5B-6

情報推薦のための意外性判定方式の提案と評価

村岡 優輔　　楠村 幸貴　　水口 弘紀　　久寿居 大
日本電気株式会社　情報・メディアプロセッシング研究所

1　はじめに

単身世帯の増加，ひきこもり問題など，社会や人とのつながりが希薄化しつつある．皆が活き活きと暮らし，支えあえるコミュニティの形成が必要であり，形成のきっかけとなるコミュニケーションの支援が今後重要となる．コミュニケーション支援においては，会話が盛り上がるように，現在の話題に関連していて興味を引く新しい話題を提供することが重要である．

本研究では，現在の話題が何に関するものかが特定できたとして（「対象物」と呼ぶこととする），対象物に関連のあるもの（「関連物」と呼ぶ）を，どういう関連があるか（「関連文」と呼ぶ）と合わせて，新しい話題として提供する方式を提案する．対象物，関連物，関連文の情報ソースとしては「Wikipedia」*を利用した．

対象物に対して，Wikipedia の記事から，以下のように関連物，関連文を取得する．

- 対象物の記事から，別の記事タイトルの言葉を関連物，関連物を含む文を関連文として取得
- 対象物を含む記事から，記事タイトルを関連物，対象物を含む文を関連文として取得

例えば，現在の話題 (対象物) が「奈良」であるとする．「奈良」の記事中に「世帯あたりのピアノの所有台数が日本一」という文があれば，それを関連文とし，「ピアノ」を関連物とする．

この方法では，リンク関係にある記事全てが関連物となり，興味を引くとは限らない．そこで，以下のような関連文に注目した．

特徴的な関連文 例えば，対象物に対する「世界一の」のようなキーワードを含む関連文は，興味を引くと考えられる．このような，関連物によらず興味を引く関連文を「特徴的な関連文」と呼ぶ．

定番の関連文 例えば，対象物が場所であるとき，その場所の有名な出身者や，有名な特産物は興味を引くと考えられる．このような，対象物の種類によって言及されやすい関係を表す関連文を「定番の関連文」と呼ぶ．

関係が未知の関連文 例えば，対象物「バス停」に対して，「ニャロメ」という関連物は一見関係がなさそうである．そのため，「ニャロメのモデルは，赤塚不二夫がバス停で見た野良猫である」という関連文は興味を引く．このような，対象物と関係があることが未知である関連物の関連文を「関係が未知の関連文」と呼ぶ．

このうち関係が未知の関連文の判定は容易ではない．[野口 09] では，単語間の関係の既知/未知を Web 文書集合中の共起頻度により判定する方法が提案された．

対象物と関連物の関係の既知/未知判定に，Web 文書集合中での単語の共起頻度ではなく，Wikipedia での記事間のリンクの多さ (リンク頻度) を共起頻度とみなして直接用いることを考える．しかし，Wikipedia でのリンク頻度は単語の共起頻度に比べて少なく，[野口 09] の方法をそのまま用いることは難しい．そこで，データ量が小さい場合での関連の既知/未知判定方法を提案する．

2　提案手法

2.1　アプローチ

興味を引く，関連物と関連文を判定するために，「特徴的な関連文」，「定番の関連文」，「関係が未知の関連文」を判定する．

「定番の関連文」と「関係が未知の関連文」の場合，関連物を既知と判定したものを興味を引くと判定する．「未知の人物が対象物の出身者である」のような，未知の関連物についての「定番の関連文」は，興味を引かないからである．また，ユーザは対象物と未知の関連物の関係の有無を想定できない．未知の関連物についての「関係が未知の関連文」は，関係があることが想定外とならず，興味を引かないからである．

また，関係，関連物の既知/未知をユーザそれぞれに対し判定するのは困難である．本研究では，Wikipedia の読者間での知名度の高さの判定により代用した．

2.2　興味を引く話題推薦の判定方法

2.2.1　特徴的な関連文の判定方法

関連文が特定のキーワードを含む場合に，特徴的な関連文と判定する．キーワードとして例えば，「驚くべきことに」，「実は」など，筆者が驚いたことを示す副詞や，「世界一」，「由来」など，特徴的な関連文を表す文字列を用いる．

2.2.2　定番の関連文の判定方法

特徴的な関連文の判定と同様に，例えば，「出身」，「特産」などのよく言及される関係を表すキーワードを含む場合に，定番の関連文と判定する．

2.2.3　関係が未知の関連文の判定方法

Wikipedia でのリンク頻度のデータ量は小さい．そのため，記事間のリンク頻度そのものによる判定は難しい．そこで，Wikipedia で各記事に付けられているカテゴリにより，似た意味の記事をまとめ上げる．まとめ上げたカテゴリ間のリンク頻度により記事間の関係の既知/未知の判定を行う．

しかし，カテゴリに含まれる記事数には偏りがある．そのため偏りを補正した比較が必要である．カテゴリ間のリンク頻度が，他のカテゴリ間のリンク頻度と比

Proposal and evaluation of the method of detecting surprise for information recommendation
Yusuke MURAOKA　Yukitaka KUSUMURA　Hironori MIZUGUCHI　Dai KUSUI
NEC Information and Media Processing Laboratories

