不満調査データセットを用いた不満グループの可視化

長谷川　徹†　　北山　大輔††

† 工学院大学工学部第二部　〒 163-8677　新宿区西新宿 1-24-2
†† 工学院大学情報学部　〒 163-8677　新宿区西新宿 1-24-2
E-mail: †c513510@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　近年，SNS や Web ページの文章を取得・分析し，様々に活用を行うテキストマイニング分野の研究が盛んである．そうしたテキストマイニング分野の研究では，何らかの事象に対する不満意見のテキストを分析，活用することを目的とする研究は数多く行われている．本研究では，国立情報学研究所（NII）が情報学研究データリポジトリ（IDR）にて提供している「不満調査データセット」を利用し，特定の事象に対して不満を持つユーザ層をグループ化し，ユーザグループ間の関係性の抽出・可視化を試みる．

キーワード　情報可視化, 不満データセット

1. はじめに

近年，掲示板や SNS，ブログなど，Web 上で文章を中心として様々な交流を行う媒体が普及している．その普及に伴い，情報学研究の分野においても，Web 上の文章を分析し，様々な活用を行うテキストマイニング分野の研究が盛んである．

それらの研究の中では，特に何らかの事象に対する不満意見が書かれた文書に着目し，不満意見の分析・活用を行おうとしているものが複数存在する．多種多様なカテゴリや意図のある文章の中で，特に不満意見だけに着目した研究が存在する理由として，不満意見というものの活用可能性の高さがあげられる．

収集した不満意見の典型的な活用例として，企業が自社の製品やサービスに対する不満を収集することで，製品やサービスの改善に活用したり，新製品の開発の参考にするというものがある．こうした目的のための不満の収集は，テキストマイニングの普及以前からアンケートなど様々な形で行われてきたものであり，不満意見を参考にした改善を実施した結果，多くのユーザが不満に思っていた事象が解決した場合，売上やユーザ満足度の向上に繋がる．

不満データをこのような目的に利用するためには,「誰が，何に対して，どのような」不満を持っているかを分析することが重要となる．本研究では，この中で特に「誰が」，つまりどのようなユーザが不満を投稿したか，という部分に着目した．

同一の対象に対する不満意見群を，年齢層や職業などのユーザの属性を利用して振り分けを行った場合，ユーザ層ごとにそれぞれ持っている不満の傾向が異なる．このようにユーザー層を基準としたグループ化を行い，どういったユーザ層が何に不満を持っているのか，どういったユーザ層同士は不満の傾向が類似しているのか，というような情報を可視化することができれば，さらなる不満の分析を行うことができ，より各個のユーザ層に合わせた不満の解決法を探るようなことに役立てられると考えられる．

以上のような分析と可視化を行うために，本研究では，後述する「不満調査データセット」を利用して研究を行う．データセットの中で，特定の事象に対して不満を述べているデータを取得し，年齢や職業などのユーザデータを基準とした不満データのグループ化を行う．その上で，複数のグループ間の不満の類似度や関係性の抽出・可視化を試みる．これにより，先に述べたような不満データの「誰が」「何に対して」という点に注目した分析を行うことが可能となる．

さらに，不満データセットが様々な対象に対する不満データを含んでいるという特性を利用することで，複数の対象事象に対する不満データのグループ化を行い，そうして生成される不満グループの間で比較を行うことで，業種や企業などを跨いだ不満ユーザーの比較が可能となる．これを行うことで，単独の企業や商品などに対する不満ユーザの分析にとどまらず，他の企業においてはどのようなユーザ層が類似する不満を投稿しているか，ということを可視化することが可能となり，さらに分析の幅が広がると考えられる．本研究では，こうした「不満データのユーザ層を基準としたグループ化」，および「グループ間の比較・類似度の可視化」を可能とするシステムを構築した．

以下，本論文の構成を示す．まず 2 章では，本研究で用いているデータ，技術，および関連研究について説明する．3 章では，本研究で用いる不満データセットについての情報を提示する．4 章では，本研究の提案手法について具体的に説明する．5 章では，実際に行った実験とその結果について提示する．

2. 関連研究

はじめに，本研究には国立情報学研究所（NII) が情報学研究データリポジトリ（IDR）にて提供している「不満調査データセット」を使用している．同データに関する研究として，データの構築自体に関わった三澤らによる，データそのものの構築時の研究 [1] [2]，および，データを利用した不満意見の分類・クラスタリングの研究 [3] が存在する．しかし，これらの研究においては，先述した不満分析の要素「誰が」「何に」「どのような」というものの中で,「誰が」という点を扱っていない．本研究は，不満を投稿したユーザグループに着目した可視化を行う点で異なっている．

その他，本研究においては，不満データの分かち書きに形態素解析エンジン Mecab [4] を，データの可視化段階では Microsoft によるグラフ作成ツール GLEE [5] を，不満データの特徴量算出・類似度算出に TF-IDF の cos 類似度推定法 [6] を使用している．

本研究は，不満が書かれたテキストデータを活用するという点において，テキストマイニング分野において不満の分析を行っているものと類似する研究であると考えられる．いわゆるテキストマイニングと呼ばれる，文章中の評価情報などを利用する研究には，小林ら [7] の研究を始めとして非常に多くのものが存在する．

