ニュース記事を利用したトピック抽出の検討

山田一郎　　金淵培　　柴田正啓　　浦谷則好

NHK放送技術研究所

E-mail: {ichiro, kimyb, shibata, uratani}@strl.nhk.or.jp

1　はじめに

ニュースには、政治、経済、社会、スポーツなどの静的なジャンルの他に、時期とともに変化する動的なジャンル＝トピックが存在する。「W杯サッカー」「長野オリンピック」などがトピックに該当する。このようなトピックは視聴者が番組を選択するときの重要な鍵となる。また、これからの放送局では、ニュースを情報として利用するために、ニュースデータの分類や索引付けが不可欠となる。トピックは、そのような索引付けにも有効である。

従来のトピック抽出の研究では、テキストデータから複数の話題語を抽出、利用する手法が提案されている[1][2]。しかし、これらはテキストに含まれる重要な単語を抽出するもので、文単位の局所的なトピックしか抽出されていない。

本稿では、大量のニュース記事を対象として、世間で話題になるようなトピックを抽出する手法を提案する。そのために、記事に含まれる単語の時期毎に変化する重要度を利用する。以下では、まず単語の重要度を定義し、この重要度を利用したニュース記事のクラスタリングについて説明する。さらに、分類された記事から、トピックとしてより理解しやすい連続した複数の単語により構成される名詞句を、月単位で抽出する手法を述べ、この手法を用いた実験を報告する。

2　単語の重要度計算

トピックの抽出は、NHKの放送で実際に利用されているニュース記事のデータベースを対象として行う。このニュース記事データベースには、1日に約２００記事が電子化されて蓄積されている。各記事の第一文はニュース内容の全貌を説明することが多く[3]、これに対して、第二文以降はトピック抽出処理では不要な要素が多い。従って本手法では、記事の第一文のみを利用する。図1にニュース記事の一部を示す。

まず、このニュース記事に含まれている名詞と動詞に、重要度を定義する。以下に、重要度計算の処理概要を述べる。

2.1　χ^2値の定義

χ^2値は、観測値と期待値がどの程度一致しているかを測る指標である。単語の出現頻度と期待値のずれが大きければ、その単語は偏って出現していることになり、χ^2値は大きな値をとる。χ^2値を利用して、単語の月単位の重要性を評価する。母集団を対象とする月から前の一年間とし、対象月の単語 t の出現頻度を $n(t)$、その期待値を $e(t)$ とした時、$\chi^2(t)$は次の式となる。

○東京都は都内に張り巡らされた下水道の有効活用のため全国の自治体で初めて、民間の電話会社などが下水道管に光ファイバーなどのケーブルを設置できるようにして、下水道の民間への開放を進めることになりました。
○電機産業の労働組合でつくる「電機連合」はきょう開いた代表者会議でマルチメディア製品の売り上げが伸びて各社の業績は回復に向かっているとしてこの冬と来年の夏を合わせたボーナスの要求を二年ぶりに前の年より引き上げる方針を確認しました。
○スイスのローザンヌで開かれているＩＯＣ・国際オリンピック委員会の理事会は一日、二〇〇〇年のシドニーオリンピックで新たにトランポリンを実施することを決めました。
○サッカーのワールドカップアジア最終予選で今月七日の日本の初戦の相手となるウズベキスタンチームは試合前日の六日に来日を予定していることが判りました。
○長野オリンピックで世界各国の放送拠点になるＩＢＣ＝国際放送センターが完成し、きょうＮＡＯＣ＝長野オリンピック組織委員会に引き渡されました。

図1.　ニュース記事の第一文（一部）

$$\chi^2(t) = \frac{(n(t) - e(t))^2}{e(t)}$$

頻度が期待値より小さいときも、大きい場合と同じ正の値をとってしまうため、実際には以下の値を利用した。

$$\chi^{2'}(t) = \begin{cases} \chi^2(t) & \cdots n(t) \geqq e(t) \\ 0 & \cdots n(t) < e(t) \end{cases}$$

