2D-06

Webコンテンツ解析に基づくPOI潜在情報検索エンジンの提案

廖　宸一†　櫻田　健†　河口　信夫‡
†　名古屋大学大学院工学研究科　　‡名古屋大学未来社会創造機構

1　はじめに

近年、インターネットの発展及びスマートフォンの普及に従い、多様なオンライン位置情報サービスが普及しつつある。 しかし、インターネット上はPOI (Point of Interest)に関する情報が多く存在する一方で、それらの情報は十分に活用されているとは言えない.

本研究ではPOIの名称や住所や電話番号や緯度経度などのような属性をPOIの固有(inherent)情報と呼び、それに対し、POIの販売する商品、提供するサービスや開催するイベントをPOIの潜在（latent）情報と呼ぶ。 既存の位置情報サービスは固有情報を用いてPOIを検索できるが、潜在情報では検索できない。 例えば、「カフェ」と「コーヒー」をキーワードとして検索すると、出力されたPOIの結果は異なる場合が多い。 あるPOIがコーヒーを販売しているが、名称に「コーヒー」という単語がない場合、マッチングされない場合がある。そのため、POI検索の網羅性が十分ではないという問題点がある。

本研究では、Webコンテンツ解析の手法を用い、POIに対応するホームページからそのPOIの扱う商品、サービスやイベントなどの潜在情報でも検索できる検索エンジンを提案する。 潜在情報を用いてPOIをベクトル空間にマッピングすることにより、POI間の類似度を計算可能にした。

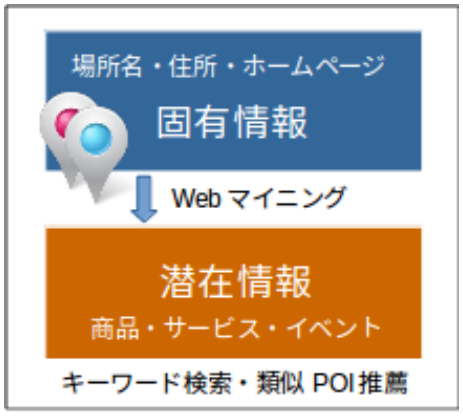

図 1: Webマイニングを用いてPOIに関する潜在情報を抽出する

2　Webからの潜在情報の抽出

Webデータ抽出技術[1]とは半構造化データのHTMLドキュメントを整理して有用な情報を構造化データとして抽出することを指す。 一般に、HTMLドキュメントは、テンプレートを用い、データベーステーブル上の構造化データ（レコード）から生成される。 そのため、類似したHTML構造のマッチングにより、構造化データが抽出可能である[2]。

本研究では、下記の手順に従ってWebページからデータを抽出する[3]。

- 枝刈り：HTML要素はIn-line要素と Block要素の二種類がある。 In-line要素は画面構造（幅、長さなど）に影響しないが、HTML構造をマッチングする場合、ノイズになる可能があるため、幅優先探索でIn-line要素を取り除く。
- HTML構造マッチング：幅優先探索で同じレベルにある兄弟要素とそのすべての子要素をマッチングした場合、レコードとして出力する。 ただし、複数の兄弟要素が一つのレコードを構成する場合、複数個ずつ比較する必要がある。
- バックトラッキング：上記のマッチング手順が完了してから、残った画面領域を抽出すると、兄弟要素のない要素、あるいは独立なレコード要素も抽出することが可能となる。

上記の手順を用いて、POIの持つウェブサイトを解析すると、POIの潜在情報が抽出されたレコードに含まれることになる。

3　類似POI検索

Word2Vec[4]は単語をn次元($n < 1000$)の単語ベクトルにマッピングする。 これにより、「コーヒー」と「カフェ」のように意味の近い単語は近くにマッピングされる。 この単語ベクトルを利用して類似POIの検索を可能にする。

まず、単語ベクトルを用いてPOIをベクトル空間にマッピングする必要がある。 各POIからは、複数のレコードからなる潜在情報を抽出できており、各レコードは含まれる単語ベクトルの平均として空間上にマッピングする。 ただし、「コーヒー」や「映画」などのPOIに重要な単語に高い重みを与え、「会社」や「採用」などの単語に低い重みを与える必要である。 そこで、IDFを単語の重みづけに利用する。

$$\theta_i = \log \frac{|D_{url}|}{|\{d : t_i \in d\}|} \tag{1}$$

数式(1)は単語t_iの重みθ_iを求める。 $|D_{url}|$は総URL数、$|\{d : t_i \in d\}|$は単語t_iを含むWebページのURL数である。 次に、(2)式を用い、レコードの

†Chenyi Liao　†Ken Sakurada ‡Nobuo Kawaguchi
†Graduate School of Engineering, Nagoya University
‡Institute of Innovation for Future Society, Nagoya University

