5ZB-5

Web上のニュースを対象にした動向情報抽出手法の提案と実装

河野安友未† 小林一郎‡
†お茶の水女子大学大学院人間文化研究科数理・情報科学専攻
‡お茶の水女子大学理学部情報科学科

1 研究背景と目的

今日ではインターネットの普及により，多くの電子情報を収集することが容易となってきている．現在のWeb検索においては，特定時点の情報，またはそれに関する情報の取得は容易に行える．しかし，経済，国際情勢などを知るには，対象の動向情報を捉える必要性が高まっており研究が進められている[1][2]．このことより，本研究では，特定対象物の動向情報抽出手法の提案と実装を目的とする．具体的には，ニュースサイトで提供している時事ニュースをまとめたRSSを利用して取得されたニュース記事の中から，特定対象物への動向情報を抽出する．

2 システムの概要

今回実装したシステムは，
工程1：同一記事追跡
工程2：追跡対象の状態抽出のためのテキスト解析
の2つからなる．図1に構築したシステムの処理の流れを示す．

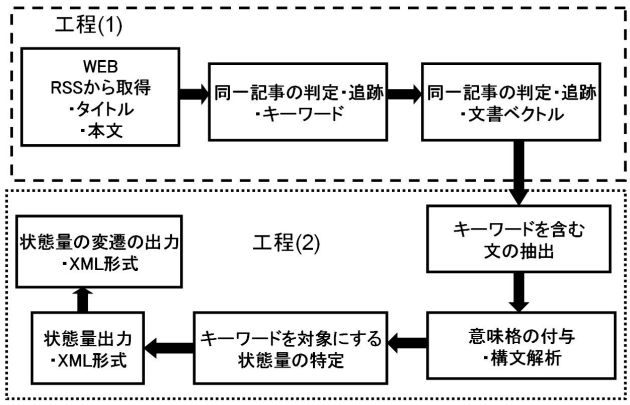

図1: システムの処理の流れ

A Method of Extracting Trends Information from On-line News Articles
Ayumi KONO†, Ichiro KOBAYASHI‡
†Graduate School of Humanities and Sciences, Ochanomizu University,2-1-1 Ootsuka Bunkyo-ku Tokyo 112-8610
‡Dept. of Information Sciences, Faculty of Science, Ochanomizu University,2-1-1 Ootsuka Bunkyo-ku Tokyo 112-8610
{kono, koba}@koba.is.ocha.ac.jp

2.1 インターネット上のニュース情報抽出

特定対象物の動向情報を抽出するために本研究では，インターネット上に公開されているRSSを利用する．今回，RSSは時事ニュースを扱っている7つのサイト，「gooニュース」,「Yahoo!JAPAN」,「nikkeiBPnet」,「REUTERS」,「YOMIURI ONLINE」,「asahi.com」,「Sankeiweb」で公開されているものを使用した．対象としたRSS中に記載されているURLのリンク先のニュース記事のHTMLファイルを自動取得し，その中から記事本文，記事タイトルを抽出し，テキストファイルとして時系列の順に番号が振られ保存される．今回は，12月4日〜1月7日の期間内に収集された全サイト合計4829記事をを使用して実験を行った．

2.2 同一記事の判定・追跡：キーワード

収集した記事の集合から同一記事を判定するために，我々は記事の内容を示す重要語が含まれる可能性が高い記事タイトルに着目する．記事タイトルから名詞だけを抽出し，その語を重要語として同一記事を追跡する．しかし，抽出される名詞すべてをそのまま重要語として利用すると，不必要な語で記事を追跡してしまうため，本研究ではtfidfを用いて重要語を判定し，さらにその中の固有名詞を最終的な追跡のためのキーワードとして記事を追跡している．

キーワードによる同一記事判定・追跡アルゴリズム

(1) 追跡期間における重要語の抽出

tfidf法を用いて蓄積された記事群を利用して語の出現頻度を計算する．出現頻度上位20 %を重要語とした．

(2) 追跡候補となる固有名詞の抽出

タイトルを茶筌を用いて形態素解析し，語の品詞情報を取り出す．その中から固有名詞のみを抽出する．

(3) 重要語を含むタイトルを抽出

(1)で抽出された重要語と(2)で抽出された固有名詞とをつき合わせ，マッチングできた語を追跡用の語としてファイルに保存する．

(4) 追跡記事番号収集

抽出された追跡用の語を含むタイトルの記事番号を，重要語をファイル名にしたファイルに格納する．

2.3 同一記事の判定・追跡：文書ベクトル

キーワードによって同一と判定された記事群の中には，別の内容の記事が混同している．そこで文書ベクトルを用いたベクトル空間法を利用して，キーワード

カテゴリの記事の中からさらに同一内容の記事を判定する．文書ベクトルの重み付けには tfidf 法を利用した．

文書ベクトルによる同一記事判定・追跡アルゴリズム

(1) 文書ベクトルに使用する語の抽出と重み付け

語の重み付けには tfidf 法を用いる．

(2) 不要語の除去

単一の記事にしか登場しない語は文書ベクトルを使用する際に不要であるため除去する．

(3) カテゴリごとの記事分類

文書ベクトルの類似度にあわせてクラスタリングを行う．全ての文書ベクトルのペアについて類似度を計算する．本研究では単語間類似度計測で一般的に使われる cosine 尺度を用いる．

