地名の関連グラフを利用した地理情報検索

椎名　宏徳† 　李　　龍†† 　上林弥彦 *†††

† 京都大学工学部
††† 京都大学情報学研究科社会情報学専攻
〒 606–8501 京都府京都市左京区吉田本町
†† サムスン綜合技術院 Ubiquitous Computing Lab.　大韓民国，郵便番号 440–600 水原市私書箱 111
E-mail: †shiina@db.soc.i.kyoto-u.ac.jp, ††ryong.lee@samsung.com, †††yahiko@db.soc.i.kyoto-u.ac.jp

あらまし　地理情報検索システムでは，地名や各種キーワードの他に利用者にわかりやすい地図インタフェースが広く利用されている．地図を利用することで距離の近い地名の分布などを容易に把握できる．しかしながら，地図では，地理的な関係以外の地名の関係を扱うことはできない．本稿では，ウェブページ上の地理オブジェクト名の共起関係に注目して (1) 地理オブジェクトの共起回数を利用する方法，および (2) 各地理オブジェクトを共起する地理オブジェクトのベクトルで表現しそのベクトルの類似度を計算する方法，により地理オブジェクトの関連度を計算し，地理オブジェクトの特性付けを行う方式を検討した．また，システム構成のおいては，これらの関係をわかりやすい形で表すことが重要であるため，オープンソースのグラフ生成ツール TouchGraph を用いて視覚化し，実装されたシステムでその有用性を確認した．

キーワード　地理情報検索，情報視覚化，WWW，ユーザインタフェース

Use of Graph Representation of Location RelationShips for Geographic Information Systems

Hironori SHIINA†, Ryong LEE††, and Yahiko KAMBAYASHI*†††

† Faculty of Engineering, Kyoto University
††† Department of Social Informatics, Kyoto University
Yoshidahonmachi, Sakyo-ku, Kyoto-shi, 606–8501 Japan
†† Ubiquitous Computing Lab. Samsung Advanced Institute of Technology
P.O.Box 111, Suwon, 440–600 Korea
E-mail: †shiina@db.soc.i.kyoto-u.ac.jp, ††ryong.lee@samsung.com, †††yahiko@db.soc.i.kyoto-u.ac.jp

Abstract　For geographic information systems, map-based user interface is commonly used since locations close-by are shown together. In order to realize advanced functions, use of relationships between locations is very important. As a relationship, we will use (1)co-occurence of location names, and (2)similarity between two vectors. Each vector shows locations co-occurring with each location. We have developed a graphical user interface using open source graph generation tool TouchGraph. We showed that the system using (2) above is better than (1).

Key words　Geographic Information Retrieval, Information Visualization, WWW, User Interface

1. はじめに

ユーザが要求する情報の種類に応じてさまざまな情報検索の手法が提案されており，その中の一つにある地域に関する情報を得ることを目的とした地理情報検索システムがある．地理情報検索システムでは，地図を利用したインタフェースが用いられていることが多い．ユーザは地図上で検索範囲を指定したり，地理オブジェクトの位置，分布などを視覚的に確認することができる．しかし，地図だけでは一部の限定された範囲しか表示することしかできない．地理的には離れていても，関連の深い地名というものは存在する．従来の地理情報検索システムではこのような関連に基づいた検索を行うことはできなかった．

この問題に対して,「概念的な地理情報システム」として，京都大学上林研究室において KyotoSEARCH [1], [2] の研究・開発

*2004 年 2 月 6 日逝去

が行われている．現段階の KyotoSERACH では，検索機能の一つとして，ウェブページを収集し，あらかじめ解析しておくことで，ある地名に関連する別の地名や一般名詞などを提示する機能が提案されている．しかし，地名や名詞の関連の強さをどのように計算するかはまだ定められていない．関連用語に関するインタフェースも関連地名，関連用語をそれぞれ羅列するのみで特に関連の強い地名・用語がどれなのか把握することはできない．

また，地理情報検索システムにおける課題の一つに観光施設，商店などの地理オブジェクトの分類が挙げられる．ユーザが自分の目的に合った地理情報を検索するためには，地理オブジェクトが分類されている必要がある．これまでの地理情報検索システムでは，地理オブジェクトの分類は人為的に行われていたため，検索できる地理オブジェクトの数が制限されていた．

本稿では，これらの問題を解決することを目的としている．なお，本研究では，地名以外の単語は利用せず，地名の関連のみを扱う．また，本稿では地名として，寺社仏閣，施設などの地理オブジェクトを主に扱う．地理オブジェクトの関連の情報源として，KyotoSEARCH と同様にウェブページ上での地名の共起関係に着目し，ウェブページを収集し地理オブジェクトの共起回数を数える．

ここで，あまり他の地理オブジェクトと共起しない地理オブジェクトとの共起は注目すべき共起関係であると考え，共起関係における地理オブジェクトの希少度という値を計算する．この希少度と共起回数から地理オブジェクトの直接共起度を定義する．この直接共起度を用いることにより，地理オブジェクトの共起関係をより正確に表現することが可能となる．

地理オブジェクトの意味関連を提示するためには，地図インタフェースに替わる視覚化を行わなければならなず，また，そのインタフェースは地図インタフェースと同様に視覚的にわかりやすいものでなければならない．一般的に情報の関係，構造を視覚化する際には，グラフが広く用いられている．地理情報検索システムにおいても，グラフをインタフェースに利用すれば，地理空間にとらわれない柔軟な情報提示が可能になると考えられる．そこで，オープンソースであるグラフ生成ツール TouchGraph [8] を利用して地理オブジェクトの共起関係を視覚化する．グラフを用いた視覚化により，色の濃淡やエッジの長さで地理オブジェクトや関係の重要性をも視覚的に把握することが可能となる．

