形式概念分析と自然言語処理による社会推論の試み

○尾山武史，秋葉健人，榊原一紀，中村正樹，本吉達郎，濱貴子 （富山県立大学）

概要 社会学の専門家自らの主張・言説について，その客観性・妥当性を，文献資料の解析を通して帰納的に検証することを試みる．具体的には，戦前期における職業婦人の社会進出に対する社会意識に関する研究を取り上げ，そこで対象としている雑誌『婦人公論』における記事を複数抽出する．専門家の作成したモデルにおける諸概念および社会意識を表す諸要素を形態素解析を用いて抽出し，その上で形式概念分析を用いて抽出要素および専門家による分類間の関係を束として表現・比較する．

キーワード: 社会推論，形式概念分析，自然言語処理

1 はじめに

社会意識を分析する際，内容分析と呼ばれる文章，音声，映像などの質的データの内容の分析がよく行われる．計算機の発達に伴って，計算機を用いた内容分析である計量的内容分析について，報告が多くなされており [1, 2, 3]，その応用分野を広げている．しかし計量的内容分析は，その分析手法自体が確立されているわけではなく，分析の対象や目的に応じて，手法を適宜設計する必要がある．そのため，従来からの専門家の洞察や知見に基づくアプローチは現在も主流であるが，分析対象の読解において研究者の恣意性が入っていないのか，あるいは誰が分析を行っても同じ結果が得られるのか，などの点において，分析結果の信頼性を疑われる場合が少なくない．吉川は文献 [4] において，「量的なデータ検証を行うことができず従来からのアプローチを行う分析の中には，印象に基づいた「空論」も含まれており，自由な分析によって，格式あった学問空間の秩序が失われてきている」という現状を述べている．

そこで本研究では，社会意識に関する研究者 (専門家) 自らの主張・言説を「専門家モデル」と呼び，文献資料の解析を通して，専門家モデルの客観性・妥当性を補強する手法を構築する．そのための一事例として本論文では，「戦前期の女性の社会進出に対する意識形成」に関する専門家モデルを研究対象とし，検証方法を提案する．その際，専門家が分析対象として用いた資料 (文献) に対して自然言語処理を施した上で，専門家が各文献に付与した属性 (あるいは分類) を自然言語処理された結果と合わせて形式概念分析 (Formal Concept Analysis: FCA)[5] の適用対象とすることにより，この分類の正当性を束 (Lattice) の上で可視化する．FCA は分析対象を，対象がもつ性質を列挙した上で，これら性質の間に成立する包含関係を束の形で提示する方法であり，本研究では，FCA により文献から抽出された名詞句と専門家が導入した分類の間に成り立つ包含関係を比較・検討する．

2 対象とする専門家モデル

本章では，分析対象とする社会意識形成過程として，戦前期における職業婦人の社会進出に対する社会意識に関する研究を取り上げる [6, 7]．

2.1 戦前期における職業婦人に対する社会意識

2.1.1 歴史社会学におけるアプローチ

日本において戦前期とは「第一次世界大戦を契機とする産業化の進展や第三次産業の拡大に伴う安価で柔軟な労働力需要の高まりによって，職業婦人の数が飛躍的に増加した」時代 [6] と捉えることができる．さらにこの時代には思想的な面において，近代的な性別役割分業を特徴とする「良妻賢母思想」と女性個人としての自由や平等 (権利) の獲得を特徴とする「女性解放思想」とが存在していた．

時代の移り変わりの中で，職業婦人のイメージがどのように形成され，変化していったのかを明らかにするために，歴史社会学を専門としている研究者である濱は，戦前期において主に中流家庭の女性たちが主な読者であった婦人雑誌『婦人公論』を資料として検討を行っている [7]．

2.1.2 分析対象

分析対象としている婦人雑誌『婦人公論』は，1916 年に中央公論社によって創刊され，1944 年 4 月に総合雑誌『中央公論』に統合されるまで継続的に発行されていた婦人雑誌である．「自由主義の旗印の下，女権拡張を主張して誕生した」[8] 婦人雑誌の中でも，「教養派」を代表する雑誌 [9] であると評されている．