1 章で述べたような不満データの活用を行っている研究の例として，磯島ら [8] の研究では，生協に寄せられたクレームデータに対してテキストマイニングを行い，野菜の種類別に頻出するクレーム語を抽出し，クレームの分析と処理への活用を試みている．坂井ら [9] の研究では，ブログに書き込まれた不満表現から商品推薦を試みている．永井ら [10] の研究では，Web から収集した文書を利用し，クレームの調査・分析を行うシステムを制作している．これらの研究は，情報系研究における不満データの分析・活用の好例といえる．反面，本研究が目的とする「不満投稿者の年齢層・職業などを基準としたグループ化，関係性の可視化」という方向の分析は行っておらず，分析の方針と目的に違いがある．

また，不満データがそれぞれ商品などに対する評価を行っている文書であることを考慮すると，各種 EC サイトのレビュー分析などを行っている研究も関連研究としてあげられる．レビュー分析関連の研究として，中野 [11] らの研究では，商品レビュー要約のためにレビューのテキストデータを分析し，商品の評価属性と，それに対する意見のペアを抽出している．平山ら [12] の研究でも，同様にレビュー文書の分析を試みているが，これらの研究はそのレビューを投稿したユーザの年齢層の分析やグループ化を行っている訳では無いという点で本研究とは異なる．

3. 不満データセット

3.1 不満データの概要

不満データセットとは，2 章で関連研究にあげた通り，国立情報学研究所（NII) が情報学研究データリポジトリ（IDR）にて 2016 年 5 月 25 日から提供しているデータセットであり，「株式会社不満買取センター」が自社の Web サービス「不満買取センター」に一般ユーザが投稿した不満データを収集・提供しているものである．

不満調査データセットの特徴として，その成り立ちから，含まれているデータが全て，最初から不満意見として投稿されたデータであると言う点があげられる．不満データに着目した先行研究においては，Twitter などの SNS や，ブログや掲示板などの Web サービスに投稿された不満文書を取得して利用しているものが多い．

SNS やブログなどから不満データを取得する場合，そうした媒体の投稿者は不満とそうでない文書を明示的に区別せずに投稿することがほとんどである．そのため，不満データの収集時に，取得した文章が不満を表している文章なのかそうでないのか，という判別を行う必要がある．この不満調査データセットを利用した場合，不満データを取得する段階における文書判別の工程が不必要となるため，判別自体にかかる手間や，判別ミスによる不適切なデータの混入などを回避することができる．

表 1　不満データの構造

項目名	データ内容
id	不満データの id
company_name	不満対象の会社名
product_name	不満の対象物
category	不満の大まかなカテゴリ *
sub_category	詳細なカテゴリ *
state	ユーザの居住県 *
birth_year	ユーザの誕生年 *
job	ユーザの職業 *
gender	ユーザの性別 *
device	ユーザの使用している端末
time	不満が投稿された時刻
user_number	ユーザ ID
fuman	不満の内容
proposals	改善のための提案
status	不満データが買い取られたか

他の特徴として，各個の不満データにそれぞれ不満を投稿したユーザの年齢層や性別，職業などのプロファイル情報が含まれていると言う点があげられる．SNS やブログなどの媒体においては，ユーザが自らのプロファイル情報を公開していない場合が多く，プロファイルを用いた分析を行うことは難しい．それと比較して，不満意見とそれを投稿したユーザのプロファイルが明確に紐付いているという「不満データセット」のデータ構造は希少かつ有用なものであると考えられる．

以下，本研究に使用する不満データの内容について述べていく．データセットには，2015 年 3 月から 9 月までに投稿されたデータのうち，「不満買取センター」のオペレータがタグ付けを行った 254,683 件の不満データが収録されている．不満を投稿しているユニークユーザ数は 25,092 人である．

各個の不満データには，表 1 に示す項目のデータが含まれている．ただし，個々の不満データによっては一部項目のデータが入力されていない場合もある．表 1 に示した項目のうち，* マークが付いている項目は選択式のものである．

3.2 不満ユーザのプロファイルに関する所見

この項では，不満ユーザのプロファイル的情報に関する統計的なデータや特性などを提示し，それに関する考察を行う．

まず，不満ユーザのプロファイルとして用いられるデータとしては，先述の不満データ項目のうち，「state：ユーザーの居住県」，「birth_year：ユーザの誕生年」，「job：ユーザの職業」，「gender：ユーザの性別」，「device：ユーザの使用している端末」の 5 つの項目があげられる．

本研究においては，これらの項目の中で，「birth_year：ユーザの誕生年」，「job：ユーザの職業」の 2 項目を重要視し，提

表 2　不満ユーザの性別ごとの年齢分布

年齢層	女性	男性	不明	投稿数	投稿比率
10s	777	362	78	8927	0.0351
20s	4877	899	467	68944	0.2707
30s	5027	818	629	86248	0.3386
40s	2691	544	369	41877	0.1644
50s	918	191	148	12275	0.0482
60s	151	62	36	2330	0.0091
70s	8	5	3	60	0.0002
none	1850	244	3935	33987	0.1334
over	3	0	0	11	0.0000

案手法によるユーザグループの作成においては，この 2 項目を変数としてユーザグループの作成を行っている．プロファイルとして利用可能な項目として 5 つの項目がある中で，この 2 項目だけを特に利用する理由として，3 つの理由が存在する．

