ネットワークニュースにおける スレッド単位のキーワードの抽出

Keyword Extraction from Threads of the Network News

毛利 隆夫　　田中 英彦
Takao MOHRI*　　Hidehiko TANAKA
東京大学大学院 工学系研究科
Graduate School of Engineering, The University of Tokyo

Because of the increase of the amount of information, technology to select or summarize information becomes more important. We propose a new view of the network news with information selection and summarization. It is based on thread, the group of articles with the same subject. By showing longer threads at first with their keyword summaries, users can easily understand what is a recent hot topic. Experimental results show that morphological analysis with augumented vocabulary from news articles is accurate enougth to extract candidates of keywords. The importance of keywords is decided by their frequencies in the thread, the newsgroup and the all of the newsgroups.

1 はじめに

1.1 本研究の背景

現在，情報の電子化やインターネットの発達などにより，大量の情報が容易に入手できるようになった．その半面，人間が処理できる情報量には限界があるため，情報の流通を促進する技術と同時に，大量の情報の中から必要な情報を取捨選択する技術や，それらを要約する技術などが強く求められている．本研究で対象としたネットワークニュースでも，ネットワークの普及や利用者数の増加によって記事数は増加し，効果的な情報フィルタが求められている．

ネットワークニュースでは，記事がニュースグループと呼ばれるカテゴリ毎にあらかじめ分類されている．利用者は一般に自分の興味を引かれるニュースグループを事前に選択してその部分のみを読むのが普通であり，ニュースグループ全体を概観するようなインターフェイスは提供されていない．そのため，他のニュースグループでの活発に議論されている話題があったとしても，気付かずに過ぎてしまう事が多いと思われる．ニュースグループを問わず，活発に議論になっている話題を知っておきたいと思うユーザは，一定数いるのではないかと筆者らは考えている．しかし，ニュースグループは多数存在し，また記事のサブジェクトだけでは内容を理解するのに不十分である場合が多いため，ネットワークニュース全体を概観するのに要する労力は少なくない．

1.2 本研究の方針

こういった利用者の要求を満たすために，我々は，従来の見方とは異なった，次のようなネットワークニュースの提示方法を提案する．

1. ニュースグループの垣根を意識する事無く，スレッドを活発に議論されている順に表示する
2. スレッド毎にキーワードによる要約を付加し，本文を読む前に記事内容をある程度推測できるようにする

以下本稿では，主にスレッド毎にキーワードによる要約を付ける方法について述べる．

2 ネットワークニュース

2.1 対象とするニュースグループ

本研究では日本語で書かれた記事を対象としたため，ニュースグループは，'fj' で始まるニュースグループ (以下単に fj と呼ぶ) に限定した．ただし，プログラムを投稿するためのニュースグループ (fj.sources, fj.archives.*) や，投稿のテストを目的としたニュースグループ (fj.test) は，あらかじめ対象から除外している．また英語などの日本語以外の記事は，本研究の対象外とした．

2.2 ニュースの記事の特徴

ネットワークニュースで配送されている記事は，その大部分がテキストであり，電子化された大規模なコーパスとみなすことができる．ニュースの記事の特徴としては，以下のような点が挙げられる．

連絡先: 毛利 隆夫　東京大学工学部 電気工学科 田中英彦研　〒113 東京都文京区本郷 7-3-1
Phone: 03(3812)2111 内線 7413　Fax: 03(5800)6922　E-mail: mohri@MTL.T.u-tokyo.ac.jp
* 現在，(株) 富士通研究所に所属

- 情報量が多い
  fj の記事量は，1996 年 3 月の時点で，一日当たり 約 1500 記事 (約 4MB) に達している [new96]．文書量の多さは，コーパスとしては大きな特長である．
- あらかじめニュースグループ毎にカテゴリ分けされている
  fj では現在，約 300 種類のニュースグループがあり，投稿者はニュースグループを指定して投稿する．あらかじめ分類されているというのは，コーパスとして見た場合に大きな利点である．

その半面，意図的に編集されたものではないため，コーパスとして見た時には以下のような扱いづらい面も多い．

- 誤字脱字等が混入している場合が多い．
- 顔マーク ((^_^) 等) や文章の引用記号 ('>' 等) など，ニュースや電子メイルに特徴的な表記が使われる
- 記事中にプログラムや表，地図などの，文章以外のものが書かれている場合がある
- 文体が統一されておらず，わざとくだけた書き方になっている記事も多い

そのため，十分な前処理を行って文章以外のものを分別しておく必要がある．また，従来よく用いられていた形態素解析プログラム等がニュースの記事に対しても十分な精度で動作するかどうか確認する必要がある．

