インタラクティブにクラスタリングを行い、構造化を行うツールの開発と実査

伊藤貴一*1・熊坂賢次*2
Email: takaichi.ito@gmail.com

*1: 慶應義塾大学 SFC 研究所
*2: 慶應義塾大学環境情報学部

◎Key Words ツール開発, クラスタリング, 情報可視化

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。

このような目的のために、以前に私は、「ひっぱるくん」を開発した[3]。本論では、共起関係のデータを用いてその関係の構造化を行うツールとして、「こうぞうくん」を開発した。

2. ツール開発のコンセプト

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、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、山崎[4]の論文において社会分析に使われた手法のツール化である。

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 べき乗分布とレイヤー

自然言語でのジップの法則[5]、マーケティングで言うロングテール[6]は、べき乗分布ゆえの現象である。べき乗分布を分析するのには、ピラミッド状にデータを分割する必要がある。このために、レイヤー分割ということを行う。頻度の最頻値と、最低値の対数を取り、それを当分割したものを区切りとして、頻度の層を作る。このようにすると、一番上のレイヤーが、メジャーなもの、一番下のものがマイナーなものとなり、配置により、データの規模感を把握できるようにしている。

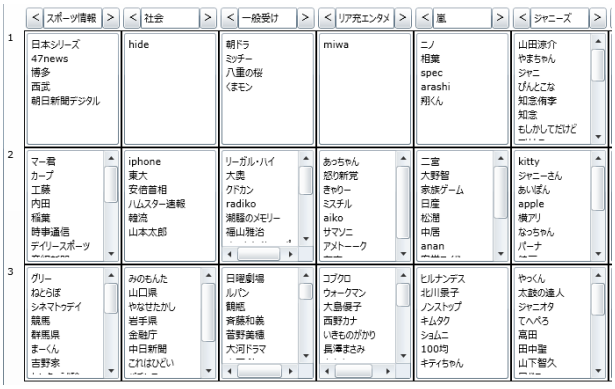

Fig.1

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 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。 ただ、クラスタリングした結果を見るというものではなく、その結果から、分析者の考えによって、違うところにアイテムがあると収まりがいいと思った時、それを移動することができるようにしている。

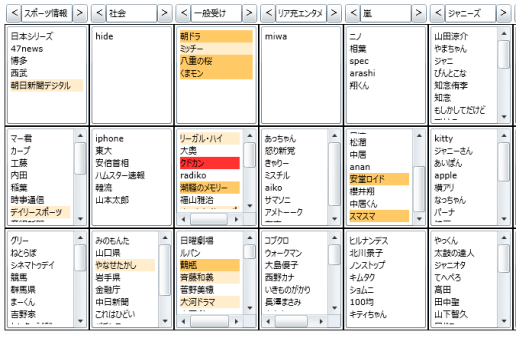

Fig.2

2.3 制約付きクラスタリング

分析者自身がアイテムの移動を行うことができることを積極的に組み込むために、制約付きクラスタリング[7]の仕組みを取り込んだ。制約付きクラスタリングとは、半教師付きクラスタリングともいわれ、少しの教師データを加えた上で、クラスタリングを行うものである。ここでは、場所を移動しないという制約を加える事をできるようにしている。このようにすることで、クラスタリング時に、必ず一緒のクラスタになるもの、必ず違うクラスタになるものを指定することができる。Fig3.のチェックを入れたアイテムが動かないと指定できる。 このようにすることで、分析者自身が考えるクラスタの塊に近づける形でデータとの対話を可能とする。

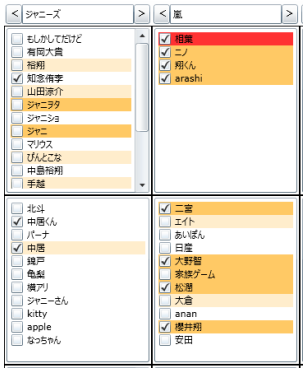

Fig.3

2.4 概念化、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[8]。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[9]。そのため、クラスタに名前をつけられるようにしている。

2.5 マップ化による可視化

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、グラフを作成できるようにしている。(Fig4)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。

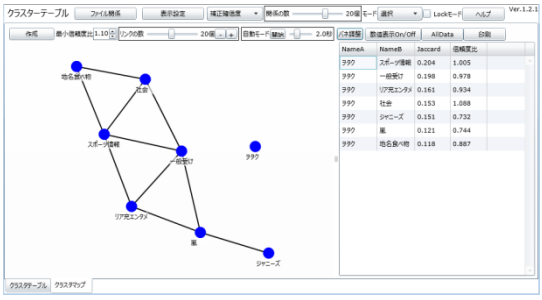

Fig.4

3. 実装

以上のような、べき乗分布、インタラクティブ性、制約付きクラスタリング、概念化を行い、概念間の関係の可視化を行うツールを作成した。プログラミングを言語は、C#で、Silverlight で作成し、Web で公開している。URL は http://kiichi.azurewebsites.net/ である。Silverlight のため、プラグインが使える状態にあれば、簡単にブラウザ上でツールを使うことができる。このため、授業などの場では、インストールトラブルなく、使うことができた。

4. おわりに

このツールを使い、実際にデータを分析した。対象としたのは、ミクシーのコミュニティのデータや、クックパッドのレシピのデータ、ツイッターでのつぶやきの分析を行った。探索的に、データから概念的な塊の把握に使えることが確認された。また、このツールを使った授業も行った。ブラウザ上で行うことができるので、統一した環境での演習を行うことができた。学生たちのツールの習熟には、時間がかからず、データの中から概念を見つけ出すという本来的な作業に集中することができた。そして、分析した結果をプレゼンさせるということを行った。各々の試行錯誤し、考えた形跡が見て取れた。しかし、授業としての満足度を上げるためには、魅力的な分析用のデータを用意することの必要性と難しさを痛感した。 この概念化のためには、簡単に答えがでてしまうのではよくないが、授業として扱うには、ある程度の簡単で理解しやすいデータである必要が有るためである。

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) 伊藤貴一,熊坂賢次、諏訪正樹、花房真理子, 自己探求する柔らかい構造化ツール『Hipparu-McS』の実装と評価, コンピュータ＆エデュケーション Vol. 29 ,2010

(4) 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

(5) Zipf, G. K, The Psycho-Biology of Language, Boston-Cambridge Mass. Houghton Mifflin.,1935

(6) Chris Anderson, "The Long tail," Wired, 2004.

(7) Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." ICML. Vol. 1. 2001.

(8) 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.

(9) 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