1 つ目は，グループ化の条件とする項目を増やしすぎると，その条件に割り当たるものとして分類されるユーザグループが細分化されすぎ，1 グループに含まれる不満データの数が少なくなりすぎてしまい，分析に不都合が生じる．そのことから，グループ化の条件として利用する項目の数を制限する必要があったこと．

2 つ目は,「gender：ユーザの性別」項目が，不満調査データセットにおいて大きな偏りが存在する項目であり，グループ化の条件として用いるのにあまり適さない項目であったこと．この点については，後述する 3.2.2 項，3.2.3 項で詳しく述べている．

そして 3 つ目は,「state：ユーザの居住県」および「device：ユーザの使用している端末」の 2 項目は,「年齢」や「職業」,「性別」などの項目と比較した場合，そのユーザの持つ意見や立場などに与えている影響が，相対的には小さいだろうと考えたこと．

この 3 つの理由から，本研究においては，ユーザの分類には職業と年齢の項目だけを用いている．次項からはプロファイルデータのうち，誕生年，職業，性別の内容について個別に述べていく．

3.2.1　不満ユーザの年齢層

まず，不満データの投稿者の年齢層についての分析を行う．注意すべき点として，不満データに含まれているのはユーザの生年のみであり，直接的な年齢データではないことがある．今回使用する不満データに関しては，不満データが取得された期間が全て 2015 年中であることから，ユーザの年齢は単純に 2015 から生年を引くことで求めている．

これを踏まえて，全不満ユーザの性別ごとの年齢分布と，投稿されている不満データの数とその分布を表 2 に示す．

表 2 を見て分かる通り，不満データ投稿者の中で最も多いのは 30 代，次いで 20 代，3 番目が 40 代となっている．この 20〜40 代の層だけでデータ全体のうち 7 割強を占めている．

一方で 20 代未満・50 代以降の数は，20〜40 代の層から離れるほどに少なくなっている．原因として考えられるのは,「不満買取センター」のサービス自体がインターネットに接続して利用するサービスであるため，ユーザ登録を行うためにパソコンやスマートフォンなどの機器を使用している必要があることがあげられる．

表 3　不満投稿アカウントの職業分布

	アカウント数	アカウント比率
専業主婦（主夫）	6373	0.253985
none	5283	0.210545
パート・アルバイト	3410	0.135900
会社員（その他）	2083	0.083015
学生	1856	0.073968
会社員（事務系）	1837	0.073211
会社員（技術系）	1467	0.058465
その他	848	0.033796
自営業	674	0.026861
無職	472	0.018811
自由業	405	0.016141
公務員	240	0.009565
経営者・役員	144	0.005739

表 4　不満投稿アカウント・不満データの男女比分布

	アカウント数	アカウント比率	投稿不満数	投稿比率
female	16302	0.649689	186583	0.732609
male	3125	0.124542	36581	0.143633
none	5665	0.225769	31519	0.123758

参考資料として，総務省の「インターネット利用動向」のデータ(注1)を参照すると，インターネットの利用率は 50 代から年齢が上がるごとに低下していることが見て取れる．こうした 50 代以降のインターネット利用者の少なさが，50 代以降の投稿者数が少ない理由の一端として考えられる．

3.2.2　不満ユーザの職業

次に，不満ユーザの職業分布について表 3 に提示する．表 3 を参照すると，ユーザの約 25%を専業主婦（主夫）層が占めており最大のグループとなっている．2 位は職業未入力のアカウントであり，3 位はパート・アルバイトの層となっている．この 1 位から 3 位がデータ全体の約 6 割を占めている．

また，職業データは，それぞれの属性ごとに年齢や性別との関係性が深い．例として,「学生」ユーザはその大半が 10 代ないし 20 代のユーザである.「専業主婦（主夫）」のユーザは女性の割合が全体の平均よりも高く，逆に「会社員（技術系）」は男性の割合が高い，というような関係が見られる．

3.2.3　不満ユーザの男女比

次に，投稿者の男女比について提示する．はじめに，不満データを投稿したアカウントの男女比分布と，実際に投稿された不満データの「gender」パラメータを集計したものを表 4 に示す．

まず，アカウント数とその中の比率を見ると，特に女性アカウントの数が多く，ユーザ全体のうち約 65%を占めていることがわかる．次いで多いのは性別不明のアカウント，男性アカウ

(注1)：http://www.soumu.go.jp/johotsusintokei/whitepaper/p/h27/html/nc372110.html

ントは最も少ない数となっている．

一方で，実際に投稿された不満データに占める割合を比較すると，女性の投稿した不満データが全体のうち約 73%を占めており，アカウントの比率よりも更に高い値になっている．このことから，女性ユーザは一人あたりの不満投稿数も男性や不明なアカウントよりも高いと言える．同時に，性別不明のアカウントと男性アカウントの占める割合も逆転しており，性別不明のアカウントよりも，性別を男性としているアカウントの方が不満投稿率が高いということがわかる．

不満データの男女比はこのように女性側に大きく偏っており，加えて前節に述べたように，職業データとも強い関係性がある．そのため，取得したデータの男女比には大きな偏りが生じることが多く，ユーザの男女による違いなどを利用した分析を行うことは難しいと言える．本研究において，不満データのグループ化に性別情報を用いないこととしたのはこの点が理由である．

