ネットニュースのダイジェスト自動生成

佐藤理史　佐藤円
北陸先端科学技術大学院大学 情報科学研究科

1　はじめに

インターネット上の電子ニュース(以下、ネットニュースと記す)は、誰もが自由に記事を投稿することができ、それがそのまま広く配布されるという特徴を持った、新しいマスメディアである。情報発信者が限られている従来のマスメディア(新聞、ラジオ、テレビ)と比べ、情報発信の機会を広くに解放した点で、ネットニュースはマスメディアの新しい可能性を開いたが、逆に、情報発信者の拡大による情報の洪水と情報(テキスト)品質の多様化という新しい現象を引き起こしつつある。このため、求める情報を簡単に見つけることができなくなりつつある。

我々は、この問題を解決する方策として、ダイジェストに注目している[1, 2]。ダイジェストとは、元となる情報の特質をコンパクトにまとめて情報の種類別に整理したものであり、我々が大量の情報に接する際に効果的なナビゲーション機能を果たす。既存のダイジェストは、人手で編集されたものがほとんどであるが、はじめからオンラインテキストとして存在するネットニュースでは、このダイジェスト作成を完全に自動化することが可能である。我々は、既に、ネットニュースのダイジェスト自動生成の1つのプロトタイプとして、fj.meetingsのダイジェスト自動生成システムを作成し、実際に運用している[1]。

本研究では、その次のステップとして、fj.wantedのダイジェスト自動生成について検討した。fj.wantedは、fj.meetingsとは異なり、かなり多様な投稿者が、多様なテキスト品質の記事を投稿しており、fj.meetingsのダイジェスト自動生成で用いた手法とは異なった手法が必要となる。

2　ニュースグループ fj.wanted

ダイジェスト自動生成システムの作成に先立ち、94年9月8日から10月18日の間にfj.wantedに流れた記事231件(フォロー記事は除く)に対する調査を行なった。特に、その中の59件については、詳細な調査を行なった。ここでは、その調査結果を示す。

[1] http://www.jaist.ac.jp/~sato/nnad/home-j.html

2.1　主題上の特徴

fj.wantedの記事の主題(目的)は、「何かを探している(求めている)ということを伝える」というものである。これらの記事の主題は、おおよそ、以下のような2段の階層的カテゴリに分類することが可能である。

0. 探しています [人, 物, 情報]
   1. 譲って下さい [物]
   2. 譲ります [物]
   3. 貸して下さい [物]
   4. 募集します [人]
   5. 教えて下さい [情報]

ここで、かぎ括弧内は、求めるものの対象が何であるかを示している。以下では、これらのカテゴリを記事のカテゴリと呼ぶ。

2.2　文章上の特徴

(1) 多くの記事において、その記事の内容を端的に表す1文(以下、サマリ文と呼ぶ)が存在する。

調査した記事59件中、54件(91.5%)にサマリ文が存在した。

(2) fj.wantedの記事で用いられる文章構造のほとんどは、単刀直入型か背景説明型である。

単刀直入型と背景説明型とは、図1に示すような文章構造(文章の流れ)をさす。調査した記事59件中、単刀直入型は47件(79.8%)、背景説明型は10件(16.9%)であった。

2.3　表現上の特徴

(1) 典型的な「求む」の表現が、多くの記事で用いられる。

これらの表現のほとんどは、典型的な動詞群と文パターン(文末表現)によって構成されている。使われる典型的な動詞は、記事のカテゴリによって異なる。

(2) 機械による言語処理を難しくする、以下のような特徴が見られる。

a. 単刀直入型

1. (あいさつ・自己紹介)
2. 要約 (1 文 or 複数の文)
3. (詳細説明)

b. 背景説明型

1. (あいさつ・自己紹介)
2. 背景説明
3. 要約 (1 文 or 複数の文)
4. (詳細説明)

図 1: 単刀直入型と背景説明型

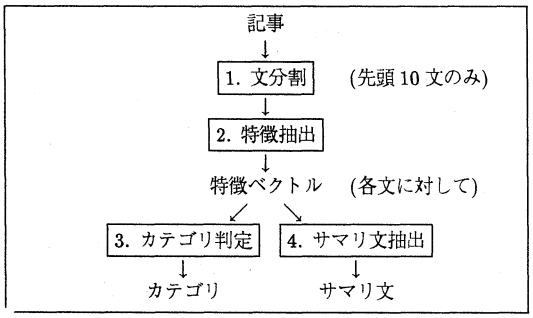

図 2: サマリ抽出の概要

- テキストが低品質である。(誤りが多い)
- 会話体が存在する。(ex. 「〜ってあるんでしょうか?」)
- 品目名として、かなり特殊な固有名詞が多数現れる。

3 サマリ抽出

上記の調査結果に基づき、fj.wanted の各記事から

1. 記事のカテゴリ
2. サマリ文

の 2 つを記事のサマリとして抽出することとし、それを行なうモジュールを作成した。その概要を図 2に示す。サマリ抽出は、以下の 4 つのステップによって行なう。

3.1 文分割

ニュース記事の本文を文毎に分割し、先頭の 10 文を取り出す。記事には、色々な表示上の工夫がされていることがあり、文を切り出すことはそれほど単純ではない。ここでは、各種のヒューリスティックを組み込んだ専用プログラムによって文を切り出す。

