人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第7回)
SIG-AM-07-01

表形式でインタラクティブにクラスタリングを行い，可視化するツールの開発と実践

伊藤貴一[1]　白土由佳[2]　熊坂賢次[3]

Takaichi Ito[1], Yuka Shiratsuchi[2],Kenji Kumasaka[3]

[1]慶應義塾大学院政策メディア研究科
[2]産業能率大学
[3] 慶應義塾大学環境情報学部

Abstract: 多数のアイテムとその関係を可視化するために，縦の列は同一クラスタ，横の行は同じくらいの頻度という表形式で Web ブラウザ上に表示するツールを開発した．このツールは，クラスタリングもでき，UI の操作により分析者の考えを反映する制約付きクラスタリングを可能にしている．このようなツールを開発し，実際にデータを分析した．

1.はじめに

インターネットが社会に浸透するに従い、多くの人が Blog や SNS を使うようになり、人々のライフスタイルを Web サービス上に表明するようになってきている。このような状況下、社会調査も、アンケート調査をするというものだけではなく、Web にある人々の声を拾う、ソーシャルリスニング[1]が一分野になってきている。ソーシャルリスニングのためのツールが求められている。

このようなツールは、あらかじめ明確な答えがないため、探索的アプローチになってしまう。そのため、探索を支援するツールであるべきだ。この探索のためには、機械処理の結果を見せるだけのシステムではいけない。人間の背景知識や分析意図を結果に反映させるようなものでなくてはならない。そのため、インタラクティブ性は重要であり、人間とデータと機械処理が融合するような、知的インタラクティブシステム[2]である必要がある。

2.ツールのコンセプト

この論文のツールで扱うのは、バスケット分析の可視化である。商品の購買履歴や、自然言語を形態素解析後のデータを用いた分析である。共起関係に基づきアイテム間の関係を可視化する。これにより、商品購買なら、購買の関係図、自然言語なら、言葉の関係図を作り、データにある構造を読み解くことができるというものである。

このようなデータを分析するために筆者は、縦列をクラスタ、横列を頻度のレイヤーの二軸を使い、表形式でインタラクティブに表すツールを作成した。これは、第二著者の論文において社会分析に使われた手法[3]のツール化である。

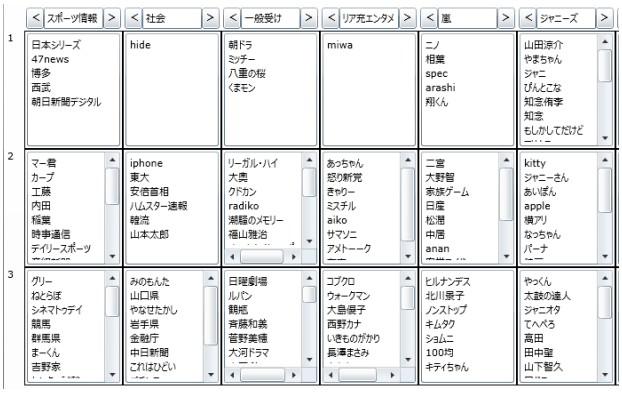

Fig.1

最終的には、Fig.1 のような可視化を行う。
以下、作成したコンセプトを示す。

2.1 べき乗分布と頻度の層（レイヤー）

商品購買履歴のようなバスケット形式のデータを分析するとき必ず発生するのは、少数の高頻度のものと、多数の低頻度のもので構成されるべき乗分布になることである。自然言語処理の世界では、ジップの法則[4]と呼ばれるものであり、マーケティングの世界では、ロングテール[5]と呼ばれるものである。このような分布は必ず発生するものとして、分析に予め組み込む必要がある。べき乗分布の性質として、両対数グラフを作ると線形に近似するというものがある。（べき乗分布は反比例に近似し、ｘｙ＝α を両対数にすると、log(y)=-log(x)+log(α）となり直線となる。）これを利用して、最大の頻度の対数と、分析に使う最小の頻度の対数の差をとり、それを等分割することで、頻度の層（レイヤー）を作るとい

うことをする。これは、上のレイヤーからピラミッド状にアイテムの個数が増えていくものとなる。これを行にして、上の行は頻度の高いものであり、下にいくと頻度が小さいものと、直感的にわかるものとした。

2.2 インタラクティブな関係の表示とクラスタリング

関係性の表示のために、アイテムのクリック時、関係が強いアイテムに色を付けるということをしている。これは、グラフにおける、エッジを、インタラクティブに見せていることに相当する。そのため、複雑な模様になってしまいがちなネットワーク図と同じ情報をすっきりと見せるようにしている。縦の列で、なるべく関係の強いもので固めるという形で、クラスタリングを行う。教師なしで、関係のみを用いで行うため、これは自己組織化させているともいえる。