濱は分析期間として

- 第 1 期 1916 年 ～ 1927 年 (創刊から『婦人公論』の大衆化が行われるまで)
- 第 2 期 1928 年 ～ 1937 年 8 月号 (大衆化が行われてから，戦争が本格化するまで)

を対象とすることにより，戦前期の「平時」における職業婦人に対する社会意識の変遷明かにしようとしている．

一方『婦人公論』は婦人向けの総合雑誌であり，職業婦人のみを扱った雑誌ではない．そこで，言説の面から職業婦人について検討するために，この期間における雑誌の記事のうち，以下の条件を満たすものを資料として抽出している:

(a) 記事名に「職業婦人」という語句を含むもの

第15回社会システム部会研究会，第59回システム工学部会研究会 (2018年3月14日-16日・沖縄) PG0002/18/0000-0024 © 2018 SICE

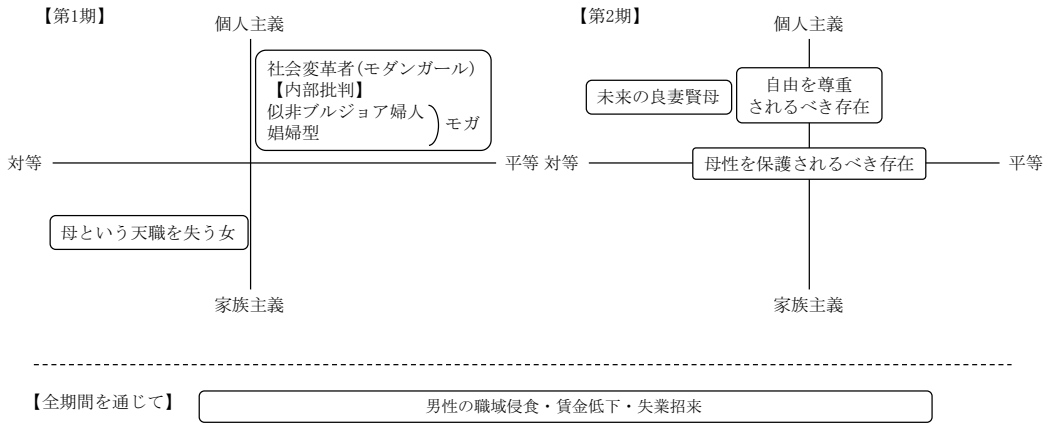

Fig. 1: An image of career women appeared in the articles and its transformation[7].

(b) 「職業婦人」という語句以外で，記事名に「女性の呼称[1]」と「職に就いていることが伺われる単語[2]」を含むもの．ただし「共働き」「共稼ぎ」については女性の呼称がないものも抽出する．

(c) 『婦人公論』における職業婦人の主要な職種 11 種について
- 当該職業を記事名に含むもの
- 当該職業に就いている著名女性が取り上げられる記事
- 当該職業に就いている著名女性が自らの半生・経験を記した記事

『婦人公論』における職業婦人の主要な職業については，

- 全期間を通じて：事務員，教師，女給，タイピスト，俳優
- 第 2 期から：美容師，医師，記者，和裁，洋裁，手芸裁縫師，看護婦，店員

のようになっている．

2.2 職業婦人イメージに対する専門家モデル

濱は論説における職業婦人イメージとその変容の分類として，Fig. 1 に示すチャート提示している．本研究ではここに登場する分類を専門家モデル例として取り上げる．

Fig. 1 において，横軸は「男」と「女」についてどのように在るべきかという「平等」を表す観点を表している．とくに「平等」は，「男も女も同じものとして，同等の権利を与えるべき」というような男女平等を支持する論調の記事が該当するのに対して，「対等」は，性別分業のように，「男と女は別のものであるので，それぞれに適した事をすべき」という論調の記事が該当する．一方縦軸は「自由」を表す観点であり，「個人主義」は，個人の自立や自己の尊重を重視する論調の記事が該当するのに対して，「家族主義」は「家族のためにどう動くのか」という「動員」や「献身」などの家族を重視する論調の記事が該当する．

