

ソーシャル情報を用いたコンテンツ間リンク付与に関する検討

宮崎 勝† 山田一郎† 住吉英樹† 宮崎太郎† 三浦菊佳† 松井 淳† 加藤直人†

†NHK 放送技術研究所

1 はじめに

VOD (Video on Demand) サービスなどの普及に伴い、ユーザの嗜好にあったコンテンツを推薦する手法への期待が高まっている。ユーザが視聴したコンテンツに関連した別のコンテンツを提示するサービスが、多くの動画サイトなどで実際に提供されている。コンテンツの内容情報を利用して推薦を行う手法は「内容ベースフィルタリング」と呼ばれ、デジタル放送の番組に付与された番組概要文を利用することで番組間の類似度を計算し、類似度が高いものを関連番組として提示する手法などが提案されている [1]。内容ベースフィルタリングは、関連コンテンツの提供という点では強力な推薦手法であるが、推薦されるのは類似コンテンツに限られるため、多様なバリエーションのコンテンツの推薦を行うことが難しい。一方、多くの e コマースの実運用サービスでは、ユーザの行動情報に基づく推薦手法である「協調フィルタリング」が利用されている。協調フィルタリングは、類似するユーザが購入した商品を推薦する手法であり、この処理により多様性のある推薦を実現している。

本稿では、ユーザ情報を用いた内容ベースフィルタリング手法を提案する。コンテンツの内容情報と、コンテンツに対するユーザの投稿文を処理対象とし、内容ベースフィルタリングだけでは獲得できない新たな情報をユーザの投稿文から抽出する。提案手法により、多様性のある内容ベースフィルタリングの推薦を実現することができる。

2 ユーザ投稿文によるコンテンツ情報付与

番組概要文などのメタデータに含まれない新しい情報でコンテンツ同士を結ぶことができれば、関連コンテンツ提示の際にも、新たなコンテンツ選択の視点をユーザに与えることができる。近年、コンテンツのメタデータを、コンテンツプロバイダーのみが付与するのではなく、ユーザも含めた第三者が付与する「ソーシャルタギング」という形態が一般的になりつつある。英

BBC が提供する World Service Radio Archive[2] では、アーカイブされたラジオ番組に対してタグ情報 (キーワード) を自動付与するとともに、ユーザが自由にタグ情報を付与することができるようになっており、付与されたタグはユーザ投票によって評価される。このタグを用いることで、番組コンテンツの検索が容易に行えるようになっている。ユーザによりタグが評価されることから、関連性の高いタグを精度よく付与できるといふ反面、意外な情報を記述したタグは淘汰されて残らない傾向があると考えられる。

一方、多くのコンテンツ提供サービスでは、コンテンツに対するレビュー投稿が可能となっている。このような投稿文には、感想などとともに、ユーザが持つ新たな関連知識が含まれることがある。たとえば、ある番組コンテンツに関して、「このキャラクターの声優は 」「このドラマのロケ地は である」といったように、番組内容情報には含まれない固有名詞をユーザが提供する場合などが考えられる。このような関連情報を自動で抽出することにより、コンテンツ同士を意外な情報をラベルとしたリンクで結ぶことができる。筆者らが提案する、ユーザ投稿文を利用した番組間リンク付与の概要を図 1 に示す。

本手法ではまず、番組投稿文および番組概要文を形態素解析し、「番組投稿文に含まれ、番組概要文に含まれない単語」(以降、投稿限定単語と記述する) を抽出する。得られた投稿限定単語を共通して持つレビューを抽出し、そのレビューの対象番組同士を投稿限定単語をラベルとしたリンクでつなぐ。次章以降では、実際の番組投稿データを対象とした投稿限定単語の特徴分析と、投稿限定単語を用いたリンク付与実験について

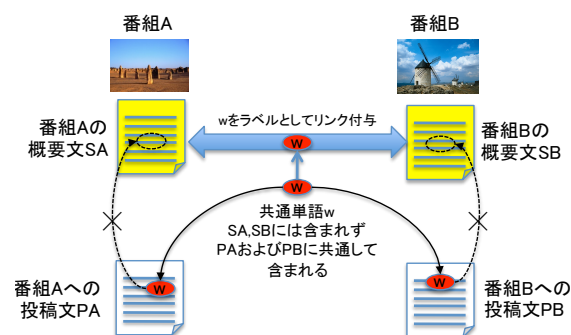


図 1: ユーザ投稿文を利用した番組間リンク付与

A study on content linkage technique using social information
 †Masaru MIYAZAKI †Ichiro YAMADA †Hideki SUMIYOSHI †Taro MIYAZAKI †Kikuka MIURA †Atsushi MATSUI †Naoto KATO
 †Science & Technology Research Laboratories, Japan Broadcasting Corporation(NHK)

