ネットニュース記事のタイプ分類

中野 貴之　村田 真樹　長尾 真
京都大学工学部 電子通信工学

1 はじめに

ネットニュースとは, インターネットにおける電子掲示板である。誰でも購読・投稿ができ, 投稿された記事はすべて世界中に配送されることが特徴である。このため, 投稿者による生の情報を得ることができる。また, 投稿に対して応答したり応答が得られたりという双方向性もある。このように, 新聞, 雑誌やテレビなど従来のメディアでは得られない情報を入手できる魅力がある。

しかし, ネットニュースには日々大量の記事が投稿され, 記事によって内容の質に大幅な差がある。このため, 必要な情報を入手するには読者が大量の記事から情報を選別しなければならない。つまり, 手間と時間を掛けないと欲しい情報が入手できないのが実情である。

より簡単に情報を入手する手法があればネットニュースがより有用なものとなる。例えば, 質問とそれに対する回答記事よりサマリ (まとめ) や FAQ (よくある質問と答え) を自動生成したり, 関連する議論記事から要約を自動生成したり, 情報提供の記事から核となる情報を抽出するという手法が考えられる。

これらの手法を実現するためには, 記事をタイプ別 (質問回答型・情報提供型・議論型など) に分類しておく必要がある。そこで, 本研究ではネットニュースの記事をタイプ別に分類することを目的とする。

2 ネットニュースの特徴

2.1 ネットニュースの長所と短所

ネットニュースの長所

記事が読者に届くまで一切編集が入らないので, 投稿者による生の情報が得られる。また, 投稿に対してフォローアップ記事を投稿できるため, 双方向的に情報をやりとりできる。このため, 従来のメディアでは得られない情報が入手できる。

ネットニュースの短所

日々雑多な情報を含む大量の記事が配送され, それらの記事は内容の質に大幅な差がある。このため, 大量の記事を読んで必要な部分を読者が選別しなければならない。つまり, 得られる情報量のわりに手間と時間がかかる。

2.2 既存の技術

以上のような短所をカバーするため, 既存のニュースリーダには記事をフィルタリングする機能が付加されているものがある。これにより, 各自で読みたい/読みたくない記事を自動選別することができる。

GNUS というニュースリーダには kill file という機能が, Gnus や slrn には Scoring という機能がある。これらの機能は, 指定したパターンによる文字列マッチングを記事ごとに行い, 条件に該当すればマークをつけて読む/読まないを区別するものである。不要記事を除去するには十分実用的であるが, 文字列マッチングの対象は主にヘッダ部分[1]であり, 本文は対象としないことが普通である。

2.3 先行研究

ニュースリーダによる記事のフィルタリング以外に, ネットニュースを効率的に利用するための研究がなされている。

佐藤らは, ニュース記事のダイジェスト作成 [1] や, ニュース記事の情報を整理して提供する手法 [2] を提案・開発した。これらは, ニュース記事という大量の記事から有用な情報を機械的に抜き出すものである。

井佐原らは, ネットニュースを読みやすくするため, スレッド (一連の記事群) で話題が転換した記事を推定する手法 [3] や, ニュースリーダが情報処理を行って関連ある記事群を抜き出す手法 [4] を提案・開発した。

3 記事のタイプ分類

3.1 記事のタイプ

本研究は, ネットニュースを利用しやすくする研究の基礎研究として, 記事をタイプ別に分類することを目的

[1] ネットニュースの記事には, 投稿者のメールアドレス (From), 記事の表題 (Subject), 投稿日時 (Date) などを情報が先頭に付加されており, これをヘッダと呼ぶ。

とする。タイプ別に分類すると，記事に対しタイプに適した情報処理を行うことができる。なお，ネットニュースは，ニュースグループが分野別に構築されているので，分野別の分類をする必要はほとんど必要ない。

