

ナイーブベイズ法に基づく SNS を利用したペルソナ推定

Persona Estimation using SNS based on Naive Bayes Method

山岡 拓生¹ 佐野 睦夫¹

Hiroki Yamaoka¹ Mutsuo Sano¹

¹大阪工業大学, 情報科学部

¹Media Information Department, Faculty of Information Science and Technology,

Osaka Institute of Technology University.

Abstract : High-capacity data with various types of information updated frequently is called big data and SNS is also one of them. In recent years, big data has improved research /product service quality and management strategy, etc. This research is aimed at estimating detailed items of persona in marketing, and we propose individual posting from big data in this paper. We propose an individual posting method from big data and hobby analysis method for users. The proposed methods can make it easy to detect the language features and analyze SNS. We apply a Naive Bayes method to create classifiers and estimate their hobbies. The result of estimating their hobbies was an average of about 93% correct answer rate. We discuss that the proposed methods could create their persona.

1. はじめに

本論文では、Twitter 上の個人アカウントの投稿情報より趣味の推定を行う。この推定を行う事によって、ターゲットの趣味や趣味となりうる趣味の推定を行う事を目的としている。

近年、様々な形でビッグデータが盛んに利用されている。その利用範囲は経済や医療、その他のサービス等でデータを利用した解析や研究などが行われている。その中でもマーケティング分野での利用は大きいと考えられる。しかしながら、ビッグデータを利用した解析などにおいてはすでに確立されたメディアに対して行うものが多く、新規事業での利用はあまり多くないといえる。そこで本研究では、マーケティング手法の 1 つであるペルソナマーケティングにおけるペルソナをビッグデータを利用し、作成を行う事を目的とし、今回はその中でも趣味の項目を作成した。

2. ペルソナ項目推定

本研究の目的にあるペルソナとその項目とは、ペルソナ/シナリオ法というマーケティング手法に利用し作成されるものであり、架空のユーザやターゲットの代表像の事である。ペルソナは実際に存在するユーザのデータを基にして作成されており、その

項目はパーソナルデータなど様々な項目が存在している。項目の例を次の図 1 に示す。



図 1 ペルソナ

本研究では、趣味項目の分類に機械学習のナイーブベイズ法を利用し、学習には Twitter 上の文章を利用した。

本研究では趣味の分析を行うが、実際の趣味の数は膨大であり抽象的な趣味もあるため、今回は統計局で公開されている平成 28 年社会生活基本調査の結果 [1]より、年代を問わず総計での人口が多い趣味をいくつか対象として分析を行った。

また本研究での分析は、次の 4 つの手順で分析を行う。

まず初めに 1 つ目は分析のためのデータ収集である。2 つ目にデータ分析はほとんどが収集したデータそのままの形式では利用ができないため、前処理が

必要となる.3つ目に2つ目の前処理を行って成形されたデータを用いて分類器を作成する.4つ目は三つ目で作成された分類器を用いて未知データの分類を行う事である.システム概要は以下の図2にまとめて記述している.

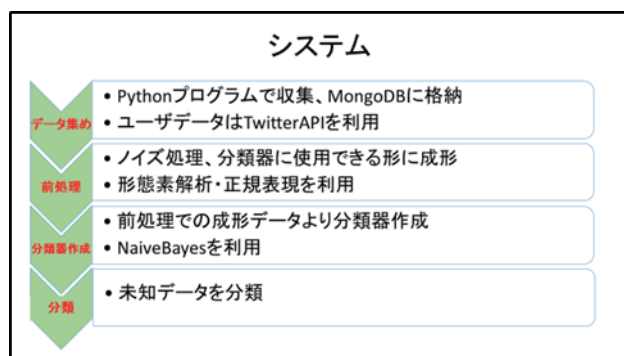


図2 本研究システム概要

次項から、システムの各手順について詳述していく。

2.1. データ収集

本手順では、TwitterDeveloper より提供されている TwitterAPI [2]と Python [3]を用いて、Twitter [4]から各趣味に適したユーザ情報を抽出し、データベースに蓄積を行う。また、本手順で Twitter [1]を利用した理由としては、TwitterAPI を利用することによってプログラムでの操作が容易となることや、リアルタイム性の高い情報を得られることなどがあり、その他にも Twitter には、自身の趣味に関しての投稿に制限して投稿を行う趣味アカウントなどが存在しており、それらは個人の趣味嗜好に沿った考えが近いデータを取り扱えることなどがあげられるためである。