この地理オブジェクトの直接共起度を用いた関連グラフを観察すると，同じ種類の地理オブジェクトが連結していることが多いという特徴が見つかる．このことから，地理オブジェクトの共起関係を利用して地理オブジェクトを分類することが考えられる．そこで，同類の地理オブジェクトは共起する地理オブジェクトも類似していると考え，各地理オブジェクトがどの地理オブジェクトと共起しているのかを特徴ベクトルで表し，その類似度を計算する．直接共起度，類似度，さらにこの二つの値を組み合わせて定義される総合共起度からグラフを作成し，地理オブジェクトがどのように連結し，分類されるかを比較する．

以下，2 章では，本研究の関連研究について触れる．3 章では，本研究の基盤となっている地理情報検索システム KyotoSEARCH について説明する．4 章では，ウェブ上での地理オブジェクトの共起関係を計算する．5 章では，地理オブジェクトの共起関係ををグラフ化する．この結果をふまえて 6 章では，地理オブジェクトの共起の類似度を利用して総合共起度を計算して，4 章の共起関係と比較する．7 章で，地図インタフェースと総合共起度グラフを組み合わせたシステムを実装し，8 章で本稿をまとめ，今後の展望等を示す．

2. 関 連 研 究

この章では，本稿に関連する研究として，ウェブなどを利用して地理情報を検索するシステムや単語の共起を利用した研究を示す．

モーバイルインフォサーチ実験におけるここのサーチ [3] は，現在地周辺に関するウェブページを検索するシステムである．ウェブページはページ中に現れる住所表記をもとに地図上の多角形として管理される．ユーザの現在地を中心とした円とこの多角形の重なりを調べることで検索を実行する．

Digital City の GeoLink [4] も地域に関連したウェブページを検索するシステム(注1) である．このシステムは京都市を対象に，登録されたウェブページを地図上の点にプロットしている．ユーザは地図上の点を選択することで，ウェブページを閲覧できる．ウェブページは観光，食事などのカテゴリに分類されており，ユーザは目的に合わせた検索を行うことができる．その他にもキーワード検索，距離検索などが実装されている．

相良氏らの研究では，文書中の地理情報を表現するタグとして <spa> 表現を提案している [5]．また，住所表記から緯度経度座標を効率的に求める分散システム [6] を考案している(注2)．これにより，ウェブページ中の住所表記から，ウェブページを地図上にマッピングすることなどが可能である．

共起データを利用した研究として，李氏らの研究 [13] が挙げられる．この研究は，任意の二つの単語集合（例えば，名詞集合と動詞集合）の間の共起データに基づいて，単語をクラスタに分類している．この方法は，単語クラスタリング問題を二つの単語集合の分割の直積上に定義される確率モデルの推定問題として捉え，情報理論や数理統計学の分野で提案されている記述長最小の原理（MDL 原理）を推定基準として用いる点が特徴である．

3. 地理情報検索システム KyotoSEARCH

この章では，本研究の基盤となっている地理情報検索システム KyotoSEARCH について述べる．KyotoSEARCH [1] は，地域に関連したウェブページを検索することを目的としたシステムとして，京都大学上林研究室において研究・開発が行われている．従来の地理情報システムと異なる点は，あらかじめ収集したウェブページを分析し，その結果発見された知識を利用し

(注1)：GeoLink Kyoto http://www.digitalcity.gr.jp/openlab/kyoto/map_guide_j.html
(注2)：
アドレスマッチングサービス http://fujieda.csis.u-tokyo.ac.jp/cgi-bin/geocode.cgi

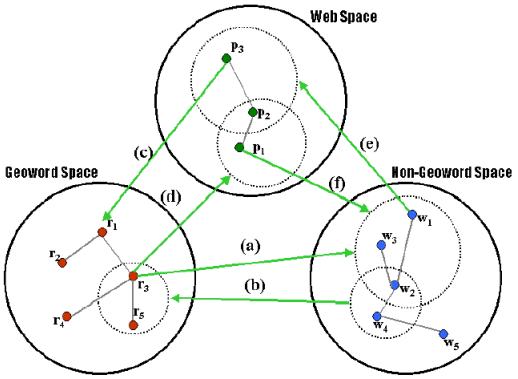

図 1　ウェブ情報空間，地名空間，非地名空間の関連

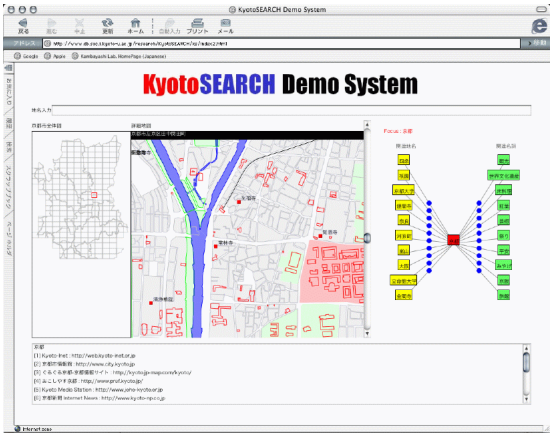

図 2　KyotoSEACH

た検索機能を備えていることである．この機能には，ユーザが一つ一つのウェブページを閲覧しなくても，ある程度ウェブから得られる知識を知ることができるという利点がある．以下に KyotoSEARCH の概要を示す．