2.4 テキスト解析

分類された記事から，キーワードが含まれる文のみを抽出し，解析に使用する．まず対象となる文を CaboCha/南瓜 [4] を用いて形態素解析，係り受け解析し，語の品詞情報，係り受けの情報を取り出す．解析した結果の中から，述語となる語の基本形を取り出し，その述語に係る文節を取り出す．文節が格を持つならば，格の種類を判別し，述語の基本形とそれがとる必須格，そしてその格に対する意味の制約（意味の選択制限）をまとめた意味格抽出用辞書を参照し，格がとる語の意味格（意味ラベル）を決定する [5]．意味格（意味ラベル）とは，格がとる語の意味的制約のことを指し，本研究では，文献 [6] に基づき，9 種類の格「ガ格」「ヲ格」「二格」「カラ格」「ト格」「デ格」「ヘ格」「マデ格」「ヨリ格」に対して，35 の意味格を用意している．

このような取り得る述語を意味格と格ごとに分類してまとめ，辞書化することによって，選択制限がなされ意味格を適切に判別することが可能になる．

意味格抽出用辞書に収録された格および意味格ごとの述語を利用し，文章中から格が抽出されれば，始めに文章中から取り出しておいた述語とマッチングをとる．述語が意味格抽出用辞書に収録されていれば，格の直前の語はその意味格と判別される．

テキスト解析により抽出された意味格は，意味格が付与される XML 形式でファイルに保存される．例として「マイクロソフトは 12 月 7 日，オンライン広告配信サービス「マイクロソフト デジタル アドバタイジング ソリューションズ」の国内提供を開始した」の文の解析結果例を示す（図 2）．

```
<?xml version="1.0" encoding="Shift_JIS" ?>
- <extract_data>
  - <sentence>
      <Predicate value="開始した" />
      <actor value="マイクロソフト" />
      <goal value="オンライン広告配信サービス「マイクロソフト デジタル
        アドバタイジング ソリューションズ」の国内提供" />
      <time value="12月7日" />
    </sentence>
  </extract_data>
```

図 2: 抽出された意味格（XML 形式）

2.5 テキストからの動向情報抽出

解析の結果，意味格が付与された文中において特に actor（主体・行為者），goal（行為の対象），Predicate（状態）が付与された部分は「状態」を示しているといえる．さらに特定の固有名詞を共有する文書を集めることにより，追跡キーワードに関するさらに詳細な動向情報を抽出する．図 3 に追跡キーワードを「マイクロソフト」とし，実験期間内におけるその動向を抽出し，その中の一つの固有名詞 (windows) について文書をクラスタリングした結果を示す．

例)「マイクロソフト - windows」

日付(time)	主体(actor)	対象(goal)	状態(Predicate)
2006年11月30日	マイクロソフト	「Windows Vista」の早期導入事例を	明らかにした
2006年12月8日	マイクロソフト	「Windows Vista Ultimate」のサポート終了期限を2017年4月11日にすると	発表した
2006年12月21日	マイクロソフト	Windows Vistaのセキュリティー関連機能について	説明した
2006年12月25日	マイクロソフト、セブン-イレブン・ジャパン、富士ゼロックスの3社	新OS「Windows Vista」向けに、デジタルカメラ写真のオンラインプリント・サービスを2007年1月30日より開始すると	発表した

図 3: 抽出された動向情報

図 3 から「マイクロソフト-windows」の時間的遷移をとらえることができ，状態がどのように変化していったのかを追うこと（動向情報抽出）が可能となっている．

3 まとめ

本研究では，Web 上のニュースを対象にした動向情報抽出手法の提案と実装法について述べた．このシステムでは，決められたカテゴリにあわせて記事を分類するのではなく，カテゴリそのものを記事群から抽出して分類を行う．そのためこのシステムでは事前知識を必要としない分類を行っている．さらにカテゴリに分類された記事本文中の状態を取り出して時系列に並べることによって，動向情報の抽出が可能となった．

参考文献

[1] James Allan, Ron Papk, and Victor Lavrenko; On-line New Event Detection and Tracking, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Pages: 37 - 45 , Year of Publication: 1998

[2] 森 幹彦，山田 誠二；Web における話題の時間変化の提示，JSAI2006.

[3] 奈良先端科学技術大学大学院松本研究室，形態素解析システム「茶筌」，http://chasen.naist.jp/hiki/ChaSen/?FrontPage

[4] 奈良先端科学技術大学大学院松本研究室，日本語構文解析器「CaboCha/南瓜」，http://chasen.org/ taku/software/cabocha/

[5] 河野安友未，小林一郎；Web 上のヘッドラインニュースを情報源とした質問応答システムの構築，FIT2005.

[6] 益岡隆志，田窪行則；“基礎日本語文法”，くろしお出版，1992.