Wikipediaを対象とした地理情報と時間情報の抽出手法の提案

岡本　章裕† 　黒井　星良† 　横山　昌平† 　福田　直樹† 　石川　博†

† 静岡大学情報学部情報科学科　〒 432–8011 静岡県浜松市中区城北 3–5–1
E-mail: †{cs05019,cs05035}@s.inf.shizuoka.ac.jp, ††{yokoyama,fukuta,ishikawa}@inf.shizuoka.ac.jp

あらまし　Web 上の地理情報は，日付や時刻などの他の情報と有機的に結びついていない場合が多く，それらの関係性を考慮した抽出と効果的な利用のための技術が必要になると考えられる．本研究では Wikipedia 上の地理情報とその地理情報に関連のある時間に関する情報の抽出を試みる．Wikipedia におけるテンプレートや記事間のリンク構造などの特徴を生かした抽出方法を提案し，抽出アプリケーションの開発を行う．また，抽出した情報を GIS アプリケーションに適用した例と記事間の地理的・時間的距離の関係を視覚的に表現するツールを示す．

キーワード　Wikipedia，GIS，情報抽出

Proposal of Extraction Technique of Geographic Information and Time Information form Wikipedia

Akihiro OKAMOTO†, Seira KUROI†, Shohei YOKOYAMA†, Naoki FUKUTA†, and Hiroshi ISHIKAWA†

† Department of Computer Science, Faculty of Informatics, Shizuoka University　Johoku 4–5–6, Nakaku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
E-mail: †{cs05019,cs05035}@s.inf.shizuoka.ac.jp, ††{yokoyama,fukuta,ishikawa}@inf.shizuoka.ac.jp

Abstract　There is a rapid growth of geographic information services on the Web Geographic information on the Web are not often directly related to other information such as date and time. There are demands to realize better extraction of relations among such geographical information and other useful information on the Web. In this paper,we propose our preliminary approach to extract information relating to geographic information and temporal information related to them from Wikipedia. Furthermore, we show a GIS application and a visualizer for relationships among geographic-time distance and articles.

Key words　Wikipedia, GIS, Information extraction

1. はじめに

Web 上には多種多様な莫大な量の情報が溢れている．その中には，地理情報システム (Geographic Information System:GIS) の普及に伴って，Web 上では地理に関する情報やそれらを利用したサービスが数多く提供されている．たとえば，Google Maps [1] などの地図情報サービスでは，地図を表示するだけにとどまらず，ユーザが地図を編集できたり，地図上に広告を掲載するサービスを提供するなどしている．このような Web における地理的な情報の発信や利用および，それらを生かしたサービスの需要が増えている．しかし，多くの Web ページは自然言語で書かれており，地理情報と日付や時刻といった他の情報は有機的に結びついていない場合が多い．そこで，地理情報とさらに別の情報を組み合わせて利用可能とする技術が必要であると考えられる．

本研究では，地理情報と時間に関する情報（年月日）の組み合わせに着目した．たとえば，地震の震源地と発生日，空港の位置と開港日，といった関係である．本研究では，このような情報の組を Wikipedia [2] 上の記事から抽出する．Wikipedia は Web 上の百科事典という位置づけにとどまらず，情報処理の分野では，Wikipedia を対象とした様々な研究が近年さかんに行われている [3]．Wikipedia は，第 2. 章で示すような他の一般的な Web ページにはない様々な特徴があり，本研究をはじめ Wikipedia を対象とした他の研究もそのような特徴に着目している．

本研究では抽出した情報に対する 2 つの利用方法を提案する．1 つは，各記事の抽出した地図座標と時間に関する情報をグラフに示す機能のツール化である．記事間の地理的・時間的距離

から記事間の関係の視覚化を行う．もう 1 つは，抽出した情報と，その情報に関連する Wikipedia 上のデータを，XML をはじめとする再利用が容易な形式として提供することである．本機能を Web サービスとして実現し，他の Web アプリケーションで利用可能とすることで，それらのアプリケーションの機能向上に利用できることを示す．

2. Wikipedia の特徴

Wikipedia は，Wiki を利用したオンライン百科事典である．誰でも無料で Web ブラウザから閲覧や編集ができることが大きな特徴である．また，記事は非常に幅広い分野を網羅しており，新しい概念や出来事に対してもユーザーが記事を次々に作成しているため，記事数は増え続けており，2008 年 12 月 1 日時点での記事数は 2,641,901 件（英語版），540,701 件（日本語版）と膨大である．

Wikipedia は，記事の膨大さ以外にも様々な特徴を持つ．本研究で特に着目した特徴には以下のようなものがあげられる．

まず，テンプレートという機能がある．Wikipedia におけるテンプレートとは，定型文の入力を簡便にするために用いられる仕組みである．テンプレートの 1 つに，coord テンプレートがある．これは座標表現（緯度，経度）で場所を示したいときに使われるものである．本研究では，coord テンプレートなどを用いた座標表現をジオタグと呼ぶ．ジオタグの例を図 1（枠で囲われた部分）に示す．このようなジオタグを抽出することで，地理情報を抽出するという目的を達成する．

