4E1-5

地域情報発見のためのキーワードへのカテゴリ付与

Keyword Categorization for Discovering Area-Specific Topics

廣嶋 伸章*1 Nobuaki Hiroshima　西岡 秀一*1 Shuichi Nishioka　鷲崎 誠司*1 Seiji Susaki

*1日本電信電話株式会社 NTT サービスエボリューション研究所
NTT Service Evolution Labratories, NTT Corporation

We have been developing a search system that navigate users to area-specific information without having them to type any words. As our system displays area specific keywords on a map, users can find useful information even when they are stranger to the area. However, the users sometimes have difficulty in selecting keywords due to the lack of knowledge. Showing cagtegories related to the keywords is an effective way to solve the problem. We propose a method for assigning categories to the keywords using concept vectors and hyponymy relations.

1. はじめに

近年，スマートフォンやタブレットなどの携帯端末の普及に伴い，外出先などでその地域の情報を調べる機会が増加してきている．地域情報を調べる方法の一つとして，興味のあるキーワードや地名を入力として検索エンジンにより検索を行うことが考えられる．しかし，例えば地名として「東京」を入力した場合に，その近くの「京橋」についての情報は検索できないという問題があった．この問題への解決策として，我々はキーワードと地域の範囲を入力として検索を行う地理情報検索の研究に取り組んできた [安田 08]．しかし，実際に地理情報検索を外出先などで利用してみると，その地域について不慣れなため，どのようなキーワードを入力すればよいかがわからない場合が多いことがわかった．そこで我々は，地図を利用した地域情報のナビゲーションを目的として，地域でよく話題となっているキーワードを提示し，そのキーワードと地域をクエリとして検索を行うことで地域情報の発見を支援する手法を提案した [廣嶋 12]．これにより，PC と比較して文字の入力に手間がかかる携帯端末において，キーワードを入力することなく地域情報を取得することが可能となった．しかし，提示されたキーワードの中には情報を取得したいユーザにとって未知のものが含まれるため，それがユーザにとって必要な情報かどうかをその都度検索して確認する必要があるという新たな問題が発生した．この問題を解決するための方法としてはいくつか考えられるが，表示するキーワードとともに「グルメ」などのキーワードの種類を表すアイコンを表示することで，表示領域に制限のある携帯端末においてそれほど表示領域を必要とせずにユーザの興味のある情報かどうかを判別するための情報を提示できると考えられる．そこで本研究では，このアイコンを表示するために必要な，キーワードに対するカテゴリの情報を付与することを目的とする．

以下では，先行研究とその問題点，問題を解決するための提案手法とその有効性を検証するための評価について述べる．

2. 先行研究

キーワードに対してカテゴリを付与する方法としては，上位下位関係を用いた方法と，文書分類を用いた方法の 2 つが考えられる．

連絡先: hiroshima.nobuaki@lab.ntt.co.jp

2.1 上位下位関係を用いた方法

WordNet などのシソーラスでは，各用語に対して上位語や下位語がどのような語であるかを表す上位下位関係の情報が付与されている．また，Web ページの構造や Wikipedia のページの構造をもとに語の上位下位関係を抽出する方法も提案されている [Shinzato 04, 隅田 09]．これらの上位下位関係をもとに，キーワードを下位語，カテゴリを上位語と考えることにより，キーワードにカテゴリを付与することが可能である．人手により付与された上位下位関係には誤りがほとんど存在せず，自動抽出された上位下位関係も比較的精度が高いため，上位下位関係を用いた場合には精度よくカテゴリの情報を付与できると考えられる．しかしながら，キーワードやカテゴリは上位下位関係として定義されたものに限られるため，任意のキーワードに対してカテゴリを付与することができないという問題や，利用するアプリケーションごとに任意のカテゴリを設定することができないという問題がある．

2.2 文書分類を用いた方法

キーワードに対してカテゴリ付与する問題は，キーワードを複数存在するカテゴリに分類する問題と言い換えることができる．分類のタスクとしては様々なものが存在するが，その中でも文書分類が最も盛んに研究されていると考えられる．キーワードから検索などを行って文書を取得することができれば，キーワードの分類を行うために文書分類の各手法が適用できる．文書分類の方法としては，クラスタリングに代表される教師なし分類と，機械学習などを用いた教師あり分類に大別される．教師なし分類では，分類のための学習データを必要としない反面，人間の直感にあったクラスタが生成できないという問題がある．そのため，アプリケーションで利用する場合には，人手によるクラスタの再構成などが必要になると考えられる．一方，教師あり分類では，利用するアプリケーションに応じて任意のカテゴリを設定できるが，各カテゴリに関する学習データを必要とするという問題がある．そのため，分類を行うための準備にコストがかかることが多い．教師あり分類の中でも，比較的コストをかけずに分類が行える方法として，概念ベースを用いた方法が提案されている．概念ベースは，単語に対してその単語の分野を表す概念ベクトルが付与されたデータベースである．概念ベースを用いて各カテゴリの概念ベクトルを生成し，文書に対しても概念ベクトルを生成して，概念ベクトル間の類似度に基づきカテゴリを選択する．このとき，各カテゴリ

