# リバースジオコーディングとオントロジーを用いたWebページ検索

2006MI136 岡村 正和
指導教員　河野 浩之

## 1　はじめに

現在, 私たちはサイトを検索する際, 検索サイトから検索することが多いが, 携帯型移動端末などの機器で検索する際, 文字入力の手間や, 目的の情報へのアクセシビティが悪い問題がある. そこで, 本研究では, ユーザーが位置情報を与える操作を行うだけで, システムがその位置情報をキーとし, そのキーから形態素解析とオントロジー検索によりその場所に適した情報が記載されたURL をユーザーへ返すシステムを提案する.

## 2　オントロジー主体の検索手法

既存のオントロジー検索エンジンとして, Swoogle がある. Swoogle には, 2009 年 3 月現在 1 万 5 千以上のオントロジーが登録されている. Swoogle では, クラス単位, プロパティ単位の検索やオントロジー内に明示的に記述されてない逆リンクの関係（あるクラスを参照しているインスタンス一覧など）を検索することが可能である. また, オントロジーを検索するための 19 種類の REST 形式の Web サービス（Swoogle Web サービス）も提供されており, プログラム上からオントロジーを検索することも可能である. しかし, Swoogle での検索結果を見る限り, 日本語での検索において満足の良く結果が得られていない. そのため, 今回では GoogleWeb 検索エンジンを用いて, オントロジー主体の検索を行うこととする.

## 3　構築システムの実装と概要

### 3.1　試作システムの全体概要

step.1: 位置情報の送信
ユーザーは GPS を利用して取得した位置情報を送信.

step.2: DB のチェック
取得した位置情報が DB に格納されていないかチェックする.

step.3: リバースジオコーディング
ユーザーから送信された位置情報をキーとしリバースジオコーディングを用いて住所を取得する.

step.4: リソース検索処理
イントラネット検索エンジンおよびインターネット検索エンジンを利用して, リバースジオコーディング住所をクエリーとする完全一致検索を行う.

step.5: Yahoo 形態素解析
step.4 の検索結果を Yahoo 形態素解析を用いて, 形態素解析を行ない, 名詞を抽出し, 最適なルールでフィルタリングを行う.

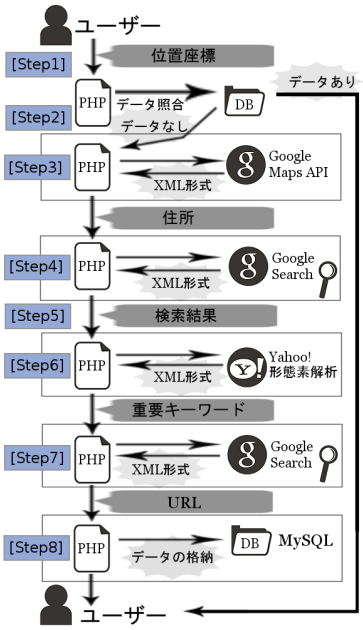


図 1　構成技術を考慮したシステムのデータフロー図

step.6: 重要キーワードの抽出
フィルタリングにより抽出されたキーワードを, 出現回数, 検索結果の順位を用いてパラメータ付けし, そのパラメータの一番高いものを重要キーワードとして決定する.

step.7: 重要キーワードをキーとした検索処理
step.6 で決定された重要キーワードをイントラネット検索エンジンおよびインターネット検索エンジンを利用して, 重要キーワードに関する URL を取得する.

step.8: URL の送信
取得した URL をユーザーに返し, DB に格納する.

位置情報による Web 検索システムのプログラムはベースを PHP とし, API でのレスポンスの形式は XML とする. **図 1** はプログラムとデータベースを考慮したシステムの全体図である. **図 1** での各 API で取得した XML 形式のデータは PHP で必要な部分だけ抽出する. **図 1** での各 API では, 文字コードを UTF8 としているため, データベースと PHP のプログラムは UTF8 で作成する.