ウェブ分析により，地域の特徴，地域の関連といった地域に関する二種類の知識が得られると考えられる．例えば,「銀閣寺」の関連ページを集めて，特徴的なキーワードを抽出すると,「世界文化遺産」,「歴史」,「観光」,「バス時刻表」などの「地域の特徴」を表す単語が抽出される．また,「銀閣寺」の関連ページによく現れる他の地名を分析すると「哲学の道」,「金閣寺」,「祇園」,「清水寺」,「東山」などの地名が挙げられ「地域の関連」を知ることができる．

こうして得られる名詞空間を地名（G) と非地名（N) に分けて，ウェブ情報空間（P-domain)，現実（地名）空間（G-domain)，非地名空間（N-domain）の関連とそこから得られる知識について考察する．これらの知識をデータマイニングでの連想ルールとして以下のように表現する．

- 地域の特徴：$G \rightarrow N*$
- 地域間の関連：$G \rightarrow G*$
- 地域ウェブページ $G \rightarrow P*$

実際の計算では，地域ウェブページ（$G \rightarrow P*$）からそれぞれ「特徴」（$P \rightarrow N*$）と「地域関連」（$P \rightarrow G*$）を計算する．この過程は，図 1 のように，三つの情報空間のリンクを辿るモデルで表現できる．

例として，あるユーザが「観光」を目的としてどこかに行きたいと思って，ウェブ検索を利用する場合を考える．まず,「観光」という非地名（図 1 の w_1）から適当なページ（P_3）を見つけて（$w_1 \rightarrow P_3$）内容を見る．その中の「京都」（r_1）という地名に興味を持ち，その地域の地図を見る（$P_3 \rightarrow r_1$）．そして，地図から「銀閣寺」という地名を見つける（$r_1 \rightarrow r_3$）．さらに,「銀閣寺」がどういった場所か知るために関連キーワードを探す（$r_3 \rightarrow w_1, w_2, w_3$）．そこから，最初の「観光」というキーワード（$w_1$）を見つけて,「銀閣寺」が「観光」にふさわしいことを知り，そこに行くことを決める．

このような三つの情報空間の間での検索を支援することが KyotoSEARCH の目的である．KyotoSEARCH はウェブページの収集・解析を行うサーバとユーザが検索・情報閲覧をおこなうためのクライアントから構成されている．図 2 にクライアントのユーザインタフェースを示す．画面はキーワード関連インタフェース，地図インタフェース，ウェブページリストから構成されている．

画面右に配置されているのがキーワード関連インタフェースで，ユーザが注目しているキーワードを中心に左に関連地名のリストが，右に関連名詞（非地名）のリストが置かれている．リスト上のキーワードを選択するとそのキーワードが新たな注目キーワードとなり，関連キーワードが更新される．また，地名が選択された場合は，画面左の地図インタフェース上にその地名の周辺の地図が表示される．地図インタフェースでは，地名の分布を確認することができ，地図インタフェースに表示されている地名を選択すると，その地名が新たな注目キーワードとなる．画面下のウェブページリストには注目キーワードに関するウェブページの URL が表示される．

この KyotoSEARCH の検索機能の精度を向上させるために，ウェブページの収集，解析，管理や検索方法などさまざまな研究が行われている [2]．本研究では，地名の関係の解析および視覚化に主眼を置く．

4.　地理オブジェクトの関連性

本稿では，地図では表せない地名の関係を示すことを目的にしている．地理的関係とは異なる関係として，ウェブページにおける地名の共起関係を利用することにする．地名は位置を示す単語であると同時に，意味を持った単語でもある．ウェブ文書における地名の共起関係を調べることで，意味的な観点で地名の関係を発見できると考えたためである．また，本稿では，地名として寺社仏閣，施設などの地理オブジェクト名を扱う．地理オブジェクト名は，住所表記のような地名に比べてページ内での出現頻度が高いと考えられる．

まず，京都市情報館というウェブページ(注3) から，URL に kyoto というフレーズを含む URL ページのみをリンクを辿って収集し（約 8500 ページを収集），ページごとに出現する地理

（注3）：http://www.city.kyoto.jp/koho/ind_h.htm

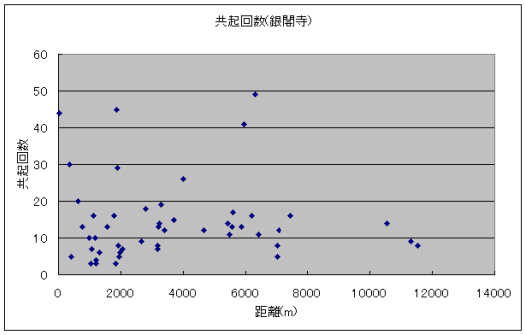

図 3　距離と共起の相関（銀閣寺）

Fig. 3　a correlation between distance and co-occurence(Ginkaku-ji Temple)

地理オブジェクト名	共起回数	距離 (km)
金閣寺	49	6.29
南禅寺	45	1.86
慈照寺	44	0.03
京都駅	41	5.93
法然院	30	0.36
平安神宮	29	1.88
清水寺	26	3.99
安楽寺	20	0.63
相国寺	19	3.28
知恩院	18	2.80

表 1　銀閣寺と共起する地理オブジェクト（回数）

Table 1　Places appearing with 'Ginkaku-ji Temple' (frequency)