単語	1	2	3	···	D
学校	-0.052	0.045	0.040	···	0.074
総理	-0.011	0.010	0.017	···	0.013
首相	-0.015	0.012	0.022	···	0.015
···	···	···	···	···	···

図 1: 概念ベースの例

の概念ベクトルを生成するための学習データが必要となるが，学習データとして各カテゴリの概要を表す数個〜数十個の単語を用意すれば十分である場合が多く，比較的コストのかからない方法であると考えられる．

このように，文書分類の方法としては様々なものが考えられるが，これらを本研究におけるキーワードへのカテゴリ付与に適用した場合には，精度の面で問題があると考えられる．キーワードを代表する文書として適切でないものが，取得される可能性があり，その場合には精度よく分類を行うことができない．精度面での問題を解決するために，我々は概念ベースを用いた方法に加えて別の学習データを利用した方法を組み合わせることにより精度を向上させる方法を提案している [廣嶋 10]．しかし，組み合わせるために利用した方法の中には人手によるチューニングを必要とするものが少なからず存在し，コストの面で問題があると考えられる．

3. 提案手法

概念ベースなどに基づく文書分類を用いた方法では，取得される文書が適切でない場合に，キーワードに対して正しくカテゴリを付与できない場合があることは先に述べた．しかし，複数のキーワードに対してカテゴリを付与することを考えると，正しくカテゴリを付与できない場合よりも正しくカテゴリを付与できることのほうが多いことが過去の研究により明らかとなっている [廣嶋 10]．このとき，正しくカテゴリが付与されなかったキーワードと関連を持つ別のキーワードに対して正しくカテゴリが付与されていれば，その情報をもとにカテゴリの誤りを訂正できる可能性が考えられる．キーワード同士が関連を持っているかどうかは上位下位関係を利用すればよく，同じ上位語を持つ下位語同士が関連を持つとみなすことができる．そこで，本研究では，上位下位関係を用いた方法と概念ベースを用いた方法を組み合わせた手法を提案する．具体的には，概念ベースを用いて各カテゴリとの類似度を算出し，上位下位関係を用いて類似度の補正を行う．以下では，それぞれの処理について述べる．

3.1 概念ベースを用いた類似度の算出

ここでは，概念ベースについて説明し，概念ベースを用いた類似度の算出方法について述べる．

3.1.1 概念ベース

概念ベースは，様々な単語に対してその単語の概念を表す概念ベクトルが付与されたデータベースである [別所 08]．

概念ベースの生成方法はいくつか存在するが，基本となるのは単語間共起に基づく手法である．単語を行，その単語と共起する語を列とする共起行列を生成し，特異値分解を行って次元を圧縮することにより，行列の各行の行ベクトルを概念ベクトルとして利用できるようになる．

生成された概念ベースの例を図 1 に示す．この例における「総理」と「首相」のように，類似する概念を持つ単語の概念ベクトルは類似するため，ベクトル間の距離が近くなるという性質を持つ．概念ベクトルの各次元は何らかの分野を表しており，概念ベクトルはそれぞれの分野に対して相関を持つかどうかを表していると考えられるため，概念ベクトルはその単語がどのような分野において用いられるかを表現するものであるととらえることができる．

3.1.2 類似度の算出

概念ベースを用いて，キーワードが各カテゴリとどの程度類似しているかを表す類似度を算出する．本研究では，以下の手順により類似度の算出を行う．

1. 各カテゴリの分野を表すカテゴリ分野ベクトルを生成
2. キーワードの分野を表すキーワード分野ベクトルを生成
   (a) キーワードに関連する文書を取得
   (b) 文書中の単語からキーワード分野ベクトルを生成
3. キーワード分野ベクトルと各カテゴリ分野ベクトルとのコサイン距離をカテゴリとの類似度として算出

カテゴリ分野ベクトルは，カテゴリに関連する単語の概念ベクトルの重心を求めることにより算出する．カテゴリに関連する単語は，特にそのカテゴリに関連の深い数個の単語を用意するだけでも比較的精度よく分類が行えることが経験的に判明しており，コストをかけずにカテゴリ分野ベクトルを生成することが可能である．

