携帯端末を考慮した詳細度と地域的スコープによるWEBページ分類

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

LD-1 山田 直治† 李 龍† 高倉 弘喜‡ 上林 弥彦†
Naoharu Yamada† Ryong Lee† Hiroki Takakura‡ Yahiko Kambayashi†

1. はじめに

携帯端末の高機能化により携帯端末ユーザはいつでもどこでもインターネットを介して特定の地域に関する情報を収集できるようになったが、一方で通信の遅延および切断、低性能な CPU 等の問題点を解決するために利用者が要求する地域情報を予め携帯端末に格納しておくことが重要である。また記憶領域に制限があるため利用者が要求する地域情報のみを選択して提供する必要があるが、従来のキーワード検索では以下のふたつの大きな問題がある。

- **WEB情報の記述形式を考慮していない**
  WEB ページは作者の意図によってさまざまな記述形式、例えばトピックに関する地名をリストで羅列したページ、地域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こ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詳細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おおまかな理解ができる要約情報を要求する。
- **WEB情報の地域性を考慮していない**
  従来のキーワード検索では位置情報を持つ名詞として地名を扱っていないため、空間的な広がりをもつ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、左京区に空間的に含まれる京都大学や銀閣寺に関する WEB ページを収集することができない。

本稿では携帯端末利用者に対して利用者の要求に応じた適切な地域情報をキャッシュに格納するために、詳細度と地域スコープという二つの尺度を用いて WEB ページを分類する。詳細度は記述形式を特定するための尺度であり、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。地域的スコープは WEB ページが着目する地域を特定する尺度であり、MBR（Minimum Bounding Rectangle）を利用して測定する。ただしここでは WEB ページに出現するすべての地名を使うのではなく、不要な地名を除去することで地域的スコープの適合率を高める。

2. 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。ここでは記述形式によってＷＥＢページを以下の３つのタイプに分類する。

- **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」などがある。このタイプ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際に有用である。
- **要約型** 特定の地域に関して数行程度で簡潔に記述したページであり、歴史的建築物の案内ページなどがこれにあたる。このタイプは特定の地域についておおまかな情報を得る場合に有用である。
- **詳述型** 特定の地域に関して詳しく記述したページであり、訪問地域の感想や歴史的建築物の解説がこれにあたる。このタイプは特定の地域についてより詳しい情報を得る際に有用である。

これらのタイプについて品詞の出現度数、ページ表記、出現する地名の種類における特徴を表１にまとめる。この特徴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、決定木を用いてページ分類ルールを作成する。

	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
地名の種類	多	少	多

表１ 各タイプの特徴

3. 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(図１)、Guttman によって提案された[1]。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

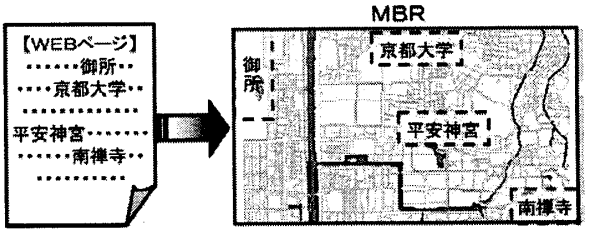

図１ WEBページのMBR

† 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

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場合、ページ内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、その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、その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 タグによって強調されている
- 単一ページに複数回出現する
- 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。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[2]、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出現数を測定する。以上３つの値を元に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。

4. 実験

京都に関するＷＥＢページを収集し、前節で述べた２つの尺度を用いてＷＥＢページを分類する実験を行った。前処理としてＷＥＢページのテキスト部分に対して茶筅を用いて形態素解析を行った。そしてページ内に出現する京都の地名と品詞毎の出現頻度、HTML タグの出現頻度を収集した。京都の地名をひとつ以上含むページを測定対象としたところ有効ページは 29784 ページであった。

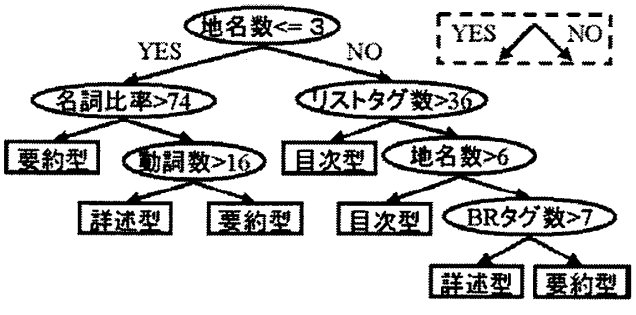

図 2　決定木によるページ分類ルール

4.1. 詳細度

無作為に 100 ページ抽出して著者の判断で３つのタイプに分類し、それを元に決定木エンジン C5.0 によって WEB ページ分類ルールを作成した（図２）。そしてこれに従って WEB ページを分類した。各タイプの特徴的なページと３０ページずつ筆者が参照して求めた適合率を以下に示す。

- **目次型**　適合率 67%。「最近行ったおいしい店」などあるトピックに基づいて地名がリストアップされているページが見つかった。一方リストタグが３６を超えるとすべて目次型と判断されるため、旅日記などリストタグも多く使っている詳述型ページが目次型と判断されてしまった。
- **要約型**　適合率 83%。特定の企業のサイトやそれらの企業に関するニュースなどが見つかった。一方登録されていない地名が多く出現する目次型ページが要約型と判断されてしまった。
- **詳述型**　適合率 73%。特定の寺社を訪れた際の感想が多く見つかった。一方リンク集は詳述型でないが、リンク数が分類ルールに明示的に含まれていないためいくつか詳述型と判断されてしまった。

4.2. 地域的スコープ

WEB ページを無作為に 100 ページ抽出し、それらに対して地名をすべて用いた場合とキーワード性の低い地名を除去した場合でそれぞれ地域的スコープを計算し、値がページの着目している地域に合致しているかどうかを調べた。ただし HTML タグによって強調されているか２回以上出現する地名がキーワードであるとした。その結果すべての地名を用いた場合には適合率が 76%であったのに対し、キーワード性の低い地名を除去した場合には適合率が 88%となった。地域的スコープが改善された主要な理由としては詳述型ページにおいて特定の地理オブジェクトに関する解説に他の地名を引用することが多く、それらの地名を除去することができたことが挙げられる。

5. おわりに

本稿では利用者の要求に応じて適切なＷＥＢページを収集するために詳細度と地域的スコープという２つの尺度を提案した。特定の地域に関する情報要求はその地域に対する利用者の興味に強く依存している。そこで利用者が興味を持っている地域とその強さを特定し、適した地域情報を二つの尺度を用いて収集しキャッシュに格納する必要がある。例えば訪問予定地域は利用者の興味が強く、訪問終了地域は興味が弱い。これらの興味に応じて、興味の強い地域に関しては詳述型ページをキャッシュに格納し、興味の弱い地域に関しては詳述型ページをキャッシュから除去するといった方法で効率的なキャッシュ管理が可能である。

参考文献

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

[2] 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.