L-022

# スレッドフロート型掲示板における情報取得支援

## Support of Information Acquisition in Thread Float Type BBS

安積庸輔 * 鈴木輝彦 † 太原育夫 ‡
Yousuke Azumi Teruhiko Suzuki Ikuo Tahara

## 1 はじめに

現在，家庭用のコンピューターと，それを用いたインターネットの普及により電子文書が増加し，氾濫する情報の中から自分にとって有用な情報を取捨選択し取得することが必要となり，それについての研究も多くなされている [1]，[2].

そこで，本研究では電子掲示板でも特に，スレッドフロート型掲示板においてユーザーが情報を取得する手助けとなるようなシステムの構築を目指した.

## 2 スレッドフロート型掲示板と話題

スレッドフロート型掲示板 (以下 TFT-BBS) とはスレッドという纏りを主とする掲示板形式の一つである.

TFT-BBS は記事の投稿を行うのに，スレッドという投稿の集まりとなる元を作成する．スレッドには以降その話題に対する投稿が行われ，その投稿は並列に表示がされる．これは，ツリーを形成するタイプの掲示板に比べ，多人数が煩雑に議論を交わすのに向いている.

そのためにツリー型掲示板のスレッドには比較的話題にまとまりが見られるのに対して，TFT-BBS では情報の範囲が分散し，更に投稿量も多いためにユーザーにとって有益な情報を取得するにあたって困難がある.

そこで，TFT-BBS において，ユーザーが情報を取得する際，その支援を行うシステムを考えることにする．一般に TFT-BBS では投稿順に時系列的に投稿が表示される．これを “話題”ごとに関連を持たせた並べ替えを行い表示することを考える． ここで，本研究では “話題”を “ある投稿による話題の提起からその投稿に対する返信を繋いだ一連の議論の流れ” と定義することにする．これはツリー型掲示板における一ツリー分に相当するものと考えられる.

## 3 情報取得支援システム

### 3.1 投稿の分類

TFT-BBS における情報取得支援の方法として文書クラスタリングを用いた投稿の分類を試みる.

情報取得支援システムにおける処理の流れは以下のようになる.

1. スレッドの指定を受ける.
2. 指定されたスレッドの情報を取得する.
3. スレッド内の投稿を各々構造化して取り込む.
4. 投稿情報を基にクラスタリングを行う.
5. クラスタリングの結果より投稿をソート.
6. 出力

### 3.2 クラスタリング手法

まず，投稿本文に対しては一般文書クラスタリングに用いられる文書ベクトルによる類似度を測定する．投稿本文 $i$ に形態素解析 (mecab[3] を用いる) を行い，単語ごとの重みから文書ベクトル $d_i$ を作成する．文書ベクトルは以下の式により表わせる.

$$d_i = (w_{i1}, w_{i2}, \ldots, w_{iN}) \quad (1)$$

ここで $w_{i1}$ とはスレッド内に存在する単語 $t_1$ の重みであり，ここでは投稿内での出現回数とする．投稿 $i$ と $j$ の本文に関する類似度は余弦尺度として定義されるので以下の式

$$s(d_i, d_j) = \frac{d_i \cdot d_j}{||d_i|| \cdot ||d_j||} \quad (2)$$

で表すことができる． 更に，ここにスレッドフロート型掲示板におけるルールによる重みを付加する．まず文頭にアンカーが存在する場合には，その記事への返信であることが明示的であるため，十分に大きな値 $A$ を設定し，これを加算する．次に投稿番号が近い物に対し関連が強まるようにしたいため，投稿 $i$，$j$ の投稿番号をそれぞれ $n_i$，$n_j (n_i > n_j)$ としたとき，以下の式

$$t(i, j) = exp(\alpha(n_j - n_i)) \quad (3)$$

を先ほどの類似度に掛け合わせることを考える．以上のものをまとめ，

$$r(i, j) = (1 + t(i, j)) \cdot s(d_i, d_j) + A \quad (4)$$

を投稿 $i$，$j$ 間の類似度とし，クラスタリングを行う.

## 4 実験による評価

スレッドに対してシステムを適用した結果と，そのスレッドを人手にてソートをした正解データとを比較することにより，本手法の評価を行う.