2.3 スレッド

ニュースの記事は，前に投稿された記事の一部を引用して，それに対して自分の意見を書き加えたものが多い．引用は何重にも繰り返される事もある．このような引用被引用の関係にある記事の集まりはスレッドと呼ばれている．一つのスレッドでは同じ話題が話されている場合が多いため，スレッドを話題の単位とし，そのスレッドを構成する記事の数によって話題の活発さを定義する (利用する記事の日数は，ユーザが指定する)．なお本研究では，サブジェクトのみを同一スレッドか否かの判断に用いている．

3 キーワードの抽出および表示

3.1 キーワードによる要約の必要性

ニュースの記事の中からスレッドを抽出し，それらを記事数の順に表示するのは困難ではない．スレッドのサブジェクトを記事数の多い順に単純に提示しても「活発に議論されている話題」を利用者に提示する点ではある程度効果があるものと思われる．

しかし，記事のサブジェクトは，必ずしも的確に本文の内容を表していない場合が多い．サブジェクトが簡潔すぎる場合や，議論の途中で話の方向が変化した結果がサブジェクトに反映されていない場合などには，サブジェクトは役に立たない．そのため，何らかの別の要約情報が必要であるといえる．本研究では，利用者の記事内容の把握を容易にするために，スレッド毎にキーワードによる要約を付加している．

3.2 キーワード抽出および表示の手順

キーワードの抽出および表示は，以下のような手順で行われる．

1. 前処理
   (a) ヘッダ，署名 (signature) の除去
   (b) ニュース特有の記号の除去，英語の除去
   署名や顔マークなどは，ヒューリスティクスて検出し除去する．また，対象言語を日本語に限定しているため，一行で半角英数字の比率が一定割合を越えた場合には，その行は削除する．これにより，プログラムなども除去される．
   (c) 引用の深さを検出
   ヒューリスティクスを用いて引用マークを除去する．引用されている文章とともに，引用レベル (何重に引用されているか) も併せて抽出する．
   (d) 一文一行にまとめる
   複数行にまたがっている一文を，句読点などを参考に一行にまとめた．
2. キーワードの抽出
3. キーワードの得点の計算
4. スレッド単位で得点の高い順にキーワードを表示

以下では，2. 3. について詳しく述べる．

4 キーワードの抽出

4.1 ニュース記事の形態素解析

本研究では，まず前処理によって一つの文章を一行毎にまとめた後で形態素解析を行い，名詞等[1]の品詞の単語を抽出して，キーワードの候補を作成する．この候補のことを，キーワード要素と呼ぶことにする．形態素解析には，フリーソフトである JUMAN ver 2.2 [松本 96] を利用した．

ニュースの記事の文章の質がそれほど高くないため，形態素解析が実用的な精度で動作するかどうかに不安があったが，以下のような実験を行った結果，ニュース記事向けの語彙を補強することで，キーワード要素を抽出する目的には十分な解析の成功率が得られることを確認した．

実験の方法は以下の通りである．まず，ある日の fj の記事[2]の中からランダムに 100 個の記事を選択し，訓練データとした．また，別の日[3]の記事からもランダムに 50 記事を選択し，テストデータとした．

訓練データの記事には，前処理の後，形態素解析を行い，誤解析を人手で検出し，これらの記事の形態素解析のために不足している語彙をリストアップした．その後，これらの語彙を追加して，訓練データを再度形態素解析した．追加した単語は 630 個で，その内訳は名詞 543 個，助詞 27 個，

[1] 名詞，英字・記号，ナ形容詞語幹，接頭辞，接尾辞，未定義語の 6 種類をキーワード要素として抽出した．

[2] 1996 年 3 月 4 日に local の news server に到着した記事

[3] 同 3 月 25 日分

記号27個，感動詞16個，その他が17個である．また，テストデータは，語彙を追加していない辞書と追加後の辞書のそれぞれを用いて，形態素解析を行った．

訓練データ，テストデータともに，人手で正解を作成しておき，正解の品詞がうまく分離できていない場合や，分離できていても品詞が誤っている場合には品詞分類の誤りとした．また，品詞が間違っており，かつキーワード要素に分類される品詞かどうかも異なっている場合には，キーワードの分類の誤りであるとした．

実験結果を図1に示す．訓練データによって得られた語彙を追加した場合には，テストデータにおいても誤解析を抑えることができるのがわかる．特にキーワードか否かの分類では，テストデータにおいても誤り率が 0.5% と非常に小さい．したがって語彙を追加すれば，従来の形態素解析プログラムもキーワード要素を得る目的では十分に使えることがわかった．

次に誤解析がどの品詞で発生していたかを調査した．その結果を図2に示す．語彙を追加した後の品詞分類の誤り率は，名詞の場合が圧倒的に多くなっている．これは，名詞の未知語が多く含まれていることを示している．しかし名詞も未知語もキーワード要素であるため，キーワードという面から見れば等価であるため，キーワードの分類の誤り率は低く抑えられている．

