ウェブ地域情報の自動要約のための特徴キーワード抽出

中戸隆一郎† 岩井原瑞穂††

† 京都大学工学部 〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院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
E-mail: †rnakato@db.soc.i.kyoto-u.ac.jp, ††iwaihara@i.kyoto-u.ac.jp

あらまし Web 上での地域情報検索において、検索結果に含まれる類似ページ集合をクラスタリングによってまとめ、要約を得るといった既存の研究が知られている。しかしクラスタとして得られるトピックは多岐にわたり、利用者が見たいトピックを探すのに手間がかかるという問題点がある。そこで本研究では、検索対象である地理オブジェクトに対して利用者の求めるトピックを多く含む補助的な地理オブジェクトを指定し、地理オブジェクト間で共通するトピックに高い重要度を与えることで、ユーザの求めるトピックを優先的に提供できるような地域情報検索システムを提案する。本稿ではプロトタイプシステムを用いた実験結果より、この重要度決定手法「トピックエンハンス法」を用いて要求するカテゴリに特化した検索結果が実際に得られていることを確認した。
キーワード GIS，Web マイニング，情報検索，キーワード抽出，クラスタリング，自動要約

Keyword extraction for automatic summarization of regional information on the Web

Ryuichiro NAKATO† and Mizuho IWAIHARA††

† Faculty of Engineering, Kyoto University Yoshida-honmachi, Sakyo-ku, Kyoto, 606–8501 Japan
†† Department of Social Informatics Graduate School of Informatics, Kyoto University Yoshida-honmachi, Sakyo-ku, Kyoto, 606–8501 Japan
E-mail: †rnakato@db.soc.i.kyoto-u.ac.jp, ††iwaihara@i.kyoto-u.ac.jp

Abstract For geographic information retrieval on the web, there is an existing research for summarizing document set by clustering similar web pages. However, this has a problem that the clustering result contains various topics and it is difficult for a user to find a topic needed from these clusters. In this paper, we propose a geographic information retrieval system that can present a requested topic by priority. By specifying not only a geographic object to be summarized but also supplementary geographic objects which contain topics belonging to a category which a user needs, this system gives high importance to a cluster which has co-occurred keywords between them. In this paper, we show that by the importance decision method named ”Topic Enhance Method,” categories which a user needs can be selected by using a prototype system.
Key words GIS，web mining，information retrieval，keyword extraction，clustering，automatic summarization

1. はじめに

1.1 研究背景

近年のインターネットの普及に伴い、ウェブを通して様々な情報を簡単に入手できるようになった。地域情報検索においても、観光情報、レストラン情報、バスの時刻表などの情報を知るといった目的で Web 情報を用いることは多いが、Web 上には膨大な数のページが存在することから、既存の検索エンジンによる Web 検索には様々な問題が存在する。

Google のようなロボット型検索エンジンを用いる場合、検索結果の適切な絞り込みの難しさが問題となる。例えば「京都大学の周辺の情報」を検索する要求に対して、「京都大学　周辺」や「京都大学　観光」といった AND 検索では周辺情報と関係ないページも多くヒットしてしまい、真に適切な検索結果は得られない。さらに、上位にランクされたページ集合内で内容が重複しているページも多く、ランクの高いものから順に閲覧する方法も検索効率は良いとは言いがたい。このように検索結果を絞り込むためのキーワードが一語でうまく表せない場合はキーワードのブール論理検索だけでは不十分である。

一方で、Yahoo!のような Web ディレクトリ型サーチエンジン

を用いれば、ユーザはカテゴリ情報を用いて効率的な Web ページの絞り込みが可能である。しかしこの場合、Web ページのカテゴリ分類作業は全て人手で行われるため、検索エンジンの維持に大きなコストがかかってしまうことが問題となる。

本研究ではこれらの問題を解決するために、前者のロボット型検索エンジンにおいて、検索結果中の類似ページをクラスタリングによってまとめ、さらに各クラスタをカテゴリに基づく重要度を与えることでランキングを行い、クラスタ内容を要約することで、ユーザの求めるカテゴリに属するクラスタを優先的に提示できるようなシステムを提案する。

類似する Web ページ群をクラスタ化し、内容を特徴づけるキーワードを抽出することにより文書集合の要約を得る手法はニュース要約などで既に実用化されているが、これらの手法は結果に現れるトピックが多岐にわたり、見たいトピックを探すのに手間がかかるという問題点がある。本研究ではこれらの手法に重要度決定を付加することで、ユーザによるトピック指定やトピック間の重要度付けのコストを軽減する。

1.2 研究概要

本研究ではクラスタリング結果の各クラスタへ重要度を与える方法として、ユーザが地理オブジェクトに対して持つ暗黙的な知識を利用することを考える。

例えば「金閣寺」という語を含むページは寺社や観光に関するトピックが多く含まれると考えられるが、「京都大学」という語を含むページには大学の学術活動に関するページが多く含まれる一方で、大学近辺での地域情報に関するページも存在していると考えられる。そこでユーザに、要約対象の地理オブジェクト O_A の他に、ユーザが知りたいカテゴリに属するトピックを多く含むであろう補助的な地理オブジェクト O_B も指定させ、O_A と O_B の双方でクラスタリングを行った上で、各クラスタより内容を表す特徴キーワードを抽出する。ここで、双方のクラスタ集合より抽出したキーワード集合のうち、共起しているキーワードは両方の地理オブジェクトに共通したトピックを特徴づけると考えられる。そこに着目すれば、共起しているキーワードを含むクラスタに高い重要度を与えることにより、ユーザの得たいトピックを含むクラスタを上位に含む要約結果が得られると考えられる。