[1]婦人，女，女子，女性，女学生，母，妻など
[2]職，仕事，働き，稼ぎ，自活，高給，丸ビル

3 検証モデル

本研究では，**2.2** で示した分類を専門家モデルと呼び，この正当性を補強する検証モデルを示す．

本研究が提案する検証モデルを Fig. 2 に示す．

Fig. 2 では，専門家モデルとそれが対象とする社会意識について，それぞれから語録と資料を用意し，それらから同じ形式の要素を抽出する．ここで専門家モデルに対しては，専門家モデルを記述した論文が用意され，社会意識に対しては，専門家モデルが対象とする資料 (文章や音声，映像など) が対応する．つまり論文には文献[7]が，資料には『婦人公論』がそれぞれ対応する．

一方，抽出される要素としては，論文からは専門家モデルを構成する属性，すなわち，4 つの分類「対等」，「平等」，「個人主義」，「家族主義」を用いる．また『婦人公論』からは，各記事に登場する名詞句を用いる．これは，『婦人公論』における記事が対象とする話題を抽出するためである．

4 形式概念分析による検証

本章では，専門家モデルを検証するために用いた形式概念分析および分析対象の選択に用いるオブジェクト抽出法について述べる．

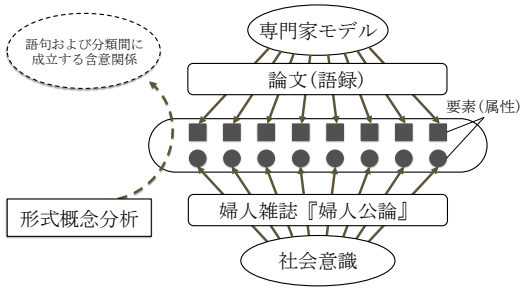

Fig. 2: A verification model for social discourses.

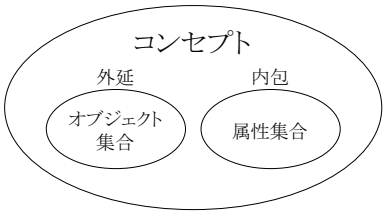

Fig. 3: Structure of concept.

4.1 形式概念分析

4.1.1 コンテクストとコンセプト

形式概念分析 (Formal Concept Analysis) は，Roudolf Wille によって提唱されたデータ分析手法の一つである [11]．分析対象を，対象がもつ性質によって表現し，これを分析することによって，性質間の関係に基づいて成り立つ，対象集合や性質集合の包含関係である含意論理 (対象が性質 A を持つならば性質 B も併せて持つ; $A \to B$ と表す) を抽出する．以下に形式概念分析で用いられる用語を定義する．

- オブジェクト G
  - 現象や事象などに出現する対象
- 属性 M
  - 対象の持つ性質
- コンテクスト $\mathbb{K}$
  - オブジェクトと属性の関係 $I \subseteq G \times M$ または $\mathbb{K} = (G, M, I)$
- 外延 A
  - 共通な属性集合を持つオブジェクト集合 $A \subseteq G$
- 内包 B
  - 外延集合が共通にもつ属性集合 $B \subseteq M$
- コンセプト (A, B)
  - 外延と内包の組
- コンセプトラティス $\underline{\mathfrak{B}}(G, M, I)$
  - コンセプトの完備束

形式概念分析におけるコンセプトとは，Fig. 3 に示すような外延と内包の組 (A, B) のことであり，コンテクストを $\mathbb{K} = (G, M, I)$ に形式概念分析を適用する事によって得られる．得られたコンセプトは，コンセプトラティス $\underline{\mathfrak{B}}(G, M, I)$ の構成要素となる．コンセプト全体は $\mathfrak{B}(G, M, I)$ で表す．

