E-035

性別・年代別の嗜好情報を基にした話題語提供システム
Offering topic word system based on preference information of gender and generation

南　光† Akira Minami　芋野 美紗子† Misako Imono　土屋 誠司‡ Seiji Tsuchiya　渡部 広一‡ Hirokazu Watabe

1. はじめに

現在，人間と円滑なコミュニケーションをとれるロボットが望まれている．最も一般的なコミュニケーション手段として会話が挙げられるが，人とロボットで円滑な会話を行うには相手の嗜好に合った適切な話題を提供する方が望ましい．そこで本稿では，各性別・年代の嗜好の傾向を利用して，ユーザに合わせた話題となる語を提供するシステムを提案する．具体的には，各性別・年代で興味のある単語を Web 上から取得し，嗜好の傾向を調べた上で，ユーザに適切な話題となる語を選出する．本稿では，Web 検索に使用する語は Web 上の膨大な情報量から自分に有用な情報を検索する特性上，ユーザにとって興味のある語だと考え，その語を嗜好の傾向の調査に使用する．また，概念ベースや関連度計算方式を使用した人間の連想を模したメカニズムによって，語と語の関連性を調べることで適切な語を選出，提供することを目指す．

2. 使用技術

2.1 概念ベースと関連度計算方式

概念ベース[1]は語（概念）の特徴を表す語（属性）を集めた知識ベースであり，属性には重要性を表す重みが定義されている．本稿では，複数の国語辞書や新聞などから抽出した概念や属性を加えた 87242 語の概念からなる概念ベースを用いる．例えば，概念「雪」は次のように定義する．

概念「雪」={(雪,0.61),(白い,0.30),…}

なお，本稿では概念ベースに登録されていない概念を未定義語と定義する．

関連度計算方式[2]とは，概念と概念の関連の強さを定量的に評価する手法である．各概念の重みを考慮した属性集合の一致度合いを計算することで関連度を算出する．

2.2 未定義語の属性獲得手法

未定義語の属性獲得手法[3]とは，未定義語 *X* の特徴を表す属性と重みの組を Web を用いて獲得する手法である．本稿では，この未定義語の属性獲得手法をオートフィードバック（*Auto Feedback : AF*）と呼ぶ．

2.3 Web-IDF

Web-IDF は Web 上にある文書のみを用いて索引語の特定性を考慮する手法である．Google が保有している日本語のページ数を *N*，索引語 *t* を Google で検索した際のヒット件数を *df(t)* とすると，*Web-IDF* は(2.2)式で定義される．

なお，Google が保有する日本語のページ数は公開されていないため，日本語の文書として最も使われている助詞「は」で検索を行ったヒット件数 7,660,000,000 (2012 年 6 月 4 日現在）を *N* としている．

$$idf(t) = \log_e \frac{N}{df(t)} + 1 \qquad (2.2)$$

3. 話題語提供システム

提案システムでは，Web 上から取得したキーワードを基に性別・年代別の嗜好情報を抽出する．次に Web 上から時事情報の見出し文を取得し，そこから固有名詞(以降，本稿では話題語と定義する)を抽出する．その話題語と性別・年代別の嗜好情報を照合することで，ユーザの嗜好に合った語を選出，提供する．

3.1 BIGLOBE サーチ旬感ランキング

旬感ランキング[4]とは，BIGLOBE サイト[5]が提供する検索エンジンによって検索された語を集計し，男女別で 10 代~50 代の急上昇ワード上位 20 位までをランキング形式にまとめたものである．本稿では，急上昇ワードそれぞれをその性別・年代の嗜好情報を示す語として取得する．

3.2 嗜好情報の抽出

過去 4 週間分の旬感ランキングのキーワードから，各性別・年代の嗜好情報を獲得する．まずキーワードに対して *AF* を行い，属性を取得する．同じ属性の重複回数が多いほどその属性を持つ語を重要視していると考えられるため，重複回数の多い属性 20 語を嗜好データと呼び，嗜好情報を形成する．図 1 に嗜好データの抽出の具体例を示す．

キーワード	属性
家政婦のミタ	ドラマ, 動画…
矢野未希子	芸能，動画…
秋元才加	芸能，アイドル…
前田敦子	うた，芸能…

↓

嗜好データ	重複回数
芸能	3
動画	2
アイドル	1
…	…

図 1　嗜好データの決定

取得した嗜好データに対して重複回数の多い順に重みを付与する．重みは 20 から順位を引いて 1 足した値とする．表 1 に重み付けの具体例を示す．

さらに，嗜好情報の特異性を求め，重みに補正を行う．ある嗜好データに対して全ての性別・年代での重複回数の順位を調べ，高い順に並べる．嗜好データである「芸能」に関して並び替えた例を表 2 に示す．並び替えた後，その順位に応じて各性別・年代の芸能の重みに補正をかける．

†同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

表 1　重複回数による重み付け

順位	嗜好データ	重み
1	芸能	20
2	動画	19
3	アイドル	18
…	…	…

表 2　「芸能」の順位が高い嗜好情報順

順位	性別・年代	「芸能」の順位
1	男性 40 代	3
2	女性 30 代	5
…	…	…

もとの重みを *W*，性別・年代で並び替えた場合の順位を *l*，補正結果を *A* とした場合，補正に使用する計算式は (3.1)で示される．

$$A = W * (2 - \frac{l}{10}) \qquad (3.1)$$

表 2 の例では，芸能の順位を見ると，男性 40 代の場合が最も高く，3 位であった．芸能の元の重みは 18 であるため，(3.1)式の *l* が 1，*W* が 18 となり，男性 40 代の芸能の重みは 34.2 となる．この結果から，男性 40 代の芸能は特に重要視されていることが表せる．これらを全ての性別・年代に関して行う．