3.2.4 不満プロファイルの正確性

以上，不満データのプロファイルに付いて述べてきたが，注意すべき点として，本研究を進める中で，不満データの「gender」項目が “female” となっているにも関わらず不満の内容を見ると「妻が〜」と妻に対する不満を書いていたり，その逆に，「gender」が “male” となっているにも関わらず，不満の内容は「夫が〜」となっているなど，ユーザの性別として登録されている性別と，投稿されている不満の内容が食い違っている不満データを複数件発見した．

こうした不満データが存在する原因として，まずアカウント登録時の設定ミスが考えられる．その他に考えられる可能性として，家族や友人など，異なる性別の人間によって構成されるグループの一人が不満買取センターに自分の情報を入力してユーザ登録を行った後，そのアカウントをグループ共有のアカウントとして利用した結果，様々な性別のユーザが同一アカウントから不満を投稿している可能性があると考えられる．

今回は少なくとも性別と不満データとの食い違いを発見したが，性別と比べると発見・判断が難しいというだけで，年齢や職業などの項目に関しても，同様の原因から食い違いが生じている不満データは存在すると考えられる．

しかし，登録している属性と不満の内容に相違があるデータを正確に判定・発見し，それを正しく修正するというのは現実的に困難である．そのため，本研究においては，不満データに含まれているユーザのプロファイルデータは全て正しいものであるという前提で実験を行っている．

このように，ユーザプロファイルの正確な判定は，本研究において今後改善すべき課題の一つである．

3.2.5 その他の不満データの情報

以下，補足として，不満データに関する統計的な情報のうち，ユーザプロファイル以外の項目で，特に注目すべき項目について説明する．

不満データの投稿期間について説明する．不満データセットに含まれている不満データの投稿期間は 2015 年 3 月 18 日から 9 月 23 日までとなっているが，3 月は 3 月 18 日以降のデータしか存在しないため，6,000 件足らずのデータ量となっている．

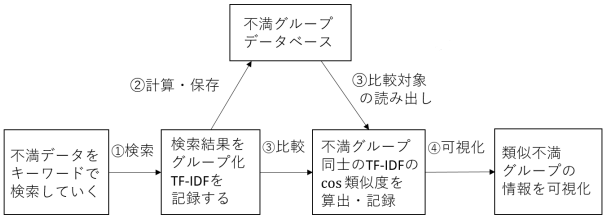

図 1　類似不満グループ可視化システムの概略

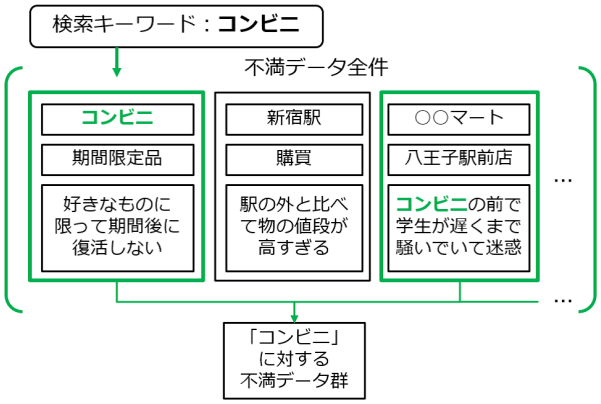

図 2　検索工程の概略

一方，4 月〜6 月は毎月 8 万件超の不満データが存在し，この期間だけで約 24 万件と，データセットに含まれる不満データ約 25 万件のうちの大半を占めるボリュームを持っている．

一方で 7 月〜9 月の不満データは数が少なく，合計して 600 件程度．加えて，この期間の不満データはほとんど 「status」値が “REJECTED” となっているものである．

4. 提 案 手 法

この章では，本研究の目標である共通する不満を持つユーザの抽出とグループ化，グループ間の関係性の可視化を行うための手法について説明していく．本研究では，図 1 のような手順で不満ユーザの抽出・関係性の可視化を行っている．各手順の詳細について，以下に記載していく．

4.1 検　　索

まず，検索の工程について説明する．検索工程の概略を図 2 に示す．検索の工程では，特定の対象に対して不満を述べている不満データを抽出するため，不満データセットの検索を行う．本研究では，特定対象に対する不満データの発見・抽出について，単純にキーワード検索を行うことによって求めている．

検索工程における具体的な処理を説明する．まず，不満データのうち「fuman」，「company_name」，「product_name」の 3 つを不満の対象が含まれうるデータであると考え，この項目に検索キーワードが含まれていた場合，キーワード k に対して生成される不満データ群 DL^k に追加していく．図に示した例の中では，「コンビニ」というキーワードで検索を行い，「company_name」にコンビニが入っていたもの，「fuman」にコンビニが入っていたものがそれぞれ不満データ群に追加されている．

