

地域情報関連テキストを対象とした 地域状況表示地図の動的生成方式

守屋 敬太[†] 佐々木史織[‡] 清木康[†]

[†] 慶應義塾大学環境情報学部 〒252-0011 神奈川県藤沢市遠藤 5322

[‡] 慶應義塾大学政策・メディア研究科 〒252-0011 神奈川県藤沢市遠藤 5322

E-mail: [†] {t06884km, kiyoki}@sfc.keio.ac.jp, [‡] {sashiori}@sfc.keio.ac.jp

あらまし 本稿では、住所情報を含む地域情報関連テキストを対象として、ユーザの与えるコンテキストに応じた「地域状況表示地図」を動的に生成するシステムの実現方式を示す。本方式は、コミュニティベースで蓄積された住所情報を含む文書の集合を、ユーザの与える任意のコンテキスト（任意のキーワードセット）によって評価し、その結果を画像として地図上に重ね合わせることで、ある場所とその周囲環境の空間的配置に関してコミュニティが持つイメージ・印象・雰囲気、地域状況表示地図として可視化する。本方式により、ある土地に関する情報をテキストや単語の観点から一望することが可能となり、また、その評価結果を地図上に視覚的に表現することにより、地理情報に応じた情報の把握や、その応用を可能にする。本稿では、ユーザの与えるコンテキストを場所に関する俗称・印象・雰囲気として設定し、実験システムを構築し、その実現可能性を検証した。

キーワード 可視化、テキスト処理、地理情報、コンテキスト抽出、コミュニティベース

A Dynamic Creation Method of Environmental Situation Maps Using Text Data of Regional Information

Keita MORIYA[†] Shiori SASAKI[‡] and Yasushi Kiyoki[†]

[†] Faculty of Environmental Information, Keio University, Endo 5322, Fujisawa, Kanagawa, JAPAN

[‡] Graduate School of Media and Governance, Keio University, Endo 5322 Fujisawa, Kanagawa, JAPAN

E-mail: [†] {t06884km, kiyoki}@sfc.keio.ac.jp, [‡] sashiori@sfc.keio.ac.jp

Keyword Visualization, Text Processing, Spatial Information, Context Extraction, Community-based Knowledge

1. はじめに

情報が急速に増加する近年にあって、ユーザが個々の情報源に個別にアクセスすることは困難な現状にある。ユーザは、各自のコンテキストに従った有用な情報の獲得を求める一方、情報の概観性・一覧性を求める傾向にある。一方で、ユーザは有用な情報獲得のため、コミュニティベースで蓄積された知識データを活用する傾向にあるが、その信頼性について疑問を寄せる声も多く、また、個々の情報が分散された状態では意味をなさない情報も少なくない。

本研究は、これら情報の概観性・一覧性に対応する可視化と、コミュニティベースの情報活用の両方の目的に対応したシステムの実現方式を示す。

現在、Google Maps [10] のような GIS API を利用した住所情報の応用範囲が広がっている。この技術を用いて、コミュニティベースの文書情報を集約し、ユー

ザに何らかの有益な情報を与えることが可能である。

既存の研究として、GIS を用いた情報検索/提供技術としては、地図インターフェースを用いたインタラクティブな都市情報検索システムの研究[1]や地図コンテンツと地図以外の異種情報源組み合わせた情報ナビゲーションシステムの研究[2]、地図インターフェース上でのユーザ操作に基づく情報提供方式の研究[3][4][5]などがある。

一方で、コミュニティベースの知識集約については、ネットワークを介した知識の形成・蓄積、共有、提供・利用を支援するコミュニティ知識ベース環境の構築に関する研究[6]、地図情報を用いた位置依存コミュニティ支援情報サイトの提案[7]、知的活動を行うコミュニティの情報空間を対象としたデータ管理方式の研究[8][9]などがある。

本方式の目的は、(1)インターネット上のコミュニテ

データベースの情報源と現実空間とのリンケージ，(2)場所・地域に関する任意の情報を緯度・経度で一つにまとめ，処理する点，(3)地域状況の2次元地図上への可視化において，状況の概観性・一覧性を高めることにある。

