

# タグクラウドを用いたメールマガジンの視覚化

宮原 良一<sup>†</sup> 村上 晴美<sup>‡</sup>

大阪市立大学大学院創造都市研究科

## 1. はじめに

メールマガジンは、ブログ、RSS、SNS などと比べると比較的古い手段となっているが、簡易に使い、仕組みも分かりやすいことからインターネット利用者の 87.9%が利用している[1]。ユーザは興味を持てるものであれば積極的にメールマガジン購読に登録するが、それらのメールマガジンすべてに目を通してユーザは少なく、購読はしたがそのまま読まずに放っておくユーザも多い。したがって、未読メールが溜まり、どのメールを読んでよいか分からなくなるなどの問題が考えられる。そこで、本研究では、各メールマガジンに記載されている内容を短時間である程度把握できるように、各メールマガジンの特徴語を抽出し、一目で分かるように視覚化する手法を提案する。

## 2. 提案手法

本研究ではメールマガジンの内容を表す特徴語をうまく抽出し、視覚化するために、以下の過程から構成される手法を提案することにした。

### (1) 特徴語抽出

- 見出しと重み付け
- URL とメールアドレスの除去
- 形態素解析
- 複合語作成
- 不要語除去

### (2) タグクラウドを用いた視覚化

#### 2.1 特徴語抽出

##### (1) 見出しと重み付け

メールマガジンには、通常は見出しがあり、“—”などの記号のみの行(野線に相当)、“●”などの記号で始まる行(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、見出しの位置を把握できる[2]。本研究では、[2]を参考に各行の見出し度を計算し、見出し度の高い行に重み付けをする。

まず、各行に対して、(a)空行、(b)“—”など特定の記号のみの行、(c)“●”など特定の記号で始まる行、(d)URL が含まれている行、

(e)上記以外の行、のいずれかの行属性を付与する。次に、各行の前後数行の行属性から見出し度を計算する(表 1)。

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

上記の方法で、メールマガジン 1 通の各行の見出し度を計算した後、見出し度 3 の行は通常行の 4 倍、見出し度 4 の行は通常行の 5 倍、見出し度 5 の行は通常行の 6 倍の重み付けをする。また、メールマガジンの件名には通常行の 10 倍の重み付けをする。

##### (2) URL とメールアドレスの除去

メールマガジンでは、本文に URL が含まれている行がある。単純にメールマガジンを形態素解析して名詞だけ抽出すると URL に含まれている単語が頻度上位に来てしまうので、URL とメールアドレスを除去する。

##### (3) 形態素解析

メールマガジンには、ニュース系のもも多く、新語や流行語も多い。形態素解析システムには、新語や流行語に比較的対応している「Yahoo!デベロッパーネットワーク」の「日本語形態素解析 Web サービス」(以下、Yahoo 形態素解析)を利用し、メールマガジンから名詞を抽出する。

##### (4) 複合語作成

本来複合語として本文内に書かれている言葉が形態素解析されることによって、各単語に分割される。名詞が 2 語以上連続して続いている部分を、複合語として結合する。

##### (5) 不要語除去

数字のみの単語、メールマガジンの区切りを表す“——”や“——”などを除去するため、文字数が 30 文字以上の名詞を、不要語として除去した。さらに、単独で意味を持たないと思われる名詞を筆者の判断で抽出し、それらの名詞をより細かい品詞に分けて、特定の品詞を不要語とし、不要語リスト(計 542 語)を作成した。

