

マイクロブログを用いた番組推薦に関する研究

袴田裕紀[†] 岡誠[‡] 森博彦[‡]

東京都市大学大学院工学研究科システム情報工学[†]

東京都市大学知識工学部経営システム工学科[‡]

1. 背景

近年、テレビのデジタル放送化に伴い、ユーザが視聴できるチャンネルは膨大になった。ユーザが番組選択する際の負担は増加する傾向にあり、視聴者は自分がどの番組を見るべきなのか迷ってしまうことも少なくない。ユーザの好みに合った番組を推薦するサービスが求められている。

一方で、twitterをはじめマイクロブログが急速に普及している。マイクロブログでは、ユーザは容易に投稿することができるため、リアルタイム性の高い情報が即座に得られる情報源として注目されている。マイクロブログに投稿される情報量は膨大であり、その中から必要な情報だけ抽出することは困難である。

マイクロブログの使用方法として、テレビ番組を視聴しながらtwitterへ番組に対する投稿が盛んに行われている。テレビ視聴時に、見ている番組に関するつぶやきや書き込みをtwitterでする人は43.4%となっている。さらに、他のSNSよりもtwitterがよく使われるという調査報告[1]もある。このような、放送時間中のツイートにはその番組の内容に関する言語的表現がリアルタイムで取得できる。これらの情報を利用することで、ユーザに対してリアルタイムな番組情報を推薦できると考える。

2. 関連研究

井川ら[2]はテレビ番組のEPGと視聴履歴を用いた番組推薦システムを構築した。EPGからキーワードと出演者を抽出し、視聴履歴からキーワード選好度と出演者選好度を決定要素として番組を推薦した。キーワードと出演者を使った推薦方法は有効であることがわかった。さらにチャンネルと時間帯にも有効であるとわかったが、番組

の進行状況を考慮していない。

澤井ら[3]は、SNSを用いた協調フィルタリングによる番組推薦手法の提案をした。代表的なユーザが好む番組、共感するユーザが視聴している番組、流行の番組を推薦した。これにより、行動履歴を取得できないユーザ・あまり書き込みをしないユーザに対して推薦可能となった。しかし、番組推薦のために、ハッシュタグを利用することでタイトルを推定するため、番組内容の考慮はしていない。

袴田[4]は現在放送されている番組内容を反映した支援をする必要があると考えた。番組状況の把握として、放送中の状況を把握するためにtwitterを用いた。この研究では、実際に被験者を設けて行うのではなく、システムとして推薦が可能かの検証を行った。システムの流れとしては、番組ツイート群の中には番組の境目が存在すると考えられるため、3ツイートごと類似度を計っていくことで、類似度が急激に変化した場所を番組内容が変化したと判断し、最新のツイートを使用する。視聴ツイート群と放送ツイート群をmecabによって形態素解析にかけ、「名詞：一般」「名詞：固有名詞」を抽出する。ツイートに出現する名詞をWikipediaのCategoryを利用することで、単語の上位概念を取得することで、単語の意味の拡張をする。拡張された語と形態素をTF-IDF法から単語の特徴量を算出し、cos類似度により類似度を求める。システムは複数の番組間の類似性を評価し、類似する別の番組を推薦する。評価実験の結果としては、適合率0.50、再現率0.41、F値0.45となった。現状の結果では被験者を設けて行うには困難である。

3. 目的

本研究は、マイクロブログの一つであるtwitterを用いて番組状況を考慮した番組推薦を行う。ユーザが視聴している番組ツイートを使用することで、そのほかの放送中の番組ツイートと近い内容の番組を推薦することを目的とし、システムとして可能かの検証をする。

Studies on program recommendation using Microblogs

[†]Yuuki Hakamata, System Information Engineering, Tokyo City University

[‡]Makoto Oka, Hirohiko Mori, Department of Industrial and Management System Engineering, Tokyo City University

