情報分布を考慮した外部リソースの位置指向情報検索*

4L－4

三浦 信幸　高橋 克巳　横路 誠司　島 健一

NTT ソフトウェア研究所†

E-mail: {miura, takahasi, yokoji, kshima}@slab.ntt.co.jp

1 はじめに

昨今，電話帳，地図，店舗情報といったデータベースがWWW等のオープンな環境で利用可能になり，これらの検索インタフェースの異種性を解消する方法として，メタサーチやデータベースwrapper[1]などがある．我々も，ある特定の場所に関する情報を検索する場面，すなわち，位置指向の情報検索の場面を想定した，モーバイルインフォサーチ(MIS)[2]という実験システム‡を構築している．このシステムでは，場所を表現する情報(位置情報)の様々な表現形態(緯度経度・住所・最寄駅名等)を相互に変換し，複数の外部リソースへの検索インタフェースの統一化をはかっている(図1)．ユーザが緯度経度・住所・最寄駅名・郵便番号・最寄ランドマーク名などの中から任意の種類のもので場所を特定し，どのWWWサイトで検索を行いたいかの要求と組み合わせて検索要求を出すと，MISサーバは検索先サイトに合わせて位置情報を変換し，検索式を生成してユーザの検索要求の中継を行う．

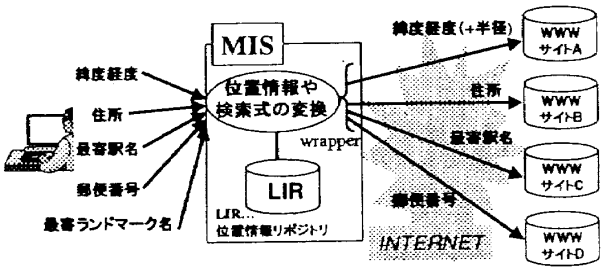

図 1: MIS でのデータベース wrapper

このような外部リソースの位置指向情報検索においては，適度な量の解が得られるよう，検索範囲を適切に与えることが必要である．本稿ではそのために複数の外部リソースに対して使用可能な一般的な位置指向の情報分布というものを算出し，それを考慮して検索範囲を決定する方法を検討する．

2 外部リソースの位置指向の情報検索における問題

位置指向の情報検索では，検索範囲を陽に与えることになる場合が多く，また，その範囲の与え方が検索結果の解の数や検索の応答時間等に影響する．中心点の緯度経度と検索範囲の半径という組はその典型的な例であるし，住所を検索条件として与える場合でも「東京都」,「東京都中央区」,「東京都中央区銀座」,「東京都中央区銀座3丁目」といった具合におのずと検索範囲を含めることになる．使い勝手を良くするためには，検索結果数が適度に得られるようにして，検索結果が0件の場合の検索再実行のコストや検索結果があまりに多い場合に検索結果の中から求める情報をユーザが取捨選択するコストを低減する必要があり，このような検索範囲を適切に与える必要がある．この時，検索する対象のデータベースの内容全体がすべて見えるような内部リソースであれば，検索範囲の自動調整は比較的容易であると考えられるが，どのような検索範囲を与えればどれくらいの個数の解が得られるのかが経験的にしかわからないような外部リソースの場合は必ずしも容易ではない．

表1は，MISにおいてwrapperをかけているサイトの中から著名な3サイトを選び，3箇所の住所について指定する住所の深さを変えて検索を行った時の検索結果の解の数である．ここでは仮に，検索結果数が100件程度が望ましいと仮定して，それを満たす検索条件の欄に○印を振った．銀座の例では，サイトAは深さ4以上，サイトBについては深さ4，サイトCについては深さ3で指定するのが望ましいということになる．しかし，このような深さの扱いは当然ながら場所依存であり，三芳町の例では，サイトAは深さ3，サイトBは深さ1，サイトCは深さ2，梅田の例ではサイトAは深さ4，サイトBは深さ3，サイトCは深さ2で指定するのが望ましい．したがって，外部リソースの位置指向情報検索においては，検索対象の場所というパラメタと検索対象データベースの特性というパラメタの2つを考慮して適切な検索範囲を決める必要がある．