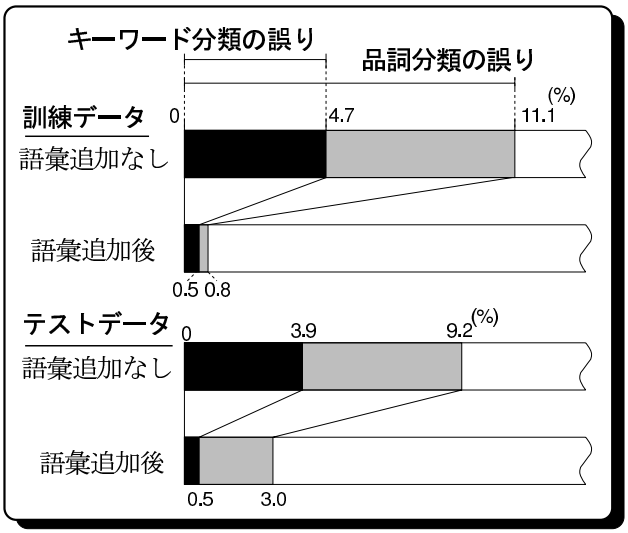

図 1: 訓練データとテストデータでの誤り率

4.2 キーワード要素の連結

つぎに，品詞に対応しているキーワード要素から，頻度や得点を計算する単位であるキーワードへの変換を行なう．キーワードとしては，キーワード要素そのものも登録されるが，キーワード要素が文章中で連続した場合には，それらの品詞を連結したものもキーワードとしている．例えば，「人工」「知能」「学会」という3つの名詞が連続して検出された場合には，キーワードとしては，「人工」「知能」「学会」「人工知能」「知能学会」「人工知能学会」という6通りの語を登録する．ただし，最終的にユーザに表示

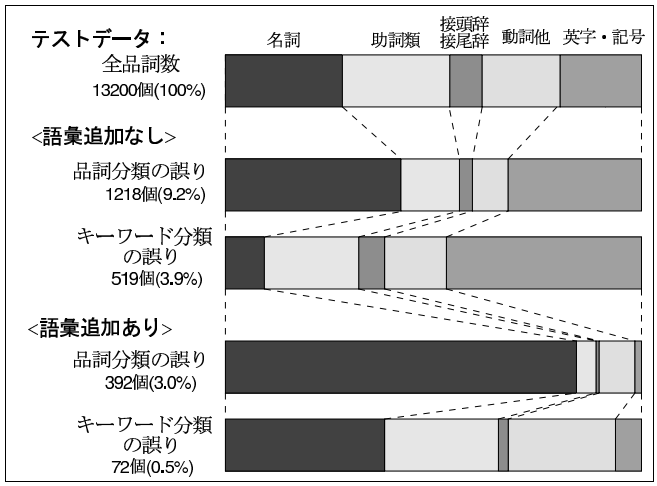

図 2: テストデータの品詞の誤解析の比率

される段階では，他に表示されるキーワードの部分文字列になっているようなキーワードは，表示しないようにしている．

5 キーワードの得点の計算

つづいて，各キーワード毎に，その頻度情報を元にして得点が計算される．まず各スレッド毎にキーワードごとに重みつき頻度が計算される．記事中に引用されている文章中のキーワードは重要度が高いと考えられるため，頻度に重みをつけている (式 (1))．

つぎに，キーワードの相対頻度を計算する．相対頻度は，同一スレッド全体の重みつき頻度の和で各キーワードの重みつき頻度を割って正規化したものである．すべてのキーワードに対して計算を行うのは得策ではない．低頻度のキーワードを避けるために，それぞれ上位 N 語[4]のキーワードに絞って特定を計算する (式 (2))．ニュースグループ単位や全ニュースグループでの重みつき頻度，相対頻度も同様に定義される．

最後に，各スレッドごとにキーワードの得点を式 (3) により計算する．あるキーワードは，そのスレッドに高い頻度で出現し，そのニュースグループでは珍しい単語であり，かつ全ニュースグループを通しても珍しい単語である場合に高い得点が与えられる．表1に，fj 全体での頻度の高いキーワードを示す．

表 1: fj 全体での頻度の高いキーワード

私, 人, もの, 何, 気, 問題, 話, ところ, 記事, 自分, ▪

- スレッドでのキーワードの重みつき頻度 $wfreq$:

$$wfreq(thread, kw) = \sum_i (1 + quote\_level(i) * weight) \quad (1)$$

- スレッドでのキーワードの相対頻度 $rfreq$:

$$rfreq(thread, kw) = \frac{wfreq(thread, kw)}{\text{上位}N\text{語の}wfreq\text{の平均}} \quad (2)$$

[4]N=min(スレッドのキーワードの個数,1000) としている

- キーワードの得点 *score*:

$$score(thread, ng, kw) = rfreq(thread, kw) - max(rfreq(ng, kw), rfreq(all, kw)) \quad (3)$$

