F-010

デジタル・コンテンツへの反応に基づく外国人の日本文化への興味抽出
Extracting Foreigner Interests of Japanese Culture from Interactive Digital Contents

レーティゴック † **植村 喜弘** † **島川 博光** ‡
Le Thi Ngoc Yoshihiro Uemura Hiromitsu Shimakawa

1. はじめに

日本は，諸外国を魅了する有形・無形の文化財を持っている．外国人に対し，日本文化への理解と関心を高める目的で，多様な文化芸術活動がおこなわれている [1]．日本在留の外国人に日本文化への理解と関心を高める利点として，２つ挙げられる．１つ目は，外国人が日本文化を理解し，魅力に触れ，日本での生活に対する精神的な充実感がより得られることである．２つ目として，これが日本の文化財の維持・継承・発展に貢献することである．しかし，異文化への不理解や言語の壁のため，外国人は，自身の興味対象が分からない．

本研究では，ソーシャルネットワークサービス（SNS）上のインスタントメッセージ（IM）の即時性の高さを活かし，デジタル・コンテンツへの反応から，日本文化の興味を発見する手法を提案する．これに基づいて，日本文化への興味対象がわからない外国人をサポートするシステムを構築する．

2. 研究背景

文献 [2] では，ツイッターにおける大量のコンテンツからユーザの興味を推定する手法が提案されている．この研究では，手法を適用するために膨大な投稿が求められる．したがって，本研究では，即時性の高い IM を活かし，膨大なデータを必要としないユーザの興味を発見する手法を提案する．これにより，日本在留の外国人が，より自身の興味のある日本文化に触れ，充実感を向上することができる．

3. 興味要素分析手法

3.1 システム概要

本研究では，チャットログを分析することで, ユーザの持つ興味を抽出し，話題を提供するシステムを提案する．本研究では，ユーザの興味を抽出するために，興味分野と興味要素をする．興味分野とは，自然・歴史・日本画のように，日本文化におけるあるジャンルを示す．それぞれの関連語句は，興味分野のうち，もっとも関連性が強いものに属すると捉える．よって，興味分野に属する関連語句は排反であると仮定する．

ユーザが最も興味を持っている可能性の高い興味分野を興味要素とする．これは，４往復のチャットログごとに推定される．

これにより，会話を通して時系列的に興味を分析することができる．システムにより推定された興味要素に基づき，関連する話題を提供することで，ユーザは興味を発見することができる．

提案システムの概要を図１に示す．本システムは４往復分のチャットログをチャットセグメントとし，これを分析することで興味要素を推定する．そして，興味要素に基づき，話題を提供する．日本文化に対する興味の発見を求める外国人をユーザ，ユーザをサポートする人物をコンサルタントとし，ユーザとコンサルタントが SNS やメッセンジャーなどを通し，本システムを利用する．

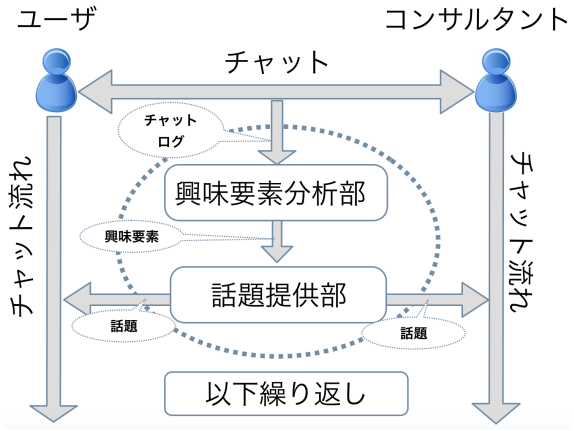

図 1: システム概要

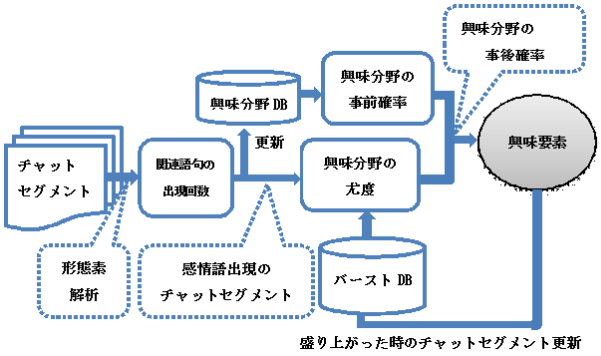

図 2: 興味要素分析部の内部構造

まず，コンサルタントが自身の経験に基づき，ユーザへ話題を提供する．会話を進めていく中で，チャットセグメントごとに，どの興味分野の関連語句がどれだけ出現したかを記録しておく．さらに、特定のチャットセグメントにおいてそのユーザが興味を示したとき、興味要素とチャットセグメント内のすべての関連語句を組にして記録しておく．

3.2 興味要素の分析

新たなチャットセグメントが発生したとき、上記の記録を使って、ユーザが持つ興味要素を興味要素分析部が推定する．話題提供部では，推定された興味要素に関する話題がユーザとコンサルタントの双方に提供される．

興味要素分析部の概要を図２に示す．まず，チャットセグメントに出現する関連語句を形態素解析する．本研究では関連語句を名詞に限定する．興味要素は，ベイズの定理に基づき算出される事後確率により決定する．これまでのすべてのチャットに現れた関連語句について，興味分野ごとの関連語句の出現回数が，全分野の関連語