表 1: 住所の深さと検索結果数

深さ	住所	サイトA	サイトB	サイトC
1	東京都	232536	5278	N/A
2	東京都中央区	17746	258	N/A
3	東京都中央区銀座	3326	150	○ 98
4	東京都中央区銀座3丁目	○ 441	○ 22	14
1	埼玉県	54695	○ 47	N/A
1.5	埼玉県入間郡	535	4	○ 1
2	埼玉県入間郡三芳町	133	1	0
3	埼玉県入間郡三芳町みよし台	○ 4	0	0
1	大阪府	82824	1709	N/A
1.5	大阪府大阪市	10095	1662	N/A
2	大阪府大阪市北区	1730	290	○ 89
3	大阪府大阪市北区梅田	324	○ 37	11
4	大阪府大阪市北区梅田1丁目	○ 39	24	5

※深さ…国土地理協会全国町字ファイルの階層付けに従った

※ N/A…検索結果数過大のため検索実行が拒否された

このような問題に対して，各データベースが検索条件とその検索結果数を外部に公開していることを前提にして解の個数を予想するアプローチ[3]やそれらを外部から参照できる仕組みの普及を行うというアプローチもあろうが，本稿では既存のデータベースには手を加えず，検索を中継する側のみで対処する方法を考える．

3 位置指向の情報分布とその構築

2章で検討したように，外部リソースの位置指向検索における検索範囲の決定に関する問題では，場所依存のパラメタと検索外部リソース依存のパラメタの2種類が存在する．本章では，前者の問題について検討する．後者については次章で検討する．

場所依存のパラメタに関しては，一般的な位置指向の情報分布というものを算出し，それをもとに推測を行う方法を提案する．

まず，位置指向の情報分布とは，次のような入出力を持つデータベースである．

入力：位置(住所や緯度経度など)

出力：該当する情報の個数

---

* Location-oriented Information Retrival of External Resource Considering Distribution of Information

† Nobuyuki Miura, Katsumi Takahashi, Seiji Yokoji, Ken'ichi Shima, NTT Software Laboratroies

‡ http://www.kokono.net/
現在，第二フェーズのMIS2実験を公開実験中

このようなデータベースの構築は，なるべく網羅的で平均的なデータベースを用いることが望ましい．例えば，我々のMISで利用することを想定した場合には，店舗情報等の検索が主であるため，NTTの職業別電話帳や通産省の商業統計地域情報等を利用することが考えられる．

このデータベースを基に，対象とする外部リソース毎に検索結果数の予測を行う．この方法は，「対象とする外部リソースが含む情報の分野と情報分布データベースを作成する際に用いたデータに含まれる情報の分野とをある程度揃えれば，情報分布データベースの情報分布と対象とする外部リソース中の情報分布には正の相関があると考えて良い．」という仮定に基づいている．表2は，表1の例について，職業別電話帳の職業分布との対比でこの仮定の検証を行ったものであり，推定相関係数から正の相関があると判断できる．

表2: 職業別電話帳の情報分布との相関係数

	サイトA	サイトB	サイトC
サンプルでの相関係数 (サンプル数)	0.946 (13)	0.926 (13)	0.812 (8)
推定相関係数 (信頼度95%)	0.85〜0.98	0.77〜0.98	0.65〜0.91

まず，検索結果数の予測の前処理として，サンプリングした，いくつかの箇所についてのみ，対象外部リソースに対して検索を実行しておき，検索結果の解の数を保持する．さらに，これらの箇所における，対象外部リソースの解の数と情報分布データの個数の比の平均を算出し，これを対象外部リソース係数とする．

この前処理を基に，検索範囲の決定は次のように行う．まず，指定された場所の情報分布と対象外部リソース係数の積で対象外部リソースのその場所における検索結果数を予測する．検索結果数の予測を基に，検索範囲を決定する．一般に，検索範囲の指定の仕方は離散的な値を取ることが多い．2章の住所の例では4〜5段階であるし，中心座標と半径が指定できるようなサイトであっても半径についてはある程度限られた離散値しか取れない場合が多い．離散的な検索範囲の候補の中からひとつを選び出す際には，検索結果数が多過ぎないようにすることと少な過ぎないようにすることとのトレードオフが存在する．我々が現在適用しようとしているMISでは，少な過ぎないことを重視する．そこで，検索結果数の予測値が目標値を下回っていない検索候補のうち，検索範囲がもっとも狭いものを検索範囲として採用する．

