話題語抽出機能を持ったニュースストリーム閲覧システム

足立 貴行　永田 昌明
日本電信電話株式会社　ＮＴＴサイバースペース研究所

１．はじめに

現在、インターネットを利用して、最新のニュースを見ることができる。しかし、複数の配信元から短時間に次々と配信される記事を、整理された状態で一覧することは困難である。

そこで、TDT(Topic Detection and Tracking)と呼ばれる研究分野では、オンラインニュースやニュース放送から話題を検出する研究が進められている(高間[1])。その他、ニュース放送記事を対象に、話題となる名詞句を抽出する手法（山田[2]）や、最新ニュース中の最新語と最新ニュース検索での検索語から短い周期で話題語を表示するシステム（川島[3]）が提案されている。

我々は、形態素解析[1]の辞書（特に、最新用語辞書）の整備を目的として、インターネットから最近の話題語を収集する研究に取り組んでいる。

本稿では、インターネット上のニュースから定期的に話題語を抽出し、その話題語を介して話題性の高い最新ニュースを閲覧できるシステムについて述べる。以下、第２節でシステムについて説明し、第３節で考察し、第４節でまとめを行う。

２．ニュースストリーム閲覧システム

本システムは、(1)ニュース収集、(2)対象文書作成、(3)単語抽出、(4)話題度計算、(5)話題語リスト表示、の５つのステップから構成される。また、(1)～(4)は事前処理であり、(5)でその結果を見ることができる。以下、各処理について順に説明する。

２．１．ニュース収集

実験のために、インターネット上のニュースサービスとして、複数の配信元のニュースを提供しているポータルサイト[2]から、１日１回の頻度で前日分のニュースを収集した。収集したサイトでは、URLの一部に年月日が含まれているので、その情報を用いて前日のニュースを収集することができる。また、HTMLに記事本文の範囲を示すタグが挿入されていたので、収集後、その情報を用いてタイトルと記事を抽出した。

[1] NTT研究所開発の形態素解析器JTAG(渕[4])
[2] goo.ne.jp

２．２．対象文書作成

収集されたニュース（タイトルと記事を連結したものでこれを１単位とする）のうち、一定期間のニュースをまとめて話題語抽出の対象文書とする。今回は一定期間として１日分とした。

また、対象文書から予め suffix array を作成しておく。suffix array は、Yamamoto[5]の方法を2byte 文字用に変更した。作成した suffix array は、後で述べる単語抽出や話題度計算で単語の出現頻度や、単語が出現する文書数（文書頻度）を求める際に利用する。

２．３．単語抽出

単語抽出では、処理の高速化のために、まず、対象文書から粗く抽出箇所を絞り込んでおき、その箇所に対して単語分割を行う。

２．３．１．抽出箇所の絞込み

対象文書からの抽出箇所の絞り込み方法は、足立[6]の語候補の絞り込みと同様の方法を用いた。
但し、機能語等を含む文字列の除外は行わず、文字種・文字数による文字列の除外の条件に、特定の文字種（カタカナ、アルファベット、アラビア数字、漢数字）が含まれ、同じ文字種で分断されているものを除外する、という条件を追加している。

２．３．２．単語分割

話題語は辞書未登録語である場合が多いと予想されるので、形態素解析を使わず、文字列統計量を用いて、絞り込んだ抽出箇所を単語分割する。単語分割は小澤[7]による方法を基本に、文字種を考慮する改良を行った。

単語分割方法

この方法は、入力文字列の任意の文字列(s)に対して単語スコア（wscore(s)）を求め、入力文字列全体で単語スコアの総和が最大となる単語並び(S)を求める（式(1)）。

$$S = \underset{s \in S0}{\operatorname{argmax}} \sum \mathrm{wscore}(s) \qquad (1)$$

