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[†]　　村上　晴美[‡]

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，簡易に使え，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いるユーザは少なく，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(罫線に相当)，“●”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，ニュース系のものも多く，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．

Visualization of e-Mail Magazine Using Tag Cloud
†Ryoichi Miyahara ‡Harumi Murakami
Graduate School for Creative Cities, Osaka City University

2.2 タグクラウドを用いた視覚化

本研究では，視覚化手法にタグクラウドを用いる．タグクラウドのタグとして，メールマガジン本文内に登場する特徴語を利用した．タグの並び順は，メールマガジン本文内で特徴語が出現する順にし，タグの個数は 70 個とした(図 1).

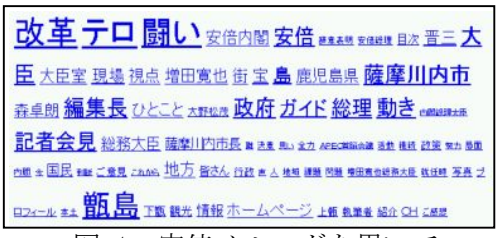

図 1　安倍メルマガを用いて
本手法で作成したタグクラウドの例

3. 実験

3.1 方法

被験者は，大阪市立大学全学共通教育「情報基礎」受講生 32 名で，「安倍内閣メールマガジン 第 46 号(2007/09/13 発行)」(以下，安倍メルマガ)に 14 名，IT ニュース系として「DragonField InternetNOW! Vol.314(2007 年 9 月 10 日発行)」(以下，IT メルマガ)に 18 名の被験者を割り当てた.

以下の①～③のアンケートを実施し，(a)～(f)のタグクラウドが，どの程度うまくメールマガジンの特徴を表していると思うかを(５：非常によく特徴を表している，４：よく特徴を表している，３：どちらともいえない，２：あまり特徴を表していない，１：まったく特徴を表していない)の 5 段階で評価させた.

①本手法を施したアルファベット順のタグクラウド(a)と単語出現順のタグクラウド(b)
②本手法を施したタグ個数 50 個(c)と 70 個(d)のタグクラウド
③メールマガジンの件名と本文から，URL とメールアドレスを除去した後，Yahoo 形態素解析で名詞のみ抽出し，アルファベット順にタグクラウドを表示したもの(以下，ベースライン)(e)と本手法を施したタグクラウド(f)

3.2 結果と考察

タグクラウドの並び順の比較結果を，表 2 に示す．なお，表内の”全体”とは，安倍メルマガと IT メルマガの両方の結果を合算したものである．また，表内の*，**はそれぞれ 5%，1%水準で有意であることを示す.

メールマガジンには，前後の文章が繋がっている読み物系のメールマガジン(以下，読み物系)もある．読み物系の場合，特徴語をアルファベット順に並べるよりも，メールマガジン本文内で出現する順に特徴語を並べた方が，本文の要約が分かる．また，前後の行に繋がりがないニュース系メールマガジンでも，同じ話題に関するニュースは同じ位置に多くあることから，特徴語が出現する順にした方が，似たような話題の特徴語が同じような位置に集まる．これらの結果から，アルファベット順より単語出現順の方が，適しているのではないかと考える.

表 2　タグクラウドの並び順の比較

並び順	全体*	安倍メルマガ	IT メルマガ*
アルファベット順(a)	2.91	2.57	3.17
出　現　順　(b)	3.44	3.07	3.72

タグクラウドのタグの個数の比較結果を，表 3 に示す．これらの結果やインタビュー調査の結果から，一応 70 個のタグの個数のタグクラウドが，メールマガジンの特徴をうまく表しているという結果が出た．しかし，20 個程度のタグ個数の差よりも，サイズの大きいタグが自分の見やすい位置にあるかということが重要であるのではないかと考える.

表 3　タグクラウドのタグ個数の比較

タグ個数	全体*	安倍メルマガ	IT メルマガ*
50 個(c)	3.16	3.00	3.28
70 個(d)	3.58	3.36	3.67

提案手法とベースラインの比較結果を，表 4 に示す．これらの結果から，ベースラインと比べると，本手法の処理を施したことで，メールマガジンの特徴をうまく表せるようになったといえる.

表 4　ベースラインと提案手法の比較

	全体**	安倍メルマガ**	ITメルマガ**
ベースライン(e)	1.97	1.64	2.22
提 案 手 法 (f)	3.66	3.57	3.72

4. おわりに

本論文では，溜まってくる未読メールマガジンに対応するために，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した．評価実験を行った結果，ベースラインよりも，本手法の処理を施した方が有効だということが分かった．また，タグクラウドのタグの並び方に関する知見を得た.

参考文献

[1] 財団法人インターネット協会：インターネット白書 2006，2006
[2] 大久保雅且，杉崎正之，森大二郎，田中一男，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，情報処理学会第 58 回全国大会，3U-1，1999.