NIFTY-Serve におけるフォーラムデータの分析

Analyzing NIFTY-Serve forum-data

成澤　克麻（Katsuma NARISAWA）[1]・比戸　将平（Shohei HIDO）[2]・海野 裕也（Yuya UNNO）[2]

松井 くにお（Kunio MATSUI）[3]・鈴木 隆一（Ryuichi SUZUKI）[3]・田代 光輝（Mitsuteru TASHIRO）[3]・

丸山 宏（Hiroshi MARUYAMA）[4]

[1]東北大学情報科学研究科　前期博士課程　・　[2]株式会社 Preferred Infrastructure

[3]ニフティ株式会社　・　[4]統計数理研究所

［Abstract］

We present our attempt to analyze a long time-span forum data to understand the dynamics of online communities. The data analyzed were salvaged from the archive that Nifty Inc., an online service, collected during 1987 - 2006. We identify various statistical parameters that show the level of activities for a given forum or thread, propose a metric of the popularity based on them, and show an experimental result of machine learning to determine the popularity.

［キーワード］

電子コミュニティ、電子掲示板、パソコン通信、可視化、盛り上がり

１．はじめに

本稿では、NIFTY-Serve データを対象としたコミュニティ分析について述べる。NIFTY-Serve とは 1987 年から 2006 年までニフティが運営していたパソコン通信サービスで、会員同士でやりとりできる電子メールや掲示板、またフォーラムと呼ばれるコミュニティサービスを提供していた。フォーラムとはある特定の趣味や話題に興味のある会員同士がコミュニケーションを取ることのできるサービスで、「スポーツ」「恋愛相談」のようなフォーラムが 1000 個以上存在していた。各フォーラムには「電子会議」「リアルタイム会議（チャット）」などといった機能があり、特に「電子会議」はテーマ別に作成できる掲示板であり、意見交換のスタイルは 1 つの発言に対して複数人がリプライ可能な、現在のインターネットにおけるツリー型掲示板のような形態であった。このデータの特徴としては、オンラインコミュニティのデータとしては比較的古くインターネット普及以前からのものが含まれること、完全会員制かつ基本的に公開されたユーザーID と紐付けられたのサービスであったため 1 つ 1 つの書き込みが誰によって書き込まれたか、またあるユーザーがどんな書き込みを行ったのかを 10 年以上の期間で追跡できること、などが挙げられる。

オンラインコミュニティを対象とした分析はこれまでにも行われているが、分析の指標は定性的なものであり、専門家が主観的に判断するものであることが多い。そのような人手による分析では、定量的な観点からの分析が行いにくいこと、また大規模なデータを用いての分析が行いにくいことなどの欠点が挙げられる。

そこで本研究では、NIFTY-Serve のそれぞれのフォーラムにおける「電子会議」のデータを用いて、コミュニティの盛り上がりの要因を定量的に明らかにする。盛り上がるコミュニティと盛り上がらないコミュニティには何らかの違いがあると考えられ、それを明らかにすることができればコミュニティを運営する際に有益であると考えられる。本稿では、盛り上がり分析を行う前のステップとして、フォーラムデータに対して基本的な分析を行い、データに対しての理解を深める。具体的には、それぞれのフォーラムやコメントから様々な統計量を抽出し、それぞれの統計量がコミュニティのどんな特徴を表しているのか分析を行い、同時にコミュニティがもつ特徴そのものを考察した。次のステップとしては、具体的に盛り上がりを定義し、相関分析を行う。最後に、機械学習を用いた簡単な盛り上がり予測による盛り上がりに有効な素性を分析する実験についても述べる。