ジオタグに加えて，infobox テンプレートにも着目する．infobox テンプレートとは，同じジャンルに対して共通の項目を表形式にして各記事に記載している形式のテンプレートである．Wikipedia の中でも特に構造化された形式で書かれている部分であり，Wikipedia からの情報抽出を試みる他の研究においても infobox テンプレートに注目したものがいくつかある．infobox テンプレートの例を図 1 に示す．infobox テンプレートは，1 つのテンプレート名と，複数のテンプレート変数と引数の組から構成されている．図 1 の例では，テンプレート名が earthquake，テンプレート変数に地震の名前，マグニチュードなどがあり，引数には各地震ごとの値が入る．具体的には，
name=Iwate-Miyagi Nairiku Earthquake in 2008
magnitude=7.2
などとなる．

また，Wikipedia は，英語のみではなく日本語やその他 200 を超える言語版がある．本研究は英語版 Wikipedia を対象として地理情報と関連する時間に関する情報の抽出を行っているが，英語以外の言語版への応用が可能となるような，言語に依存しない要素を考慮した抽出手法を提案する．

さらに，Wikipedia の記事は適切な見出しが与えられ，他の多くの Web ページに比べて文書が構造化しているという特徴もある．本研究では，見出しの位置を考慮することで，地理情報に関連した時間に関する情報抽出の精度向上を目指す．

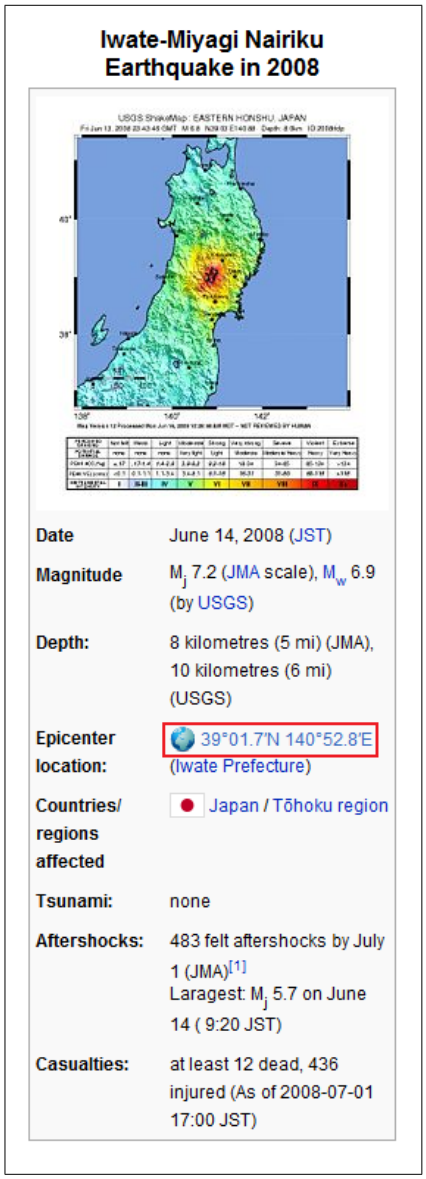

図 1　infobox テンプレートとジオタグの例

3. 本研究の提案手法

本研究では，Wikipedia から地理的な情報と関連する時間に関する情報を抽出することを試みる．このような場合，一般的には自然言語処理を用いた方法が頻繁に用いられる．しかし，本研究では，自然言語処理は用いずに，Wikipedia の構造を手がかりとしてパターンマッチングにより必要な情報を抽出することを考える．これには，構造のみに着目した手法を用いることにより，言語に依存しないというメリットがある．

本研究では，1 つのものや出来事と 1 つの Wikipedia 記事が対応していると考えて，各記事において最も重要と考えられる地理情報と関連する時間に関する情報の組 1 つを抽出することを考える．たとえば，空港の記事において，開港日と利用者数が 1,000 万人を超えた日の 2 つの時間に関する情報がある場合，より空港の位置という地理情報に関連があるのは前者の開港日である．そこで，限られた範囲ではあるが関連の強い地理情報と時間に関する情報が抽出できると期待される infobox テンプ

レートからの抽出と，記事のすべての部分を対象とした本文中からの抽出手法の提案を行う．

3.1 infobox テンプレートからの抽出

本研究の目的である関連のある地理情報と時間に関する情報の組を抽出するために，第 2. 章の Wikipedia の特徴のひとつである infobox テンプレートを利用することを考えた．ひとつの infobox テンプレート中に地理情報と時間に関する情報があれば，その 2 つの情報には関係があり，どのような関係を持つかは，テンプレート名やテンプレート変数から推測できる．そこで，各記事からテンプレート名，テンプレート変数と引数を取り出し，取り出された引数から地理情報と時間に関する情報を抽出することで，関連のある地理情報と時間に関する情報の組を得ることを考える．この手法では，テンプレート中に地理情報と時間に関する情報の両方が記載されていなければ抽出できないが，その条件に該当する場合であれば，非常に高い精度で関連のある情報の抽出が行えると考えられる．さらに，抽出した地理情報と時間に関する情報に加えて，他の関連する情報（図 1 の地震の例ではマグニチュードなど）も容易に抽出することができる．このような情報も共に抽出することで，他のアプリケーションに応用した場合の機能向上に役立つことが期待される．