Fig.2

Fig.2 において、赤は選択したアイテムであり、橙色は、関係しているアイテム。薄い橙は、弱い関係である。ネットワークを可視化しているといえる。

2.3 概念化とメタ認知

このようなクラスタリングによる、データの可視化だけでは不十分である。社会学的分析にするためには、概念化が必要である[6]。概念化とは、分析の全体像を考えることであり、それにしたがって、クラスタを表す言葉を探し、名付け、その塊を分析者が把握することである。そのような概念化をすることで、事象の解釈が可能になる。この行為は、認知科学的に言えば、メタ認知による言葉での外化である[7]。そのため、クラスタに名前をつけられるようにしている。

2.4 制約付きクラスタリング

名前をつける時に困るのは、データに忠実で機械的な処理に基づくクラスタリングの結果では、人間が考える概念とは、しばしばズレることである。このズレを解決するためには、人間の背景知識をクラスタリングの結果に反映する制約付きクラスタリング[8]という手法を用いる。制約付きクラスタリングとは、MustLink、CanNotLink を予め指定し、その情報を付加した上でクラスタリングする手法である。ここでは、高間[9]のように、グループ指定による、制約情報を UI の上で加えるようにする。ユーザが直感的に行えるように「固定する」というメタファで説明している。UI により固定化したアイテムは、クラスタリング時には、動かなくなる。当然、動かないとしても、そのアイテムの関係情報は使って、他のアイテムには影響している。

2.5 クラスタリングの失敗の可視化

このクラスタリングは、必ずどこかのクラスタに所属させるハードクラスタリングのため、うまくクラスタスタリングできていないアイテムが発生することがある。このようなものは、複数のクラスタと関係をもつものであり、ネットワーク構造的にはハブである。ネットワーク分析ではハブの重要性はしばしば指摘される。特に、低頻度のハブは、KeyGraph で言う、赤ノードに相当し、KeyGraph の考案者である大澤の主張では、そのようなものにはチャンスが眠っているとされる[10]。このようなクラスタリングに失敗しているハブ的なものは、縦列に並んでいるものは、同じクラスタであるという、可視化のルールから外れるため、それには、赤丸をつけ可視化する。

2.6 クラスタ間の関係の可視化

縦と横の二次元の可視化では、クラスタ間がどのような関係になっているかがわからない。概念化により、概念同士の関係がどのようになっているかを知るためにも、クラスタをノードとして、グラフとして可視化した。この際、クラスタに対して適切な名前を与えていないと、機械的な名前になり、イメージ出来ないものになってしまう。このことでも、名前をつけることを促進させている。

2.7 属性情報の付加と可視化

データには、いつどこでだれがといった、５Ｗ１Ｈの情報が本来的にある。このような情報をテーブルに重ねあわせる仕組みを用意する。

3.実装

実装は、C#で行い、Silverlight というブラウザのプラグイン上で実行できるようにした。そのため、Windows と Mac のブラウザ上で実行できる。

Silverlight にしたのは、Windows と Mac 両方で実行できるということと、最新版への更新が簡易なこと、ブラウザ実行とはいえ、ローカルファイルを扱え、通常のアプリケーション同様のことができるからである。次のアドレスで公開している。
http://goo.gl/VuHWa6

3.1 入力ファイル

リレーショナル・データベースからの出力を入力に仮定している。そのため、入力ファイルが１つでは済まない。すべては、UTF8 でエンコードされたテキストファイルでなければならない。

- 構造用ファイル
- UserId と ItemId のデータ
- UserId と属性のデータ

これらの TSV（タブ区切りデータ）が必要である。さすがに３つのファイルを用意するのは大変なので、一つで済むような仕組みを検討している。

3.2 画面の説明

ツールの基本画面はこのようになっている。(Fig.3)

Fig.3

画像に振った番号にそって機能を説明すると、

1. ファイル関係。ファイルの入出力を行う。
2. 表示設定。表の大きさなどの設定。
3. アイテムのクリック時、表示する関係の数と指標の設定。
4. マウスのモード設定。デフォルトでは、選択であるが、移動に変更すると、アイテムが移動可能になる。削除にすると、削除できる。
5. 固定化モードの ON/OFF。ON にした時、アイテムの横にチェックボックスが現れ、チェックしたものは、クラスタリング時に動かない。制約付きクラスタリングための制約を与えることができる。
6. クラスタリングパネル。クラスタリングの設定と実行、経過の表示を行う。