オブジェクト名を数える．ここで用いた地理オブジェクト名は，株式会社ゼンリンの住宅地図電子データ（TOWN2）が持つ文字情報のうちレイヤ 107（目標物レイヤ）に含まれるものを利用した．具体的には，京都市内の代表的に寺院や大学などのランドマーク名で数は約 2500 である．そのうち約 1400 の地理オブジェクトが収集したページ内に現れた．HTML タグを利用して，ページ中の地名の重要性を考慮する手法 [9] やページ自体の地域性の強さを考慮する手法 [10] も考案されている．しかし今回は計算の高速化のために HTML タグを用いたりページの内容を考慮したりせずに，ウェブページ中に何らかの形で地理オブジェクト名が現れれば，すべて出現回数 1 回とみなし，地理オブジェクトの出現回数を数える．この解析結果から地理オブジェクトの共起数を計測する．一つのページに二つの地理オブジェクトが現れた場合，各地理オブジェクトの出現回数にかかわらず，共起 1 回として，共起の回数を数える．

この結果の中から，図 3 に「銀閣寺」について，地理オブジェクトの共起回数と地理的距離の相関図を示す．この図によると，共起回数の多い地理オブジェクトは，地理的に近い距離にある地理オブジェクトであることが多いことが読み取れる．これらは特に意味的な関連があるというよりも，周辺地域について述べたウェブページ上でまとまって共起しているものと考えられる．

一方，地理的に遠い距離にあっても，共起回数の多い地理オブジェクトは存在し，これらの地理オブジェクトは地図からでは把握することができない．表 1 は「銀閣寺」と共起する回数の多い地理オブジェクトを上から順に 10ヶ所具体的に挙げたものである．地理的に遠い距離にありながら，共起回数の多い地理オブジェクトに「金閣寺」と「京都駅」が挙げられる.「金閣寺」は「銀閣寺」と並ぶ室町時代の代表的な建築物であり，宗派も同じであることから，実際に関連の深い地理オブジェクトであると考えられる．一方の「京都駅」は，京都を訪れる観光客の出発点であり，そこからの交通方法を示すために共起しているに過ぎず,「銀閣寺」に特別関連が深いわけではない.「京都駅」は同様の理由により，多数の地理オブジェクトと共起していると考えられる．文書の特徴づけを行う手法である tfidf 法では，多くの文書に現れる単語は文書を特徴付けるという点ではあまり役に立たないという考えから，idf(Inverse Document Frequency) という指標が用いられている．この考え方と同様に，多くの地理オブジェクトと共起する地理オブジェクトは特徴的な関係を表さないと考え，式 1 に示すように，共起関係における地理オブジェクトの希少度（$rare_i$）を地理オブジェクト i について定義する．この希少度（$rare_i$）と共起回数（$cf_{i,j}$）を利用して，地理オブジェクト i, j に対して式 2 示す直接共起度 ($dco_{i,j}$) を共起関係の指標とする．

$$rare_i = \log \frac{\text{地理オブジェクトの総数}}{i \text{と共起する地理オブジェクトの数}} \quad (1)$$

$$dco_{i,j} = cf_{i,j} \times rare_i \times rare_j \quad (2)$$

「銀閣寺」について直接共起度の高い地理オブジェクト上位１０ヶ所を表 2 に示す.「京都駅」の直接共起度が小さくなっていることが確認できる．

地理オブジェクト名	直接共起度	距離 (km)
慈照寺	24.97	0.03
金閣寺	17.39	6.29
法然院	12.69	0.36
南禅寺	12.64	1.86
京都駅	9.76	5.93
安楽寺	9.71	0.63
平安神宮	9.64	1.88
大豊神社	8.3	0.99
清水寺	8.03	3.99
霊鑑寺	7.98	0.75

表 2　銀閣寺と共起する地理オブジェクト（直接共起度）

Table 2　Places appearing with 'Ginkaku-ji Temple' (Direct co-occurence value)

5.　共起関係の視覚化

この章では，4 章で求めた地理オブジェクトの共起関係を視覚化する．

視覚化には，グラフを用い，ノードが地理オブジェクトを表し，エッジで共起関係を表すことにする．情報視覚化にはさまざまな技術があるが [7]，グラフの生成は TouchGraph の Touch Graph Layout Ver1.21 を利用する．このツールはオープンソー

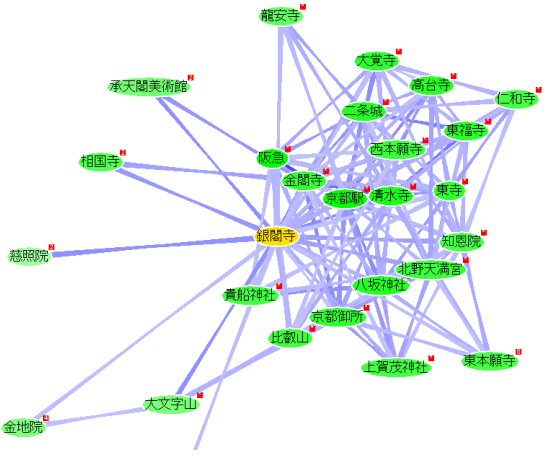

図 4　直接共起度グラフ

Fig. 4　A part of the direct co-occurrence graph

スで利用でき，ノードとエッジのデータを入力することでグラフが作成される．本稿のグラフでは，ノードはパラメータとしてウェブ上での出現回数を持ち，出現回数の多い地理オブジェクトほど色が濃くなっている．ウェブ上での出現回数の多い地理オブジェクトは重要性の高い地理オブジェクトであると考えられるため，このグラフでは地理オブジェクトの共起関係に加えて，地理オブジェクトの重要性も視覚化している．エッジのパラメータには，4 章で求めた直接共起度を利用し，値が大きいほど枝は短く，色が濃くなっている．つまり，関係の強いノード同士は近い距離に配置される．共起回数の少ない地理オブジェクト同士までつなげてしまうと特徴がつかみづらくなってしまうため，直接共起度の値が一定値以下の共起関係は視覚化していない．ここでは，直接共起度が高い順に上位 3000 のエッジを利用する．