3.2 記事の本文からの抽出

3.1 節の手法が適用できるような，1 つの infobox テンプレート中にジオタグと時間に関する情報が含まれる場合は限られる．そこで，infobox テンプレート以外の本文中からジオタグと関連する時間に関する情報を抽出することが必要になる．1 つの記事中には時間に関する情報が複数ある場合が多く，どの時間に関する情報がジオタグと関連があるのかを判別することが課題である．本研究では，本文中のジオタグと時間に関する情報の距離と，Wiki 記法における見出しから関連のある 2 つの情報を抽出する手法を提案する．

本手法では，記事の上位にあるものほど重要度が高く，また，ジオタグと時間に関する情報の本文中での距離が短いほど関連が強いという 2 つの仮定を元にアルゴリズムの構築を行った．しかし，本文中での距離が短い場合でも，それぞれの情報が異なる見出しの下にある場合では関連度が低くなると考えられ，見出しの位置を取得し，それを考慮する必要がある．そこで，記事本文からの抽出では，ジオタグと時間に関する情報のそれぞれの本文中での位置，および，ジオタグのある記事の見出し位置の取得が必要である．

本手法では Wikipedia のレベル 2 およびレベル 3 の見出し（レベル 1 の見出しは本文中にはないため）を取得した．ただし，3.3 節ではどちらのレベルの見出しも同じ条件で扱うことにする．

3.3 ジオタグと時間に関する情報の組み合わせ

本手法では，ジオタグと時間に関する情報が共に 1 つ以上ある記事について，それらの情報の組み合わせを抽出する．その組み合わせは，各記事について 1 組とする．そこで，3.1 節，3.2 節の手法で抽出したジオタグと時間に関する情報から，各記事について関連性が最も高い 1 組のペアを抽出するための手法を以下に示す．

まず，ジオタグ，または，時間に関する情報のいずれかが抽出できなかった記事は該当なしとする．また，ジオタグと時間に関する情報が 1 つずつ存在する場合は，その組み合わせが該当するペアとする．

ジオタグが複数ある場合は，infobox テンプレート内がもっとも重要度が高く（infobox テンプレート内に複数ある場合は，最も上位にあるものを優先），次に，テンプレート外のうち，記事のより上位にあるものが重要度が高いと考え，このような優先度でジオタグを 1 つに決定する．

時間に関する情報は，ジオタグとの本文中での距離を用いる．ジオタグと時間に関する情報が同じ infobox テンプレート中に共にある場合には，最も上位にある時間に関する情報を選択する．ジオタグが本文中にあり，時間に関する情報が infobox テンプレート中にある場合には，infobox テンプレートの項目のより上位にある時間に関する情報を選択する．ジオタグが本文中にあり，時間に関する情報が infobox テンプレート中になく本文中にのみにある場合には，第一に同じ見出しの中で距離が近いもの，それに該当しなければ見出しを考慮せず，単に距離が近いものを選択する．ジオタグが infobox テンプレート中にあり，時間に関する情報が本文中のみにある場合には，時間に関する情報は本文中で最も上位のものを選択する．

また，距離が同じ値になった場合には，より記事の上位にあるもの（infobox テンプレートならば，より項目の並び順で上位のもの）を選択するものとする．

以上の手法により，1 記事につき 1 組の地理情報と時間に関する情報の組み合わせを抽出する．

4. 実　　験

4.1 目　　的

Wikipedia から関連のあるジオタグと時間に関する情報の組を抽出する．抽出の対象とするデータは Wikipedia（英語版）の全ダンプである．

infobox テンプレート内での抽出では，関連が高い地理情報と時間に関する情報の組が得られるが，ひとつの infobox テンプレート内にそれら 2 つの情報がある場合は限られる．そこで，infobox テンプレート内での抽出の有効性を確かめることを目的とし，実際に抽出を行うと共に，適用可能なジオタグが Wikipedia 中の全ジオタグに占める割合を算出する．

さらに，テンプレート以外の本文中からの抽出も試みる．本文中からの抽出では，同一テンプレート内での抽出より範囲が広く，多くの情報が抽出できると考えられるが，本文中に複数の時間に関する情報がある場合が多く，関連が高い時間に関する情報を抽出することが難しい．そこで，提案する手法で実際に抽出を行い，その精度を確かめることを目的とする．

4.2 抽出プログラムの実装

Wikipedia のダンプは XML 形式で各記事の最新版のタイトルや本文がテキストデータとして格納されている．本論文で使用したダンプは 2008 年 12 月に取得した．本研究では，抽出プログラムを，XML を読み取る部分（XML パーサー）と，本文

