E-007

ニュースを関連記事毎にダイナミックに分類収集し閲覧可能とするシステムの開発
Development of Dynamical News Crawling and Classification System

宮下 学† 納富 一宏† 速水 治夫†
Gaku Miyashita Notomi Kazuhiro Haruo Hayami

1. はじめに

近年，オンラインニュースが情報収集の手段として主要となっており，日々多くの記事が掲載されている．掲載されているニュース記事には，発信側の意図が含まれている場合や誤った情報である場合がある．読者が公平性を保った考えや正確な情報を得るためには，複数の情報源からニュース記事を取得し，同一出来事のニュース記事を比較しながら閲覧することが求められる．

複数の情報源からニュース記事を取得し，同一出来事のニュース記事を収集することは，日々の記事掲載数が多いことから大変困難である．また，同一出来事の記事を比較しながら閲覧することは，多くの記事を閲覧しなければならないため手間となる．

日々のオンラインニュースを確認する手段として，グノシー[1]や SmartNews[2] といったニュースアプリがある．これらのニュースアプリには，最新のニュースを通知する機能があり，利用者は日々報道されるニュースを知ることができる．しかし，ニュース記事はカテゴリ毎にまとめられているが出来事毎にまとめられていないため，同一出来事の記事を収集することは難しい．また，同一出来事の記事を比較できるような機能はない．

本稿では，複数の情報源からニュース記事を取得し，同一出来事の記事を収集する際にかかる手間の削減を目的とする．解決策として，複数の情報源のニュース記事を確認でき，同一出来事でニュース記事を分類収集し閲覧を可能とするシステムを開発した．システムの機能とその評価について報告する．

2. 試作システム

解決策を確認するため，試作システムを開発した．試作システムでは複数の情報源を産経ニュース [3] と朝日新聞デジタル [4] とした．ニュース記事のデータは産経ニュースと朝日新聞デジタルが公開する XML サイトマップを基に，クローラーにより取得し，データベースに格納する．試作システムは日々のニュースを確認するために記事の一覧表示や記事を閲覧できる機能を持つ．関連記事の分類・提示は，同一出来事の記事かどうか判定する処理と再帰処理による分類を行いユーザに提示する．同一出来事かどうかの判定には，機械学習アルゴリズムであるサポートベクターマシン（SVM）を用いた分類器を使用する．分類器には，2 つの記事タイトルを文字列レベルで比較し，一致する要素数と要素の総数を与え，似ているか似ていないかの 2 クラス分類を行っている．

試作システムは，日々報道されるニュースの確認を支援するものとして，記事のタグ検索機能や記事の閲覧履歴を提示する機能を持つ．

試作システムは Web アプリケーションとして開発し，主に使用しているプログラミング言語は Python である．

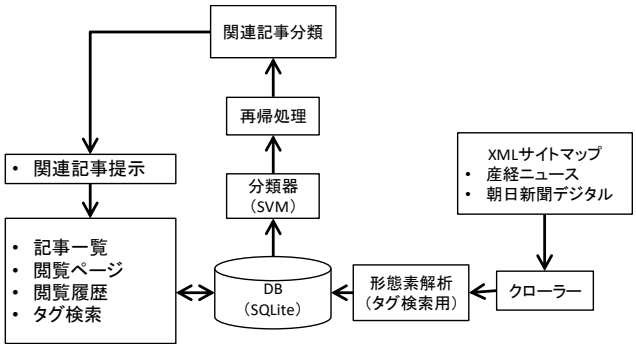

図 1　試作システムの概要図

試作システムの概要図を図 1 に示す．

2.1 関連記事の分類・提示

試作システムは，ニュース記事を出来事毎に関連記事として分類・提示する機能を持つ．以下に詳細を示す．

2.1.1 使用方法の種類

関連記事の分類・提示機能には，以下の 2 種類の使用方法がある．

(1)　記事一覧画面から任意の記事を選択して使用する

記事一覧画面でニュース記事を選択し，選択された記事の関連記事を収集・分類し提示する．記事は複数選択が可能であり，その場合は選択された記事それぞれの関連記事を収集し，提示する．

(2)　記事を選択しないで使用する

記事を選択しないで関連記事の分類・提示機能を使用する場合は 24 時間分のニュース記事を出来事毎に分類し提示する．ニュース記事は関連記事が存在する関連記事グループと，関連記事が存在しない単一記事グループにわかれて提示される．

2.1.2 関連記事分類の処理

関連記事として出来事毎に記事を分類する処理は，以下の 3 つの処理で構成されている．

(1)　24 時間分の記事をデータベースから取得

分類する対象の記事をデータベースから取得する．24 時間分の範囲の取り方は，関連記事分類・提示の使い方によって異なる．記事一覧画面から記事を 1 つ選択した場合は，ニュース記事の更新日時を基準とし，基準より前後 12 時間を範囲とする．記事一覧画面から記事を複数選択した場合には，選択された記事の更新日の 0 時を基準とし，23 時 59 分 59 秒までを範囲とする．記事を選択しないで使用する場合は，現在日時から過去 24 時間を範囲とする．