形式概念分析におけるコンテクストとは，オブジェクトと属性の関係のことである．コンテクスト $\mathbb{K} = (G, M, I)$ と表したとき，G と M はそれぞれオブジェクトと属性の集合を表す．I は付随関係と呼ばれ，G の要素であるオブジェクト g が M の要素である属性 m を持つことを表す (g, m) の集合である．これを表にまとめたもの (コンテクスト表と呼ぶ) の一例を，Table 1 に示す．コンテクスト表では，行にオブジェクトを，列が対象が内包する属性が並ぶように配置される．各行に示されたオブジェクトが各列の属性を持つ場合，対応する行と列にクロス (× 印) が記入される．例えばオブジェクト g1 に注目すると，属性 m2，m4，m5 を保有しており，そのコンセプトは ({g1}，{m2，m4，m5}) と表される．

Table 1: An Example of the Context Table

	m1	m2	m3	m4	m5
g1		×		×	×
g2	×				×
g3	×	×			×
g4	×		×		×
g5				×	×
g6	×	×	×	×	×

4.1.2 アソシエーションルールと含意論理

コンテクスト表において成立する，「属性 A を持つオブジェクトは属性 B も持つ」といった関係はアソシエーションルールと呼ばれ，属性 A を持つオブジェクトのうち，属性 B も持つオブジェクトの割合を信頼度と呼ぶ．例えば，Table 1 のコンテクスト表において，属性 m1 を持つオブジェクトは属性 m2 を持つといったアソシエーションルールは，属性 m1 を持つオブジェクトの 50%のオブジェクトにおいて成立している．このときの信頼度は 50%となる．一方，信頼度が 100%のアソシエーションルールを含意論理と呼ぶ．

4.1.3 コンセプトラティス

4.1.2 で示した，含意論理を束 (Lattice) として図したものをコンセプトラティス (Hasse 図) と呼ぶ．コンセプトラティスの一例を Fig. 4 に示す．

Fig. 4 は，Table 1 に成立する含意論理に対応している．コンセプトラティスにおける各ノードは Fig. 3 に示すようなコンセプトを表す．ここでノードの配置には，以下の特徴を持つ:

- 上方に配置するノードほど
  - コンセプトの外延に含まれるオブジェクトが多い
  - コンセプトの内包に含まれる属性は少なくなる
- ノードは以下に示すオブジェクトおよび属性を持つ:
  - 上方向にアークを辿ったときに通るノードのコンセプトが持つ全ての属性を内包に持つ
  - 下方向にアークを辿ったときに通るノードのコンセプトが持つ全てのオブジェクトを外延に持つ

さらにコンセプトラティスにおいては，上半分が青いノードは，省略されない新たな属性がコンセプトに含まれ，下半分が黒いノードは省略されない新たなオブジェクトがコンセプトに含まれることを示す．また，ノードの大きさは新たに含まれたオブジェクトの数による．

これらのことから，Fig. 4 において，「g3」と表示されているノードは，コンセプト ({g3，g6}，{m1，m2，

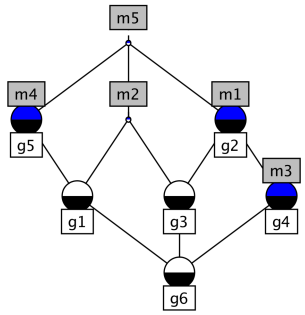

Fig. 4: An example of the concept lattice.

m5}) であることから読み取ることができる．また，前述の特徴を踏まえると，ノードを見ることによってオブジェクト間および属性間の包含関係を読み取ることができる．例えば，「g4」と表示されているノードを見た場合，オブジェクト g4 はオブジェクト g2 の持つ属性を全て持ち，オブジェクト g6 の持つ属性の一部を持っている．別の捉え方をすると，属性 m3 を持つオブジェクトは必ず属性 m1 を持つということが読み取ることができる．

4.2 オブジェクト抽出によるコンセプトラティスの簡素化

形式概念分析の問題点として，コンセプト数が多くなるとコンセプトラティスが複雑化し，描画することができなくなる点が挙げられる．たとえ描画できたとしても，含意関係を読み取ることが非常に困難である．複雑化する関係性を読み取るためにコンセプトラティスを簡素化する一手法として，氷山概念束 [12] がある．氷山概念束は，概念束の全体構造の概略と特徴を把握するために，ある個数以上の対象を含む外延のみで構成されるコンセプトラティスである．そのため，少数オブジェクト間で成立する関係性がコンセプトラティス上で失われてしまう可能性がある．

