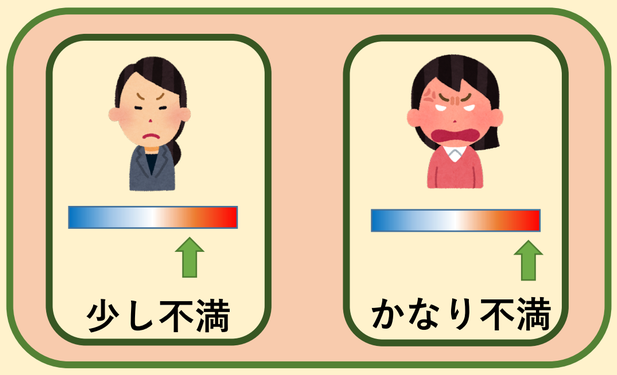
不満調査データセットを用いたコメントの不満度評価システムの提案

沼尾研究室　開発　知斗

背景と目的

　世の中の多くの企業や組織が, アンケートや掲示板上の書き込みなどからVoC(Voice of Consumer, 顧客の声)を収集し活用しようとしているが, それらの解析が上手く行かず十分に活用できていないケースも多く見られる．本研究では, VoCの中でも否定的な文章が多く含まれている不満調査データセット中の文章を対象に, その中でも特に筆者の不満が高いと推測される文章を自動的に抽出することを目的とする.



不満調査データセットとは

　不満調査データセットは株式会社Insight Techが提供　するサービス「不満買取センター」内で収集されたデータセットで, 投稿データは約520万件となっている.データセット内では不満文章以外にも年齢や性別などの投稿者の情報, 不満の対象となる製品や業界の種類, 構文解析結果や意見タグなどの多くの情報が含まれている. ただ, 今回は以前に提供されたバージョンである約25万件の投稿データを元に実験を行った.


```
{
    "normalized_company_name": null,
    "product_category": "その他",
    "user_number": 5,
    "fuman": "医者が処方する薬が病院によって量が違いすぎる。",
    "state": "埼玉県",
    "product_name": null,
    "birth_year": 1976.0,
    "status": "ANNOTATED",
    "company": null,
    "job": "会社員（技術系）",
    "gender": "male",
    "industry": "美容・健康",
    "proposals": "統一のためのガイドラインなど制定すべき。過剰処方には罰則を設けるべき。",
    "time": "2015-03-18 22:59:24"
},
```


図1: 不満調査データセットのレコードの例

提案システム構成

　主に提案手法は, **①不満表現辞書の構築** と **②不満度の算出** の２つのプロセスに分かれている.

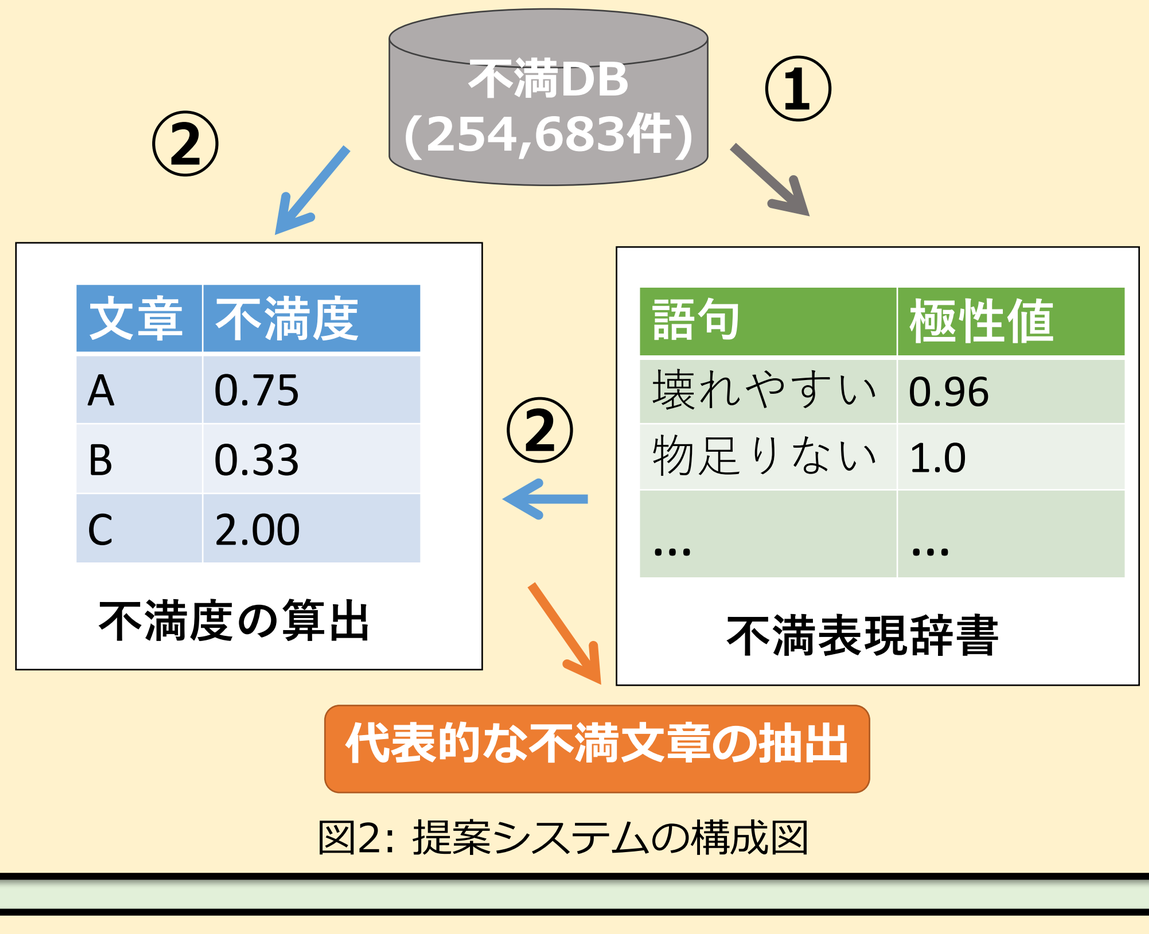


図2: 提案システムの構成図

実験

　不満調査データセットのうち, 年齢が10代から40代の複文となっている文章に対して不満度を計算した. また, 得られた結果を元に横軸が不満度, 横軸が頻度となるヒストグラムを作成した.