一般に，情報検索/情報提供技術は，ユーザの要求に最も適合する情報・メディアデータの選択を目的としている．確かに，ユーザの検索要求が特定の明確かつ専門的なものである場合，可能な限りその要求に合致する情報・メディアデータを選択することは重要である．しかし，ユーザの要求が漠然としたものである場合，必ずしも提供される情報・メディアデータが特定のものである必要はない，と考えることが可能である．したがって，ユーザの要求が不明確で不特定なものである場合は，おおまかな情報源・メディアデータ集合から情報を汲み取り，メタ情報として概要・概観を示すシステムが有効なのではないかと考えられる．情報爆発時代の膨大なデータ量を前提にすれば，ユーザも個別の情報・メディアデータの信頼度に拘らずに，メタ情報として概要・概観を相対的に捉えることができると考える．

本論文は，以上の問題意識に基づき，コミュニティベースで蓄積された住所情報を含む文書の集合を，ユーザの与える任意のコンテキスト（任意のキーワードセット）によって評価し，その結果を画像として地図上に重ね合わせることにより，ある場所とその周囲環境の空間的配置に関してコミュニティが持つイメージ・印象・雰囲気を，ユーザが直感的に概要を把握できるような地域状況表示地図として可視化するシステムの実現方式について述べる．

本方式の基本アイデアを図1に示す．

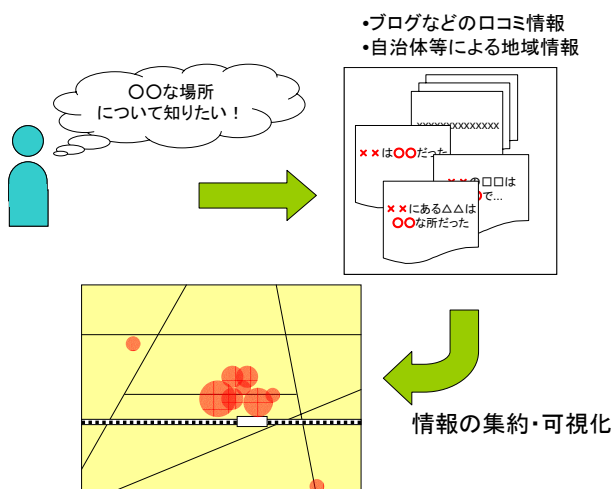


図1：本方式の基本アイデア

2. 基本方式

本システムは，(1)データ取得・変換プロセスと，(2)データ評価・描画プロセスの二つのプロセスによって構成される．システム構成図を図2に示す．

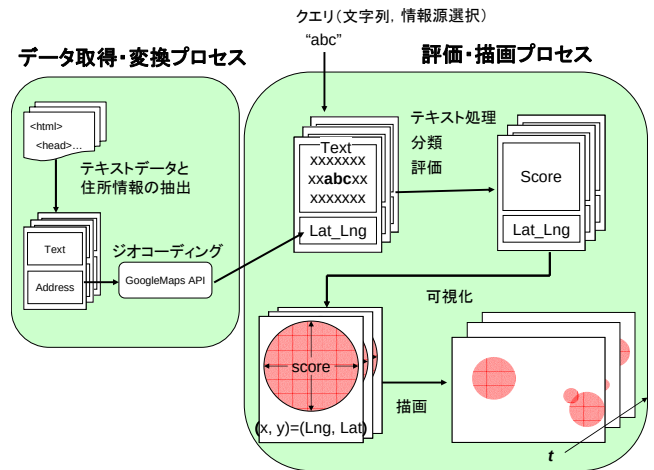


図2：システム構成図

2.1. データ取得・変換プロセス

データ取得・変換プロセスは，情報源であるコミュニティベースで蓄積されたHTML文書集合よりテキストデータ部分と住所情報をメタデータとして抽出し，住所情報を緯度経度に変換した後，テキストデータ部分と緯度経度とを1セットとしたものを地域状況データベースに格納する．このプロセスは，検索・描画プロセスに必要な前処理プロセスである．