2.2. 前処理

前処理の手順では、収集されたデータを後述する本研究で用いられる分類器の作成に利用するために、データの成型を行う。詳細は後述するが、本研究では NaiveBayes 法を用いてデータの分類を行う。NaiveBayes 法を適用するためには、ユーザのテキスト情報を形態素に分解する必要がある。よってそれに伴いデータの成型を行うために形態素解析を行う。形態素解析には Python で利用しやすい形態素解析ライブラリである Janome [5]を利用する。

2.3. 分類器作成

分類器作成では、前処理で形態素解析によって分解された文章の語彙の中から名詞と動詞を抽出して分類器を作成する。また今回の分類器作成には前述したように NaiveBayes 法を用いて作成する。NaiveBayes 法については以下に詳述する。

2.3.1. NaiveBayes 法

NaiveBayes 法とは教師ありの機械学習の1つであり、テキスト分類を行う手法の中でも単純な仕組みであるが高速で精度が高い手法である。この手法は、ベイズの定理を応用した次の式(1)で表すことができます。式中の C はカテゴリを表し、 D は文書を表しています。ここでの文書とは単語の集合のことです。

$$P(C | D) = (P(C)P(D | C))/P(D) \propto P(C)P(D | C) \quad (1)$$

$$P(D|C) = P(W_1 \wedge \dots \wedge W_k | C) = \prod_i P(W_i | C) \quad (2)$$

この式(1)における $P(C|D)$ は事後確率とよばれ、本手法ではテキストのカテゴリが未知の文書では、事後確率が最も高いカテゴリへと分類をします。この分類推定を MAP 推定と呼びます。

ここでの文書 D は bag-of-words であり、単語間に繋がりが無い単語 W の集合であり、その単語間の独立性を仮定した時に以下の式(3)は成り立つ。

ただし、実際の文書では単語間で出現率に共起が起こるため単語間の独立性が成り立つことはない。そのためあくまで仮定を行った上で式(2)の様に文書の確立の積で表す式が NaiveBayes 法である。

また、式(2)における $P(W_i | C)$ は単語の条件付き確率とも言い、カテゴリ中の単語の出現のしやすさを表している。これは訓練データ中の特定カテゴリ C におけるある単語 W の出現数をカテゴリ C 中の全単語数で割ることによって求めることができる。

以上の結果よりまとめた式が次の式(3)である。

$$M = \operatorname{argmax}_C \log P(C|D) = \operatorname{argmax}_C \log P(C) \prod_i P(W_i | C) \quad (3)$$

NaiveBayes 法での実装式は上記のとおりだが、実装に関してゼロ頻度問題が発生する。このゼロ頻度問題とは訓練時に存在しなかった未知の単語が含まれている文書を予測する場合に、計算式に席を用いているために結果がゼロになってしまうという問題がある。この問題には学習にない未知の単語の出現回数に $\alpha=1$ を加算するラプラススムージング等が代表的な対策であり、本研究ではこちらのラプラススムージングと出現回数に $0 < \alpha \leq 1$ の範囲で値を加算する Lidstone スムージングを採用した。今回の α には $\alpha = 0.1, 0.5, 1$ の 3 つの値で検証を行った。

2.4. 分類

本手順では、前手順で作成した分類器を用いて未知の文書を分類した。分類結果には K 分割交差検証と混同行列を用いて分類器の性能評価をおこなった。

3. 結果

表 1 K 分割交差検証結果

α	0.1	0.5	1
平均正解率	0.8704	0.8600	0.8496
正解率の標準偏差	0.0657	0.0587	0.0636

表 2 混同行列 性能評価 $\alpha=0.1$

$\alpha=0.1$	Class0	Class1	Class2	Class3	Class4
正解率	0.9547	0.9342	0.9136	0.9259	0.9506
検出率	0.9167	0.8776	0.8140	0.8039	0.8125
精度	0.8049	0.8113	0.7292	0.8367	1.0000
真陰性	0.9614	0.9485	0.9350	0.9583	1.0000

表 2 混同行列 性能評価 $\alpha=0.5$

$\alpha=0.5$	Class0	Class1	Class2	Class3	Class4
正解率	0.9588	0.9300	0.8848	0.9300	0.9342
検出率	0.9429	0.8462	0.7174	0.8810	0.7647
精度	0.8049	0.8113	0.7292	0.8367	1.0000
真陰性	0.9614	0.9485	0.9350	0.9583	1.0000

