社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－DBS－130 (11)
2003－ＦＩ－ 71 (11)
2003／5／22

ドキュメント・データを対象としたジオ・コーデイング手法

細川 宜秀† 高橋 直久†

本稿では，ドキュメント・データを対象としたジオ・コーディング手法について述べる．ここで，ジオ・コーディングとは，地名から対応する緯度経度を算出する手法を表す．本手法の主要な特徴は，完全な形で記述されていない地名表現をドキュメント・データの持つ文脈から適切な緯度経度を抽出するための機構を実現する点にある．本手法の実現によって，空間オペレータをドキュメント・データ検索機能として利用することが可能になる．さらに，実験によって本手法の有効性を明らかにする．

A Geo-coding Method for Document Databases

Yoshihide Hosokawa† and Naohisa Takahashi †

In this paper, we present a geo-coding method for document databases. The main feature of our method is to fix locations of a document data by recognizing the contents. Our method makes it possible to apply spatial operators to information retrieval for document databases. We clarify availability and effectiveness of our method by showing several experiments.

1. はじめに

近年の広域コンピュータ・ネットワーク技術の発展・普及にともない，地理情報を含む情報源は増大している．それらの情報源に含まれる地理情報を介した新しい情報獲得技術の確立が重要となっている．

本稿では，ドキュメント・データを対象としたジオ・コーディング手法について述べる．ここで，ジオ・コーディングとは，地名から対応する緯度経度を算出する手法を表す．本手法の特徴は次の点にある．

特徴-1 ドキュメント・データにメタデータとして緯度経度データを埋めこむことなく，ドキュメント・データに含まれる地名の緯度経度を確定するための機構を実現する．

現在，位置情報を介した情報システムを実現するために，位置情報を，ドキュメント・データのメタデータとしてそのデータに埋めこむための手法[3),4)]（ここではこの手法を比較対象手法と呼ぶことにする.）が提案されているが，提案方式は，そのような手法と一線を画する．すなわち，比較対象手法との比較における提案方式の最大の特徴は，ドキュメント・データと関連する位置情報を動的に計算するためのメタデータを，その位置情報に埋めこむ点にある．これによって，位置情報に対応するメタデータを埋めこみさえすれば，どんなドキュメント・データからも関連する位置情報を動的に計算することが可能になる．

特徴-2 完全な形で記述されていない地名を，ドキュメント・データの内容認識を伴って対応する緯度経度を計算するための機構を実現する．

ドキュメント・データに現れる地名には，対応する緯度経度を確定するのに十分な表現を有さないものがある．提案方式では，そのような地名を対応する緯度経度に変換するために，その地名を含むドキュメント・データの内容認識を伴って，対応する緯度経度を確定するのに不足している情報を補完し，その上で対応する緯度経度を計算するための機構を実現する．

本手法の実現によって，空間オペレータをドキュメント・データ検索機能として利用することが可能になる．さらに，実験によって本手法の有効性を明らかにする．

2. 内容認識を伴うジオ・コーディング方式

本節では，内容認識を伴うジオ・コーディング方式（提案方式）について述べる．提案方式の主要な特徴は，完全な形で記述されていない地名表現をドキュメント・データの内容認識を伴って，適切な緯度経度に

† 名古屋工業大学大学院工学研究科情報工学専攻
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

変換するための機構を実現する点にある．提案方式によって，ドキュメント・データの検索を緯度経度を介して行うことが可能になる．

提案方式の実行手順は，次のとおりである．**図 1** は，その実行例を表す．

準備： 地図データに登録されているランドマーク毎のメタデータ作成

提案方式を実行する前に，地図データに登録されているランドマークにその特徴を表す単語群を割り当てる．図 1 の 4 属性を持つリレーションは，地図データに登録されているランドマークにメタデータを割り当てて作成されたものである．

緯度経度，住所，ランドマークの 3 データについては，提案方式では，市販されている数値地図データから自動抽出する．昭文社の数値地図 [11] などの地図データは一般に，これらのデータを含むためである．