S ：単語スコア最大となる文字列の集合

S0：入力文字列中の任意文字列の集合
s ：S0 中のある文字列

単語スコアは、式(2)のように、ある文書における文字列統計量（文字列の出現頻度と文書頻度）から計算した残差 IDF(RIDF(s))と文字列長(length(s))と文字種条件(w(s))の項からなる。

wscore(s)＝RIDF(s)×(length(s)-1)×w(s) (2)

残差 IDF は式(3)のように定義される。

RIDF(s)＝-log(df/D)＋log(1-e^(-tf/D)) (3)

tf：文書中の文字列 s の出現頻度
df：文書中の文字列 s を含む文書数
D ：文書中の総文書数

式(2)の文字種条件 w(s)として、下記の２点とした。下記以外は、w(s)＝1 とする。

(A) 文字列の先頭か末尾が、カタカナ、アルファベット、アラビア数字、漢数字であり、同種の文字種で分断されている場合、および記号が２文字以上続く場合、その文字列の w(s)＝0 とする
(B) ひらがなを含む文字列に対する残差 IDF が閾値（＝1.1）未満のものは単語らしくないと判断し、w(s)＝0 とする

上記の分割方法により、単語候補を得る。但し、名詞でないものの多くは１文字に分断されるので、１文字の語を除外する。

単語スコア計算用の文書

この単語分割方法では、単語スコアの計算で用いる文字列統計量の求め方が最も重要である。

我々は、予め、別の文書の文字列統計量を計算しておき、実行時に比較的小規模な対象文書の文字列統計量を計算して足し合わせるようにする（足立[6]）。これにより、より信頼性の高い統計量が得られ、かつ、処理時間も短時間で済む。今回は別の文書として毎日新聞 94 年版の 1 年分を用いた。

２．４．話題度計算

抽出された各単語に対し、対象文書中の文字列統計量に基づいて話題度の計算をする。

我々は、話題度を対象文書中で内容をよく表わす、もしくは内容との関係が高い語であって、さらに、ある期間とそれ以前のある期間とを比べると出現状況が劇的に高まっているものと考えた。そこで、単語 w の話題度を、(4)式のように定義した。但し、past_tfidf(w)＝0 のとき、話題度(w)＝now_tfidf(w)とした。

話題度(w)＝now_tfidf(w) / past_tfidf(w) (4)

now_tfidf(w) ：
　現在のある期間における単語 w の tfidf
past_tfidf(w)：
　以前のある期間における単語 w の tfidf

tfidf はある文書に偏って出現し、特徴的な単語を得ることができるだけでなく、出現頻度(tf)も考慮しているので、ある期間とそれ以前の期間を量的に比較するのにも適している。今回は、ある期間を１日とし、それ以前のある期間をその前日１日とした。

２．５．話題語リスト表示

CGI を用いて、話題度の大きい順に単語を並べた話題語リストを表示する。以下、図１の表示例を基に説明する。

２．５．１．話題語リスト

図 1 の左上のページは話題語リストの例である。利用者が、年月日、ジャンル（全て、社会、経済など 9 種類）、表示件数（10、100 件）を選択すると、該当する文書集合中の話題語リストが表示される。また、話題語だけでなく、話題度(score)、出現頻度(tf)、文書頻度(df)も表示しており、例えば、ジャンルを絞って対象文書数が少ない場合、本当に話題語かどうかを利用者が判断する材料となる。図１の例では、2004/1/12 の全てのジャンルにおける、上位 10 件の話題語リストであり、１位は「鳥インフルエンザ」であった。

利用者は、「鳥インフルエンザ」をクリックすると、話題語に関する統計量などの情報（図１の右のページ）が表示される。

２．５．２．話題語の情報

図１の右のページは、特定の話題語（「鳥インフルエンザ」）に関する情報（関連語、出現状況の推移、タイトル、元ページへのリンク、引用）を表示した例である。

関連語

ページの上段には、「鳥インフルエンザ」と同じ文書に現れた別の語（話題語の上位 100 位中の語）を列挙しており、例えば、「養鶏場」や「ニワトリ」といった「鳥インフルエンザ」に関連性のある語が分かる。関連語を見ることで、話題語の意味の限定や、文脈の推測ができる。また、別の語をクリックすると、同じ年月日、同じジャンルの別の語に関する情報のページを見ることができる。