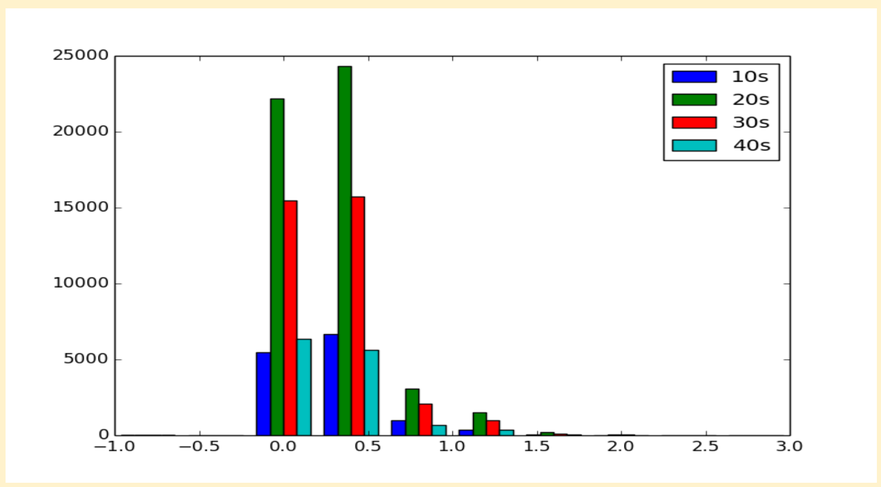


図3: 年齢ごとの不満度ヒストグラム

　上図の結果より, どの年代においても全体の大部分の文章の不満度が0以上1.0未満に集中し, 各年代毎の不満度の分布に大きな差は見られなかった。

　また, 各カテゴリ毎に不満度の最大値・最小値・中央値を算出し, それらを元にカテゴリ毎の不満度の箱ひげ図を作成した.

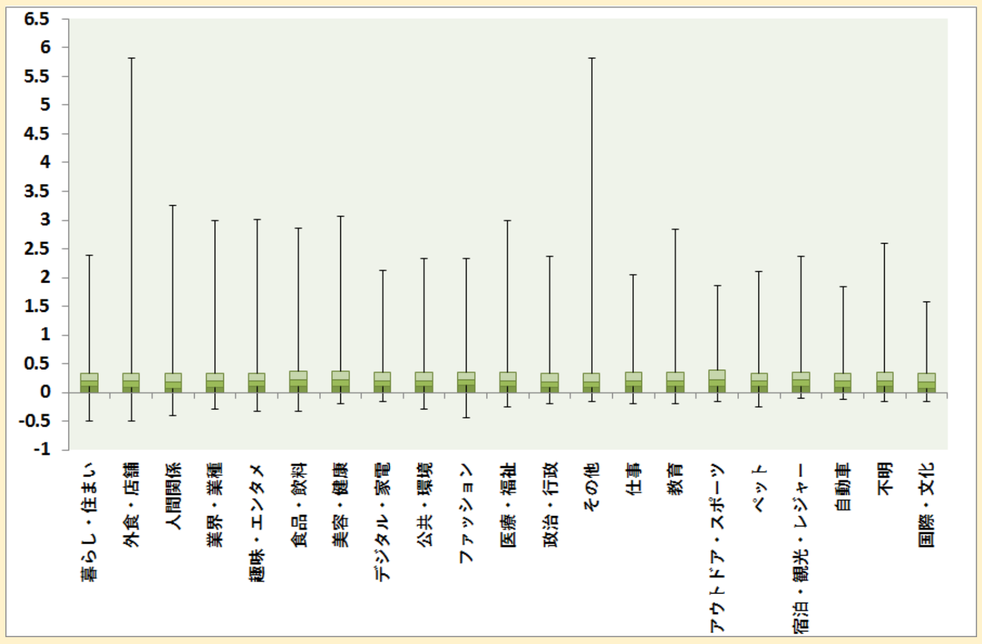


図4: カテゴリ毎の不満度の箱ひげ図

　不満度の最低値は「暮らし・住まい」「外食・店舗」のカテゴリに含まれる文章が, 最高値については「外食・店舗」「その他」のカテゴリに含まれる文章がそれぞれ記録した。

考察

　不満度の分布に大きな差が見られなかった理由としては, 収集された否定表現が多くの文章中に1, 2語程度しか現れなかったことやそれぞれの評価表現の重みを均一にしたことが挙げられる. また, 否定表現が多く含まれる文章は不満度が高くなるが多く含まれない複数の文章を区別するためには, 「悪い」や「最悪」などの評価表現自体の重み付けや「とても」や「非常に」などの評価表現を強調する語彙や評価表現の活用形を考慮する必要がある.

今後の課題

不満表現辞書に登録される評価表現の収集精度, 語彙数の向上や不満度の式の改良, 評価表現の重み付けが挙げられる.