キーワードに関連する文書の取得には，一般的な検索エンジンを利用できる．取得した文書中の単語の概念ベクトルの重心である文書分野ベクトルを求め，文書分野ベクトルの重心をキーワード分野ベクトルとする．このとき，キーワードとはあまり関係のない文書が取得される場合があり，キーワード分野ベクトルが正しく生成できない可能性がある．

キーワード分野ベクトルと各カテゴリ分野ベクトルとの類似度としては，コサイン距離を用いる．

3.2 上位下位関係を用いた類似度の補正

本研究では，概念ベースを用いて算出した類似度を初期値として，Co-HITS に基づき類似度の補正を行う．ここでは，Co-HITS について述べ，Co-HITS を上位下位関係に適用することにより類似度を補正する方法について述べる．

3.2.1 Co-HITS

Co-HITS は，二部グラフの 2 つの集合にスコアの初期値を与え，エッジを通してスコアを相互に伝搬させることによりスコアの更新を行うリンク解析手法である [Deng 09]．集合 U, V 中の各ノードにスコアの初期値を与え，U の各ノードのスコアからエッジを通して V の各ノードのスコアを求め，V の各ノードのスコアからエッジを通して U の各ノードのスコアを求めるという計算を反復することによりスコアを更新していく．具体的には，以下の式によりスコアの更新を行う．

$$x_i^k = \lambda_U x_i^0 + (1-\lambda_U) \sum_{j \in V} w_{ji}^{VU} y_j^{k-1}$$

$$y_j^k = \lambda_V y_j^0 + (1-\lambda_V) \sum_{i \in U} w_{ij}^{UV} x_i^{k-1}$$

ここで，k は更新回数，x_i^k, y_j^k は k 回の更新が行われた後の集合 U, V の i, j 番目のノードのスコア，x_i^0, y_j^0 は集合 U, V の

i, j 番目のノードのスコアの初期値である．また，λ_U, λ_V は，集合 U, V の各ノードのスコアの初期値をどの程度重視するかを表す 0 から 1 までの値であり，1 に近いほどスコアの初期値をより重視することを表す．また，w_{ji}^{VU}, w_{ij}^{UV} は重み係数であり，ノード間のエッジの有無から次のように求める．

$$w_{ji}^{VU} = \frac{a_{ij}^{UV}}{\sum_{i \in U} a_{ij}^{UV}}, w_{ij}^{UV} = \frac{a_{ji}^{VU}}{\sum_{j \in V} a_{ji}^{VU}}$$

ここで，a_{ij}^{UV} は集合 U の i 番目のノードから集合 V の j 番目のノードへのエッジが存在する場合に 1 となり，存在しない場合に 0 となる値である．

3.2.2 類似度の補正

Co-HITS は，以下のような性質を持つ集合 U, V が存在する場合にスコアの補正を行うことができる．

- 集合 U のあるノード u のスコアが高ければ u とエッジでつながっている集合 V のノードのスコアが高くなる
- 集合 V のあるノード v のスコアが高ければ v とエッジでつながっている集合 U のノードのスコアが高くなる

佐藤らは，ユーザの専門性を推定するために Co-HITS を利用している [佐藤 12]．ユーザの集合を U，文書の集合を V とすると，専門性スコアの高い (専門的な知識を持つ) ユーザは専門性スコアの高い文書を閲覧し，専門性スコアの高い文書は専門性スコアの高いユーザによって閲覧されやすいため，集合 U, V は上記の性質を満たし，Co-HITS が適用可能である．佐藤らは実際には Co-HITS に修正を加えた手法を適用している．重み係数の算出にはクリックスルーログを利用している．

ここで，本研究の目的であるキーワードのカテゴリ付与について考える．キーワードの集合を U，上位語の集合を V とし，あるカテゴリとの類似度をスコアとすると，あるキーワードのカテゴリとの類似度が高ければその上位語もカテゴリとの類似度が高くなり，ある上位語のカテゴリとの類似度が高ければその上位語に属するキーワードもカテゴリとの類似度が高くなると考えられる．例えば，キーワード「カレーライス」についてカテゴリ「グルメ」との類似度が高ければその上位語である「洋食」についても「グルメ」との類似度が高くなり，その逆も成り立つ．よって，集合 U, V は上記の性質を満たし，Co-HITS が適用可能な問題であることがわかる．重み係数の算出には，上位下位関係がそのまま利用できる．そこで本研究では，カテゴリごとに Co-HITS を用いて類似度の補正を行う．キーワードのカテゴリとの類似度の初期値には，概念ベースを用いて求めた類似度の値を用いる．上位語のカテゴリとの類似度の初期値にはすべての上位語に対して一定の値を用いる．

