学位請求論文

ジオワード・マイニングを用いた ローカルサーチの研究

東京大学
大学院 情報理工学系研究科
電子情報学専攻

高橋　克巳

2005 年 12 月 16 日提出

目次

図目次

表目次

第 1 章

序論

1.1 はじめに

自分の住んでいる地域，自分の通勤通学先，週末に買い物に出かける街，これから旅行に行く土地，これらに関した「ローカルな」情報を知りたいということは古くからある情報検索のニーズである．これらに答えるメディアとしては，日本全国の地域に対して出版されている地図があり，観光地にはガイドブックなどの書籍も流通している．さらに生活地域であっても新聞のチラシや電話帳の広告，あるいは市役所が発行する生活案内の冊子等が存在し，知らず知らずのうちに我々はこれらを利用して生活している．

一方，現代の我々の多くが情報を調べる手段として最も身近に感じているメディアがインターネットである．インターネットは「何に関する情報でも探せばある」ことが一つの利点として認識されており，ローカルな情報もインターネットで利用されている．しかしながら，ポータルサービスやサーチエンジンサービスが社会的に認知される中，ローカル情報に特化したサービスは一般の情報に埋もれる形で，必ずしもインターネットの重要な構成要素として認識されて来たわけではなかった．我が国ではローカル情報はカーナビゲーションシステムや，地図 CD-ROM の低価格化による普及（1990 年代中頃),，さらには携帯電話の基地局の位置情報を使った情報サービスの開始（2001 年）などの例の通り，ローカル情報を提供するコンピュータシステムは身近な存在であった．しかしローカルサーチが広くインターネットのキラーアプリケーションとして認識されるには，2004 年に米国 google，Ask Jeeves，Yahoo!らが一斉に位置依存情報サービスをローカルという単語を使って開始したことまで待たねばならなかった．2004 年にローカルサーチが普及した背景には，利用者ニーズの変化，情報提供社者のニーズ変化，環境の変化などがあると考えられる．

- 利用者ニーズ

インターネットが本格的に普及したことにより，より日常生活に密着した情報が求められるようになった

- 情報提供者ニーズ
  インターネット広告の方法がバナー広告からキーワード広告へ移った．その結果として，ロングテールと呼ばれる多種小利用頻度の情報群の代表格である地域情報が魅力ある広告市場であることがわかり，その媒体としてのローカルサーチ開発に拍車がかかった
- 環境
  オンライン地図情報が一般に普及し，地図会社以外でも地図アプリケーションの提供が容易になった

このように発展の背景には複数の要因があり，ローカルサーチは技術，産業共にこれからさらなる発展が期待できる分野である．

ここで改めてローカルサーチとジオワードの定義をする．

ローカルサーチ（Local Search）とは，地理的な条件で情報を検索する機能を提供するシステムである．地理的な条件には，住所，駅名，ランドマーク名，郵便番号，緯度経度など地上の位置を示す文字列や値が含まれる．狭義のローカルサーチはインターネット上の情報を地理的な条件で検索することであり，クローラが収集したウェブページを地理的な条件と検索キーワードで検索することが典型的な例となる．ローカルサーチの例は前記以外にも，イエローページやレストラン検索で扱われる店舗情報の検索（タウン情報検索とも呼ばれる），地図，天気予報，乗換案内などの専門のコンテンツの検索などがある．なお，ローカルサーチの定義には含まれないが，地理的な条件で情報検索を行うことには，単なる地名キーワードの有無ではなく，暗黙に地理的な検索が行われることが期待されている．地理的な検索とは，地理位置間の遠近／包含などの関係で行われる検索で，検索条件および検索対象を図形として表現した上で，検索条件と対象間の両者の重なり，距離，方角といった関係に応じて対象を出力する検索処理のことを呼ぶ．

ローカルサーチにおいて，場所や地域といった地理位置を表現する重要な方法の一つがジオワード（geo word）である．ジオワードとは地上のある領域に関連付けられた固有名詞で，住所，地名，ランドマークなどが例になる．「山」は地表の形状を表すが，特定の領域を示さないのでジオワードではない．緯度経度は特定の位置を示すが，固有名詞ではなくジオワードではない．地理位置を緯度経度などの数値で表現することは可能であるが，人間が地理位置を表現する場合はジオワードを用いるのが一般的であり，ローカルサーチを実現するためにはジオワードの利用が不可欠である．

ジオワードにはいくつかの特徴がある．一つは，それぞれが地理位置と結びついていることである．情報検索にジオワードを用いることは，情報検索に地理位置を持ち込むこと

であるが，ジオワードの地理的な性質をシステムが正しく扱えば，従来のキーワード検索では得られなかった効果が生まれる可能性がある．ジオワードの第二の特徴は，ジオワードは人間が地上の空間に付けた名前であるので，何らかの人間の活動を反映している可能性があることである．ジオワードとそこから導かれる人間の活動の情報を解析することにより，人の行動に関する知見を効果的に得られる可能性がある．

本論文の目的は以下の２つである．

1. ジオワードを活用してローカルサーチを実現する方法を明らかにすること
2. ジオワードを活用して大量のデータから地域の特性を提示する方法を明らかにすること

本論文では，まずインターネットのウェブ文書等コンテンツ中に含まれるジオワードに着目する．ジオワードを使って文書に緯度経度情報を付与することができれば，ウェブ文書が地理的に検索可能になり，さらに地図上に表示することも可能になる．本論文の前半ではウェブ文書中からジオワードを発見して地理的な索引付けを行った上で，検索システムとして提供する方法について述べる．この技術は現在インターネットで提供されているローカルサーチの基礎をなす方法であるが，本研究はいち早くその性質に着目し，インターネットでの実証実験を通じて有用性を確認してきた技術である．

ジオワードは利用者がローカルサーチで検索を行うときにも用いられる．本論文の後半ではローカルサーチサービスのアクセスログをジオワードでマイニングする方法を定義しながら，地域に関する知見を提示する方法について述べる．この技術では，どこの地域においても地理的な情報の推薦ができること，すなわち，情報の分布が粗な地域があっても一定数的の基準を満たした地理的な相関ルールを提供することが可能になる．

ジオワード・マイニングとは大量のデータからジオワードを用いて有用な情報を発見することであるが，本論文で考えるジオワード・マイニングとは，ウェブ文書等コンテンツの中やあるいは利用者の検索ログデータ等の利用者の履歴情報に現れるジオワードを役立てるプロセス全体を対象としている．はじめにローカルサーチの実現に関して，次の貢献を報告する．

1.2 コンテンツ中のジオワードを用いた地理的検索の実現

インターネット上のウェブ文書を地理的に検索をする手法を明らかにした．

本研究は 1990 年代後半に着手したものである．インターネットの情報源としての豊かさは既に認識されており，ローカル情報も数多く存在していたが，これらを正しく検索する方法がなかった．ウェブ文書には住所等のジオワードを持つものがあり，この文字列を検索条件として全文検索システムへ入力して検索を行うことはできたが，キーワード検索

は次の理由で正しい地理的な検索といえない.

- 検索が文書中のジオワードに依存してしまう
  キーワード検索では検索条件の文字列と，文書中の文字列が一致することが検索の前提となる．しかし駅の近くの地域が住所だけでなく駅名で呼ばれることがあるように，同じ地理位置を表現するジオワードは複数存在する場合がある.
- 任意の領域を指定した検索が困難である
  検索者の探したい領域は，必ずしも単一のジオワードで表現できるとは限らない．例えば，県境上の地理位置を一つのジオワードで表現することが困難なように，ローカル情報をキーワードのみで検索することには限界がある.

このような問題は，文書に地理属性（緯度経度情報）を付与して，地理的な検索を行うことで解決できる．しかし一般にウェブ文書中には緯度経度属性は記述されていない．そこでシステム側で，自動的に緯度経度を付与する方法を考案した.

開発した手法は (1) ウェブ文書の内容を予測し位置に関連した情報を選択的に収集するクローラ，(2) ウェブ文書からジオワードを取り出し，緯度経度を付与する構造化モジュール，(3) さらに付与された緯度経度情報を使って地理的検索を行う検索モジュールを有する．この構成によって任意の地理的領域に属するウェブ文書を，任意の緯度経度を条件として検索することが可能になった．この方法を近隣住所の検索もれの問題として評価すると，従来の住所文字列を使用するキーワード検索では少なくとも 25% 存在していた検索もれを解消できることが明らかになった.

この提案は，住所辞書やイエローページといった構造化情報をレポジトリとして使用し，ウェブ文書を解析し構造化して検索を可能とするという今日のローカルサーチの基本原理を明らかにしたものである.

このことで地理的な検索が可能になったが，ここでもう一つの目標に目をむける.

1.3 アクセスログから地理特性を提示するためのジオワード・マイニング

ジオワードを含んだ大量の利用者の検索履歴から，地理特性を提示するジオワード・マイニング法を明らかにした.

ローカル情報の取得は利用者が目的を明確に持っている対象を精度よく検索するというタスクだけでは定義できない．むしろ現地に赴いたときに目に飛び込んで来たり，耳にしたりする地域の特色を受け入れることも重要な側面である．この地域に応じた情報の推薦を実現するために，利用者の履歴情報を活用する方法を報告する.

本手法は，ローカルサーチシステム利用者の検索履歴には，何らかの地理的特性が反映

されているという考えにもとづいている．すなわち，ある利用者が検索しようとしている地域に関して，過去の別な利用者からの同じ地域に対する検索履歴から知見を得ることを試みる．

まずアクセスログを緯度経度情報に対応させながら地理的に分析するジオワード・マイニング方法を定義し，地理的な相関ルールを求める方法を明らかにし，さらに求めたルールから興味深いルールを取り出す方法について考察した．この方法を実際のログデータに適用した結果，ジオワード・マイニングで得られるルールはジオワードによる検索頻度に偏りがあり，地域によっては非常に粗な状態になるため，取り出されるルールは当たり前のものであるか，あるいは何も取り出せないということがわかった．この問題に対してジオワードを頻度に応じてクラスタリングする方法を提案し，このクラスタリング法を用いてサポート値を調整しながら一般化相関ルールを求めることにより，提案手法は従来の単純な住所階層を用いる手法に比べ，多様な地域でより多くの興味深いルールの抽出が可能であることを明らかにした．

本論文は以上の 2 つの提案を軸として，背景で用いられる固有名詞処理技術の提案を加えて，ジオワード・マイニングを用いたローカルサーチの技術に関する報告を行う．

1.4 本論文の構成

本論文は以下の章より構成される．

第 1 章は序論であり，本研究の背景および目的について概観し，本論文の構成を述べている．

第 2 章は「関連研究」と題し，ローカルサーチ実現技術の問題点を指摘するとともに，データマイニングにおける空間マイニングとログマイニングの研究をまとめている．

第 3 章は「ジオワードを中心とした情報の統合」と題し，様々なコンテンツに緯度経度情報を持たせるための方法を提案している．イエローページデータに対し，ジオワードを用いてデジタル地図データを統合し緯度経度情報を持たせる手法，さらにウェブコンテンツに対し，イエローページデータを統合し同様に緯度経度情報を持たせる手法を述べている．前者は住所属性を持つコンテンツの検索結果を地図に表示する仕組みとして実システムで長く利用され有効性が実証されている．また，後者は次章で述べる多様なローカルサーチ実現の基盤技術をなすものといえる．