２．NIFTY-Serve フォーラムデータについて

NIFTY-Serve は、各フォーラム（例：「スポーツ」「恋愛相談」）の下に、複数の電子会議室（掲示板）（例：「1999

年夏甲子園」）を持つという構成になっている。過去のデータ整備プロジェクト[6]の成果により、実際のデータもそのようなディレクトリ構造になっており、電子会議室のデータは1つ1つ分割されてcsvファイルで記述されている。csvファイルの中身は、1つの行毎に「(電子会議室内での）発言ID」「リプライ先のID」「発言者のID」「発言者のハンドルネーム」「発言のタイトル」「発言した時刻（「yymmddHHMMSS」という記述方式。例：2012年9月21日17時51分22秒→120921175122)」「発言の内容」が記述されている。我々はこれらのcsvファイルを、通常のツリー型掲示板のようなUIでデータを見ることができるように処理をし、htmlファイルに変換した。また、次節で紹介する統計量毎にソートして電子会議室やコメントを閲覧することができるようにした。

NIFTY-Serveデータ[6]はニフティ内の古いシステムからサルベージされた過去データの一部であり、データと会議室名の食い違いや欠損を含んでいる（電子会議室のデータが最初から最後まで揃っておらず、途中の一部分しかない、など)。今回用いたのはデータの総数は40フォーラム分、全540MBである。実際にNIFTY-Serveには1000を超えるフォーラムがあったとされ、分析可能なデータは今後順次追加される予定である。

３．既存研究

会話やオンラインコミュニティの「盛り上がり」について述べた論文としては[1],[2],[3]などが挙げられる。[1]ではネット掲示板の盛り上がりとメッセージの参照関係の可視化に焦点を置き、時間軸を横軸、コメント数を縦軸としたグラフに、参照関係にあるそれぞれのノードを半円で接続するという可視化の手法を紹介している。[2]では対話における盛り上がりの自動判定を行うことを目的とし、CRFを用いた判定手法を提案している。ここで判定機の学習に用いる素性は、発話の長さや、発話間の語彙的結束性（相手が話した内容と同じような内容を自分が話したか）などである。

一方、本研究で対象としている「盛り上がり」とは何かを具体的に述べている研究は少ないが、これを考える際に、[1],[2]における「コメント数」や「発話の長さ」「発話の関係性」などに着目した分析が参考になる。しかし、盛り上がりがこういった要素のみで定義できるとは考えにくい。盛り上がりを定量的客観的に定義する事は難しく、またそもそも主観的にも何が盛り上がりで何が盛り上がりでないかを定義するのは意見が分かれる(例えば「一人のユーザーが延々と書き込みを続ける掲示板は盛り上がっているのか？」あるいは「複数のユーザーが機械的にイベント等の告知を書き込む掲示板は盛り上がっているのか？」など)。しかしながら、厳密な定義は難しいにしても、盛り上がりに少なからず関係する指標は様々あると考えられ、定量的客観的な盛り上がりの分析のためには、それらについて分析・考察することは重要である。本研究では、まずは盛り上がりに関係する/しないに関わらず、様々な統計量を計量し、データがどんな特徴を持っているのか、また統計量がデータのどんな特徴を表しているのかを分析し、最終的に盛り上がりとは何なのか、どの統計量が盛り上がりの指標となるのかを考察する。

また[3]では「２ちゃんねる」を対象とした盛り上がりの分析を行い、２ちゃんねるを特徴付ける指標として、「1コメント辺りのサイズ」「1スレッドあたりの投稿数」「引用府つきで返信された割合」「1日辺りの投稿数」などといった8つの指標を使い、共分散構造分析を用いてそれらの因果モデルの構築を試みている。指標には2ちゃんねるに特有な指標も含まれ、例えば「1スレッドあたりの無名ハンドルネームでの書き込みの数」などがある。本研究でも、「コメント数」のような基本的な指標以外に、NIFTY-Serve特有の「ユーザー名が確実に判定できる」という特徴を用いた、「誰が誰に何回返信したか」などの指標を用いている。

４．統計量の計量、分析

本節では、フォーラム内の会議室単位や投稿位において様々な統計量を抽出し、それぞれの統計量がデータのどんな特徴を表しているのか分析を行い、それと並行してデータがどんな特徴を持つのか、データに対しての理解を深める。統計量は、会議室単位での統計量と、コメントツリー（ある投稿と、それに対して返信したコメント、それらに対して返信したコメント…の全体を称してコメントツリーと呼ぶ）単位での統計量の2つを算出した。会議室単位での統計量については、グラフを用いて可視化したものも含めて紹介する。

