義している。

山田らの方法は、「銀閣寺」の例のように特定のランドマークそのものを対象にしたウェブページについてはよく機能するが、住所リストや地理情報システム上に存在しない地域知識まで扱う場合にはうまく機能しない。我々の方法は、地名がウェブページのどの範囲で有効かを見て判断するため、ウェブページの書式・種別を選ばず対応ができる。

4.2　追加検索条件の抽出

本章では 3.2 章で議論した再帰的検索で用いる追加検索条件について述べる。再帰検索は、上位を特定のトピックが埋め尽くしてしまった場合に、そのトピックを取り除く検索条件を加える**「除外検索」**、検索結果の上位に対象地域のトピックが現れなかった場合に地名を付加することで目的の情報を絞り込む**「絞込検索」**に分類される。

a） 除外検索条件の生成

検索結果の上位ページが特定のトピックで埋められていた場合、これらを除外して新しいトピックを取得することを目的として２回目の検索を行う。２回目の検索では、最初の検索の上位のページを除外するように、最初の検索結果を網羅するようなキーワード群が要求される。

そのためまず最初の検索で得られた上位数ページを特徴付けるキーワード群を取得する。各ウェブページから前章で述べた日本語形態素解析により単語を切り出し、それぞれ TF/IDF 法によりページを特徴付ける重要なキーワードを上位 n 個選択する。このようにして得られたキーワード群が候補となる。

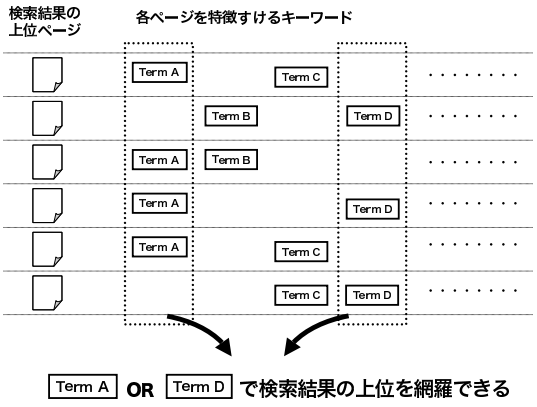

図 18　除外検索条件の生成

除外検索は、このようにして得られたキーワード群の中から２，３単語選択して、検索結果の上位ページを網羅する組み合わせを探す処理に置き換えられる（図 18）。これは貪欲法により計算可能である。

b） 絞り込み検索条件の生成

対象地域の情報が検索結果にほとんど含まれていない場合、絞込検索により、対象地域の情報だけに絞り込む処理が要求される。これは地域を指定するキーワードを追加で指定することで実現する。ここでは最初の検索条件に用いたランドマークの所在地を示すキーワードを追加する。

「熊野神社」の例ならば、京都の熊野神社の所在地である「京都市左京区聖護院」用い、「熊野神社 and (京都 or 左京区 or 聖護院)」と住所型の地名を OR 条件で結んで検索条件として追加する。

5.　実験システム

我々は現在図 19 のようなシステムを構築している。本システムは、ユーザがブラウザ上の地図インターフェースで指定した地域の情報を、キーワードベースのウェブ情報検索システム

である Google [2] を用いて取得し、得られた地域情報を分類・整理して利用者に提示するシステムである。

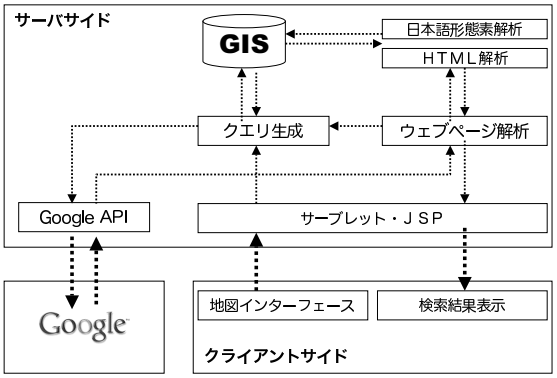

図 19　実験システムの構成

本システムの動作は以下のようになる。まず、ユーザがクライアント側の**地図インターフェース**（図 20）を操作して最小短径領域 (Minimum Bounding Rectangle, MBR) で表現される検索対象領域を指定する。地図インターフェースは Java Applet で実装される。