第 4 章は「ジオワードを用いたウェブローカルサーチの実現」と題し，前章の提案手法の有効性を明らかにすべく，ウェブ文書から抽出したジオワードを介し文書に緯度経度情報を付与することにより空間索引付けを可能とし，更に，ウェブ文書に対する空間問い合せの実現手法について詳述している．実ウェブ文書データを用いた評価実験により，地理的検索条件の領域と住所が示す領域の差異により従来の住所方式では 25% 程度の検索も

れが発生する場合があったのに対し，提案方式では当該問題を解消出来ることを明らかにした．実証システムは1998年6月から2003年までインターネット上において実験サービスとして公開され，世界に先駆け，いち早く大規模なローカルサーチの利便性を体感できる場を提供した．

第5章は「アクセスログから地理特性を提示するためのジオワード・マイニング」と題し，アクセスログ内のジオワードに注目した新しいマイニング手法を提案している．すなわち，ジオワードの地理的関係を用いることにより，空間的相関ルールをマイニングする方法を，地域に関連する単語を推薦することを目的に考察している．検索回数が少ない地域に対し，ジオワード間の地理的距離を用いジオワードのクラスタリングを行いサポート値の調整を可能とする一般化空間相関ルールのマイニング手法を提案している．9,000万件の実検索ログを用いた実験により，提案手法は従来の単純な住所階層を用いる手法に比べ，多様な地域でより多くの興味深いルールの抽出が可能であることを明らかにしている．

第6章は「ジオワード・マイニングのための固有名詞解析手法」と題し，固有名詞を処理する際に生じるあいまい性の解消法に関して，住所の省略および固有名詞の表記のゆれの問題について論じている．省略のある住所文字列の正式住所名への変換は，同名異所住所の場合は困難であるという問題があった．アクセスログを詳細に解析し，頻出ジオワード10,000の3%に達する同名異所住所問題を，セッション情報を用いた上位住所推定手法により解消可能となることを明らかにしている．更に，固有名詞の表記ゆれに関しては，ゆれの同値規則から自動的に正規化規則を作成する方法を考案し，日本人姓のかな表記で分析した結果，約9万通りのゆれの単位に分類可能であること，加えて，完全一致検索時に1検索あたり15%存在していた検索もれを，93%という高い適合率を達成しつつ解消できることを示している．

第7章は結論であり，本研究の成果と今後の課題について総括している．

第 2 章

関連研究

2.1 ローカルサーチ実現技術の研究

本論文が対象とするローカルサーチ (Local Search) とは，ウェブ文書を地理的な位置情報を条件として検索することである．ローカルサーチとは，検索システムの利用者の立場を想定した呼び方である．利用者の地元で役に立つ情報，すなわちローカルな情報を提供するというサービス上の特徴を強調した呼び方となっている．ローカルサーチは，Geographic Search Engine，Location Based Search，タウン検索などの名前でも呼ばれることがある．Gepgraphic Search Engine は検索手法を中心にした呼び方である．ローカルサーチといっても，検索者の地元の情報のみ提供されるわけではなく，任意の地理位置に関して，地理的な検索を可能とするので，この名称は正確な名称である．Location Base Search は検索条件である位置情報に焦点を当てた呼び方である．特にモバイル環境で利用者の現在位置にもとづく情報サービスは Location Based Service (LBS) と呼ばれるため，この名称は LBS に呼応した呼び方である．タウン情報検索は我が国独特の呼び名である．ローカルサーチの主な対象は市街地に存在することを考えると，対象を的確に表現した名称である．

2.1.1 ローカルサーチの構成

本論文におけるローカルサーチとは，ウェブ文書を地理的な位置情報を条件に検索することで，例えば以下の検索がある．

- 東京都にある本屋を探しなさい
- 青山道理沿いにあるレストランを探しなさい
- 東経 135 度 10 分北緯 35 度 10 分近辺の大学を探しなさい

上記のような検索サービスを実現するためには，以下の技術要素が必要である．

1. ウェブ文書の地理位置情報特定
2. 地理的検索と地理的なランキング
3. 位置情報を入力するための利用者インタフェイス
4. 情報源の入手

まずコンテンツに関しては，コンテンツとなるウェブ文書の地理位置情報の特定，すなわち緯度経度等の地理位置を定義できるデータをウェブページに付与することが必要である．地理位置に関連したウェブページは多数存在するが，一般にウェブページは地理的属性を陽には持っていない．そのため，ローカルサーチ提供者が何らかの方法で，その文書の地理位置情報を特定する必要がある．

ウェブページに地理位置が付与されると，コンテンツの地理的属性と検索条件の地理的属性の関係に応じた検索を提供するデータベースシステムが必要である．この性質を満たす検索では，コンテンツと検索条件双方の地理的属性を図形として表現して，その図形間の関係により検索を行うことが一般的に行われている．この検索を行うデータベースは地理空間データベースと呼ばれる．地理空間データベースは「図形のための特殊な保管構造，図形のコレクション，属性，属性間の関係および図形間の関係を含む空間および属性データのための保管機構」[Childs 2001] と定義されている．

さらに利用者に対しては，ローカルサーチに適した検索インタフェイスが必要である．検索インタフェイスは，少なくとも利用者が検索したい地域を表現できることが必要である．現在ローカルサーチの利用者インタフェイスとしては，住所文字列を入力させ，さらに検索目的の文字列（例えば，レストラン）を付加条件として入力させることが一般的であるが，利用者が現在地周辺の検索を行うのであれば，GPS 等の測位システムを使って検索条件を自動的に取得するなどの利便性の向上が可能である．

最後にローカルサーチを実現するためには，情報源となるデータの入手方法も一般のサーチエンジン実現とは異なることを指摘しておく．ローカルサーチを通常のサーチエンジンと同様に，インターネットで公開されているコンテンツを利用する場合も，ローカルサーチを実現するためにはなんらかの外部データが必要であることが知られている．外部データには地図データ，住所データ，イエローページデータなどがある．

2.1.2 ウェブ文書の地理位置特定技術

ローカルサーチ実現技術の中で，ウェブ文書への地理位置特定は中心かつ必須の要素である．例えば地理空間データベースはウェブのローカルサーチ以前から研究が進められ，商用のツールも存在しているが (例えば [Oracle 2005])，地理位置特定はウェブ文書を地理的に検索したいというニーズがあって初めて認識された課題である．なお，このプロセスはジオコーディングとも呼ばれている [McCurley 2001].

ローカルサーチが実現されるまで，ウェブ文書への地理位置の付与は手動で行われていた（例えば，Yahoo.com の地域カテゴリ）．位置の付与を自動的に行うためには，なんらかの情報を使う必要がある．位置付与に利用可能な情報には大きく２つある．一つはウェブ文書の内容であり，もう一つはウェブ文書が格納されるサーバの所在地である．前者は文書が東京都新宿区のことを記述していると判断できれば，その文書が東京都新宿区と関連すると決めることであり，後者はウェブ文書が設置されている場所が沖縄県であれば，沖縄県の情報が格納されており，沖縄と関連すると決めることが考え方の基礎となっている．ローカルサーチはこれらの方法を選択して検索を実現している．

GeoSearch は Columbia 大学のプロジェクトで，主要な学会で報告された最も古いローカルサーチの研究である [Buyukkokten 1999]，[Ding 2000]．彼らの研究ではウェブリソースに地理スコープ（Geograhical Scope）という考えを定義している．地理スコープとは「情報の作成者が，情報を届けたい地域」と定義されている．この考えに従うと，ピザ屋の地理スコープはピザ屋の所在地の近隣地域，また新聞「USA Today」の地理スコープは全米となる．GeoSearch の提案は，地理スコープを求めるウェブリソースに対する入りのハイパーリンクを調べ，そのハイパーリンクを張っているサイトの所在地から地理スコープを決めるという手法を提案した．ハイパーリンク元の所在地はインターネットドメイン管理データベースである whois [whois] を検索して得ている．この手法により新聞サイトのスコープが全国紙と地方紙で違いがあることを報告している．但しこの手法は，十分な数の入りリンクを必要とする欠点がある．なお GeoSearch という名称は Northen Light 社と Vicinity 社によって提供された，最古よばれている商用ローカルサーチエンジンと同じであるが直接の関係はない．

一方ウェブ文書の内容を用いる方法に関しては，McCurley が初めてまとまった報告を行った [McCurley 2001]．この論文では住所，郵便番号，電話番号，地名を利用した地理位置付与に言及している．また同時に対象のページのハイパーリンク元のページを利用する方法も述べている．郵便番号と電話番号に関しては実際に実験を行っており，収集したウェブページの 4.5% には米国の郵便番号が含まれていたとしている．ただし具体的な抽出方法への言及はなく，例えば住所に関しても，Chicago はイリノイ州だけでなく他の 27 州にも存在していて，あいまい性の解消が必要であるとのみ述べている．

続く IBM の Web-a-Where[Anitay 2004] はコンテンツに含まれる住所を用いた地理位置付与の課題と方法を初めて詳しく述べた報告である．この研究は世界中の住所（国名／州名／市名の階層データ）から約 40,000 を集めて，Gazetteer と呼ばれる住所データベースを作成し，文書への地理位置付与の実験を行っている．位置付与は以下の三つの処理を順に行うことによって実現される．

1. 住所発見処理

2. 住所間のあいまい性解消
3. 複数住所から主住所の決定

最初の住所発見処理は，Gazetteer データベースを検索して，コンテンツ中の文字列から地名候補を抜き出す．続くあいまい性解消処理はいくつかのヒューリスティックスを用いて，住所間のあいまい性を解消する（例えば同名異所の場合，最も人口の多い市を採用する）．あいまい性を解消された住所が，一つのコンテンツに複数存在した場合は，主住所の決定を行う．決定は住所間で数の多さや互いの関係を考慮して行っている．この方法により 82% のウェブコンテンツに正しく位置付与ができたとしている．ただしこの方法で，地名／非地名（人名等）のあいまい性の解消は課題であるとしている．

Markowetz らは，ドイツの情報を収集してローカルサーチシステムのプロトタイプを構築しているプロジェクトを行っている [Markowetz 2005]．ここでの位置特定手法は Web-a-Where の手法を改善したものとなっている．住所発見処理において，地名／非地名を見分ける処理を導入したこと（Mr. Dr. などの敬称があるものは人名と扱う），主住所決定はそのページにあるハイパーリンク元の属性も使うことなどの改善が行われている．さらにローカルサーチの結果のランキングは，検索語に対する文書のスコアと，地理的検索条件に対する地理的なスコアと，文書の全体におけるスコア（例えば PageRank [Brin 1998]）の和で定義する必要があるとした上で，これらの独立したスコアリングによるランキングを高速に解決する検索を実現するために，地理的検索のクエリ処理の効率化問題を論じている [Chen 2006]．

SPIRIT (Spatially-Aware Information Retrieval on the Internet) [Jones 2004] [Vaid 2005] は欧州委員会の研究プロジェクトで，ローカルサーチ全般に関してプロトタイプを作成しながら研究を行っている．このプロジェクトはローカルサーチのためのオントロジを作成して，「チューリッヒ中心から 10 キロ以内の学校」などといったクエリを扱うために，ジオワードと地理指示語（距離，位置関係，方角等）の関係を定式化している特徴がある．また [Gaihua 2005] もオントロジの観点から，地理的なオントロジを用いて地理的にクエリを拡張する方法を提案している．