ランドマークのメタデータについては，人手によって作成するものとする．

図 1 では，米大リーグに参加しているチームの本拠地球場に，‘リーグ名’，‘本拠地名’，‘チーム名’を表す単語を割り当てている．

Step-1： ドキュメント・データとランドマークの内容に関する類似性評価によるランドマーク・ソーティング・リストの生成

ここでは，入力として与えられたドキュメント・データと地図データに登録されている各ランドマーク間の類似度を計算し，類似順に並べたランドマークのソーティング・リストを生成する．この手続きによって，入力されたドキュメント・データに関連するランドマークの候補を生成する．

提案方式では，入力として与えられたドキュメント・データと地図データに登録されている各ランドマーク間の類似度を，主要なドキュメント・データ検索方式として位置付けられるベクトル空間モデル [12] を用いて計算する．具体的には，そのドキュメント・データとランドマークの特徴ベクトルを生成し，そのベクトル間の余弦を計算する．また，ドキュメント・データとランドマークの特徴ベクトルは，TF/IDF[12] による重みづけ手法を用いて生成する．

図 1 の Step-1 では，‘ヤンキースタジアム’ など入力されたドキュメント・データに高い相関を持つランドマークが上位にランキングされる．

Step-2： ドキュメント・データに含まれる地名が指すランドマーク候補の生成

ここでは，入力されたドキュメント・データに含まれる地名が指すランドマークの候補の生成を行う．この手続きは，入力されたドキュメント・データと高い相関があるランドマーク群（Step-1 の結果）から，そのドキュメント・データが指すランドマークを確定するための本質的手続きとして位置付けられる．

提案方式において，入力されたドキュメント・データに含まれる地名が指すランドマーク候補の生成を，次の手順によって行う．

Step-2.1 入力されたドキュメント・データを形態素に分割する．

Step-2.2 Step-2.1 で抽出された形態素の集合から名詞と未知語に分類される形態素を抽出する．

Step-2.3 Step-2.2 で抽出された形態素を含む住所を持つランドマークを，その形態素のランドマーク候補とする．

提案方式において，地名を構成する単一の形態素からその地名が指すランドマークの候補を生成する理由は，提案方式の目的がジオ・コーディングのための特別な辞書構築のオーバヘッドを最小限に抑えることにあるからである．提案方式では，既存の形態素解析技術を利用することによって，この手続きを実行するための特別な辞書構築を一切行わない．また，提案方式の目的が，緯度経度を介したドキュメント・データ検索のためのインデックス作成にあるので，厳密な地名の抽出は提案方式の本質的課題ではない．

図 1 は，この手続きによって，入力されたドキュメント・データから，地名に対応する 2 つの形態素（‘ニューヨーク’ と ‘フロリダ’）とそれが表すランドマーク候補が生成される様子を示す．

Step-3： 入力されたドキュメント・データと高い相関があるランドマーク群からの，そのドキュメント・データが指すランドマークの確定

ここでは，入力されたドキュメント・データと高い相関があるランドマーク群から，そのドキュメント・データが指すランドマークを確定する．具体的には，Step-2 で抽出したドキュメント・データが指すランドマーク候補の中から，そのドキュメント・データに最も内容が類似しているランドマークを選択することによって，その確定を行う．