4. システム概要

本研究では番組状況の把握として、EPG の情報は番組全体の情報を示すもので、EPG の情報だけでは放送中の状況を把握することは難しいと考え、番組に対するツイートを利用する。ユーザが視聴している番組とその他の放送中の番組、それぞれのツイートを取得する。視聴している番組ツイートについては、興味ありと判断するために、関連研究[4]より、20秒以上視聴した番組は興味があると判断する。視聴時間から興味があると判断した番組は、実際に視聴していた時間帯の情報を利用する。視聴している番組ツイートと裏で放送している番組ツイートを比較することで近い番組を推薦する。使用するツイートはリアルタイム検索「テレビなう」から番組ごとに取得する。「RT」「URL」のあるツイート文は取り除いた。本研究では、実際に被験者を設けて行うのではなく、システムとして推薦が可能かの検証を行う。

5. 分析

旧システムでは、名詞を分割しすぎたために、意味の拡張が十分に機能しなかった。そこで、ツイート文から正しい形態素を作成し、システムで拡張した単語と人手で作成した正解データを比較することで、意味の拡張の分析をする。

方法としては、旧システムより Wikipedia の Category を使うためには正式な単語でなければシステムで反映されなかった。ただし、ツイート文には mecab の辞書で処理できない単語が含まれていた。人名といった固有名詞がわかれてしまう場合がある。固有名詞は番組の特徴を表す重要なワードだと考えたため、正式なものに変換し、意味の拡張を行う必要がある。そこで、形態素の正解データを作成した。作成方法としては、mecab の辞書に「Wikipedia」と「はてなキーワード」のタイトルを固有名詞として追加する。さらに、わかれてしまう語については、取得したツイートと比較し、連続した名詞は形態素解析後に人手で結合させる。加えて、意味の拡張の正解データを作成した。分析方法は、形態素の正解データを使用した意味の拡張データと意味の拡張の正解データを比較する。使用データは60番組、うち20番組ずつ「バラエティ」「ニュース」「情報・ワイド」。1つのシステムの拡張に対して、1以外の拡張から一致するキーワードと1つの正解データに対して、1以外の正解データから一致するキーワード、これらを使って12分割交差検定を行う。システムと正解データを比較し、適合率、再現率、F値を求めて評価した。

5. 1. 結果

分析の結果を表1に示す。

表1 分析の結果

	適合率	再現率	F値
全体	0.07	0.39	0.12
バラエティ	0.06	0.32	0.10
ニュース	0.07	0.42	0.12
情報・ワイド	0.08	0.45	0.13

結果より、現状の Wikipedia の Category を利用した方法では、意味の拡張をすることで不必要な語を多く抽出されてしまっている。これらの結果と関連研究[4]の類似度の値を含め、番組ごとの区別ができていないと考えられる。推薦対象となる番組もそうでない番組も類似度が変わらないために関連研究[4]の結果になったと考えられる。

6. おわりに

今回の分析では、すべての単語に対して、カテゴリズを行ったが、現状の拡張方法では不必要な単語を多く抽出した。今後は、番組内容の区別する方法として、クラス分類する必要があると考えた。分類としては、テレビ番組では共通して登場人物が存在する。人物は番組状況の手がかりとなるため、登場人物を「音楽家」「スポーツ」「政治家」「俳優」「お笑い芸人」「タレント」の6つに分類する。従来は意味の拡張を元の単語に対して、二段階で拡張を行った。クラス分類の方法としては、6つのクラスが出るまで拡張し、分類をする。そしてシステムと正解データを比較し検証する。

参考文献

- [1] ネットエイジア株式会社/株式会社シオシオンジャパン: “テレビとソーシャルメディアの関係性” (2013)
- [2] 井川一樹, 福原知宏, 藤井秀樹, 武田英明: ” テレビ番組の視聴履歴と電子番組表を用いた番組推薦システムの構築と評価”, 人工知能学会全国大会論文集 24th, 3C4-3, 2010
- [3] 澤井里枝, 有安香子, 藤沢寛, 金次保明: “SNS を利用した協調フィルタリングによる番組情報推薦手法”, 情報処理学会報告, Vol. 2010-DBS-151, No. 43
- [4] 袴田裕紀: ” 非計画視聴時におけるリアルタイムな番組情報の提供に関する研究”, 卒業研究 (2016)