E-012

Q&A サイトを対象にした地域別土産物情報収集ツール

A tool for collecting local souvenir information from a question and answer

川野 覚† 溝渕 昭二‡
Satoru Kawano Shoji Mizobuchi

1. はじめに

観光などの目的で旅行した際，土産物を購入することがある．土産物とは，知人や縁者に贈るために出先で入手する，その土地の産物のことである．旅行に関するインターネットのアンケート調査[1]によると，旅行時に土産物を全く購入したことがない人の割合は 1%であることから，旅行者にとって土産物の購入は，旅行時の主要な行為の一つであると考えられる．したがって，旅行時に適切な土産物情報を提示することができれば，旅行者の満足度の向上が期待できる．そして，それを受けて，現在著者らは，土産物情報を提示するアプリケーションの開発を行っている．本アプリケーションは，滞在場所周辺で入手できる土産物情報を提示する機能を設ける予定であるが，その機能を実現するには，土産物名と入手地域の情報が最低限必要となる．

そこで，本稿では，Q&A サイトから土産物名を地域別に収集するツールを提案する．本ツールは，指定された地域の土産物名が登場する見込みの高い Q&A サイトのエントリから，土産物名となりそうなキーワードを，スコア付きで抽出するものである．

2. 関連研究

土産物に関する情報を含む観光情報全般の抽出，あるいは，広く固有名詞について取り扱った研究は多く行われている．

奥ら[2]は，グルメ情報サイトから IDF と WebPMI を組み合わせて算出される地域限定スコアを基に地域特有のスポットを抽出する手法を提案している．遠藤ら[3]は，人手によりコストをかけて生成する辞書は利用せず，形態素 N-gram と残差 IDF (Residual IDF; RIDF)による重み付けを利用して，地域サイトの情報から対象地域の観光キーワードを抽出する手法を提案している．西川ら[4]は，CGM より求めた形態素 N-gram に対して，ルールによるフィルタリングと頻度によるスコアリングを行うことにより，固有名詞を抽出する手法を提案している．

これらの研究では，観光情報を利用，または抽出しているが，土産物情報はその中の単なる要素として取り扱われているだけであり，地域も限定されたものとなっており，十分に検討されている状況にはない．

3. 土産物情報収集ツール

本ツールは，次の(1)から(5)までの手順に従って，Q&A サイトから土産物名となりそうなキーワードを抽出するものである．

(1) Q&A サイトからテキストの取得

土産物に関する質問のベストアンサーを Q&A サイトから，都道府県毎にテキストで取得する．今回利用する Q&A サイトは，Yahoo!知恵袋[1]とする．ベストアンサーの例を図 1 に示す．

図 1 ベストアンサーの例

(2) 形態素列の生成

手順(1)で取得したテキストに対して形態素解析を行う．形態素解析器には Mecab[2]を用いる．Mecab で使用する辞書は標準 API 辞書のみとする．テキストを形態素解析した例を図 2 に示す．

```
三重　　名詞,固有名詞,地域,一般,*,*,三重,ミエ,ミエ
県　　名詞,接尾,地域,*,*,*,県,ケン,ケン
と　　助詞,格助詞,引用,*,*,*,と,ト,ト
いえ　　動詞,自立,*,*,五段・ワ行促音便,仮定形,いう,イエ,イエ
ば　　助詞,接続助詞,*,*,*,*,ば,バ,バ
やはり　副詞,一般,*,*,*,*,やはり,ヤハリ,ヤハリ
「　　記号,括弧開,*,*,*,*,「,「,「
赤　　名詞,一般,*,*,*,*,赤,アカ,アカ
福　　名詞,一般,*,*,*,*,福,フク,フク
」　　記号,括弧閉,*,*,*,*,」,」,」
が　　助詞,格助詞,一般,*,*,*,が,ガ,ガ
オススメ　　名詞,サ変接続,*,*,*,*,オススメ,オススメ,オススメ
です　助動詞,*,*,*,特殊・デス,基本形,です,デス,デス
```

図 2 テキストを形態素解析した例

(3) 形態素 N-gram の生成

手順(2)で生成した形態素列から 1 から 5 までの形態素 N-gram を生成する．生成された形態素 N-gram の例を表 1 に示す．

表 1 生成された形態素 N-gram の例

N-gram	生成される形態素 N-gram の例
1-gram	「三重」，「県」，「と」，「いえ」，「ば」
2-gram	「三重　県」，「県　と」，「と　いえ」
3-gram	「三重　県　と」，「県　と　いえ」
4-gram	「三重　県　と　いえ」,「県　と　いえ　ば」
5-gram	「三重　県　と　いえ　ば」

(4) 土産物名と判断されない形態素 N-gram の除去

手順(3)で生成した形態素 N-gram の中から，明らかに土産物名となり得ない形態素 N-gram を除去する．除去する形態素 N-gram のパターンの一例を以下に示す．

† Graduate School of Science and Engineering, Kinki University
‡ Faculty of Science and Engineering, Kinki University

[1] http://chiebukuro.co.jp/

[2] https://code.google.com/p/mecab/

- 最初の形態素が，「助詞」，「助動詞」，「副詞」，「名詞-接尾」，「動詞-接尾」，「動詞-非自立」，「形容詞-接尾」，「形容動詞-接尾」，「記号-句点」，「記号-読点」，「記号-一般」である形態素 N-gram
- 最後の形態素が，「接頭詞」，「助詞」，「副詞」である形態素 N-gram
- 同一形態素に，四つ以上の連続する形態素が含まれていない形態素 N-gram
- N/2 以上の記号が含まれている形態素 N-gram
- 括弧の整合性に誤りがある形態素 N-gram
- 「名詞-非自立」，「名詞-数」以外の名詞が存在しない形態素 N-gram
- URL が含まれる形態素 N-gram