また，4 章でも述べたように，ここでの地理オブジェクトの関係は以下の二つが考えられる．

- 地理的に近い距離にある関係
- 地理的に遠いが，意味的関連のある関係

地図インタフェースは，距離，方位などを正確に表現できるため，地理的に近い距離にある地理オブジェクトは地図上に表示するのが適している．一方，グラフは，地図で表示できない，地理的に遠いが，意味的関連のある関係を示すのに適していると考えられる．そこで，グラフ上では，地理的距離が一定距離以下の地理オブジェクトの共起関係は無視することにする．今回はこの距離の閾値を 2km に設定している．作成されたグラフの一部を図 4 に示す．

このグラフを検証すると,「銀閣寺」と「金閣寺」,「下鴨神社」と「上賀茂神社」(注4)のような歴史的背景による関連が確認できる．

その他に注目すべき点としては，グラフ上では，同じ種類の地理オブジェクトが連結していることが多いということがあげられる．例えば，図 5 は，直接共起度グラフ上で「同志社大学」の周辺のノードを表示したものであるが，大学同士がつながっていることが確認できる．つまり，直接共起度グラフにより，地理オブジェクトが種類ごとに分類されていると考えることができる．だが，大学に関していえば名称が「大学」というフレーズで終わる地理オブジェクトをまとめてしまった方が効率がよいであろう．しかし，図 6 を見ると,「博物館」「美術館」などのさまざまな文化施設が，グラフ上でまとまっていることがわかる．このような分類を名称のみから行うことは困難である．この結果から，地理オブジェクトの共起関係を，地理オブジェクトの分類に利用することが考えられる．次の章では，共起関係から地理オブジェクトの類似度を計算し，地理オブジェクトの分類への応用を検討する．

(注4)：「下鴨神社」,「上賀茂神社」は，二つを合わせて賀茂社と呼ばれ，京都三大祭の一つ，葵祭が催される．

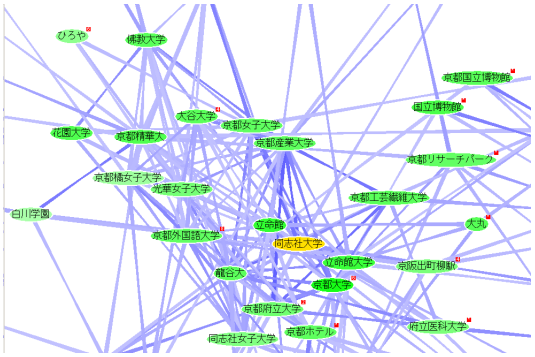

図 5　直接共起度グラフにおける大学の関連

Fig. 5　The relationship between universities on the direct co-occurence graph

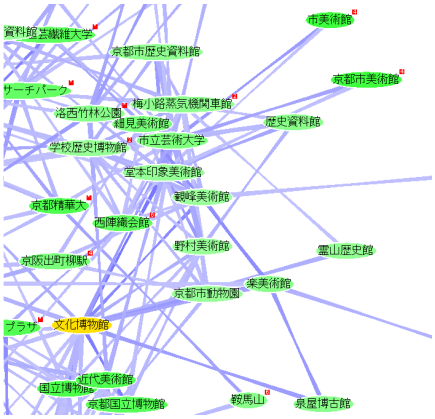

図 6　直接共起度グラフにおける文化施設の関連

Fig. 6　The relationship between museums on the direct co-occurence graph

6.　共起の類似度を用いた地理オブジェクトの分類

5 章でのシステムの観察から地理オブジェクトの共起関係を利用して地理オブジェクトを分類することを考える．図 7 が直接共起度グラフの概観である．ある程度同類の地理オブジェクトがまとまっているが，全体としては，多くのノードがつながっており，地理オブジェクトが分類されているとはいえない．

地理オブジェクトの分類されたグラフを生成するためにさらに地理オブジェクトの共起関係を解析する.「京都駅」は京都の交通の起点として，多くの地理オブジェクトと共起している．しかし,「銀閣寺」や「清水寺」などの寺社が京都駅と同じ分類