$quote\_level(i)$: 文章 i の引用レベル
(0: 非引用，n: n 重に引用)
$weight$: 引用された語への重み付け
$rfreq(ng, kw)$: ニュースグループ ng での
キーワード kw の相対頻度
$rfreq(all, kw)$: 全ニュースグループでの
キーワード kw の相対頻度

次に，実際のスレッドにおいてキーワードの得点を計算する例を示す．対象としたスレッドは，fj.rec.autos という，自動車に関連した話題が中心のニュースグループにおいて，「Fog Lamp」というサブジェクトで投稿されていたものである．実際の記事の内容は「どういう場合にライトをハイビームして走るのがよいか」というもので，サブジェクトとは関連が薄くなっている．キーワード抽出には，このスレッドに属する 109 個の 記事[5]を使用した．ニュースグループおよび全体のキーワードの相対得点の計算には，別の日の 1 週間分の記事[6]を使用した．

キーワードの得点を計算の様子を表 2に，単純な重みつき頻度によるキーワードの順位と，得点によるものとの比較を表 3に示す．「自分」「私」といった，全ニュースグループで広く使われている単語や，「運転」といった，ニュースグループで頻出している単語の順位が下がり，このスレッド特有のキーワードが上位になっているのがわかる．

表 2: キーワードの得点の計算

キーワード	相対頻度 (NG:newsgroup) スレッド	NG	全 NG	得点
自分	28.329	4.446	6.316	22.013
ハイビーム	27.217	0.000	0.000	27.217
光	21.132	0.000	0.537	20.595
ライト	20.740	0.577	0.000	20.163
私	19.955	11.527	25.047	-5.092
自動車	16.553	1.606	0.000	14.947
運転	15.375	4.405	0.921	10.970
自転車	13.870	0.000	0.000	13.870

表 3: 重みつき頻度と得点でのキーワード順位の比較

重みつき頻度:
自分, ハイビーム, 光, ライト, 私, 自動車, 運転, ▪
得点:
ハイビーム, 自分, 光, ライト, 自動車, 自転車, ロービーム, ▪

[5] '96 年 3 月 24 日〜30 日に local な news server に到着した記事
[6] '96 年 3 月 3 日〜9 日に local な news server に到着した記事

6 関連研究

佐藤ら [佐藤 94] は，ニュースの記事のダイジェストを作成する研究を行っているが，その研究は特定のニュースグループでの記事のスタイル情報や言語表現パターンを手がかりにダイジェストを行なうものであり，本研究とは目的が異なる．

キーワードの抽出では，シソーラスを用いる方法 [木本 91] があるが，ニュースの記事では分野が幅広くシソーラスの入手・作成が困難であるために，本研究ではシソーラスを利用しない方法をとった．

また，本研究ではニュース記事の口語的な表現に対処する際，簡単のために語彙を追加する方法をとったが，抜本的には [竹元 94] のように形態素解析プログラムを改良すべきであると考えられる．

7 まとめと今後の課題

本研究ではネットワークニュースを対象に，スレッドと呼ばれる同一サブジェクトの記事の集まりを単位として，キーワードによる要約を付加してユーザに提示する方式を提案した．

ニュース記事から得られた語彙を新たに追加することにより，形態素解析の精度は，キーワードの候補を抽出するという意味では十分であることを確認した．また，ネットワークニュースでは記事がニュースグループ毎に分類されていることを利用して，キーワードの得点を計算する方式を示した．

今後は本システムを定常運転して一般に公開し，ユーザからのフィードバックを元にシステムを改良していきたい．

謝辞

本研究は文部省科学研究費補助金 (特別研究員奨励費, No. 06004134) の援助による．

参考文献

[new96] news@TokyoNet.AD.JP. Daily traffic status. *tnn.netnews.stats*, 3 1996.

[佐藤 94] 佐藤 円, 佐藤 理史, 篠田陽一. 電子ニュースにおけるダイジェスト機構の実現. In 第 *49* 回情報処理学会全国大会講演論文集 *(3)*, pages 221–222, 1994.

[竹元 94] 竹元 義美 福島 俊一. 口語的表現を含む日本語文の形態素解析の実現と評価. In 第 *48* 回情報処理学会全国大会講演論文集 *(3)*, pages 37–38, 1994.

[木本 91] 木本 晴夫. 日本語新聞記事からのキーワード自動抽出と重要度評価. 電子情報通信学会論文誌, J74-D-I(8):556–566, 8 1991.

[松本 96] 松本 裕治, 黒橋禎夫, 宇津呂武仁, 妙木 裕, 長尾 真. ftp://ftp.aist-nara.ac.jp/pub/nlp/tools/juman/juman2.2.tar.gz. 平成 8 年 1 月