4 対象外部リソース毎の特性への対応

本章では，3章の冒頭であげた二つのパラメタのうちの後者，すなわち，検索対象外部リソース依存のパラメタについて検討する．

外部リソースが職業別電話帳のように様々なジャンルについて平均的にデータが整備されている場合には3章のような方法で予測できるが，保持しているデータのジャンルについて何らかの特色を持っているような場合には，3章の方法では誤差が大きい可能性があるかもしれない．例えば，パソコンショップについての情報に強い外部リソースの場合には，東京秋葉原，大阪日本橋，東京新宿での情報分布は他の場所での一般的な情報分布と大きく異なることが予想される．このようなジャンルに応じた外部リソース依存性については，3章で構築した情報分布データの場合，業種毎といった情報の種類毎の情報分布を考慮することである程度解消できると考える．

3章で構築した情報分布データは全業種合計の分布であるので，まず，業種毎の情報分布データベースを作成する．個々の業種毎の分布データベースについて，3章同様，サンプリングした，いくつかの箇所について対象外部リソースに対して検索を実行し，リソース係数を求める．この係数が3章で求めた全業種平均と大きく離れている場合に，対象外部リソースはこの業種に関するリソースだと判断して，この外部リソースについては全業種平均での分布データ・リソース係数を用いるのではなく，この業種の分布データ・リソース係数を用いることにする．

5 試行実験

3章で考えた検索範囲決定法に従って，表1でとりあげた3つのサイトに対して試行実験を行った．ここでは次のような評価条件を採用した．

- 評価条件
  検索結果数が1件以上あり，かつ，なるべく100件を越えないこと．例えば，検索結果が，30000,3000,300,30,3となるような5種類の検索範囲があれば，検索結果数が30になる場合と3になる場合が望ましい検索条件とする．同様に，30000,3000,300,0,0となる場合がある時には，検索結果数が300となる場合を望ましい検索条件とする．

まず，住所を日本中からランダムに100箇所抽出し，深さを変えながら，各サイトに対して検索を行い，検索結果数を調べ，評価条件に合致する住所の深さを調べた．

一方，職業別電話帳のデータを用いて情報分布データベースを作成し，表1のサンプルをその情報分布と比較し，各サイト毎にリソース係数の平均を算出し，3章で考えた検索範囲決定法に従い，検索結果数の予測を行って，100箇所について検索範囲を決定した．

これらを基に，従来法に相当する深さ固定の検索範囲決定法を用いた場合，提案手法を用いた場合のそれぞれについて，評価条件を満たす検索結果が得られた割合を調べたのが表3である．例えば，深さ1固定，つまり常に住所は都道府県名までのみの住所を検索条件とした場合，サイトAでは3%，サイトBでは57%，サイトCでは39%であった．

なお，サイトCはそもそも100件を越えると検索結果数過大で検索実行を拒否するような挙動を示し，また，検索結果数も常にあまり多くは得られないため，サイトCについてのみ評価条件中の上限値を100件ではなく，20件に設定した．

表3から，提案手法を用いて情報分布を考慮して検索範囲を決定した方が有効であると言える．

表3: 評価条件を満たした検索の割合

	サイトA	サイトB	サイトC
深さ1固定	3 %	57 %	39 %
深さ1.5固定	4 %	55 %	2 %
深さ2固定	62 %	57 %	55 %
深さ3固定	29 %	7 %	7 %
深さ4固定	29 %	1 %	1 %
提案手法	64 %	63 %	76 %

6 おわりに

本稿では，検索条件と検索結果数の対の情報が外部からは参照不可能な位置指向の外部リソースに対して，一般的な位置指向の情報分布を用いて検索結果数が適切な範囲になるように検索範囲を決定する方法を提案した．

今後は，提案手法に加えて，リソース係数決定のためのサンプル位置の抽出方法とリソース係数の位置依存性の取扱，サイト依存性への対応のための業種毎分布の考慮の具体的手法などの問題をさらに検討していく．

最後になりましたが，MIS2実験にあたって，御協力下さっている，ユーザの皆様ならびに，関係各位に深く感謝致します．

参考文献

[1] "The Standard-IBM Manager of Multiple Information Sources (TSIMMIS)". http://www-db.stanford.edu/tsimmis/.

[2] 三浦ほか. "モーバイルインフォサーチ: 移動環境下でのユーザ指向型WWW検索". 第3回 モーバイルコンピューティング研究会. 情報処理学会. Dec. 1997.

[3] 須藤昌徳. 横山和俊. 井上潮. 木谷強. "分散環境における情報検索を支援するデータベース選択方式". データベース研究会. pp. 9-16. 情報処理学会. May 1997.