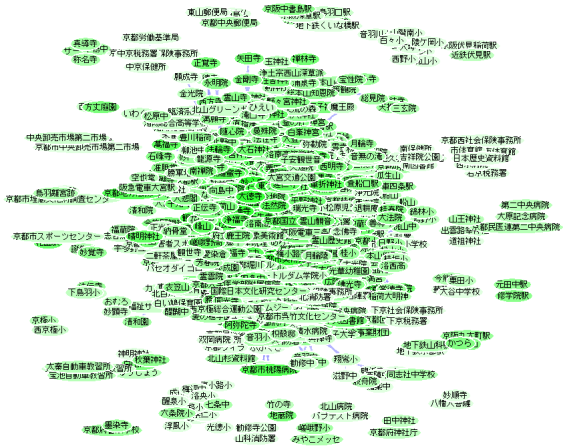

図 7　直接共起度グラフの全体像

Fig. 7　The overview of the direct co-occurence grpah

図 8　総合共起度グラフの全体像

Fig. 8　The overview of the total co-occurrence graph

というわけにはならない．そこで，同じ分類にあてはまる地理オブジェクトは共起する地理オブジェクトも類似していると考え，地理オブジェクトの共起状況の類似度を計算する．まず，地理オブジェクト i の共起特徴ベクトル $\mathbf{x}_i$ を式 3 のように定義する．

$$\mathbf{x}_i = \frac{1}{\sqrt{\sum_{j=1}^{n} dco_{i,j}^2}}(dco_{i,1}, \ldots, dco_{i,n}) = (x_{i,1}, \ldots, x_{i,n}) \quad (3)$$

続いて，二つの特徴ベクトル $\mathbf{x}_i, \mathbf{x}_j$ のなす角の余弦を二つの地理オブジェクトの類似度 $sim_{i,j}$ として式 4 のように定義する．

$$sim_{i,j} = \mathbf{x}_i \cdot \mathbf{x}_j = x_{i,1}x_{j,1} + \ldots + x_{i,n}x_{j,n} \quad (4)$$

この特徴ベクトルの余弦を用いた類似度計算は情報検索の分野で広く用いられている手法である．すべての地理オブジェクトの特徴ベクトルの組に対して，この類似度を計算する．さらに，地理オブジェクト同士の共起関係の強さを表す直接共起度（$dco_{i,j}$）と，地理オブジェクト同士の類似関係の強さを表す類似度（$sim_{i,j}$）から，地理オブジェクトの総合共起度（$tco_{i,j}$）を式 5 のように定める．

$$tco_{i,J} = dco_{i,j} \times sim_{i,j}^{\alpha} \quad (5)$$

α は定数で，類似度の重みを調節するのに用いる．

TouchGraph を用いて，地理オブジェクトをノード，一定値以上の総合共起度を無向エッジにしてグラフ化する．こちらでもエッジは類似度をパラメータに持ち，総合共起度が高いほど短く，色が濃くなっているので，関係の強い地理オブジェクト同士は近づいて配置される．

$\alpha = 3$ とし，総合共起度の高い上位 3000 のエッジを利用して作成された総合共起度グラフは図 8 のような概観をしている．図 7 に比べて，クラスタ状のグラフがいくつも散らばっている．関連度を用いると，類似した地理オブジェクト同士の間のエッジの数が増加し，クラスタが形成されることが伺える．

α の値とエッジの本数を変化させてグラフを作成し，どのようなクラスタが形成されるかを計測する．エッジは総合共起度の値が高いものから利用している．図 9 は，α の値と形成されたクラスタの総数との関係をエッジの本数ごとに示したものである．$\alpha = 0$ は直接共起度を用いたことを意味している．また，α の個所が sim となっているのは，式 5 の総合共起度ではなく，式 4 による類似度によって作成されたグラフのデータである．図 10 は，縦軸に二つ以上の地理オブジェクトを含むクラスタの数をとったグラフであり，図 11 は，一つのクラスタに含まれる地理オブジェクトの数の分散を縦軸にとっている．また，表 3 にエッジを 3000 とした時，クラスタに含まれる地理オブジェクトの数を上位 10 のクラスタについて示す．

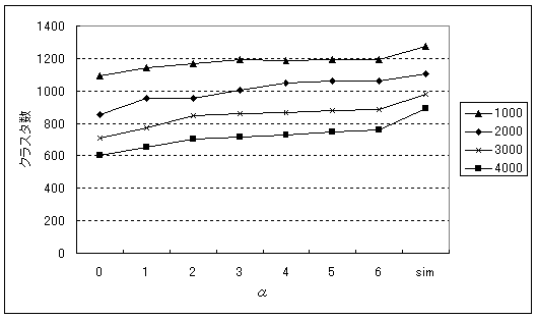

図 9　クラスタの総数

Fig. 9　The number of clusters

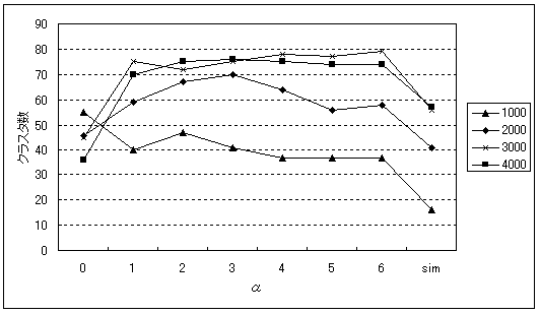

図 10　サイズ 2 以上のクラスタの数

Fig. 10　The number of clusters containning more than one places

4000 本のエッジを利用した場合，クラスタの総数が 600〜700 程度であるのに対し，二つ以上の地理オブジェクトを含む

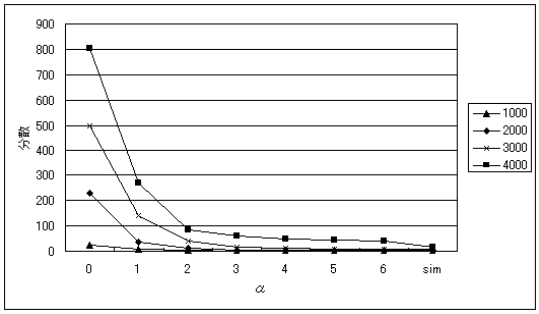

図 11　クラスタサイズの分散
Fig. 11　The standard deviation of size of clusters

クラスタの数は 100 にも満たない．この二つの数の差は他の地理オブジェクトとの間にエッジを持たない地理オブジェクトの数を表している．この場合，500〜600 程度の地理オブジェクトが他の地理オブジェクトとの間にエッジを持たず，実際に他の地理オブジェクトと結びつきクラスタを形成したのは約 1400 の地理オブジェクトのうち，800 程度であるということになる．