出現状況の推移

ページの中段には、指定した年月日の前後数日分の話題度(score)、出現頻度(tf)、文書頻度(df)の値が表示される。この表を見ることで、話題語がいつ出現し、出現状況はどう変化しているかを詳しく確認できる。また、年月日をクリックすると、その年月日における、話題語に関する情報のページを見

ることができる。

タイトル、元ページへのリンク、引用

ページの下段には、「鳥インフルエンザ」が出現するニュースのタイトル、元ページの URL へのリンク、ニュース中の話題語（強調表示）とその前後の文脈の引用を列挙している。ページの上段で説明したような関連語がない場合や、あっても話題語の意味が分からない場合は、まず、タイトルや引用を見ることで実際の文脈から理解することができる。この例では 2 番目のニュースの引用中に「…鳥の病気の一種「鳥インフルエンザ」のウィルス…」のような説明がされている。それでも分からない場合は、元ページの URL をクリックすることで、図１の左下にあるように元のページを開いて、全体を読むことができる。

３．考察

３．１．話題語

2004/1/12 において、話題語となった上位３例に対し、当日とその前後６日間の出現頻度、文書頻度、話題度を調べた（図２、３）。但し、上位１位の「鳥インフルエンザ」と３位の「養鶏場」は同じ文書に現れることが多く、結果もほぼ同じであったので、「養鶏場」に対する説明は省略する。図２は、１位の「鳥インフルエンザ」、図３は、２位の「移植」の出現状況の推移であり、X 軸は日付、左の Y 軸は出現頻度および文書頻度、右の Y 軸は話題度となっている。また、X 軸下の各矢印は同じ話題であったニュースの出現期間を表す。

図２、図３とも、前日と比べて tf、df が急激に増加する日（１２日）にはその単語の話題度が高くなっていることがわかる。

ただ、「移植」は一般的な単語のため、図３では、異なる話題のニュースが次々と出現している（矢印）。このように、異なるニュースが連続する場合は、同じ語でも内容ごとに分けて話題度を計算する必要があり、今後の課題である。

３．２．辞書未登録語

形態素解析用の辞書の未登録語がどの程度得られるか調べた。

図４は、システムが抽出した話題語の上位 100 件における、①既知語、②複合語（既知語から構成される）、③未登録語、④非単語、の割合（2004/1/11～13 の３日間を平均したもの）である。既知語と複合語を合わせると 84%は既知といえる。未登録語、非単語はともに 8%であった。

未登録語の多くは、固有名詞であった。文字種はカタカナ、アルファベットが大半を占める。他には、「よりきり」のようなひらがな語や「厚労省」のような略語も得られた。

一方、非単語は、単語抽出誤りである。多くは、「席大使」のような名詞の途中で分断されたもので、他は、「フラムは」や「悪意ある」のような助詞などの付加や、名詞でないものであった。