一方形式概念分析では，コンセプトラティスで示される属性間の関係は含意論理のみであり，信頼度が100%未満のアソシエーションルールは Hasse 図上で見ることはできない．ここで属性 A を持つオブジェクト集合の中で，属性 B を持っていないオブジェクトが一つでも存在すると，含意関係 $A \to B$ の信頼度が 100%を下回ることに注意されたい．つまり，多数のオブジェクト間で成立する含意論理が，少数のオブジェクトの存在のために見えなくなることを意味する．

本研究では，上記 2 点の問題点を踏まえ，コンセプトラティスの簡素化を行いながら，少数で成立するアソシエーションルールの可視化を実現する方法について検討する．つまり，オブジェクト (すなわち記事) の数が減ることで，コンセプトラティスは簡素化され，要素間の関係を読み取りやすくなることを実現しながら，それと同時に，記事の中には専門家モデル作成における諸概念や社会意識を表していないものがあると仮定し，このような記事を除外することで，より多くの含意論理を含意関係を可視化することを試みる．

具体的には，成立するアソシエーションルール数が最大となるように，式 (1) および式 (2) で表されるの最適化問題を求解する:

$$\text{maximize} \quad w_1 g^{\mathrm{C}}(\hat{O}) + w_2 g^{\mathrm{A}}(\hat{O}) \tag{1}$$

$$\text{subject to} \quad \hat{O} \subseteq O \tag{2}$$

$$|\hat{O}| \geq \bar{o} \tag{3}$$

ここで，O はオブジェクトの集合，$\hat{O}$ は部分オブジェクト集合，$g^{\mathrm{C}}(\cdot)$ は $\cdot$ の下で成立する含意関係数，$g^{\mathrm{A}}(\cdot)$ は $\cdot$ の下で成立するアソシエーションルール数，および $\bar{o}$ はオブジェクト数の上限をそれぞれ表す．また w_1，w_2 はそれぞれ，重み係数を表す．

5 検証結果

本章では，4 章で述べた手法を専門家モデルに適用し，妥当性を検証するとともに，得られたコンセプトラティスに基づく新たな分析結果をを示す．

形態素解析によって抽出した普通名詞および専門家モデルを用いて，コンテクスト表を作成する．このとき属性すなわち全記事中に登場する普通名詞数は 832 個となる．コンテクスト表は，オブジェクトとして婦人雑誌『婦人公論』の記事を，その属性として記事中の形態素および専門家モデルによる記事分類として作成する．検証に用いた『婦人公論』の記事は 31 件，そのうち **2.2** で示した分類に該当する記事は「平等」16 件，「対等」9 件，「個人主義」21 件，「家族主義」6 件となる．

5.1 コンセプトラティスの生成 (簡素化の無い場合)

Fig. 5 および Fig. 6 に，**4.1.3** に示す方法により，生成されたコンセプトラティスを示す．Fig. 5 と Fig. 6 は互いに同じコンセプトラティスを示しているが，前者には「#平等」(*equality*) のノードと，後者には「#対等」(*coordination*) のノードと，それぞれ接続するアーク群を青色で表記している．つまり，青色で記されているノード群は下方から上方に対して含意関係を有していることが分かる．Fig. 5 と Fig. 6 を比べると，互いに青色のノード群がコンセプトラティス上で別領域に配置されており，つまり，両分類に属する記事で用いられる語句は互いに異なることがわかる．このことは分類「家族主義」および「個人主義」間でも確認されている．

一方で，両図においても記事中の語句 (普通名詞) は下部の行に多く登場している．属性が下部の行に登場していることは，少数のオブジェクトに登場し，他の要素に包含されやすい要素であることを示す．加えて，専門家の分類の下部に登場し，分類と含意関係のある

