1C1-R-5-6

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

Related Term Extraction for Keyword Input Support in Geographic Information Retrieval

廣嶋 伸章[*1] 安田 宜仁[*1] 藤田 尚樹[*1] 片岡 良治[*1]
Nobuaki Hiroshima Norihito Yasuda Naoki Fujita Ryoji Kataoka

[*1]日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

We have been developing an geographic information retrieval(GIR) system that accepts geographic queries in addition to ordinary keywords. Such a system is useful, for example, when a user want to make a decision about where to go on a trip. However, the user sometimes have difficulty inputting both keywords and geographic queries. Showing representative terms related to the query is an effective way to solve the problem. We propose a method that selects related terms based on the poisson probability that expresses the probability of an actual number of occurrences if a certain number of occurences is expected.

1. はじめに

外出先でその地域の情報を調べたり，旅行の計画を立案したりする際には，その場所に関して様々なキーワードで検索を行い，その場所に関する情報を知ることができれば有益である．そこで，我々はこれまで，様々なキーワード・場所に関する検索を行う地理情報検索の研究について取り組んできた [安田 2008, 戸田 2009]．しかし，地理情報検索を実際に利用するためには，キーワードと場所の両方を指定しなければならなかった．そのため，場所が決まっていてもそこでどのような事物が話題であるかわからず，キーワードを指定できない場合が多くあった．また，キーワードを指定できる場合でも，スマートフォンなどの携帯端末から入力を行うのは手間がかかるという問題があった．この問題を解決するためには，クエリの入力を支援する仕組みとして指定した場所に特徴的な語の提示を行い，場所の指定だけでキーワードを想起させることが必要であると考えた．これにより，その場所に関してよく知らない場合でも，場所を指定して提示されるキーワードを選択することにより検索を行うことができ，その場所に関する情報を知ることができるようになる．そこで本研究では，場所に関する特徴的な語を獲得することを目的とする．

以下では，提案手法および評価について述べる．また，提案手法を用いて，様々な場所に関するキーワードを提示し，検索を行うことを可能としたスマートフォン向けのアプリケーションについて述べる．

2. 提案手法

特徴的な語は，地域全体での出現の割合に比べて対象とする地域での出現の割合が高い語であると考えられる．語がすべての地域で一様に分布すると仮定すれば，地域全体での出現の割合をもとに，対象とする地域での期待頻度が算出できる．その期待頻度と比べて，実際に出現した頻度で出現するという事象の起こりやすさを調べ，この事象が起こりにくい語を特徴的な語として獲得するということが考えられる．そこで，本研究では，平均で λ 回発生する事象が k 回発生する確率を表すポアソン確率を用いて語の出現が起こりやすいかを算出する方法を提案する．

連絡先: hiroshima.nobuaki@lab.ntt.co.jp

地域全体でのすべての語の出現頻度を n，地域全体での着目する語の出現頻度を s，対象とする地域でのすべての語の出現頻度を k，対象とする地域での着目する語の出現頻度を r とすると，地域全体でのその語の出現の割合は s/n と表せ，そこから対象とする地域でのその語の出現頻度の期待値は sk/n と表すことができる．この値に比べ，実際の頻度が r 未満であるような確率はポアソン確率により以下の式で表すことができる．

$$P(x < r) = \sum_{x=0}^{r-1} \frac{e^{-\frac{sk}{n}} (\frac{sk}{n})^x}{x!} \quad (1)$$

この値を地域内で出現したすべての語に対して算出し，値の大きい，すなわち頻度 r 以上で出現することはめったに起こらないと考えられる語を特徴語として獲得する．

出現頻度としては，文書内頻度 (TF) と文書頻度 (DF) が考えられるが，本研究では，文書内でどれだけ多く言及されているかよりも多くの文書でどれだけ多く言及されているかのほうが重要であると考え，DF を用いてポアソン確率を算出することとする．

3. 評価

提案手法の有効性を検証するため，従来手法と提案手法との比較評価を行った．

3.1 実験条件

ブログ記事には様々な地域に関する記述が多いと考えられるため，文書データとしては，2010 年 8 月から 2012 年 1 月までの 1 年半の間に投稿されたブログ記事を利用した．各ブログ記事から特徴語候補の抽出および地点の特定を行った．抽出する特徴語候補は Wikipedia の見出し語となっている語とした．地点の特定のための地名表現の抽出および緯度経度への変換の方法としては平野らの手法 [平野 2008] を用いた．

東京都新宿区，岡山県岡山市，富山県富山市，山形県山形市，三重県津市の 5 つの地域を評価対象の地域として，ベースラインと提案手法により特徴語の獲得を行った．ベースラインとして，以下の 2 つの手法を用意した．