表 1　メインシステムの検索精度結果

| ジャンル | 精度 |
|---|---|
| 映画館 | 0.77 |
| 商業施設 | 0.73 |
| 観光地 | 0.57 |
| レジャー | 0.46 |
| 飲食店 | 0.45 |
| 大学 | 0.75 |
| 役所 | 0.7 |
| すべて | 0.61 |

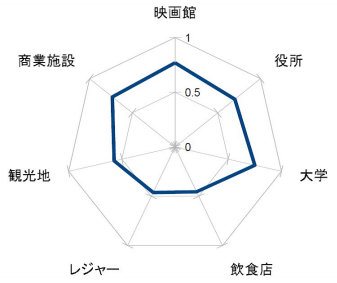

図 2　メインシステムの検索精度結果のグラフ

### 3.2　性能向上のための不用語の排除

不用語の排除を行うプログラムを PHP で用意する．文字列のエスケープのプログラムでは, 形態素解析に渡す文章の中で, 数字や”-”, ”丁目”といったワードが入っていると, 形態素解析の結果が数字ばかりになってしまう可能性があるため, エスケープ文字として排除する処理を行なっている．また, 住所で検索しているため検索結果 summary には多くの検索した住所が入っていると考えられる．そのため, 形態素解析に渡す前に解析する文章内から検索に用いた住所をエスケープする処理を行う．

### 3.3　サブシステムによる性能向上

試作システムでランダムな位置座標でテストを行ったところ, 余分な情報が多く出てくるため, フィルタリングを行うことを検討したが, これにより必要な情報までも削り取られてしまう可能性があった．そこで, 本研究では, 1 つのメインシステム（main）とメインシステムを主体とした, メインシステムとは異なる検索をする 3 つのサブシステム (alfa, beta, gamma) を作り, サブシステムの結果を加味し, ユーザーに結果を返すシステムを構築こととした．

## 4　評価実験結果

レジャー, 飲食店を集約したポータルサイトが数多く存在するため, レジャー, 飲食店で 50 %以下と低い精度となっている．検索結果が適切でない例としては, ぐるなび, Hotpeppr などのポータルサイトのページが表示されてしまう場合が存在する．ポータルサイトの情報では, 現状その施設にいるユーザーに対しては有用な情報とはいえないと考えられる．また, 観光地では, インターネット上に住所を登録している地点が少なく, 検索結果自体が 0 である地点も存在した．しかし, 役所, 大学のジャンルにおいては, インターネット上に住所のデータを持つページが多く存在した為, 検索精度が高いものとなった．Main の弱い精度を 3 つのサブシステムが補い, Main は 3 つのサブシステムでは検索結果として表示されないようなページを収集することができたため, システム全体としては, 全く関係の無い情報はほとんど見受けられなかった．

## 5　今後の課題と改善点

本研究では, 位置座標からオントロジー主体の検索を用いて Web ページを検索するシステムを試作した．本システムは, 既存のウェブ上の施設検索サイトと異なり, 施設の情報をデータベースではなくロボットにウェブをクロールさせることで取得する．そのため, 既存のデータベースに登録されていない施設に関する情報も取得でき, また低コストな環境での利用が可能である．本システムが出力する住所のクエリに対する適合性の評価実験を行った．評価実験の適合率は施設のジャンルごと集計し, 結果をグラフとしてまとめた．また, オントロジーキーワードでの検索の際に 3 つの異なるフィルタリングを持つシステムを作り, 検索精度の向上を図った．その結果, 適合率は平均して 60.75 % となった．また本システムの大きな欠点として API に依存してしまう点が挙げられる．外部システムに依存することで, 外部のシステムの変更に敏感に対応しなければならないなどの問題点がある．今後の課題としては, 現在のアルゴリズムでは適合率が低くなってしまう施設のジャンルに対して, 高い適合率得られるようにアルゴリズムを改善することが挙げられる．また, ユーザーの行動や, 使用されるであろう環境の絞込みなど, 別のアプローチからのアルゴリズムの改善も必要であると考える．

## 参考文献

[1] 大沼他:“ Web コンテンツの分析に基づくオントロジ構築および属性抽出の試み ”, 第 72 回情報学基礎研究会, pp.49-54, 2003.

[2] 間瀬, 山田:“ Web ページ集合からの階層的知識の構築, 人工知能学会全国大会論文集 (17) ”, pp.46-47, 2003.

[3] 松平他:“ Web コンテンツの分析に基づくオントロジー構築および情報整理の試み ”, 人工知能学会研究会資料 SIG-SW&ONT-A302-08, pp.1-8, 2003.

[4] 森田他:“ オントロジー検索エンジンを用いた領域オントロジー構築支援環境 DODDLE-OWL の拡張 ”, 人工知能学会研究会資料 SIG-SWO-A603-07, pp.1-8 2007.