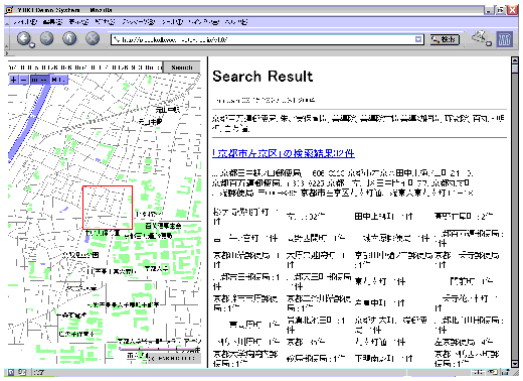

図 20　実験システムのユーザインターフェース

検索対象領域はサーバー側で動作する **JSP** を通じて、**クエリ生成モジュール**に送信される。クエリ生成モジュールは、**地理情報システム (Geographic Information System, GIS)** から対象地域の地名（ランドマーク名）を取得し、ウェブ情報検索システムは現在最も広く使われている Google を、また Google との仲介には Google 自身が配布する **Google API [12]** を用いた。

Google 検索結果の上位１０ページのＵＲＬなどの情報が**ウェブページ解析モジュール**に送られる。各ウェブページを **HTML 解析・日本語形態素解析**したのち、その結果を用いて、ノイズと思われるウェブページのフィルタリング、重要なキーワードの切り出しが行われる。日本語形態素解析には、奈良先端大学院大学の MeCab [14] を用い、一般的な辞書 [13] のほかに地域情報を判断するために地理情報システムのもつ地名辞書を用いた。解析結果として得られた重要なキーワードは、再びクエリ生成モジュールに渡され、再検索に利用される。

上記のようなサイクルを通じて、ウェブページ解析モジュールに集まったウェブページはキーワードを利用して類似度を計算、クラスタリングされる。クラスタリングによって得られた各トピックごとに代表ページを定め、**JSP** で HTML 形式で整形された後、クライアントサイドのウェブブラウザで表示される。

6. 結　　論

本論文では、既存のキーワードベースのウェブ情報検索システムを最大限に活用しながら、効果的かつ効率的にに地域情報を行い、地域情報を網羅的に取得する手法について述べた。また、取得した地域情報をトピックごとに分類して利用者に提示することで、地域情報の要約を容易に取得できるようなシステムについても提案してきた。

PDA や携帯電話など非ＰＣでのインターネット利用が一般的になるにつれ、ウェブ情報検索を行う場所は、机の前から街中へと次第に移り変わっていくと思われる。その中で、ウェブから地域情報を取得したいという要求は、ますます大きなものになっていくだろう。ウェブの情報検索技術は日々進化している。この検索技術の進歩を効率的に利用することで、ウェブを最大限に活用できるだろう。

謝辞　本研究の一部は科学技術振興機構 (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ました。ここに記して謝意を表すものとします。

文　　献

[1] M. Arikawa and K. Okamura. Spatial media fusion project. In *Proc. of Kyoto International Conference on Digital Libraries: Research and Practice*, pages 75–82, 11 2000.
[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:1–7, 1999.
[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *WWW9*, 2000.
[4] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *VLDB2000*, pages 545–556, 2000.
[5] Y. Inoue, R. Lee, H. takakura, and Y. Kambayashi. Computation of page popularity by restricting topics and locations. In *DBWeb 2001*, 12 2001. (in Japanese).
[6] Y. Inoue, R. Lee, H. Takakura, and Y. Kambayashi. Web locality based ranking utilizing location names and link structure. In *Second International Workshop on Web and Wireless Geographical Information Systems*, 12 2002.
[7] R. Lee, H. Takakura, and Y. Kambayashi. Visual query processing for giss with web contents. In *VDB6*, 5 2002.
[8] K. McCurley. Geospatial mapping and navigation of the web. In *WWW10*, 2000.
[9] N. Yamada, R. Lee, H. Takakura, and Y. Kambayashi. Classification of web pages with geographic scope and level of detail for mobile cache management. In *Second International Workshop on Web and Wireless Geographical Information Systems*, 12 2002.
[10] X. Zhou, J. Yates, and G. Chen. Searching the web using a map. In *Internation Conference on Web Information Systems Engineering.*, volume Vol. 1, pages 117–124, 2000.
[11] Google. http://www.google.com/.
[12] Google web apis (beta). http://www.google.com/apis/.
[13] Ipadic. http://chasen.aist-nara.ac.jp/.
[14] Mecab: Yet another part-of-speech and morphological analyzer. http://cl.aist-nara.ac.jp/ taku-ku/software/mecab/.