なお，今回構築したシステムでは，検索時には単純な文章の

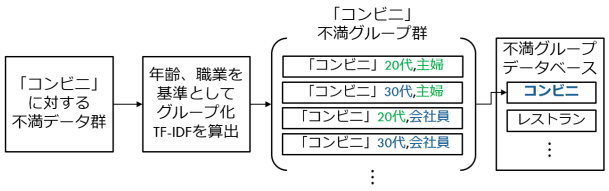

図 3　保存工程の概略

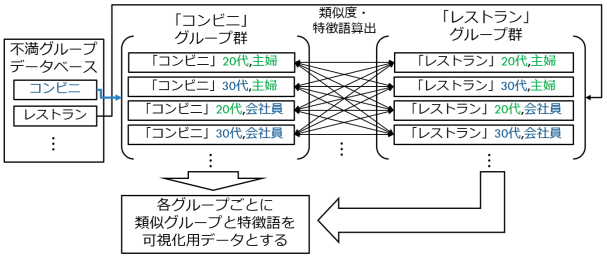

図 4　比較工程の概略

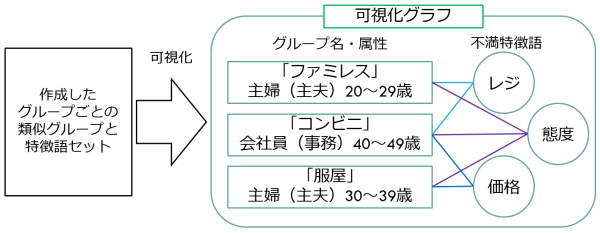

図 5　可視化工程の概略

一致だけを見ているため，例えば「シャープ」と検索した際には，電機メーカーの「シャープ」の他,「シャープペンシル」に対する不満なども検索結果に含んでしまう．この点は，今後の課題の一つである．

4.2　保　　存

次に，保存工程について説明する．保存工程の概略について図 3 に示す．保存の工程では，検索結果のグループ化，それぞれに対する TF-IDF 値の算出，結果のデータベースへの保存という 3 つの作業を行っている．

まず，求めた不満データリスト DL^k について，含まれている個々の不満データの「job」「birth_year」項目を利用し，職業・年齢を基準としたグループに分割する．例として，職業を「“主婦”，“会社員”」のどちらか，年齢を「“20〜29 歳”,“30〜39 歳” 」のどちらか，とした場合，不満データリスト DL^k に対しては $2 \times 2 = 4$ 通りの不満グループが生成されることになる．

ユーザープロファイル情報である職業と年齢の組を p とした時，DL^k に対して生成される不満グループ群を Gr^k，それに含まれる個々のユーザ属性ごとの不満グループを Gr_p^k と表記する．

次に，各不満グループ Gr_p^k について，不満データの「fuman」項目，つまり不満本文の文書データを分かち書きし，その中から名詞のみを取得して TF-IDF 値を算出していく．この際，不満グループ群 Gr^k を文書集合，それに対して存在する不満グループ Gr_p^k を一文書として TF-IDF を計算している．

まず，TF 値は各文書，つまり Gr_p^k における名詞の出現回数を集計することにより，それぞれの文書に対して求める．数式で表せば，

$$tf(t,d) = w_t^d \tag{1}$$

となる，この式における d は文書，つまり Gr_p^k を示し，t は索引語，つまり文書中に出現する名詞を示す．

IDF 値は，特定の単語が文書集合の中でいくつの文書に出現しているかを求め，その対数を IDF 値として利用するものである．数式で表すと，

$$idf(t) = log\frac{N}{df(t)} \tag{2}$$

となる．この時，t は先と同じく索引語を示す．N は文書の数，つまり文書集合 Gr^k のサイズである．$df(t)$ は索引語 t が含まれる文書 Gr_p^k の数を示す．

このようにして求めた TF と IDF を掛け合わせることにより，文書特徴量 TF-IDF が求まる．

$$tf\text{-}idf(t,d) = tf(t,d) \times idf(t) \tag{3}$$

最後に，このようにして算出したグループごとの TF-IDF 値をデータベースに記録していく．結果の保存は不満グループ群 Gr^k 単位で行っている．

4.3　比　　較

次に，比較工程について説明する．比較工程の概略について図 4 に示す．比較の工程では，データベースに記録されている他の不満グループ群 $Gr^{k'}$ を読み出し，それぞれ含まれている不満グループ間の類似度を算出してゆく．類似度算出には TF-IDF の cos 類似度推定法を用いている．

比較対象とする不満グループについてはデータベースに保存済みのものから任意に選ぶことができる．つまり，実験において「検索・保存」の工程を繰り返すほどに，比較対象として選択できる不満グループ群が増えていくこととなる．

具体的な動作内容としては，総当り方式で類似度を算出していく方式となっている．仮にグループの属性 p が $p_1, p_2, p_3 ... p_x$ とある場合，最初は $Gr_{p_1}^k$ と $Gr_{p_2}^k$ の cos 類似度を求め，次は $Gr_{p_3}^k$ へと推移し，Gr^k を比較し終えたら次は $Gr_{j_1}^{k_2}$ との比較を行う，という形で類似度を算出していく．

この際，グループ間の類似度が一定の値を超えた場合，その二つのグループは類似グループであるとする．類似グループに関しては二つのグループの間で TF-IDF 値の平均値を求めていき，値が高かった単語から順に「グループ間の特徴語」として記録していく．この特徴語は後の可視化において使用する．またこの際，類似度の高さに応じて抽出する特徴語の数は変動させる．

4.4　可　視　化

最後に，可視化工程について説明する．比較工程の概略について図 5 に示す．なお，図 5 に示した可視化グラフの形状はあくまで概念図としてのものであり，実際に本研究で試作を行ったシステムの出力結果とは異なる．

可視化の工程では,「比較」の工程で発見した類似する不満グ

ループと特徴語を可視化していく．具体的には，各不満グループと特徴語をそれぞれノードとし，グループから特徴語に対してエッジを接続していく．この時，「検索キーワード」「職業」「年齢層」をセットとしたものをグループ名とし，可視化に用いる．この時，一つのグループが複数のグループに対して類似度が高く，それぞれの特徴語に同一の単語が含まれていた場合，ノードを複数本作ることはせず，一本のノードにまとめている．