3.2 特徴抽出

各文に対して、42 個の特徴が存在するかどうかを調べ、特徴ベクトル (42bit のビット列) を作成する。ここでの「特徴」とは、例えば、

特徴 2 「譲って下さい」に類する表現が存在する

といったものであり、これは、表 1 に示すような表現が存在するかどうかを、文字列照合によって調べることによって判定する。42 個の特徴の概要を表 2に示す。

3.3 カテゴリ判定

カテゴリ判定では、35 個の規則を用いる。このうち、31 個の規則は、特徴 1–16, 18-32 に直接対応する規則で、これらの特徴の存在がそのままカテゴリの候補に対応する (表 2中の、「以降の処理での利用」欄を参照)。残りの 4 つの規則のうちの 3 つは、特徴 33–35 に対応する規則で、他の規則によって「探しています」というカテゴリが候補となっている場合にのみ使用する。残りの 1 つの規則 (特徴 17 に対応) は、他の規則によってカテゴリの候補が得られない場合にのみ使用する。

具体的には、以下の手順によってカテゴリを決定する。

1. カテゴリ候補リストを空とする。
2. 先頭の文の特徴ベクトル、次の文の特徴ベクトル、という順に調べていく。
   (a) 規則を適用し、その特徴ベクトルから得られる全てのカテゴリの候補をカテゴリ候補リストに追加する。
   (b) そのリストの中に、「譲って下さい/譲ります/貸して下さい/募集します」のいずれかが含まれる場合は、それを最終的なカテゴリとし、処理を終了する。
3. カテゴリ候補リストに「教えて下さい」が含まれている場合は、それを最終的なカテゴリとする。
4. カテゴリ候補リストに「探しています」が含まれている場合は、それを最終的なカテゴリとする。
5. カテゴリは不明とする。

表 1: 特徴 2 の表現

譲って	{欲しい | 下さい | もらいたい | 頂きたい}		
	もらえ 頂け	ると	
		ます ません ない	か でしょうか
お譲り	{欲しい | 下さい | もらいたい | 頂きたい}		
	頂け 願え	ると	
		ます ません ない	か でしょうか

表 2: 42 個の特徴

ID	特徴	以降の処理での利用
1	探しています	→ 探しています
2	譲って下さい	→ 譲って下さい
3	売って下さい	→ 譲って下さい
4	買って下さい	→ 譲ります
5	貸して下さい	→ 貸して下さい
6	教えて下さい	→ 教えて下さい
7	知らせて下さい	→ 教えて下さい
8	紹介して下さい	→ 教えて下さい
9	ダビングして下さい	→ 譲って下さい
10	譲ります	→ 譲ります
11	売ります	→ 譲ります
12	募集します	→ 募集します
13	知りたいの	→ 教えて下さい
14	買いたいの	→ 譲って下さい
15	欲しいの	→ 譲って下さい
16	求めています	→ 探しています
17	希望します	→ (譲って下さい)
18	存在しますか	→ 教えて下さい
19	はあるのでしょうか	→ 教えて下さい
20	はいらっしゃいますか	→ 探しています
21	可能でしょうか	→ 教えて下さい
22	知りませんか	→ 教えて下さい
23	質問です	→ 教えて下さい
24	譲って下さる	→ 譲って下さい
25	売って下さる	→ 譲って下さい
26	買って下さる	→ 譲ります
27	貸して下さる	→ 貸して下さい
28	ダビングして下さる	→ 譲って下さい
29	知っている	→ 教えて下さい
30	情報を待っています	→ 教えて下さい
31	情報をお持ちの	→ 教えて下さい
32	届きません	→ 譲って下さい
33	価格	→ 譲って下さい (+探)
34	1 万円	→ 譲って下さい (+探)
35	どのように/誰か/どこか	→ 教えて下さい (+探)
36	疑問文	
37	です文	→ skip
38	あいさつ	→ skip
39	自己紹介	→ skip
40	代理投稿	→ skip
41	境界線	→ skip
42	コメント	→ skip

3.4 サマリ文抽出

サマリ文の抽出では、以下の 2 つの方法を実装した。

a. 表現パターンによる方法
特徴 1–32 を持った最初の文をサマリ文とする。但し、その前の文が「疑問文 (特徴 36)」である場合は、その文をサマリ文とする。

b. 文章構造による方法
特徴 37–42 を持たない最初の文をサマリ文とする[2]。

4 実験

2 節での調査の対象とした 231 件の記事 (テストデータ, TD) と、それとは異なる 80 件の記事 (ブラインドデータ, BD) に対してサマリ抽出の実験を行なった。ここでは、その実験結果について述べる。