結果画面は Fig.4 のようになっている。

行が、頻度のレイヤーを示し、上から頻度が大きいものから並んでいる。列が、クラスタを示し、基本的に、塊を形成している。列の上には、自分でクラスタの名前が書き込めるようになっている。また、列は左右に移動できるようになっており、解釈に最適な並びを探索することをできるようになっている。

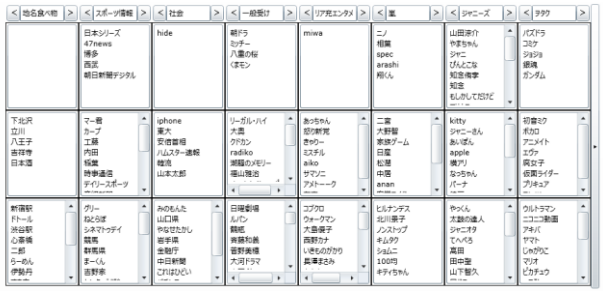

Fig.4

3.3 クラスタリングのアルゴリズム

クラスタリングのアルゴリズムは、可視化に合わせて作成した。K-Means 法の改変である。
クラスタ数は予め UI で指定する。固定化アイテムも予め指定してもいい。

1. すべてのアイテムを頻度レイヤーに沿ってランダム配置する。固定化されたアイテムは別である。
2. すべての非固定化アイテムにおいて、指定個数分の補正信頼度の高い順にアイテムを抽出し、それぞれのクラスタごとに補正信頼度の平均をとり、もっとも高いのを勝者クラスタとし、そこに移動させる。（ただし、移動は同一頻度レイヤー間で行い、移動情報は一時表に保存）
   (ア) 補正信頼度は、信頼度-支持度。Lift 値が割り算である代わりに引き算である。Lift 値は、割り算を使うので、値域が０から無限大までとるが、補正信頼度は、-1 から１の間で実データでは、-0.1~0.3 ぐらいの値に収まるため、扱いやすい。
3. 一時表に保存したものを本表にアップデート。固定化アイテムは同じ位置のままである。
4. 移動したアイテムが０なら終了、あるなら、2 に移動。
5. 1~4 を、指定回数繰り返し、評価値が最も高いものを表示させる。

このようなアルゴリズムにした。重み付きグラフのクラスタリングである。

3.4 クラスタリングの評価指標

クラスタリングの評価指標にはジニ係数を使う。ジニ係数は、格差を示す経済指標として有名だが、機械学習でも使われている。（例えば決定木[11]）。性

質としては、値の格差が大きいと１に近づき、格差が小さいと０に近づく。このジニ係数を使い、２つの軸を持って評価する。

1. すべてのアイテムが縦列で、まとまっていることがいいクラスタリングである。
   (ア) すべてのアイテムで、指定個数分の補正信頼度の高い順にとり、それをクラスタごとに総和を求める。これを変数としてジニ係数を求める。
   (イ) 求まったそれぞれのジニ係数の平均値を出す。１に近いほどいい。
2. 可視化として、それぞれのクラスタに入っているアイテムの数が均等に近いのがいいクラスタリングである。
   (ア) 頻度レイヤーごとに、クラスタごとのアイテム数を数え、それを変数としてジニ係数を求める。
   (イ) 求まった各頻度レイヤーごとのジニ係数の相乗をだす。０に近いほどいい。
3. １と２の２つを掛けあわせたのを最終的な指標とした。ただし、１と２は向きが違うので、向きを揃えた。

１は、ツールでアイテムそれぞれをクリックした時、関係しているアイテムが表示されるところが、縦の列でなるべくまとまっているというのを表現している。しかし、１だけでは、不十分だった。この指標を最大化するには、クラスタの空欄を増やせば増やすほど高くなるため、クラスタ数の指定が意味を成さないことが判明した。そのため、なるべく空欄をださないように、それぞれのクラスタに均等になったものがいいクラスタリングと評価されるように２の指標を追加した。

3.5 クラスタリングの失敗の検出

１の評価指標は、一つ一つのアイテムにおいて、１に近づけば近づくほど、クラスタとしてまとまっていることを意味する。逆に、０に近いものは、複数のクラスタと関係を持っているアイテムであるということ、すなわち、ネットワーク構造上、ハブになっているものだと思われる。縦の列でクラスタを作っていることが可視化のルールなので、そのルールから外れているので、このようなものを可視化する。具体的には、１の平均値を求める前のデータで、ジニ係数が低い物順に指定個に対して、赤い丸をつける。

