特定分野のリソース収集を行うWWWロボットの性能評価*

4L－3

横路誠司† 三浦信幸‡ 高橋克己§ 島健一¶
NTTソフトウェア研究所‖

1 はじめに

情報検索に必要なリソースの収集のために、多くの情報収集ロボット（以下ロボットと呼ぶ）がインターネット上で動作している。これらのロボットは、情報収集先サーバの負荷分散および、リソース提供者の意志[1]等を考慮して、リソースの収集方法を決定しているが、リソースの分野を限定して、収集する手法は、まだ確立されていない。

本稿では、内容に応じて選択的にリソース収集を行う手法[2]を実装したロボットについて、性能評価を行った結果について述べる。

2 情報検索システムとロボット

近年のWWW（World Wide Web）の急速な発達により、インターネット上に提供される情報量は膨大になっている。このような状況下では、ユーザがインターネット上で必要なリソースを見つけ出すことは、非常に困難である。これらの問題に対処するために、情報検索の分野で様々な研究が行われ、情報検索システムが実装、公開および利用されている。

特に、ロボット型情報検索システムと呼ばれる、リソースの収集およびキーワード抽出を自動的に行う情報検索システムは、提示できるリソースが大容量であることと、リソースの更新に素早く追従できる利点があるため、活発に研究が行われている。現在は、広範囲のリ

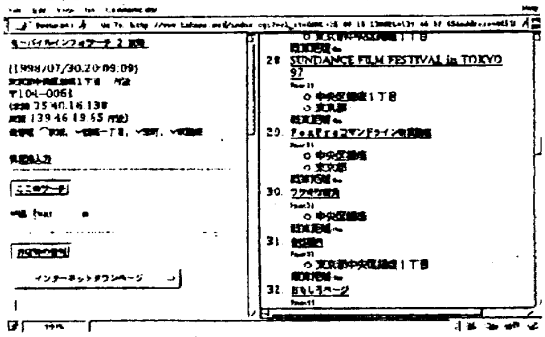

図 1: MISの使用例

ソースを網羅的に検索出来るロボット型検索システムが主流であるが、その検索の精度は、一般的にあまり良くない。そこで、我々はロボット型検索システムの新しい形態として、分野を特定した精度の高い検索システムを提案する。今回我々は、分野を限定した検索システムとして、携帯端末からの有用性が高い、位置に関連した情報の検索に分野を限定した、位置情報指向の検索システムMIS（Mobile Info Search）[3][4](図1)を構築した。

位置に関連した情報とは、住所、ランドマーク（駅名や目印となる建造物名）、店舗名、電話番号、郵便番号、緯度経度等の「位置情報」を使用して、位置の特定が可能な情報である。

MISでは、位置に関連した情報以外は必要としないため、リソースの内容を判断しつつ、必要なリソースを収集するロボットLIG（Location Information Gatherer）を使用して、効率的なリソース収集を行っている。

次章以降では、LIGの、位置情報抽出手法とリソースの収集優先度決定法を紹介した後に、リソース収集実験の結果およびその評価について述べる。

3 LIGにおけるリソース収集手法

図2に示すような流れに従って、LIGは、位置に関連するリソースを収集している。

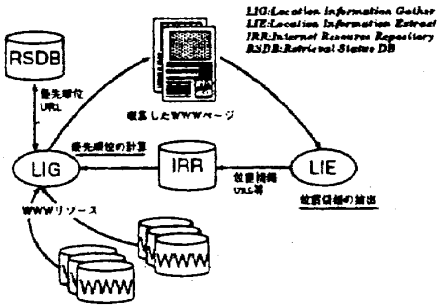

図 2: LIGにおけるリソース収集の流れ

- 収集したリソースをLIE(Location Information Extractor)に渡し、LIEは、位置情報の抽出を行い、その結果をIRRへ格納する。
- IRRの情報を基に優先度を計算した結果を、RSDBに格納し、優先度に基づいて収集を行う。

現在、位置情報としてLIEが抽出するものは、住所、郵便番号、電話番号である。住所の抽出は住所辞書を用いて、郵便番号および電話番号の抽出は、パターンマッチを使用して行っている。

一方、収集優先度としては、次の2つのものが考えられる。

- 未収集リソースに対する優先度
  未収集のリンク先のリソースが、位置に関連した情報か否かを予測するためのもの。
- 収集済のリソースに対する優先度
  収集したリソースが、再収集すべきものか否かを判断するためのもの。

今回の実験では、未収収集のリソースに対する優先度のみを使用した。未収集のリソースに対する収集優先度については、前回の実験[2]により、収集済のリソースのリンク中の位置情報を調べることにより、リンク先が位置に関連するリソースであるか否かを、予測可能なことがわかっているので、[2]の手法に従って基本的に次のように決定する。

1. リンク中に位置情報を含むものの優先度を上げる。
2. リンク中には位置情報を含まないが、リンクの抽出元に、位置情報を含むリンクが存在する場合には、優先度を変更しない。

---

*Performance evaluation of WWW robot that dynamically changes its retrival strategy using contents of WWW resources.

† Seiji Yokoji , yokoji@slab.ntt.co.jp
‡ Nobuyuki Miura , miura@slab.ntt.co.jp
§ Katsumi Takahashi , takahasi@slab.ntt.co.jp
¶ Ken-ichi Shima , kshima@slab.ntt.co.jp
‖ NTT Software Laboratories.