4.1 カテゴリ判定

カテゴリ判定の実験結果を以下に示す。

	TD	BD
全記事数	231	80
正しくカテゴリを判定	204	65
	88.3%	81.3%

どちらのデータに対しても、80%以上の高い正解率を示した。

4.2 サマリ文抽出

サマリ文抽出結果を以下に示す。

	TD	BD
全記事数	231	80
サマリ文が存在	198	69
	85.7%	86.3%
表現パターンによる	185	60
方法で正しく抽出	80.1%	75.0%
	(93.4%)	(87.0%)
文章構造による方法	156	44
で正しく抽出	67.5%	55.0%
	(78.8%)	(63.8%)

括弧内は、サマリ文が存在する場合の成功率を示す。2 つの方法のうち、表現パターンによる方法は、高い精度でサマリ文を正しく抽出できた。

[2] これは、単刀直入型に対応した方法である。

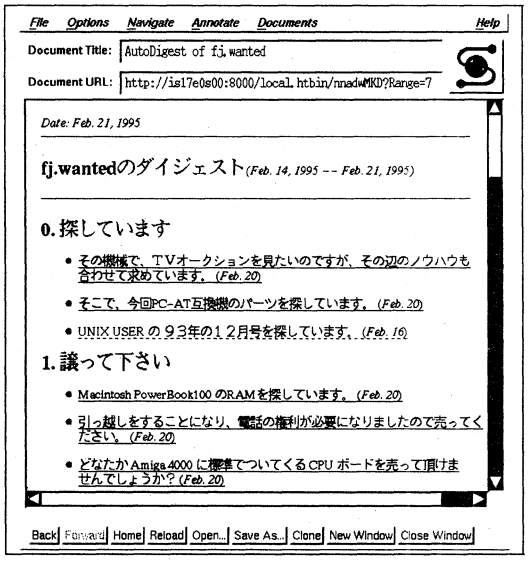

図 3: fj.wanted のダイジェスト

5 ダイジェスト生成システム

上記のサマリ抽出モジュールを用いて、fj.wanted のダイジェスト自動生成システムを試作した。システムは、現在、WWW において試験運用している[3]。図 3 にダイジェストの表示例を示す。

6 議論

(1) 本研究により、fj.wanted に関しても実用的なダイジェストの自動生成が可能であることが明らかになった。

当初、我々は、テキストと投稿者がかなり多様であるため、fj.wanted のダイジェスト自動生成は、難しいのではないかと考えていた。しかし、本研究の結果は、この予想を覆すものであった。fj.wanted のダイジェスト自動生成が可能であった最大の理由は、「fj.wanted の記事が、自分の求めるものが何であるかを読み手に伝えるという明確な目的を持った文章であり、そのような情報を伝達するために使われる文章構造と文章表現はかなり限定される」ということにあるだろう[4]。

(2) さらなる精度向上を目指すならば、サマリ文がない記事 (15%) のサマリ生成が必要となる。

サマリ文がない記事の多くは、照応や省略といった現象が現れているためにサマリ文となるような文が存在しない記事である。このため、照応、省略の処理が十分な精度で実現できなければ、適切なサマリ文を生成できないと考えられる。

(3) 投稿者によるサマリ作成は非現実的である。そのため、サマリの自動抽出は重要である。

サマリを自動生成するのではなく、あらかじめサマリを付けて投稿してもらうという解も存在する。しかし、現在の記事の Subject に書かれている情報からみて、我々は、それは非現実的だと考える。

(4) 本方法は、他の質問応答型のニュースグループの質問記事のダイジェストにも応用できると考えられる。

応答記事のサマリを含んだ形で、ダイジェスト (あるいは、FAQ) を自動生成することも考えられるが、その重要性は低いと考える。なぜならば、質問記事のリストを、質問のサマリとそれへの応答記事へのポインタという形で示すことができれば、十分にダイジェストの役割を果たすと考えられるからである。

(5) テキストの主題による分類は重要である。

テキストには、主題 (目的) と分野 (内容)[5] という 2 つの直交する分類が存在し、この 2 つが、いわば情報の取捨選択の縦糸と横糸となっている。このうち、分野による分類はいままで多くの研究があるが、主題による分類は、それほど注目されていなかった。この主題による分類も、分野による分類と同様に、求める情報に到達することを支援するナビゲーション機能の実現において、強力な道具となると考えられる。

[3] http://www.jaist.ac.jp/~sato/nnad/home-j.html

[4] 逆の側面から見れば、このように文章構造や文章表現が限られているため、我々は明確にその文章の主題 (目的) を理解することができるとも言えよう。

[5] 例えば、新聞記事では、経済、政治、スポーツといった分類が、この「分野による分類」に相当する。

参考文献

[1] 佐藤円. 電子ニュースにおけるダイジェスト機構の提案と実現. 修士論文, 北陸先端科学技術大学院大学情報科学研究科, 1994.

[2] 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. 情報処理学会第 49 回全国大会講演論文集, Vol.3, pp211-212, (3K-3), 1994.