3.3 処理の流れ

以上を踏まえ，本研究では以下の手順によりキーワードに対してカテゴリを付与する．

- カテゴリごとに以下を実行
  - 各キーワードのカテゴリとの類似度の初期値を概念ベースを用いて算出
  - 各キーワードのカテゴリとの類似度を正規化
  - Co-HITS によりカテゴリとの類似度を補正
  - 各キーワードのカテゴリとの類似度を非正規化
- キーワードごとに，カテゴリとの類似度を比較し，類似度が高いカテゴリを付与

表 1: 付与するカテゴリの種類

グルメ	買い物	自然	文化
歴史	レジャー	エンタメ	スポーツ
交通	健康	組織	宿泊

表 2: 従来手法との比較評価結果

手法	適合率 (%)
従来手法	76.6
提案手法	85.4

正規化は Co-HITS のスコアの初期値の合計を 1 とする必要があるために行うものであり，非正規化は正規化したスコアをカテゴリごとに比較可能とするために行うものである．

4. 評価

提案手法の有効性を検証するために，評価を行った．以下では，従来手法との比較評価およびサービス適用時の評価について述べる．

4.1 従来手法との比較評価

はじめに，カテゴリ付与の精度について従来手法との比較を行った．

4.1.1 評価手順

Wikipedia の見出し語約 128 万件に対し，従来手法と提案手法それぞれによりカテゴリを付与した．従来手法としては，概念ベースによる手法のみでカテゴリを付与するものを用意した．キーワードに関連する文書は，1 年半分のブログ記事から文書検索を行うことにより取得した．付与するカテゴリは表 1 に示す 12 種類とした．

文書が取得できて何らかのカテゴリが付与された 12 万語のうち，どのカテゴリにも属さず分類のノイズとなりうるスコアの低い 3 万語のキーワードを除去した．各カテゴリに対して 100 語ずつのキーワードを選択し，キーワードに付与されたカテゴリが正しいかどうかの適合率を求めた．

4.1.2 評価結果

評価結果を表 2 に示す．提案手法は従来手法より 8.8 ポイント高い適合率となった．この差は，除去するキーワードの量を変化させてもほぼ同じであった．これにより，提案手法の有効性が検証できた．

次に，カテゴリごとの適合率を評価した．その結果，12 カテゴリのうち，従来手法のほうが高い適合率となったものは「文化」の 1 種類，同じものは「健康」の 1 種類，残りの 10 種類については提案手法のほうが高い適合率となった．適合率が下がった「文化」について誤り事例を確認したところ，正解が「レジャー」である事例が多いことがわかった．従来手法において「レジャー」に対する適合率は 28.0%と最も低いが，これはこのカテゴリにおける類似度の初期値として多くのキーワードに適切な値が付与されていないことを意味している．このような状況でスコアを伝搬させると，他のカテゴリとの類似度が高くなり，本来は「レジャー」であるべきキーワードが他のカテゴリへと流出するためにこのような結果となったことが考えられる．この問題を解決するためには，類似度の初期値が

表 3: 発見機会の増加につながるかの評価結果

手法	不明の割合 (%)
カテゴリなし	34.7
カテゴリあり	29.9

表 4: 手戻りをなくせるかの評価結果

手法	キーワード選択率 (%)	適合率 (%)
従来手法	35.1	69.6
提案手法	34.2	80.0

適切に与えられるように，適切なカテゴリ分野ベクトルを生成する必要があると考えられる．今回の評価では，カテゴリに関する数個のキーワードを用意してその概念ベクトルの重心をカテゴリ分野ベクトルとしたが，関連する文書から生成するなどの方法についても検討する必要があると考えられる．

4.2 サービスに適用した場合の評価

我々は，[廣嶋 12] の手法の有効性を検証するために，地図を用いた情報の発見を支援することを目的とした Android アプリケーション「発見探地図エリアダス」を Google Play 上で公開し，2013 年 3 月現在，実験サービスを提供している*1．このアプリケーションでは，表示された地域でよく話題となっているキーワードを提示し，そのキーワードと地域をクエリとして検索を行うことで地域情報を発見することが可能である．このアプリケーションにおいて，提案手法によりキーワードに対してカテゴリの付与を行い，キーワードをカテゴリつきで提示したり，特定のカテゴリに属するキーワードを提示することにより，より効率よく情報の発見が行えることが考えられる．そこで，カテゴリ分類・絞り込み機能を本アプリケーションに組み込んでサービスに適用した場合の評価を行った．