*東京理科大学大学院 理工学研究科 情報科学専攻
†東京理科大学 理工学部 情報科学科
‡東京理科大学 理工学部 情報科学科

### 4.1 実験環境

実験に用いたスレッドは 2 ちゃんねる [4] 大学受験板の東京理科大学を第一志望にしている人のためのスレ part4 であり，各パラメータは以下の通りである．
スレッドの総投稿数：1000
減衰関数の係数：$\alpha$=0．10
接続閾値：0．15

### 4.2 実験結果

各投稿は正解と比較することによって以下のように分類することができる．

- 話題の起点となる投稿
  - 他の投稿に属さない
  - 誤って他の投稿に属している
- 他の投稿へ返信する投稿
  - 正解と同じ投稿の下に属している
  - 正解の投稿より下位の投稿に属している
  - 正解の投稿より上位の投稿に属している
  - 誤った投稿に属している
  - どの投稿にも属していない

この規則に沿って正解データとの比較を行った結果を表 1 に示す．

表 1: 実験結果

| | | 件数 |
|---|---|---|
| 起点となる投稿 | 属さない | 65 |
| | 属している | 20 |
| 返信となる投稿 | 正解に属している | 380 |
| | 下位投稿に属している | 32 |
| | 上位投稿に属している | 7 |
| | 誤った投稿に属している | 6 |
| | どの投稿にも属していない | 113 |

## 5 考察

実験結果より，話題の起点となる投稿については，正解通りどの投稿にも属さないものを，他の投稿へ返信する投稿については，正解通りの投稿に属するもの，正解投稿の下位投稿に属しているもの，正解投稿の上位投稿に属しているものを，話題に即したソートに成功しているとみなし，話題を構成する投稿の全体数からの割合をみると，77．69% という結果になった．本実験のパラメータ設定では，この数値を大きくすることよりも，誤ったソートを起こす割合を少なくすることに重点を置いた．これは，スレッド内の投稿に，ソートが行われていない場合よりも，誤ったソートが行われている場合のほうが，よりユーザーに違和感を与えるためである．誤ったソートが行われている件数は，話題の起点となる投稿について 20 件，他の投稿へ返信する投稿について 6 件であり，この割合は全体の 4．17% となる．

また，他の投稿へ返信する投稿について，失敗となるどの投稿にも属していないもののうちこの要因が予測できるものを下表にまとめる．

表 2: 予想される原因

| | 件数 |
|---|---|
| 直前の投稿への返信による情報不足 | 45 |
| 投稿番号が遠いため | 22 |
| アンカーの記述ミス | 10 |

このことから，この失敗の最大の原因はクラスタリングに対する情報の不足と考えられる．これについて，クラスタリングに以下の指標を加えることによって改善が見込めるのではないかと思われる．

- 投稿者名,ID による人物の同定
- アンカーに近い記述や, 投稿番号での考慮
- TF-IDF 等のキーワード抽出手法を用いることやクラスタリング手法の改善
- 属した投稿, 属された投稿に対する考慮

## 6 おわりに

現状のシステムでは，話題を高い精度でソートを行えているとは言い難い．特に，アンカーの指標などの無い状態では投稿番号差が 10 程度のものまでしか正しいソートを見込めなかった．係数の設定を緩和すると，より目立つ誤った接続が起こりやすくなるためである．この点については，前項で挙げたことからクラスタリングに際し別観点の特徴を考慮する必要があるといえる．

しかし，ユーザーが掲示板の内容を読み取りやすくするという目標については，有効であると考えられる．特に，離れた記事に返信を行う場合にはアンカーが使われることが多いが，このアンカーをつけた投稿が現れることにより，同一の話題が投稿番号の離れた位置で再燃する場合が見られる．この様な場合において，アンカーを通して二つの同一の話題群を一纏めにして表示できる点で，このシステムの情報取得支援としての立場での有用性が伺える．

## 参考文献

[1] 松尾 豊，大沢 幸生，石塚 満，“電子掲示板における会話からのハイライト部分の抽出，”第 46 回人工知能基礎論研究会，2002.

[2] 岸田 和明，“文書クラスタリングの技法：文献レビュー Techniques of Document Clustering: A Review”，Library and Information Science,No.49, pp.33-75，2003.

[3] MeCab（和布蕪），http://mecab.sourceforge.net/

[4] 2 ちゃんねる，http://www.2ch.net/