2.2 *idf*値の定義

単語の逆文書頻度(*inverse document frequency*)を*idf*値と呼ぶ。*idf*値は、記事中に頻繁に出現する単語ほど一般的な単語と見なし小さな値をとる。一ヶ月のニュース記事の総数をN、一ヶ月のニュース記事中で単語tが出現する記事数を$df(t)$としたとき、$idf(t)$は以下の式となる。

$$idf(t) = log(\frac{N}{df(t)})$$

2.3 単語の重要度の定義

χ^2値により月ごとに変化する単語の重要性を評価でき、*idf*値により一般的な単語を処理対象から除外できる。この2つの値を相乗的に利用して、月ごとに変化する単語tの重要度$weight(t)$を以下の式で定義した。

$$weight(t) = \chi^{2'}(t) \times idf(t)$$

1998年9月のニュース記事に含まれる単語の重要度を図2に示す。「台風」「ホームラン」「ミサイル」など、この月のキーワードとなりうる単語に大きな値が付けられ、良好な結果が得られている。

単語, 重要度	
台風,	18898.15
修正,	9995.42
防衛庁,	8425.61
ホームラン,	8291.00
マグワイア,	7320.54
調達,	7241.97
ミサイル,	6879.87
再生,	6781.90
ウー,	6015.50
ロン,	5853.95
備品,	5712.11
発射,	5454.62
背任,	5198.64
茶,	5143.28
長銀,	4593.75
便,	4322.70
カーディナルス,	4240.01
ソーサ,	3916.74
カブス,	3746.62
東洋通信機,	3668.33
サミー,	3548.49
域,	3539.06
暴風,	3403.42
欠航,	3123.57
日本長期信用銀行,	3100.78

図2. 単語の重要度上位（1998年9月）

3 トピック抽出

トピック抽出のために、まず、ニュース記事のクラスタリングを行う。同内容のニュース記事を一つのクラスタにまとめ、各クラスタから一つのトピックを抽出することで、抽出されるトピックは排他的な性質を持つ。

3.1 ニュース記事のクラスタリング

クラスタリングは、前章で定義した単語の重要度を利用し、一ヶ月毎のニュース記事を対象として行う。記事の特徴ベクトルVは、単語t_iが記事に含まれるときは$w(t_i)=weight(t_i)$、含まれなければ$w(t_i)=0$として、以下の式で定義する。

$$V = (w(t_1), w(t_2), \cdots w(t_n))$$

ここでは、データベース中の記事に出現するすべての単語数をn、単語を$\{t_1,\cdots,t_n\}$としている。記事Vの重要度$weight(V)$は以下の式とする。

$$weight(V) = \sum_{i=1}^{n} w(t_i)$$

また、m個の記事($V_1 \sim V_m$)が集まり形成するクラスタの重心ベクトルCVは以下の式とする。

$$CV = \frac{1}{m}(\sum_{i=1}^{m} w_i(t_1), \cdots, \sum_{i=1}^{m} w_i(t_n))$$

このとき、記事ViとクラスタCVjとの類似度$Sim(Vi,CVj)$は以下の式とする。

$$Sim(Vi, CVj) = \frac{Com(Vi, CVj)}{\sum_{k=1}^{n} w_i(t_k) + \sum_{k=1}^{n} w_j(t_k) - Com(Vi, CVj)}$$

$$Com(Vi, CVj) = \sum_{k=1}^{n} min(w_i(t_k), w_j(t_k))$$

ここで $Com(Vi,CVj)$ は、ベクトル Vi , CVj の共通する要素の値の和を示している。Vi と CVj が一致するとき、$Sim(Vi,CVj)$ の値は1となり、逆に全く関係ないとき0となる。