記事のタイプは以下の通りである。かっこ内は，本研究で用いた略称である。

- 土台記事
  質問 (ques), 情報提供 (info), 各種提案 (prop), 各種募集 (want)
- 土台から発展した記事
  質問への回答 (ans), 情報の追加 (corr), 情報の訂正 (add), 討論・議論 (diss), 感想・返事 (impr), 単なる雑談 (chat)
- 文書
  サマリ・FAQ (summ), プログラムソース (src)
- 不要記事 (garb)
  テスト投稿，ねずみ講など

土台記事とは，一連のスレッドのきっかけとなる記事である。質問記事，情報提供の記事，提案記事，募集記事が該当する。

多くの場合，土台記事にフォローアップ記事が投稿されて話題が発展する。質問に対しては回答記事が，情報提供記事に対しては追加情報や訂正情報が，提案記事に対しては議論記事が投稿される。質問とその回答記事から議論に発展したり，情報追加記事からさらなる疑問が生じて便乗質問記事が投稿されることもある。また，単なる感想や茶々が入ることもあり，それが雑談に発展することもある。

文書とは，固定化した情報をまとめた記事である。回答記事が出尽くせば，サマリという形で質問と回答のまとめを投稿するのがよい習慣とされる。また，同じような質問がくり返し投稿された場合は，FAQ という形で質問と回答をまとめることもよく行われている。

最近は少なくなったが，プログラムのソースやバイナリコードが投稿されることもある。これらが議論に発展することは少ない。

テスト投稿やねずみ講募集は，一般的には情報のない記事とみなせるので，不要記事と分類している。

3.2 タイプ分類の利点

ニュース記事をタイプ別に分類すると，それぞれのタイプに適した情報処理を行うことができ，ネットニュースを情報源として活用することが容易になる。例えば，次のようなことが可能となると思われる。

- FAQ の自動生成
  質問とその回答記事を取り出すことで，サマリや FAQ の自動生成が可能になる。質問と回答記事からなるスレッドをとりだせばいいように思えるが，それでは，便乗質問や茶々入れなどの記事も含まれてしまう。そこで，質問とその回答記事を的確に取り出すことが必要になる。
- 議論の要約
  ネット上の議論はしばしば「フレーム」とよばれる不毛の議論になり，記事数が多いだけで読む意味がないものになりがちである。そこで，議論記事だけを取り出して要約を作成すれば，膨大な数の議論記事を読まなくて済むようになる。
- 情報の抽出
  情報提供記事の核となる情報だけを取り出せば，短時間で質の高い情報を得ることができる。

4 タイプ分類の手法

記事のタイプ分類は，C4.5[5] の決定木学習機能を用いた。あらかじめ，多量の記事について人手でタイプ分類を行い，これらを訓練事例およびテスト事例とした。

本研究では，属性を作成するために次のような情報を利用した。

- 引用元記事のタイプ
- 記事の行数
- ヘッダ中の特定の文字列パターン
- 本文中の特定の文字列パターン
- 5-gram 統計で得られた文字列

文字列パターンについての属性を作るため，ニュース記事をヘッダ・引用文[2]・孫引用文[3]・残りの本文の 4 つに分解した。この 4 部分それぞれについて，タイプの特徴を表す文字列パターンを人手で作成した。これら特定の文字列パターンが含まれる回数 (属性値は整数値であるから continuous [4]) や含まれるか否かの情報 (属性値は T または F) が属性となる。

[2] フォローアップ記事には，元記事の一部を引用することが多く，引用部分は引用符 (> や >> など) を行頭につけて区別される。

[3] 引用部分に含まれている引用文。

[4] C4.5 では，連続数値を属性値にとる場合 “continuous” という属性値を用いる。

また, 人手で作成した属性では見つけられないものを拾うため, 訓練事例において引用文・孫引用文・残りの本文それぞれの 5-gram 統計を取り, 上位 25 位までの文字列について, その文字列を含む回数と文字列を含むか否かという情報を属性として追加した。