の Wiki 記法を解析する部分から構成した．

抽出プロセスは以下の段階に分けて実装を行った．

まず，全記事の本文中からジオタグを抽出し，ジオタグのある記事名をデータベースに格納する．また，infobox テンプレートからの抽出のために，各 infobox テンプレートを定義している記事から，テンプレート名とテンプレート変数を取得し，本文からの抽出のために，ジオタグのある記事における見出しの本文中での位置を取得する．ジオタグのある記事のテンプレートを抽出して，それが infobox テンプレートであり，かつそのテンプレート変数が定義されたものである場合，その引数をデータベースに格納し，テンプレート以外の本文からジオタグと時間に関する情報をその本文中での位置と共に抽出する．最後に，抽出した情報の中から 3.3 節の手法に基づき，関連のあるジオタグと時間に関する情報の組を取り出す．

4.3 抽出結果

実装を行った抽出プログラムで実際に抽出を行った結果を以下に示す．最初に，infobox テンプレート内での抽出と本文中からの抽出結果をそれぞれ示す．

infobox テンプレート中に存在するジオタグと時間に関する情報の抽出結果を表 1，2 に示す．表 1 では，同一テンプレートからの抽出実験で着目したジオタグと infobox テンプレートが Wikipedia 中でどの程度出現しているのかを示している．ジオタグや infobox テンプレートが広く利用されていることが分かる．

表 2 では，ジオタグの出現する記事に infobox テンプレートがあった場合に，その infobox テンプレート内の引数を抽出した結果を示している．表 2 中の「ジオタグと時間に関する情報のある記事」が，同一テンプレート内での抽出で求める地理情報と関連のある時間的な情報の組み合わせの抽出が行えたものである．提案した infobox テンプレートからの抽出手法により Wikipedia 中のジオタグのある記事 0.57%から，関連する時間的な情報と共に抽出できた．

次に，表 3 では，抽出範囲を広げジオタグのある記事の本文中からジオタグと時間に関する情報を抽出した結果を示す．2,641,901 記事中 61,128 記事から，ジオタグと時間に関する情報の組を抽出できた．

4.4 抽出結果の検討

本手法では，地理情報を収集するのに Wikipedia のジオタグを用いた．Wikipedia は地理情報に特化した百科事典ではないが，ジオタグは表 1 の通り Wikipedia 中に多く使用されており，地理情報を収集する手段として有効であると考えられる．また，2008 年 7 月に取得したダンプに対して同じ手法を用いてジオタグを抽出した結果を表 4 に示す．数ヶ月間の間にジオタグが増えていることが分かる．このように，ジオタグが Wikipedia で地理情報を示すのに有効な手段として広く利用されており，その利用はますます増加すると考えられる．

ジオタグはテンプレートとして形式が定まっているため抽出は容易であるが，時間に関する情報（年月日）はさまざまな形式があるため，抽出が難しい．本研究では英語版 Wikipedia で用いられている標準的と考えられる 5 種類の記法を正規表現を用いて抽出した．用いた記法を図 2 に示す．しかし，同じ年の日付が連続する場合では年を省略するなどの記法が多くあり，本研究で用いたような年月日がすべて記載されているパターンだけでは，すべての時間に関する情報を抽出するのは不可能であることがわかった．表 3 で示すとおり，時間に関する情報を抽出できたのはジオタグのある記事のうち 24.8%のみであった．残りの記事は時間に関する情報がまったくない記事が含まれていると考えられるが，抽出ができなかった記事も相当数あることが見込まれ，時間に関する情報の抽出を改善することが，地理情報と時間に関する情報を組み合わせるという本手法の性能向上のための重要な課題だと考えられる．

表 1　抽出記事数（infobox テンプレート）

対象	記事数	全記事数に対する割合
全記事数	2,641,901	——
ジオタグが含まれる記事数	246,561	0.093
ジオタグと infobox テンプレートの両方が含まれる記事数	103,677	0.039

表 2　ジオタグのある記事の infobox テンプレートの引数からの抽出結果

対象	記事数
ジオタグのある記事数	17,276
時間に関する情報のある記事数	7,069
ジオタグと時間に関する情報のある記事数	1,404

表 3　地理，時間情報の抽出記事数（本文および infobox テンプレート）

対象	抽出できた件数	記事数
ジオタグのある記事数	302,345	246,561
時間に関する情報のある記事数	190,640	61,128

本手法では，本文中からジオタグと関連のある時間情報を結びつけるのに，本文中からの距離を用いた．それに加え，本手法では見出しの位置にも着目した．本文中の距離のみであると，2 つの情報が見出しを超えてしまう場合があり，見出しを超えて近いものより，同じ見出し内の情報の方が関連が高いと考えられるからである．今回の抽出では，328 記事においてこの手法に該当するパターンがあった．これはジオタグと時間に関する情報が両方抽出できた記事の 0.54%にあたる．原因として，ジオタグは段落内の本文中には少なく，記事の上下の端にある場合が多いことと，前述の通り時間に関する情報の抽出が不十分なため，ジオタグと同じ段落の時間に関する情報が抽出できていないことがあるため，ごく少数の記事でしかこの手法が適用できなかったと考えられる．

5. 英語版以外の言語の記事への適用

本手法の特徴のひとつに，他の言語の記事に簡単に適用できるということがある．Wikipedia は最大の記事数を有する英語版をはじめ 200 以上の言語版があり，世界中で利用されている．