例えば、要約対象を「京都大学」とした時、補助的に「金閣寺」を与えることにより、京都大学の要約でありながら、金閣寺と共起するであろう京都大学及びその周辺の観光情報、歴史についてのトピックが重視されることになる。この手法によれば、ユーザは要約対象となる地理オブジェクト（以下、ターゲットオブジェクトと呼ぶ）と補助的な地理オブジェクト（以下、ランキング制御オブジェクトと呼ぶ）の２つを与えるのみであり、トピックの指定や重要度づけは一切必要ない。また、本手法のキーワード共起は、京都大学と金閣寺のページ内共起といった伝統的な検索方式ではなく、クラスタリング結果に現れるキーワード群の共起であり、ページ内共起とは異なる。

地理オブジェクト間でクラスタのトピックが共通するということは、言い換えれば地理オブジェクトの持つ概念が共通するということである。すなわち本研究で提案するシステムは、「京都大学の検索結果のうち、金閣寺の概念と共通するもの」といった、共通概念による絞り込み方式ということができる。本研究では、このような地理オブジェクトの共通概念に基づくクラスタの重要度決定法を「トピックエンハンス法」と名付けた。本稿では京都府内の地域情報を表す Web ページ集合を用いて実験を行い、トピックエンハンス法に基づくカテゴリ分類の妥当性・有効性を示す。

以下 2 章では本研究に関連する研究を紹介する。3 章でアルゴリズムの詳細を述べ、4 章ではキーワード集合の共起関係の観点からトピックエンハンス法の有効性を検証する。5 章では実際に作成したプロトタイプシステムによる実験結果よりトピックエンハンス法によってユーザの要求するカテゴリに特化した検索結果が得られることを示し、6 章でまとめと今後の課題について述べる。

2. 関連研究

2.1 カテゴリ分類を用いた Web 情報検索システム

検索結果を自動的にクラスタ化して提示する検索システムとして Vivisimo [8] がある。Vivisimo は検索要求に対し、検索エンジン、ディレクトリサービス、ニュース検索サイトなどを横断的に検索し、重複を整理した上で検索結果を階層的にソートされたカテゴリフォルダに自動分類し、ユーザに提示する。本研究の優位点としては、各カテゴリを並列に提示するのではなく、ターゲットオブジェクト O_A と別にランキング制御オブジェクト O_B を指定することで要求されるカテゴリを優先的に表示できることが挙げられる。ユーザは「金閣寺」「京都駅」「京都大学」といった地理オブジェクトそれぞれにどのようなトピックが存在するかについて大まかに想像できることから、O_A に対して得たいと思っている話題を多く持つであろう地理オブジェクト O_B を与えるといった少ない負担で、得たいトピックを扱うクラスタを優先的に参照することができる。

また大槻 [5] は、地域情報に関する Web ページを８種類のカテゴリに自動分類することにより、Web 地域情報ディレクトリの自動編集を行うシステムを提案している。このシステムは、地域情報に関する Web ページをあらかじめ収集し、カテゴリに自動分類するという点で本研究と共通点が存在するが、ここで扱われる地名は都道府県名及び市町村名のみであり、観光地や大学等の地理オブジェクトは含まれていない。また、カテゴリを人手で作成していること、カテゴリの自動分類に辞書を用いていることも本研究と異なる点である。本研究では検索結果に対してカテゴリや概念を直接指定するのではなく、そのような概念を持つであろう他の地理オブジェクトを指定することのみによって検索結果を希望のカテゴリに絞り込む。このような手法は絞り込みのための概念が漠然としていてはっきり表せない場合などには特に有効であると考えられる。

2.2 Web からの地理オブジェクトの関係抽出

Web 空間から地名集合の相関関係を抽出する研究も存在する。

地域情報検索システム KyotoSEARCH [2], [3] におけるキーワード空間では、地域情報に関連した Web ページから抽出した特徴キーワードを「地名」と「非地名」に分けた上で関連の

強さを計算することで、各地名についての特徴及び関連地名を知ることができる。ここで関連の強さの計算にはキーワードのページ内共起を用いているため、ページ内で共起しない地名や単語については関連の強さを知ることができない。

これに対して椎名ら [6] は、ページ内の地名共起の度合いを表す「直接共起度」の他、「共起する地名の共起」の度合いを表す「類似度」を定義し、これらを組み合わせることによって得られる「総合共起度」から地名間の意味的関連の強さを求める手法を提案している。この総合共起度を用いて地名集合をクラスタリングすることにより、「大学」「ホテル」などの属性ごとに地名を分類することに成功している。この手法には、未知の関連の強さに対してはその要因を特定できないという問題点があるが、本研究では地理オブジェクト間の関連を「共通概念」という形で具体的にクラスタ及び特徴キーワードとして抽出することができるため、直感的にわかりにくい関連であった場合も意味関連の強さを示す要因を特定することが可能である。

3. ウェブ地域情報の自動要約アルゴリズム

本研究で用いるアルゴリズムは、以下の４つの段階から成る。

（1） 地理オブジェクトに関する Web ページを収集
（2） 収集した Web ページ集合をクラスタ化
（3） トピックエンハンス法による各クラスタの重要度計算
（4） クラスタを重要度順に表示し、内容を要約して提示

以下、3.1 節〜3.4 節でそれぞれの段階の詳細を述べる。

3.1 Web ページ収集

まず、検索クエリとなる地理オブジェクトに関する Web ページをあらかじめ収集する。本来的には Web 空間全体より Google Web API などを用いて収集することが想定されるが、本稿の実験においては手塚 [7] により実装されている Focused Web Crawler を用いて収集された京都に関するページ 15 万ページを用い、そこから再検索する形で Web ページを収集する。

Focused Web Crawler は、京都関係の特定ページをスタートページとしてそこからリンクを辿るかたちでクローリングを行い、京都地名(注1)の存在しないページに辿りつくとそこで停止する。これにより「(金沢市) 寺町」など、京都に同名の地名が存在するために含まれてしまう京都以外の地名に関するページを収集してしまうことを避けられる。