続いて，我が国におけるローカルサーチの研究について述べる．

著者らのモーバイルインフォサーチプロジェクトではいくつかの地理位置特定の実験を行い，1998 年にジオワード（住所）を使う地理位置特定方法によってウェブページの地理的検索を可能にし，日本中の地理位置に関するコンテンツが緯度経度で検索可能なプロトタイプをインターネットで公開した [高橋 1998] [横路 2000]．このシステムはウェブページの全文検索機能は提供しなかったものの，後述する最初の商用ローカルサーチ GeoSearch よりも 1 年半ほど早く，本格的なシステムとしては世界で最も古いものと考えられる．この内容は 3 章，4 章で詳しく述べる．

平松らは京都デジタルシティプロジェクトの中で，独自のローカルサーチを実現している [平松 2000]．ここではウェブページを位置特定した上で，ウェブページ間に地理的なリンク（地理的ジェネリックリンク）を動的に生成し，ウェブ情報空間を拡張している．この拡張された空間では「大学の近くにあって土曜日も営業している郵便局」という問い合わせを行うことができる．

東京大学，相良らの研究は位置特定に用いるアドレスマッチング技術，およびウェブ上のローカル情報の高度な統合によって利用者の情報選択（お店選び）を強く支援するシステムの実現に特徴がある．アドレスマッチングとは不完全なものも含む住所文字列に対して，正しい住所文字列と緯度経度情報に変換する方法で，[相良 2003] では日本の住所表記体系に即した方法が提案されており，日本語特有の住所表記の多様性に対応した効率のよいマッチングを可能としている．後者は，「レストランのウワササーチ」というプロトタイプとして実現されているもので以下の特徴を持つ [Sagara 2004]．まず地理位置の特定はイエローページデータを用いて飲食店などの情報を集積している．イエローページデータを用いることで，地図表示が可能になり，さらに複数の情報源からのコンテンツが統一されるので，複数のレストランのレビューページが並列的に閲覧できる，レビューから特徴となる情報の抜粋が提示される，などの機能を有する．このシステムは著者の知る限りウェブページを地理的に統合したものの中で，最も利用者の利便性が高いものである．

筑波大学の研究 [Zhang 2005] は，ローカル情報を効率よく収集するためのクローリング手法を研究している．彼らの提案する LocalRank 値はそのウェブページと地域との関連度を表すスコアである．

京都大学の研究 [Tezuka 2006] は，地図とウェブページへのハイパーリンクと，ウェブページの地図上表示からなる現在のローカルサーチの利用者インタフェイスでは陳腐であるとして，ブログ情報を収集した上で，場所，時間，行動，目的を要素とした相関ルールを作成してその結果を地図上に表示する試みや，ウェブコンテンツを地理な関係を保ったまま連続的に表示する方法の提案などを行っている．

2.1.3 商用のローカルサーチ

インターネットでローカルサーチが一般に浸透したのは，google，Ask Jeeves，Yahoo!ら既に通常のサーチエンジンで成功している各社が，検索高度化の位置付けで相次いでローカルサーチを開始した 2004 年になってからのことである [google 2004][Ask 2004][Yahoo 2004]．これらローカルサーチの例を図 2.1 に示す．何れのシステムにも共通することは，利用者の検索したい位置のジオワードと検索の目的を表すフリーワードを入力として，検索結果のウェブページを地図に表示する機能を持つことである．

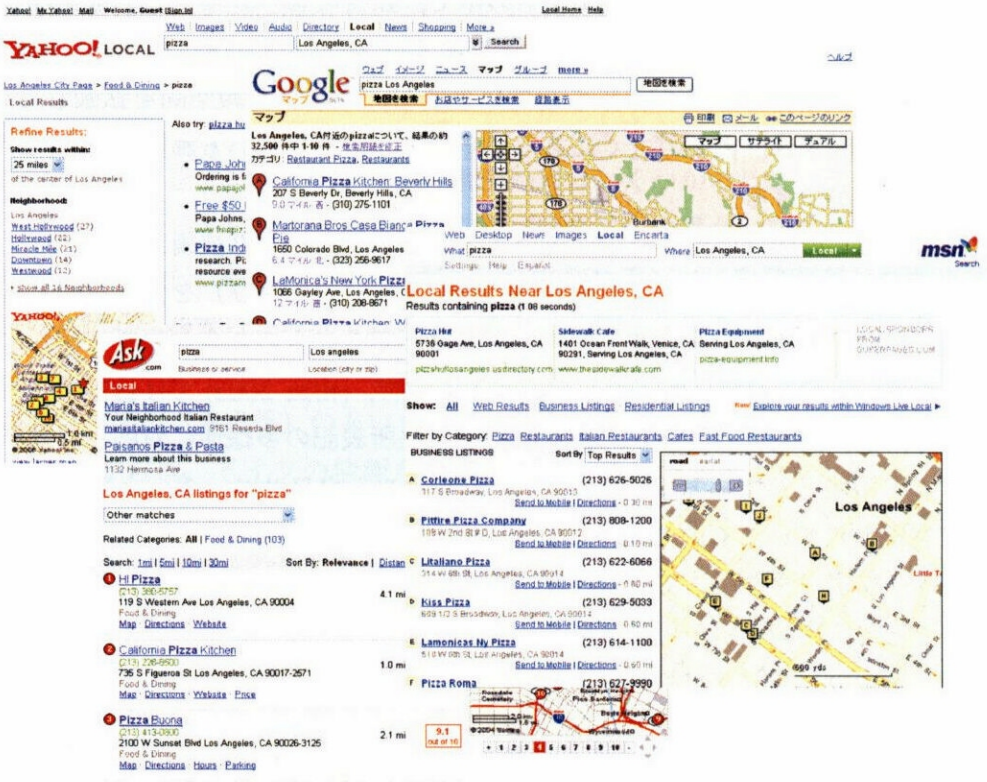

図 2.1 インターネットのローカルサーチ

商用のインターネットローカルサーチの最古として知られているものに Northern Light 社と Vicinity 社が 2000 年にサービスを開始した Geosearch がある [Geosearch 2000]. Geosearch の初期のプロトタイプは 1998 年に Vicinity 社によって開発されている [Himmelstein 2005]. 彼らは北米のウェブページの約 20% に米国／カナダの住所または電話番号が含まれていることを明らかにしている.

また Google のローカルサーチ初期の取り組みに Geographic Search [google 2002] がある. Google は 2002 年に新技術をプログラミングコンテストという名で公募したが, Geographic Search という提案に, 優勝者が与えられた. 提案はウェブページの地理位置をメッシュ単位で管理し, 大量のコンテンツであっても地理的な検索を可能としており, Google ローカルサーチのコンセプトを先行して体現していた.

これらの商用システムは, 技術詳細が論文の形では明らかになっていないが, 検索対象のウェブ文書の地理位置付与には, イエローページデータが用いられている. すなわち, 各システムではクローラで収集したウェブ文書に対してイエローページデータが参照され, イエローページデータと一致が認められたウェブ文書は店, サービス等の単位で統合

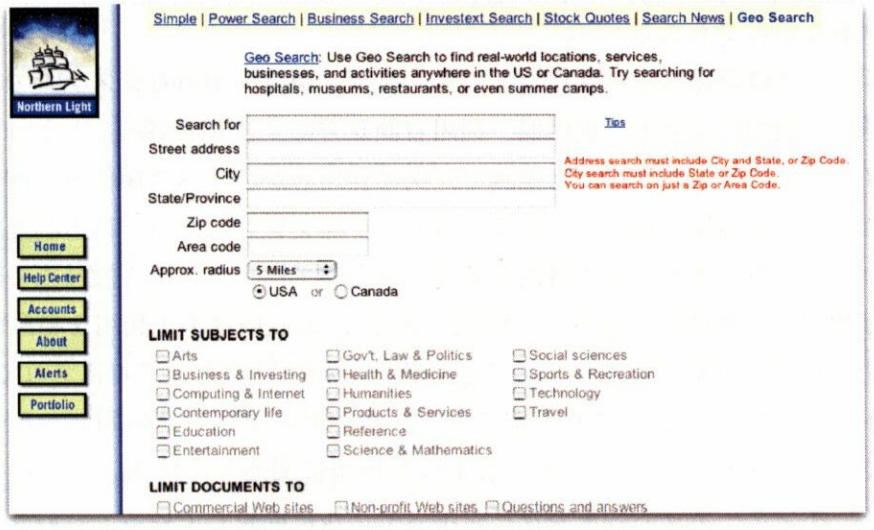

図 2.2 Northern Light の Geosearch

される．検索結果はイエローページ検索と類似しており，地図と店のリストが表示され，店を選択することにより，インターネット上の文書をさらに参照することができる．このことにより，ウェブページの地理的検索や地図表示が可能になっているが，一方で処理の単位がイエローページデータに存在するものに限られるという問題がある．

2.2 データマイニングの研究

本研究は，Agrawal らが提唱した相関ルールマイニング [Agrawal 1994] の考え方を基本とした上で．地理的な解析を行う空間マイニングと利用者の履歴データの解析を行うログマイニングに特徴を持つ．この 2 分野は独立に研究が進められてきており，それぞれの分野の技術を活かすことも本研究の重要なテーマとなりうる．

空間マイニングは地理的属性をもった事象間の関係を調べるために行われてきた分野で，気候と植生といった自然現象や，公害と疾病といった人工的な現象の因果関係を明らかにしてきている．これらの対象は測量などの手段で明確に地理的属性をもっている．しかし本論文が対象とするジオワードは地理的属性を暗に有するもので，既存の有効な空間マイニングを利用するためにジオワードと空間の関係をシステム側で定義する必要がある．そのツール等はまだ一般的ではないので，現在はインターネットのコンテンツを気軽に空間マイニングする状況にはなっていない．しかし前述のジオコーディングなどの手法がツールとして普及するにつれて，インターネットコンテンツ空間マイニングはますます盛んになると考えられる．本研究は，ジオワードに関して，空間マイニングが有用である

ことを示す取り組みでもある．

ログマイニングはなんらかの利用者とのインタラクションがあるシステムであれば成り立つ技術で，一般的にシステム提供側，利用者側双方にメリットがあるとされている．利用者側のメリットで代表的なものは推薦で，次に利用者が行うべき操作を，前もって提示してくれれば利用が容易になる．このことはサーチエンジンにおいては単語推薦として成果が出ている．一方システム提供者側のメリットとしては，システム設計上の問題点の把握（これは利用者の利便性の裏返しの問題である），あるいは不正利用を自動的に発見するという使われ方もされている．ここで近年新たな利用目的として着目されてきたのが，キーワード広告というビジネスモデルである．サーチエンジンへの問い合わせキーワード，もしくはインターネットの文書に含まれた単語に関連した情報（広告）を提供するサービスで，インターネットの主要なビジネスモデルに成長している．このサービスを実現するためには，単語間の関連をより広い対象で定義する必要がある．単語空間は非常に広く，人手で無数にある単語間の関連を記述することには限界があるが，この課題を解決する有力な技術がログマイニングである．このロングテールとも呼ばれる，一見対象とするセグメント（利用者／顧客の集まり）が小さい事象を正しく必要な人に伝える問題は，ローカルサーチにおいても特徴的かつ深刻な問題である．例えば，ある事象に興味のあるセグメントは地理的に分割することにより，より小さくなってしまうためである．この問題を利用者側から見ると，任意の地域で自分に有益な，興味深い情報を発見的に探す行為とも重なる．もしインターネットを全世界を均一に最適化する道具であると定義すると，地域最適化は必要でないかもしれないが，実世界に遍在しているローカルな事物を局所的な価値を見いだしながら検索することは興味深い作業であろう．ログマイニングを地理的に行う目的は，ローカルサーチシステムを改善し，また新たなビジネスモデルをもたらし，かつ今までになかったローカル情報を提示するシステムを構築できる可能性がある．現状ではログマイニングを地理的に行う研究は，著者の知る限り報告されていないが，少なくとも上記のシステム改善を超えた目的の下でログマイニングの研究が行われている．

