3K-02

土産情報 DB 構築に向けた品名候補の抽出

長尾 哲志† 安藤 一秋‡

香川大学大学院工学研究科† 香川大学工学部‡

1．はじめに

近年，オンラインショップの増加により，多種多様な商品が時間と場所を選ばずに購入できるようになった．それに伴い，現地でしか購入できないという特徴が旅行時の土産選択・購入時に注目されるようになった．しかし，現在，現地でしか購入できない土産に関する情報を一元的に提供しているサイトやサービスは存在しない．

現地でしか購入できない土産情報は，QA サイトや口コミサイト，ブログなどに散在している．そこで本研究では，現地でしか購入できない土産の各種情報（品名，店名，評判など）を Web 上から収集・整理して，ユーザに提示するシステムの構築を目的とする．観光情報を対象とする研究[1-3]はいくつか存在するが，土産情報に特化した研究[4]はほとんどない．

本稿では，まず土産情報提示システムと土産情報 DB の概要を説明する．そして，ブログ記事から品名を自動抽出する手法を提案し，評価した結果を述べる．

2. 土産情報提示システム

提案システムは，ユーザからの入力情報（場所，土産ジャンル，金額など）に基づき，商品のレア度（現地でしか購入できない度）や最寄りの店舗を付与した土産情報を，評判や金額，現在地からの距離などを基に地図やランキングを使って提示する[5].

提案システムでは，Web 上から収集した様々な土産情報を土産情報 DB に登録して利用する．図 1 に土産情報 DB 構築法の流れを示す.

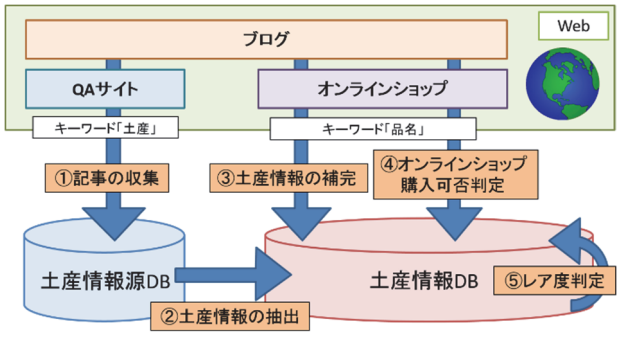

図 1: 土産情報 DB を構築する処理の流れ

以下，土産情報 DB の構築手順を示す.

Step 1: 記事の収集

ブログと QA サイトから，キーワード「土産」を含む記事を収集し，土産情報源 DB に登録する.

A Method to Extract Candidates of Souvenir Names toward Constructing a Database for Souvenir Information
† Graduate School of Engineering, Kagawa University
‡ Faculty of Engineering, Kagawa University

Step 2: 土産情報の抽出

土産情報源 DB 内の各記事から品名とその他の土産情報（会社名/店舗名，評判など）を自動抽出して，土産情報 DB に登録する.

Step 3: 土産情報の補完

Step 1 で収集した記事に含まれていなかった土産情報（例えば，販売場所，ジャンル，個数情報，土産の写真，賞味期限など）をオンラインショップと検索エンジンを利用して自動収集し，土産情報 DB を補完する.

Step 4: オンラインショップで購入可否の判定

オンラインショップと検索エンジンを利用して，土産情報 DB 内の各商品がオンラインショップで購入可能であるかの判定を行い，その結果を DB に追加する.

Step 5: 商品のレア度の計算

土産情報 DB 内の各商品に対し，オンラインショップでの購入可否，ブログや掲示板などの口コミ情報などを基に，レア度（現地でしか購入できない度）を計算する.

以降，本稿では，土産情報の抽出に必要不可欠となる「品名」をブログ記事から自動抽出する手法を提案する.

3. ブログ記事からの品名抽出手法

現在，現地以外で買える／買えないにかかわらず，土産情報を一元的に管理・提供する情報源は存在しない．また，現地でしか購入できない土産は，オンラインショップで取り扱われないため，そのサイトから商品情報を抽出することができない．そこで本研究では，ブログと QA サイトから商品情報の自動収集を目指す.

本稿ではブログ記事から品名を自動抽出する手法を提案する．提案手法は品名候補の抽出と妥当性判定の 2 つの処理で構成される．まず，品名候補の抽出手順を示す.

① ブログ記事からの品名候補の抽出

手順 1：ブログ記事の本文から切り出した各文に対して，36 個の正規表現規則（例えば“この[品名候補]を購入”）を適用し，品名候補（文字列）を抽出する．これらの規則は，Yahoo!ブログの「お菓子・デザート」カテゴリ内の品名が出現する 76 件のブログ記事から 105 文を抽出し，一文内での品名の出現傾向を分析し，人手で作成した.

手順 2：適用した規則に従い，次の (a) or (b) を実行する.
(a) 適用規則が括弧のみで構成される場合，括弧内の文字列を品名候補として抽出する．(b) (a)以外は，抽出した文字列を形態素解析して「複合名詞」or「複合名詞+“の”+複合名詞」となる部分を候補 *c_name* として抽出する.

人間が品名候補で Web 検索した際，スニペットを参照することで，その候補が品名かどうかを判断できる．そこで，スニペットには品名候補の妥当性を判断できる情報が潜むと仮定し，品名候補の妥当性判定に，品名候補で Web 検索したときのスニペットを利用する．以下，妥当性判定の手順を示す.