3.2 クラスタリング

収集した Web ページ集合をクラスタリングし、クラスタ集合とする。クラスタリングには namazu を用いたクラスタリングツール gnmz [1] を用いる。gnmz は法情報学の分野で類似文書や関連文書を検索するための実験ツールとして開発されたシステムであり、テキスト集合をクラスタ化し、さらに各クラスタから特徴キーワードを抽出する機能を持つ。以下詳細を述べる。

(1) namazu を用いたテキスト集合のクラスタリング

対象となるテキスト（ここでは Web テキスト）の集合に対して、namazu によるインデキシングを行う。gnmz はそのインデックス情報を用いて Web ページの文書間近接度を計算し、クラスタ化する。文書間近接度の計算には Kullback-Leibler divergence (KLD) を用い、クラスタリングアルゴリズムには K-Means 法及び Ant Colony Optimization (ACO) [4] を用いている。

(2) 各クラスタからの特徴キーワード抽出

クラスタ化した後、各クラスタのテキストより tf・idf 法を用いて特徴キーワードを抽出する。（本実験においては、抽出する特徴キーワード数を 1 クラスタにつき 5 個としている。）

gnmz を用いた例として、テキストに「金閣寺」を含む Web ページ集合から抽出されたキーワード集合の一部を表 1 に示す。

地理オブジェクト名	キーワード集合
金閣寺	相国寺、義、炎上、関西、 足利義満、金閣、株式会社、 京料理、系統、展、観光、引用、 古墳、米、返信、イラク、…

表 1　金閣寺のキーワード集合（一部）

3.3 トピックエンハンス法による重要度計算

生成されたクラスタ集合及びキーワード集合を用いて、トピックエンハンス法による各クラスタの重要度計算を行う。

重要度計算には、ランキング制御オブジェクトとのトピックの共通関係を利用する。共通するトピックを扱っているかどうかは、地理オブジェクト間のキーワード集合の共起関係から判定する。以下に詳しい考え方を述べる。

キーワードが複数の地理オブジェクトのキーワード集合に共起する要因としては以下の２つが考えられる。

（1） **地理オブジェクト名自体がテキスト内に共起する場合**

地理オブジェクト A と地理オブジェクト B の２つがテキスト中に現れる Web ページが存在した場合、そのページは A について収集された Web ページ集合と B について収集された Web ページ集合の両方に含まれることになる。このようなページからクラスタが形成された場合、クラスタから抽出されるキーワードも同じものになる可能性がある。

（2） **地理オブジェクト間に共通する概念がある場合**

地理オブジェクト A と B それぞれの Web ページ集合が共通するページを含まなかったとしても、例えば「地理オブジェクト A の紅葉情報について書かれたページ群から成るクラスタ」と「地理オブジェクト B の紅葉情報について書かれたページ群から成るクラスタ」が存在すれば、「紅葉」というキーワードが共起する。これは、地理オブジェクト A と B が共に紅葉に関係している、すなわち「紅葉」という共通概念を持っていることを表している。このように、クラスタ間の特徴キーワードの共起関係に着目することで、単純な地理オブジェクト同士のページ内共起からは知ることのできない共通概念を発見できる可能性がある。

具体例として、「金閣寺」と「清水寺」の間には「寺」「観光地」「紅葉」などといった共通概念が存在するが、「金閣寺」と「立命館大学」の間には、「京都市北区に存在」という以外の共通概念が直感的には思いつかない。このような地理オブジェクト間の共通概念の有無がキーワード集合の共起関係の傾向とし

(注1)：京都地名データベースには、ゼンリン [9] の電子住宅地図を用いている。

て表れているとすれば、そのキーワードが現れるクラスタを抽出することによって、それらの概念に即した内容を表しているクラスタを抽出することも可能である。

以上の考え方に基づき、各クラスタの重要度を以下のように定義する。

ターゲットオブジェクト O_A について、O_A のクラスタ集合を C、それに含まれる各クラスタを c_i、クラスタ c_i のキーワード集合を k^{c_i}、1つのクラスタから抽出される特徴キーワード数を n_e、クラスタ c_i に付されたキーワードを $k_j^{c_i}$ $(1 \leqq j \leqq n_e)$、O_A のキーワード集合全体を K_A とする。

ここで O_A に対して指定する r 個のランキング制御オブジェクトのうち p 個が AND クエリ、$q(=r-p)$ 個が NOT クエリであるとし、それぞれを $O_{B_1^+},\ldots,O_{B_p^+}$、$O_{B_1^-},\ldots,O_{B_q^-}$ とすると、クラスタ c_i の重要度 $Imp(c_i)$ は以下の (1) 式で表される。

$$Imp(c_i) = \mid k^{c_i} \cap K_{imp} \mid \tag{1}$$

ここで K_{imp} は

$$K_{imp} = K_{AB_1^+B_2^+\ldots B_p^+\overline{B_1^-B_2^-\ldots B_q^-}} \tag{2}$$

$$= K_A \cap \bigcap_{m=1}^{p} K_{B_m^+} - \bigcup_{n=1}^{q} K_{B_n^-} \tag{3}$$

である。上式より、$Imp(c_i)$ は区間 [0,n_e] の間の整数値を取る。O_B を変化させることによって K_{imp} も変化するので、重要度 $Imp(c_i)$ の値は O_B を利用者が選択することにより決定される。

3.4 クラスタ内テキストの自動要約

3.3 節で定義した重要度に基づいてクラスタ集合をランキングし、クラスタのテキストを自動要約してユーザに提示する。

プロトタイプシステムでは、テキストの自動要約に HTML parser モジュールを用いる。クラスタに含まれる Web ページテキストを HTML parser によって最小タグ領域に分解し、分解されたテキストの中でクエリとなる地理オブジェクトを含んでいるものを元となる Web ページのページタイトルと共に要約文として提示する。

