

話題推薦システムのためのモデル構築手法

富坂 亮太*1 鈴木 崇史*2 相澤 彰子*1*3

東京大学大学院 情報理工学系研究科 コンピュータ科学専攻*1

東洋大学 社会学部 メディアコミュニケーション学科*2

国立情報学研究所 コンテンツ科学研究系*3

1 はじめに

「会話」を実現するためのシステムとして、これまで数多くの研究がなされてきた。たとえば、会話システムという研究分野は、ユーザからの「要望」(質問など)をうけ、それに対して適切な応答を探したり、知識ベースから回答を構築したりして、ユーザとのインタラクションを実現する。

このように、ユーザの発言に対して適切な応答を探す類の研究は数多く存在するが1-2)、システム自らがユーザに対して適切な話題を振るという研究はあまりなされていない。

この話題を振るというタスクの肝は話題の選択にある。この、話題の選択に失敗したならば、ユーザとの会話がはずまず、会話が成立しない。

話題の選択において重要なことは、ユーザが「いま」、「何を」求めているかということである。この「いま」「何を」求めているかを探すタスクは、情報推薦に通じる。情報推薦では、ユーザの求めるものを、さまざまな情報から判別し、実際にユーザに提示する。

本稿ではこの情報推薦と会話における話題選択の類似性に注目し、情報推薦の分野で得られた知見を使い、それを会話の切りかえ文の選択に適用する方法を提案し、ユーザ実験により、その有用性を示す。

2 関連研究

Liuらは、文献3)において、ユーザの現時点での興味はユーザの本来の興味に加えて、ユーザが住んでいる地域における話題性に左右されて決定されるものであるということを、大量のGoogle Newsのクリックログの解析によって示し、これに基づきユーザの興味モデルと話題性のモデルを重ね合わせたニュース・カテゴリの選択に基づく記事推薦法を提案している。ここで、ニュース記事の推薦では、複数の記事を同時にユーザに提示するため、モデル構築に用いる記事カテゴリを詳細化しすぎると、同じ話題の記事ばかりが推薦されてしまうことになる。

しかし、会話における話題選択ではユーザに提示する話題はただ1つであり、むしろ、ユーザの現時点での興味に、より特化した推薦が有効であると考えられる。これより本研究では、明示的なカテゴリを想定せず、ユーザの選択した文章全体から直接、詳細なユーザモデル獲得する方法を検討する。

3 提案手法

本稿では、以下の、話題選択のための評価モデル R_{ut} を提案する。

$$R_{ut} = l_u \cdot F_t$$

ただし、ここで、 l_u はユーザの本来の興味モデル、 F_t は話題性のモデルをそれぞれ単語の頻度ベクトルで表したものである。添え字 u は対象のユーザ、 t はある時点を表す。この

R_{ut} に対象の文書の単語ベクトル S を掛け合わせたものが、その文書の評価値となる。それぞれ、ユーザの興味モデル l_u と話題性のモデル F_t の作成方法については以下で述べていく。

3.1 ユーザの興味モデルの作成

本稿では、ユーザが選択した興味のある文章群から単語の頻度ベクトルを作成し、ユーザの興味モデル l_u として用いる。なお、本稿の実験では便宜的にユーザにあらかじめ文章を選択してもらったが、インタラクティブに文章を選択してもらいそのたびにユーザモデルを更新していくことも可能である。

ただし、ユーザが選択する文章は数に限りがあり、そのままでは単語ベクトルが、スパースになりすぎて、正しくユーザの興味を反映できないと考えられる。このため、さらに言語資源を用いて、このベクトルを拡張し、スパースネスを解消することを考える。

この操作は、クエリ拡張4)や、オントロジー構築5)の操作と類似している。これらの研究でよく用いられるWikipediaやWordNetを利用して、モデルのスパースネス解消を図った。

また、本稿では、これらの言語資源ではカバーできない新語や、複雑な略語などを対応する必要がある。これを実現するためにYahoo!の検索エンジンをモデルの拡張の資源として用いた。WikipediaとWordNetについては含まれる用語による最長一致によって、Yahoo!検索エンジンについては、Nグラムの統計情報に基づいて用語きりだしを行った。

3.1.1 Wikipediaによる用語拡張

Wikipediaの各用語 w_1 の記事のアブストラクト部分から用語 w_2 にアウトリンクが張られている場合、用語 w_2 の頻度を用語 w_1 の頻度で拡張する。

3.1.2 WordNetによる用語拡張

用語 w_1 と用語 w_2 が同じ同義語集合に所属する場合、用語 w_1 の頻度を用語 w_2 の頻

Model Construction Method for Topic Recommendation

*1 Department of Computer Science, Graduate School of Information Science and Technology, The University of Tokyo

