新聞記事の特徴を利用した推薦話題検索のためのキーワード抽出

濱崎 雅弘† 小作 浩美†† 河野 恭之† 木戸出 正継†
†奈良先端科学技術大学院大学 ††通信総合研究所

我々は新聞記事を用いて，注目すべきキーワード (話題) を見つけだし，そのキーワードを元にオンライン上の情報をフィルタリングする手法を提案する．この手法を実現するために，まず新聞記事からの注目すべき話題の抽出実験を行った．本稿では，抽出実験とそこから得られた知見について報告する．
実験では一年間のデータから χ^2 分布で時間軸に対して出現に偏りがある語を見つけ，その単語すべてについて共起情報を調べ，クラスタ化する．その単語クラスタから，新聞中の出現頻度を元に話題抽出を行い，一年間の中で注目された話題を発見した．次に，一ヶ月間だけのデータから，出現頻度の変化，および紙面特性を考慮した評価法を用いて話題を抽出した．そして，良好な結果を得ることができた．

1 はじめに

オンライン上の情報は，一般に無数の人々によって，無計画に制作・編集されたものが多い．これらの情報は，多くの有益な情報を含んではいるが，全体として整理されていないため，一般に利用者にとって検索・収集が容易である情報源とは言い難い．そこで，利用者は何らかの方法で無数の情報の中から，有益な情報を取り出す必要がある．

我々は新聞記事を用いて，注目すべきキーワード (話題) を見つけだし，そのキーワードを元にオンライン上の情報をフィルタリングする手法を提案する．この手法を実現するために，まず新聞記事からの注目すべき話題を抽出する必要がある．そこで，新聞記事の特徴を利用し，抽出実験を行った．本稿では，その実験結果と，そこから得られた知見について報告する．

2 推薦話題とは

本研究の目的は，雑多なオンライン上の情報から，ユーザに対して推薦できる話題を提示するシステムを構築する事にある．推薦話題とは，情報選択の指針となるような，注目されていると考えられる情報とする．

さらに，ユーザにとって，目的があまり定まっていないあいまい検索を行っている状態では，何の手がかりも無しに目的を明確化し情報検索を行うことは困難であり，仮の目的や指針であっても，手がかりがある方が，より具体的な情報検索が出来るという報告がある [1]．我々が提示したい推薦話題とは，そのような目的の曖昧なユーザに対して，必ずしもユーザの要求する情報と一致するものではないが，検索のきっかけとなるような情報とする．

推薦話題となる情報には，色々な種類がある．知的ニュースリーダ HISHO[2] に推薦機能を付加させた小作らの研究では，推薦話題となる情報を，以下のように3つに分類している [3]．

- プロファイルにあった情報
- 先端的な情報
- 一般性のある情報

目的が曖昧なユーザに対して，検索のきっかけとなるような情報を提示するためには，多くの人に容易に受け入れられるような一般性のある情報をする事が必要であると考える．よって，この3種類の中で，本研究で目標とした推薦話題とは，一般性のある情報である．

一般性のある情報を推薦するにあたり，その知識源として新聞記事を用いる．新聞は多くの人に読んでもらうために，一般性のある情報を多く集めていると言える．だが，新聞の情報量は膨大であり，その中からさらに推薦話題を抽出する必要がある．

3 新聞からの抽出

膨大な新聞記事の中から，特に推薦できる話題を取り出すために，キーワード (単語列) のの出現情報を用いて，推薦話題と関係するような単語列を抽出する．ここで，推薦話題と関係する単語とはどのようなものであるかを考える．

新聞は，時間と紙面（社会面，スポーツ面など）という軸で分類・整理された文書の集まりである．この軸を元に，新聞に出現する単語を，以下の様に分類した．

一般語とは，その名の通り新聞記事中に一般に用いられる語で，紙面や時間で出現頻度があまり変化しない語を指す．例として「日本」や「政治」などが挙げ

† Masahiro HAMASAKI (masa-ha@is.aist-nara.ac.jp)
†† Hiromi Itoh OZAKU (hiromi-i@is.aist-nara.ac.jp)
† Yasuyuki KONO (kono@is.aist-nara.ac.jp)
† Masatsugu KIDODE (kidode@is.aist-nara.ac.jp)
Graduated School of Information Science, Nara Institute of Science and Technology (†)
Communications Research Laboratory (††)