- ベースライン 1(TF-IDF)
  対象とする地域に関して書かれている文書集合をまとめ

表 1: 従来手法との比較評価結果

手法	MAP
ベースライン 1(TF-IDF)	0.831
ベースライン 2(カイ 2 乗値)	0.951
提案手法	**0.963**

表 2: 提案手法によって得られた特徴語の例

地域	特徴語
新宿	東京ラーメン, とげぬき地蔵, 新宿御苑
岡山	岡山弁, 苫田温泉, 岡山後楽園
富山	富岩運河, 呉羽丘陵, 富山ブラック
山形	済生館, 霞城公園, 文翔館
津	榊原温泉, 専修寺, 津城

て 1 つの記事と考えたときの TF と，それ以外の地域に関して書かれている文書での DF をもとに TF-IDF の値を算出し，値の大きいものを特徴語として抽出する．

- ベースライン 2(カイ 2 乗値)
  特徴語候補が出現するという事象と，対象とする地域で出現するという事象に相関があるかどうかを表すカイ 2 乗値を算出し，値の大きいものを特徴語として抽出する．

3.2 評価指標

獲得された特徴語だけを見て，その地域に関して特徴的であるかどうかを直接評価することは難しいと考えられる．そこで，獲得された語を直接評価するのではなく，その語に関するその場所での記事を検索し，その記事の評価をもとに語を評価した．本研究では，検索された文書に対して，その文書を読んでその場所に行ってみたいと思うかどうかを有用性として 5 段階で人手により付与することとした．地域ごと，手法ごとに 10 個の特徴語を獲得し，特徴語ごとに最大 100 件の記事を地理情報検索 [安田 2008] を用いて取得した．取得された記事に対して有用性を付与し，最大 100 件の記事の中で有用性が 4 以上となった記事が存在する場合に，その特徴語は適合していると考え，獲得された特徴語の順位を考慮して平均適合率を算出した．この平均適合率の各地域ごとの平均である MAP の値を評価指標とした．

3.3 実験結果

実験結果を表 1 に示す．提案手法は高い MAP 値を示しており，提案手法により特徴語が適切に獲得できる可能性が示唆された．

提案手法によって得られた特徴語の例を表 2 に示す．特徴語として適切と思われる語が獲得されていることがわかる．

提案手法により獲得された特徴語のうち適合しないと判定されたものについて分析を行った．適合しないと判定された語の中には，地元で開催されるイベントの会場や，地元で応援されているサッカークラブのようなその地域に居住している人々によってよく話題にされていると思われる語が多く含まれていた．これらの語は，その地域でのみ言及されている割合は非常に高いが，頻度はあまり高くないという共通する特徴がみられた．今回の評価では，その地域に行ってみたいと思うかという点に着目したが，そのような場合には地元の人たちの間でよく話題にされている語は獲得しないように手法を改良することにより，より適切な特徴語が獲得できると考えられる．

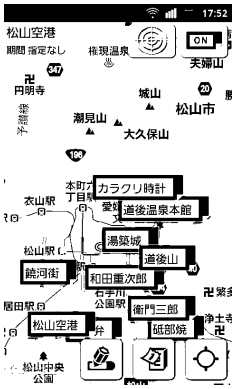

図 1: 特徴語提示画面の例

4. アプリケーション

全国の様々な地域において適切な特徴語を提示できるかを検証することは重要である．そこで，提案手法が旅行計画の立案などの際の地域情報の取得に有効かどうかを確認することを目指して，様々な場所に関する特徴語を提示し，その特徴語を用いて検索を行うことを可能としたスマートフォン向けのアプリケーション「発見探地図エリアダス」を作成した．本アプリにおける特徴語提示画面の例を図 1 に示す．

本アプリは 2012 年 4 月現在，Android アプリとして Google Play 上に公開しており，無料で利用することが可能である[*1]．

5. まとめ

本研究では，場所に関する特徴的な語を獲得する手法を提案した．提案手法の有効性を検証するための評価を行い，提案手法により特徴語が適切に獲得できる可能性が示唆された．

今後は，地元でよく話題になっている語の扱いについて改善したいと考えている．また，地域情報を取得する際には，地元以外の人がその地域に旅行に行くという状況だけではなく，地元の人が周囲の役立つ情報を探すという状況も考えられ，地域情報取得の状況としては後者のほうが頻度が高いと考えられる．ユーザの居住地に応じて獲得する語を変更するような手法についても検討していきたい．また，作成したアプリの利用者からの要望などをもとに提案手法やサービスの改善を図っていきたいと考えている．

参考文献

[安田 2008] 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).

[戸田 2009] 戸田浩之, 安田宜仁, 奥村学, 松浦由美子, 片岡良治: 地理情報検索のためのスニペット生成法, 人工知能学会論文誌, Vol.24, No.6, pp.494-506 (2008).

[平野 2008] 平野徹, 松尾義博, 菊井玄一郎: 地理的距離と有名度を用いた地名の曖昧性解消, 情報処理学会全国大会, pp. 3D-7 (2008).

*1 http://areadas.jp/