ユーザに適合した匿名掲示板スレッドの提示システム

木幡 徹† 東 基衛†

†早稲田大学大学院理工学研究科経営システム工学専攻

1. 研究の背景と目的

WWW(World Wide Web)が急速に普及し、人々が容易に情報を発信できるようになった。そのため、WWW上にはレビュー記事や個人の日記、blog、新聞記事のように様々な文書が混在している。情報発信先サイトの一つとして「匿名掲示板」注目されている。その代表的な例が「2ちゃんねる」である。

この2ちゃんねるは、日本屈指のアクセス数を持つ電子掲示板群である。掲示板は、カテゴリと呼ばれる大きな分野単位で区切られており、さらに分野ごとに多くのジャンルの掲示板が存在している。また、一番の特徴として投稿時に必ずしも名前を書かなくて良いという点がある。匿名投稿にはデメリットが多い一方で、匿名であるが故の投稿の中には情報媒体として有用な内容も多々含まれている。

匿名掲示板の投稿の量は膨大であるにもかかわらず投稿内容は玉石混淆であるため、投稿が有用であるかどうかは実際に内容を読まないと判定できない。そこで、匿名掲示板の投稿に対し、ユーザに適合したスレッドを提示する事を本研究の目的とする。

2. 従来研究

匿名掲示板解析に関して、松村らは様々な因子を設定し、共分散構造分析を行い、因果モデルを構築し、それを元に2ちゃんねるが盛り上がるメカニズムを解き明かしている[1]。しかし、システムには落としこんでおらず、モデルの構築に留まっている。

次に、深谷らはユーザの行動履歴と掲示板のスレッドの特徴をベクトル化し、比較することでユーザの興味があるスレッドを推薦している[2]。しかし、匿名掲示板においてこのアルゴリズムを適用することを考えると、投稿履歴等のデータが使えないという問題がある。

3. 研究内容

3.1 提案システム要件

提案システムの要件は従来の掲示板解析手法と異なる、匿名掲示板の特徴を捉え、それに対応したアプローチが求められる。

(1)本研究では投稿者の投稿履歴や投稿者の性質などの情報が使えない。なぜなら、研究対象が匿名掲示板だからである。

(2)本研究では悪しき投稿、無用な投稿への対応を考えなくてはならない。匿名掲示板は他の掲示板と異なり、そのような投稿が圧倒的に多いからである。

なお、今回は研究対象を2ちゃんねるに限定して研究を行った。

3.2 研究アプローチ

3.2.1 特徴単語ベクトルの生成

まず、各スレッドをベクトル化する。これには、2ちゃんねるから提供されているキーワード機能を用いる事にする。このキーワード機能は、2ちゃんねるの各スレッドを表す代表的な単語をスレッドごとに上位7件ずつ提示する物である。

得られた単語7件をスレッドTの特徴を表す単語ベクトル $\vec{F}_{Tword}$ とし、(1)式で示す。

$$\vec{F}_{Tword} = (W_{ij1}, W_{ij2}, W_{ij3} \ldots W_{ij7}) \quad (1)$$

3.2.2 属性情報ベクトルの生成

次に、匿名掲示板の特徴を表している幾つかの要素を用い、各スレッドごとに属性情報ベクトルを生成する。具体的には以下のような指標がある。

C…メッセージごとに交わされる議論の量を測る指標。
A…スレッドの盛り上がりを投稿数によって測る指標。
I…参加者同士のインタラクションの程度を測る指標。
S…スレッドの盛り上がる早さを測る指標。
V…スレッドの2ちゃんねるらしさを測る指標。
AA…スレッドにアスキーアートがどれくらい使われているかを測る指標。

また、2ちゃんねるの利用アンケートをとった結果、情報を収集する目的にしている人が多い一方で、単なる息抜きの為に閲覧している人が一番多かった。

そこで、2ちゃんねるの属性情報とアンケートの結果から、スレッドを議論深化スレッド、議論発散スレッド、息抜きスレッドと分類し、それぞれ先の指標を用いて計算できるようにする。分類法を以下に述べる。

まず、A、Sが高いスレッドは投稿数が多く投稿のペースも速いので議論発散傾向とする。次に、C、Iが高いスレッドは量のあるメッセージが交わされ、参加者間のインタラクションも活発なので議論深化傾向とする。最後に、V、AAが高いスレッドは、2ちゃんねる語やアスキーアートを用いた定型的な2ちゃんねるならではの表現なので息抜き傾向とする。

User Adaptive Presentation System for Anonymous Bulletin Board Threads

Toru KOHATA†, Motoei AZUMA†

† Dept. of IMSE, Graduate School of Sci. & Eng., Waseda Univ.