3. それ以外の場合は、優先度を下げる。

項目1の場合には、更に細かく優先度を決めるために、位置情報の精度を使用する。以上から、未収集のリソースに対する優先度を次のように定義する。

$$Priority = \begin{cases} \frac{\sum_{k=1}^{n} prec_k}{n}\alpha & (1の場合) \\ 0 & (2の場合) \\ -\alpha & (3の場合) \end{cases}$$

ここでは、位置情報の精度は、住所の階層の深さであると定義し、nはリンク中の位置情報の数、$prec_k$は位置情報の精度を表す。αは位置情報の精度を考慮しない場合の、優先度の上げ幅である。

4 実験

はじめに、3章で示した、優先度の妥当性を調べるために予備実験を行った。予備実験では、優先度別に300ページを収集して、ページの評価を行った。ただし、優先度が75のものは、300ページに満たないため、24ページ収集し、優先度が100のものは発見できなかったため、実験を行っていない。

収集したリソースの評価は、リソース中の位置情報の精度を用いて行った。但し、FORM中の位置情報[1]は、リソースの内容には反映されていないので、除いた。残った文字列から位置情報を抽出し、各位置情報の精度を求め、その最大値をリソースの得点とした。尚、今回の予備実験および実験では、αを100とし、表1の様な、住所階層の深さに対する精度の値を用いた。

表 1: 住所階層の深さと精度の値

住所	精度
東京都	0.25
東京都武蔵野市	0.50
東京都武蔵野市緑町	0.75
東京都武蔵野市緑町1丁目	1.00

予備実験の結果を、表2に示す。

表 2: 優先度毎の収集結果

優先度	位置に関連したリソース数	平均得点
-100	46(15.3%)	47.3
0	103(34.3%)	45.6
25	219(73.0%)	42.7
50	231(77.0%)	55.5
75	16/24(66.7%)	70.3

表2でわかるように、位置に関連したリソースの割合では、優先度の符号による差異がみられる。一方、平均得点においては、優先度が25以下では、あまり変化がないが、50以上では、リソースの得点は高くなっている。この予備実験の結果により、未収集のリソースに対する優先度を用いた収集法が妥当であることが、再確認できた。

そこで、LIGの特定分野リソースの収集性能を調べるために、幅優先探索[2]ロボットとの比較実験を行った。比較実験の手順は次の通りである。

1. LIG, 幅優先探索ロボットともに、同じスタートポイントから10,000ページを収集する。
2. 双方が収集したリソース中の位置情報に関する評価を行い、得点をつける。

実験結果を以下に示す。

[1]例えば、< select > < option value = ''hokkaido.html'' > 北海道 < / select > の下線部

[2]ロボットの収集戦略の一つで、情報提供サーバに対する負荷を減らすために、出来るだけ異なるサーバからリソースを収集する手法

表 3: 位置に関連したリソースの収集性能

	位置に関連したリソース数	平均得点
LIG	3749(37.5%)	52.0
幅優先探索	1480(14.8%)	51.6

表3を見ても分かる通り、LIGは、幅優先探索ロボットの2.35倍の確率で、位置に関連したリソースを収集した。更に、図3を見てもわかるように、LIGは、収集リソース数が数100ページ程度と少ない時点から安定して、35 ~ 40%程度の位置に関連したリソースを収集している。このことから、収集するリソース数が増大しても、位置に関連するリソースの収集効率は、落ちないと予想される。

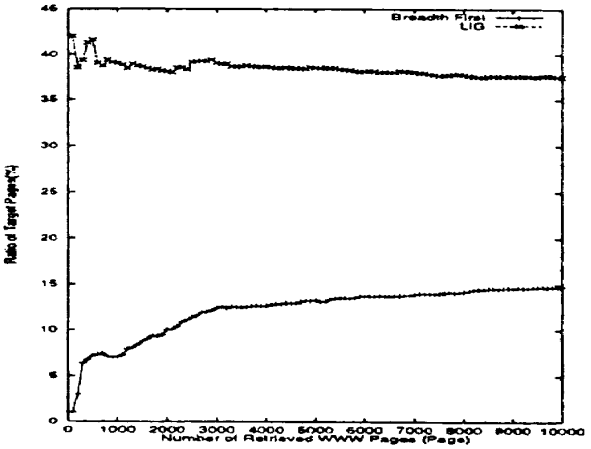

図 3: 位置に関連したページの割合の推移

5 まとめ

特定分野のリソース収集を行うロボットについて性能の評価を行った。[2]による、優先度付き収集法を実装したロボットは、幅優先探索ロボットの倍以上の高確率で位置に関連するリソースを収集した。また、収集リソース数にあまり依存せずに、安定した効率で位置に関連するリソースを収集した。一方、優先度と位置に関連するリソースの含有率および、平均得点には、予想した程の関連がなかったので、優先度の決定には、まだ改善すべき点があることがわかった。

今回の実験で、未収集のリソースに対する優先度が、特定分野のリソース収集に、有効なことを確認出来たので、今後は、収集済のリソースの優先度の追加、および優先度決定手法の改良を行いたいと考えている。

日頃から貴重な意見を頂いている、ソフトウェア研究所知的ソフトウェア研究グループの皆様に深く感謝致します。

参考文献

[1] Martijn Koster: A Method for Web Robots Control, *Internet Draft* , Dec. 1996., http://info.webcrawler.com/mak/projects/robots/norobots-rfc.html

[2] 横路、高橋、鷲坂、三浦、島: 情報内容を考慮した情報収集方法, 情報処理学会 第56回 全国大会, Mar. 1998.

[3] 三浦、高橋、横路、島: 位置指向の情報統合 - モーバイルインフォサーチ2実験 - , 情報処理学会 第57回 全国大会,Oct. 1998.

[4] 高橋、三浦、坂本、島: 位置指向の情報統合, *Japan W3 Conf. '97* , 日本インターネット協会, 1997. , http://www.kokono.net/w3c/