以下空間マイニングとログマイニングに関して，事例を紹介しながら説明する．

2.2.1 空間マイニング技術

空間マイニング手法は一般的なデータマイニング手法に基づく．データマイニング代表手法の一つが相関ルールである [Agrawal 1994]．相関ルールの研究は数多くあるが，おもに処理性能に関するもの [Agrawal 1994] [Han 2000]，ルールを一般化するものがある．後者はルールに階層を入れるもの [Srikant 1995] や，時系列のデータに時間的な制約を入れたルールにするもの [Srikant 1996] などがある．地理的な相関ルールは，地理的なデータから興味深い知識を発見することであるので，概念を一般化のする手法に地理的な性質

を使うことに特徴がでる．地理的な一般化には住所階層を一般化する方法，地域をクラスタリングする方法，さらには“intersecting”や“near-by”などの関係を導入して「海の近辺」などといった地域の関係で一般化するものがある [Koperski 1995][Han 1997].

空間相関ルールは次のように定義できる [Koperski 1995].

定義 空間相関ルール

$P_1 \wedge ... \wedge P_m \rightarrow Q_1 \wedge ... \wedge Q_m$

但し，述語 $P_1...P_mQ_1...Q_m$のうち少なくとも一つが地理に関する表現を持つ.

空間相関ルールには以下の例がある.

空間相関ルール 相関ルールにジオワードが含まれる

例：札幌 → ラーメン

地理的一般化相関ルール 相関ルールの一般化に地理的情報を使うもの

例：上野 → 美術館，渋谷 → 美術館というルール集合から得られるより一般的な東京 → 美術館

Co-location **ルール** 同一の地理位置に共起するもの間の相関ルール

例：(伊豆，温泉)，(伊豆，土産物店) から 温泉 → 土産物店

位置関係にもとづく相関ルール ジオワードとの関係から導かれるルール

例：(湘南，レストラン)，(鎌倉，レストラン) から 海に近い → レストラン

上記の例にもとづいて空間相関ルールを次の通り定義する.

定義 相関ルール $r : x \rightarrow y$

$x \rightarrow y; x \subseteq W, x \subseteq W, x \cap y = \phi.$

W をアイテム集合

支持度 $supp(r)$ は x と y を同時に含むトランザクションの割合

確信度 $conf(r)$ は x を含むトランザクションが y を含む割合

$conf(r) = support(r)/sup(x).$

定義 空間相関ルール $r_g : x \rightarrow y$

$x \rightarrow y; \exists g \in$G,$g \in x$ または $g \in y.$

G は地理空間を表現するアイテム

定義 $co-location$ ルール $r_{cl} : x \rightarrow y$

$x \rightarrow y; \exists gt \in$ GT, $x \in gt$ かつ $y \in gt.$ GT はジオワードタプル集合.

定義 ジオワードタプル gt

$$gt = \{g, w_k, ..., w_n \mid \exists g \in \mathrm{G}, \mathrm{w} \in \mathrm{W}\}$$

また，多くのクラスタリング手法も地理的な情報を扱うことに適している [Everitt 1993]．例えば2次元クラスタリングを行える手法は，距離の定義にジオワード間の地理的距離を用いることにより，ジオワードを扱うことができる．

2.2.2 ログマイニング技術

アクセスログの検索語を解析する研究は既にいくつか行われている．文献 [Baeza-Yates 2004a] は，サーチエンジンログの検索語を検索語を入力した後に閲覧された URL 等を用いて検索語の分類を行う手法である．同様の手法に Lycos の [Beeferman 2000]，Microsoft の [Wen 2002] がある．本研究は閲覧 URL を前提とせず，検索語間の地理的な関係を元に解析をおこなうため手法が異なる．

日本語を対象にしたものでは [大久保 1998] がある．本研究はある一定期間における検索語の頻度等をもとに時期に依存した単語を分類する方法で，時期と地理を置き換えると本研究のモチベーションを共にする．また [大塚 2005] は特定のサーチエンジンに限定されない大域なアクセスログを，ウェブコミュニティ等を使って解析する研究である．この研究はキーワードをインターネットのコンテンツ全体で解析するもので，ジオワードをグローバルに分析するためにも重要なフレームワークである．

Yahoo の [Jones 2006] は，サーチエンジンにおけるリクエスト単語（クエリ）の置き換え方法を提案している．利用者の入力単語の置き換えを利用者からの正解の提示とみなし（疑似 relevance feedback），そこから単語置き換えを行うとしている．この研究はキーワード広告を意識したもので，より多くのクエリの入力機会に対して，適切な単語置き換えができることを明らかにしている．

本研究で扱うアクセスログはローカルサーチのログ，すなわち（ジオワード, フリワード）の組であるという特徴がある．一般にはサーチエンジンの検索では1回の検索に1単語しか入力されないといわれているが（1検索あたり 1.05 単語 [Baeza-Yates 2004b]），ローカルサーチログは検索ごとに最低ジオワードとフリーワード2単語が期待できるので，閲覧 URL などを前提としなくても検索単位でジオワードと他の単語の関連が観測できるという特徴がある．そこで本研究は，ユーザの検索もしくは連続的な検索であるセッションの情報を使ってそこから地理的な関係と取り出すことを基本とする．

なお，ローカルサーチのログマイニングの報告は著者の知る限り存在しないが，[Sanderson 2004] は，一般のサーチエンジンの検索の中のローカルサーチ存在を調べたもので，Excite のアクセスログを使って，検索の 18.6% にジオワードが含まれるこ

とを報告している．続く microsoft の [Wang 2005] はサーチエンジンのクエリの位置情報についてはじめて大規模な分析を行なったもので，クエリの位置情報を QDL (Query Dominant Location) と名付け，クエリを地名の有無と，実際の QDL の有無の 4 つのマトリックスで分類し，それぞれに精度の高い解析を行なう方法を提案している．著者らの先の [Iko 2002] はローカルサーチポータルのアクセスログを解析した報告であるが，ジオワードは単にアイテムとして扱い，通常の相関ルール分析を行ったものである．

ウェブマイニング [Kitsuregawa 2001] は大量のウェブデータを処理するマイニングで，その対象から大きくコンテンツマイニング，ログマイニング，リンク構造マイニングにわけることができる．ログマイニング [大塚 2005] およびリンク構造マイニング [Toyoda 2001] は筆者の研究室で集中的に行われており，様々なインターネットでの動向が統計的に明らかにされて来ている．

ログマイニングを地理的にな立場から行う研究はこれからの分野であり，本論文で手法と有用性を明らかにして行く．

第 3 章

ジオワードを中心とした情報の統合

本章では本研究の 2 つの課題の 1 つである，ローカルサーチの実現，すなわち様々なコンテンツに緯度経度情報を持たせるための方法を報告する．第一の取り組みでは，イエローページデータに緯度経度情報を持たせるために，デジタル地図データと統合する実験を，第二の取り組みでは，ウェブページに緯度経度情報を持たせるために，イエローページデータを統合する実験の結果を述べる．前者は住所属性を持つコンテンツに地図に表示する方法として，実システムで長く使われることができ，有用性が自称された．後者はイエローページデータとウェブのローカルコンテンツを互いに関連づけできることを示して，現在のローカルサーチサービスで実施されている基本概念を明らかにしたことである．

3.1 地図データとイエローページデータの統合

3.1.1 地図検索インタフェイス

店やサービス，あるいは行政機関などを網羅したイエローページデータ（職業別電話帳）は，ローカル情報の代表的なコンテンツである．本節ではイエローページデータを地図データと統合して地図検索サービスを実現する方法について述べる．

ローカルサーチにおいて，結果の地図表示や地図で位置を指定して検索する地図インタフェイスは必須の要素である．しかしながら一般的に名簿や電話帳データは地図サービスを実現するための緯度経度情報を持たない．すなわちこれらの店や人のデータは基本属性として名義や住所を持つことは明らかであるが，その所在地を示す緯度経度情報は，店や人自身が把握していることは稀である．これらの情報を集積したイエローページデータもその例外ではなく，地図サービスを実現するためには，別途緯度経度情報を基本属性に付与する必要がある．

緯度経度情報を管理している代表的なデータがデジタル地図データである．デジタル地図データは，行政区界や地形等の線情報，住所やランドマークと呼ばれる土地を代表する建造物などの位置を示す点情報，線情報／点情報の文字属性を表す文字情報（住所情報）などを有している．緯度経度情報を測量などの方法で取得して，住所情報とともに管理することは地図作成会社の得意とする分野である．

図3.1に，筆者らが実現したイエローページ検索システムの地図インタフェイスを示す[島 1997a][高橋 1997][島 1997b]．地図インタフェイスを用いると，検索結果が地図上にプロットできるが，それだけでなく以下のような検索が実現できる．

1. はじめに利用者は住所を条件に指定して検索を行う（「東京都港区六本木」の「東京大学」）
2. 検索結果が地図上に表示される
3. その地図を位置情報の検索条件として別な検索を実行する（地図で表示した地域の近くの「銀行」を検索）
4. その結果近くの銀行が検索される

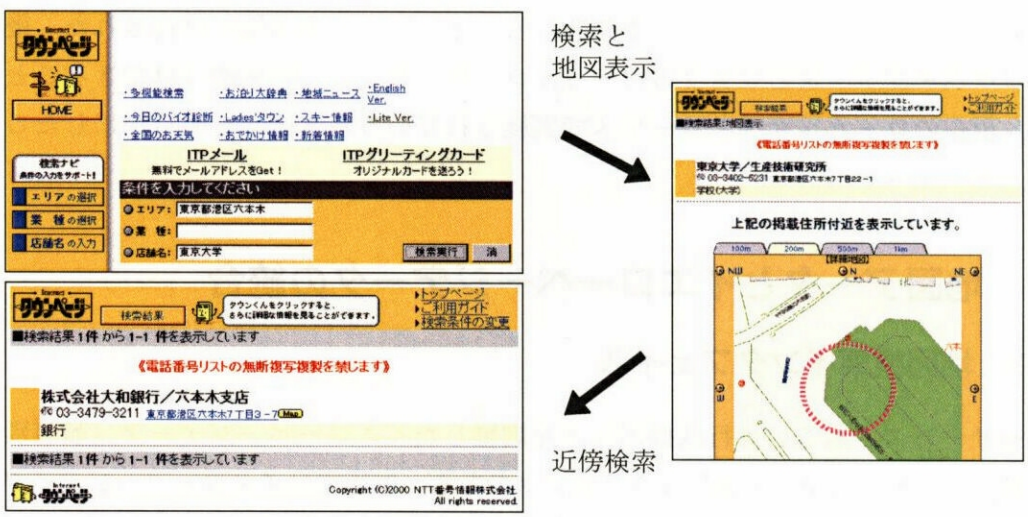

図 3.1 イエローページの地図インタフェイス

3.1.2 異なった固定データのジオワードを用いた統合

地図検索インタフェイスを実現するために，イエローページデータと地図データとの統合を行った．この作業はイエローページデータから見た場合は，新たに緯度経度属性を付与するため，ジオコーディングと呼ばれる作業を行うことということもできる．両者を統