て述べる。

3 番組投稿文の分析

NHKのWebサービス「NHK TV60 特選コレクション」ではアーカイブ番組の情報検索や一部番組の視聴、コメントの投稿機能を提供している。筆者らは、このサイトの番組概要文および投稿文(2013年1月から2013年10月の間に投稿された479番組に対する873個のレビュー投稿)を調査対象とした。まず、各投稿文に対し、その投稿に「番組概要文だけを見てもわからない新しい情報」が含まれているかどうかを手作業でアノテーションし、50%の投稿に概要文にはない情報が含まれていることを確認した。更に、各投稿に対し、形態素解析エンジン MeCab* を用いて投稿文の長さおよび投稿限定単語数(名詞)を計算し、検定を行うことによって、「概要文だけを見てもわからない新しい情報」を含む投稿は、「投稿長」および「投稿限定単語数(名詞)」が多い傾向があることを検定により明らかにした。

投稿限定単語から、コンテンツを結ぶ情報に相応しいと考えられる4種類の名詞を抽出した、名詞の種別を表1左に、その中の固有名詞の詳細種別を表1右に示す、

表1: 投稿限定単語数(名詞)の種別

名詞種別		固有名詞種別	
一般	2233 (61.1%)	人名	385 (72.6%)
サ変接続	610 (16.7%)	地域	86 (16.2%)
固有名詞	530 (14.5%)	組織	27 (5.1%)
形容動詞語幹	282 (7.7%)	その他	32 (6.0%)

4 番組間リンクの付与

得られた投稿限定単語を利用し、異なる番組間にリンクを付与する検討を行った。別番組のレビュー間に共通する投稿限定単語(固有名詞)があった場合に、それらの間に該当する投稿限定単語をラベルとしたリンクを付与する処理を行ったところ、そのうち780個のレビュー対に固有名詞をラベルとしたリンクを付与することができた。表2に、リンクがはられた番組対の例を示す。この例では共通する投稿限定単語である「坂本九」という人物名をリンクのラベルとして2つの番組を結ぶことができる。

一方で、番組レビューでは当該番組ではない他番組に言及することがあり(表3)、この情報が原因で誤ったリンクを付与してしまう例が多数見られた、このような場合、当該番組に対するコメントか否かを識別する処理が必要と考えられる。

表2: 投稿限定単語によってリンクが付与された番組例

番組名	レビュー投稿文
連続人形劇 新八犬伝	ナレーションの軽妙さ、人形の造形からくる独特の世界観に魅了されました。坂本九さんによるエンディング・テーマは当時の愛唱歌でした。
連続テレビ小説 梅ちゃん先生	最終回で高橋克実さんが、NHKのど自慢に出演して坂本九さんの「上を向いて歩こう」を歌うシーンがあまりにも印象深かったです。しかも合格。こういう前向きになれる最終回はとてもいいですね。

表3: 他番組への言及が含まれた例

番組名	レビュー投稿文(一部抜粋)
ドラマスペシャル 炎熱商人	「ドラマスペシャル」では「日本の面影」が好きだった。山田太一さんの本で、小泉八雲がジョージ・チャキリスで、セツさんを <u>檀ふみ</u> さんが演じてた。

また、同じ姓あるいは名の別人に対して誤ってリンクを付与してしまった例なども見られた、適合率向上のためには、同一人物を特定する名寄せ技術なども必要と考えられる、

5 まとめ

投稿テキストに含まれて、内容情報に含まれない単語を抽出することにより、コンテンツ同士をユーザが想定できないリンク情報で結びつけることができる可能性を示した。また、リンク付与の際には、固有名詞の正確な同定や、他コンテンツへの言及部分を特定する必要がある、といった課題も明らかになった。今後は、大規模な番組レビューデータを処理対象とするとともに、適合率向上に向けて、投稿長や投稿限定単語数、単語重要度などの特徴を考慮した処理を取り入れる、また、「有益な情報を含む投稿文」や「情報を含んだ投稿を行う傾向があるユーザ」を自動抽出する機械学習手法を検討し、多様性のある番組推薦システムの構築を目指していく、

参考文献

- [1] J. Goto, H. Sumiyoshi, M. Miyazaki, H. Tanaka, M. Shibata, and A. Aizawa: Relevant TV Program Retrieval using Broadcast Summaries, Proceedings of ACM on Intelligent User Interfaces(IUI), pp.411-412, 2010
- [2] Yves Raimond and TristanFeune, "The BBC World Service Archive Prototype," Sementic Web Challenge, The 12th International Semantic Web Conference (ISWC2013)

*<http://mecab.googlecode.com/svn/trunk/mecab/>