Wikipedia には同じ出来事やもの，概念を表す記事について各言語の記事を結ぶ言語間リンクがある．たとえば日本語版の

表 4　異なる時期に取得した Wikipedia ダンプの比較

取得時期	全記事数	ダンプの大きさ	ジオタグ の件数	ジオタグの 含まれる記事数
2008 年 7 月	2,435,638	33.00GB	178,272	132,214
2008 年 12 月	2,641,901	37.52GB	302,345	246,561
増加割合（%）	108.5	113.7	169.6	186.5

a) [[1986]] [[July 15]]
b) [[1986]] [[15 July]]
c) 1986-7-15
d) July 15, 1986
e) 15 July 1986
※すべて1986年7月15日を示したもの
※[[　　]]で囲まれた部分はリンクになる

図 2　本手法で抽出の対象とした Wikipedia の時間に関する情報（年月日）の記法

「岩手・宮城内陸地震」からは，英語版の「2008 Iwate-Miyagi Nairiku earthquake」，ハングル，中国語版など 7 つの言語の記事にリンクされており（2009 年 2 月現在），リンク先には各言語の記事での岩手・宮城内陸地震について書かれた記事がある．各記事から地理情報と時間情報を抽出する場合，たとえば地震の場合は震源地と発生日であるが，それらはどの言語の記事についても同一のことが記載されているはずであり，複数の言語の記事から抽出を行う意味はないと考えられる．しかし，英語版や他の言語の記事にはないが日本語版では作成されているという，各言語においてのローカルな記事が多くあるのも事実である．ジオタグのある記事でかつ他の言語の記事にないものの例として「イオン浜松市野ショッピングセンター」などの記事がある．このような記事を抽出するためには，各言語の記事ごとに抽出を行う必要がある．また，同じ出来事を示す記事でも，ある言語版ではうまく抽出できないが，他の言語の記事では抽出できる場合があり，他の言語版を抽出することで補完できることがあると考えられる．そこで，本手法を日本語版に適用し，その過程と結果，英語版との比較を示す．

5.1　日本語版への適用

本手法のアルゴリズムそのものは言語版に依存しないが，ジオタグや時間に関する情報の抽出は各言語に依存する．英語版においてもジオタグにはいくつか種類があるが，日本語版のみに使われているジオタグ（「日本の位置情報」など）があり，それらを抽出対象に加える必要がある．時間に関する情報は英語版 Wikipedia での抽出と同様に，Wikipedia 記事を閲覧して，年月日のパターンを探し分類して，正規表現を構成する．本手法で抽出の対象とした年月日のパターンを図 3 に示す．これらを実際に行い，英語版の抽出プログラムを改編し実装した．

5.2　日本語版での抽出結果と英語版との比較

同時期（2008 年 12 月）に取得した Wikipedia の英語版と日本語版で比較を行う．表 5，6，7 に結果を示す．表 5 からは，英語版と日本語版の規模の差が記事数で 5 倍程度あることがわかる．表 6 からは，ジオタグは日本語版よりも英語版で積極的に使われていることがわかる．表 7 からは，時間に関する情報の抽出の程度がわかる．抽出した範囲での 1 記事あたりの時間に関する情報は，日本語では 4.3 件，英語では 3.1 件とやや差があるが，抽出の対象となるジオタグのある記事のうち約 25%の記事から抽出ができた点ではほぼ同じであった．しかし，全記事数との割合で見ると英語版 2.31%，日本語版 0.57%と大きく差がある．つまり，時間に関する情報の取得には言語間の差は大きくないが，ジオタグの量の差が大きいので，最終的な抽出結果数に大きく影響しているといえる．

表 5　日本語版と英語版との比較（全記事数）

言語	全記事数	ダンプの大きさ
日本語	540,701	4.24GB
英語	2,641,901	37.52GB

表 6　日本語版と英語版との比較（ジオタグ）

言語	ジオタグの件数	ジオタグが 含まれる記事数	全記事数に 対する割合
日本語	9,014	7,776	0.014
英語	302,345	246,561	0.093

表 7　日本語版と英語版との比較（ジオタグのある記事の時間に関する情報）

言語	時間に関する情報が 含まれる件数	時間に関する情報が 含まれる記事数
日本語	13,366	3,108
英語	190,640	61,128

a) [[1986年]][[7月15日]]
b) [[1986年]]（昭和61年）[[7月15日]]
c) 1986年7月15日
d) 1986年（昭和61年）7月15日
e) 1986-7-15
f) 1986/7/15
※すべて1986年7月15日を示したもの
※[[　　]]で囲まれた部分はリンクになる

図 3　本手法で抽出の対象とした Wikipedia 日本語版の時間に関する情報（年月日）の記法

6.　抽出したデータの利用

本手法の応用事例として，抽出した緯度，経度，時間の 3 次元の情報から各記事の地理的・時間的距離を視覚的に示すツールと GIS アプリケーションに適用した例を示す．

6.1　Wikipedia 記事の可視化

Wikipedia の記事間の関係をグラフなどに示して視覚化をする研究がいくつか行われている．中山らは [4]，Wikipedia から抽出した連想関係辞書を利用して，どの概念とどの概念が関連が強いかを可視化して表示するアプリケーションである Wikipedia シソーラスビジュアライザーを開発している．新井ら [5] は，Wikipedia の言語間リンクに着目して，翻訳語やカテゴリの可視化を行っている．