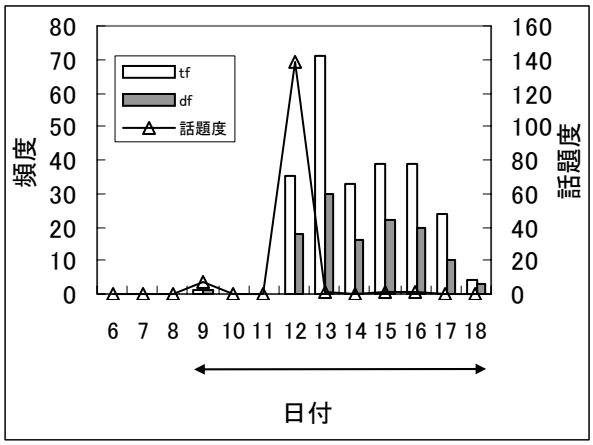

図２　「鳥インフルエンザ」の出現状況の推移

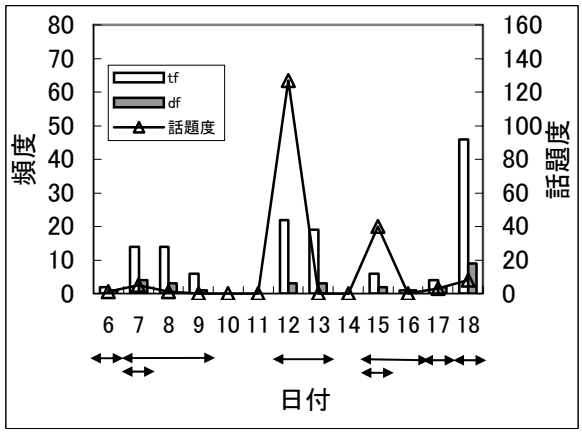

図３　「移植」の出現状況の推移

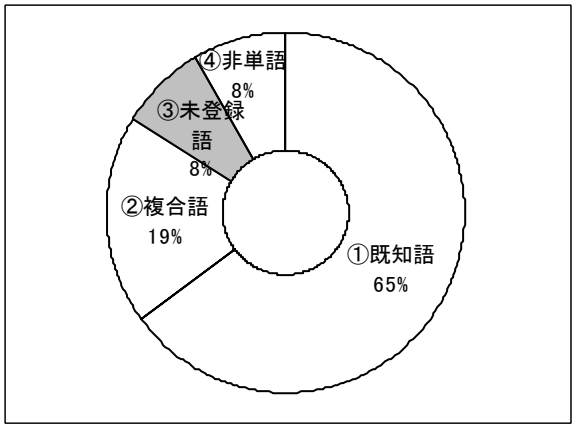

図４　上位 100 件中の未登録語の割合

４． おわりに

本稿では、辞書の整備を目的として、インターネット上のニュースから定期的に話題語を抽出するとともに、その話題語を介して話題性の高い最新ニュースを一覧できる閲覧システムについて述べた。

現在、本システムは、既存辞書に存在する語を話題語リストに表示しない機能を追加して、形態素解析用辞書の人手による整備のための支援ツールとして用いている。

今後は、考察で問題点となった箇所を改良するとともに、掲示板やブログなど、ニュース以外へのコンテンツへの適用を検討したい。

参考文献

[1]高間康史:Web 情報ストリーム,情報処理, vol. 44, No.7, pp.720-725, 2003.

[2]川島晴美,大橋二大,佐藤吉秀,安部伸治,大久保雅且:HotWindow:最新情報の話題性に着目した情報取得支援システム,インタラクション 2004, B28, 2004.

[3]山田一郎,金淵培,柴田正啓,浦谷則好:ニュース記事を利用者したトピック抽出の検討,言語処理学会第 5 回年次大会,pp.116-119,1999.

[4]渕武志,松岡浩司,高木伸一郎:保守性を考慮した日本語形態素解析システム,情報処理学会自然言語処理研究会,NL117-9,pp.59-66,1997.

[5]Mikio Yamamoto and Kenneth W. Church:Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus, Computational Linguistics, Vol.27, No. 1, pp.1-30, 2001.

[6]足立貴行,山田節夫,永田昌明:小規模な文書集合からの語彙獲得法,言語処理学会第 9 回年次大会,pp.274-277, 2003.

[7]小澤智裕,山本幹雄,山本英子,梅村恭司:情報検索の類似尺度を用いた検索要求文の単語分割,言語処理学会第 5 回年次大会, pp.306-308, 1999.

