# チャットルームにおける発言履歴に着目したトピック抽出システムの構築

川端　聖† 佐藤　喬† 村山　隆彦††　多田　好克†
†電気通信大学大学院情報システム学研究科　††NTT情報流通プラットフォーム研究所

## 1　はじめに

チャットとは、ネットワーク上に用意された1ヵ所のスペースにおいて、複数のユーザがリアルタイムに会話を行うシステムのことである。会話が行われている各スペースのことをチャットルームと言う。

チャットルームには、テーマが静的に付けられている (ex. 食べ物、スポーツ、政治)。ユーザは、興味のあるテーマのチャットルームを選択する。しかし、実際のチャットルームでは、そのチャットルームのテーマに沿った自分の興味のある会話が行われているとは限らない。つまり、テーマによってチャットルームを選んでも、自分に最も適しているチャットルームに行けるわけではないという問題がある。このことから、チャットルームにおいて、実際の会話の内容にふさわしいテーマを知りたいという要求がある。

これに対し、本研究では、主にユーザの特徴とチャットルームの発言履歴 (ログ) を用い、それらを現在のチャットルームのログの一部として扱うことで、実際の会話にふさわしいテーマを示す。ここでは、会話のトピック (上位概念のカテゴリに含まれるような名詞群) をテーマとして代用する。

## 2　トピック抽出の課題

チャットルームのトピックを抽出する際の課題は、チャットの特徴から、主に以下の二つが考えられる。

- チャットでは、話題が短時間で変わってしまうことがあるので、短期間の少ない情報量のログで、会話の特徴を抽出しなければならない。
- チャットでは、ニュース記事のトピック抽出 [1] 等と比較すると、特徴が捉え辛いような、整っていない文がよく見られる。このような文でも、うまく特徴を捉えなければならない。

## 3　アプローチ

トピック抽出の課題に対して、本研究では、ログとユーザの特徴を抽出し、それらをマッチングするという手法をとり、主に以下の二つのアプローチを行う。

- 過去に固有の分野の会話を行ったユーザは、現在もその分野の会話を行う可能性が高いと考えられる。短期間の少ない情報量のログで、会話の特徴を捉えるために、これを利用し、過去のログから、各ユーザを特徴付け、そのログを現在の会話を構成する一部として扱うことで、現在のログにユーザの特徴を取り入れる。
- 直前に行われた会話の話題は、現在の会話の話題にもなりやすい。よって、短期間の少ない情報量のログで会話の特徴を捉えるために、直前に行われた会話を、現在の会話を構成する一部として扱うことで、現在のログを拡大する。

## 4　システム概要

アプローチを実現するには、まず、現在のログ、過去のログの特徴として名詞を抽出する。次に、これらのログをマッチングし、マッチングしたデータからトピックを抽出する。これを反映し、本研究では図1のようなシステムとした。

### 4.1　入力データ

入力とするデータは次の二つである。一つは、現在のチャットルームの会話の名詞を抽出するためのトピック推定用データ (現在のログ) である。このデータによって、現在のチャットルームの会話のトピックを推定する。もう一つは、各ユーザの特徴を抽出するために用いる過去ログデータ (過去のログ) である。このデータによって、情報量の少ない現在のチャットルームのログを拡大し、トピックをより強く特徴付けることができる。

### 4.2　システムの各処理

システムは主に、名詞を抽出する三つの処理、ログのマッチング処理、トピック抽出処理の三つに分かれている。

#### 4.2.1　名詞を抽出する三つの処理

発言の中に、どのような上位概念の名詞が存在するかによって、トピックが推定できると考えられる。よってこの処理では、現在のログ、直前のログ、各ユーザが過去に発言したログの名詞を抽出する。名詞は、形態素解析器 MeCab[2] によって抽出する。

---

The construction of the topic extraction system Base on statement history in the chat room
† Takashi KAWABATA
† Takashi SATOU
†† Takahiko MURAYAMA
† Yoshikatsu TADA
The Graduate School of Information Systems,University of Electro-Communications (†)
NTT Information Sharing Platform Laboratories (††)