合するためのキーは両者に共通して存在する属性である住所文字列を用いる．この統合作業は，管理者／形式の異なる２つのデータを，住所をキーとして統合する作業である．

実験に用いたデータ

タウンページデータ（NTT） 12,758 件 (1998 年東京都新宿区)

住宅地図（ゼンリン）(1998 年東京都新宿区)

データの例

都道府県	市区郡町村	町	丁目	番地	号
東京都	新宿区	歌舞伎町	１丁目	２番地	３号

マッチング実験結果

全部 (町，丁目，番地，号) が合致したデータ数	9246 / 12758	72.5%	
町，丁目，番地までが合致したデータ数	1875 / 12758	14.7%	(累計 87.2%)
町，丁目までが合致したデータ数	1602 / 12758	12.6%	(累計 99.7%)
町までしか合致しなかったデータ数	31 / 12758	0.2%	(累計 99.97%)

図 3.2 イエローページデータと地図データのマッチング実験

図3.2は，電話帳データ 12,000 件を住宅地図データと照合した結果である．実験結果からわかるように，同時期のデータを照合させても必ずしも全てのイエローページデータの緯度経度を解決することができない．原因としては登録誤り，地名の表記ゆらぎによる異なりなどが存在した．なお，解決しなかったデータは上位の合致した住所による緯度経度で代用する，あるいは人手で正確な緯度経度を付与することで対処することができる．

イエローページデータは住所，名前，カテゴリ，電話番号を属性とし，一方地図データは，住所，名前，緯度経度を属性に持つ．これら２つのデータの統合は単に情報検索を改善するだけでなく，双方の属性を補完的に結びつけて表現することにより，データに別の見方を与えることができる．

図3.3 は上記マッチングの結果を応用して，カテゴリ毎に掲載データの存在を緯度経度で地図にプロットしたものである．赤い箇所により多くの店が存在している．実際にはクリッカブルマップとして実装し，カテゴリごとの地図を選んで，好きな位置をクリックすることにより，その位置の該当カテゴリの検索を行うことを実現した．

以上のようにイエローページデータと住所をテキストエントリとして持つ地図データは統合が可能で，検索サービスを高度化しただけでなく，従来にない，情報の発見的なユーザインタフェイスの提供が可能であることを示した．

なお，本技術にもとづいて開発されたインターネットタウンページ（現，ｉタウンページ）に関しては付録 A.2 節を参照されたい．

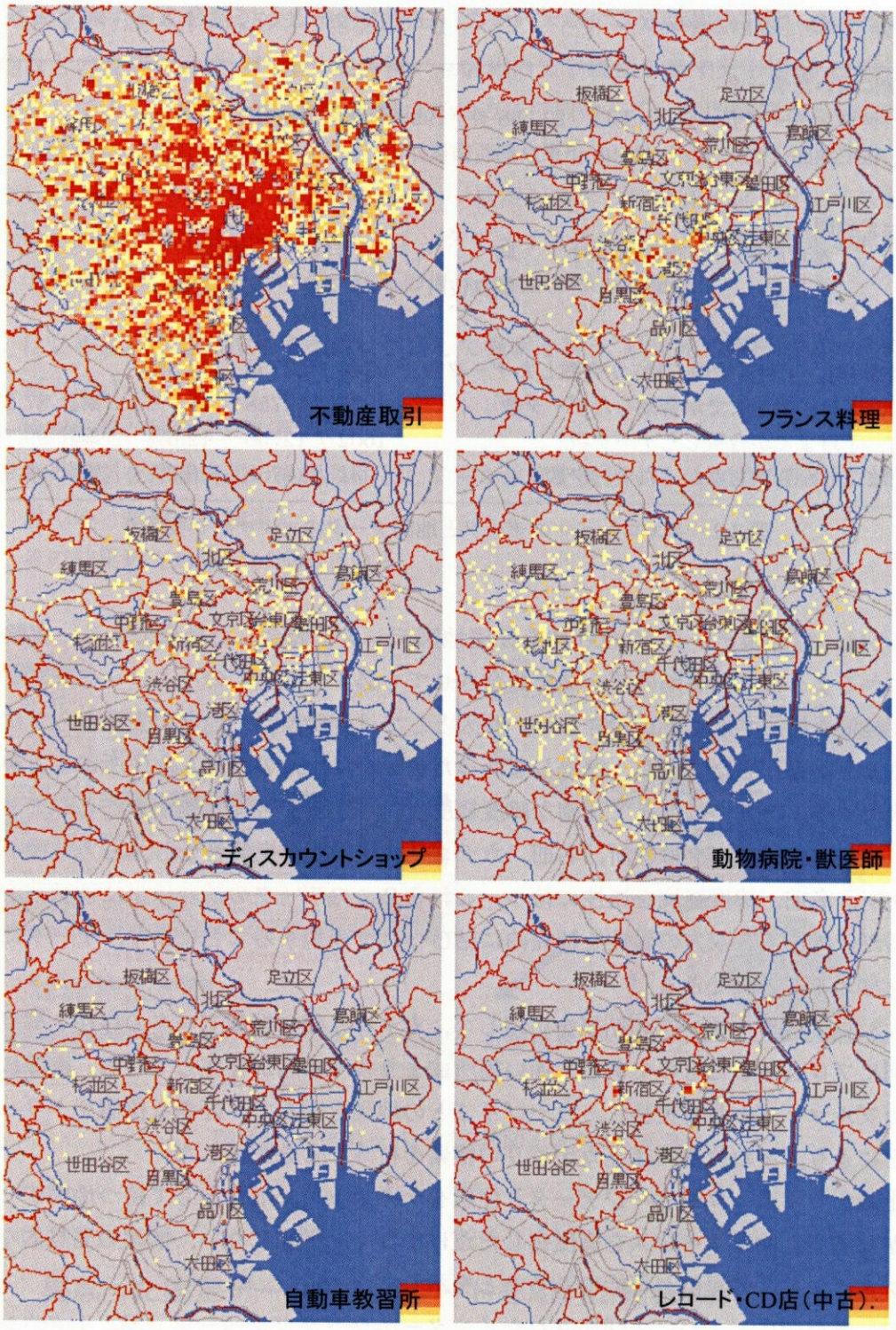

図 3.3 イエローページデータの地図表示による情報分布の可視化（東京都）

3.2 ウェブページとイエローページデータの統合

本節では，ウェブページ（テキスト情報）とイエローページデータを統合する方法について述べる．

モバイル環境での情報検索，すなわち人が携帯情報端末を持ち歩いて，外出先，移動先にて必要な情報を検索する状況を想定した時，その移動先の周辺情報を検索するローカルサーチは最も必要とされるサービスの一つである．しかしながら，モバイル環境での情報検索には次の制約がある．

- 携帯情報端末で利用者が行うことができる操作には制限がある
- モバイル環境で利用できる通信回線の接続性や帯域には制限がある

モバイル環境で多用なローカル情報をインターネットから検索するためには，上記問題点を解決するための工夫が必要である．本論文で扱う情報の地理的統合は一般的な課題であるが，本節ではウェブページとイエローページデータの統合をモバイル環境での情報検索という視点考える

なお本節で想定した携帯情報端末はいわゆる PDA と呼ばれるもので，携帯電話でのインターネット接続サービス（ i モード等）はまだ発表されていない．従って本取組みの対象は，小型の端末に携帯電話等でオンライン接続を行う環境での情報検索である．

このような環境への対策として登場したのがモバイルエージェント，ソフトウェアエージェントと呼ばれるもので，その一つの実現例であるモバイルエージェント言語 Telescript とそのグラフィカルユーザインタフェイス (GUI) を操作できる OS 環境である Magic Cap は以下のコンセプトをプラットフォームとして提供する試みを進めていた．

- 利用者の煩雑な操作をモバイルエージェントという形で切り出し，極力エージェントに操作を行わせる
- そのエージェントタスクの単位（あるいはクラスの単位）に GUI を協調的に対応させることでユーザビリティを向上させる
- 通信量，回数を減らすために，モバイルエージェントも端末と情報源間の通信の制御を行う（複雑なタスクの場合は回線を切断し，結果をメールで送る）

本エージェントの取り組みは，必ずしも現代の情報検索環境においてインパクトがあるものではないが，実現のコアコンセプトにはウェブ情報への地理情報を含む構造化を行いたいということがあった．本取り組みはジオワードを活用したローカルサーチ実現へ，ターゲットをウェブページに移して具体的に踏み出した初期の例である．

3.2.1 Intelligent Page システム

以下に 1995 年に行った Intelligent Page と呼ばれるプロジェクトの実験結果を報告する [Takahashi 1997]. 想定した利用シナリオは以下の通りである.

シナリオ 複雑な検索を行えない環境の利用者（例，モバイル利用者）のために，ソフトウェアエージェントが利用者の望む情報をインターネットから集めて提示する．対象はローカル情報とする（例，銀座のレストランをインターネットのウェブページから集めなさい).

上記のタスクをソフトウェアエージェント [Genesereth 1994] に行わせるためには，雑多なウェブコンテンツに何らかの構造化が必要である．このために考えたのが図 3.4 に示すアーキテクチャである．この仕組みは [Wiederhold 1994] [Neches 1991] らによって議論されていた，メディエータならびに連邦アーキテクチャを参考にしている.

図中 Interface Agent は利用者の検索の代行を行うソフトウェアエージェントの役割を，Yellow Pages Server は利用者の抽象的なリクエスト（銀座のレストラン）を具体的な固有データ（店名，電話番号等）に解決する役割を，Source Index Server は利用者リクエストを解決できるエンティティ（Yellow Pages Server）のアドレスを解決できる役割を担う．各エンティティ間でやり取りさせるメッセージを図 3.5 に示す.