図 1 は，この手続きによって，入力されたドキュメント・データに含まれる 2 つの地名（‘ニューヨーク’ と ‘フロリダ’）が指すランドマークが確

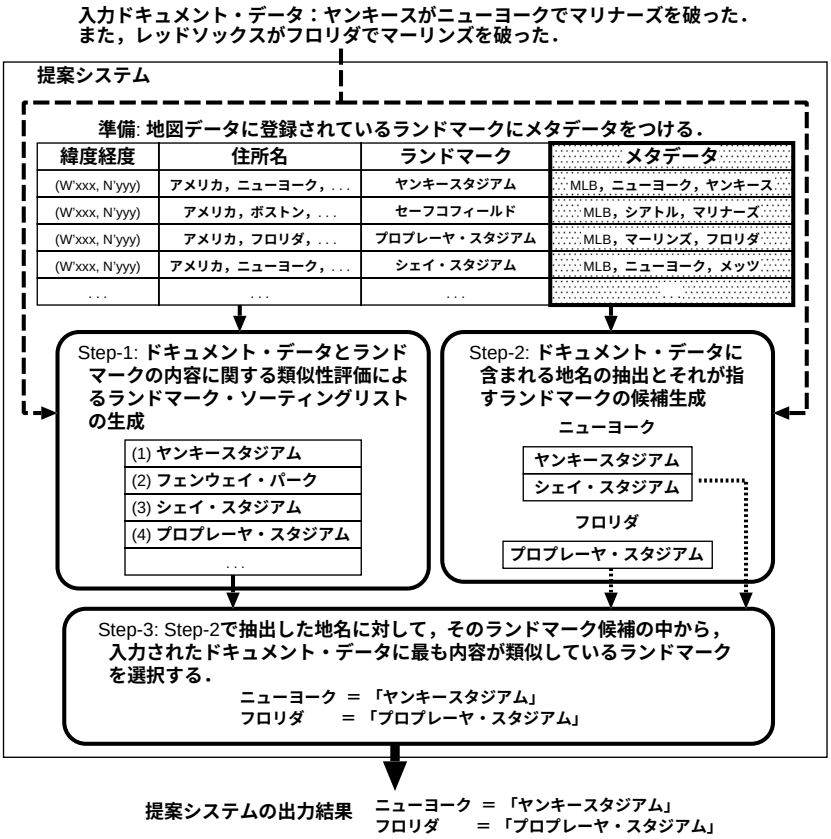

図 1 提案方式によるジオ・コーディング手続き
Fig. 1 The geo-coding procedures of our system

‘茶筅’ による形態素の抽出		‘Namazu’ に発行するキーワード列
ヤンキース, が, ニューヨーク, で, ..., レッド, ソックス, が, フロリダ, ..., 破っ, た	→	ヤンキース or が or ニューヨーク or で or ... or レッド or ソックス or が or フロリダ or ... or 破っ or た

図 2 日本語形態素解析器 ‘茶筅’ の結果と住所の照合による地名の識別手順
Fig. 2 Our computation method of the correlation between a document data and a landmark by using the ‘ChaSen’ and the ‘Namazu’

定される様子を示す．

2.1 ドキュメント・データとランドマークの内容に関する類似性評価によるランドマーク・ソーティング・リストの生成方式

ここでは，Step-1 の実現方式について述べる．提案方式では，ドキュメント・データとランドマークの内容に関する類似性評価を全文検索システム Namazu[6] を利用して行う．

提案方式では，ランドマークのメタデータを検索対象とする．入力されるドキュメント・データからランドマーク検索のためのキーワード列を生成する．これによって，入力ドキュメント・データに類似するランドマークを抽出する．その実行手順は次のとおりである．

Step-1.1 ランドマークのメタデータを形態素解析し，形態素を特徴とする特徴ベクトルを生成する．提案方式の実験システムでは，主要な形態素解析器として位置付けられる茶筅 [1] を利用して特徴ベクトルの生成を行う．

Step-1.2 入力されたドキュメント・データから，ランドマーク検索のためのキーワード列を生成する．ここで生成するキーワード列は，そのドキュメント・データを構成する形態素によって構成する．これは，キーワード列の特徴ベクトルを，ランドマークの特徴ベクトルと同じ要素を持つベクトルとして構成するためである．キーワード列の生成には Step-1.1 で利用した形態素解析器を使用する．

図 2 は，入力されたドキュメント・データから Namazu が解釈可能なキーワード列への変換例を

日本語形態素解析器 '茶筅' による形態素の抽出 (Step-2.1)		名詞と未知語の選択 (Step-2.2)	住所名に含まれる形態素の選択 (Step-2.3)
ヤンキース	名詞-固有名詞-組織	選択	
が	助詞-格助詞-一般		
ニューヨーク	名詞-固有名詞-地域-一般	選択	選択
で	助詞-格助詞-一般		
…			
レッド	名詞-一般	選択	
ソックス	名詞-一般	選択	
が	助詞-格助詞-一般		
フロリダ	名詞-固有名詞-地域-一般	選択	選択
…			
破っ	動詞-自立		
た	助動詞		

図 3 日本語形態素解析器 '茶筅' の結果と住所の照合による地名の識別手順
Fig. 3 Our selection method of location names included in a document data

表す．ドキュメント・データを構成する形態素がランドマークのメタデータに含まれていれば，そのドキュメント・データとランドマーク間に相関があるので，提案方式では，ドキュメント・データの形態素を 'OR オペレータ' によって連結した検索式を Namazu に発行する．