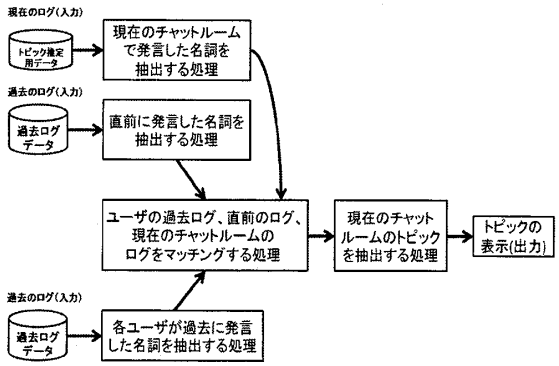


図 1: システムの概要

#### 4.2.2 ログのマッチング処理

名詞の抽出を行った三つのログをマッチングすることで、現在のログの意味を拡大し、トピックをより強く特徴付けることができる。よってこの処理では、名詞を抽出した三つのログをマッチングする。マッチングは、現在のログの名詞に、直前のログの名詞を付加する。また、現在参加している各ユーザの発言比率によって、そのユーザが過去に発言した名詞を、現在のログに存在する名詞数に正規化し、現在のログに付加する。これらを、現在の名詞:直前の名詞:過去に発言した名詞 = 2:1:1 となるようにする。

#### 4.2.3 トピック抽出処理

この処理は、マッチングしたログが、どのトピックかを表す処理である。トピックの抽出には、以下の式を用いる。

$$variance_{max} = max_{i \in C}\{(c_{ti} - \frac{\sum_{i=1}^{C} c_{ti}}{C})^2\} \quad (1)$$

C はトピックの種類の数、$c_{ti}$ は時間 $t$ でのトピックカテゴリ $i$ に属する名詞数である。

## 5 実験と評価

本システムの有用度を調べるために、実際のチャットログを用いて、トピックの抽出を行った。

第一の実験は、7 つのチャットルームを 5 分ごとに区切ったログを用いて、上位 30 個の抽出されたトピックの適合率と上位 10 個の $MAP$(Mean Average Precision) 値を測った。トピックは、食べ物、アダルト、政治経済、趣味、テクノロジの 5 つの中から、1 つ選択される。適合率とは、推定トピックが目視でログにつけたトピックと一致している割合である。$MAP$ とは、以下の式で表される。

$$MAP = \frac{\sum_{m=1}^{M}\{\sum_{r=1}^{R}(\frac{h(r)}{r} * H(r))\}}{M} \quad (2)$$

$M$ はトピックを抽出するチャットルームの数、$R$ は各チャットルームの式 (1) の上位 R 位の数、$h(r)$ は各チャットルームの式 (1) の r 位までの正しい推定トピックの数、$H(r)$ は各チャットルームの式 (1) の r 位の推定トピックが正しければ 1、間違っていれば 0 になる数である。比較対象は、各ユーザの発言ログの有無 (user effect)、直前のログの有無 (just before)、二種類のトピック抽出法 (algorithm)(本手法 (v) と最大の名詞数が属するカテゴリをトピックとした方法 (m)) の組合せで行った。

表 1: トピックの適合率と MAP 値

| user effect | ○ | × | ○ | × | × |
|---|---|---|---|---|---|
| just before | ○ | ○ | × | × | × |
| algorithm | v | v | v | v | m |
| 適合率 | 0.890 | 0.895 | 0.900 | 0.895 | 0.900 |
| MAP | 0.887 | 0.889 | 0.904 | 0.872 | 0.831 |

表 1 から、適合率では、どの手法にも大きな差は見られなかった。MAP 値では、本研究のトピック抽出法が優れている。これにより、信頼性の高いとされる推定トピックでは、本手法の精度が高い。

第二の実験は、直前のログの効果を調べる。実験は、第一の実験と同じ 7 つのチャットルームで、上位 30 個の推定トピックの直後のトピックの適合率を測った。比較対象は、直前のログの有無 (just before) によって行った。その他のパラメタは、本手法を用いる。

表 2: 直後のログの適合率

| just before | ○ | × |
|---|---|---|
| 適合率 | 0.667 | 0.638 |

表 2 から、直前のログを考慮した方が、より継続性のあるトピックを抽出できた。このことから、本手法のトピックを目安に、チャットに参加すれば、抽出されたトピックの会話ができる可能性がより高い。

## 6 まとめ

本研究では、ユーザの過去の発言ログや直前のログを考慮したチャットのトピック抽出システムを提案した。現在のトピックの適合率では、大きな差が見られなかったが、本手法では、各ユーザの過去の発言ログを考慮している。そのため、他の手法より、抽出したトピックの会話を行う土壌が整っていると考えられる。このことは、実際のチャットを行う際に、役立つと考えられる。

### 参考文献

[1] 野美山 浩, 新聞記事データベースからの話題の抽出, 情報処理学会全国大会講演論文集, 第 50 回平成 7 年前期 (4), pp.45–46 (1995).

[2] Mecab:Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net, (accessd 2010-01-12).