属性の例を表 1にあげる。

表 1: 属性の例

	特徴	属性値
1	引用元記事が分類されたタイプ	ques, info, prop, want, ans, corr, add, diss, impr, chat, summ, garb
2	引用文の行数	連続 (continuous)
3	References があるか	continuous / T,F
4	Subject が「Re:」で始まるか	continuous / T,F
5	Subject が「Q:」で始まるか	continuous / T,F
6	引用文が「せんか」「しょうか」「ですか」を含むか	continuous / T,F
7	残りの本文が「教えて下さい」「教えてください」を含むか	continuous / T,F
8	残りの本文が「書きました」を含むか	continuous / T,F

表 1中の 1 は, 引用元記事のタイプが何であるかという属性である。引用元記事がない場合, 属性値は不定を表す「?」とする。質問記事には回答記事が, 議論記事には議論記事がフォローアップされることが多い。今回の実験では人手によるタイプ分類がつけられているため利用できる属性であり, 実際には元記事のタイプは不明である。しかし, 各記事を古いものから順に正しく解析できるようになると利用できる情報なので, 本研究では利用できることを仮定して用いることとした。2 は, 引用文の行数を用いた属性である。引用文の長さは記事のタイプに関連があると思われる。3 は, References ヘッダがあるか否かという属性である。ある記事のフォローアップ記事である場合は, フォローの元記事を特定する情報として References ヘッダが付加されている。4,5 は, Subject の文字列に関する属性である。フォローアップ記事の Subject は「〜の件」という意味の接頭辞「Re:」で始まることが多く, 質問記事の Subject はしばしば「Q:」という文字列で始まる。6,7,8 は, 引用文・孫引用文・残りの本文に関する属性である。「せんか」「しょうか」「ですか」という文字列は疑問文に含まれることが多く, これらが引用文にあれば該当記事は回答記事であると思われる。また, 記事の本文に「教えて下さい」「教えてください」という文字列を含めば質問記事であると思われる。8 は 5-gram 統計より拾った情報である。引用文の前に「○○さんは〜の記事で書きました」という文字列をつけることがあるが, そのうちの「書きました」を属性として拾っている。

5 タイプ分類の実験

実験は, fj.rec.rail の記事より訓練事例を 788 記事, テスト事例を 196 記事用意して行った。人手によるタイプ分類は訓練事例・テスト事例ともに行い, テスト事例を決定木で分類して人手による分類と比較した。

fj.rec.* という趣味の話題を扱うニュースグループを対象としたのは, これらのニュースグループでは質問と回答・情報提供・議論・茶々入れの記事がまんべんなく投稿されているからである。本研究では, 記事数の多い fj.rec.rail を対象とした。

属性の数は, 人手により作成した属性が 108 個, 5-gram 統計により作成した属性が 150 個である。

表 2は 788 記事の訓練事例において, 人手によるタイプ分類と決定木による分類を比較したものである。左から順に, 人手による分類, 再現率, 適合率, 総数, 決定木による分類結果を表す。

表 2を見ると, diss (議論) と add (情報追加) と間で分類の間違いが多く生じていたことが分かる。これは, 情報追加記事で「と思います」「じゃないでしょうか」など意見を表す文字列があることや, 議論記事で新たな事実を述べることがあることが理由であると思われる。これらのタイプへの分類精度を上げるためには, それぞれのタイプ固有の特徴を調査して属性として追加する必要がある。

prop (提案), summ (サマリ) はテスト事例での該当記事が 0 通であった。また, これらに分類された記事もなかった。want (募集), impr (感想), chat (雑談), garb (不要) は, 訓練事例での該当記事が少なく, テスト事例ではこれらに分類された記事はなかった。これは, 分類するための情報が不足していたと思われる。これらのタイプに分類されるためには, 該当記事の数を増やすことが必要である。

訓練事例の数による分類精度の変化を調べるため, 訓

表 2: タイプ分類の結果 (訓練記事は 788 記事)