このような形で，類似したグループの間で TF-IDF 値が高かった単語を表示し，グラフ化を行う．これにより，類似する不満グループは同じ単語に対してエッジを伸ばしていることで判別できるようになる．

5. 実　　験

5.1 実験データおよびパラメータ

本章では，実際に構築したシステムによって行った不満グループの関係性の可視化実験について説明・考察する．

本研究で実施した実験においては，まず，4.1 節「検索」の工程における不満データのグループ化の条件として，「job」を「“パート・アルバイト”，“会社員（事務系）”，“専業主婦（主夫）”」のどれか，年齢を「20〜29 歳，30〜39 歳，40〜49 歳」のどれかとしている．

それぞれの選択基準として，まず職業については 3.2.2 項の表 3 で示した通り，「パート・アルバイト」「専業主婦（主夫）」の 2 つは，未入力を除けば上位 2 つを占めていることが選択の理由である．「会社員（事務系）」に関しては，「パート・アルバイト」の次に来る「会社員（その他）」が区分けとして曖昧であること，「学生」についても，年齢分布が 20 代以下に偏っていることから，どちらもグループ化に用いるデータとしてはあまり適切でないと考え，その次点である「会社員（事務系）」を選択した．年齢については，3.2.1 項で述べた通り，不満データセットは 20 代〜40 代が最もデータ量が多く分布しているため，単純にこの部分を実験に用いた．

次に，4.3 節「比較」における，類似グループと判断する類似度の条件として，今回の実験においては類似度 0.15 以上のものを類似不満グループと認定し，基本的な特徴語の抽出数は 2 個，そこから類似度が 0.1 高くなるごとに特徴語の抽出個数を 1 個増やしている．また，類似度が 0.15 を超えるグループが多数存在した場合，全てを類似グループと判断して特徴語の抽出・可視化を行うとグラフが複雑化しすぎることから，今回は「0.15 を超えたもののうち上位 3 件」を類似グループと判定している．

なお，可視化にあたって，各グループの各属性の正式名称を表示した場合，グラフ内で非常にスペースを消費してしまうことから，今回の実験ではグループ名・職業・年齢の全てについて，文頭から 3 文字を取得したものを可視化時に表示している．例えば，キーワードとして「工学院大学」と入力していた場合は「工学院」に，職業が「パート・アルバイト」の場合は「パート」にするという具合である．

5.2 実験の概要

実際に行った実験として，事業内容が類似する企業同士のグループ類似度算出と可視化を行った．実際にキーワードとして利用したのは「東芝」「パナソニック」の二つで，「東芝」には 442 件，「パナソニック」には 883 件の不満データが存在した．この二つの企業はどちらも日本を代表する大手電機メーカーであり，その展開商品は多岐にわたる．そのことから，不満の対象も幅広いものになることが考えられるため，グループ化を行う上で特徴的な結果が出るのではないかと想定し，キーワードとしてこの 2 社を選択した．

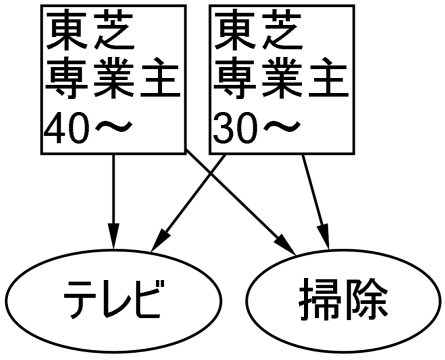

図 6 「東芝」の可視化結果

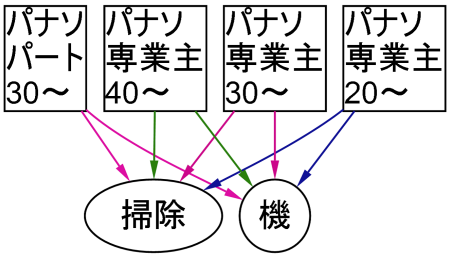

図 7 「パナソニック」の可視化結果

5.2.1 単一キーワード内でのグループ可視化

実際のグループ化実験の結果を述べる前に，まず「東芝」「パナソニック」それぞれのキーワードについての結果単体でグループ化と可視化を行い，それぞれのキーワードごとのグループの傾向を確認する．

最初に，「東芝」について可視化した結果を図 6 に示す．東芝については，各グループごとに不満の傾向が大きく別れており，可視化されるほど類似度の高いグループが 1 組しか存在しないという結果になった．類似度が高かったのは「専業主婦 30 代」「専業主婦 40 代」という 2 グループであり，特徴語として抽出されたのは「掃除」「テレビ」という 2 語であった．

次に，「パナソニック」について可視化した結果を図 7 に示す．

パナソニックについては，東芝と比べるとグループ間の類似度が高く，4 つの類似グループが可視化されたが，4 グループの間で特徴語はバラけず，4 グループで共通して「掃除」「機」という単語が特徴語として抽出された．また，可視化手法の関係で表示されなかったものの，これらのグループで TF-IDF 平均値が上位に来たものとしては，「洗濯」「乾燥」などが見られた．