表 1: 新聞中の単語の分類

一般語	時間・紙面が出現頻度に影響を与えない語
分野語	時間は出現頻度に影響を与えないが，紙面の影響がある語
話題語	時間・紙面が出現頻度に影響を与える語

られる．これらの単語は，話題という観点から見た場合，特徴が無く，話題を特定出来ない．

分野語とは，その分野特有の語であり，その分野では非常に一般的に用いられる語を指す．例えばスポーツ面における「野球」や「サッカー」などがあげられる．これも同様に何らかの話題を示すには，概念的に広すぎる単語である．だが，話題を補足的に説明する単語ではある．

話題語とは，何らかの話題に強く関係する単語であり，その出現パターンに何らかの特徴があるものである．例えば，1999 年では NATO によるユーゴスラビアの空爆が大きな話題となったが，「NATO」や「ユーゴスラビア」といった語は，空爆が行われた日の前後に多く現れ，普段はあまり現れない．我々が抽出したいと考える「推薦話題」と関係する単語列は，この話題語の中でも，特に偏りの大きいものであると考えられる．

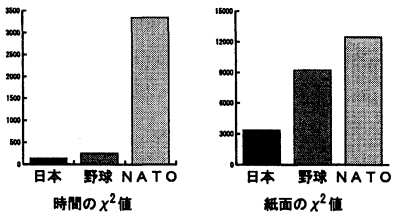

図 1: χ^2 値の比較

図 1 は，「日本」(一般語)，「野球」(分野語)，「NATO」(話題語) の 3 つの単語の，χ^2 値を示したものである (χ^2 値については次章参照)．χ^2 値は偏りの大きさを示しており，「NATO」という話題語が極めて偏りが大きいことがわかる．特に時間軸での χ^2 値は顕著である．

4 予備実験

我々は，出現に偏りを持つ単語を見つける事が，話題語を見つける事に繋がると考えた．よって，出現に大きな偏りのある単語を見つけ出すことによって，推薦話題に関係する語の抽出を行う．しかしながら，抽出された話題が推薦話題であるかどうか確かめる方法はない．

本来ならば，データとして用いた毎日新聞の 1999 年における 10 大ニュースの記事を用いるのが理想的であったが，存在しなかったため，今回は重要文抽出として有効な手法である有効な χ^2 法 [?] を用いて，仮想的な解を求める．そこで，仮想的な解を求めるための予備実験を行う．

仮想的な解を得るために，χ^2 法を通年データに適用し，各月ごとの推薦話題を抽出する．母集団サイズが定義されてる場合，χ^2 法を用いる事によって単語の出現の偏りを抽出することはできる．

一年間の新聞記事データから，推薦話題を抽出する処理の流れを図 2 に示す．

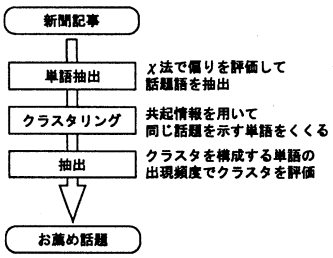

図 2: 処理の流れ

新聞記事　対象とするデータは 1999 年の毎日新聞の全紙面の記事である．自然文である新聞記事を，茶筅 [5] で形態素解析し，名詞と未知語のみを利用した．

単語抽出　χ^2 検定の χ^2 値は，出現頻度の偏りを示す指標として用いることが出来る．χ^2 値は以下の式によって求めることが出来る．

$$\chi_i^2 = \sum_{j=1}^{n} \frac{(x_{ij} - m_{ij})^2}{m_{ij}} \tag{1}$$

$$m_{ij} = \frac{\sum_{j=1}^{n} x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}} \times \sum_{i=1}^{m} x_{ij}$$

m : 異なり単語数
n : その月の日数
x_{ik} : 単語 i の j 日における頻度
m_{ik} : 単語 i の j 日における理論度数

χ^2 法によって単語の出現頻度の一ヶ月単位の偏りを求めた．ある一ヶ月間における単語の偏りは，一日の単語の χ^2 値の和で求める．つまり，その期間における出現頻度と理論度数とのずれを用いて，偏りを求める．

$$\chi^2_{im} = \sum_{j=F}^{L} \frac{(x_{ij} - m_{ij})^2}{m_{ij}} \qquad (2)$$

χ^2_{im} : 単語 i の m 月における χ^2 値
mF : m 月の最初の日
mL : m 月の最後の日
x_{ij} : 単語 i の j 日における頻度
m_{ij} : 単語 i の j 日における理論度数