*http://ja.wikipedia.org/

表 1: 使用する記号

記事 w の属するカテゴリの集合	$C(w)$
記述回数評価に用いるカテゴリの集合	C_{ALL}
カテゴリ c に属する記事数	$n(c)$
c_1 の記事と c_2 の記事のうち，リンクが存在した組み合わせの数	$\mathrm{link}(c_1, c_2)$

較して少ないことを判定する指標を計算する．指標として，リンク頻度の確率分布のもとでの p 値を用いる．

計算方法を表 1 の記号を用いて説明する．w_1 のカテゴリ c_1 と w_2 のカテゴリ c_2 の記述回数の p 値を計算する．c_1 と他のカテゴリ c との記述回数が，サンプルサイズ $n(c_1)n(c)$，パラメータ θ_{c_1} の二項分布に従うと仮定する．パラメータ θ_{c_1} は以下の式で推定する．

$$\hat{\theta}_{c_1} = \frac{\sum_{c \in (C_{\mathrm{ALL}} \setminus c_2)} \mathrm{link}(c_1, c)}{n(c_1) \sum_{c \in (C_{\mathrm{ALL}} \setminus c_2)} n(c)} \quad (1)$$

$\mathrm{link}(c_1, c_2)$ を評価するために，推定した二項分布のもとでの p 値を求める．

$$p_{c_1} = \sum_{l=0}^{\mathrm{link}(c_1,c_2)} \left(\begin{array}{c} n(c_1)n(c_2) \\ l \end{array} \right) \hat{\theta}_{c_1}^{l} (1 - \hat{\theta}_{c_1})^{n(c_1)n(c_2)-l} \quad (2)$$

c_1 と c_2 の役割を入れ替えて上記の計算を行った結果を p_{c_2} とする．w_1, w_2 間の関係の既知/未知を表す指標は，以下で計算する．

$$p_{\mathrm{cc}} = 1 - \frac{1}{|c(w_1)||c(w_2)|} \sum_{c_1 \in c(w_1)} \sum_{c_2 \in c(w_2)} \frac{1}{2}(p_{c_1} + p_{c_2}) \quad (3)$$

2.2.4 関連物の既知/未知判定方法

Wikipedia の多くの文書で記述のある関連物は，被閲覧回数が多く読者にとっての知名度が高いと考えられる．Wikipedia での関連物の単語頻度がある閾値より高ければ既知と判定する．

3 実験

種類が多く，旅行などで話題になることも多いので，対象物としては地名や建物など場所に関する Wikipedia の記事を選んだ．複数の対象物に対し，関連物，関連文となりうる候補を 500 組取り出し，興味を引く関連物，関連文かを判定した．判定には，2 章で説明した方法を以下の条件のもとで用いた．

- 特徴的な関連文の判定のためのキーワードは 78 個
- 定番の関連文の判定のためのキーワードは 48 個
- 関係が未知の関連文は，式 (3) の指標を用いて上位 25 個を判定

表 2: 提案する方法の精度と再現率

	精　度 (%)	再現率 (%)	サンプル中の正解の割合 (%)
特徴的な関連文	83.3	10	8
定番の関連文	81.2	23.6	8.6
関係が未知の関連文	28	21	6.4

候補とした関連物と関連文の組全てに対して，2 名の評価者が正解を決定した．興味を引くものであり，かつ，関連文が特徴的な関連文，定番の関連文，関係が未知の関連文のいずれかであるものを正解とした．精度，再現率は，表 2 のようになった．判定できた正解例を，表 3 に示す．

4 考察

実験の結果，関係が未知の関連文の判定は難しく，これだけで十分な話題推薦が行えるとはいえない．一方，特徴的な関連文と，定番の関連文の判定精度は高い．しかし，これらはキーワードによる限られた種類の関係しか含まない．限られた種類の関係のみの話題推薦は，ユーザに飽きられてしまうため，話題推薦のためにはこれらを組み合わせて用いるのがよいと考える．

実際，会話を想像しても，関係が未知である意外性のある話題だけでなく，「ここには世界一の ~ がある」「ここの特産品は ~ である」のような定番な話題が含まれてよいし，また，多くの推薦された話題のうちいくつかで会話が盛り上がれば十分である．

例えば，特徴的な関連文，定番の関係文の判定結果から 2/5 ずつ，その他のものを 1/5 の割合で選び，推薦するシステムを考える．その他のものとして，ランダムな取得を考えると，推薦結果全体として精度は 67.1%である．関係が未知の関連文の判定結果を用いることで，精度が 71.8%に上昇するという効果がある．

5 まとめ

話題の推薦システムの実現のため，現在の話題に対して関係があり，興味を引くような関係である記事を判定する方法を提案した．記事間の関係が未知であり，興味を引くような関係かの判定については，それ単独で用いるにはまだ十分な精度は得られていない．しかし，関係の種類が限定されてしまうが精度の高い他の判定方法と組み合わせることで，精度高く，多様な種類の関係の話題の推薦が実現可能となる．

参考文献

[野口 09] 野口大輔：Web 上の HTML 文書を用いた意外性のある情報の獲得支援, 2009.

表 3: 判定できる正解例

	対象物	関連物	関係文 (抜粋)
特徴的な関連文	摩周湖	ジンクス	晴れた摩周湖を見ると出世できない，結婚できないというジンクスが語られることがある
定番の関連文	奈良市鼓阪小学校	明石家さんま	さんまの出身小学校である
関係が未知の関連文	渋谷駅	三島由紀夫	渋谷駅ホームから転落