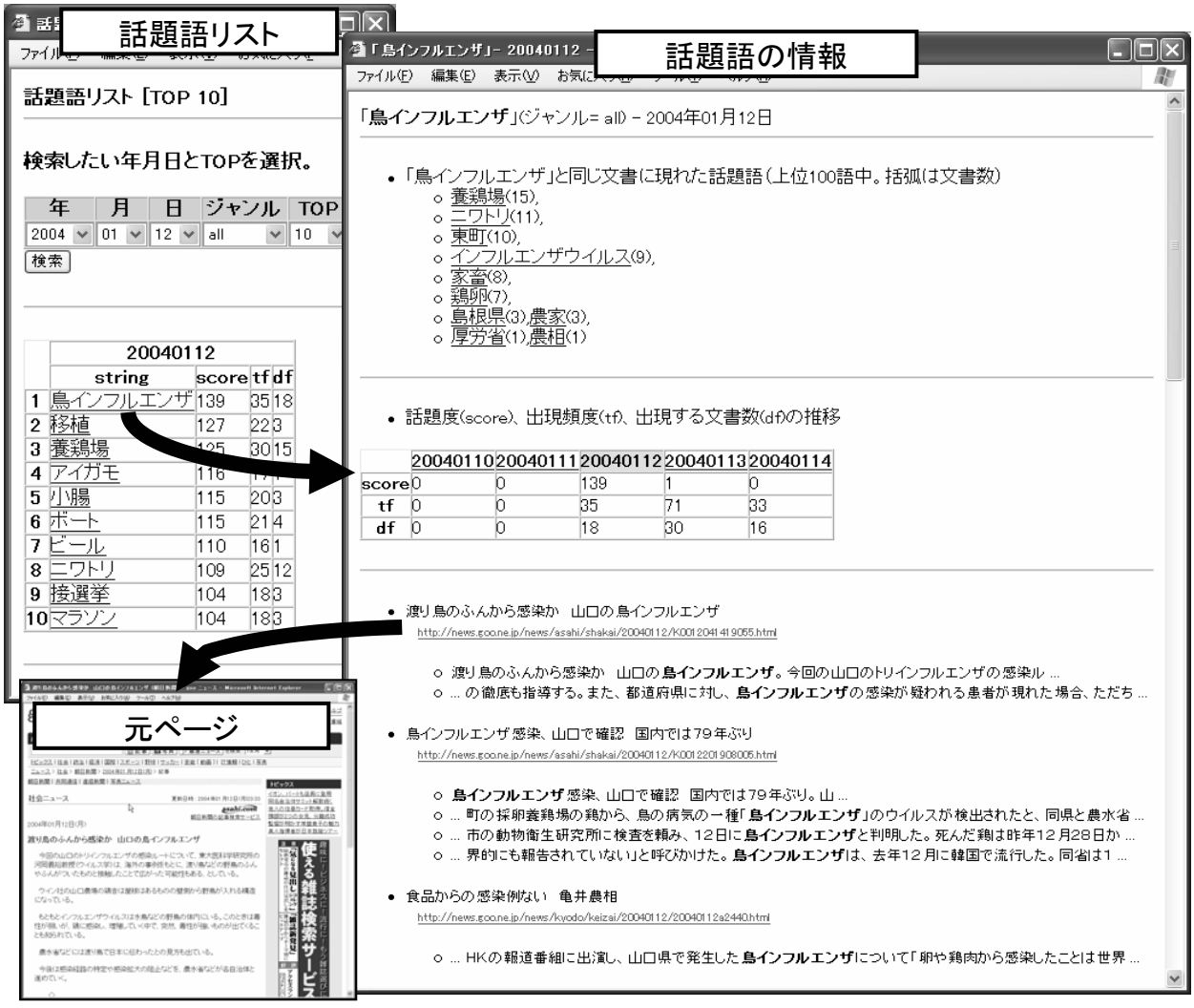

図１　表示例（「鳥インフルエンザ」）