† 立命館大学大学院情報理工学研究科
‡ 立命館大学情報理工学部

句の出現回数に占める割合として，ある興味分野にユーザがもっとも強い興味を示す確率，すなわち，ユーザの特定の興味要素が発生する事前確率を求める．話題が盛り上がった，過去のチャットセグメントに現れた関連語句を使い，ある興味要素がわかっているときの，話題が盛り上がるチャットセグメントの発生確率を尤度と考える．事前確率と尤度の積に，事後確率は比例する．事後確率は，話題が盛り上がったチャットセグメントが得られたときに，それが特定の興味要素が原因であることを示している．よって，事後確率がもっとも大きい興味分野が，興味要素であると推定できる．

3.3 興味分野と話題の盛り上がり

本研究では，過去のすべてのチャットセグメントで，各興味分野に対し，それに属する関連語句が出現した回数を記録した興味分野データベースをあらかじめ準備する．興味分野データベースは，興味分野とその関連語句の出現回数からなるテーブルである．

過去のチャットセグメントにおいて，興味を持っているときに発せられる感情語が現れたものを収集する．本研究で扱う感情語は，“wow”，“definitely”，“uhm”などである．これらのチャットセグメントは，過去のユーザの興味を示したものである．そこに現れた語句の集合を，新しいユーザが発生させた場合，その新しいユーザも過去のユーザと同じような興味を持っているとみなす．

さらに，本研究では，これらの感情語が現れたときのチャットセグメントを，ユーザの興味要素とともにバースト・データベースに記録しておくものとする．バースト・データベースは，話題が盛り上がったチャットセグメントの集合である．このチャットセグメントは，興味要素と，そこに現れた関連語句の集合の組である．

3.4 ベイズモデルの応用

本研究では，ベイズモデルを応用することで，チャットセグメントを得たときの，事後確率が最も高い興味分野を興味要素とする．チャットセグメントを doc，興味分野を cat とする．doc を得たときに cat の事後確率 $P(cat|doc)$ はベイズの定理に基づき，以下の式によって与えられる．

$$P(cat|doc) = \frac{P(doc|cat)P(cat)}{P(doc)} \tag{1}$$

ここで，$P(doc|cat)$ は過去のチャットで，特定の cat に得たチャットセグメント doc が含まれる確率，$P(cat)$ は特定の cat が発生する確率，$P(doc)$ は特定の doc が過去のチャットで発生する確率である．また，興味要素は興味分野の事後確率の大小関係から求められるため，式 1 に比例する以下の式から求める．

$$P(cat|doc) \propto P(doc|cat)P(cat) \tag{2}$$

$P(cat)$ は，過去のチャットデータを訓練データとし，各興味分野のチャットセグメントが総チャットセグメントに占める割合である．訓練データは興味分野データベースからのデータである．

興味分野データベースでは，i 番目の興味分野 cat を cat_i とする．それに属する各単語はもっとも関連が深い興味分野にしか属さないとし，それの出現回数を f_i を格納する．よって，$P(cat)$ を算出できる．

本研究では，bag-of-words 法に基づき，単語間の独立性を仮定する．すなわち，チャットセグメント doc を解析から得た関連語句の集合 $word_1$, $word_2$, …, $word_k$ と捉え，単語の順序を無視する．よって，$P(doc|cat)$ を以下の式で求められる．

$$\begin{aligned} P(cat|doc) &= P(word_1 \wedge ... \wedge work_k|cat) \\ &= \textstyle\prod_i P(word_i|cat) \end{aligned} \tag{3}$$

ここで，$P(word_i|cat)$ は，訓練データの興味分野 cat で，そこに含まれる全チャットセグメントの単語数 V に対する，単語 $word_i$ の出現回数の比である．興味分野 cat での $word_i$ の出現回数を $T(cat, word_i)$ とし，$P(word_i|cat)$ は以下の式で計算する．

$$P(cat_i) = \frac{T(cat, word_i)}{\sum_{word_j \in V} T(cat, word_j)} \tag{4}$$

実用上，興味分野 cat に出現しない単語があると，確率が 0 になるため，分母は対象興味分野 cat に出現する単語に限定する．得たチャットセグメントから求める興味要素を cat_m とし，以下の式で算出する．

$$\begin{aligned} cat_m &= \mathrm{argmax}_{\mathrm{cat}} P(cat|doc) \\ &= \mathrm{argmax}_{cat} P(cat) \textstyle\prod_i P(word_i|cat) \end{aligned} \tag{5}$$

ここで，チャットセグメントの中に多くの単語が含まれる場合，$\prod_i P(word_i|cat)$ がアンダーフローを起こす．これを解決するために，式 5 の対数をとり，以下の式で計算する．

$$\begin{aligned} cat_m &= \mathrm{argmax}_{cat} \log P(cat|doc) \\ &= \mathrm{argmax}_{cat} (\log P(cat) + \textstyle\sum_i \log P(word_i|cat)) \end{aligned} \tag{6}$$

ここで，cat_m を大小関係から求められるため，対数をとっても結果は変化しない．式 6 の結果を興味要素とする．

4. おわりに

本研究では，IM におけるユーザとコンサルタントとのチャットログを分析し，ユーザの日本文化に対する興味を発見する方法を提案した．今後の課題としては，バースト・データベースに更新するチャットセグメントの盛り上り度を評価することが挙げられる．これにより，興味を抽出する精度の向上が期待される．今後は，システムがコンサルタントの役割を担当することを目指す．

参考文献

[1] 文化庁:文化発信戦略に関する懇談会の報告 (online), http://www.bunka.go.jp/seisaku/bunkashingikai/sokai/sokai_9/48/pdf/shiryo_10.pdf.

[2] Parantapa hattacharya, Muhammad Bilal Zafar, Niloy Ganguly, Saptarshi Ghosh, Krishna P.Gummadi: Inferring User Interest in the Twitter Social Network, RecSys '14 Proceedings of the 8th ACM Conference on Recommender Systems, 2014.