ベクトルを計算するため、レコードにある各単語のベクトルの重み付き平均を取る。

$$X = \frac{\sum_{i=1}^{k} \theta_i \omega_i}{1 + \sum_{i=1}^{k} \theta_i} \tag{2}$$

(2)式により、レコードのベクトルXが生成される。 kはレコードにある単語ベクトルの数を表す。 ω_iはレコードにある単語である。 次に、POIにあるレコードベクトルXを行ベクトルとし、POIの潜在情報行列$A_{m\times n}$を作る(式3)。 Aの行がレコードにある各単語のベクトルであり、列は各レコードを表す。 mはPOIにあるレコードの数を表す。その行列の各行をベクトル空間にマッピングすると、POIはn次元空間のベクトル列で表示される。

2つのPOIベクトル列AとBの類似度を求めるため、Wang[5]らの手法を用いる (式(4))。 なお、$A_{m\times n}$は正方行列ではない場合が多いので、特異値分解(SVD)を利用して、主成分分析(PCA)を行う(式3)。

$$A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n} \tag{3}$$

$$\begin{cases} N = min(m_A, m_B, n) \\ S_{AB} = |\sum_{i=1}^{N} \frac{\lambda_i^A + \lambda_i^B}{2} \cdot \frac{<\mu_i^A, \mu_i^B>}{|\mu_i^A| \cdot |\mu_i^B|}| \\ S_{AB} \in [0, 1] \end{cases} \tag{4}$$

式3により、POI行列AはU、Σ、V^Tの積になる。その行列Σの対角成分は式4の$\lambda = (\lambda_1, \lambda_2, ..., \lambda_n)$になる。 行列$V^T$の各列ベクトルは基礎ベクトル$\mu$になる。 それぞれ式（4）に代入して類似度$S_{AB}$を計算する。 NはPOI Aのレコードベクトル数m_A、POI Bのレコードベクトル数m_Bまたはベクトル次元数nの最小値を取る。 S_{AB}が小さいほど、2つのPOIの潜在情報の類似度が高いと言える。

4 アプリケーション提案

図 2: POI類似度によって場所推薦する例

以上に述べたアルゴリズムにより、2つのアプリケーションを提案する。 まず、従来のキーワード検索したPOIの結果を用い、各POIと類似度・距離の近いPOIを検索可能である。 図2に示すように、従来には「コーヒー」を検索した場合「XXコーヒー」という店が表示されるが、POI類似度計算により、「XXコーヒー」に近い「XX喫茶店」と

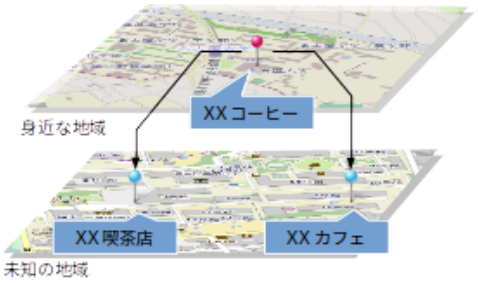

図 3: 身近な地域にあるPOIによって未知の地域にあるPOIを検索する例

「XXカフェ」を提示する。 これにより、検索結果のカバレッジ向上が期待できる。

次に、地域についての情報が不足する場合、従来のキーワード検索が使えないこともある。 ユーザに身近な地域にあるPOIにより、未知の地域にある似ているPOIが推薦できる。 図3に示すように、ユーザの身近い「XXコーヒー」により、行ったことのない場所にある似ている「XX喫茶店」と「XXカフェ」を推薦する。

5 まとめ

本研究では、Webコンテンツ解析の手法を用い、POIに対応するホームページからそのPOIの扱う商品、サービスやイベントなどの潜在情報でも検索できる検索エンジンを提案した。 潜在情報を用いて類似したPOIの検索機能により、POI検索結果のカバレッジ向上が期待できる。 ユーザに身近な地域にあるPOIにより、未知の地域にある似ているPOIも推薦可能にした。

References

[1] Emilio F. Pasquale D.M. Fiumara.G Baumgartner R. Web data extraction, applications and techniques: A survey. *arXiv - Social Media Intelligence*, pages 301–323, 2014.

[2] Zhai Y. Liu B. Web data extraction based on partial tree alignment. *Proceedings of the 14th international conference on World Wide Web*, pages 76–85, 2005.

[3] Liao C. Kaji K. Hiroi K. Kawaguchi N. An event data extraction method based on html structure analysis and machine learning. *COMPSAC 2015*, pages 217–222, 2015.

[4] Tomas T. Kai C. Greg C. Jeffrey D. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint*, pages 1–12, 2013.

[5] Wang W. Sakurada K. Kawaguchi N. Incremental and Enhanced Scanline-Based Segmentation Method for Surface Reconstruction of Sparse LiDAR Data. *Remote Sensing*, 8:1–22, 2016.