F-024

対話からの興味をもとに情報を推薦するボットの作成

Implementation of a Dialog Bot to Recommend Information of Interest to Users by Extracting their Preferences through Dialog

五十島 志織 † 冨永 和人 亀田 弘之 ‡
Shiori Isoshima Kazuto Tominaga Hiroyuki Kameda

1. はじめに

インターネットの発展によりウェブ上にある情報は，年々増加しているが，その中から興味ある情報を的確に取得することは容易ではない．この問題を解決する方法の一つに情報推薦システム [1][2] がある．このシステムは，利用者の興味に合う情報を選別，提示するものの，一方的に情報推薦をする．そのため，満足度の高い情報を推薦するためには，利用者からのフィードバックが必要である．また，個人の興味は流行や情報の鮮度によっても変化するので，流行や鮮度の高い情報を提供しつつ，利用者からのフィードバックを学習するシステムが必要である．

マイクロブログの一種である Twitter[3] は，手軽に情報発信を行うことができる．そのため Twitter では年々利用者数や投稿数が増え，様々な情報が蓄積されている．特に近年では，ブログなどの様々なウェブアプリケーションと連動し，鮮度の高い情報が発信されやすくなっている．Tiwtter には Retweet という再投稿機能があり，この機能から人気や流行の度合いを調べることができる．また，チャットのように特定の個人とやり取りを行うリプライ機能もある．

これらの特性に着目して，筆者らは Twitter から URL を含むつぶやきを収集・分析・分類・蓄積するサブシステムと，利用者との対話から興味を抽出し，その結果を利用して利用者の興味にあった情報を推薦するサブシステムの 2 つを構築する．

2. 提案システムの概要

情報の収集から蓄積までを行う「情報収集サブシステム」と会話を行う「ボットサブシステム」の 2 つを提案する．システムの全体像を図 1 で示す．

まず，情報収集サブシステムでは，言語設定が日本語である Twitter 利用者のつぶやきから，URL を含むものを収集する．URL 先とつぶやき本文を分析後，その結果をもとに URL を独自にタグ付け，分類し，データベース（DB）に蓄積する．

ボットサブシステムでは，利用者とシステムが会話を行う．利用者の会話文を分析し，そこから興味（タグ）を抽出する．その興味をもとに DB から URL 情報を取得し，利用者に返信する．

3. つぶやき情報の分析

Twitter から取得したつぶやき情報を下記の手順で分析する．

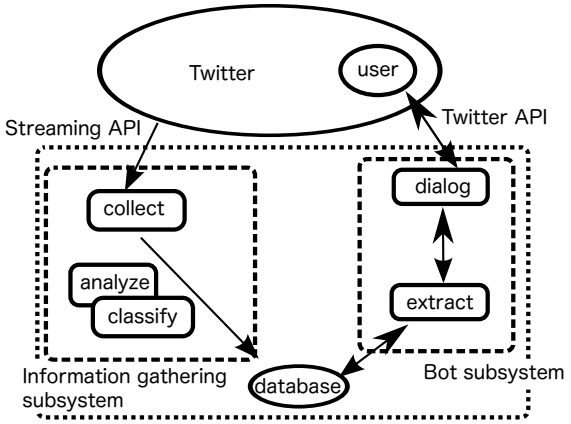

図 1: システムの全体像

手順 1. 本文から URL を分離

手順 2. 本文と URL 先を句読点，記号で分割

手順 3. 形態素解析

手順 4. 合成語計算

手順 5. シソーラスを利用して類義語を抽出

取得したつぶやき情報から本文を抜き出し，本文と本文中に含まれる URL を分離する．その後，URL 先とつぶやき本文を一文単位で形態素解析する．さらに形態素解析された文に対して，図 2 の様に，i, j の位置をシフトさせながら式 (1) を用いて合成語の計算を行う [4]．単語 t_k に対して，直前直後の単語 t_{k-1} と t_{k+1} が出現する確率をそれぞれ求めていく．連結スコア $C_{i,j}$ が低い場合や単語の出現数が 1 以下の場合は合成語の候補から外す．

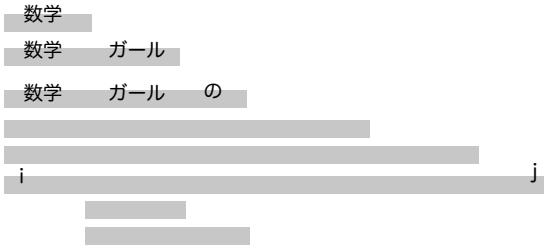

図 2: 複合語の候補生成

$$C_{i,j} = \prod_{k=i}^{j} P(t_{k-1}|t_k)P(t_{k+1}|t_k) \quad (1)$$

また，合成語に含まれる部分文字列も合成語候補とはしない．例えば，合成語「東京工科大学」には，そ

†東京工科大学大学院,Tokyo University of Technology Graduate School
‡東京工科大学,Tokyo University of Technology