4. キーワード分類によるトピックエンハンス法の評価

この章では 3.3 節の (1) 式で定義した重要度によってどの程度クラスタ集合のカテゴリを分類できるかをキーワード集合の観点から検証する。すなわち、ターゲットオブジェクト O_A 及びランキング制御オブジェクト O_B について、共起キーワード集合 K_{AB} $(=K_A \cap K_B)$、非共起キーワード集合 $K_{A\overline{B}}$ $(=K_A - K_B)$ がそれぞれ実際に O_A と O_B の共通概念、非共通概念を表したものとなっているかを検証する。

4.1 節では、検証方法の詳細について説明する。4.2 節では共起キーワード集合 K_{AB} を抽出した場合の結果、4.3 節では非共起キーワード集合 $K_{A\overline{B}}$ を抽出した場合の結果について述べる。最後に 4.4 節では、キーワード集合の割合分布の変化が実際にクラスタのランキングの変化に結びついていることを示す。

4.1 検証手法

前準備として、ターゲットオブジェクトとするサンプル地理オブジェクトのキーワード集合を手作業でカテゴリ分類する。自動的に抽出されるカテゴリを利用することが本研究の目的であるが、本章では出力されるクラスタの概念が実際にどのように変化しているかを検証するため、各クラスタをあらかじめ手動でカテゴリ分類し、検索結果に含まれるカテゴリの全体に対する割合の変化を計算する。

分類するカテゴリは、「総合・仏教・行事・文化・掲示板・ニュース・リンク集・生活情報・名所・宿泊・交通・ショッピング」の 12 種類とする。各カテゴリにどのようなトピックのテキストが分類されるかについて、表 2 にその例を示す。これらのカテゴリは、関連研究 [5] で用いられているカテゴリ及び実際の京都情報サイトで使用されているカテゴリを元に作成した。

カテゴリ名	含まれる Web ページ
総合	京都全体を扱う総合情報サイト (KyotoNavi など)
仏教	寺院一覧・宗派別紹介など
行事	祭、イベント情報、歳時記など
文化	地理オブジェクトそのものの紹介、歴史など
生活情報	不動産・銀行・郵便など
名所	観光地としての見所、散策コースの紹介など

表 2　手動分類によるカテゴリ一覧 (一部)

カテゴリ分類は、まず各クラスタ c_i に対して、12 個あるカテゴリのうちどのカテゴリに分類されるかを付されたキーワード及びテキストの内容に基づいて人間が主観的に決定する。全てのクラスタのカテゴリが決定されたら、各クラスタについて、そのクラスタに付されているキーワードを全てそのカテゴリに分類することにする。すなわち

$$c_i \mapsto g_p \qquad \Rightarrow \qquad k_j^{c_i} \mapsto g_p \ (1 \leqq j \leqq n_e) \tag{4}$$

となる。ここで g_p はカテゴリの要素であり、$c_i \mapsto g_p$ はクラスタ c_i がカテゴリ g_p に属することを表す。

1 つのキーワードが複数のクラスタに現れる場合は、1 つのキーワードが複数のカテゴリに属することを許した上で、属するカテゴリを延べ回数で表す。すなわち、例えばキーワード「京都市」がカテゴリ「総合」に属するクラスタ c_i、同じくカテゴリ「総合」に属するクラスタ c_j、カテゴリ「文化」に属するクラスタ c_k の計三回現れるとすると、キーワード「京都市」は「総合 2、文化 1」の延べ 3 のカテゴリに属することになる。

以上の方法により、キーワード集合をカテゴリ分類した。表 3 は各カテゴリに現れる代表的なキーワードを、いくつかのカテゴリについて例示したものである。

本章の実験では、ターゲットオブジェクトのサンプルとして

カテゴリ名	代表的なキーワード
仏教	仏壇、法名、浄土真宗、…
文化	文化遺産、新撰組、舞妓、…
掲示板	送信、返信、表示、…
交通	市バス、時刻表、系統、…
ショッピング	京料理、グルメ、あぶらとり紙、…

表 3　各カテゴリのキーワード（一部）

「金閣寺」と「西本願寺」の２つを用いる。

金閣寺は正式名称を鹿苑寺と言い、世界文化遺産として登録されている。足利義満の別荘として建設されたこの寺は京都を代表する観光地の一つであり、Web 上においても仏寺としてではなく観光地として紹介されることがほとんどである。また「金閣寺道」「金閣寺町」などと近隣地名にも用いられているため、不動産物件などの Web ページにも多く現れる。

西本願寺は浄土真宗本願寺派の本山であり、Web 上においても観光地としてではなく仏教の本山として紹介されていることが多い。また金閣寺と同じく世界文化遺産として登録されており、最も重要な建造物である御影堂の修復工事が 1998 年から行われているため、関連したニュース記事などのページも多い。

4.2　共起するキーワードのみを抽出する場合

「金閣寺」と「西本願寺」について、3.1 節、3.2 節で説明した手法によって Web ページを収集しクラスタリングした結果、金閣寺について 1138 ページ 71 クラスタ、西本願寺について 712 ページ 52 クラスタが生成された。

実験１：金閣寺

まず金閣寺について見る。概念を共起させるランキング制御オブジェクトとして、金閣寺と並んで京都の代表的な観光地である「清水寺」、同じく観光地であり地理的に金閣寺と近い「仁和寺」、地理的に近いが直接的なつながりはない「立命館大学」の３つを選ぶ。金閣寺に対して、それぞれの地理オブジェクトが異なった共通概念を持つので、共起キーワード集合の割合変化は異なるはずである。

4.1 節の手法によってキーワード集合のカテゴリ分類を行い、金閣寺全体のキーワード集合 K_A と、各ランキング制御オブジェクトとの共起キーワード集合 K_{AB} のそれぞれについてカテゴリ別に属するキーワードの総数を集計し、割合を計算し比較したものを図 1 にまとめた。