クラスタリング処理は以下の手順で行う。

1. 重要度が最大の記事を初期クラスタとする。
2. どのクラスタにも属していない記事に対し、
   - 記事の重要度の大きさの降順に、記事とクラスタとの類似度を計算する。
3. 類似度があるしきい値以上であれば、記事を最も類似したクラスタに統合。すべてしきい値以下であれば、新たなクラスタを生成。

全ての記事に対して上記2、3の処理を繰り返し行う。ここでは、しきい値を0.25として実験を行った。

3.2 クラスタを代表する名詞句の抽出

クラスタリング処理により、内容が似ている記事は同じクラスタに分類される。形成された全てのクラスタに対して、そのクラスタを代表する名詞句を抽出し、その名詞句をラベルとする。

まず、各クラスタを代表する記事を抽出する。クラスタ中の記事に含まれる単語のクラスタへの寄与度を、そのクラスタでの単語の出現率と単語の重要度の積とし、記事に含まれる単語の寄与度の合計がクラスタ中で最大のものを、そのクラスタの代表記事とする。

次に、この代表記事に含まれる全ての名詞を対象に、連続する名詞群、助詞「の」で接続した名詞群を抽出する。抽出した名詞群について、そのクラスタにおける寄与度の和が最大のものをクラスタを代表する名詞句とする。下記の例では、「台風十九号」が代表する名詞句、つまりクラスタのラベルとして抽出されている。これにより、すべてのクラスタにラベルが付けられる。

クラスタに含まれる記事数と重心ベクトルとの積をそのクラスタの重要度とする。大きな重要度を持つクラスタに付けられたラベルを、重要度の降順にトピックとして抽出する。

4 実験

前章で述べたトピック抽出法を検証するために、1995年3月～1998年10月までのニュース記事を対象にクラスタリンク、トピック抽出実験を行った。

4.1 実験システム

実験システムは、SGIのワークステーション上にシステムを構築した。そのユーザインターフェース画面を図3に示す。ユーザが年月を指定すると、システムはその月のトピックと記事を提示し、TVML[4]のインターフェースが音声合成装置を利用して記事を読み上げる。このシステムにより、ニュース記事を効果的に管理、検索することが可能となる。