「パナソニック」の可視化がこのようになった理由としては，類似度の高くならなかったグループである「会社員」の「20 代〜40 代」および「パート・アルバイト」の「20 代・40 代」と

比べて，抽出された 4 グループが共通して「掃除」「洗濯」など，家事に関連する機器にまつわる不満を多く述べていたことが原因だと考えられる．類似度が高かったグループのうち 3 つが「専業主婦（主夫）」であることと，これらのグループが家事関連の単語を不満として述べているということは，直感的にも納得できる関係である．

5.2.2 2 つのキーワードでのグループ可視化

以上の結果を踏まえて,「東芝」「パナソニック」について可視化したものが図 8 である．

キーワードを 2 つ用いた可視化の結果として，それぞれ個別に可視化した際には現れなかったグループが可視化されている．例として,「パナソニック 会社員 20 代」の層を見てみると，単体で見た場合の特徴語，つまり TF-IDF 値の高い単語としては「放送」「録画」「翌週」などの単語があったが，これらは同じパナソニックの他のグループとは類似しない単語だった．一方で，東芝グループとも類似度計算を行った場合,「東芝 パート 30 代」層とは「録画」が,「東芝 専業主婦 30 代」層とは「時間」が類似している，という結果が可視化された．このように，複数のキーワードを用いた可視化を行うことにより，初めて他のグループとの類似度が可視化されたり，単体での可視化を行った際には特徴語として現れなかったグループの特徴などが可視化されることがわかった．

次に，同一対象にエッジを伸ばしているグループの属性に着目すると，異なるキーワードであっても，職業か年齢層が一致するグループの類似度が高くなる傾向があることがわかった．一例としては，東芝・パナソニックそれぞれの「専業主婦（主夫）・20 代」層の類似度が高く，共通して「番組」「表」を特徴語としている部分や，同じくそれぞれの「専業主婦（主夫）・30 代」層が共通して「テレビ」「洗濯」「音」「掃除」という単語を特徴語としている部分などは特に顕著である．

次に，エッジ数が多い特徴語に着目すると,「掃除」「機」「テレビ」に関しては個別に可視化を行った際にも抽出されていたが,「録画」「番組」といった語は，個別の可視化時には出現しなかった語である．これらは，個別に可視化を行った際にはそれぞれグループごとの独立した特徴として存在していたものが，別のキーワードの不満グループにも似たような形で存在していたことにより，可視化されたと考えられる．

一例として,「東芝 パート・アルバイト 40 代」層は東芝の個別の可視化の際には類似度の高い別グループが存在しなかったが,「パナソニック・専業主婦」の「30・40 代」層との類似度は高く，可視化の対象に含まれた．このことから，一つのキーワードに対して作成された不満グループで他に類似するものがないグループであっても，別のキーワードに対して作成されたグループとは類似する場合があることがわかる．

6. まとめと今後の課題

本研究では,「不満データセット」を利用し，不満を投稿したユーザの年齢・職業を基準とした不満データのグループ化と，グループ間の関係性を可視化するシステムを試作した．

試作したシステムでは，実際にキーワード 2 つを用いて類似グループの可視化実験を行った．結果として，複数キーワードで作成したグループの中で比較を行うことにより，横断的に類似関係を確認することができるようになった．結果としては，異なるキーワードのグループで類似度が高いものはそのユーザ属性の類似度が高い場合が多いということ，単独キーワードの中で独立しているグループが他のキーワードのグループと比較すると類似度が高くなる場合があることを確認した．

今回試作したシステムによる実験の結果を踏まえて，今後の課題をあげていく．

まず，可視化システムの「検索」段階において，現状では企業名の表記ゆれや，検索単語が複数の企業名などに含まれている場合などの検索に対応できないため，検索時に取得しているデータの正確性には少なからず疑問がある．この点については，正規表現を利用可能な検索システムに改良するなどの解決策が考えられる．

また，現状では複数キーワードでの可視化を行う際に使用する単語の組は純粋な人手によって検討・入力しているが，この場合，可視化を行っても類似度の高いグループが存在しなかったり，あまり有効な可視化とならない場合が多くある．この点については，一つのキーワードを人手で入力すると，もう一つのキーワードとして入力した場合に有効な結果が出やすいと推測される別のキーワードを推薦するなど，検索をアシストする機能を追加する必要があると考えている．

次に，可視化における最終的な出力結果については大いに改善の余地がある，現状では類似グループと特徴語が多くなるにつれてノード数・エッジ数が増加していくため，複数キーワードの可視化結果はしばしば各グループの関係性を把握することが難しい複雑なものとなってしまう．可視化において特徴語として抽出している単語についても，現状では単純な形態素解析の結果から計算された TF-IDF の値だけを基準としているため，意味の分かりづらい単語や，分かち書きに失敗した結果と思われる不適切な単語などがしばしば特徴語として登場している．これらの点については，何らかの手法により，特徴語として抽出する単語についてフィルタリングなどを行ったり，ノードを統合するなど，可視化の結果をシンプルかつ理解しやすいものとする必要がある．

一方，先に述べた結果の単純化と相反する事柄ではあるものの，現状の可視化では，特徴語として抽出された単語がどのようなニュアンスで言及されているのか分かりづらいという問題が存在する．例として，図 8 では「テレビ」という単語が特徴語になっているものの,「テレビ」の何について，どのように不満を持たれているのかは把握することができない．これを解決するためには，特徴語を抽出する際に，その語と同時に現れている動詞や形容詞などを同時に抽出・可視化に利用することで，不満の発生したシチュエーションやニュアンスをより深く理解できるのではないかと考えられる．