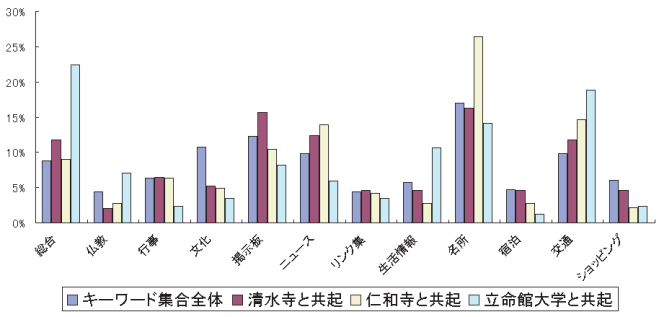

図１　共起キーワード集合のカテゴリの割合分布

この結果を見ると、仁和寺の場合、「名所」の割合が増加し、「生活情報」の割合が減少する傾向にある。これは観光地同士でキーワード集合の共起を取ったため、観光案内的なトピックが共起し、不動産などの生活情報は共起しなかったことが考えられる。また「交通」の割合も増加しているが、これは金閣寺と地理的に近く、金閣寺と仁和寺を結ぶ道が「きぬかけの道」と呼ばれ有名な観光ルートとなっていること、「ショッピング」の割合が減少しているのは仁和寺の周辺にみやげ物店などのお店があまり存在していないことから直感的に理解できる。

金閣寺と地理的に近接しているが観光地ではない立命館大学の場合、「総合」「交通」「生活情報」「掲示板」「ニュース」カテゴリの占める割合が増え、「行事」や「宿泊」といったカテゴリの割合が減少する傾向が見られる。これは、観光関連のトピックが共起しないために、市バスや不動産などの地理的なカテゴリや、どのような地理オブジェクトにも存在する掲示板や新聞記事、多くのカテゴリを網羅的に扱う総合サイトのクラスタがクローズアップされたことを示している。

清水寺との共起キーワード集合の割合は、他の２つの場合と比べてカテゴリ分布の割合がキーワード集合全体の場合と最も近い。この理由として、金閣寺と清水寺の共通概念が非常に多く存在し、この 12 種類のカテゴリ分類では金閣寺固有の概念が見つけられないことが考えられる。しかし清水寺についても、今回と違ったカテゴリ分類を行うことで共通しない概念が結果として出てくる可能性がある。

実験２：西本願寺

２つめのサンプルとして西本願寺を用いる。

西本願寺は金閣寺と比べて仏教カテゴリに属するクラスタの割合が多く、キーワード集合にも「浄土真宗」「住職」「法名」「念珠」といった仏教特有のものが数多く含まれていた。

西本願寺に対するランキング制御オブジェクトには、西本願寺と並び浄土真宗のもう一つの本山である「東本願寺」、西本願寺と異なる宗派の臨済宗相国寺派に属する「金閣寺」、西本願寺と距離的に近く観光地でない「JR 京都駅」、西本願寺と特に直接的な関係を持たない「京都国立博物館」の４つを用いた。

金閣寺の場合と同様に、これらのランキング制御オブジェクトと共起するキーワードを抽出した場合のカテゴリの割合分布のうち、特徴的だったカテゴリの変化を図 2 に示す。

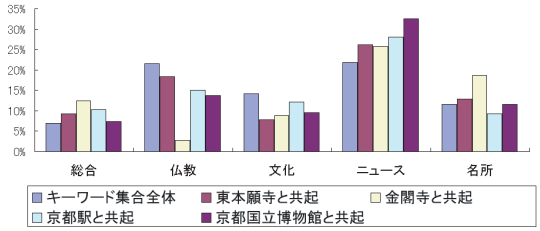

図２　キーワード集合を共起させた場合のカテゴリ割合の特徴的な変化

「仏教」カテゴリについて見ると、東本願寺の場合が最も割合が多く、逆に金閣寺の場合が最も割合が少ない。これは、西本願寺の「仏教」カテゴリを表すキーワードの中で出現回数の多かった「浄土真宗」「真宗」と言った単語が、同じ浄土真宗である東本願寺では共起するが宗派の違う金閣寺では共起しなかったことが理由として挙げられる。また金閣寺の場合、「名所」カテゴリの割合の増加率が最も多いことからも、この結果は金閣寺は Web 上で仏寺ではなく観光地として認識されていることを裏づけていると言える。

京都駅と共起させた場合、キーワード集合全体と比較して、全体的に割合分布が似通っているが、これは京都駅が西本願寺と距離的に近く、また京都駅は交通の起点として引用されるため結果的に西本願寺と京都駅が頻繁にページ内共起しており、

3.3 節の (2) ではなく (1) の理由によってキーワードが共起していることが原因と考えられる。

京都国立博物館の場合、「ニュース」カテゴリの伸びが特に大きい。理由としては、西本願寺と特に共通する概念が無いために、どの地理オブジェクトにも存在し得るニュース関連のキーワードが多く共起し、共起キーワード集合の中で見た場合に割合として増加することがある。このような傾向は、関連が薄い地理オブジェクトと共起させた場合に共通して起こっていた。

4.3 共起しないキーワードのみを抽出する場合

次に、非共起キーワード集合 $K_{A\overline{B}}$ について考える。

4.2 節の実験を見ると、キーワード集合 K_A と共起キーワード集合 K_{AB} を比較した場合に、「文化」「ショッピング」のようにどのような地理オブジェクトと共起させても割合が減少する傾向のあるもの、「総合」「ニュース」のようにどの地理オブジェクトと共起させても割合が増加する傾向のあるものが存在する。

共起させた場合に常に割合が減少するカテゴリは、他の地理オブジェクトと共起しにくいキーワードを多く含むカテゴリであると言える。そのようなキーワードには「足利義満」「浄土真宗本願寺派」といった固有性の高いキーワードや、「京菓子」「キムチ」「工芸品」などの細かい商品名などが相当する。

