レイアウトに着目したメールマガジンからの話題抽出方式

3U－1

大久保 雅且　杉崎 正之　森 大二郎　田中 一男

NTT ヒューマンインタフェース研究所

1. はじめに

近年、インターネットの発展に伴い、新聞社や出版社などが電子メールによるニュースの配信サービスを開始し[1]、個人の情報収集手段として定着してきた[2]。それぞれの内容や量は様々だが、いずれの場合でも、多くの記事や広告などから成るテキストを1つのメールとして配信している。受信者は、紙の新聞と同様に、メールを読むだけでなく、個人用データベースとして、検索、分類、関連付けによる傾向把握[3]などに活用できる。しかし、そのためには、紙の新聞からスクラップブックを作成するときに記事を切り抜くように、電子メールニュースでも、各記事を切出すことが重要なステップとなる。

1ページ内に複数の記事が掲載されている新聞では、各記事には「見出し」が必ずついている。それぞれの見出しは、人間の目につきやすいように、文字の大きさ、罫線、地紋、空間などをうまく工夫しており、各記事の意味内容を理解しなくても、レイアウト的な要素によって、見出しの位置や記事の範囲を一覧できる。

同様に、電子メールニュースでも、通常は見出しがあり、"-"などの記号のみの行（罫線に相当）、"●"などの記号で始まる行（地紋に相当し見出しの前などによく使用される）、文字のない行（空間に相当）、などによって、見出しの位置や記事の範囲が視覚的に把握できる（図1の本文テキスト参照）。本稿では、これらのレイアウト的な要素に着目して、各記事や広告を抽出する方式を提案する。

2. 話題抽出方式

2.1 行属性の付与

まず、各行に対して、(a)空行、(b)特定の記号のみの行、(c)特定の記号で始まる行、(d)URLのみからなる行、(e)上記以外の行、のいずれかの行属性を付与する。

(b)の特定の記号とは、1行全体をその記号のみとすることにより、見出しを強調したり、他の話題と区別するためによく用いられる記号で、例えば、"-"，"="，などである。(c)の特定の記号とは、"●"，"○"，"◆"，など、見出しの前によく使用される記号である。また、紙媒体の新聞と異なり、サイズの小さな電子メールニュースでは、詳細な内容や出典をURLで参照することが多いため、URLのみの行に対して属性(d)を与える。例えば、図1の1行目は、"-"のみからなる行なので行属性は(b)、2行目は"●"で始まっているので(c)とする。

2.2 見出し度の計算

次に、各行の見出し度を計算する。見出し度とは、その行の「見出しらしさ」を示す値で、前後数行の行属性によって決定する。見出し度を計算するためのルールの例を表1に示す。例えば、ルール1は「行属性が(c)の見出し度を2増加する」、ルール2は「行属性が(a)または