STEP1: HTML文書集合からのテキストデータと住所情報の抽出

STEP2: 外部APIを用いた住所情報の緯度経度への変換

STEP3: テキストデータと住所情報の地域状況データベースへの格納

2.2. データ評価・描画プロセス

本プロセスはまず，ユーザが指定した任意の文字列（キーワード）をクエリとして，各文書におけるキーワードの出現回数を算出する．この出現回数は，各文書のキーワードに対するスコアとして扱われる．ここでは，1テキスト中の全単語数で割るなどの正規化を行わず，単純に出現回数のみを扱う．また，形態素解析やTF*IDF計算は行わず，単語の活用については，シソーラスを用いたクエリ拡張と，ヒューリスティクスによる動詞・形容詞の語尾処理により，対応する．これは，各文書の正確な特徴抽出やメタデータ抽出が目的ではなく，文書中にある情報を最大限に利活用することを目的とするためである．

各文書の地図上における描画ポイントは，緯度経度情報に応じて指定され，各文書の入力キーワードに対

するスコアは、描画ポイントの円周の大きさに反映される画像が描画し、それを Google Maps[6]上にオーバーラップさせることで、クエリに応じた地域情報表示地図を生成する。

STEP1：シソーラスを用いたクエリ処理・クエリ拡張

STEP2：文書集合におけるクエリ（任意の文字列）の出現回数（スコア）を測定

STEP3：地理情報を用いて、クエリに対するテキストデータの地図上における描画ポイントを特定

STEP4：各テキストにおけるクエリの出現頻度（スコア）を描画ポイントの円周に反映

STEP5：描画された画像を、2次元地図上にオーバーラップ

3. 実現方式

基本方式において設定した、データ取得・変換プロセス3STEP、データ評価・描画プロセス5STEPによって、本システムを実現した。

3.1. データ取得・変換プロセス

STEP1 対象となる情報源から、PerlのLWPモジュール[11]を用いて逐次WebページHTML文書を取得する。取得したHTML文書集合から、本文、住所、ページタイトルを、また記述されていれば時刻情報を、正規表現によって取得する。ここで取得する住所情報のなかには、番地に続いて施設名や、〇〇ビルx階といった情報が付加されているものが少なくない。これらは人間には役に立つが、緯度経度変換の際にはノイズとなり、変換が失敗する原因となり得る。そこで、住所情報に以下の正規表現をかけ、番地以降に現れる不要な文字列を削除する。

```
my $dash = "[――]";  
my $num = "(?:[0-9])[一二三四五六七八九十][0-9])+";  
my $regxp = "(.+?)(?($num(?:(:丁目?)$dash(?:番地?)))(?($num(?:(:番地?)$dash|号)?)?($num(?:号)?)?(.*)";  
$address =~ s/$regxp/$1$2$3$4/
```

STEP2 ノイズを除去したのちに、Google Maps APIを用いて緯度経度へ変換を行う。

STEP3 STEP2において得られた緯度経度データをPerlのStorableモジュール[12]を用いてデータを保存する。

3.2. データ評価・描画プロセス

STEP1 データ評価・描画プロセスは、ユーザのク

エリ発行に応じて起動する。クエリは任意の文字列をとるが、2つの拡張オプションがある。

(1) シソーラスによるクエリ拡張と語尾処理

クエリ拡張には、Weblio[13]の提供するオンライン・シソーラスを用いる。WeblioはプログラムからアクセスするためのAPIを提供していないため、対象のページを取得し、そこから正規表現で類義語を抽出する。

語尾処理に関しては、予備実験過程において得られたヒューリスティクスを用いて、語尾に「い,な,る,と,の,く」のいずれかの文字があった場合、これらを削除するというプロセスをとる。この語尾処理プロセスにより、例えば「楽しい」は語尾の「い」が削られ「楽し」となり、「楽しかった」「楽しく」などの語にマッチするようになる。ただし、クエリが2文字だった場合はこの操作は行わない。例えば「広い」を「広」にすると、ユーザの意図しないものにまでマッチしてしまう可能性が大きくなるためである。

