情報内容を考慮した情報収集方法*

2Y－4

横路誠司 高橋克己 鷲坂光一 三浦信幸 島健一†
NTTソフトウェア研究所‡
{yokoji,takahasi,wasisaka,miura,kshima}@slab.ntt.co.jp

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）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等を考慮して、リソースの収集方法を決定しているが、収集したリソースの内容に応じて、動的に収集方法を変化させる方法はまだ確立されていない。

本稿では、ロボットが、自ら収集したリソースから収集目的とするリソースを予測および学習することで、必要なリソースの選択的な収集を実現する方法を、提案する。

2 情報検索システムとロボット

最近の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。この様な状況下では、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。

情報検索システムは、ディレクトリ型とロボット型に大きく分類される。ディレクトリ型検索システムでは、通常WWWページの収集および分類は人手で行うため、正確できめ細かな分類が出来るが、提示できるリソース数が少ない。一方、ロボット型検索システムではリソースの収集及び特徴抽出を自動的に行うので、広範囲の情報を提示できるが、リソースの内容に関係なく収集を行うので、情報検索システムの性質によっては、そのサービスにとって、不要な情報も多量に収集してしまう可能性がある。

現在我々は、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[1]を構築中である。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。リソースの収集には収集したリソース中の位置情報を使用した選択的な収集方法を検討中である。

以下では、位置情報検索システムで用いられているロボットの情報収集方法について述べる。

3 内容を考慮した収集方法

本章では、我々の提案する位置情報検索システムにおけるWWWページの内容を考慮したリソース収集方法を説明する。

3.1 情報収集アルゴリズム

リソース収集は基本的に、以下に示すリソースの内容を考慮した、WWWページへの重み付けによる方法で行う。ただし、重み（収集の優先順位）が同じであった場合には、幅優先探索[1]を使用して行う。

位置情報を含むリソースを選択的に収集するためには、既に収集したWWWページ中に含まれるハイパーリンク(アンカー)から収集すべきものを予測および選択する必要がある。そこで次のような実験を行った。

- サンプルページ（WWW上からランダムに選択した20ページ）に含まれる位置情報を抽出し、位置情報の現れ方により分類する。
- サンプルページ中に含まれるアンカーについて、アンカーラベル中の位置情報の有無による分類を行い、更に、リンク先ページの位置情報の有無を調べる。
- 各分類毎に、位置情報を含むWWWページの割合（＝位置情報含有率）を求める。

実験結果を図1に示す。図1中の上段の数字の分母は、

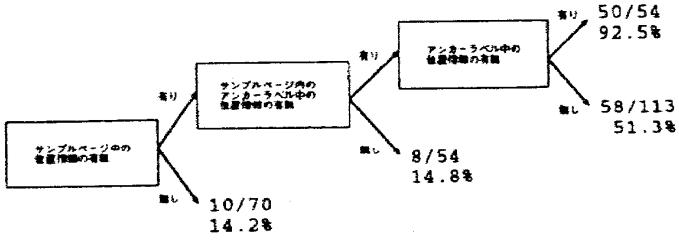

図 1: 位置情報の現れ方によるリンク先ページの位置情報含有率

条件にあてはまる全アンカー数を示し、分子はそのうちリンク先のページが位置情報を含むものの数を示している。下段の数値は、位置情報含有率である。

実験の結果、アンカーラベルに着目することで、WWWページを収集せずに内容の推測が可能であることが分かった。この実験を基に、以下に示す、未収集のWWWページに対する重み付けのアルゴリズムを提案する。

- アンカーラベルを形態素解析[2]し、地名やランドマークの情報を含むものの重みを大きくする。
- ただし、抽出元のページに位置情報を含むアンカーラベルがある場合は、そのページに含まれる未収集のWWWページに対しする重みは変更しない。どちらにもあてはまらない場合は、WWWページの重みを下げる。

*Information retrieving method on the Internet by weighting to the WWW pages.

†Seiji Yokoji, Katsuumi Takahasi, Mitsukazu Washisaka, Nobuyuki Miura, Ken-ichi Shima
‡NTT Software Laboratories.