3.6 クラスタマップの作成

表形式では、クラスタ間の関係がわからない。そのため、クラスタをノードとした、グラフを作成した。クラスタ間の関係には Jaccard 係数を使い、エッジの足切りには、Lift 値を使った。Lift 値の足切りにより、エッジの数を増やしても、グラフは完全グラフにはならないようにした。また、ノードの名前は、名付けていないと、素っ気ない機械的な名前にすることで、積極的にクラスタに名前をつけることを促進している。 また、関係性の実データが見えるようにもしている。

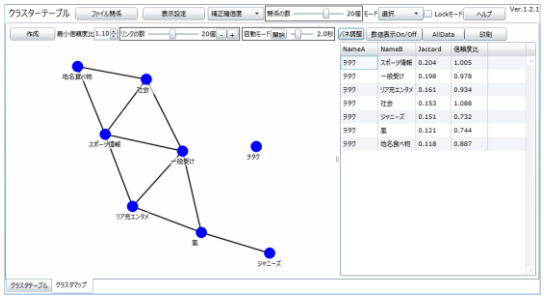

Fig.5

4.分析事例

2013 年、テレビドラマの「半沢直樹」は、最終話の視聴率が 42.2%（ビデオリサーチ調べ）という空前のヒットを飛ばした。この「半沢直樹」についての調査を行った。

分析データは、ツイッターで、「半沢直樹」の公式アカウント(@Hanzawa_Naoki)をフォローしているユーザ（45,315 人）のツイートを 2013 年 11 月に取得した。クリーニングとして、オープンであり、言語が日本語であり、ツイート数が 2000 以上のユーザを使った。約 11000 人に絞られた。そのツイートの中から、形態素解析を行い、頻出語 250 語を抽出し、その頻出語を用いて、バスケットを作成した。

また、そのユーザが特につぶやいた「半沢直樹」の俳優名と役名を属性とした。

このデータを用いて、本ツールを使い分析を行った。

まず、はじめにクラスタリングを行った時、ジャニーズと嵐が混ざった感じの大きいクラスタを形成していた。この２つは当然結びつきが強いが、数として大きすぎるので、２つを分けるために、次のように固定化行った。嵐のメンバーと、中居くんなどの他のジャニーズのメンバーを分けるようにした。

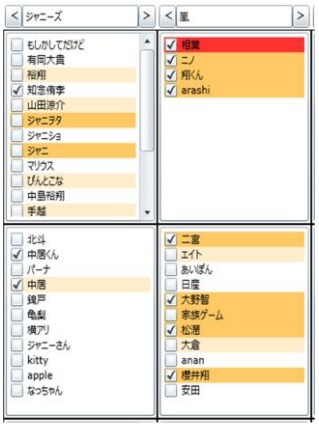

Fig.6

Fig.7

最終的には、クラスタに名前をつけて、Fig.7 のような結果になった。「半沢直樹」はテレビドラマであり、テレビドラマ的な要素が大きいことがわかる。その中でも、「おっさん向けテレビ」の大河ドラマ、朝ドラが好きな層と「若者向けテレビ」の若手のお笑いタレントが集まるクラスタと、「嵐」「ジャニーズ」のクラスタが発生したことがわかる。

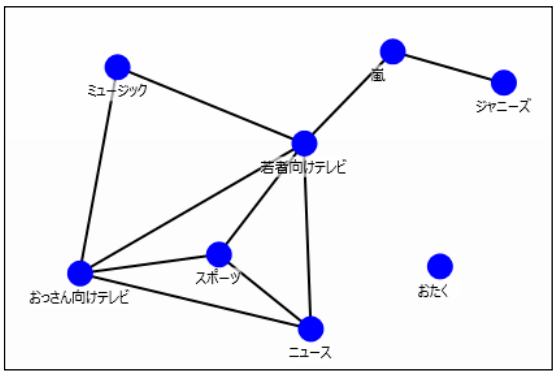

Fig.8

クラスタマップを作るとこのようになる。全体がつながるようにと Lift 値 1.1 で作成した。しかし、「おたく」のクラスタは、テレビ関係のクラスタとは強い関係性はなく、独立関係となった。ジャニーズも嵐を媒介項として全体像とつながっていることがわかるのも面白い。

次に、属性で見る。これらのクラスタは、どの俳優役名と関係が強いか？をみる。赤いところが、その俳優で特化しているところで、青色が特化していないところである。堺雅人、壇蜜、大和田常務の結果を示す。主役の半沢直樹は、タイトル名であるため、分析にはそぐわない。