４．１．会議室毎の統計量

1. 総コメント数

最終的にコメントがいくつあったか表す統計量である。最も単純に盛り上がりを示す指標だと考えられる。最小コメント数は1、最大は999コメント（上限）であった。

2. ユーザー数

何名のユーザーが会議室に書き込んだかを表す統計量である。同じ総コメント数でも、少人数だけが書き込

んでいるのか、大人数が書き込んでいるのかで、会議室の特徴は異なると考えられる。

3. 期間

会議室が存続した期間を表す統計量である。

4. 時間あたりのコメント数

コメント数を期間で割ったものである。この統計量で、同じコメント数でも、短い間に多くのコメントがついた掲示板と、長い間に少しずつコメントがついた掲示板を区別できると思われた。実際は、短期間に盛り上がった掲示板も、その後も尻すぼみになりながらコメントがつくことが多く、有意な期間を算出することが難しかったため、区別は難しかった。

5. 常連数、非常連数

定期的に書き込みを行っているユーザーの数を可視化するために用いた統計量である。「常連」の定義として、1 週間に 1 回以上書き込んだ場合に 1 ポイントとし、10 ポイント以上となったユーザー（単純には、10 週以上書き込んだユーザー）を常連とした。逆に、10 ポイントに満たなかったユーザーを非常連とした。非常連の数が常連の数より遥かに多い会議室は、「カメラ購入の相談室」のような、頻繁に新規ユーザーが訪れ、一部の常連ユーザーが質問に答えるといった形態をもつ会議室であった。逆に、非常連の数が常連の数より十分に少ない会議室があれば、新規ユーザーの数が少なく常連だけが常に話しているような会議室であることが予想されたが、実際は非常連の数が十分少ない会議室は見られず、どの会議室も一定数の新規ユーザーの参加があることが確認された。

6. 圧縮率

生の csv データに、gzip コマンドをかけた時の圧縮率を示す値である。gzip では、同じ文字列が繰り返される場合は圧縮率が高くなる。これを用いて、gzip によりその会議室の発言全体がもつ情報量のようなものをごく簡単に近似できるのではないか、そして情報量の差が盛り上がりに関連しているのではないか、と考えこの手法を用いた。実際に圧縮率が高かったデータを見てみると、決まったフォーマットをもつ会議室が上位に見られた。決まったフォーマットをもつ会議室とは、例えば以下のような大会の結果報告が、同じ形式で連続的に書き込まれ続ける会議室である。

＝＝＝＝＝＝＝第 n 回○○記念大会＝＝＝＝＝＝＝
１位：○○○○○○○○
２位：××××
（略）
＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

圧縮率が低かったデータには、既に圧縮されたテキストデータで書き込みを行うという特殊な会議室が見られた。このような例外を除くと、圧縮率が低いデータはたしかに様々な話題が見られるようにも見えたが、人目で圧縮率の高いデータと低いデータを比べても、あまり有意な差は見られなかった。

以上より、圧縮率を指標として、定型フォーマットをもつ文章かどうか、といった判定はできそうだが、それ以上の指標にはならなそうであると言える。

7. 返信先の分布

会議室の全てのコメントに対して、それぞれが誰に対して返信したコメントなのかの分布を算出した。具体的には、「返信先なし」「自分に対しての返信」「誰かに対しての返信」の 3 タイプで分類した。「自分に対しての返信」が多い会議室は、自分の小説を投稿する会議室やスポーツなどの試合の結果を一人がひたすら報告する会議室で、ユーザー同士の交流が少ないことが確認された。