の部分文字列「東京」,「東京工科」などが含まれている．合成語の部分文字列になっている文字列は，スコアとして計算しない．ただし,「東京工科」のように単体のみで出現する場合は 1 回と数える．さらに,「の本」や「本の」の様に先頭や末尾が助詞の文字列は出現確率が高いため，先頭もしくは末尾が助詞の場合も，合成語と見なさない．なお，抽出した合成語は，以後も出現する確率が高いため MeCab 辞書に自動登録する．

シソーラスを使い，すでに得られている名詞や合成語の類義語を抽出する．それぞれの単語をシソーラス検索にかけ，類義語を取得する．情報取得サブシステムでは，類義語とシソーラス検索に利用した単語を分類タグとする．

4. 興味抽出処理

利用者の会話文を前述 **3** の手法で分析し，興味を抽出する．シソーラスで類義語を抽出し，単語と類義語を興味情報として蓄積する．一定量興味が蓄積されたら，その興味をキーとして DB 検索にかける．より多くの興味と一致し，なおかつ投稿日時が新しいつぶやきの URL 情報を返す．

5. システムとの対話

人工無脳機能を用いてシステムと利用者が対話を行う．利用者との会話は Tiwtter の返信機能を使い，利用者からは「@ボット ID　本文」，システムからは「@利用者 ID　本文」という形で会話を行う．Twitter とシステムとの会話文の取得や送信は TwitterAPI を使用する．

本システムの会話の種類は,「情報要求」,「フィードバック」,「それ以外」の 3 種類とした．

情報要求とは,「xxx が欲しい」などと直接的な要求をする文のことである．この場合，即時に情報を推薦できるように「xxx」部分を抜き出し，興味抽出を行う．抽出した興味をキーとして DB 検索を行い，URL 情報を取り出す．

会話文が情報要求ではない場合，興味を引き出すような会話を行う．会話文を対象として逐時興味抽出を行い，興味を蓄積していく．ある程度まで蓄積した興味をキーとして DB 検索を行い，URL 情報を取り出す．取り出した情報は,「@利用者 ID　説明文 + URL」という形で利用者に提供する．

6. 動作例

試作段階のシステムを実際に動かして得られた例を図 3，4 に示す．

現在のボットサブシステムは,『「xxx」ほしい』などと「」内に書かれた単語を抜き出し，直接 DB 検索を行い，DB 内に「xxx」と同じタグが存在すれば，その URL 情報を返す仕組みである．図 3 は「Windows」関連の URL を返しているため，上手く利用者の興味と一致する URL 情報を提示することができている．一方，図 4 の興味情報は「農林水産省」であるが，提示した URL 情報は「セシウム」関連の話題であった．これは，Twitter より取得した URL 情報を分析した結果から，

図 3:「Windows」情報の結果

図 4:「農林水産省」情報の結果

出現頻度の高い名詞を分類タグとしているため，分類したタグと URL 情報が一致しているとはかぎらないことが原因である．今後，シソーラスを使い，名詞の上位概念を抽出し，それをタグとすることでより正確な分類ができると考えられる．

7. おわりに

本稿では，Twitter から URL を含むつぶやき情報を収集・分析・分類・蓄積するサブシステムと，利用者と対話を行い，興味に合う情報を提示するサブシステムの 2 つからなる，情報推薦システムを提案した．

今後，開発実験を行うことで，システムの有効性を検証すると共に，情報収集サブシステムの分析方法や分類方法についても検討し，さらに，ボットサブシステムの対話手法や興味と合う情報の抽出方法の検討も行う．現在判明している問題点に，Streaming API の取得漏れがる．これは，Twitter 全体の 1%のつぶやきをランダムに取得するために発生する．収集部を並列化し，取得漏れを少なくする必要がある．

参考文献

[1] 桑原雄, 稲垣陽一, 草野奉章, 中島伸介, 張建偉．マイクロブログを対象としたユーザ特性分析に基づく類似ユーザの発見および推薦方式．情報処理学会研究報告．データベース・システム研究会報告, Vol. 2009, No. 18, pp. 1–3, 2009.

[2] 澁谷翔吾, 廣安知之, 三木光範, 横内久猛．対話的なキーワード抽出によるブログ推薦システム．情報処理学会研究報告．BIO, バイオ情報学, Vol. 2008, No. 126, pp. 155–158, 2008.

[3] Twitter. `https://Twitter.com`.

[4] buzzter. `http://buzztter.com/ja`.