それぞれの指標の求め方だが、ある掲示板 B について、A、S、C、I、V、AA の値を各スレッドごとに出し、平均を 0、分散を 1 に標準化し、議論発散傾向、議論深化傾向、息抜き傾向の3つに対応した指標をそれぞれ取り、2 次元のベクトルを生成する。

これらのベクトルの大きさが大きければ、そのスレッドが各傾向に属しているということが言える。

3.2.3 ユーザベクトル生成

ユーザは初め掲示板とキーワードを入力し、そのキーワードが含まれるスレッド一覧を得る。その中から興味がありそうなスレッドを幾つか選択し、それらのスレッドの平均ベクトルをユーザベクトルとして得る。

3.2.4 無用投稿の判定

辞書構築により、無用な投稿の判定を行う。これは、2ちゃんねるの無用な投稿は単語の意味に依存しないものが多く、文書解析による判定が難しいからである。

今回は、無用な投稿の数が閾値を超えたスレッドを無用スレッドと判定する。

3.2.5 適合スレッド判定

先ほど生成されたユーザベクトルと掲示板のその他のスレッドを比較し、近いものを順にユーザに提示していく。この際、(2)式の類似度 S を用いる。

$$S(\vec{F}_{user}, \vec{F}_{T\,word}) = \frac{\vec{F}_{user} \cdot \vec{F}_{Tword}}{|\vec{F}_{user}||\vec{F}_{Tword}|} \quad (2)$$

また、属性情報ベクトルの大きさによるソートも可能で、各ベクトルの大きさが大きければにスレッドの性質を色濃く反映しているものと言える。

また、無用投稿が閾値を越えたスレッドは無用投稿として処理され、ユーザに提示されない。

4. 提案システム

提案システム概要図を図に示す。input、output 及び各モジュールの説明をする。

input

最初にユーザに閲覧する掲示板を選択させる。そして選択された掲示板のスレッド郡から興味ベクトルを選択させ、ユーザベクトル生成に用いる。

特徴単語ベクトル生成モジュール

選択された掲示板を input とし、各スレッドごとにスレッドを特徴付ける単語ベクトルを生成し、ユーザベクトル生成モジュールに渡す。

属性情報ベクトル生成モジュール

選択された掲示板を input とし、各スレッドごとに $|\vec{F}_{hr}|, |\vec{F}_{st}|, |\vec{F}_{ir}|$ の値を出し、属性情報ベクトルを生成し、それを適合スレッド判定モジュールに渡す。

無用スレッド判定モジュール

選択された掲示板及び無用語辞書を input とし、選択された掲示板のスレッド内の無用な投稿を判定し、適合スレッドから除外する。

ユーザベクトル生成モジュール

キーワード入力により提示されたスレッド郡からユーザが興味のありそうなスレッドを選択し、そのスレッド郡の特徴単語ベクトルの平均値を取り、ユーザベクトルとする。

適合スレッド判定モジュール

得られたユーザベクトルとスレッドベクトルの比較を行い、類似度 S が近いものから順に提示する。属性情報ベクトルを用いた、スレッドの性質を選択することによるスレッドのソートも可能である。無用スレッドは切り捨てられる。

output

ユーザに適合したスレッド(郡)が得られる。

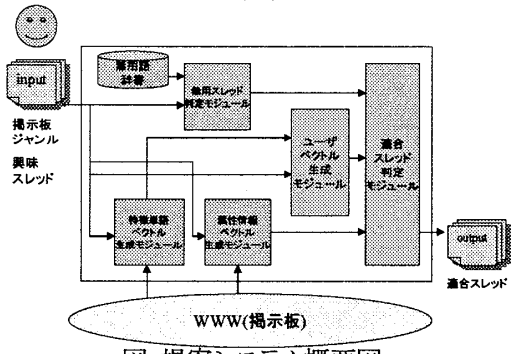

図 提案システム概要図

5. プロトタイプ・評価実験

プロトタイプには実装言語に PHP5.2.5(Hypertext Processor)、Web サーバに Apache2 を用いた。

プロトタイプを用いた評価実験として、提示された上位10件のスレッドがユーザに適合しているか、と言う評価を 5 段階評価で行った。結果、従来手法であるキーワード検索よりも良い値が得られ、本システムの有効性が示せた。

6. 今後の課題

今後の課題として、スレッドのベクトル化においてより多くの要素を取り入れる事、他の掲示板への応用、input の多様化とそれに対応したアプローチの考慮などが挙げられる。

参考文献

[1]松村ら:2 ちゃんねるが盛り上がるダイナミズム、情報処理学会論文誌 Mar.2004 Vol.45 No.3,pp1053-1061

[2]深谷ら:電子掲示板における行動履歴を用いたユーザにとって興味あるスレッドの推薦手法、電気情報通信学会研究報告 Vol.HIP2006-96, pp149-154 (2006)