NIFTY-Serveにおける フォーラムデータの分析

成澤克麻[1]　比戸将平[2]　海野裕也[2]
松井くにお[3]　鈴木隆一[3]　田代光輝[3]　丸山宏[4]
[1] 東北大学　　[2] 株式会社Preferred Infrastructure
[3] ニフティ株式会社　　[4] 統計数理研究所

第5回知識共有コミュニティWS　2012/11/11

NIFTY-Serveとは

- ニフティが運営していたパソコン通信サービス
  - 1987年~2006年
- フォーラムと呼ばれるコミュニティサービスが特徴
  - それぞれのフォーラムに電子会議室（掲示板）を自由に設置可
    - 例：「スポーツ」フォーラムの「☆1999年夏甲子園☆」
  - フォーラム数は最大1000個以上
- オンラインコミュニティのデータとしての特徴
  - 比較的古い（インターネット普及以前のデータ）
  - 完全会員制で、あるユーザーがいつ書き込んだか完全に把握可能

- 40フォーラム（540MB）を対象

分析の目的

コミュニティの「盛り上がり」の要因を明らかにする

例：疑問系多めの敬語で返信をしてやると盛り上がりやすい

応用例：コミュニティの運営に役立てる

※大量のデータを用いた定量的な手法で分析を行う

タスク

- 1. 「盛り上がり」の定量的な定義
  - 盛り上がりとは何か？ を考える
  - 例：コメントが◯件ついていて△△な会議室は盛り上がっている
- 2. 「盛り上がり」の要因分析
  - 何が盛り上がりに影響するか？ を考える
  - 例：レスの早さが、盛り上がりに◯◯%影響する〜

今回：
1. 関係しそうな統計量をデータから抽出 & 考察
2. 簡単な分析

統計量の計量、分析（前処理）

- NIFTY-Serve フォーラムデータについて
  - 1つ1つの電子会議室のデータがcsvになっている
  - 内容：「（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」

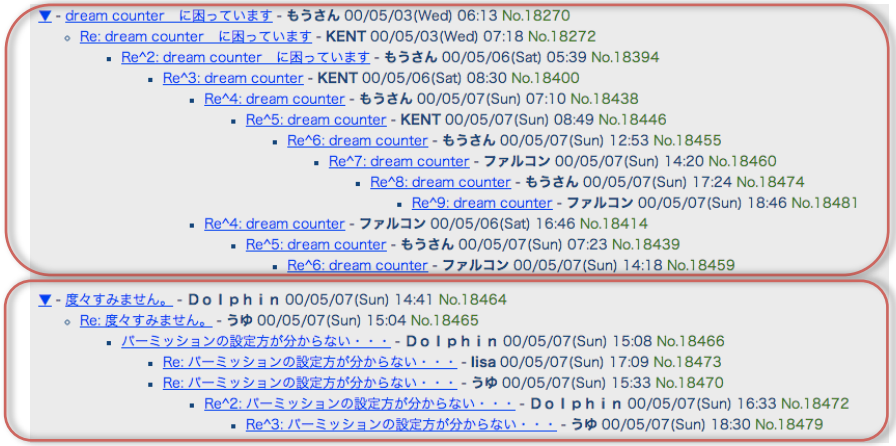
- 統計量を抽出しつつ、htmlデータに変換
  - 各統計量でソートできるような形にして分析を行った

conf-id	会議室名	コメント数	ユーザー数	常連の数	非常連の数	コメント数 / 日	期間
frm0054-conf10	☆1999年夏甲子園☆	331	30	10	20	2.1	124
frm0031-conf03	カメラ購入相談室	999	125	15	110	1.2	234

統計量の計量、分析

- 以下の2つの単位で、統計量の抽出を行った
  - 会議室単位
  - コメントツリー単位

- ※コメントツリー：
  - あるコメント以下にあるコメント全体



統計量の計量、分析（会議室単位）

- 以下の統計量を抽出した
  - 1. 総コメント数
  - 2. ユーザー数
  - 3. 期間
  - 4. 時間あたりのコメント数
  - 5. 常連数、非常連数
  - 6. 圧縮率（gzip）
  - 7. 返信先の分布
  - 8. 返信先、返信された先の分布
  - 9. 各ユーザーのコメント数
  - 10. 各ユーザーの常連度
  - 11. 2ユーザーの会話数
  - 12. 単語頻度