(2) 式で求めた χ^2 値によって，各月に偏って現れる単語を発見した．そして，χ^2 値でソートを行い，各月ごとに上位 100 個の単語を抽出した．

クラスタリング　単語クラスタを作成するために，記事を対象とした共起情報を元に，相関ルールを用いる．相関ルールとは単語 A が現れたときに単語 B が現れる確率を指す．相関ルールが互いに 0.5 以上の単語を同士をクラスタ化する．

抽出　単語クラスタを構成する単語の出現頻度の和を，その単語クラスタのスコアとする．この際，クラスタを形成しなかったもの，つまり相関ルールが互いに 0.5 以上である単語が，1 つも無かった単語は評価から外した．

4.1 抽出結果

一年間のデータから χ^2 値を元に評価をつけ，各月で最も評価の高かった単語クラスタを表 2 に，それぞれの単語クラスタが示す話題を表 3 に示す．

8 月の「来春の国公立入試」は記事中に入試情報が書かれており，同じ単語が大量に現れるため，抽出されてしまった．内容がデータであるため，単語数が極めて多くなり検出されてしまった．読売新聞*，共同通信社†による 1999 年の 10 大ニュースと比較すると，ほぼ，どれも推薦話題としてよい記事と考えられた．

* http://www.yomiuri.co.jp/
† http://www.kyodo.co.jp/

表 2: 発見された単語クラスタ

1 月	単一，円，通過，ユーロ，ドル，欧州
2 月	八木，オリンピック，五輪，IOC，‥
3 月	脳死，判定，脳波，手術，心臓，臓器，‥
4 月	ユーゴ，難民，空爆，コソボ，NATO，‥
5 月	ユーゴ，コソボ，ユーゴスラビア，‥
6 月	部隊，ユーゴ，ミロシェビッチ，撤退，‥
7 月	日の丸，法制，歌詞，国旗，天皇，‥
8 月	前期，工，その他，医，後期，若干
9 月	部隊，治安，ディリ，インドネシア，‥
10 月	東海，茨城，臨界，原子力，燃料，‥
11 月	保険，介護
12 月	燃料，MOX，プルサーマル，‥

表 3: 単語クラスタが示す記事タイトル

1 月	欧州通貨統合，スタート ユーロ交換レート決定
2 月	長野五輪招致買収疑惑
3 月	国内初の脳死移植
4 月	ＮＡＴＯが空爆を実施
5 月	ＮＡＴＯ，ユーゴ全域で空爆を続行
6 月	コソボ停戦合意へ
7 月	国旗・国歌法案，衆院委で可決
8 月	来春の国公立入試
9 月	東ティモール独立運動
10 月	東海村臨界事故
11 月	保険介護制度の制定
12 月	プルサーマル実施大幅延期

5 推薦話題の抽出実験

5.1 紙面データを用いた抽出

前節の予備実験で，通年データを利用し，各月の推薦話題を χ^2 法を用いて抽出した．その結果を回答データとし，次に通年データではなく，もっと短期間の情報のみで推薦話題の抽出を行う．

本研究は，新聞記事を知識源として，推薦話題を抽出することを目標としている．新聞は毎日発行されるため，母集団サイズは変動的である．また，推薦話題となる情報は，ある程度，鮮度が無くてはならなため，過去一年間の中から推薦話題を取り出していたのでは意味が無い．よって，短期間のデータのみを用いて，一年間のデータを用いた場合と同様の性能でお勧め話題の抽出が出来る必要がある．