Step-1.3 ドキュメント・データの特徴ベクトルとランドマークの特徴ベクトルの内積を計算し，その類似度を求める．

2.2 ドキュメント・データに含まれる地名が指すランドマーク候補の生成方式

ここでは，Step-2 の実現方式について述べる．

Step-2.1 では，茶筅を利用して入力されたドキュメント・データを形態素に分割する．

Step-2.2 では，Step-2.1 の結果の品詞情報を参照して，名詞と未知語に分類される形態素を抽出する．ここで，名詞を地名の候補とする理由は，地名が名詞であること，ならびに，複数の形態素からなる地名からもランドマークの候補を生成するためである．これによって，地域を表す名詞から構成される地名からのランドマーク候補の生成を可能にする．また，未知語を地名の候補とする理由は，形態素解析器が使用する辞書に登録されていない地名からのランドマーク候補の生成を可能にするためである．

Step-2.3 では，Step-2.2 において抽出した形態素群から，ランドマークの住所名に出現する形態素を選択する．これによって，地名を表す形態素群を入力されたドキュメント・データから抽出する．

図 3 は，ドキュメント・データから地名を抽出する提案手法の実行例を表す．この図では，2 つの地名が抽出される．

3. 実　　験

本実験では，提案方式（内容認識を伴うジオ・コーディング方式）の妥当性を，ドキュメント・データからの正しい緯度経度の抽出率を検証することによって明らかにする．

3.1 実 験 方 法

本実験では，提案方式に複数のドキュメント・データを適用し，正しい緯度経度の抽出率から統計的に検証する．ここで，正しい緯度経度の抽出率（変換成功率）とは，全ドキュメント・データに出現する地名に対する正しい緯度経度に変換された地名の割合を表し，次式によって定義する．

$$\text{変換成功率} = \frac{\text{正しい緯度経度に変換された地名数}}{\text{全ドキュメント・データに出現する地名総数}}$$

提案方式の妥当性検証のために，実験データとして，同じ表現だが異なる緯度経度を表す地名を含む新聞記事を使用した．具体的には，米 4 大スポーツ（MLB, NBA，NFL，NHL）に関する新聞記事各リーグ 20 件（合計 80 件）を提案方式の適用対象とした．また，ランドマークとして，米 4 大スポーツに参加する全チームの本拠地スタジアムを使用した．その総数は 118 である．各チームの本拠地スタジアムに 'リーグ名'，'本拠地名'，'チーム名' を表す単語群をそのメタデータとして付与した．

3.2 実験結果と考察

ここでは，実験結果より，次の項目について考察することによって，提案方式の妥当性を明らかにする．

(1) ドキュメント・データベースへの提案方式の適用性
(2) 提案方式の適用範囲
(3) 位置情報にメタデータを埋め込む方式の優位性

表 1 実験結果
Table 1 The experimental result

	地名総数	変換が成功した地名数	変換成功率
MLB 新聞記事	48	39	81.3 %
NBA 新聞記事	36	35	97.2 %
NFL 新聞記事	47	34	72.3 %
NHL 新聞記事	29	28	96.6 %
合計	160	136	85.0 %

3.2.1 ドキュメント・データベースへの提案方式の適用性に関する考察

表 1 は，80 件の新聞記事に対する提案方式の変換成功率を表す．この結果より，提案方式によって，どの新聞記事に対しても高い変換成功率を得たことを確認した．

一方，入力されたドキュメント・データから正しい緯度経度に変換されなかった理由は次のとおりである．

理由-1 地名から正しい緯度経度に変換するための十分な情報が，ドキュメント・データに含まれていない．

例えば，「ヤンキースがニューヨークでメッツを破った.」というドキュメント・データは，地名から正しい緯度経度に変換するための情報を十分に含んでいない．具体的には，ヤンキースとメッツがともにニューヨークにある異なる球場を本拠地とするが，そのドキュメント・データには，ニューヨークが指す球場を特定するための情報を含んでいない．

そのため，提案方式では，そのドキュメント・データが両チームの本拠地球場に最も相関があることを計算することはできたが，両球場の相関量が同等であったため，1 つの球場に特定することができなかった．

