6ZA-1

ニュース記事の自動分類による学習教材作成支援に関する研究

中村健二† 田中成典‡ 池辺正典† 吉村智史† 熊本真也‡
関西大学大学院総合情報学研究科† 関西大学総合情報学部‡

１．はじめに

近年，総合学習の時間[1]の実施に伴い，新聞教材を用いた学習（NIE：Newspaper in Education）[2]が注目されている．総合学習の時間では，情報や環境など複数の科目にまたがる横断的な学習活動を行う．しかし，NIE による教育活動の現場では，教師が大量の記事を閲覧し，手動で教材を作成している．そのため，教材作成は教師にとって大きな負担となっている．大量の新聞記事を活用するための既研究としては，新聞記事間の関係を出現単語と時系列情報に基づいて抽出する研究[3]や新聞記事を記事のテーマに基づいて効率的に検索する研究[4]がある．しかし，前者は記事の関係抽出，後者は記事の検索を主目的としており，記事が教育に利用できるかの判断や教材利用に適した記事の結び付けを行っていない．そのため，抽出した記事群は，記事の読み比べ[5]を行うための分類がされておらず，NIE に使用する新聞教材に適していない．そこで，本研究では，入力記事と類似したニュース記事を抽出し，類似記事から複数の話題を抽出することにより，NIE に利用可能な教材の作成を支援することを目指す．

２．システムの概要

本研究では，入力記事に類似する記事を抽出し，教材としての利用が容易になるように話題毎にまとめて導き出すことを目的とする．本システムは，図１に示すようにニュース記事を入力とし，話題別記事群を出力する．本システムは，１）カテゴリ特定機能，２）類似記事の抽出機能，３）話題の分類機能の３つの機能により構成される．

２．１ カテゴリ特定機能

本機能では，ベクトル空間モデルを使用して入力記事のカテゴリを特定する．ベクトル空間モデルとは，文書を単語の出現頻度に基づいたベクトルとして表現し，ベクトル間の内積から類似度を算出する手法である．入力記事の単語の出現頻度を基に作成した記事ベクトルと各カテゴリにおける記事群の単語の出現頻度から作成したカテゴリベクトルの類似度を算出し，入力記事を最も類似度の高いカテゴリに分類する．

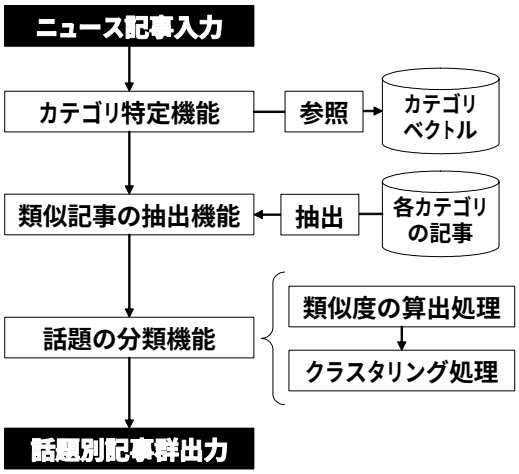

図１ システムの流れ

２．２ 類似記事の抽出機能

本機能では，入力記事のタイトルから名詞を抽出する．そして，入力記事と同一カテゴリの記事群から，入力記事のタイトルから抽出した名詞をタイトルに含む記事を抽出し，類似記事群として出力する．

２．３ 話題の分類機能

本機能では，入力記事と類似記事の類似度から類似記事群のクラスタリングを行う．本機能は，類似度の算出処理とクラスタリング処理の２つの処理により構成される．

２．３．１ 類似度の算出処理

本処理では，記事本文に出現する名詞を基に，入力記事の記事ベクトルと類似記事の記事ベクトルを作成し，ベクトル空間モデルを使用して記事間の類似度を算出する．

２．３．２ クラスタリング処理

本処理では，入力記事と類似記事の類似度を基にクラスタリングを行い，類似記事を話題毎に分類する．クラスタリングには，クラスタ数が未知の場合においても適用可能な階層的クラスタリングの最長距離法を使用する．

Research on Classification of News Articles for Making Prepare Teaching Materials
†Kenji Nakamura, Masanori Ikebe, Satoshi Yoshimura
Graduate School of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan
‡Shigenori Tanaka, Shinya Kumamoto
Faculty of Informatics, Kansai University, 2-1-1 Ryouzenji-cho Takatsuki-shi, Osaka 569-1095, Japan