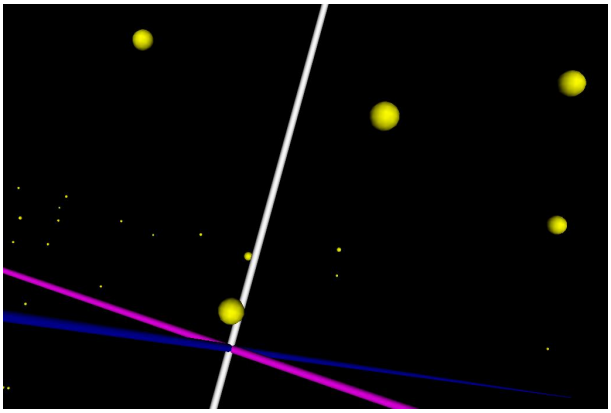

図 4　3 次元散布図スクリーンショット（原点拡大）

本研究では各記事における地理情報と時間に関する情報から，各記事の地理的・時間的距離を視覚的に示すツールを作成する．

6.1.1　3 次元散布図の試作

3 次元散布図で Wikipedia 記事を可視化したスクリーンショットを図 4，5 に示す．緯度，経度，時間の 3 次元座標に，各記事をプロットし散布図を作成した．Wikipedia 記事の可視化機能の実装には，Web3D Consortium [6] が開発を推進している VRML という，Web ブラウザ上で 3 次元物体を表現できるファイルフォーマットを用いた．ブラウザ上で，視点の移動や拡大，縮小，表示されている 3 次元の物体をマウスでクリックする動作などが可能である．図 4 に散布図の原点付近の拡大を示す．青い軸が緯度，赤い軸が経度，白い軸が時間を示し，黄色の球が各記事を表している．この可視化ツールでは原点は緯度，経度が共にゼロで，時間軸が 2001 年 1 月 1 日の地点とした．図 5 に散布図の全体図を示す．黄色い球が密であるところや疎である部分が確認できる．このように，Wikipedia の記事間の地理的，時間的な関係を視覚化できるようなツールを作成した．

6.1.2　Google Earth による可視化

本研究での抽出結果をわかりやすく示し，利用するために，アプリケーションを用いて可視化することを考え，Google Earth [7] を用いて実現を試みた．Google Earth を利用して Wikipedia 記事を可視化したスクリーンショットを図 6，7 に示す．Google Earth は Google 社が提供している，世界中の衛星画像をまるで地球儀を回しているかのように閲覧できるアプリケーションである．Google Earth では，3 次元地理空間情報の表示を管理するための XML ベースのマークアップ言語である KML（Keyhole Markup Language）を用いることにより，ユーザが任意の地点に名前を付けるなどの拡張を行うことができる．本研究での Google Earth による可視化では，データベースからデータを読み出し，それらのデータから KML を動的に生成し，それを Google Earth に読み込ませることで可視化を実現した．図 8 に本研究で生成した KML の一部を示す．

緯度・経度からの Wikipedia 記事へのリンク機能は，Google Earth に元から備わっている．本研究ではそれに時間の要素を加え，図 6 のように時間をアイコンの高さ（上にあるものほど新しい）で示した．また，図 7 のように「同じ年月の記事のみを表示」などの関連のある Wikipedia 記事の閲覧を補助する機能を実装した．さらに，Google Earth 5.0 より過去の衛星画像の閲覧機能が追加され，生成する KML に本研究で抽出して得られた時間情報を付加し，ビルの工事の様子，空港の整備の様子などの衛星画像のアニメーション表示が可能となった．

図 7　Google Earth による可視化（閲覧補助機能）

このように，Google Earth を用いて抽出結果を可視化し，Wikipedia 記事や Google Earth の衛星画像の閲覧を補助するツールの実装を行った．

6.2　GIS アプリケーションへの適用

本手法の応用事例として，当研究室で開発している Web GIS アプリケーションフレームワークである rinzo.ma [8] での使用例を示す．

本事例では，地震に関する情報を抽出したものを rinzo.ma のプラグインに適用した．地震の場合の地理的な情報は震源地の緯度経度であり，時間的な情報は発生日である．それらに加えて，地震名と地震に関する情報のうち infobox テンプレートから抽出できるマグニチュード，津波の有無，被害状況なども追加した．これらの情報を加えることで，ユーザーがマグニチュードの大きい地震のみを閲覧したり，簡単な被害の状況が他の Web ページを参照しなくても確認することができる．データ形式は rinzo.ma が JavaScript で開発されているため，JSON（JavaScript Object Notation）[9] を用いた．JSON は JavaScript におけるオブジェクトの表記法をベースとした軽量のデータ交換フォーマットであり，特に本事例のような JavaScript で実装されたウェブアプリケーションとのデータの受け渡しなどに利用される．rinzo.ma で利用するために出力した JSON 形式のデータの一部を図 9 に示す．実際にこの JSON