8. 返信先、返信された先の分布

会議室の全てのコメントに対して、それぞれが誰に対して返信したコメントであり、そして誰から返信があったかを算出した。具体的には、「返信先なし、返信もなし」「A に対して返信、A から返信あり」「A に対して返信、B から返信」という 3 つのタイプを抽出した。「A に対して返信、A から返信あり」と「A に対して返信、B から返信」の 2 つのタイプの数を見る事で、2 ユーザー間で盛り上がっているだけの会議室なのか、複数のユ

ーザー間で盛り上がっている会議室なのかが区別できた。

9. 各ユーザーのコメント数

その会議室で1回以上書き込んだユーザーのコメント数を、ソートした上で棒グラフで可視化した。これにより、その会議室で中心的な役割を果たすユーザーがわかる。

10. 各ユーザーの常連度

5で紹介した「常連度」をソートした上で棒グラフで可視化した。9でのコメント数では、短期間に集中して書き込んだユーザーと長期間に渡って書き込み続けたユーザーが区別できなかったが、このグラフではそれらを区別できる。

11. 2ユーザー間の会話数

「AがBに返信した回数」を全てのユーザー間で求め、ソートした上で棒グラフで可視化した。例えばAとBの会話が非常に多い会議室では、棒グラフの上位が「AがBに返信した回数」「BがAに返信した回数」と並び、他のユーザーの会話数は非常に低い、というグラフになる。逆に複数ユーザーがまんべんなく会話している会議室では、ランキング上位は様々なユーザー間の会話がみられる。またユーザー間の交流があまりない会議室では、「AがAに返信した回数」「BがBに返信した回数」などがランキングの上位にあがる。

更にこのグラフを分かりやすく可視化するために、グラフ構造として可視化することも考えられたが、今回はそこまでの実装は行わなかった。グラフ構造として可視化することができれば、例えばAとB、CとDが仲良く会話しているだけの会議室はAとBの間とCとDの間にのみ太いエッジが接続される、Aが複数ユーザーと仲良くしているがA以外のユーザー同士ではやりとりをしていない場合はいわゆるスター型のグラフ構造となる、などと会議室の特徴が更にわかりやすく可視化できるものと考えられる。

12. 単語頻度

会議室の全てのコメントに対して形態素分析を行い、登場した単語の頻度を算出し、上位10件を表示させた。その会議室のキーワードとなるような単語が頻度の上位に来る事を確認した他、助詞や助動詞（通常、単語頻度分析を行う場合は除外するが、今回はあえて除外しなかった。記号やアルファベットは除外している）が頻度上位に入っているかいないかで、決まったフォーマットをもつ会議室（「6.圧縮率」で挙げたようなもの）かどうかが区別できることを確認した。

４．２．コメントツリー毎の統計量

1. 総コメント数

コメントツリー全体のコメント数を表す統計量である。

2. 直接の返信数

コメントツリーのルートとなるコメントと、直接の返信関係にあるコメントの数である。直接の返信数が多いコメントツリーには、オフ会の告知が多く見られた（一人の告知者に対して、多くのユーザーがそのコメントに対する返信という形で参加希望を表明するため)。

3. ユーザー数

コメントツリー全体の、参加ユーザー数を表す統計量である。

4. ツリーの深さ

コメントツリーの最大の深さを表す統計量である。

5. ツリーの分岐の数

コメントツリーのそれぞれのコメントに対して、返信がn回来たとき、n-1回の分岐がそこで発生した、とカウントし、コメントツリー全体で分岐数を合計した統計量である。自分の小説を自分のコメントに対する返信という形でひたすら書き込んでいるコメントツリーなどでは分岐数は低く、複数のユーザーがどんどん話題を膨らませていくようなコメントツリーや、前述のオフ会の告知のようなコメントツリーでは分岐数は高くな

る。

５．盛り上がりの分析

盛り上がりについて深く考察する前に、簡単な統計量で盛り上がりを近似した上で、いくつかの分析を行った。以下ではいくつかの説明変数に対して単純相関をとった結果と、単純な二値分類モデルにおいて形態素を素性として、盛り上がっているか/いないかを学習した結果に関して述べる。