逆に「第」「回」「日」「email」などのように、一般的でありどのようなトピックのクラスタからも抽出される可能性のあるものや、「掲示板」や「京都新聞」のように、どの地理オブジェクトにもヒットするページから成るクラスタから抽出されるキーワードは共起キーワードに含まれることが多く、従って相対的なカテゴリの割合が増加する。

ここでランキング制御オブジェクトを NOT クエリとし、ターゲットオブジェクトのキーワード集合のうちランキング制御オブジェクトと共起しない非共起キーワード集合 $K_{A\overline{B}}$ を 3.3 節の (1) 式における重要度 Imp 値の計算に用いれば、掲示板などの雑多なクラスタの重要度が下がり、固有度の高いクラスタを抽出できる可能性がある。ここで西本願寺と京都国立博物館の場合を例にとり、そのような傾向が実際に現れるかを検証する。

西本願寺をターゲットオブジェクト O_A、京都国立博物館をランキング制御オブジェクト O_B とし、キーワード集合 K_A、共起キーワード集合 K_{AB}、非共起キーワード集合 $K_{A\overline{B}}$ それぞれのカテゴリ割合を図 3 に示す。

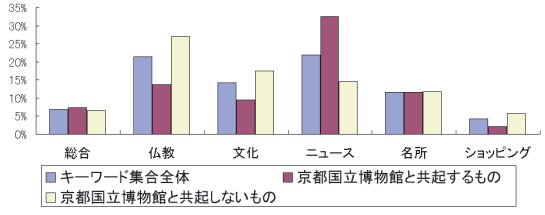

図 3　共起・非共起キーワード集合のカテゴリ別割合分布の比較

西本願寺のキーワード集合に対して京都国立博物館を共起させた場合、「ニュース」カテゴリが増加し、「仏教」「文化」が減少するという傾向にあったが、非共起キーワードを抽出すると、逆に「ニュース」カテゴリのキーワードが減少し、「仏教」「文化」が増加する。これより、「仏教」や「文化」についてのトピックに興味があり、「ニュース」関連のトピックは必要ないというユーザは、例えば「京都国立博物館」を NOT クエリとして指定すれば良いということが言える。

このように共通しない概念に属するクラスタを抽出することで、欲しいクラスタの抽出のみならず、不要なクラスタの除外も可能となる。さらに、どの地理オブジェクトとも共起しないようなキーワードを抽出することで、固有性の高いトピックを扱うクラスタを抽出することも可能であると考えられる。

4.4 キーワード集合のカテゴリ割合の変化による重要度ランキングの変化

本節では、キーワード集合の割合分布の変化が実際にクラスタのランキングの変化に結びつくことを示す。

4.2 節の金閣寺の場合を例にとり、それぞれの共起キーワード集合を用いて 3.3 節の (1) 式により重要度を決定した場合、上位 10 位にどのようなカテゴリのクラスタがランクされるかを表 4 に示した。これを見ると、それぞれのランキング上位には、清水寺の場合は「掲示板」「ニュース」、仁和寺の場合は主に「名所」「ニュース」、立命館大学の場合は「総合」「生活情報」のカテゴリに属するクラスタが多く現れている。この上位にランクされるカテゴリの違いは、4.2 節の図 1 で示した各キーワード集合のカテゴリ割合の分布の差と一致する。

これより、トピックエンハンス法におけるクラスタの重要度ランキングは実際にキーワード集合のカテゴリ別割合分布に依存しており、従って本章の実験結果は 3.3 節での重要度決定式の有効性を示すものであると言える。

ランキング	清水寺と共起	仁和寺と共起	立命館大学と共起
1 位	掲示板	名所	総合
2 位	ニュース	交通	総合
3 位	ニュース	名所	交通
4 位	生活情報	名所	総合
5 位	掲示板	ニュース	生活情報
6 位	ニュース	仏教	総合
7 位	総合	ニュース	総合
8 位	交通	ニュース	生活情報
9 位	リンク集	ニュース	ニュース
10 位	交通	名所	リンク集

表 4　上位 10 クラスタの分類カテゴリ一覧

5. プロトタイプシステムでの地域情報自動要約

4 章での実験結果より、本研究で提案するトピックエンハンス法の有効性が確認された。本章では実際に作成したプロトタイプシステムを使って、トピックエンハンス法を用いたクラスタの重要度ランキングの有効性を示す。

5.1 節でプロトタイプシステムの概要について説明し、5.2 節では AND クエリ・NOT クエリを適切に指定することによって要求するカテゴリに特化した検索結果が得られることを示す。

