マイクロブログを対象とした話題チャンクの抽出法に関する研究

新谷　歩生

近年, Twitter に代表されるマイクロブログが注目を浴びている．マイクロブログは，文字数制限が設定されているため，まとまった長い文章を記述することはできないが，今していることや，考えていることを，即座にタイムライン上に短い言葉で手軽に書き込めるという特徴がある．また，ユーザは関心を持った他ユーザを選択し，彼らの投稿記事を自身のタイムラインに取り込んで閲覧することができる．タイムラインを用いて複数ユーザの投稿記事を時系列順に表示することで，ユーザ間でリアルタイムに知識や体験を共有することができる．

しかし, 簡潔で思いつきの短文が次々と投稿される状況では, 一つの話題が複数の記事に分割され，散発的に投稿されることも少なくない．このような場合，話題を十分に理解するには，関連する前後の記事を閲覧しなければならない．そのため，ユーザが話題の理解を容易にするために，同一話題の複数記事を集約して提示する手法が必要とされてきた．また，マイクロブログの記事を検索するサービスも活用されているが，検索条件に合致した記事だけでなく，その記事と関連する記事を出力したいという要求も高まっている．

本論文では, マイクロブログからある話題について記述された複数の記事の塊を話題チャンクとして抽出する手法を提案する．提案手法の第 1 の特徴は，連続して投稿される前後の記事間に共起する単語の有無で，話題の連続性をを判断することである．しかし，文字数制限のあるマイクロブログでは，一つの文章を前後に分割して投稿するなどの場合があり，単語の共起判定だけでは話題チャンクを抽出することは困難である．そこで，提案手法の第 2 の特徴として，マイクロブログの即時的な投稿スタイルに着目し，ユーザが短期間に連続して投稿する記事は同一話題である可能性が高いと判断する，時間間隔に基づく抽出手法を提案する．

Twitter から異なる特性を持つユーザ 3 名を選定し，提案手法による話題チャンクの抽出実験を行った．*Jaccard* 係数を用いて算出する抽出精度の評価手法を考案し，評価した結果，共起語による抽出手法で約 40 %の精度を示したのに対して，時間間隔に基づく抽出手法は，2 名のユーザが共起語に基づく抽出精度を上回る結果を示した．また，いずれのユーザにおいても，平均投稿間隔の 0.1 倍弱の時間間隔で抽出したときに，最も精度が高くなることを明らかにした．以上のことから，共起語による抽出を手法に時間間隔に基づく抽出手法を適用することで，抽出精度を向上できるとの知見が得られた．

今後の課題は，抽出結果を，抽出精度を向上できている場合と低下させている場合に分けて詳細に分析し，更なる精度の向上を目指していくことなどがある．

（指導教員　佐藤 哲司）