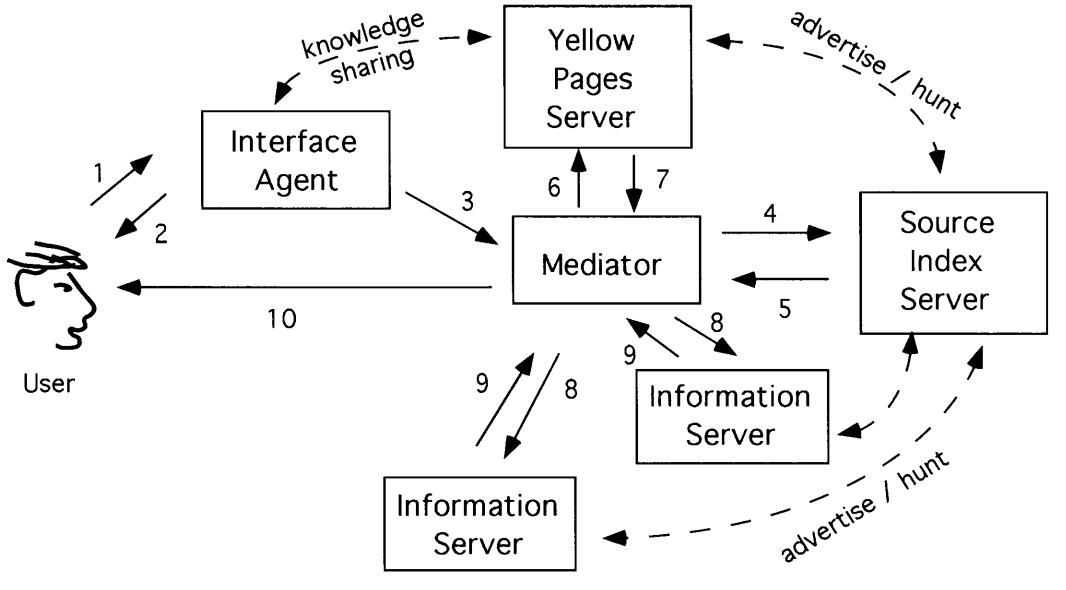

図 3.4 イエローページを用いた情報統合システム

このアーキテクチャは基本的に実験当時でもインターネットで実現可能な情報源を想定して設計をした．この設計に基づく実装を図 3.6 に示す．以下の環境を用いて実装を

1 "I'd like to find <restaurants in Ginza, Tokyo>."	2 Inspection of the user query
3 "Find shops <category = restaurants, address = Ginza, Tokyo>."	
4 "Find the Information Server profile that provides relevant information to the query given in message 3."	5 A list of Information Server profiles
6 "Search telephone listings that have <category = restaurants, address = Ginza, Tokyo> from the Yellow Pages Server."	7 Phone listings (name, address, category, phone no.)
8 "Search any information that has <"ABC restaurant" or "1234-5678"(as a phone no.)>."	9 A review of <ABC restaurant> from a Web review page.
10 A list of organized restaurant information with reviews or a menu	

図 3.5 エンティティ間で通信されるメッセージ例

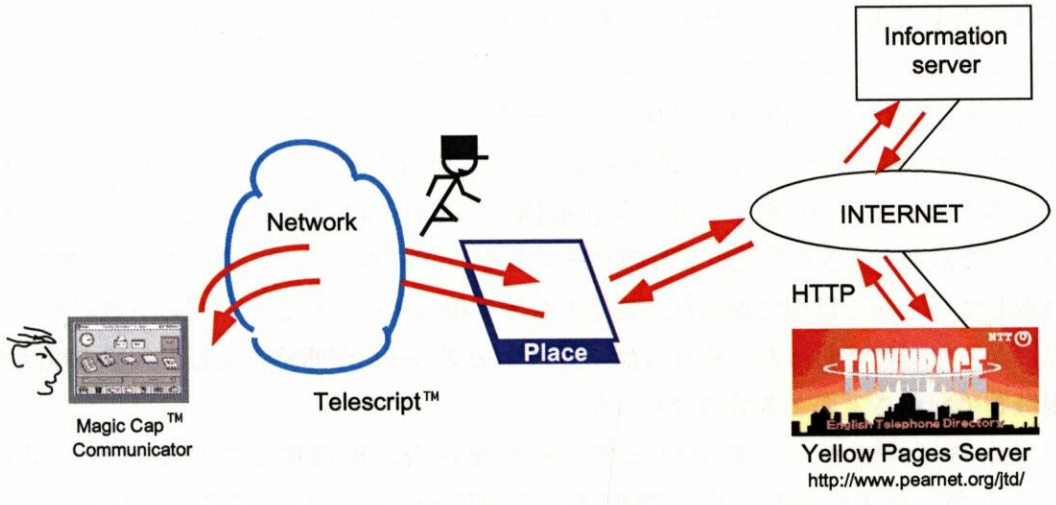

図 3.6 実装のアーキテクチャ

行った.

- 利用者とのインタラクション：MagicCap と Telescript
- Yellow Pages とのインタラクション：Telescript と http プロトコル
- 実際の情報源とのインタラクション：Telescript とウェブクローラの収集結果

MagicCap は Telescript が有するメソッドに対応した GUI を豊かに提供するプラットフォーム（端末の OS でもある）である. Telescript[White 1994] はオブジェクト指向の言語で通信をともなうソフトウェアエージェントを記述することを目的に設計されていた. Yellow Page は実際にインターネットで提供を始めたタウンページサーバを用い, Telescript エージェントが http 疑似クライアントとして動作した. また最終的な情報源はウェブクローラ [林 1996] が収集したウェブページをデータベースに格納して用いた. なおこのプロトタイプを Intelligent Page と呼んだ.

3.2.2 実験と考察

実験の内容は以下の通りである.

- 以下の情報を統合する
  ウェブコレクション：50,000（クロールした JP ドメインのページ）
  イエローページ：15,000（English TOWNPAGE）
- イエローページデータのレコード（店）がウェブコレクションに存在するかどうかを, イエローページの電話番号がウェブコレクションに存在するかで判定する

統合結果を表3.1に示す. 表からもわかる通り, 大学や博物館等のカテゴリでは既に高い割合でウェブページが制作され, この手法でウェブページが一定の高い割合の量で収集できること, さらには結果としてウェブページが構造化できることが明らかになった.

次に得られたウェブページの評価を行った. 評価は例題のクエリに該当するイエローページのレコードを 10 件ずつサンプル抽出し, それらに対応するウェブページが見つかったかを計算し, さらにそれぞれのページがイエローページのレコード記載の対象（店）と合致しているかを目視で調べた. 結果として学校のページで 2 ページの不適切なページが見つかっただけであった. それらは学会のウェブページが連絡先として大学の電話番号を掲載していたことによるものであった.

以上からイエローページでウェブページを統合, 構造化していく試みが可能であることが, 小規模ながらも実際のデータで確かめられた. 実験の当時は Bargain Finder[Krulwich 1995] や ShopBot[Doorenbos 1996] などの同様のインターネット情報収集の取り組みがなされていた. これらの取り組み, すなわち各種の情報源を統合して

表 3.1 ウェブページで見つかったイエローページ情報とそのカテゴリ

イエローページ中の情報数	ウェブページ集で見つかった数	カテゴリ
108	57	UNIVERSITIES & COLLEGES
411	32	HOSPITALS
318	32	SCHOOLS
636	29	RESTAURANTS
477	19	ASSOCIATIONS
94	17	MUSEUMS
49	15	CABINET
174	12	PUBLISHERS
313	8	HOTELS
83	6	COMPUTERS-SOFTWARE & SERVICES
39	6	INFORMATION SERVICES
29	6	HALLS
81	5	ART GALLERIES
28	5	DATA PROCESSING SERVICES
154	5	EMBASSIES & CONSULATES

表 3.2 情報統合結果の精度

問い合わせ	見つかった店舗・組織数 (見つかったウェブページ数)	不適合なページ数
"restaurants in Tokyo"	7 (11)	0
"schools in Tokyo"	4 (14)	2
"museums in Tokyo"	10 (21)	0

仮想的に一つの新しい情報源を作る営みは，現在ソフトウェアエージェントではなく，Web2.0 などとして知られるウェブの Open Interface によってかなりの部分が解決されている．当時の予測では情報源は多用で複雑であることが前提で，その解決のためにより賢いエージェントを開発すべきという考えがこの分野では支配的であった．現在のウェブ情報の構造化は当時の予想を上回るもので，数々のデファクトを含む標準化がすすみ情報の統合を容易にしている．しかしながら本節で指摘したローカル情報の標準化は必ずしも

行われておらず，結果として提案のイエローページを使ったアーキテクチャの価値は失われていない．なお Intelligent Page プロトタイプに関する情報は付録 A.4 も参照されたい．

3.3 情報統合ディレクトリ

本節では，多様な情報を構造を持った情報のコレクションに統合することにより，多様な情報に構造を持たせる方法について論じる．この方法でウェブページをイエローページに統合して，ウェブページに緯度経度情報を持たせることができる．

3.3.1 情報統合ディレクトリの考え方

オープンネットワークに分散して不均一な形式で存在している情報を集め，それらに構造を与え，統合するためのアーキテクチャー，情報統合ディレクトリ（IID:Information Integrate Directory）を提案する．IID の基本思想は，明確な構造を持たないウェブなど情報と，構造情報の集まりである電話帳情報などの「目録」情報を結び付けることにより，不均一な情報に構造を与えることである．IID は，広くインターネットから情報を収集する情報収集モジュール，情報統合の基本となる目録情報の Base Directory（BsDir），および不均一な情報と目録情報を結合する Integrate Module からなる．続いて IID の有効性を検証するために，店やサービス，イベントなどの「ローカル情報」を対象分野にした BsDir を用いて，雑多な情報を統合する実験を報告する．BsDir には実際の職業別電話帳情報 100 万件を用いた結果，集めた雑多な情報のうち，少なくとも 10 から 20% 程度は統合が可能であることがわかった．これにより IID のアーキテクチャの有効性を確認した．さらにこのことは，ローカル情報を対象分野にすると，不均一な情報に，高い精度で，カテゴリや位置座標などを付与できることを示すので，インターネットでのきめ細かな情報サービスに適用可能である見通しを得た．

ウェブに掲載されたローカル情報を含む様々な雑多で不均一な情報がある一方，構造情報集すなわち構造を予め定義し，定義に従って，おもに人手で集められている情報もある．「目録」と呼ばれるものがそれで，固有名詞などの固有情報を扱う分野には目録が存在し，電話帳だけでなく図書目録，製品カタログなどがある．

目録の一情報当たりの情報量は，ネットワークで見つかる一般の図書情報と比べると少ないかもしれない（例えば書評が含まれることはあまりない）．しかしこれらは網羅的に数多く集まることによって重大な価値を持つ．しかもこれらの目録は電子化されているものが多数存在し，ネットワークからも参照可能になりつつある．

我々のアプローチは，一般の不均一な情報に，構造の明確な目録を結び付けることにより，一般情報に統一した構造の定義とその属性値を与え，その属性を介した情報の統合を行うことである．

我々の提案する情報統合ディレクトリ（IID : Information Integrate Directory）は，

ネットワークの不均一で分散した情報を統合するためのアーキテクチャで，広くネットから情報を収集する情報収集モジュール（Information Collecting Module），情報統合の基本となる Base Directory，および一般情報と BsDir を結び付ける Integrate Module からなる．

3.3.2 不均一で分散した情報の統合

問題点

情報の統合には，情報分類，情報抽出，情報組織化などがある [武田 1996]．本論文が目標とする情報の統合とは，情報の対象が同一であることの見きわめ（情報の同定），ならびに情報の対象が持ち得る構造の定義とその属性値を与え（情報の構造化）その属性値で情報を関連付けることである．ここで構造の定義とは，存在単位や属性の定義のことでここからはスキーマと呼ぶことにする．

通常上記統合を行うためには，統合したい情報群にまず構造化を行い，その結果を使って同定を行うことになる．例えば，まずそれぞれの情報から情報の対象の名前などを取り出し，次にその取り出した名前を元に複数の情報のつき合わせを行う．これには以下の問題点がある．

この構造化は，与えられた情報（テキスト）から必要な情報を取り出すという方針で行われるので，情報抽出（Information Extraction）[Cowie 1996] の問題と等価である．すなわち与えられたテキストに自然言語処理を行い，必要な情報を取り出すことが基本である．この流れではどちらかというと意味解析などを行わず，言語パターン情報などを用いて予め定めた情報を効率良く取り出す方法がさかんに行われている（例えば佐藤らの電子ニュースのダイジェスト [佐藤 1995]）．しかしこの方法では，同定に必要な情報が抽出される保証はないので，情報の同定まで確実に行うことは難しい．

さらには同定あるいは情報検索などに必要な属性を確実に取り出すためには，抽出のため何らかのスキーマを使って自然言語処理を行って抽出を行う必要がある．以上から，現状のアプローチではスキーマの自動抽出／効率の良い構築方法が問題と，処理性能はスキーマを使ってテキスト情報を解析する言語処理能力に依存するという 2 点が問題である．

アプローチ

我々は前記2点に依存しない情報統合を実現するために，構造情報である目録との関連付けを行う．そもそも情報の同定を行うためには，存在の単位（店であるのか，図書であるのか）が合意されている必要がある．これに対応するスキーマを定義して情報抽出を行っても，言語処理能力がボトルネックとして存在し，さらに処理が成功したとしても同

定や構造化に使いたい属性値を解析対象の情報が全て含んでいる保証はない.