5.1 検索インタフェース

3 章で説明したアルゴリズムに基づき設計した自動要約シス

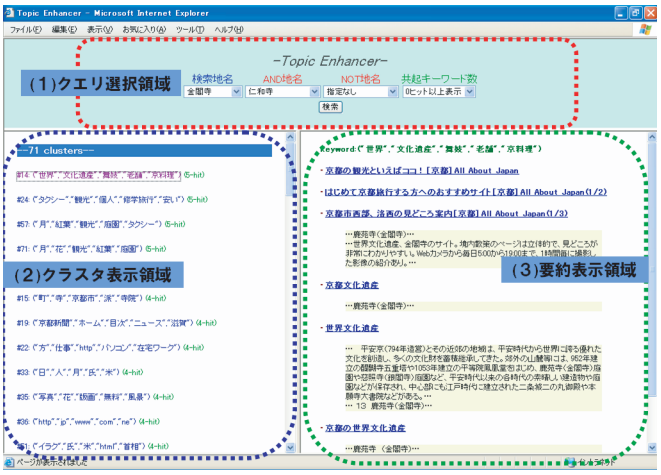

図 4　自動要約インタフェースのプロトタイプシステム

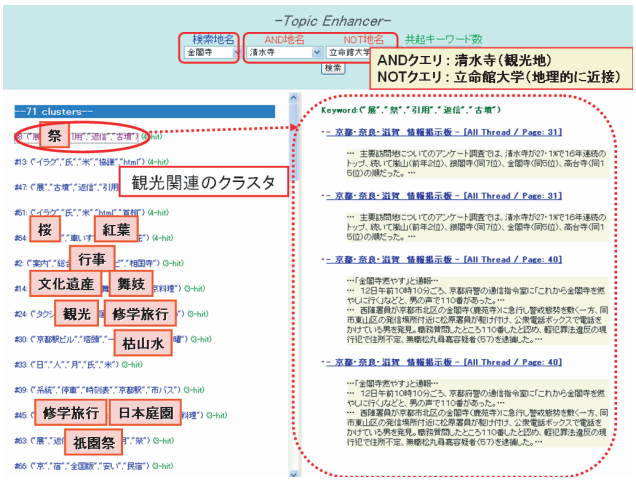

図 5　金閣寺の観光情報を検索

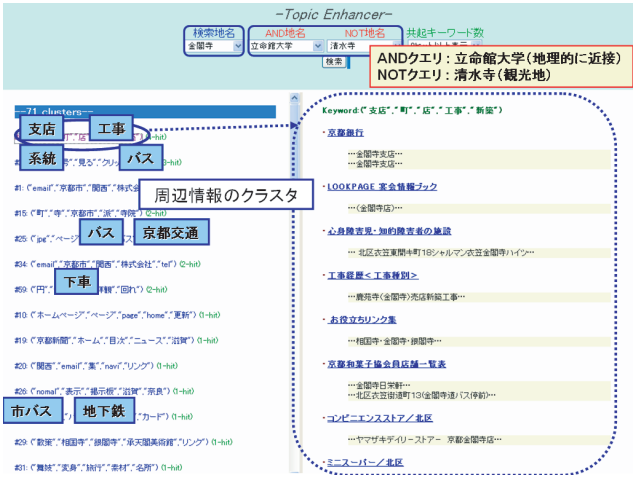

図 6　金閣寺の周辺情報を検索

テムのプロトタイプの検索画面を図 4 に示す。開発言語としては、プログラム部分に perl 及び PostgreSQL、インタフェース部分には HTML 及び PHP を用いている。

インタフェースは、以下の 3 つの部分で構成される。

- クエリ指定を行う検索フォーム (1)
- クラスタ及びキーワード一覧の表示領域 (2)
- クラスタの要約テキスト表示領域 (3)

ユーザが (1) の検索フォームよりターゲットオブジェクトを指定すると、(2) のクラスタ表示領域にクラスタ一覧が表示される。ランキング制御オブジェクトを指定すると、トピックエンハンス法に基づいて各クラスタの重要度を計算し、ランキングして表示する。ユーザは興味のあるキーワードを持ったクラスタを選択することによって、(3) の要約表示領域に表示される要約文からクラスタの内容を知る。要約文中のページタイトルをクリックすれば、元となる Web ページに飛ぶこともできる。

5.2　共通概念の抽出による検索結果の変化

検索結果例として、「金閣寺」「同志社大学」をターゲットオブジェクトに用いた場合の検索結果を図 5〜8 に示す。

- **金閣寺について検索**

（1）　観光情報について知りたい場合を想定（図 5）

金閣寺の観光情報を知りたいユーザを想定し、観光地として知られる清水寺を AND クエリ、地理的に近いが観光地ではない立命館大学を NOT クエリとして検索を実行した。

その結果、「祭」「紅葉」「桜」「舞妓」といったキーワードを持つクラスタが上位にランキングされた。

（2）　周辺情報について知りたい場合を想定（図 6）

金閣寺の周辺の情報が欲しいとする。そこで、地理的に近い立命館大学を AND クエリとし、清水寺を NOT クエリに指定することで観光情報を持つクラスタの重要度を下げるようにする。

その結果、銀行、周辺のスーパー、市バスの時刻表などを表すクラスタが上位にランキングされるようになった。

- **同志社大学について検索**

（1）　大学の情報について知りたい場合を想定（図 7）

同志社大学の大学としての情報が知りたい場合を想定する。そこで同じ大学である京都大学を AND クエリ、清水寺を NOT クエリとしてランキングした。

その結果、「大学図書館」「産学連携」「大学入試」といった内容のクラスタが上位にランキングされるようになった。

（2）　観光的な情報について知りたい場合を想定（図 8）

同志社大学の観光的な情報を知りたい場合を想定する。そこで、観光地である清水寺を AND クエリとし、大学関係のクラスタを除去するため京都大学を NOT クエリとして検索を実行した。

すると図 7 で抽出されていた大学関連のクラスタのランクは下がり、代わりに「祭」「劇場」「クラーク記念館」といった、同志社大学の学生が参加するイベントや大学の観光案内などのクラスタが上位に来るようになった。

- **その他の場合**

その他、特徴的な結果が現れたものを簡単に紹介する。

「金閣寺 AND 銀閣寺 NOT 清水寺」とすると、金閣寺、銀閣寺の総本山である相国寺関連のクラスタ、金閣寺の建てられた歴史を表すクラスタが上位に来た。（足利義満、炎上など）