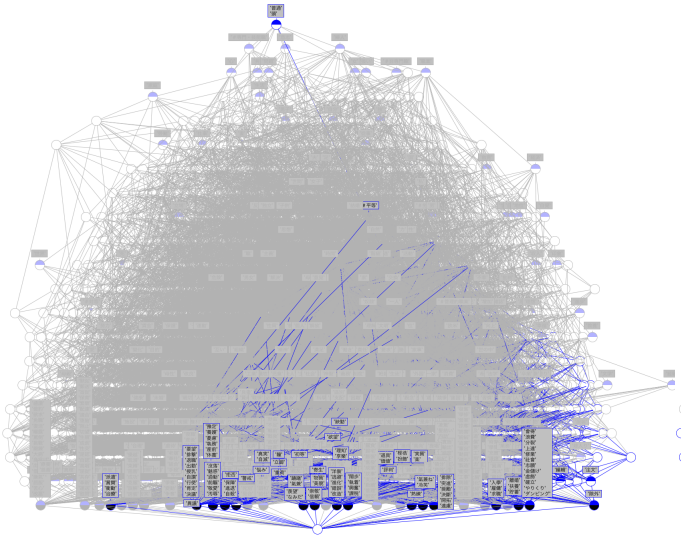

Fig. 5: The result of the concept lattice about *equality*.

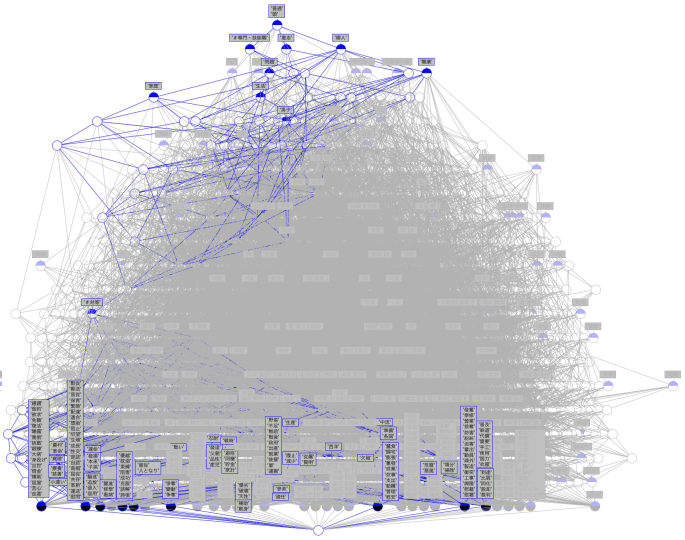

Fig. 6: The result of the concept lattice about *coordination*.

語句は，その分類にしか現れない語句である．例えば，「平等」の下部に登場し，「平等」と含意関係にある語句は，「対等」との含意関係には現れない．一方，上部に登場する語句であるほど，多くの記事に登場し，他の要素を包含しやすい．加えて，専門家の分類の上部に登場し，分類と含意関係のある語句は，他の分類にも登場し，含意関係を持ちうる．例えば，「個人主義」の上部に登場し，「個人主義」と含意関係にある語句は，「家族主義」の上部にも登場し，「家族主義」と含意関係を持ちうる．本検証においては，下部に現れる形態素は，記事特有の話題を示し，上部に現れる形態素は用いられやすい話題であると言える．

5.2 コンセプトラティスの生成 (簡素化の有る場合)

4.2 にて示したオブジェクト抽出法による，コンセプトラティスの簡素化の結果を記す．オブジェクト抽出法で最適化問題を解く際に，$w_1 = 0.5$，$w_2 = 0.5$ とし，アソシエーションルール数の計算には信頼度は90%以上のものを用いた．オブジェクト抽出法により，抽出したオブジェクト，すなわち記事数は 18 となった．このうち「平等」および対等に分類される記事は 8 件，「対等」に分類される記事は 8 件，個人主義 10 件，家族主義 5 件となった．一方，属性すなわち語句 (普通名詞) 数は 773 個となった．このときのコンセプトラティスを Fig. 7 および Fig. 8 に示す．**5.1** と同様に，分類「# 平等」および「# 対等」にそれぞれ接続関係を示すアーク群を青色に記してある．