短期間のデータを用いると，母集団サイズが小さく

なるため．χ^2 法などは性能が発揮されない．そこで，紙面や記事といった新聞特有のデータを利用する必要があると考える．

新聞は，注目すべき話題を取り上げる際，複数紙面で同時に取り上げるという傾向がある．この傾向を利用すれば，全体の出現情報を知らなくとも，瞬間的な変化を見るだけで，データが少なくなっても，問題なく抽出できるのではないかと考えた．

$$m_{ij} = \sum_{k=j-2}^{j} x_{ij} - \sum_{k=j-9}^{j-3} x_{ij} \qquad (3)$$

m_{ij} : 単語 i の j 日における紙面・記事・単語変化量
x_{ik} : 単語 i の j 日における紙面・記事・単語数

紙面数や記事数，語数の変化量を用い，式 (3) によって，単語 i の j 日におけるスコアを求める．前後数日間のデータの平均の差分によって求める事で，単純な変化量だけでなく，話題の大きさも測れると考えた．

期間は前日 7 日間・後日 3 日間と設定する．これは予備実験の結果との比較より，前日を長く取る方が重要であるという結果が得られていたためである．

5.2 抽出結果

前日 7 日・後日 3 日の合計 10 日間のデータから算出した変化量で単語を抽出した．語数の変化量を用いたものの，各月で抽出した単語クラスタの中で評価値の最も高いものを表 4 に示す．

表 4: 変化量を用いて抽出した単語クラスタ

	語数	記事数	紙面数
1 月	×	×	×
2 月	○	×	×
3 月	◎	◎	◎
4 月	×	×	×
5 月	○	○	○
6 月	○	○	×
7 月	◎	×	×
8 月	△	×	×
9 月	○	×	×
10 月	◎	◎	◎
11 月	○	○	×
12 月	×	×	×

χ^2 法で見つけた解が
◎ : 1 番が同じ，○ : 3 番以内，△ : ある，× ない

語数の変化量を用いた抽出結果は，比較的良い結果が得られたが，記事数・紙面数を用いたものは良い結果を得られなかった．また，語数，記事数，紙面数の順に作られるクラスタ数が減少した．これは，多くの記事・紙面に現れる単語が，それだけ意味の強いものであり，各話題の代表的な語が集まったような状態になってしまったため，単語同士の相関ルールが小さくなってしまったのだと思われる．

6 結論

新聞記事を元に，推薦話題を検索する方法について実験を行った．結果，χ^2 法による年間データを用いた推薦話題抽出は良好な結果を示した．データ数を少なくした場合も，語数の変化量に注目した抽出法ならば，良い結果を得ることが出来たが，紙面および記事という新聞の特性に着目した抽出法では良い結果は得られなかった．

7 おわりに

単語の出現情報だけでは抽出できなかった重大ニュースに「警察官の不祥事相次ぐ」というものがあった．「相次ぐ」という情報自体は，ニュースとして取り上げられない．だが，複数のニュースが折り重なって「相次ぐ」という話題を生み出している．

「相次ぐ」というニュースは存在しないが，「相次いだ」という事実は話題として注目される．このような話題を抽出するには，単純に単語を追いかけるだけでは不十分であり，語数だけでなく，紙面や記事といった新聞特有の情報を利用することが解決の糸口になると考える．

謝辞

この論文を書くにあたって，通信総合研究所の内山将夫氏をはじめ，多くの方々にご協力して頂きました．心より感謝いたします．

参考文献

[1] 斉藤真理 他『ネットワーク空間における目標が不明確な検索行動の目標明確化と満足度』(インタラクション'98, 1998)

[2] 小作浩美 他『話題関連性に着目した知的ニュースリーダの提案』(電気関係学会関西支部連合大会, 1995)

[3] 小作浩美 他『対話型ネットニュースにおける「お薦め話題」の自動抽出』(2000 年情報学シンポジウム講演論文集, 2000)

[4] 渡邉靖彦 他『χ^2 法を用いた重要漢字の自動抽出と文書の自動分類』(日本機会学会, 1999)

[5] 松本裕治 他『日本語形態素解析システム「茶筅」version 2.2.0 使用説明書』(http://chasen.aist-nara.ac.jp/)

[6] 柳瀬隆史 他『メールマガジンを利用した注目ニュースの自動抽出』(情報処理学会研究報告 自然言語処理, 2000)