話題が混在するテキストからの話題セグメントの抽出方式

1V－5

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

インターネットの普及に伴い、様々な手段での情報取得が可能となった。特に、電子メールによるニュースなどの配信サービスは90%以上が利用しており、10種以上の利用者も約12%と、個人の情報収集手段として定着してきている[1]。これらのメールは、配信時に読むだけでなく、蓄積して個人用データベースとしての活用が期待される。しかし、1つのメールには、通常複数の話題や広告が混在している。このため、検索や分類、関連付けなどを有効に行うには、各記事への分解が必要である。本稿では、配信型メールからの話題セグメント抽出方式について述べる。なお、これらのメールは、内容、趣旨、発行形態など様々であるが、本稿では「メールマガジン」と呼ぶ。

2. メールマガジンからの話題セグメント抽出

メールマガジンは、見出し、本文、広告など様々な部分から構成されており、そのいくつかがまとまって1つの記事や広告（話題セグメントと呼ぶ）を構成している。各セグメントの境界を示す方法はメールマガジンごとに異なり、さらに1つのメールマガジンの中で異なる場合もある（図1参照）。

このように様々な形式が用いられていても、人間は適切に各セグメントを認識できる。これは、

(2.1) 情報伝達のための論理構造やレイアウト的手段が情報の送受信者間で暗黙のうちに共有されており、情報内容を知らなくてもレイアウトから情報構造を把握できる[2]。

(2.2) テキストの意味的な連続性／不連続性によって、各話題セグメントを判定できる。

という2つの理由からと考えられる。すなわち、視覚的な不連続性による話題の始まりや切れめ（の候補）の知覚と、意味的関係の理解の相乗効果によって話題セグメントを認識している。

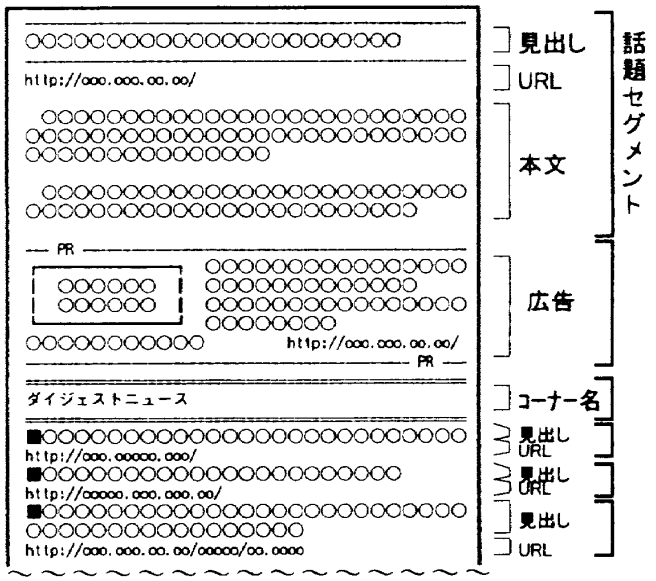

図1 メールマガジンの典型例とその構造

そこで本稿では、改行や記号などの視覚的要素に着目して、メールマガジンを細かい部分（これをフラグメントと呼ぶ）に分解し、言語的な関係に基づいてフラグメントを結合する、という2段階によって話題セグメントを抽出する方法を提案する。

3. 視覚的要素に基づく分解

メールは基本的には、一定の大きさの文字のみの使用を前提としている[注1]。このため、利用できる視覚効果は限られる。数種類のメールマガジンを解析した結果、各話題セグメントの区切りには、空行や記号が用いられていることがわかった。そこで、以下の条件によってメールを分解する。

(3.1) 空行、記号のみからなる行、1行がある程度長くそのほとんどが記号からなる行、をフラグメントの終わりとし、その次の行をフラグメントの始まりとする。

[注1] 近年、HTMLを用いたメールにより、多様なレイアウトや図の挿入などが可能になってきているが、まだ一般的ではない。

Topic Segmentation for Multitopic Text

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

NTT Human Interface Laboratories

メントの始まりとする。

(3.2) ●や★などの記号で始り(3.1)に該当しない行をフラグメントの始まりとし、その直前の行をフラグメントの終わりとする。

4. 言語的要素に基づく結合

次に、隣接するフラグメント間の言語的な類似性によってフラグメントを結合していく。3.で得られた各フラグメントを上から順に見ていき、

(4.1) 直後のフラグメントとの類似度が閾値を越える。

(4.2) 直後のフラグメントとの共通単語の割合が閾値を越える。

のいずれかが成り立つときにフラグメントを結合する。連続する2つのフラグメントF_kとF_{k+1}の類似度（ただし$k=1,2,\ldots,N-1$）は、文献[3]の*tf·idf*を基本とする以下の式によって求める。

$$\sum_i tf(i, \min(F_k, F_{k+1})) \cdot \log \frac{N}{ff(i)}$$

ここで、$\min(F_k,F_l)$は、2つのフラグメントF_kとF_lのうち短い方、$tf(i,F)$はフラグメントFに含まれる単語iの出現頻度、Nは総フラグメント数、$ff(i)$は単語iを含むフラグメント数である。また、共通単語の割合は$\min(F_k,F_{k+1})$を基準に計算する。

隣接セグメントの結合操作によって上記の各値が変わるので、結合が行われなくなるまで繰り返す。最終的に得られたフラグメントを話題セグメントとする。

5. 実験および考察

以上の基本方針を実現し、毎日配信されるメールマガジン6種類29通に対して実験を行った。メールマガジンには様々なコンテンツが含まれるが、そのうち記事（見出しを含む）と広告に対してセグメント抽出の精度を調べた。実験結果を表1に示す。

表1 実験結果

	ｾｸﾞﾒﾝﾄ数	正解数	正解率
広告	104	68	65.4%
記事	886	780	88.0%

3.で述べた視覚要素により、話題の切れめは100%抽出できた。一方、4.の結合の過程で、

(5.1) 1つの話題が複数に区切られたまま結合されなかった。

(5.2) 複数の話題が1つに結合された。

ために抽出誤りが起きた。(5.1)の誤りは、比較的長い記事が空行で区切られている場合や、「●お問合せはhttp://…へ」という行を1つのセグメントとした場合などが多かった。また、(5.2)の誤りは、プレゼントのお知らせや、同種のソフトのバージョンアップ情報など、同じような内容が連続しているときに1つの話題とみなす場合が多かった。これらに関しては、メールマガジンに頻出する特徴的な言い回しを考慮したり、長文を対象とした話題構造抽出[4]等の利用により精度の向上が可能と考えられる。

一方、広告では記号が多用されるために各行をフラグメントとしてしまい、また、それぞれの行での共通単語がほとんどないことから(5.1)の誤りとなることが多かった。逆に(5.2)の誤りは0であった。頻出する言い回しを考慮する以外に、言語としての文の連続性の判定を正確に行うことにより、広告に対する(5.1)の誤りは減らせると考えられる。

6. おわりに

メールマガジンを対象とした話題抽出方式について、その基本方針と実際のメールマガジンに対する実験について述べた。今後は、精度の向上と、特集記事やコラムなど比較的長いセグメントの構造化について検討を加えていく。

参考文献

[1] 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/

[2] 大久保ほか, "AV情報構造化技術とその情報要約への応用", 情処研報94-IM-15, 1994, pp.25-32.

[3] Salton,G., "Automatic Text Processing", Addison-Wesley Publishing, 1989, p.280.

[4] 竹下ほか, "テキストの概要把握支援のための話題構造抽出",情処論, Vol.37, No.11, 1996.