Fig. 7 および Fig. 6 を Fig. 5 および Fig. 8 と比較すると，ノードおよびアーク数が減っており，また全体的に比較的縦長の形状になっていることから，簡素化が達成されつつ，多くの含意論理が抽出できていることが確認される．

これらの結果から，以下の点が確認される:

分類の下位に登場する語句

分類のノードと関わりがある語句が出現する Fig. 7 および 8 中では，分類の下位に位置する語句が複数みられる．分類の下位に出現する語句は，必ずその分類に属しているということが言える．例えば，Fig. 8 において，「#対等」のノードの下位に「良妻」が現れていることが見て取れる．これは，「良妻」が登場するならば，必ず，「#対等」に分類されることを意味する．すなわち，「良妻」について話題を出している記事は，「対等」に属する記事である．「#対等」の下位に「良妻」が出現するのは，「対等」が「男と女は別のものであるので，それぞれに適したことをすべき」という論調であると解釈できる．

分類の上位に登場する語句

分類の上位に出現する語句は，その分類に属するならば必ず出現する語句である．分類の下位に出現する語句とは異なり，分類の上位に出現する語句は他の分類においても出現しうる語句であるが，その分類の議論を展開するにあたって前提となる語句である．また，上位に出現する語句が多いほど，論旨を構成する語句が固定化していると見て取れる．例えば，「#家族主義」のノードの上位に「家庭」が現れるていることが見て取れる．これは，「#家族主義」に分類されるならば，必ず，「家庭」が登場するということを表す．すなわち，「家族主義」について論じるとき．必ず「家庭」についての話題を出しているといえる．「# 家族主義」の上位に「家庭」が出現するのは，「家族主義」が「家族のためにどう動くのか」という論調であると解釈できる．

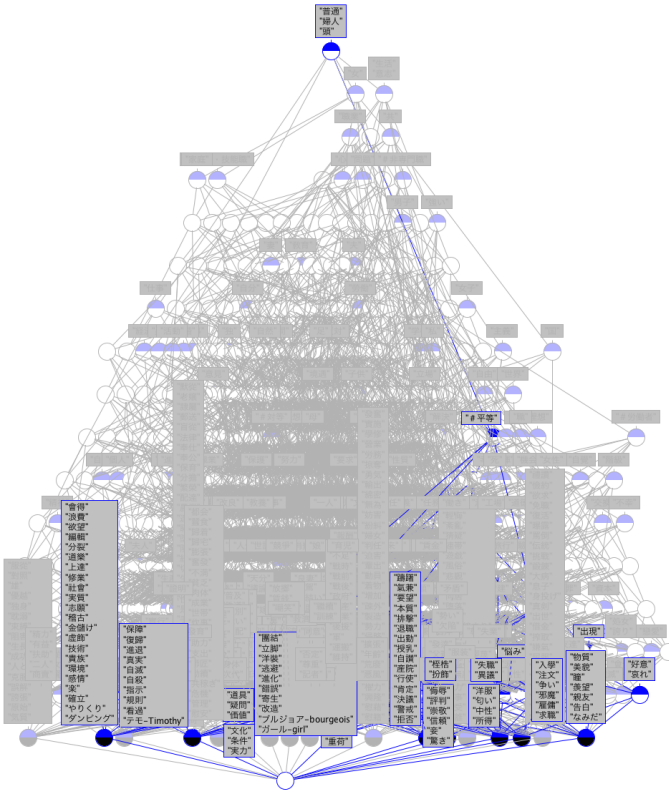

Fig. 7: The result of the concept lattice about *equality* (simplified).