直接共起度を用いた場合（$\alpha = 0$），二つ以上の地理オブジェクトを含むクラスタの数が少なく，クラスタに含まれる地理オブジェクト数の分散が非常に大きくなっている．このことから，直接共起度を用いて作成したグラフでは表 3 からもわかるように，非常に大きなクラスタが一つだけ形成され，残りのクラスタはごく少数の地理オブジェクトを含むのみであるということが伺える．

一方，類似度を用いた場合（図中で α が sim となっているもの）は，クラスタに含まれる地理オブジェクト数の分散は小さく，比較的均等な大きさのクラスタが生成されていることがわかる．しかし，クラスタの総数が大きい反面，二つ以上の地理オブジェクトを含むクラスタの数は小さくなっている．このことは，他の地理オブジェクトと結びつかない地理オブジェクトの数が多いことを示している．

以上により，直接共起度のみでは，地理オブジェクトが広く結びついてしまい，非常に大きなクラスタが形成されてしまい，類似度のみでは，小さなクラスタが少数形成されるのみになってしまうことがわかる．総合共起度を利用した場合は，直接共起度のみ，類似度のみの場合と比較して，二つ以上の地理オブジェクトを含むクラスタが多く形成されることがわかる．また，α の値が 3 以上になるとクラスタに含まれる地理オブジェクト数の分散もだいぶ小さくなり，クラスタのサイズも均等になると考えられる．

図 8 に示した関連度を利用したグラフ（$\alpha = 3$，エッジ 3000 本）の一部を拡大したものを図 12，13 に示す．図 12 はホテル名がまとまっている．名称の中に「ホテル」を含むもの以外にも,「ホリデイ・イン京都」や「ルビノ京都堀川」といった地理オブジェクトが含まれていることがわかる．単純に「ホテル」というフレーズが含まれているかどうかで分類しただけでは，これらの地理オブジェクトは無視されてしまう．図 13 では，美術館などの文化施設のクラスタが形成されていることが確認できる．一対一の共起関係に加えて，他の地理オブジェクトとの共起の類似度を考慮することで，より同類の地理オブジェクトを結びつけることができると考えられる．

$\alpha=0$	$\alpha=1$	$\alpha=2$	$\alpha=3$	$\alpha=4$	$\alpha=5$	$\alpha=6$	sim
594	326	171	97	78	53	45	42
9	22	25	30	30	32	32	33
6	17	22	26	26	31	27	24
4	15	20	26	25	26	26	23
4	13	18	22	22	25	25	22
4	13	18	21	21	23	22	22
4	13	17	17	19	21	21	21
3	12	16	15	18	18	20	19
3	12	15	15	17	17	17	17
3	11	14	14	15	15	15	14

表 3　上位 10 のクラスタに含まれる地理オブジェクトの数（エッジ 3000）
Table 3　The number of places in clusters in the top 10 of size

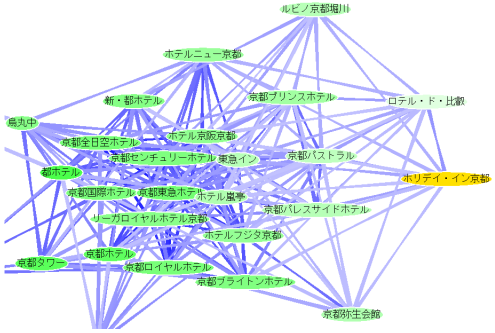

図 12　ホテルのクラスタ
Fig. 12　A cluster of hotels

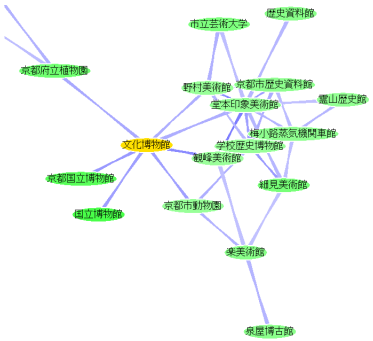

図 13　文化施設のクラスタ
Fig. 13　A cluster of museums

7.　地図インタフェースとの融合

総合共起度グラフ（$\alpha = 3$，エッジ 3000 本）を地図インタフェースと組み合わせて地理オブジェクトの検索システム実装する．地図インタフェースには，KyotoSEARCH の地図インタフェースを利用する．

ある場所を中心に，地図インタフェースにより，周辺の地理オブジェクトとの地理的関係を，グラフインタフェースにより，地理的には遠いが意味的に関連のある地理オブジェクトを検索できるシステムになっている．図 14 にこのユーザインタフェースの図を示す．グラフ中のノードを選択すると，地図上にその