*2 Department of Media and Communications, Faculty of Sociology, Toyo University

*3 National Institute of Informatics, Digital Content and Media Sciences Research Division

度で、逆に用語 w_2 の頻度を用語 w_1 の頻度で拡張した。

3.1.3 Yahoo! 検索エンジンによる用語拡張

用語 w_1 による検索結果のサマリの部分に含まれる各用語の頻度を用語 w_1 の頻度で拡張した。

3.2 話題語モデルの獲得

本稿では、Twitter のつぶやき総数を話題性の度合いとして見て、用語 w の Twitter でのつぶやきの推移を現在のその用語の話題性の度合いとしてモデルを作成した。

$$F_{wt} = \frac{freq(w, t)}{freq(w, t - \Delta t)}$$

ただし $freq(w, t) < freq(w, t - \Delta t) + 1$ のとき $F_{wt} = 1$

$freq(w, t)$ は用語 w のある時点 t でのつぶやき総数を示す。 Δt は予備実験の結果から6カ月とした。つまり上式は用語 w の6カ月前と当月とのつぶやき総数の変化量を示す。

4 ユーザ実験

提案手法の有効性を検証するために、被験者4名の簡単なユーザ実験を行った。

図1のような画面インターフェースを介して、ユーザに興味のある文章をクリックしてもらい、ユーザの興味のある文章群を獲得した。1被験者ごとに368記事についてデータを作成し、364記事を学習に、4記事をテスト用に用いるクロスバリデーションを行い、得られる92個の F 値について相加平均を求めた。その結果を4名の被験者について相加平均をとったものを最終的な評価値とした

表1が実験結果である。それぞれ、

1. ベースライン：記事をランダムに選んだ場合の期待値
 2. R1：ユーザモデル l_u を単位ベクトルとした場合(話題性のみを考慮した場合)
 3. R2： F_t を単位ベクトルとし(話題性を考慮しない)、ユーザモデル l_u の拡張に WordNet を用いた場合
 4. R3： F_t を単位ベクトルとし(話題性を考慮しない)、ユーザモデル l_u の拡張に Wikipedia を用いた場合
 5. R4： F_t を単位ベクトルとし(話題性を考慮しない)、ユーザモデル l_u の拡張に Yahoo!検索エンジンを用いた場合
 6. R5：ユーザモデル l_u の拡張に WordNet を用いた場合
 7. R6：ユーザモデル l_u の拡張に Wikipedia を用いた場合
 8. R7：ユーザモデル l_u の拡張に Yahoo!検索エンジンを用いた場合
- 以上を表す。

結果から、ユーザモデル、話題性のモデルを考慮した場合に、ベースラインより良い結果が得られることがわかる。また、両方を考慮した場合にさらに良い結果が得られていることがわかる。このことから、提案手法により、ユーザのモデルおよび話題性のモデルを適切に取得できていることがわかる。

表 1: 実験結果

	評価値
ベースライン	0.631
R1	0.641
R2	0.671
R3	0.657
R4	0.665
R5	0.677
R6	0.680
R7	0.68

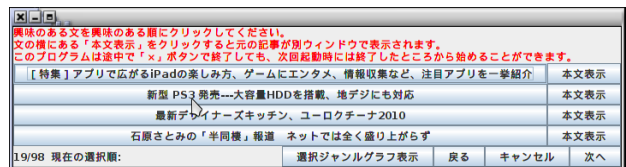


図 1: 画面インターフェース

5 おわりに

本稿では、会話における話題選択と情報推薦の類似性に注目することで、ユーザモデルならびに話題性のモデルから、現在のユーザの興味モデルを作成、それによる話題推薦の手法について提案し、ユーザ実験によりその有効性を示した。今後は、さらなる精度の向上と、より会話らしい文章生成ができるシステムの構築を進めていきたい。

参考文献

- 1) Shilin Ding, Gao Cong, Chin-yew Lin, Xiaoyan Zhu: Using Conditional Random Fields to Extract Contexts and Answers of Questions from Online Forums, ACL2008, pp.710-718 (2008).
- 2) Wen-Yun Yang, Yunbo Cao, Chin-Yew Lin: A Structural Support Vector Method for Extracting Contexts and Answers of Questions from Online Forums, EMNLP2009, Vol. 2, pp. 514-523 (2009).
- 3) Jiahui Liu, Peter Dolan, Elin Rønby Pedersen: Personalized News Recommendation Based on Click Behavior, IUI2010, pp.31-40 (2010).
- 4) Ai Yokoyama, Vitaly Klyuev: Search Engine Query Expansion Using Japanese WordNet, IEEE2010, pp.1-5 (2010).
- 5) Fei Wu, Daniel S. Weld: Open Information Extraction Using Wikipedia, ACL2010, pp. 118-127 (2010).