†神奈川工科大学，Kanagawa Institute of Technology

1 グノシー，https://gunosy.com/
2 SmartNews，https://www.smartnews.com/ja/
3 産経ニュース，http://www.sankei.com/
4 朝日新聞デジタル，https://www.asahi.com/

(2)　同一出来事の記事かどうかの判定

同一出来事の記事かどうかの判定はニュース記事タイトルを文字列レベルで比較し，テキスト間の類似度を用いて行う．テキスト間の類似度を算出する方法は，レーベンシュタイン距離の算出や N-gram などのアルゴリズムを使用することが知られているが，今回は比較的使用が容易な Python の標準ライブラリである SequenceMatcher を用いる．SequenceMatcher は，ratio メソッドによりテキスト間の類似度を得ることができる．

SequenceMatcher により求められる類似度は，ある類似度以上を類似していると定める閾値が不明瞭である．そこで，2 クラス分類を可能とする機械学習アルゴリズムの SVM を用いる．SVM に SequenceMatcher で類似度の算出に用いられている 2 つの要素，テキスト間の一致した要素数（M）と要素の総数（T）を与え，似ているか似ていないかの 2 クラス分類を行う分類器を生成した．分類器の性能は，学習データの精査やパラメータチューニングを行うことで向上が見込めるが，試作システムではこれらを行わないものとした．

学習データには 2017 年 10 月 30 日の記事データを用いており，産経ニュースが 445 件，朝日新聞デジタルが 88 件である．2 つの記事タイトルを比較した際の M と T の組み合わせのうち 36187 件を学習させており，正解のラベルには筆者の判断による正解が与えられている．

判定処理後に再帰処理による分類を行うためのデータを生成する．再帰処理用のデータは，記事タイトルの組み合わせと似ているかどうかの判定結果をセットにしたデータである．

(3)　再帰処理による関連記事の分類

再帰処理用に生成したデータを基に，再帰処理を行う．再帰処理の終了条件として，既に処理をした値である場合と似ている記事が存在ない場合の 2 つを設定している．

3. 評価実験

試作システムの関連記事分類・提示機能が同一出来事の記事を収集する手間を削減できているか，関連記事の提示ができているかを確認するために評価実験を行った．

3.1 評価実験の方法

評価実験の方法は，実際のニュースサイトから人手で同一出来事の記事を収集する時間と試作システムが関連記事として同一出来事の記事を提示するまでの時間を比較する．また，人手で収集した同一出来事の記事集合を正解とし，試作システムが関連記事として提示した記事集合の精度を適合率，再現率，F 値で求める．提示した記事集合の精度を比較するために，試作システムで同一出来事かどうかの判定を ratio メソッドにより得られる類似度により判定し，閾値を 0.5 とした場合の適合率，再現率，F 値を併せて求める．

評価実験は情報系大学生 6 名を対象として行った．実際のニュースサイトは産経ニュースと朝日新聞デジタルとし，使用する記事は 2018 年 1 月 18 日の記事とした．

3.2 評価実験の結果

時間の比較についての結果を表 1 に，関連記事提示の精度についての結果を表 2 に示す．

表 1　時間の比較（秒）

項目	人手	試作システム	時間の差
平均値	251.3	22	230.8
標準偏差	85	3	84.0
最大値	445	32	413
最小値	128	18	106

表 2　関連記事提示の精度（平均値）

手法	適合率	再現率	F 値
試作システム	0.90	0.72	0.74
閾値 0.5	1.00	0.59	0.71

表 1 より，試作システムが人手に比べて平均値で 230.8 秒速いことから同一出来事の記事を収集する際の手間を削減できているといえる．また，人手での作業と試作システムでの処理の時間が最大値となった記事は共通しており，「相撲」に関する記事であった．対象とした 2018 年 1 月 18 日は相撲に関する記事が多く掲載されており，人手では同一出来事かどうかの判断・選択に時間がかかり，試作システムでは再帰が深くなったため時間がかかったと考えられる．

表 2 より，試作システムが閾値 0.5 とした場合と比べて，再現率と F 値が高くなったことから概ね同一出来事の記事を提示できているといえる．適合率が低くなり再現率が高くなったことから試作システムの分類器は閾値を低く設定していると考えられる．理由として，産経ニュースと朝日新聞デジタルの記事タイトル文字数の違いが考えられる．各ニュースサイトの特徴を考慮することで分類精度の向上が期待できる．

4. おわりに

複数の情報源から記事を取得し，同一出来事の記事を収集する際に手間がかかることを問題点として挙げた．問題点を解決するために，複数の情報源のニュースを確認することができ，同一出来事の記事を提示する機能を持つシステムを開発した．評価実験により試作システムは，同一出来事の記事を収集する手間を削減でき，概ね同一出来事の記事を提示できているといえる．今後は，同一出来事を比較しながら閲覧する際にかかる手間の削減に着手し，皿海らの研究[1]を参考とする機能の考案と検討を行うことで，より記事の閲覧を支援するシステムを目指す．

謝辞

本研究を行うにあたり，ご助言をいただいた方々ならびに，評価実験にご協力いただいた方々に厚く御礼申し上げます．

参考文献

[1] 皿海宏明, 湯本高行,新居学,上浦尚武, “同じ出来事についての記事からの共通点と差異の抽出”,情報処理学会第 78 回全国大会, pp.551-552 (2016).