(5) 形態素 N-gram の重み付け

手順(4)によって残った形態素 N-gram に対して，残差 IDF による重み付けを行う．残差 IDF は，IDF からポアソン分布を用いて推測される IDF を差し引いたものである．任意の形態素 N-gram をX，収集した文書数をZ，Xの文書中の出現回数を$CF(X)$，Xの出現する文書数を$DF(X)$とするとき，Xの残差 IDF は次式で計算される．

$$RIDF(X) = log\frac{Z}{DF(X)} - log\frac{1}{1-e^{-\frac{CF(X)}{Z}}}$$

形態素 N-gram に対して求めた残差 IDF の例を図 3 に示す．

```
安永 三重 2.900771647456427
安永餅　三重 2.866112608457404
永餅 三重 2.704986306393894
おかげ横丁三重 2.6072000324626394
志摩 三重 2.6072000324626394
津市 三重 2.552716138330584
御福 三重 2.3599617624974063
絲印 三重 2.3599617624974063
笹井 三重 2.3599617624974063
笹井屋　三重 2.3599617624974063
```

図 3　形態素 N-gram に対して求めた残差 IDF の例

４．　性能調査

本ツールによる土産物名抽出性能について，三重県の土産物を対象にして調査した．

本調査では，まず，Yahoo!知恵袋から，解決済みの質問に対するベストアンサーを都道府県毎に取得した．取得する際に使用した条件として，Yahoo!知恵袋のカテゴリは「地域、旅行、お出かけ ＞ 国内 ＞ おみやげ」とし，検索クエリは都道府県名とし，取得件数は 100 件とした．取得した 100 件のベストアンサーは，都道府県毎に一つの文書にまとめた．すなわち，ベストアンサーが 100 件記載された，47 文書を作成した．

三重県のベストアンサーが 100 件記載された文書から，土産物名になりそうなキーワードを出力した．この際，生成された形態素 N-gram は，34,242 個となり，そのうち，6751 個が，最終的に候補として出力された．それらを，人出で確認したところ，69 個の N-gram が土産物名と判定された．

最後に，出力された 6,751 個の形態素 N-gram を RIDF の順にソートを行い，土産物名の分布を確認した．その結果を表 2 に示す．表 2 の横軸は，RIDF のスコア順にソートした際の各形態素 N-gram の順位である．縦軸は，その順位までに登場した，土産物名と判定された形態素 N-gram の個数である．

表 2　調査結果

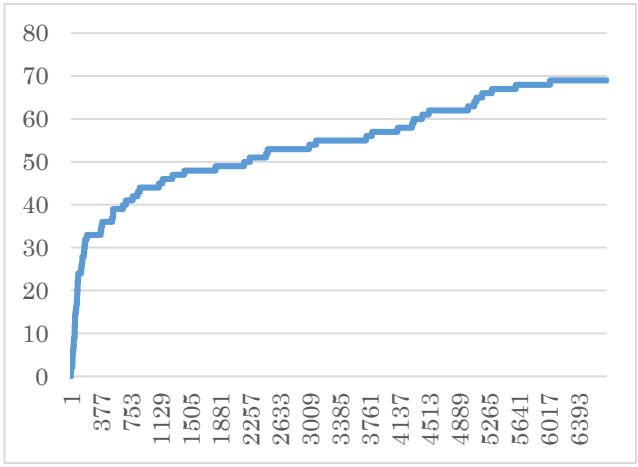

表 2 によると，土産物名と判断された形態素 N-gram の半数が，上位 10%の範囲に出現しているが，残り半数は，それ以降に，散在する結果となった．その理由として，土産物名と判断された形態素 N-gram の出現頻度が，今回対象とした三重県の文書では，低かったことが考えられる．また，不要な形態素 N-gram が多く存在することも，理由としてあげられる．

５．　おわりに

本稿では，都道府県毎に，ベストアンサーが 100 件記載された，47 文書を作成することで，土産物の名称と入手地域の情報を抽出するツールを提案した．

本ツールの性能を調査したところ，一つの文書に出現頻度が多く，土産物だと判定される形態素 N-gram は上位に集中して分布するが，一つの文書で出現頻度が少ない形態素 N-gram では，下位の範囲に渡り分布することがわかった．また，不要な形態素 N-gram の比率が依然として高いことがわかった．

今後の課題としては，土産物名の出現に敏感なスコアリングや，不要な形態素 N-gram の除去を高度化することが挙げられる．

参考文献

[1] 何でも調査団: お土産についてのアンケート・ランキング, http://chosa.nifty.com/travel/chosa_report_A20140221/

[2] 池野 篤司, 濱口 佳孝, 山本 英子, 井佐原 均: Web 文書集合からの専門用語獲得, 情報処理学会論文誌 Vol. 47, No, 6, pp. 1717-1727, (2006)

[3] 西川 侑吾, 伊藤 直之, 田村 直之, 田中 慶之, 中川 修, 新堀 英二: 形態素 N-gram を用いた助詞を含む固有名詞抽出, 第 16 回言語処理学会 A1-3, (2010)

[4] 奥 健太, 西崎 剛司, 服部 文夫: 地域限定性スコアに基づく位置情報付きコンテンツからの地域限定語句の抽出, 情報処理学会論文誌 データベース Vol. 5, No. 3, pp. 97-116, (2012)

[5] 遠藤 雅樹, 大野 成義, 石川 博: 地域サイトからの観光キーワードの自動抽出と関連情報の融合, 第 158 回データベース研究発表会 A1-2 (2013)