３．システムの実証実験と考察

本システムの有用性を検証するために，記事のカテゴリ特定精度と話題分類の評価に関する実験を行った．

３．１　実証実験

記事のカテゴリ特定の実験では，ベクトル空間モデルを使用して，ニュース記事のカテゴリ分類精度について評価を行った．ベクトル空間モデルは，ベクトル作成に使用する品詞の種類によって精度が変化するため，本実験では，ベクトルの作成に使用する品詞を変更して，精度の比較を行った．分類カテゴリは，総合的な学習の時間で行う学習活動の具体例として挙げられている分野を基に，スポーツ，経済，教育，環境，国際と社会保障の６カテゴリを採用した．精度の評価には，カテゴリが特定されている記事を入力記事とし，適合率，再現率と F 値を使用した．F 値とは，カテゴリ分類手法の精度判定に用いられる尺度で，適合率と再現率の調和平均である．

また，記事の話題分類の実験では，類似記事の抽出とクラスタリングを行い，各クラスタの記事の内容について評価を行った．話題分類の実証実験では，国際カテゴリに分類されている京都議定書に関する記事を入力記事とした．入力記事から抽出した単語を表１示す．分類された話題の内容についての評価は，各クラスタに分類された記事の本文に含まれている単語の出現頻度上位単語の比較と目視による記事の確認によって行った．

３．２　結果と考察

カテゴリ分類の実証実験結果を表２に示す．表２から，全名詞を使用して作成したベクトルがニュース記事のカテゴリ分類には最適であることがわかった．ここでの全名詞とは，一般名詞，固有名詞とサ変名詞を指す．入力記事に対して類似記事の抽出を行った結果，69 件の記事を抽出した．抽出した記事を人手で確認したところ，半数以上の記事は入力記事と関連のある内容であった．それ以外の記事は，京都で開催されたイベントや国際会議に関する記事であった．話題分類の実証実験では，抽出した 69 件の類似記事に対して入力記事との類似度を算出し，クラスタリングを行った．クラスタリングを行った結果，６つのクラスタに分類した．各クラスタの出現上位単語から，複数のクラスタに共通して出現した共通語と単一のクラスタにしか出現しなかった特徴語を表３に示す．表３のクラスタの共通語から，入力記事では出現していなかった重要単語が新たに抽出できていることがわかった．また，表３のクラスタの特徴語から，各クラスタの特徴語が異なっていることがわかった．これらの結果から，クラスタによって話題が分類されていると考える．

表１　入力記事の出現頻度上位単語

	出現頻度の上位単語
入力記事	先進，森林，議定，排出，削減，ルール

表２　カテゴリ分類の実証実験結果

品詞	適合率	再現率	F値
全名詞	0.8092	0.7601	0.7765
名詞と自立動詞	0.8072	0.7538	0.7708
一般名詞のみ	0.7889	0.6930	0.7292
固有名詞のみ	0.6762	0.4205	0.4876

表３　各クラスタの出現上位単語

	クラスタの共通語	クラスタの特徴語
クラスタ１	国際，世界，議論	政府，検討，提案
クラスタ２	国際，世界，議論	各国，対策，生活
クラスタ３	国際，研究，対策	影響，負担，CO2
クラスタ４	国際，研究，対策	経済，科学，会合
クラスタ５	国際，世界，参加	首脳，規制，誘致
クラスタ６	国際，世界，参加	宗教，指導，停戦

４．おわりに

本研究では，入力記事と類似した記事を抽出し，話題の違いによる分類を実現した．これにより，新聞教材を用いた学習に必要な記事の抽出を可能にし，NIE で利用可能な教材の作成支援ができたと考える．今後は，より利用し易い学習教材の作成支援のために，記事の難易度算出，時系列による話題の移り変わりや学習者にあわせた教材作成を支援するなどの発展研究を行う予定である．

参考文献

[1] 文部科学省：平成 14 年度文部科学白書，国立印刷局，2003.5.

[2] 朝日新聞 NIE 委員会：総合的な学習に NIE を，朝日新聞社，1999.1.

[3] Rafael Llavori，Maria Cabo，Fernando Miralles：Discovering Temporal Relationships in Database of Newspapers，Lecture Notes in Artificial Intelligence，Springer，Vol.1416，No.2，pp.36-45，1998.7.

[4] Maria Nuno，Silvia Mario：Theme-based Retrieval of Web News，Lecture Notes in Computer Science，Springer，Vol.1997，pp.26-37，2001.5.

[5] 影山清四郎：学びを開く NIE-新聞を使ってどう教えるか，春風社，2006.7.