５．１　盛り上がり分析（単純相関）

本節では、簡単な盛り上がり分析の結果について述べる。盛り上がり分析の第一歩として、前節で述べたような深い盛り上がりの定義や、複雑な盛り上がりの説明変数・分析手法は用いず、まずは簡単な盛り上がりの指標と簡単な盛り上がりの説明変数の間の相関関係を導いた。

盛り上がりの分析は、会議室単位での分析（コミュニティの盛り上がりの分析）と、コメントツリー単位での分析（コメントに対する盛り上がりの分析）が考えられたが、まず後者についての分析を行う。盛り上がりに対する指標としては「総コメント数」「直接の返信数」「ユーザー数」を、盛り上がりを表す説明変数（素性）としては以下の３つを考え、これらの単純相関をとった。

1. 疑問符の数
   これは、疑問系が多いコメントの方が、返信が多くなるのではないかという仮説に基づいた素性である。
2. 自分の返信先に対して返信の早さ
3. 子供の自分に対しての返信の早さ
   2,3 は、できるだけ早く返信をすることで、その後の返信は多くなるのではないかという仮説に基づいた素性である。

フォーラムに含まれる全てのコメントを対象として、それぞれについて単純相関をとった結果、どの結果も 0.2 以下の値となり、相関はほぼ見られなかった。ここから、疑問符の数や返信の早さといった単純な素性では、盛り上がりは分析できないことがわかる。今回の定義における「盛り上がっている」コメントを人手でいくつか見て分析してみても、これといって共通する特徴はみられず、そのコメントが扱う話題や、時間とユーザーの偶然の一致のような、分析するのが非常に難しい部分に盛り上がりの要因は隠れているように思われた。

５．２　盛り上がり分析（形態素を素性とした二値分類）

形態素を素性として、フォーラムの各投稿がその後のコメントの盛り上がりに繋がっている/いないを学習する実験を行った。ある投稿の特徴量としてはタイトルと投稿本文の Mecab[4]による形態素解析結果を用いた。分類対象とする二値を Boom（盛り上がっている）と Normal（特に盛り上がっていない）を定義し、実際には上記盛り上がり分析の特徴の１つである総コメント数、総ユーザー数、返信数の値に単純なしきい値を設け、いずれかの値がしきい値を超えたものを Boom クラスの属するとした。学習に用いたデータ数は合計で 3209 サンプルであり、内訳は Boom クラスが 1716 サンプル、Normal クラスが 1493 である。

分類器の学習には Bazil を用いた。Bazil は株式会社 Preferred Infrastructure で開発した機械学習ライブラリであり、現在は特に自然言語処理に特化した前処理機能と、線形分類学習機能、予測結果の可視化機能、およびこれらをコントロールする Web ブラウザベースの UI を備えている。学習モデルは正則化付きオンライン線形分類器の一種である AROW[5]を用いた。以下の図 1 は学習結果表示のスクリーンショットである。なお、著作権の問題から表示されている投稿例はダミーであるが、モデルの学習には実際のニフティサーブのデータを用いている。この投稿は Boom クラスによく現れる、オフ会参加募集を模倣しているが、実際に Boom クラスだという予測が得られていることが左下の Predicted label のスコア順から判る。

