平成23年度 情報工学コース卒業研究報告要旨

渡邉 研究室	氏　　名	田 渕　義 宗
卒 業 研 究 題 目	話題の時間的遷移による関連記事の推定	

近年，コンピュータやインターネットの普及により手軽にオンラインニュースを得ることが可能となった．イベントや事件の経緯を知りたい利用者はこのニュース記事を話題ごとに整理して必要な情報を見つけなければならない．しかし，ニュース記事の数は膨大であるため人手で整理することは困難である．本研究では語の組合せで話題を表し，話題の時間的変移により関連記事を推定する手法を提案する．

本研究では話題を，ある対象に関する特定の事象とする．話題を表すために名詞の 2 語の組合せと時間を使用する．同じ語の組合せで表現される事象が複数の事象を指す場合があるため，時間を用いて区別する．また，ニュース記事では重要語であっても記事中に 1 度しか出現しないことが多いため，ニュース記事の重要語を判断するのに語の出現回数を用いることは適さない．一方，ニュース記事のタイトルには記事の内容を端的に表すために重要語が記述されていることが多い．そのため，ニュース記事がどのような話題に対して記述しているかを判断するのにタイトルの名詞を用いる．

提案手法では，ユーザが少なくとも一方が固有名詞である名詞 2 語を入力する．システムは入力された 2 語で表現される話題が含まれる記事を抽出し，関連話題を抽出する．抽出された関連話題から関連話題の期間を判定し，関連話題の補間を行う．判定した期間のうちユーザが入力した名詞 2 語で表現される話題の期間を出力する．提示された期間のうちユーザが日付を選択する．ユーザが選択した日付に含まれる関連話題の記事候補を抽出する．抽出された記事候補から適切な記事を抽出し，時間を利用しまとまりごとに分割する．分割された記事集合のうち，選択された日付が含まれるまとまりの記事を関連話題の記事として抽出する．関連話題の記事から特徴ベクトルを作成し，関連話題のコサイン類似度を計算する．コサイン類似度が閾値 θ_{th} 以上であり，一番類似している関連話題同士を併合する．この処理を値 θ_{th} 以上のものが存在しなくなるまで行い，クラスタリングした関連話題をユーザに提示する．ユーザがその中から関連話題をひとつ選択すると，システムはその関連話題の記事を出力する．

評価実験では，提案手法を評価するため比較手法との比較を行った．ここで比較手法として，tf-idf を用いた特徴ベクトルで記事を表現し，コサイン類似度によりクラスタリングする手法を用いた．その結果，提案手法は比較手法より関連記事を多く抽出可能であることが明らかになった．