例. 代表する名詞句の抽出

大型で強い 台風十九号 は 鹿児島県の奄美地方 を 暴風域 に巻き込みながら・・・

368.5　4259.6　337.7　577.2　← クラスタへの寄与度

抽出

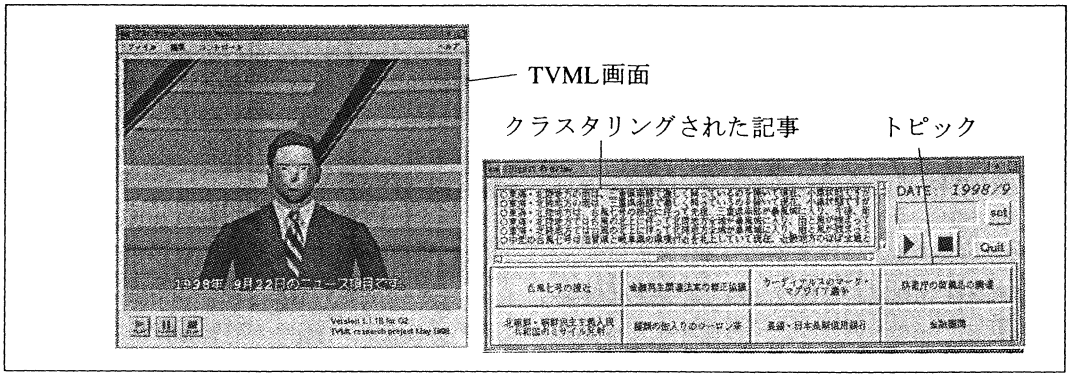

図3. トピック抽出実験システムのユーザーインターフェース画面

1997年9月	適合率	再現率
TOPIC[1]:台風十九号 -- 463個 3843780	79.7%	90.4%
TOPIC[2]:佐藤氏の辞任 -- 119個 1440906	96.6%	92.7%
TOPIC[3]:山口組の幹部の射殺事件 -- 87個 881859	92.0%	100%
TOPIC[4]:ダイアナさんの葬儀 -- 91個 719166	100%	84.3%
TOPIC[5]:サッカーのワールドカップアジア最終予選 -- 94個 657501	96.8%	94.0%
TOPIC[6]:ガルーダ・インドネシア航空の国内線のジェット旅客機 -- 56個 519650	100%	80.0%
TOPIC[7]:一連の商法違反事件 -- 74個 443503	100%	64.9%
TOPIC[8]:日米の防衛協力の指針 -- 88個 422016	100%	96.7%

1998年9月	適合率	再現率
TOPIC[1]:台風七号の接近 -- 703個 17622772	51.5%	94.8%
TOPIC[2]:金融再生関連法案の修正協議 -- 259個 7327851	98.8%	81.3%
TOPIC[3]:カーディナルスのマーク・マグワイア選手 -- 170個 4501151	100%	95.0%
TOPIC[4]:防衛庁の装備品の調達 -- 164個 3963743	100%	96.5%
TOPIC[5]:北朝鮮・朝鮮民主主義人民共和国のミサイル発射 --193個 3407358	100%	100%
TOPIC[6]:種類の缶入りのウーロン茶 -- 55個 1367057	100%	74.3%
TOPIC[7]:長銀・日本長期信用銀行 -- 59個 693348	-------	-------
TOPIC[8]:金融機関 --167個 567042	100%	78.4%

上記は、TOPIC[]:トピック ----- クラスタに含まれる記事数　クラスタの重要度
1998年9月のTOPIC[7]はTOPIC[2]と同一内容のため適合率、再現率は計算していない

図4.　トピック抽出結果

4.2　実験結果

1997年9月と1998年9月のトピック上位8項目を図4に示す。97年9月では、「台風十九号」がトップ項目として挙げられ、これに関連する463個の記事が一つのクラスタを形成している。他に「ダイアナさんの葬儀」「サッカーのワールドカップ最終予選」など、この月のトピックに相応しい内容が抽出されている。しかしながら、助詞「の」で表層的に繋げてラベル付けを行っていることにより,98年9月の結果には「種類の缶入りのウーロン茶」といった不自然な名詞句も存在する。そのため、記事の構文情報、意味情報を考慮した手法も検討している。

また、この2ヶ月の上位8項目に対して、クラスタリング結果を、人手により抽出した結果と比較し、それぞれの再現率、適合率を求めた。結果を図4に示す。その平均は、適合率94.4%、再現率88.1%であり、良好な結果が得られた。

5　おわりに

本稿では、単語の出現頻度のχ^2値と*idf*値を利用してニュース記事のクラスタリングを行い、月単位でトピックを抽出するシステムを構築、実験を行った。その結果、良好な結果が得られた。この手法は、これから急激な増加が予想されるマルチメディア情報の管理に有効となる。

今後は、クラスタを代表する名詞句の切り出し手法について検討を進める。さらに、テレビ視聴者が番組を選択する際の補助機能として、世間で話題となっているニュースや番組を自動的に知らせる機能を持った知的なテレビ[5]へと応用させる予定である。

【参考文献】

[1]大附ほか「ニュース音声を対象とした大語彙連続音声認識と話題抽出」信学技報, SP97-27, pp67-74 (1997)

[2]今井ほか「ニュース番組自動字幕化のための音声認識システム」情報処理学会研究会, SLP-23-11,pp59-64(1998)

[3]加藤「ニュース文を対象にした局所的要約知識の自動獲得」言語処理学会第4回年次大会論文集,pp542-645(1998)

[4]林ほか「テレビ番組記述言語TVMLの言語仕様とCG記述方法」第3回知能情報メディアシンポジウム論文集,pp141-148(1997)

[5]金ほか「エージェントTV」信学技報, AI98-54, pp1-8(1998)