統計量の計量、分析（会議室単位）

- 以下の統計量を抽出した
  - 1. 総コメント数
  - 2. ユーザー数
  - 3. 期間
  - 4. 時間あたりのコメント数
  - 5. 常連数、非常連数
  - 6. 圧縮率（gzip）
  - 7. 返信先の分布
  - 8. 返信先、返信された先の分布
  - 9. 各ユーザーのコメント数
  - 10. 各ユーザーの常連度
  - 11. 2ユーザーの会話数
  - 12. 単語頻度

統計量の計量、分析（会議室単位）

- 5. 常連度、非常連度
  - 常連：ある一週間に1回以上書き込むと1ポイントとしたとき、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

考察

- 非常連の数が極端に多い会議室：
  「カメラ購入相談室」のような質問・相談系の会議室
  - 新規ユーザーが頻繁に出入りする会議室
  - これは盛り上がりとは違うような？
- 常連の数が極端に多い会議室：あまり見られず

統計量の計量、分析（会議室単位）

- 6. 圧縮率
  - 生のcsvデータに、gzipコマンドをかけたときの圧縮率
    - gzip：同じ文字列が繰り返されると、圧縮率高

仮説：

- 盛り上がっている ≒ 様々な話題がでる ≒ gzipの圧縮率低？
  - かなり強引
  - 厳密にやりたければ、単語頻度などをカウントするべき

統計量の計量、分析（会議室単位）

- 6. 圧縮率

  結果：

  - 圧縮率高：定型フォーマットをもつ会議室

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

（この形式の書き込みが延々と続く）

  - 圧縮率低：特に特徴は見られなかった
  - 仮説は実証されなかった

統計量の計量、分析 (コメントツリー単位)

- 以下の統計量を抽出した
  - 1. 総コメント数
  - 2. ユーザー数
  - 3. 直接の返信数
  - 4. ツリーの深さ
  - 5. ツリーの分岐の数

盛り上がりの分析

- 以下の２つの分析を行った
  - 対象はコメントツリー単位

- 1. 単純相関
- 2. 形態素を素性とした二値分類

盛り上がりの分析（単純相関）

- 以下を盛り上がりの指標、また説明変数として、単純相関をとった

指標：1. 総返信数（総コメント数）
2. 総ユーザー数
3. 直接の返信数

説明変数：1. 疑問符の数
2. 返信の早さ

- 対象：フォーラムに含まれる全てのコメントツリー

- 結果：どれも相関は見られず（0.2以下）
- 考察：簡単な素性では、盛り上がりは分析できない
  - もっと深い素性（コメントの内容etc）を考えた分析を行う必要

盛り上がりの分析（二値分類）

- 形態素を素性として、分類器を学習
  - コメントがその後のコメントの盛り上がりに繋がっている (Boom) / いない(Normal) を分類する

皆さん, こんにちは, ！, ◯◯, って, どう, 思い, ます, か？　(Boom)

- 学習データ
  - それぞれのコメントをBoom / Normal クラスに分類したもの
  - Boomクラスの定義…以下の値のいずれかが閾値を超えたもの
    - 「総コメント数」「総ユーザー数」「直接の返信数」
    - ※単純相関で用いた指標と同じ
  - 計3209サンプル（Boom 1716, Normal 1493)

- 学習結果のスクリーンショット

- ※学習にはBazilを使用
  - 株式会社Preferred Infrastructureで開発した機械学習ライブラリ
- 学習モデル：AROW（正則化付きオンライン線形分類器）

評価実験・考察

- 評価実験
  - 3-fold交差検定を行った
  - precision 99.42　　recall 99.33
    - 十分に学習できている

- よく効いた素性
  - タイトルに含まれるもの　例：「【】」「オフ」「！」
  - オフ会告知etcでよく表れるもの　例：「下さい」「お願いします」
  - 今回の盛り上がり定義の結果としては妥当
    - しかし、これが「盛り上がり」なのかは疑問

まとめ・今後の課題

まとめ

- 様々な統計量を抽出・考察
  - 新規ユーザーの参加具合、定型フォーマットをもつフォーラム、ユーザーの交流具合etc… などを統計量で可視化できた
  - 結局何が盛り上がりなのか？ まで深く考察できていない
- 簡単な手法で「盛り上がり」を分析
  - 相関分析：疑問符や返信の早さと、相関は見られない
  - 形態素：オフ会やタイトルによく見られる形態素
  - 単純な素性で「盛り上がり」が分析できるのかは疑問。深い分析を行うためにはどうすればいいのか？