(2) 時間軸を用いた絞込み

文書の時間情報が付加されていた場合、これを用いた対象の絞り込みが出来る。13時から15時のあいだのもの、という指定や、月や日、曜日なども条件に指定することができる。

STEP2 発行されたクエリを用いて、文書の評価を行う。クエリである文字列（キーワード）が各文書中に何回出現するかをカウントし、出現回数をその文書の評価値（スコア）とする。

STEP3 STEP2において算出された評価値をもとに、画像を描画する。そのためには、緯度経度をピクセル単位のx,y座標に変換する必要がある。そこで、以下の関数を用いて変換を行う。

```
my $lat_unit = -0.0000347622506703975;  
my $lng_unit = 0.00004284381866455;  
sub latlng2pix {  
    my ($lat, $lng) = @_;  
    return(($lng-$lng_start)/$lng_unit,  
           ($lat-$lat_start)/$lat_unit);  
}
```

ここで用いられている\$lat_unit, \$lng_unitとは、それぞれ、ある拡大度で表示したGoogle Maps上で、1ピクセル移動した際の緯度経度の変化量である。\$lat_startと\$lng_startはそれぞれ生成中の地図の開始点を示す。ピクセルへと変換したい緯度経度を\$latと\$lngに入れることで、その画像上での座標を得ることが出来る。

STEP4 STEP3によって得られた座標と評価値をもとに、画像の描画を行う。描画には、PerlのGDモジ

STEP5 最後に，生成した画像を GoogleMaps 上にオーバーラップさせる．これには，外部ライブラリである TPhoto[15]を用いる．

ここでは、ユーザの与えるコンテキストを場所に関する俗称・印象・雰囲気として、検索条件にクエリ拡張と時間帯を設定し、3章で記述した実現方式により構築した実験システムを用いて、本方式の実現可能性を検証する。

実験 2：時間的推移についての有効性の評価実験

本実験では、本システムを用いたユーザによる情報の把握・分析についての有効性を検証する．対象とする地域情報関連テキストとして、京王グループの提供するコミュニティベースの沿線情報サイトである「街はび」[16]を用いる．街はびでは、「ライター」と呼ばれる一般の人々が施設や店舗、イベント等についての情報を投稿しており、累計の情報件数は 4000 を越えている．

4.1.1. クエリに名詞を用いた実験

一見して、幡ヶ谷、笹塚、下北沢から北、また下高井戸駅の周辺に商店街が広がっていることを見て取ることができ、これは実際の商店街の所在の通りである。

4.1.2. クエリに形容詞を用いた実験

表れているものであると考えられる．また，赤と青の円の分布に差が認められないことから，クエリによって文書が有意に分類されていないものと判断した．

そこで、それぞれの円の大きさに上下左右 10 ピクセル範囲の文書数の逆数をかけて再度描画を行った。その結果が図 5 である。たしかに文書の配置の偏りは解消されているように見えるが、文書数の偏りもその情報源が持っている重要な情報の一つと考えると、その情報を損なう結果であり、クエリに対する文書スコア計算の際に文書長による正規化を行わないという当初の方針に反した結果であるとも考察できる。この点については、今後の検討課題である。



図3：京王線沿線の地域状況について“商店街”をクエリとした場合の出力結果(部分). 商店街が自動的に評価され、地図上に表示されている。



図4：京王線沿線の地域状況について“楽しい”“静か”をクエリとし、前者を赤，後者を青で出力した結果。“楽しい”“静か”というキーワードを含む文書が反応し、その位置に円が描画されている。