Fig.9 堺雅人

Fig.10 壇蜜

Fig.11 大和田常務

主役を演じた、堺雅人が特化しているのは、ドラマクラスタだとわかる。一方、女優の壇蜜が、特化しているのは、ニュースクラスタであり、普段ニュースについてつぶやいている層に壇蜜は受けたというのが想像できる。当然、ジャーズクラスタには不人気である。

クラスタマップにおいて、独立だった、「おたく」クラスタのみ特化していたのは、大和田常務である。大和田常務は、その顔芸がネットでヒットして、ネット上のまとめサイトで、いろいろな形でまとめられており、その影響だと思われる。

このように考えると、ドラマ「半沢直樹」は、国民的ヒットなっていったことは、おぼろげながら見えてくる。つまり、普段からドラマを見ている人たちを惹きつけ、ジャニーズ出演でジャニーズ好きな人たちを惹きつけ、壇蜜で、普段、ニュースをつぶやいている人たちを惹きつけ、大和田常務で、テレビドラマを見ない、ネットだけを見ている層を惹きつけることに成功したことが、大ヒットに繋がった、ということが推察される。

このような分析ができるのも、このツールだからこそのものである。

5.議論

クラスタリングは、ランダムな初期配置から作成していくアルゴリズムであるため、同じデータであれば、同じ結果を必ず保証するものではない。また、クラスタリングとしては、大雑把なクラスタリングのため、精密なクラスタリングのために、制約付きクラスタリングの枠組みを利用しているものとなっている。クラスタリングには正解はないとはいえ、分析者の能力に依存するところが大きい。

知的インタラクティブシステムのために、システムとして最小のユーザフィードバックで済むことが望ましいとされる[2]。このツールの場合、機械的な仕組みとして、最小のユーザフィードバックをサポートするような仕組みは存在しない。しかし、意味的な側面と可視化としてのサポートはある。それは、クラスタに名前をつけるのだから、意味合いとして大きい高頻度のアイテムを固定化すべきという意味的な要請と、アイテムをクリック時の関係の表示で、すでに相互に関係があって塊を形成しているものに対して固定化をしてもナンセンスであるということである。そのため、固定化すべきものは、UI 的におぼろげながら示していると言える。しかし、これも、分析者の能力への依存が大きく、初めて使うユーザにとっては不親切であり、何かしらの改善の余地はあるだろう。

とはいえ、ツールの使用者に聞くと、初めに出力される結果にある程度満足してしまうようである。そのため、固定化による制約付きクラスタリングは、アドバンストな機能であるといえる。しかし、このような手段が存在するかしないかでは、分析者の選択肢が増えることなので有効である。

本論文で示した、インタラクティブなツールは使わないと良さがわからない。ツールは Web ブラウザ上で動かせるので、ぜひ使って分析してみて欲しい。

参考文献

[1] 萩原 雅之,次世代マーケティングリサーチ, ソフトバンククリエイティブ, 2011

[2] 岡部正幸 山田誠二,知的インタラクティブシステムにおけるインタラクションデザインとは何か,2013,JSAI

[3] 山崎 由佳,熊坂 賢次,共有化と生活化から生成される2つの"かわいい":4ファッションスタイルをめぐるネットコミュニティ分析 ファッションビジネス学会論文誌 1348-9909 ファッションビジネス学会 ,2012

[4] Zipf, G. K, The Psycho-Biology of Language, Boston-Cambridge Mass. Houghton Mifflin.,1935

[5] Chris Anderson, "The Long tail," Wired, 2004.

[6] 熊坂賢次, 山崎由佳,"ソーシャルな時代・柔らかい構造化手法 そしてライフスタイル論", AD・STUDIES, Vol.40 Spring, 2012.

[7] 諏訪正樹 身体知獲得のツールとしてのメタ認知的言語化, 人工知能学会誌, Vol. 20, No. 5, pp. 525-532..2005

[8] Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." ICML. Vol. 1. 2001.

[9] 高間康史, 三宅遼祐,グルーピング操作に基づくインタラクティブなクラスタリング対制約生成手法の考察, 第 27 回人工知能学会全国大会, F4-OS-04-3,2013

[10] 大澤幸生 チャンス発見の情報技術、東京電機大学出版局,2003

[11] L.Breiman, J.H.Friedman, R.A.Olshen and C.J.Stone, "Classification and Regression Trees", Wadsworth ,1984