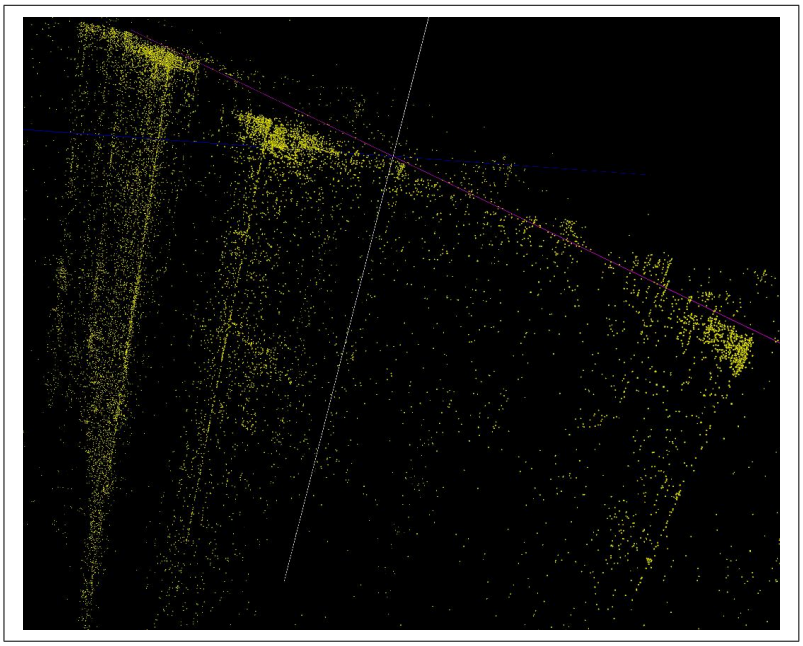

図 5　3 次元散布図スクリーンショット（全体）

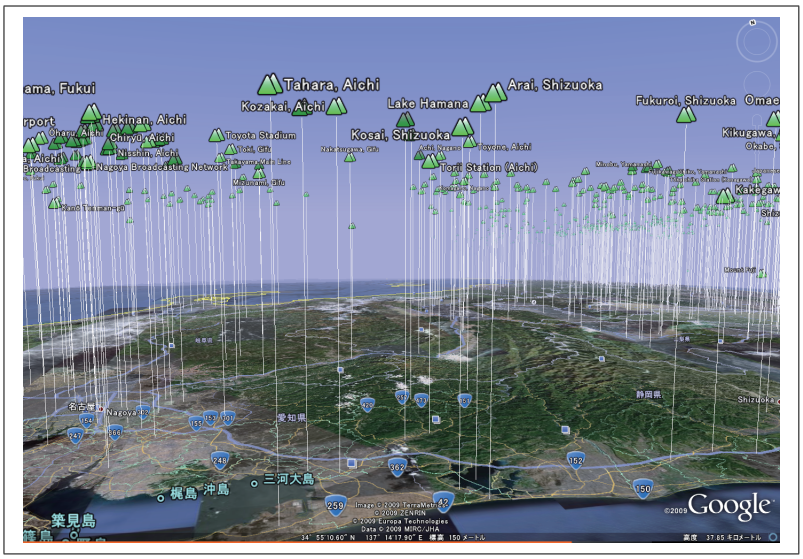

図 6　Google Earth による可視化（アイコン）

形式のデータを用いることにより rinzo.ma では，震源地の緯度経度と時間からその地震の被害を示すような衛星画像を検索し提示することが可能となった．

7. 関 連 研 究

Web 上にある位置情報を抽出する研究としては，相良らの空間情報抽出システムがある [10]．このシステムでは，Web ページから地名や住所などの地理情報に関する記述を抽出し，アドレスマッチング手法により抜き出した地理情報記述を緯度経度表現に変換している．さらに，これらの情報を地図上にリンクするアプリケーションの開発を行っている．Web 上の地理情報を抽出して処理を行い，アプリケーションに応用するといった目的は同じであるが，本研究では直接，緯度経度情報の取得を行うため，地名から緯度経度を求めることは行わない．さらに，本研究では地理情報に加えて，時間に関する情報も抽出している点で異なっている．

Wikipedia は様々な研究分野で用いられている．そのひとつに，中山ら [11] の研究がある．中山らは Web オントロジの構築を目的にしており，本研究の目的とは異なるが，Wikipedia の記事の膨大さや，記事間のリンクの多さなど Wikipedia の特徴に着目したという点では共通である．Wikipedia から Web オントロジの構築を行う研究は他にも DBpedia [12] などいくつかある．特に，DBpedia は本研究で着目した infobox テンプレートなどの半構造情報を解析することで，人物などの属性を抽出し，RDF に変換し大規模なデータベースを構築し Web で公開している [13]．このように Web 全体を対象とするわけでなく，特に Wikipedia に着目し，情報や知識を取り出し活用しようとする研究が盛んに行われている．