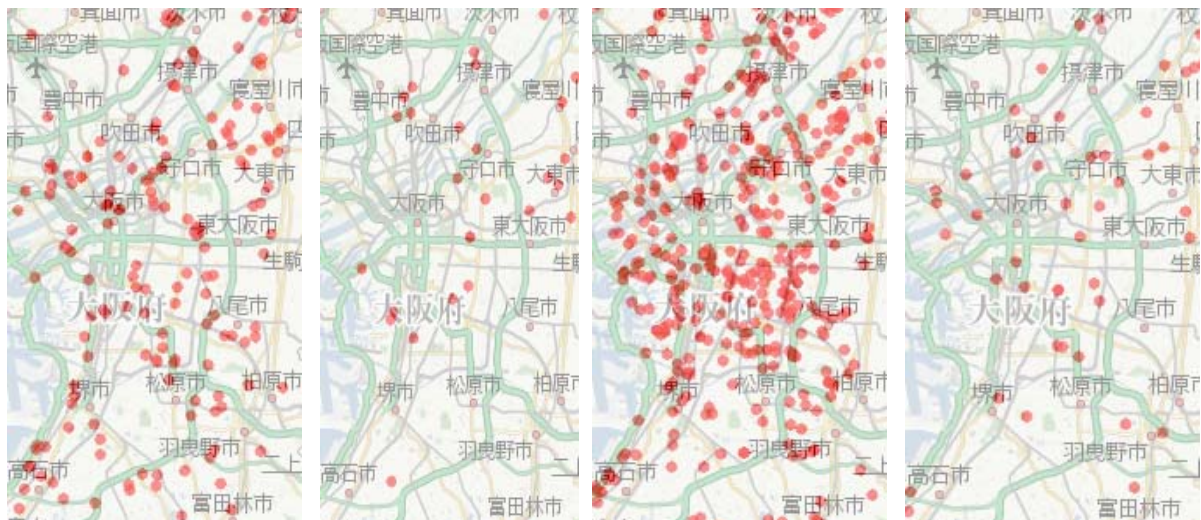
4.2.2 実験24

同じ実験について、文書分布数に基づき、本実験の重み付けをおこなった場合の結果に関連する情報の時間的推移についてどの程度表示可能であり、ユーザによる情報の把握・分析にどの程度有効であるかを考察する。対象とする地域情報関連テキストとして、大阪府警察安まちメール[17]を用いる。大阪府警察安まちメールとは、大阪府内で発生した軽犯罪の情報を、種別、詳細事項、場所、発生時刻を記述したかたちで提供している Web サイトである。

このサイトでは個別事件に関する情報提供の他に、事件の発生件数の 2 次元地図上へのマッピングと可視化表示[11]を独自に行っている。これは事件の起こった地点に、発生時期に応じたマーカを表示する手法をとっている。しかし、(1)大量の情報をマーカによって表示していることで視認性に欠けること、(2)提供者の指定した分類による地図しか閲覧できないこと、(3)時間情報を任意に活用できないこと、(4)他地図との比較が困難であることなどから、本システムの適用により、利用者に新たな価値を提供できると考えた。

図 6 は、クエリ「子供、被害」で一時間刻みに時刻指定をして出力したものからの抜粋である。登下校時刻、とくに下校時刻に事件件数が増加すること、登校時間直後は発生件数が著しく現象すること、などの情報を表出させることに成功している。

また地図を詳細に見ていくと、図 7 で示すような一部分だけ事件の発生していないスポットの存在が見て取れる。本システムによってこのようなスポットを発見し、更に時間帯や曜日ごと、または入力するキーワードを変えた出力により「なぜここでは子供の被害がないのか」という分析を多面的に行い、他地域での防犯に役立てるといった応用が考えられる。



07:00

子供の登校時刻であるため、事件件数が多い

09:00

登校時間直後は事件件数が著しく減少する

15:00

下校時刻が最も事件件数が多い。

20:00

夜間は 09 時とほぼ同水準であることがわかる

図 6：大阪府の治安状況について、「子供」「被害」

をクエリとした場合の一時間ごとの出力結果(部分)



図7: 図6と同じ実験における15時の地図における、局所的に事件が発生していないスポット。

5. 考察

単純な単語のパターンマッチによっても、クエリが名詞の場合は地域情報関連テキストから有意な地域状況を抽出可能なことがわかった。とくに商店街の例では、「人々が商店街だと認識している範囲」を可視化できたことから、コミュニティベースのテキスト集合から地域状況表示地図の生成に成功したと言える。