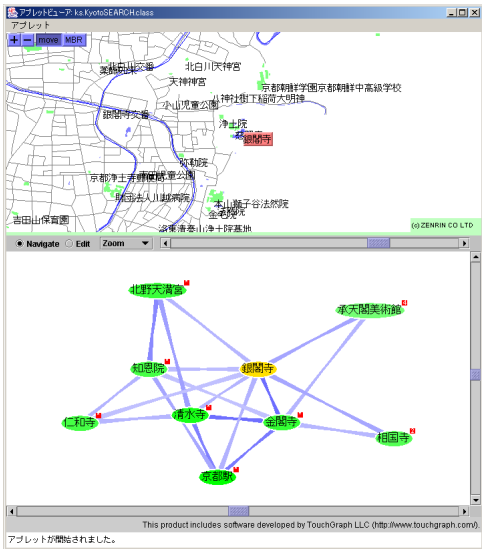

図 14　地理オブジェクトの検索システム

Fig. 14　the geographic object searching system

地理オブジェクトの周辺の様子が表示される．逆に，地図上の地理オブジェクトを選択すると，グラフ中でその地理オブジェクトのノードを中心に関連オブジェクトが表示される．地理オブジェクトの重要性や関連の強さを視覚的に把握しながら，情報検索を行うことができる．

8. おわりに

本稿では，地理情報検索システム KyotoSEARCH の地名間のナビゲーション機能を改良するために，希少度，類似度，総合共起度などの値を，地理オブジェクトの関連を対象に定義している．これらの値を利用することは，単純な地理オブジェクトの共起回数を使用するよりも歴史的背景などによる珍しい共起関係や発見したり同類の地理オブジェクトをまとめたりすることに役立つと考えられる．また，地理オブジェクトの関係を TouchGraph を用いて視覚化することで，重要な地理オブジェクトや強い関連を容易に把握することができる．

しかし，ただ単に関連する地理オブジェクトを提示するだけでは，不十分である．ユーザにとっては，地理オブジェクトがどのような意味で関連しているかという情報が重要であると考えられる．KyotoSEARCH では，ある地名に関して関連する地名と関連する一般名詞を抽出することを提案しているが，地理オブジェクト間の関連がどのようなものなのか，関連を特徴付けるキーワードを抽出することが考えられる．これは，二つの地理オブジェクトが共起するウェブページにどのような特徴的な名詞が多く現れるかを調べることなどが考えられる．

また，今回は，京都市内の代表的な地理オブジェクトのみを利用したが，これを商店名などに拡張すると膨大な量の地理オブジェクトを扱うことになり，特徴ベクトルを用いた類似度の計算は，多大な計算時間を要するため現実的でなくなってしまう．大量の地理オブジェクトを対象とするときには，計算量の面での考慮が必要になる．

総合共起度グラフにおいて，完全に他と切り離されたグラフになっていなくても，他の部分に比べて，密な部分グラフになっている個所が見られた．このような密な部分グラフを抽出する手法も考案されており [11], [12]，これらの手法を利用してグラフを解析することも考えられる．

謝　辞

本研究の一部は科学技術振興機構 (JST) 戦略的創造研究推進事業・CREST における「デジタルシティのユニバーサルデザイン」による支援を受けています．ここに記して謝意を表します．

また，本研究は京都大学情報学研究科上林研究室で開発されている地理情報検索システム KyotoSEARCH のモジュール，データを利用しています．開発メンバーである手塚太郎氏，井上陽介氏に謝意を表します．

また，本稿を改善する上で有益なご意見をいただいた査読者の方々に感謝いたします．

文　献

[1] 李龍，高倉弘喜，上林弥彦，“ 地域ウェブ情報を利用した地域情報検索と地域分析，”第 2 回空間情報 IT ワークショップ (特集：「デジタル認知空間」)，Dec. 2001.

[2] R. Lee, Y. Inoue, T. Tezuka, N. Yamada, H. Takakura and Y. Kambayashi, ”KyotoSEARCH: A Concept-based Geographic Web Search Engine,” Second IRC International Conference on Internet Information Retrieval, pp.139-147, Koyang, Korea, Nov. 2002.

[3] 横路誠司，高橋克巳，三浦信幸，島健一，”位置指向の情報の収集，構造化および検索手法，”情処学論，Vol.41，No.7，pp.1987-1998，2000.

[4] 平松薫，小林堅治，Ben Benjamin，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，”情処学論，Vol.41，No.12，pp.3314-3322，2000.

[5] 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，”情処学論，Vol.41，No.SIG6(TOD7)，pp.69–80，2000.

[6] 相良毅，有川正俊，坂内正夫，”分散位置参照サービス，”情処学論，Vol.42，No.12，pp.2928–2940，2001.

[7] 増井俊之，”情報視覚化技術，”UNIX MAGAZINE，1998 年 6 月号，pp.161–167，1998.

[8] TouchGraph LLC, http://www.touchgraph.com/

[9] 山田直治，李龍，高倉弘喜，上林弥彦，”地理的スコープと詳細度による WEB ページ分類とモバイルキャッシュへの応用，”信学技報，Vol.102，No.209，pp.109-114，2002.

[10] 馬強，松本知弥子，田中克己，”ページ内容と位置情報に基づく Web コンテンツのローカル度検出とその応用，”信学技報，Vol.102，No.209，pp.115–120，2000.

[11] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, ”Self-Organization and Identification of Web Communities,” IEEE Computer, Vol.35, No.3, pp.66–71, 2002.

[12] M. Girvan and M. E. J. Newman, ”Community Structure in Social and Biological Networks,” http://arxiv.org/abs/cond-mat/0112110/, 2001.

[13] 李航，安倍直樹，”共起データに基づく単語クラスタリング法，”自然言語処理シンポジウム「実用的な自然言語処理に向けて」，Nov. 1997.