分類	再現率	適合率	総数	分類結果の内訳											
				ques	info	prop	want	ans	corr	add	diss	impr	chat	summ	garb
ques	33%	37%	18	**6**	8	0	0	0	1	2	1	0	0	0	0
info	58%	33%	12	4	**7**	0	0	0	0	1	0	0	0	0	0
prop	–%	–%	0	0	0	**0**	0	0	0	0	0	0	0	0	0
want	–%	–%	0	0	0	0	**0**	0	0	0	0	0	0	0	0
ans	44%	34%	18	0	1	0	0	**8**	0	7	1	0	1	0	0
corr	20%	33%	10	0	0	0	0	0	**2**	3	3	0	2	0	0
add	53%	47%	63	1	0	0	0	8	2	**34**	17	0	1	0	0
diss	51%	51%	52	4	0	0	0	3	1	17	**27**	0	0	0	0
impr	0%	–%	6	1	0	0	0	0	0	3	2	**0**	0	0	0
chat	27%	42%	11	0	0	0	0	3	0	4	1	0	**3**	0	0
summ	0%	–%	1	0	0	0	0	1	0	0	0	0	0	**0**	0
garb	0%	–%	5	0	5	0	0	0	0	0	0	0	0	0	**0**
合計	44%	44%	196	16	21	0	0	23	6	71	52	0	7	0	0

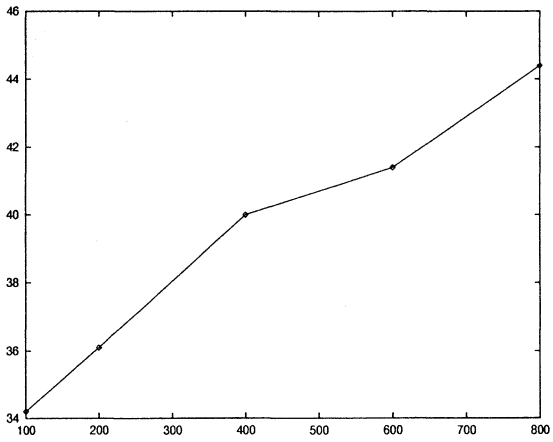

図 1: タイプ分類の結果 (縦軸: 精度, 横軸: 記事数)

練事例の記事数を 約100, 約200, 約400, 約600, 約800 と増やしつつ実験した。テスト事例の数は 196 記事で固定させた。図 1の通り, 訓練事例が増えると精度も向上していることが分かる。現時点の精度は, 訓練事例が約800 記事のとき 44% とよくないが, 訓練事例を増やすことで精度向上が期待できる。

6 おわりに

ネットニュースの記事のタイプ分類はいままでにない試みである。タイプ分類ができると, ニュース記事からの FAQ の自動生成や議論の要約などの研究に役立てることができる。現在は精度が低く, これらの研究に利用できないような状況であるが, 訓練事例を増やしたり属性を増やすことで精度を向上させ, タイプ分類を有用なものにしていきたい。

参考文献

[1] 佐藤理史, 佐藤円, ネットニュースグループ fj.wanted のダイジェスト自動生成, 言語処理学会論文誌, Vol. 3, No. 2, (1996).

[2] 佐藤円, 佐藤理史, ネットニュース記事群の自動パッケージ化, 情報処理学会論文誌, Vol. 38, No. 6, (1997).

[3] 内元清貫, 小作浩美, 井佐原均, 対話型ネットニュースグループにおける話題転換記事の推定, 言語処理学会第 3 回年次大会発表論文集, (1997), pp. 377–380.

[4] 井佐原均, 小作浩美, 内元清貫, 討論型ニュースグループを対象とする知的ニュースリーダーの開発, 自然言語処理研究会 119-3, (1997), pp. 13–18.

[5] J.Ross Quinlan, *C4.5 : Programs for Machine Langueage*, (Morgan Kaufmann Publishers, 1993), (邦訳: 『AI によるデータ解析』 J.R. キンラン, 古川康一監訳, トッパン, 1995 年).