「金閣寺 AND 同志社大学 NOT 銀閣寺」とすると、金閣寺についての新聞記事や掲示板のクラスタが上位にランクされた。（記事、掲示板など）

同志社大学と地理的に近い京都御所を AND クエリとし、京

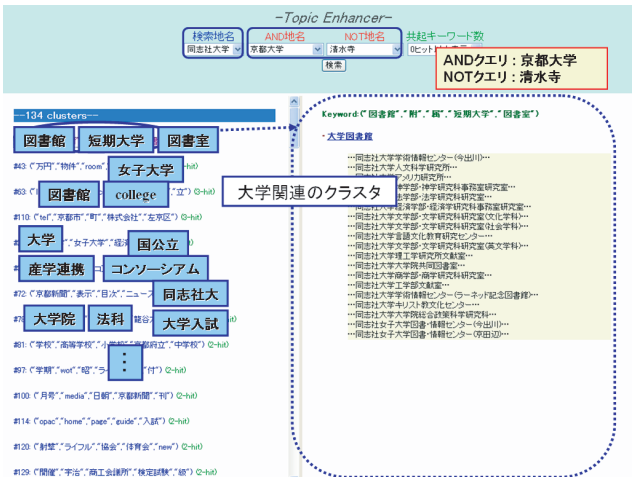

図 7　同志社大学の大学自体の情報を検索

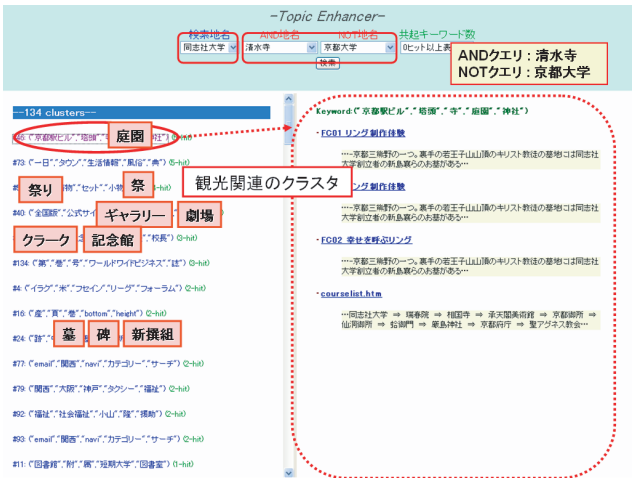

図 8　同志社大学の観光的な情報を検索

都大学を NOT クエリとすると、幕末についての歴史のクラスタが上位に来た。（新撰組、薩摩藩など）

このように、さまざまなトピックを持つクラスタが混在していたクラスタ集合をトピックエンハンス法によってランキングすることにより、検索結果を特定の概念に基づいてソートすることが可能となっていることがわかる。

AND クエリのみを指定した場合、「ニュース一覧」など、どの地理オブジェクトにも入るような雑多なクラスタも抽出してしまうが、NOT クエリを同時に指定することによってそのようなクラスタを除去し、固有度の高い有用なクラスタが得られるようになった。これらの結果は、適当なランキング制御オブジェクトを指定することで、ユーザのさまざまな目的に即したクラスタの重要度決定が可能であることを示すものと言える。

6.　まとめ、及び今後の課題

本稿では、検索対象となる地理オブジェクトに対して、補助的に指定した地理オブジェクトとの共通概念を利用した重要度計算法「トピックエンハンス法」を用いた、トピックの自動カテゴリ分類による地域情報検索システムを提案した。

本稿では、共起キーワード集合 K_{AB} を抽出した場合、非共起キーワード集合 $K_{A\overline{B}}$ を抽出した場合それぞれについて、キーワード集合のカテゴリ別割合分布に一定の変化が表れることを示した。またその変化は直感的に理解しやすいものであった。さらに実際に作成したプロトタイプシステムによる実験では、いくつかの地理オブジェクトを指定することにより、要求するカテゴリに特化した検索結果を得ることができた。

「金閣寺の周辺情報」や「同志社の観光情報」といった要求に対しては既存のロボット検索エンジンのブール論理検索で有効な検索結果を得ることは難しいが、本システムではそういったカテゴリ的な絞り込みを手作業によるカテゴリの作成やカテゴリ分類などを一切行うことなく達成している。さらにカテゴリ自体も地理オブジェクト間の共通概念という形で自動生成するため、対象となる地理オブジェクトに最も適したカテゴリ分類が可能となるだけでなく、未知の意味的関連を持つ２つの地理オブジェクトに対してその関連の要因を抽出されるクラスタ及び特徴キーワードから特定するといった利用法も考えられる。

今後の検討課題としては、周辺情報の抽出のための地理的に近接する地理オブジェクトの選択を自動化し、地図検索インタフェースに組み込むといった利用法や、ターゲットオブジェクトと意味的関連の強い地理オブジェクト等をランキング制御オブジェクトとしてユーザに推薦することで、その地域にあまり詳しくないユーザの検索時の負担を軽減するなどの手法が考えられる。また、今回は地域情報検索について本手法の有効性を確かめたが、地域情報検索以外への応用についても考えていく。

謝　　辞

本研究を進めるにあたり、クローリングシステムをはじめとして様々な形でお世話になりました京大情報学研究科田中研究室の手塚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] CLUSTERING NAMAZU DOCUMENTS, http://icrouton.as.wakwak.ne.jp/pub/kks/cnamazu.html

[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] 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)

[4] Marco Dorigo，Gianni Di Caro，Luca M.，”Ant Algorithms for Discrete Optimization Cooperation of Distributed Agents for Optimization,” (1999)

[5] 大槻洋輔，佐藤理史，”地域情報ウェブディレクトリの自動編集，” 情報処理学会論文誌，vol.42，no.9，pp.2310–2318，(Sep. 2001)

[6] 椎名宏徳，李龍，上林弥彦，”地名の関連グラフを利用した地理情報検索，” DEWS2004 4-B-04, (2004).

[7] T. Tezuka, R. Lee, H. Takakura and Y. Kambayashi, ”Integrated Model for a Region-Specific Search Systems and Its Implementation,” in Proceedings of 2003 IRC International Conference on Internet Information Retrieval, pp. 243-248, Koyang, Korea, (Oct. 2003)

[8] Vivisimo, http://vivisimo.com/

[9] Zenrin, http://www.zenrin.co.jp/