理由-2 提案方式が，地名からランドマークへの変換時に利用する文脈を正確に認識することができなかった．

今回使用した実験データには，2 つのリーグについて記述しているドキュメント・データがあった．そのドキュメント・データの特徴は，地名を特定するための文脈が複数存在し，段落間でその文脈の変化があった点である．一方，提案方式におけるドキュメント・データの文脈認識の単位がドキュメント・データであるので，提案方式は，ドキュメント・データ内の段落間・文間での文脈の変化を認識できない．そのため，このようなドキュメント・データに対して，提案方式は，地名から正しいランドマークに変換することができなかった．

理由-3 ランドマークのメタデータが時間経過にしたがって変化する．

今回使用した新聞記事は，3 年間に渡って発行されたものである．その間に本拠地球場が移転したチームがあった．一方，提案方式では，時間経過に伴うランドマークのメタデータの変化を扱っていない．そのため，このようなドキュメント・データに対して，提案方式は地名から正しいランドマークに変換することができなかった．

しかし，全体として 85%という高い変換成功率が得られたことから，多くの場合において，ドキュメント・データベースへの提案方式の適用性があると判断できる．

3.2.2 提案方式の適用範囲に関する考察

地名から緯度経度への変換が失敗した理由の解決方法を検討することによって，人間のどのような位置認識プロセスを再現したかについて考察し，提案方式の適用範囲を明らかにする．

まず，理由-1 の解決方法について考える．この理由の解決方法は，ランドマークを特定するための十分な文脈を提案方式に与えることである．これによって，提案方式は，ドキュメント・データと同等の相関量を持つ 2 つのランドマークから，追加された文脈と高い相関があるランドマークを選択することが可能になる．理由-1 の状況は，米 4 大スポーツに詳しくない人が，ニューヨークが指すランドマークを確定できない状況に対応する．このとき，提案方式の解決方法と同様に，ニューヨークが指すランドマークを確定させるのに十分な文脈をその人に与えることによって，その人は，ニューヨークが指すランドマークを確定することができる．これより，提案方式は，人がドキュメント・データに含まれる文脈のみを認識して地名に対応する緯度経度を決定するプロセスを再現していることを明らかにした．

理由-2 を解決するためには，提案方式において，ドキュメント・データの文脈認識をその全体だけでなく，段落単位，および，文単位で行うための機構を実現することが必要である．この機構は，人間がドキュメント・データの全体の文脈だけでなく，ドキュメント・データ内の文脈の変化を認識して地名に対応する緯度経度を確定することに対応する．これより，理由-1 の考察と同様に，提案方式は，人がドキュメント・データに含まれる文脈のみを認識して地名に対応する緯度経度を決定するプロセスを再現していることを明らかにした．

理由-3 を解決するためには，ドキュメント・データ

に含まれる時間表現から対応する時間を抽出するための機能，ならびに，ドキュメント・データの時間に対応して，ランドマークのメタデータを選択する機能を実現することが必要である．これによって，地名から緯度経度への変換を時間経過に応じて実行することが可能になる．これより，提案方式は，時間経過に応じてランドマークのメタデータが変化しない状況において適用可能であることを明らかにした．

3.2.3 位置情報にメタデータを埋め込む方式（提案方式）の優位性に関する考察

ここでは，メタデータの更新の観点から提案方式を比較対象方式（ドキュメント・データが指す緯度経度を，ドキュメント・データに埋め込む方式[3),4)]）と比較することによって，提案方式の優位性を明らかにする．

提案方式において，メタデータの更新が必要となる場合は，ランドマークに地理的変化が生じた場合である．しかし，本拠地の移転のようなランドマークの地理的変化は緩やかであるので，多数のメタデータの更新要求が同時に起こることは希である．したがって，提案方式のメタデータの更新のオーバヘッドはそれほど高くない．

一方，比較対象方式では，ランドマークに地理的変化があった場合，そのランドマークを参照するドキュメント・データを更新しなければならない．ここで，ある 1 つのランドマークにおいて発生した出来事について複数の新聞記事が存在するように，1 つのランドマークを参照するドキュメント・データは，一般に複数存在する．したがって，比較対象方式におけるメタデータの更新に要するオーバヘッドは，提案方式のそれに比べて大きい．