表 3 混同行列 性能評価 $\alpha=1$

$\alpha=1$	Class0	Class1	Class2	Class3	Class4
正解率	0.9630	0.9218	0.8724	0.9300	0.9177
検出率	0.9706	0.8148	0.6889	0.9211	0.7222
精度	0.8049	0.8302	0.6458	0.7143	1.0000

真陰性	0.9617	0.9524	0.9141	0.9317	1.0000
-----	--------	--------	--------	--------	--------

4. 考察

前章の結果より、表 1 の結果から $\alpha=0.1$ が最も高い平均正解率であり、最も低い $\alpha=1$ と比較してもその差は 3%程度ということがわかる。このことより、 α が小さいほど本実験での分類器は精度が高い傾向にあるということがわかる。 α はテストデータに未知語が存在している場合に影響があるため、まだ今回のデータには未知語が多い。または未知語に対しての対応力がまだ低いということがわかる。これに関しては収集されるデータ量が増加した場合に今回より結果がよくなると考えられる。これは今後の課題になると考えられる。

5. まとめ

本研究では、背景として近年多様な利用範囲とニーズがあるビッグデータの利用方法の 1 つである小売分野のレコメンデーションに着目した。その中でもユーザー重視のマーケティング方法であるペルソナ/シナリオ法に注目し、その多様な項目の作成を行う事を目的とした。本研究ではその中でも趣味の推定を行う事を目的とした。

本研究の実現方法には、ビッグデータの Twitter を対象に投稿情報を利用して趣味の推定を行う分類器を教師あり機械学習の 1 種類であるナイーブベイズ法を用いて作成を行う事により実現させた。

実現させた分類器の性能評価を行うことによってその有効性を明らかにすることを目的として実験に K 分割交差検証と混同行列の 2 種類の手法を利用した。

結果として、表 3.2 より全体の分類は最も高い平均として約 93%を示したが、検出率と精度を確認すると正しい分類以外にも誤分類も同時に行っていることも考えられた。全体として分類性能は約 80%程度とやや低い性能であるといえる。この要因として、クラス毎のデータ数によってはデータ数が少ないため分類に必要な語彙を発見できていなかったのではないかと考えられる。

以上の事より今後の課題として、各クラスのデータの増量が必要であることが挙げられる。また、今回は 5 クラスの分類と実際の趣味で考えた場合ではかなり狭い範囲にのみ対応する形となっているためより多くのクラスを作成する必要がある。これによって増加したデータでの分類器の性能評価も重要であると考えられる。

また、本研究での分類器作成によって趣味の分類を行う事は可能であることが証明されたため、他のペルソナ項目にも応用が可能であるかの検証も今後の課題として挙げられる。

今後の展望としては、前述した課題の解決とその他の機械学習法との比較や他の項目での適用方法を考案していく。

謝辞

本研究を行うにあたって、ご指導をいただきました指導教員の佐野睦夫教授に深謝の意を表する。また、本研究を行うまでの中で知識や示唆をいただいた自然言語処理研究室の平博順准教授に深謝の意を表する。その他にも、日常の議論を通じて多くの知識や示唆をいただきましたインタラクシオンデザイン研究室の大井翔氏をはじめとした同研究室の皆様へに深謝の意を表する。

参考文献

- [1] “統計局ホームページ/平成 28 年社会生活基本調査の結果,” [オンライン]. Available: <http://www.stat.go.jp/data/shakai/2016/kekka.htm>. [アクセス日: 20 12 2017].
- [2] Twitter, Inc., “Twitter Developer Platform - Twitter Developers,” Twitter, Inc., [オンライン]. Available: <https://developer.twitter.com/>. [アクセス日: 25 01 2018].
- [3] Python ソフトウェア財団, “Welcome to Python.org,” [オンライン]. Available: <https://www.python.org/>. [アクセス日: 31 01 2018].
- [4] Twitter Inc., “Twitter,” Twitter, Inc., 15 07 2006. [オンライン]. Available: <https://twitter.com/>. [アクセス日: 25 01 2018].
- [5] T. Uchida, “Welcome to janome’s documentation! (Japanese) – Janome v0.3 documentation (ja),” 2015. [オンライン]. Available: <http://mocobeta.github.io/janome/>. [アクセス日: 31 01 2018].