これに対して網羅的な目録の存在があれば，目録の要素を介することで統合の可能性は広がる．例えば，互いに欠落箇所の異なる不完全な情報同士を同定することよりも，それぞれを構造の明確な目録の情報と同定し，その結果で同じであると判断することにより，両者を結び付けることができる．また，目録と関連付けられた情報は，欠落した属性値を目録の要素から補完することができる.

我々の提案する手法の特徴は以下の通りである.

- 既存の目録に一般情報を関連付ける
- 関連付け方法は目録の要素と一般情報との照合
- 目録の持つスキーマを結合した情報に付与
- 目録の要素の属性値を結合した情報に付与

このアプローチは，情報の自動分類，特に既存の分類体系に，一般情報を追加分類する方法と似ている．しかしこれはキーワードの存在割合などを元に，分類体系に対して比較的緩やかに分類を行うが，我々のアプローチは，分類を行うにも，属性値を得るにも，具体的な目録の要素と一般情報を結合することによって行う点が異なる．すなわち目録情報は，情報の存在単位と存在そのものを定義するオントロジとして存在する．オントロジを使った情報の分類や抽出をする研究に岩爪らの IICA[岩爪 1997] があるが，IICA がオントロジを概念単位で分類などに用いているのに対し，我々のアプローチはさらに細かい個々の存在の単位で用いる点が異なる.

このアプローチの性能は，目録の網羅性と，目録の要素と一般情報との照合の精度の 2 つに依存する．前者は，全ての分野を網羅した目録は考えにくいが（百科辞典の極端なものであろうか)，分野を限定すると，かなり網羅的なものが得られる可能性がある．前述の図書に対する図書目録やタウン情報に対する電話帳などがそうである．目録的データベースのサービスは今後ネットワークでさらに充実するであろうことが予測され，それらを積極的に他の情報と関連で使っていこうという思想である．後者の照合精度は，目録の要素である固有情報を一般のテキスト情報などから見つける問題であるので，言語処理の問題というよりは，情報検索（全文検索）の問題としてとらえることができる.

これら単独で存在している目録的データベースや情報検索などのアプリケーションソフトウェアを連携させること，および情報検索などの利用者の情報活動の支援をするために，「情報を集め統合しておく」という粒度のタスクを提供するためのアーキテクチャを提案することが本研究のねらいである.

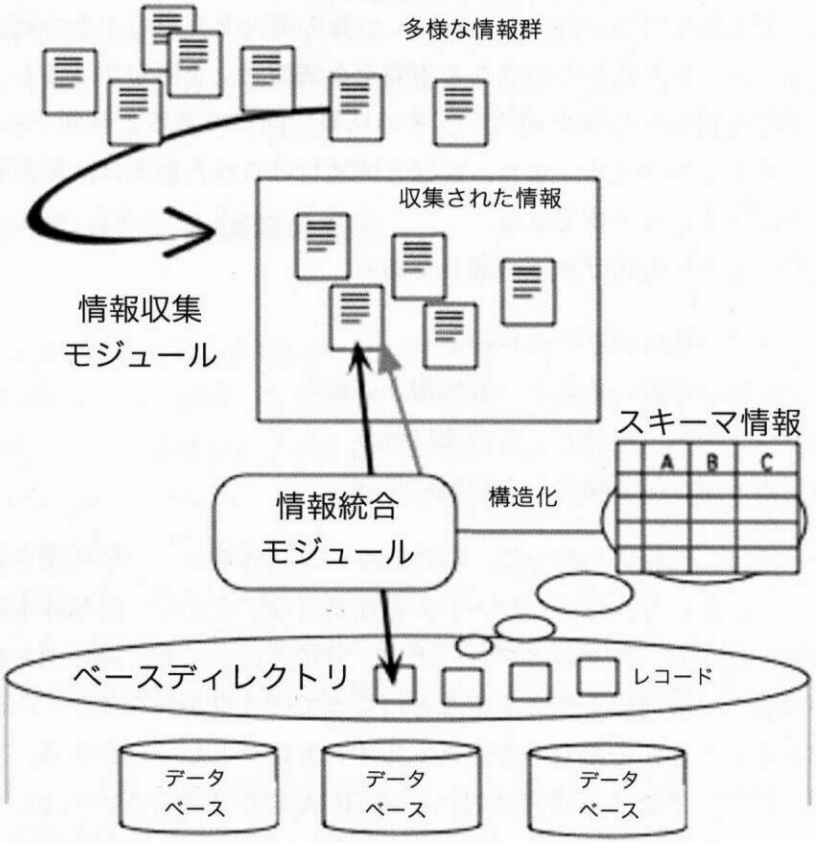

図 3.7 情報統合ディレクトリのアーキテクチャ

3.3.3 情報統合ディレクトリのアーキテクチャ

不均一な情報の統合を行う情報統合ディレクトリ（IID）のアーキテクチャを情報統合ディレクトリのアーキテクチャに示す．IID は大きく以下の 3 つのパートからなる．

Base Directory

ベースディレクトリ (BsDir) の役割は次の 4 つである．

1. Module や情報の存在単位（店など）と具体的なインスタンス（「港区の○×レストラン」など）を規定する
2. 統合する情報のスキーマを規定する（店名，電話番号，住所，メニューなど）
3. 概念体系（分類体系）を規定する
4. 対象分野の具体的な存在の記述を網羅する

1 から 3 の役割は今回は人手で定義を行った．また 4 番目の役割は実用的な立場から，現在手に入るデータベースを利用することを前提とする（後の章で電話帳の例で詳しく述べる）．網羅性を高めるため，あるいは必要な属性を得るために，複数のデータベースを統合して使ってもよい．この場合の異種性の解消は，マルチデータベース [sheth 1990] の考えに従う．

情報収集モジュール（Information Collecting Module）

ネットワークから情報を収集すること，及び収集した情報のインデクシング／検索機能の提供が目的である．さらに収集，維持，保管のコストを下げるために，BsDir を参照しながら BsDir の対象分野に関する情報を選択的に探すことが求められる．

収集は，いわゆるサーチエンジンで使われているウェブクローラ [Cheong 1996] で，基本的なインデクシングは全文検索用のソフトウェアで実現ができる．加えて構造情報に基づくインデクシングを行う．すなわち集めた情報一つひとつに対して Integrate Module (IM) に後述の「固有情報照会」をかけ，その結果得られたスキーマと属性情報を個々の情報に付与する．

収集の効率化は，IM を呼び出すことにより可能となる．すなわち収集を行いながら IM に後述の「固有情報照会」を行わせ，統合が行われるファイルの割合を調べ，その値がある閾値を越えた時点でそのホストからの探索を打ち切る／優先順位を下げる．この閾値は対象のコンテンツと IM の照合法に依存して決まる．

統合モジュール（Integrate Module）

収集モジュールによって集められた一般情報と構造情報である BsDir の要素を関連付けることが目的である．関連付けは両者の照合によるが，その照合方法は一般情報と構造情報のどちらを中心に考えるかによって 2 種類に分けられる．

■**《固有情報全文検索》** 任意の構造を持った固有情報に対して，一般情報の集まりの中から対象が同じである情報を探す．同一性の判定は，構造情報の各属性値をキーワードとして全文検索した結果で行う．属性値の扱いは対象分野の性質によるが，固有性の高い属性値を単独で用いるか，低い属性値を組み合わせて用いることになる．例えばお店ならば《電話番号》を含むもの，または《名前》と《住所》を共に含むものを同一と判断する．BsDir を中心とした統合に用いる．

■**《固有情報照会》** 任意の一般情報に対して，構造を持った固有情報の中から対象が同じである情報を探す．同一性の判定は，一般情報から抽出したキーワードを構造情報集 (BsDir) を検索した結果で行う．詳しく述べるとまず IM は一般情報から BsDir で規定されたスキーマに対応する情報抽出を行い，その結果を条件として検索を行う．例えば

お店ならば《電話番号》や《名前》などを抽出した上でBsDirに照会する．IMがBsDirのスキーマとそれに関する情報抽出をする知識を持っていることを前提とする．集めた情報を中心とした統合に用いる．この時IMはBsDirから見つかった情報を情報収集モジュールに渡し，スキーマと属性値は各一般情報に付与される．

照合の難易度は，情報抽出が入る分「固有情報全文検索」より「固有情報照会」の方が難しい．一方前者はオンライン処理を考えた時BsDirへの通信が常に保証されている必要があるが，後者はBsDirをオフラインで利用し，統合した結果を情報収集モジュール側だけで再利用できる特徴がある．

3.3.4 統合されたローカル情報の検索

ローカル情報と呼ばれる，お店，企業やサービス主体に関する情報がある．これらは不均一な形式で多数ネットワーク上に分散して存在しているが，これを構造化し，日常生活に手軽に利用したいなどというニーズがある．ローカル情報はIIDを使うのに適した分野である．それは以下の理由による．

- 《名前》《住所》などの構造を持つ
- BsDirとして網羅的で構造の明確な目録である電話帳情報が存在する

IIDを用いると，例えば次のような検索システムを作ることができる（図3.8）．数字はメッセージの流れを示す．

外出中の利用者が最寄りのレストラン情報をウェブから調べたくなったと仮定する．しかし一般にウェブの情報はカテゴリ（レストラン）や位置情報といった属性情報を持っているとは限らない．このような時にIIDを次のように使う．

ユーザエージェントはIIDを利用者から隠蔽することが目的であり，利用者の抽象的なリクエスト（「最寄りのレストラン」）を受け，IIDを使った検索結果を利用者に返却する．このために必要な仕事は，利用者のリクエストをBsDirの概念体系に翻訳し，統合モジュールに一連の仕事を委託することの2つである．前者を実現するために，User AgentはBsDirの概念体系（分類語彙の体系）とスキーマを知っている必要がある．