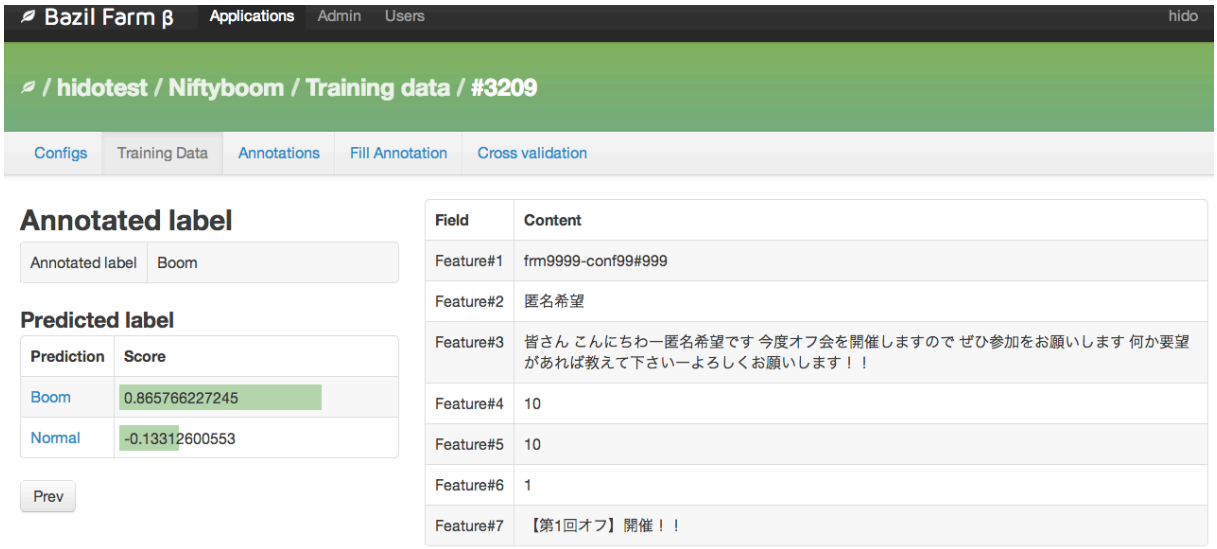

図 1　Bazil における学習結果表示のスクリーンショット

学習を行った後、評価実験を行った。3-fold 交差検定を行い、結果は Boom クラス予測の Precision が 99.42%、Recall が 99.33%となった。ただし、Normal クラスには特徴的な定型形式を持つ情報共有型の投稿が多数含まれていることや、サンプル数が十分ではないことを考慮すると、過学習している可能性があり、統計的有意な結果を得るためにはさらに多くのサンプルを用いた実験が必要である。

よく効いた素性はタイトルに含まれる「【】」「オフ」「！」などのイベント情報やオフ会告知によく現れると考えられるもの、及び本文中の「下さい」「お願いします」など要望や依頼を表す表現となっており、実際の盛り上がり度の定義から考えられる直感に比較的合致する傾向が見られた。

６．おわりに

本稿では NIFTY-Serve のフォーラムデータを対象としてデータの基本的な分析を行い、そこで得られたデータの特徴について述べた。分析の際に用いた統計量としては、「コメント数」「ユーザー数」のような基本的な統計量の他に、「常連数」「2 ユーザー間の会話数」などの固有の統計量も用い、データの特徴を出来る限り可視化することができた。これらの手法は、あるコミュニティにおける発言のデータに対して可視化を行う際、有効に使える手法であると考えられる。またその他に、簡単な盛り上がりの分析を行い、単純な素性では盛り上がりは分析できないことを示した。

今後は、今回の分析を基に盛り上がりの定義を行い、その上で深い分析を行っていく予定である。

［参考文献］

[1] 塩澤秀和『ネット掲示板の盛り上がりとメッセージ参照関係の可視化』
[2] 稲葉通将，鳥海不二夫，石井健一郎『語の共起情報を用いた対話における盛り上がりの自動判定』　電子情報通信学会論文誌 2011　Vol. J94-D No. 1 pp. 59-67
[3] 松村真宏，三浦麻子，柴内康文，大澤幸生，石塚満『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 Vol. 45 No. 3 Mar. 2004
[4] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto, Applying Conditional Random Fields to Japanese Morphological Analysis, EMNLP, 2004
[5] K. Crammer, A. Kulesza, M. Dredze, Adaptive Regularization of Weight Vectors, NIPS, 2009
[6] 宇田周平，三浦麻子，森尾博昭，折田明子，鈴木隆一，田代光輝，佐古裕，NIFTY-Serve におけるフォーラムデータの分析と整形，第 4 回知識共有コミュニティワークショップ，2011