しかしクエリが形容詞である場合には、必ずしも意味のある結果が出たとは言えず、こちらは文書の評価方法等について再検討が必要である。

また、シソーラスでクエリの拡張を行った際に、ユーザ意図とかけはなれた意味の単語にまで拡張されてしまう例があり、拡張後にクエリに用いる単語を選別できるインターフェースの構築、または、クエリの拡張方法自体の再考の必要性を感じた。

6. 結論および今後の展望

本稿では、住所情報を含む地域情報関連テキストを対象として、ユーザの与えるコンテキストに応じた「地域状況表示地図」を動的に生成するシステムの実現方式を示した。本方式は、コミュニティベースで蓄積された住所情報を含む文書の集合を、ユーザの与える任意のコンテキスト（任意のキーワードセット）によって評価し、その結果を画像として地図上に重ね合わせるにより、ある場所とその周囲環境の空間的配置に関してコミュニティが持つイメージ・印象・雰囲気、地域状況表示地図として可視化するものである。本稿では、ユーザの与えるコンテキストを場所に関する俗称・印象・雰囲気として設定し、実験を行い、本方式によって、ある土地に関する情報をテキストや単

語の観点から一望することが可能となり、また、その評価結果を地図上に視覚的に表現することにより、地理情報に応じた情報の把握や、その応用を可能にすることを検証した。

今後は、本システムを、多様な情報源に対応可能な、より汎用的な検索エンジンとして実装を進め、定量的実験を行うことにより、本システムの実用可能性について検証する。

文 献

- [1] 平松 薫, 小林 堅治, 石田 亨, 赤埴 淳一, “デジタルシティにおける情報検索のための地図インタフェース,” 情報処理学会, Vol.41, No.12, pp. 3314-3322, 2000.
- [2] 北角 智洋, 池田 哲夫, 田辺 弘実, 星 隆司, “分散した地理情報を統合利用する情報ナビゲーションシステム,” 情報処理学会, Vol.42, No.SIG_15(TOD_12), pp. 30-39, 2001.
- [3] 高木 悟, 松本 一則, “地図情報を利用した情報検索,” 情報処理学会, Vol.41, No.4, pp.357-362, 2000
- [4] 平元 綾子, 角谷 和俊, “地図インタフェースにおけるユーザ操作からの Web 検索方式,” 日本データベース学会, Vol.5, No.1, pp. 85~88, 2006
- [5] 廣瀬正義, 平元綾子, 角谷和俊, “GeminiMap: ユーザ操作を用いた地図および写真の連動方式,” 電気情報通信学会第 18 回データ工学ワークショップ(DEWS2007), 2007.
- [6] 池田 文人, 山本 恭裕, 高田 眞吾, 中小路 久美代: コミュニティ知識ベース環境の構築へ向けての知識の形成と利用に関する調査と分析, 情報処理学会論文誌, Vol.40, No.11(19991115) pp. 3887-3895.
- [7] 大須賀志保, 有川正俊, 田中浩也: 風景写真シークエンスを用いた位置依存コミュニティ支援情報サイトの提案, 電子情報通信学会第 13 回データ工学ワークショップ(DEWS2002), 2002 年 3 月.
- [8] 望月祥司, 児玉麻莉子, 石川憲一, 森嶋厚行, 田島敬史: 大規模ディレクトリ空間視覚化のための論理構造抽出, 電子情報通信学会第 18 回データ工学ワークショップ(DEWS2007), 2007 年 3 月.
- [9] 三森祐一郎, 森嶋厚行: コミュニティ情報空間管理のための制約記述と処理手法の提案, 電子情報通信学会第 19 回データ工学ワークショップ(DEWS2008), 2008 年 3 月.
- [10] Google Maps: <http://maps.google.com/>
- [11] LWP Ver. 5.814
- [12] Storable Ver. 2.18
- [13] Weblio: <http://thesaurus.weblio.jp/>
- [14] GD Ver. 1
- [15] TPhoto: <http://gmaps.tommangan.us/tphoto.html>
- [16] 街はび: <http://www.happy-town.net/>
- [17] 大阪府警察安まちメール: <http://www.info.police.pref.osaka.jp/>
- [18] 大阪府警察犯罪発生マップ: <http://www.map.police.pref.osaka.jp/Public/index.html>