Google Maps や Google Earth では，Wikipedia の記事を地図や衛星画像上にリンクするというサービスを行っている．本

```
<kml>
 <Folder>
  <name>GROUP:2008(1)</name>
  <Placemark>
   <description>
    ID:190124<br>
    <a href=
    'http://en.wikipedia.org/wiki/The Tokyo Towers'>
    The Tokyo Towers</a><br>
    2008年1月30日<br>
     ・・・・・
    <a href='http://localhost/make_kml.php?
    year=2008&month=1'>
    同じ年月の記事をみる</a><br>
   </description>
   <name>The Tokyo Towers</name>
   <visibility>1</visibility>
   <TimeSpan>
    <begin>2003</begin>
    <end>2013</end>
   </TimeSpan>
   <Style>
    <IconStyle>
     <Icon>
      <href>palette-2.png</href>
     </Icon>
    </IconStyle>
   </Style>
   <Point>
    <extrude>1</extrude>
    <altitudeMode>
     relativeToGround
    </altitudeMode>
    <coordinates>
     -69.3333333333,48.0166666667,1
    </coordinates>
   </Point>
  </Placemark>
  <Placemark>
     ・・・・・
  </Placemark>
 </Folder>
</kml>
```

図 8　出力した KML 形式のデータの一部

```
{"earthquake":
 [
  {"name":"2005 Ruichang earthquake",
   "data":"2005/11/28",
   "lat":"+29.657","lon":"+115.717",
   "magnitude":"5.2","tsunami":"none",
   "casualties":"at least 14 killed"},
  {"name":"2005 Hindu Kush earthquake",
   "data":"2005/12/12",
   "lat":"+36.332","lon":"+71.130",
   "magnitude":"6.7","tsunami":"Not_Found",
   "casualties":"5 dead"}
 ]
}
```

図 9　出力した JSON 形式のデータの一部

研究とは，地理情報のみではなくさらに時間という概念を加える点において違いがある．

8. おわりに

Wikipedia から地理的・時間的情報を抽出手法の提案と，さらに抽出したデータの応用を示した．また，提案した手法を実装し実際に抽出を行い，その結果を示した．

課題としては，抽出結果の評価，また 1 つの記事に時間情報が複数ある場合の抽出があげられる．抽出結果が膨大であり，また正解とも不正解ともとれる結果が多くあり，効果的な評価方法が定まっていない．評価を行い，その結果からさらに抽出精度を上げる方法の検討をしていきたい．また抽出結果から，1 つの記事に複数の時間情報が含まれる場合が多くあることがわかる．たとえば，地震の記事であれば発生日以外にも余震のあった日や鉄道が再開した日などがある．1 つの地理情報に複数の時間情報を結びつけることが出来れば，より応用範囲が広がると考えられる．

さらに今後の発展として，抽出範囲の拡大と，地理情報と時間に関する情報以外の抽出の可能性があげられる．

本研究の提案は Wikipedia に特化した手法であり，Wikipedia 以外の Web ページから情報を抽出することを想定していない．Wikipedia 以外の Web ページからも情報を抽出することで，抽出できる情報の量の増加が見込まれる．しかし，Web ページの構造や品質は多様であり，有用な情報の効率的な収集には多くの課題がある．本研究では抽出範囲を Wikipedia に限定することで，これらの課題を解決した．

また，本研究では地理情報と時間に関する情報の組み合わせに着目した．これ以外にも地理情報と別の情報との組み合わせ，または地理情報以外の情報の組み合わせなどが考えられる．このような，本研究が扱わなかった情報の組み合わせを抽出し，他の情報やアプリケーションと組み合わせることで，新たな応用が可能となるのではないかと考えられる．

謝辞　本研究の一部は科学研究費補助金基盤研究（B）（課題番号 19300026），基盤研究（A）（課題番号 20240010），若手研究（B）（課題番号 20700104）の助成による．

文　献

[1] Google Maps
http://maps.google.com/
[2] Wikipedia
http://www.wikipedia.org/
[3] 中山浩太郎, 原 隆浩, 西尾章治郎, “ 人工知能研究の新しいフロンティア：Wikipedia ”, 人工知能学会誌, Vol. 22, No. 5 (2007).
[4] Nakayama, K., Pei, M., Erdmann, M., Ito, M., Shirakawa, M., Hara, T. and Nishio, S.“ Wikipedia Mining - Wikipedia as a Corpus for Knowledge Extraction - ”, in Proceedings of Annual Wikipedia Conference (Wikimania) (2008).
[5] 新井嘉章, 福原知宏, 増田英孝, 中川裕志, “ Wikipedia の言語間リンクに関する分析 ”, 人工知能学会第 22 回全国大会 (JSAI 2008).
[6] Web3D Consortium
http://www.web3d.org/
[7] Google Earth
http://earth.google.com/
[8] rinzo.ma
http://rinzo.ma/
[9] JSON
http://json.org/
[10] 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 7) pp.69-80 (2000).
[11] 中山浩太郎, 原隆浩, 西尾章治郎, “ 自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築 ”, 日本データベース学会論文誌 Vol.7, No.1, pp.67-72 (2008).
[12] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. ”DBpedia: A Nucleus for a Web of Open Data”, International Semantic Web Conference (ISWC2007), pp. 722-735 (2007)
[13] DBpedia
http://dbpedia.org/