これより，メタデータの追加・更新の観点から，比較対象方式との比較において，提案方式は有効である．

以上の考察より，提案方式の妥当性が明らかとなった．

4. 関 連 研 究

文献 9)，10) は，住所を対応する緯度経度に変換するための方式を示している．文献 9) では，県名や番地が省略されるような住所の部分を構成する地名を入力として受け取り，その地名と住所の部分一致を行うことによって，対応する緯度経度を検索する方式を示している．文献 10) は，与えられた住所を大規模な住所の集合から検索し，対応する緯度経度に変換するための分散検索方式を示している．

文献 5) は，ドキュメント・データから住所を自動的に抜き出すための手法を示している．この手法の特徴は，形態素解析を用いずに実現している点にある．形態素解析を用いる場合には，住所の切り出し精度が，形態素解析の辞書の作成方法によって左右されるが，この方式は，そのような辞書に依存しない形で実現している．

文献 8) は，Web ページに含まれる住所，郵便番号，電話番号，Web ページ間のリンク，ならびに，作成者によって付けられたメタデータから，ドキュメント・データが指す位置を特定するための手法を示している．

文献 2) は，Web ページに含まれる地理情報と Web ページ間のリンクから，地理空間における各 Web ページの有効範囲を計算するための手法を示している．

文献 7) は，Web ページの地域依存度を計算し，その依存度に応じて Web ページを検索するための方式について述べている．Web ページの地域依存度は，そのページに含まれる地名群を含む最小領域を求め，その領域の面積，ならびに，地名数に応じて計算される．

提案方式は，これらの方式と異なり，ドキュメント・データの内容認識を伴って対応する緯度経度を確定するための機構を実現する．

5. お わ り に

本稿では，ドキュメント・データを対象としたジオ・コーディング手法について述べた．ここで，ジオ・コーディングとは，地名から対応する緯度経度を算出する手法を表す．本手法の特徴は次の点にある．

特徴-1 ドキュメント・データにメタデータとして緯度経度データを埋めこむことなく，ドキュメント・データに含まれる地名の緯度経度を確定するための機構を実現する．

特徴-2 完全な形で記述されていない地名を，ドキュメント・データの内容認識を伴って対応する緯度経度を計算するための機構を実現する．

本手法によって，空間オペレータをドキュメント・データ検索機能として利用することが可能になる．さらに，実験によって本手法の有効性を明らかにした．

今後の課題は，提案方式の実際の数値地図への適用性評価，ドキュメント・データの内容認識技術の向上，ならびに，地図インタフェースを有するドキュメント・データ検索システムの実現が挙げられる．

参 考 文 献

1) 茶筅: http://chasen.aist-nara.ac.jp
2) Ding, J., Gravano, L., and Shivakumar, N.: Computing Geographical Scopes of Web Resources Proc. 26th Int'l Conf. on Very Large

Databases, pp.545–556 (2000)
3) Dublin Core Metadata Initiative: http://dublincore.org/
4) G-XML:http://gisclh.dpc.pr.jp/gxml/contents
5) 井ノ上直己, 平田育大, 米山正秀: Web テキストからの住所情報自動抽出手法, 情報処理学会第 65 回全国大会講演論文集 CD-ROM (2003)
6) Namazu: http://www.namazu.org/
7) 松本知弥子, 馬強, 田中克己: Web ページの地理情報と話題の日常性を考慮したローカル度検出とフィルタリング機構, 情報処理学会 データベースと Web 情報システムに関するシンポジウム (DBWeb2001) 予稿集, pp.193–200 (2001)
8) McCurley, K. S.: Geospatial Mapping and Navigation of the Web, WWW10 (2001)
9) 相良毅, 有川正俊, 坂内正夫: ジオリファレンス情報を用いた空間情報抽出システム, 情報処理学会論文誌:データベース, Vol.41, No.SIG 6 (TOD), pp.69–80 (2000)
10) 相良毅, 有川正俊, 坂内正夫: 分散位置参照サービス, 情報処理学会論文誌, Vol.42, No.12, pp.2928–2940 (2001)
11) 昭文社: MAPPLE デジタルデータ, http://www.mapple.co.jp/
12) 徳永健伸: 情報検索と言語処理, 東京大学出版会 (1999)