[1]実際の収集では、この他に、WWWサーバへのアクセス間隔の管理、テキストデータ以外のWWWページの排除、robots.txt[4]への対応も行う必要がある。

[2]形態素解析エンジンには「すもも」[3]を使用する。

更に正確な収集を行うために、位置情報検索システム用ロボットでは、既に収集したWWWページに対する重み付けも行う。この重み付けは2回目以降の収集に有効である。この場合、WWWページは既に収集済みで、位置情報抽出には、WWWページの内容を使用できるため、上記の抽出に加え、電話番号や郵便番号等の抽出を行うことで、より正確な位置情報抽出を行う。これらの重み付けの様子を図2に示す。

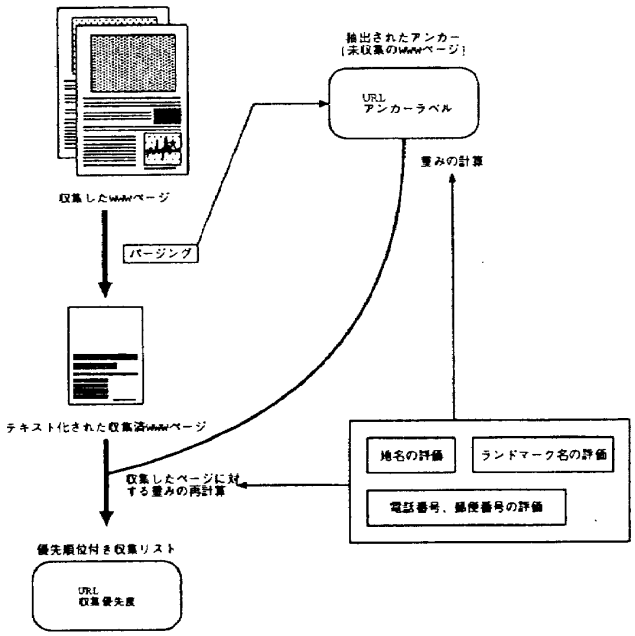

図 2: WWWページに対する重み付け

3.2 リソース内容の評価方法

地名等の位置情報をWWWページ中から正確に抽出するためには、WWWページを形態素解析し、品詞情報を利用する方法が有効である。そのため、内容の評価は、主に形態素解析を使用して行うが、地名やランドマーク名には固有名詞が多く使用されており、形態素解析に使用される辞書は全てを網羅しきれていない。そこで、地名やランドマークに関する固有名詞を辞書に登録して形態素解析をおこなう。

また、位置情報として有用な電話番号、郵便番号は形態素解析では抽出できないため、表1のようなパターンによる抽出を行う。表1に示したパターンは、実際にWWWページを見て得られたものである。表1には、WWWページ中からの電話番号抽出実験の結果も示しており、各パターンの出現回数は表1通りである。パターンの出現文書数は、1369ページ[10.3%]で、このことからも電話番号等パターンマッチによる情報抽出が有効なことが分かる。実験には、YAHOO Japan,NTTディレクトリ,Infoseek Japanを始点URLとした13313ページ(幅優先探索のみを使用して収集したもの)を用いている。

4 おわりに

位置情報検索システムにおける内容を考慮したインターネットからの情報収集法について述べた。インターネット上にあるWWWページ中に含まれる地名やランドマーク、電話番号及び郵便番号等の位置情報によるWWWページへの重み付けを行うことによって、選択的な情報収集を実現する方法を提案した。今回述べた項目以外にも、URLの構造による重み付け(位置情報を含む文書が多いディレクトリの優先順位を上げる)や文書に含まれるテキストで書かれた地図、相対的な位置(例:東京西部→多摩)の抽出等による重み付けを検討している。

また、今回はリソース収集の速度については述べなかったが、多くのロボットでは、WWWページを収集している間に多くのアイドル時間が存在するため、今回示した手法をロボット内で使用してもわずかな負荷しかかからず、収集の目的とするリソースの収集速度は向上すると思われる。

最後に、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。

表 1: 電話番号の抽出結果(総ページ数:13313)

マッチパターン	出現回数
0x-xxxx-xxxx	3522
0x[xxxx]xxxx	145
[0x]xxxx-xxxx	52
0x,xxxx,xxxx	126
0x(xxxx)xxxx	709
(0x)xxxx-xxxx	258
0x－xxxx－xxxx	47

参考文献

[1] 高橋、三浦、坂本、島:位置指向の情報統合,Japan W3 Conf. '97 オンラインプロシーディングス,http://www.kokono.net/w3c/

[2] 岩爪,白神,畑谷,武田,西田:オントロジーに基づく広域ネットワークからの情報収集・分類・統合化,情報処理学会論文誌, Vol.38,No.3,pp606-615

[3] 鷲坂,山崎,広津,尾内:情報検索のための高速日本語形態素解析システム「すもも」,情報処理,Vol38,No.9,pp830-831

[4] Martijn Koster:A Standard for Robot Exclusion, http://info.webcrawler.com/mak/projects/robots/norobots.html