3Q-06

残差 IDF と WebPMI を用いた土産物名の抽出

川野覚†　溝渕 昭二‡

近畿大学大学院総合理工学研究科†

近畿大学理工学部情報学科‡

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，旅先などで，知人や縁者に贈る，その土地の産物のことである．旅行に関するアンケート調査[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主要な行為の 1 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，旅行者の満足度向上が期待できる．

土産物情報を含む観光情報全般の抽出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 および地域関連重みと呼ばれる尺度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．上原ら[3]は，Web 上に存在する観光情報から，複数の特徴ベクトルを構築し観光地間の類似度を推定するシステムを提案している．遠藤ら[4]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．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，十分に検討されている状況にはない．

そこで，本稿では，残差 IDF と WebPMI を用いて Q&A サイトから土産物名を抽出する手法を提案する．また，提案手法の効果について検証する．提案手法は，奥らの手法と遠藤らの手法を組み合わせたものであり，前提知識なしで網羅的に土産物名の抽出を目指すものである．

Extraction method of names of souvenirs using Residual IDF and WebPMI
†Satoru Kawano・Graduate School of Science and Engineering, Kinki University
‡ Shoji Mizobuchi ・ Department of Informatics, Kinki University

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

2.　提案手法

提案手法は，次の(1)から(4)までの手順に従って，Q&A サイトから土産物名を抽出するものである．これらの手順をまとめたものを図 1 に示す．

(1) Q&A サイトからのテキストを取得

土産物に関する質問に対するベストアンサーを Q&A サイトからテキストとして取得する．今回利用する Q&A サイトは Yahoo!知恵袋[1]とする．

(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．また，辞書は標準 IPA 辞書のみを使用する．

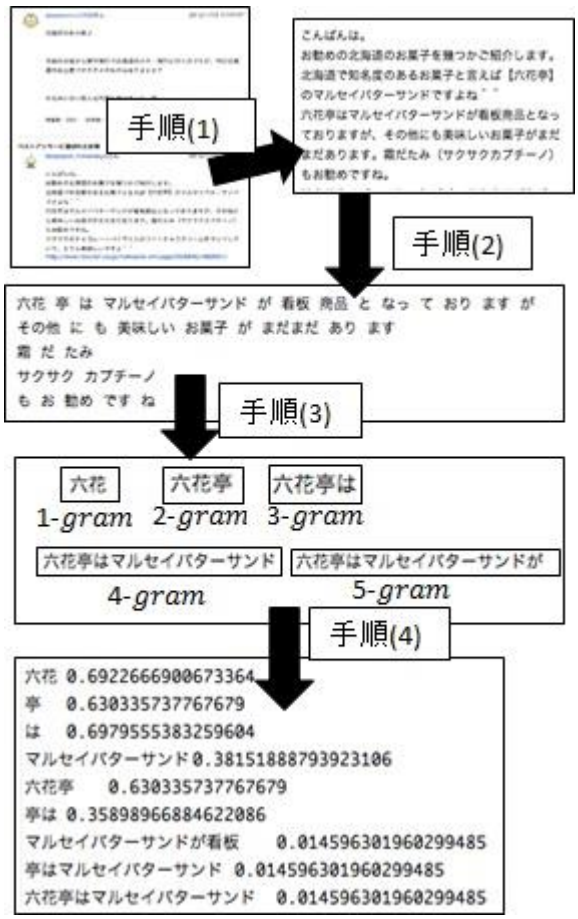


図 1　提案手法の処理手順

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

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.

(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．

(4) 形態素 N-gram の重み付け

手順(3)で生成した1から5までの形態素N-gramに対して，残差 IDF と WebPMI 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定された 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．

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

また，WebPMI は，検索エンジンを用いて語の共起度合いを測る尺度である．2 つの語 p と q に対して，$H(p)$，$H(q)$，$H(p \cap q)$をそれぞれ「p」，「q」，「p AND q」をクエリとして検索した時のヒット件数とするとき，p と q の WebPMI は次式で計算される．

$$WebPMI(p,q) = \begin{cases} 0 & if\ H(p \cap q) \leqq c \\ log\frac{\frac{H(p \cap q)}{N}}{\frac{H(p)}{N}\frac{H(q)}{N}} & otherwise \end{cases}$$

これらの値から，Xの重みは，次式で計算される．

$$\omega = RIDF(X) \times (1 + WebPMI(X,L))$$

ここで，WebPMI に 1 を加えるのは，WebPMI が 0 の時，ωが 0 になってしまうのを防ぐためである．また，式中のLは，場所を表す語とする．

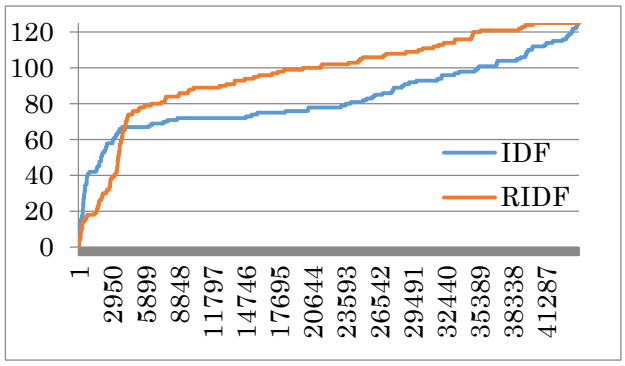
3. 検証

提案手法の効果を検証するために，文献[2]で提案されているIDFと WebPMI を用いた手法との比較を行った．

今回，検証するにあたり，北海道の土産物に関する質問のベストアンサーを Yahoo!知恵袋から 100 件取得した．それらの文書に対し形態素解析を行い，1 から 5 までの形態素 N-gram を生成した結果，形態素 N-gram の個数は 44,224 個となった．さらに，その中から人手で土産物名を判定した結果，土産物名と判定された形態素 N-gram の個数は 125 個となった．WebPMI の計算には，楽天市場の商品検索 API[3]を使用した．この際，N は楽天市場の全商品数である 166,375,075 に，cは 0 に設定した．ωを算出する際のLには，「北海道」を使用した．

このような設定のもとで，各形態素 N-gram に対して，残差 IDF と WebPMI を利用した重みと，

表 1　実験結果



IDF と WebPMI を利用した重みを算出し，それぞれの場合について順位を求めた．この結果を表 1 に示す．表 1 の横軸は，各形態素 N-gram に対して算出した重みの順位である．また，縦軸は，各順位までの形態素 N-gram のうち，土産物名であった形態素 N-gram の個数である．

表 1 によれば，上位は IDF を用いた手法の方が，土産物名を多く検出しているが，中位以降は残差 IDF を用いた手法のほうが，土産物名を多く抽出している．つまり，全体的に見れば，提案手法のほうが，土産物名の分布を上位に押し上げるような結果となっている．

4. おわりに

本稿では，残差 IDF と WebPMI を用いて Q&A サイトから土産物名を抽出する手法を提案した．提案手法の効果を検証したところ，網羅的に土産物名を抽出した場合，IDF を使った手法より，土産物名の分布を底上げする効果があることを確認した．

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] 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, 5(3), pp. 97-116 (2012)

[3] 上原 尚，嶋田 和孝，遠藤 勉: Web 上に混在する観光情報を活用した観光地推薦システム，電子情報通信学会技術研究報告, 112(367), pp. 13-18 (2012)

[4] 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)

[3] https://webservice.rakuten.co.jp/api/ichibaitemsearch/

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.