②品名候補の妥当性判定

手順 1：①で得られた品名候補 *c_name* が店舗名の場合，品名候補から除外する．なお，店舗名の判定には，タウンページ等からエリア別に収集した情報を利用する．

手順 2：残りの *c_name* をクエリとして検索エンジンで検索し，検索結果を取得する．

手順 3：検索結果から上位 *n* 件のスニペットを抽出し，BOW（bag-of-words）を素性とした SVM（Support Vector Machine）で，*c_name* が品名判定に有用なスニペットか否かを判定する．学習にはスニペット単位で正誤ラベルを付与したデータを利用する．

手順 4：スニペット *n* 件に対する各々の判定結果 $eval_k$ と検索結果における順位 *k* を用いて，*c_name* の妥当性判定値 *score* を計算し，その値が閾値 α 以上の場合，*c_name* を品名として出力する．*c_name* に対する *score* は，式(1)で計算する．

$$\mathrm{score}(c\_name) = \frac{\sum_{k=1}^{n} log(n-k+2)*(eval_k)}{\sum_{k=1}^{n} log(n)*1} \quad (1)$$

ここで，*n* はスニペットの総数，*k* は順位，$eval_k$ は SVM の判定結果が品名の場合は 1，さもなければ 0 である．

4. SVM によるスニペット判定の予備実験

②品名候補の妥当性判定は，手順 3 の SVM によるスニペット判定の性能が影響する．そこで，スニペット判定の性能を予備実験で確認する．

実験には，ブログ記事から抽出した品名 35 件を利用する．各品名に対する上位 10 件のスニペット 350 件を収集し，品名判定に有用か否かに基づき人手で正例（153 件），負例（61 件），ニュートラル（136 件）に分類した．このうち，正例と負例のスニペットを評価データに利用する．SVM の素性には，スニペットの BOW（1,600 次元）を利用し，10 分割交差検定で評価する．また，次元削減による効果も確認する．SVM は scikit-learn の LinearSVC，次元削減には TruncatedSVD を利用する．

実験結果を表 1 に示す．次元削減なしの場合，F1 値が 0.816 となった．そして，50, 100, 150, 200, 300, 400, 500 次元に削減した結果，100 次元まで削減したときの F1 値が最も良く，0.868 まで向上した．

表 1:スニペット判定の実験結果

	次元削減なし	次元削減あり						
次元	1,600	50	100	150	200	300	400	500
precision	0.867	0.875	0.888	0.874	0.883	0.888	0.874	0.867
recall	0.844	0.867	0.876	0.857	0.849	0.856	0.853	0.846
f1	0.816	0.859	0.868	0.839	0.830	0.834	0.831	0.821

次に，品名ではない 15 クエリを使って収集したスニペットから品名判定に有用ではないスニペット 60 件を抽出し，先の実験で利用した評価データに負例として追加し，以下の 2 つのテストデータに対して性能を調査する．

テストデータ 1 は，品名と品名ではないクエリを 10 クエリずつ検索した結果から上位 10 件のスニペットをそれぞれ抽出して合計 200 件のスニペット集合とする．テストデータ 2 は，“スイーツ”や“商品”などの一般名と店名（フィルタリング漏れを想定）に対する判定性能を調査するため，この種の 10 クエリを検索した結果から上位 10 件のスニペットをそれぞれ抽出して合計 100 件のスニペット集合とする．なお，SVM の素性には 100 次元まで圧縮した BOW を利用する．

テストデータ 1 と 2 に対する実験結果を表 2 に示す．表 2 に示すように，いずれのテストデータに対しても 100 次元まで圧縮した結果が高く，F1 値がそれぞれ 0.895 と 0.742 となった．以上の実験より，スニペット判定の有効性が確認できたといえる．

表 2: クエリ単位での実験結果

	学習データ+テストデータ1		学習データ+テストデータ2	
	2,890次元	100次元	2,805次元	100次元
precision	0.800	0.810	1.000	1.000
recall	1.000	1.000	0.520	0.590
f1	0.889	0.895	0.684	0.742

5. 品名候補の妥当性判定の評価

テストデータ 1 と 2 を用いて，②品名候補の妥当性判定の有効性を評価する．なお，SVM の素性には 100 次元まで圧縮した BOW を利用する．

閾値 α を 0.6 として，品名候補毎に妥当性判定した結果を表 3 に示す．表 3 に示すように，テストデータ 1 に対して，正解クエリの 1 クエリ以外は全て正しく判定できた．正しくされなかった「ごま蜜団子」は，漫画に登場する名前であったため誤った判定がされてしまった．テストデータ 2 について，誤って判定した 4 クエリのうち 3 クエリは店名であった．店名がフィルタリングできない場合，品名と判定されてしまう可能性が確認できた．残りは「バームクーヘン」という一般的な品名であったが，品名として成立する場合もあるため，問題ないと考える．

表 3: 品名候補毎に妥当性判定した結果

	品名である	品名でない
テストデータ1	9/10	10/10
テストデータ2	-	6/10

6. まとめ

本稿では，土産情報 DB 構築するため，ブログ記事から品名を自動抽出する手法を提案し，実験により有効性を確認した．今後の課題として，閾値 α を 0.6 としたが，この値を精査した後，学習モデルを未知のデータに適応した結果について調査する．また，使用するスニペットの件数を変化させた場合の性能を評価する．

参考文献

[1] 石野他，“旅行ブログエントリからの観光情報の自動抽出”，日本知能情報ファジィ学会誌，Vol.22, No.6，pp.667-679，2010.

[2] 上原他，“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”IEICE 技術研究報告, 112(367), pp.13-18, (2012).

[3] 奥他，“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”, IPSJ 論文誌データベース, 5(3), pp. 97-116, (2012).

[4] 川野他，“Q&A サイトを対象にした地域別土産物情報収集ツール”，FIT2015 講演論文集，pp.221-222，(2015).

[5] N. Nagao, et al., “Extraction of Product Names for Constructing a Database of Souvenir Information”, Proc. of the fifth International Conference on Informatics and Applications, pp.88-96, (2016).