4.2.1 発見機会の増加につながるかの評価

提案手法によりカテゴリを付与することで情報発見の機会の増加につながるかの評価を行った．300 名の被験者に対し，3 つの地域 (横浜，浅草，鎌倉) を用いた．それぞれの地域に関して興味のあるカテゴリを 7 種類 (グルメ，買い物，自然，文化，歴史，レジャー，エンタメ) の中から 3 種類選択してもらい，各地域のキーワードをカテゴリつき／なしの場合でそれぞれ 10 個ずつ提示して，各キーワードについて検索をしてみたいと思うかどうかを理由とともに回答してもらった．回答結果から,「キーワードが何か不明のため検索をしてみたくない」と回答した割合を算出した．その結果を表 3 に示す．

表 3 より，カテゴリを付与することでキーワードに対する理解を助けることが確認できた．これにより，情報発見の機会が増加することが考えられる．

4.2.2 手戻りをなくせるかどうかの評価

提案手法により精度よくカテゴリを付与することで従来手法と比較して手戻りをなくせるかの評価を行った．上の評価と同様に，興味のあるカテゴリを 7 種類の中から 3 種類選択してもらい，各地域の各カテゴリのキーワードを従来手法と提案手法の場合でそれぞれ 10 個ずつ提示して，各キーワードについて検索をしてみたいと思うかどうかを回答してもらった．各キーワードに対して，その地域で正しいカテゴリであったかどうかの正解を付与し，回答結果と正解データからキーワードが選択された割合と選択されたキーワードの適合率を算出した．その結果を表 4 に示す．

表 4 より，精度よくカテゴリを付与することで無駄なキーワードの選択が抑止できることが確認できた．また，適合率も高くなるため，カテゴリをもとにキーワードを選択したが必要な情報でなかった場合の手戻りを抑止できると考えられる．

*1 http://areadas.jp/

5. おわりに

概念ベースを用いて各キーワードのカテゴリへの類似度の初期値を算出し，上位下位関係を用いて Co-HITS により類似度の補正を行うことで，様々なキーワードに対しカテゴリを付与する方法を提案した．提案手法の有効性を検証するための評価を行い，従来手法と比較して精度よくカテゴリの付与が行えることを確認した．また，提案手法をサービスに適用した場合の評価を行い，精度よくカテゴリを付与することで効率のよい情報の発見につながることを確認した．

今後は，カテゴリ付与の精度が対象カテゴリによってばらつきがある問題を解消するため，カテゴリ付与に適したカテゴリ分野ベクトルを生成する方法について検討していきたい．

参考文献

[安田 08] 安田 宜仁, 戸田 浩之: 検索位置のごく周辺を対象とした地理情報検索, 人工知能学会論文誌, Vol.23, No.5-C, pp.364-373 (2008).

[廣嶋 12] 廣嶋 伸章, 安田宜仁, 藤田 尚樹, 片岡 良治: 地理情報検索におけるクエリ入力支援のための特徴語の提示, 人工知能学会全国大会, 1C1-R-5-6, (2012).

[Shinzato 04] Shinzato, K. and Torisawa, K.: Acquiring Hyponymy Relations from Web Documents, Proc. of COLING 2004, pp. 73-80, (2004).

[隅田 09] 隅田 飛鳥, 吉永 直樹, 鳥澤 健太郎: Wikipedia の記事構造からの上位下位関係抽出, 自然言語処理, Vol. 16, No. 3, pp. 3-24 (2009).

[廣嶋 10] 廣嶋 伸章, 戸田 浩之, 松浦 由美子, 片岡 良治: 概念ベースに基づく Web 検索のクエリタイプ判定手法とその評価, 情報処理学会論文誌, データベース 3(3), pp.33-45, (2010).

[別所 08] 別所 克人, 内山 俊郎, 内山 匡, 片岡 良治, 奥 雅博: 単語・意味属性間共起に基づくコーパス概念ベースの生成方式, 情報処理学会論文誌, Vol. 49, No. 12, pp. 3997-4006 (2008).

[Deng 09] Deng, H., Lyu, M. R. and King, I.: A Generalized Co-HITS Algorithm and Its Application to Bipartite Graphs, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining Pages 239-248 (2009).

[佐藤 12] 佐藤 大祐, 安田 宜仁, 小池 義昌, 片岡 良治: 検索システムユーザの特定分野に関する専門性推定のためのクリックスルーログの利用, 情報処理学会論文誌, Vol.4, No.3, pp.12-21 (2012).