根本的な可視化の方針についても，今回の研究においては様々な職業・年齢のユーザをグループ化し，グループごとの類似度を比較する方式とした．これとは異なる方針として，例えばユーザの傾向が専業主婦・女性などに偏っていることを逆に

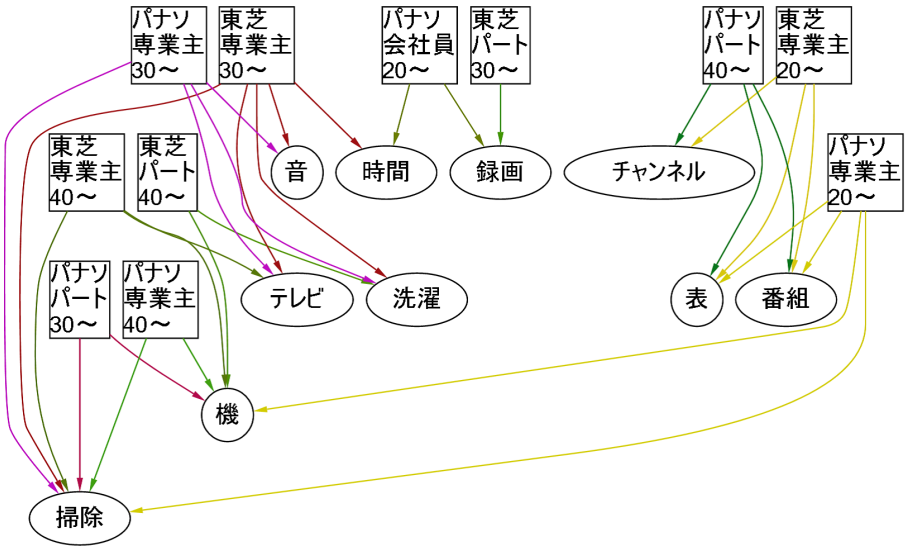

図 8 「東芝」「パナソニック」の可視化結果

利用し,「専業主婦」ユーザだけに注目して，特定のユーザ層が不満の対象ごとにどのような傾向の不満を持っているかの違いなどを可視化するという方針や，一旦作成した不満グループに含まれるユーザが，他にどのような対象に不満を持っているかを可視化するなどの方針も考えられる．

こうした点を踏まえ，今回作成した可視化システムの改善や，異なる方針での可視化などを今後実施・検討していく必要がある．

謝　　辞

本研究では、株式会社不満買取センターが国立情報学研究所の協力により研究目的で提供している「不満調査データセット」を利用しています．ここに謝意を表します．

本研究の一部は，平成 28 年度科研費若手研究 (B)(課題番号：15K16091) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, and Tomoya Mizumoto. FKC corpus: a japanese corpus from new opinion survey service. In *proceedings of the Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results*, pp. 11–18, 2016.

[2] 三澤賢祐, 田内真惟人, Mathieu Domoulin, 中島正成, 水本智也. ネガティブ評判情報に特化したコーパスの構築と分析. 言語処理学会第 22 回年次大会 発表論文集, pp. 501–504, 2016.

[3] 三澤賢祐, 田内真惟人, Mathieu Domoulin, 中島正成, 水本智也. 意見投稿プラットフォームにおける意見クラスタリングの試み. 言語処理学会第 22 回年次大会 発表論文集, pp. 1037–1040, 2016.

[4] 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析. 情報処理学会研究報告自然言語処理（NL）, Vol. 2004, No. 47, pp. 89–96, may 2004.

[5] Lev Nachmanson, George Robertson, and Bongshin Lee. Drawing graphs with glee. In *Graph Drawing 2007*, p. 12, June 2007.

[6] 相澤彰子. 語と文書の共起に基づく特徴度の数量的表現について. 情報処理学会論文誌, Vol. 41, No. 12, pp. 3332–3343, dec 2000.

[7] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. テキストマイニングによる評価表現の収集. 情報処理学会研究報告自然言語処理（NL）, Vol. 2003, No. 23, pp. 77–84, mar 2003.

[8] 磯島昭代, 野中章久, 清野誠喜. テキストマイニングによるクレームデータの分析. 農業経営研究, Vol. 42, No. 1, pp. 148–152, jun 2004.

[9] 坂井俊之, 藤村考. ブログに記述された不満表現からの潜在ニーズの発見. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3806–3816, dec 2011.

[10] 永井明人, 増塩智宏, 高山泰博, 鈴木克志. インターネット情報監視システムの試作. 情報処理学会研究報告自然言語処理（NL）, Vol. 2003, No. 23, pp. 125–130, mar 2003.

[11] 中野裕介, 湯本高行, 新居学, 佐藤邦弘. 商品レビュー要約のための属性-意見ペア抽出. 研究報告データベースシステム（DBS）, Vol. 2014, No. 15, pp. 1–7, nov 2014.

[12] 平山拓央, 湯本高行, 新居学, 佐藤邦弘. 語の共起と極性に基づく商品レビュー閲覧支援システム. 研究報告データベースシステム（DBS）, Vol. 2012, No. 3, pp. 1–9, nov 2012.