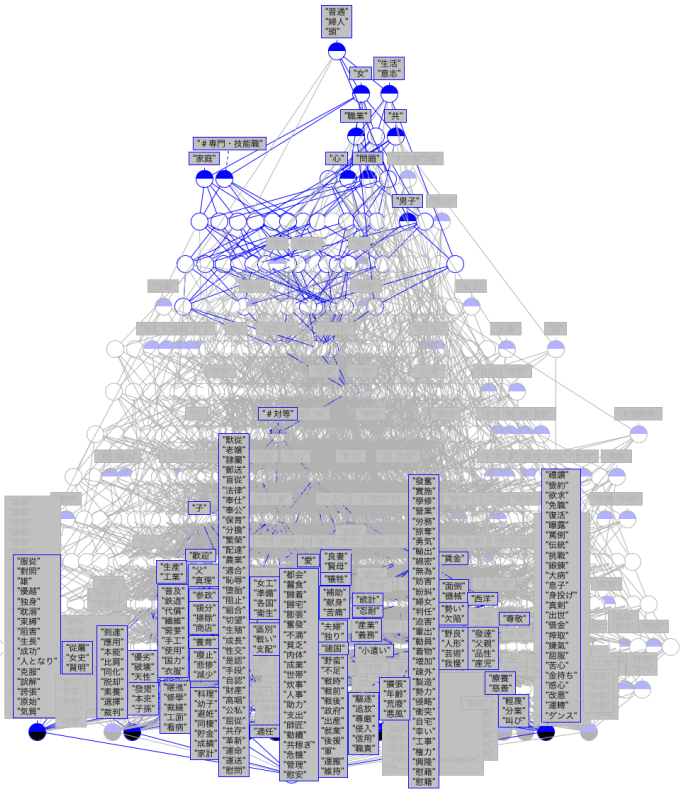

Fig. 8: The result of the concept lattice about *coordination* (simplified).

6 おわりに

社会意識に関する研究者 (専門家) 自らの主張・言説を，文献資料の解析を通して，その客観性・妥当性を補強する手法を構築した．具体的には，専門家が分析対象として用いた資料 (文献) に対して自然言語処理を施した上で，専門家が各文献に付与した属性 (あるいは分類) を自然言語処理された結果と合わせて形式概念分析の適用対象とすることにより，この分類の正当性をコンセプトラティス上で可視化した．この方法について，一事例として「戦前期の女性の社会進出に対する意識形成」に関する研究を取り上げ，適用結果を考察した．

今後の課題としては，形態素のみではなく，対象記事の文脈を考慮した解析手法の構築が挙げられる．とくに注目している語句が肯定的・否定的な係り受けがなされているかについて解析対象に加えることが考えられる．

参考文献

1) 工藤拓，松本裕治: チャンキングの段階適用による係り受け解析，情報処理学会研究報告情報学基礎 (FI)2001，97/04 (2001)

2) 小木曽智信，小椋秀樹，近藤明日子: 近代文語文を対象とした形態素解析辞書の開発，言語処理学会，第 14 回年次大会発表論文集，225/228 (2008)

3) 樋口耕一: 社会調査のための計量テキスト分析，ナカニシヤ出版 (2014)

4) 吉川徹: 現代日本の「社会の心」，有斐閣 (2014)

5) 鈴木治，室伏俊明: 形式概念分析 — 入門・支援ソフト・応用 —，知能と情報，19(2), 103/142 (2007)

6) 山﨑貴子: 戦前期における職業婦人の葛藤，日本教育社会学会大会発表要旨収録 (62)，216/217 (2010)

7) 濱貴子: 戦前期『婦人公論』における職業婦人イメージの形成と変容，富山県立大学紀要，第 26 巻，56/82 (2016)

8) 木村涼子: 「主婦」の誕生—婦人雑誌と女性たちの近代，吉川弘文館 (2010)

9) 岡満男: 婦人雑誌ジャーナリズム—女性解放の歴史とともに，現代ジャーナリズム出版会 (1981)

10) 松村明: 大辞林，三省堂，第三版 (2006)

11) R. Wille: Restructuring lattice theory: an approach based on hierarchies of concepts, *Springer Berlin Heidelberg*, 445/470 (1982)

12) G. Stumme: Computing Iceberg Concept Lattice with TITANIC, *Journal on Data and Knowledge Engineering*, 42(2), 189/222 (2002)