以上の処理によって作成した嗜好データと補正後の重みの組を集めたものを，その性別・年代の嗜好情報とする．

3.3 話題語の抽出

本稿では人に提供する話題語として時事情報中の固有名詞を用いる．Web から獲得してきた 1 日分のニュース記事の見出し文に対して形態素解析を行い，時事情報の文中に含まれる固有名詞を抽出する．次に *Web-IDF* を用いてその固有名詞の重要性を調べ，閾値を用いて話題語の選別を行う．閾値 3.0 は既存研究[6]によって報告された値である．

3.4 照合

旬感ランキングから得られた嗜好情報を使用して，推薦すべき話題語を選別する．推薦すべき話題語とは，嗜好データそれぞれと関連のある語と考えられる．よって，関連度計算方式により話題語と嗜好データの関連性を調べ，推薦すべき話題語を選別する．

話題語と嗜好データそれぞれについて関連度計算を行い，その関連度と嗜好データの重みをかけ合わせたものを加算することでその話題語の重みを決定する．話題語への重み付けの具体例を図 2 に示す．

話題語	嗜好データ		関連度
	語	重み	
ナウシカ	動画	3	0.369
	株式会社	2	0.018
	…	…	…

➡ナウシカの重み=0.369*3+0.018*2+…

図 2　話題語への重み付け

図 2 では嗜好データとして重みがそれぞれ 3 の動画，2 の株式会社が存在する．関連度の算出を行った結果，話題語であるナウシカと動画の関連度が 0.369，株式会社との関連度が 0.018 となった．それぞれの重みと関連度を掛け合わせた値を足していき，結果をナウシカの重みとする．同様に 3.3 で選出された話題語全てに対して重み付けを行い，最終的に重みが高い上位 20 語をその性別・年代に推薦すべき語として出力する．

4. 評価

評価には 2011 年 12 月 17 日から過去 4 週間分の旬感ランキングから抽出した嗜好情報と 2011 年 12 月 18 日の時事情報から選出した話題語を用いた．各性別・年代の被験者に話題語 20 語を提示し，それぞれに対して興味を惹くか否かの判断を行ってもらった．興味を惹けば○，惹かなければ×と評価し，○の割合をその性別・年代の精度とした．各性別・年代の被験者数と評価結果をそれぞれ表 3，表 4 に示す．

表 3　性別・年代別の被験者数

	10 代	20 代	30 代	40 代	50 代
男性	5 人	10 人	2 人	4 人	4 人
女性	3 人	10 人	4 人	2 人	4 人

表 4　評価結果

	10 代	20 代	30 代	40 代	50 代
男性	36.7%	36.5%	21.2%	32.5%	26.3%
女性	24.0%	39.0%	47.5%	47.5%	15.0%

女性 30 代の嗜好情報にはドラマやテレビ，芸能といった語が存在した．結果，ドラマの女王や柄本明，西田敏行といったドラマのタイトルや芸能人の名前が多く出力され，評価が高くなった．対して女性 10 代の嗜好情報には歌やライブといった語が存在したが，DA PAMP といった歌手以外にもベルリン・フィルなどのオーケストラ名が多く出力された．その結果，評価が低くなってしまった．

5. おわりに

本論文では，Web から各性別・年代の嗜好情報を獲得し，時事情報に対するユーザの興味を満たす話題語を選出する手法を提案した．具体的には，時事情報中の単語と旬感ランキングから得られた嗜好情報の関連性を求め，話題となる語に重みを付与することで，各性別・年代の興味を惹くような話題語を出力する手法を提案した．評価結果より提案手法の精度向上が望まれる．そのためには，情報源の追加や，話題語の選出手法のさらなる厳密化が必要だと考えられる．

謝辞

本研究の一部は，科学研究費補助金(若手研究(B) 24700215)の補助を受けて行った．

参考文献

[1] 奥村紀之, 土屋誠司, 渡部広一, 河岡司, “概念間の関連度計算のための大規模概念ベースの構築”, 自然言語処理, Vol.14, No.5, pp.41-64, 2007.
[2] 渡部広一, 奥村紀之, 河岡司, “概念の意味属性と共起情報を用いた関連度計算方式”, 自然言語処理, Vol.13, No.1, pp.53-74, 2006.
[3] 辻泰希, 渡部広一, 河岡司, “www を用いた概念ベースにない新概念およびその属性獲得手法”, 人工知能学会全国大会, 2D1-01, 2003.
[4] Biglobe サーチ旬感ランキング, http://search.biglobe.ne.jp/ranking/, 2012/6/11 参照
[5] Biglobe, http://www.biglobe.ne.jp/, 2012/6/11 参照
[6] 藤田晴樹, 渡部広一, 河岡司, “コンピュータ日常会話のための Web からの時事情報獲得技術”, 情報処理学会研究報告, 2007-ICS-147(22), pp.145-150, 2007.