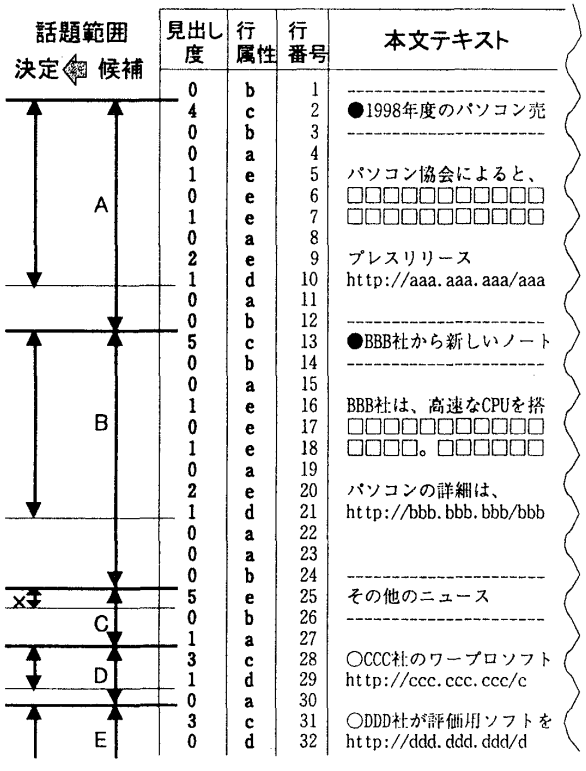

図1　本方式による話題抽出の例

Topic extraction from e-mail magazines based on their layout structure

Masaaki OHKUBO, Masayuki SUGIZAKI, Daijiro MORI, and Kazuo TANAKA

NTT Human Interface Laboratories

(b)の次の行は見出し度を1増加する」、ことを示している。ただし、行属性が(a)および(b)の行は見出しとはなり得ないので見出し度は0とする。これらのルールを図1に適用すると、例えば2行目は、ルール1, 2, 3に当てはまる。3行目はルール3に当てはまるが、行属性が(b)なので見出し度は0とする。

表1 見出し度計算ルール

ルール	行属性				見出し度増減値
	前々行	前行	対象行	後行	
1	-	-	(c)	-	+2
2	-	(a), (b)	-	-	+1
3	-	-	-	(a), (b), (d)	+1
4	(a)	(a), (b)	-	-	+1
…					

2.3 話題範囲の決定

行属性と見出し度に基づいて話題範囲を決定する。

まず、見出し度が閾値以上の行を話題の始まりとし、次の話題の始まりの直前の行までを話題範囲候補として抽出する。図1で、閾値を3とすれば、(A)2～12行、(B)13～24行、(C)25～27行、(D)28行～30行、(E)31行以降、の5つが話題範囲候補として切出される。

次に、各話題範囲候補の終わりから見ていって、空行や記号のみの行を除く。図1で、例えば(A)の話題範囲候補では、12行目から順に前へ見ていき、12行目と11行目の行属性がそれぞれ(b)と(a)なので、その2行を除く。その結果、各話題範囲候補は、(A)2～10行、(B)13～21行、(C)25行、(D)28行～29行、(E)31行以降となるが、このうち2行以上の、(A),(B),(D),(E)の4つを話題として抽出する。

3. 実験

以上のアルゴリズムを実現し、発行者の異なる電子メールニュース3種類、約1ヶ月分の71通について話題抽出の評価実験を行った。

話題抽出精度は、適合率と再現率によって評価する。記事や広告の始まりをうまく検出できたときに正解、それ以外を検出したときに誤検出、始まりを検出できなかったとき未検出とし、

適合率 = 正解／(正解＋誤検出)

再現率 = 正解／(正解＋未検出)

として求める。実験結果を表2に示す。

表2 実験結果

	総数	適合率	再現率
記事	3513	95.7%	100.0%
広告	340	91.9%	80.0%
合計	3853	95.4%	98.2%

4. 考察

記事に関しては、未検出はほとんどなかった(再現率100.0%)。誤検出については、特集記事やコラムのような長い記事にサブ見出しがついているとき、それを見出しと判断したことによる場合が多かった。これについては、例えば、見出しに続く本文が、「ところで」で始まっているなど、直前の話題との連続性を考慮することによって減らすことが可能である。

広告は、5行程度の中に様々な記号や空間を配置していることが多い。このため、例えば最初の行がすべて記号で始まっている場合には、見出しのみの話題、すなわち1行の話題として除かれた。また、空間を多くとっている場合には、見出しと判断されずに直前の記事や広告と結合された。これらによる未検出によって再現率が低くなった。前後数行の記号の割合などを判定に加えることによって、上記原因による未検出を回避できると考えられる。一方誤検出については、広告の途中の行を見出しと判断した場合が多かったが、記事の場合と同様の手法によって精度向上が可能と考えられる。

5. おわりに

メールニュースからの記事や広告の切出し方式を提案した。本方式は、そのレイアウトのみを利用しているため、発行者や内容に依存しない。さらに実験により、記事や広告を高精度に抽出できることを確認した。

参考文献

[1] 安井, "ネット・ビジネス情報源50", 日経マルチメディア, No.42, 1999.1, 50-65 (1999).

[2] 第12回インターネット利用に関する調査結果,情報通信総合研究所, 1998.
http://www.commerce.or.jp/minfo/enq/report12/

[3] 杉崎ほか, "情報分類を用いたトレンド・アウェアネスの支援", 情処研報97-DD-7, 23-30 (1997).