統合モジュール(IM)はユーザエージェントからのリクエストに従って利用者のリクエストを遂行する．BsDirでは通常の検索が行われ，レストランの集合を得る．BsDirではマルチデータベースの工夫をしてもよい．例えばカテゴリと位置情報を条件として検索を行うために，電話帳DBと地図データベースと結合したマルチデータベースを作っておき，位置情報を補間しておく．

検索が行われたら，BsDirのレストランのリストはIMに渡される．IMはリストのそれぞれの店ごとに「固有情報全文検索」，すなわち属性値を取りだし検索式を作成して全

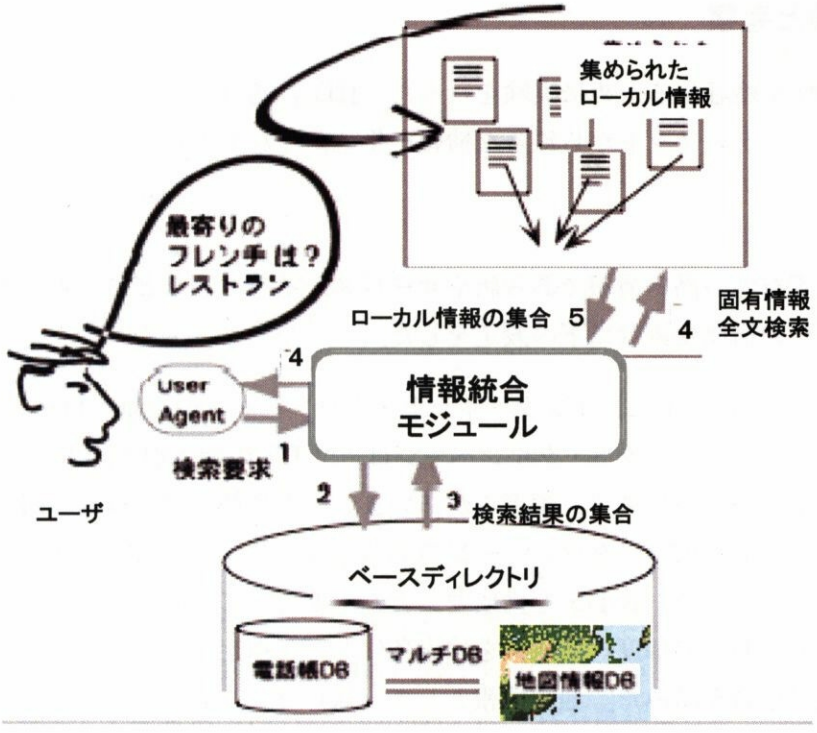

図 3.8 情報統合ディレクトリに基づく情報の検索

文検索を行い，関連ある情報の集合を得る．IM はこうして作られた店単位の情報の集合をユーザエージェントに返却する．このようにして，集められた一般情報に対して構造を意識した検索が可能になる．

なおこの構成のアプリケーションの処理時間は，BsDir に対する 1 回の検索時間と，その検索結果全てに対して繰り返し行われる固有情報全文検索の時間で決まる．データの入手可能性などから判断して，前者をインターネット上の検索サーバにアクセスし，後者はローカルな計算機に収集したものに対して行うと仮定する．後者は通常の全文検索にかかる時間と同等なので，例えば収集したファイル数が 100 万でそれを全文検索するのに 1 秒かかるとしても前者で n 件の検索結果が得られた場合は，BsDir にインターネット経由でアクセスする時間に n 秒のオーバーヘッドがかかることになる．なお次節で説明する実験から，情報収集エージェントが収集した情報にある固有情報が多量に合致することは少ないので，処理時間はインデックスを走査する時間と僅かな実ファイルアクセス時間で見積もることができる．

3.3.5 実験と考察

IIDの有効性を検証するために実験を行った．IIDの適用可能な領域と割合を明らかにするために，インターネットから雑多な情報を集め，それをIIDを用いて統合を試みた．

実験構成

対象分野は固有性の高い情報である店やサービス，イベントなどの「ローカル情報」とし，IIDの3つの構成要素は以下の設定をした．

■Base Directory BsDirにはインターネット・タウンページの東京23区の全て，約100万件を用いた．インターネット・タウンページは，ＮＴＴの職業別電話帳（タウンページ）情報を提供しているサービスで，筆者らがあいまい検索や個人適応技術の研究の対象として構築している [島 1997a]．タウンページの掲載数は全国で約1,100万件である．電話帳情報は網羅性が高いので，BsDirを構成要素として期待している．

このサーバでいわゆる緯度経度情報を検索条件として指定可能とするために，地図情報と電話帳情報の結合を試みた．地図情報として（株）ゼンリンの住宅地図を用い，主として住所情報の形式の異種性を解消して結合を行った．広い意味でのマルチデータベースに含まれる．その結果このBsDirの持つ属性値は《名前》《カテゴリ》《住所》《電話番号》《位置座標》の5つである．

■情報収集エージェント ウェブクローラである情報収集エージェントはlibwww-perl [Fielding] を用いて，ウェブからHTMLファイルを以下の基準で収集した．(1) 形式の偏りを避けるため，数多くのサイトから集めること，(2) 分野を限定した情報と，分野を限定しない情報をそれぞれを集めること．このため，いわゆるリンク集を起点として，幅優先，同一ホスト内の取得ファイル数に上限をつけて収集した．集めた情報は以下の3種類である．

1. 分野を限定しないもの（日本の新着情報）
   NTT Home Page (http://www.ntt.co.jp/) の日本の新着情報から3日分．
2. ローカル情報（レストラン）
   Yahoo! JAPAN (http://www.yahoo.co.jp/) の「東京:エンターテイメント:飲食店」と「東京:エンターテイメント:飲食店:リンク集と総合情報」から．
3. ローカル情報（ビジネス）
   Yahoo! JAPANの「東京:ビジネス」から．

■Integrate Module 今回の役割は，集めた情報に BsDir のエントリを対応付けることなので，「固有情報照会」を行う．

- 集めた HTML ファイルから情報抽出する
- 抽出結果と BsDir の要素を照合
- 照合の結果一致情報を統合されたとする

ここでは，統合可能を保証する最低値を知りたいので，照合基準として多少のもれが存在しても精度の高い電話番号を用いた．電話番号を含まない情報はあり得るが，含んだ情報で一致したものは適切なものであることによる．

結果

実験の結果を表3.3に示す．

統合率は収集した HTML ファイルが BsDir の要素に一致した割合である．照合一致の基準が電話番号であるので，統合率は「電話番号を含有する割合 [C]」と「含有電話番号が BsDir で見つかる割合（照会された割合）[D]」の積で与えられる．[D] は実験に使ったタウンページのデータが東京 23 区のみであるので，「含有電話番号のうち東京 23 区に該当するものが実験データ中で見つかる割合」で代用した．電話番号を抽出する方法は，数字列の桁数，および区切り記号（ハイフン，括弧など）のヒューリスティックスを用いた．

実験結果の考察

電話番号含有率 [C] は，分野を限定しない「新着情報」に関して 14.9% が得られた．この「新着情報」は個人や団体が自分の製作したウェブのページを自己紹介する目的のリンク集であるため，比較的無作為なウェブページの集まりである．さらに，分野を限定したレストラン情報で 27.5% が得られた．

そもそも情報収集には探索の開始点の URL しか与えておらず，しかも幅優先探索を行っているため，例えばレストラン情報をたどっても，全くタウン情報と関連のない情報も含まれている．この条件下で特に「新着情報」に電話番号が含まれる割合が高かった．しかもここで統合されたものには，例えば BsDir の分類カテゴリを付与することが可能である．例として表3.4に統合された情報の BsDir における分類を示す．

Integrate Agent が行う，BsDir のレコードと雑多な情報との照合は，全文検索に基づくが，より高精度の照合を検討する必要がある．特に固有情報の表現の多様性と，集められた情報の内容の多重性に注目している．

前者であるが，今回は情報統合のキーとして電話番号を使ったが，当然タウン情報には電話番号を含まないものもあり，例えば名前を使えばより多くの情報を見つけることがで

表 3.3 ウェブページがベースディレクトリに統合された割合

分野:新着情報		
収集 HTML ファイル数 [A]（ホスト数）	10420	(306)
電話番号を含む HTML ファイル数 [B]	1548	
電話番号含有率 [C](=B/A)	14.9	%
電話番号照会率 [D]	53.7	%
統合率 (=CxD)	8.0	%
分野:レストラン		
収集 HTML ファイル数 [A]（ホスト数）	5017	(157)
電話番号を含む HTML ファイル数 [B]	1380	
電話番号含有率 [C](=B/A)	27.5	%
電話番号照会率 [D]	84.7	%
統合率 (=CxD)	23.3	%
分野:ビジネス		
収集 HTML ファイル数 [A]（ホスト数）	9715	(345)
電話番号を含む HTML ファイル数 [B]	1837	
電話番号含有率 [C](=B/A)	18.9	%
電話番号照会率 [D]	69.5	%
統合率 (=CxD)	13.1	%

表 3.4 統合された情報の Base Directory における分類

分野:レストラン		分野:新着情報	
レストラン	505	文具・事務用品店	58
レストラン（各国料理）	376	情報提供サービス	52
フランス料理店	155	事務用機械器具販売	24
日本料理店	121	旅行業	21
バー・クラブ	106	外国公館	20
イタリア料理店	88	屋形船	17
中国料理店	85	市区町村機関	16
居酒屋	85	電気通信業	14

きる．これはあいまい検索と呼ばれる照合法，すなわち複合名詞である固有名詞から，接頭語などの複合語処理を行った後，標記のゆれを正規化する照合 [高橋 1997] を用いればよい．表3.5に表3.4で使ったレストランの情報をサンプル抽出して人手により分析した結果を示す．照合技術の充実により，レストランの分野では最大約 50% までの統合が可能である．

後者の情報の多重性，すなわち一つのファイルが複数の対象を記述している場合に対する問題であるが，集めたファイルから，多重性には並列記述（例：お店のリスト），包含記述（例：主情報の関連情報の付記）などが見受けられた．なお今回多重性には，現れた全ての対象に同等に尊重して対応付けを行った．

表 3.5 レストラン情報ウェブページの分析

《名前》を含む	34%
《住所》を含む	26%
《電話番号》を含む	34%
《名前》または《電話番号》を含む	48%

統合率は BsDir の網羅度に依存する．今回は単一の電話帳を用いたが，分野ごとに網羅性の高い BsDir の構築は，先述のマルチディレクトリなどの方法論の問題だけでなく，内容的にいかに充実したものが構築できるかが重要である．このためには情報源の中から自動的に BsDir のエントリを作り出す技術も必要である．さらに BsDir を統合の目的に応じて，より抽象的なものを選んでやることにより，より網羅的な統合が可能となる可能性がある．例えばイエローページデータをローカルサーチという観点で抽象化する場合，住所などを集めたジオワード集が代わりとなる抽象的な BsDir を与える．

オープンネットワークで IID を実際に成立させるには，BsDir を構成する実際のデータベースのデータをプライバシや課金の観点から保護する配慮が必要である．