Vol.2017-MPS-113 No.30
Vol.2017-BIO-50 No.30
2017/6/24

意外性を持つ検索・推薦のための多重ラベル分類を用いた形式概念探索

大久保 好章[1]　原口 誠[1,a)]　劉 赫宇[1]

概要：ラベルが未知のクエリに対し，多重ラベル分類の手法を用いて決まる属性部分空間における近接性を用いて (暗に) ラベル推定を行う．それにより，クエリとデータベース中のオブジェクトは，推定ラベルに関する近接性と元の属性空間における近接性という，２種類の「距離」を持つことになる．本報告では，オブジェクトと属性の形式概念で，ラベルに関して遠く，その意味で意外性を持ち，かつ（元の）属性に関して近く，その意味でクエリーと関連性を持つオブジェクトを含む形式概念を算出する方式を与える．

Formal Concept Analysis for Finding Interesting Objects with respect to Label Information

Yoshiaki OKUBO[1]　Makoto HARAGUCHI[1,a)]　HeYu LIU[1]

1. はじめに

我々を取り巻く膨大なデータを有効に活用するための最も身近な手段である**情報検索** [1] では，通常，データベース中のオブジェクトを特徴付ける属性空間での近接性により，クエリとの類似性や関連性を判断するが，実際にはこうした主要な属性に加え，付加的な属性も利用できることが少なくない．例えば，映画や音楽といった対象には，(主に) 商用・営利目的でのカタログ的なラベルが付与されることが多く，さらに，インターネットを介したメディア配信サービスや，それらを専門に扱う SNS の普及により，ユーザが自由に付与したタグ情報も現在では入手可能となっている．こうした**ラベル情報**は，検索や推薦を行なう上で極めて有用なものであり，それら情報を取り込むことで，より柔軟な検索・推薦処理が実現できるだろう．本稿ではその一提案として，所与のクエリに対する，**ラベル情報の意外性**を考慮した検索・推薦手法を考察する．具体的には，データベース中のオブジェクトに付与された多重ラベルの類似性を反映した属性部分空間での近接性に基づき，(ラベルが未知の) クエリのラベルを (暗に) 推定し，元の空間ではクエリと近接，すなわち，関連性がある一方で，部分空間では離れている，すなわち，異なるラベルを有するという意味での意外性を有するオブジェクトの検出を試みる．

そこでは，ラベルに関する意外性 (遠さ) を測る物差しの設計が重要になる．このために本稿では，多重ラベルを持つ対象-属性データ行列から，対象のラベルを学習・推定する手法を与えた Sun 等の研究 [3] の考えた方を採用する．すなわち，対象のラベル情報とその近接性を最もよく近似できる属性部分空間を使うことである．ただし，[3] では，そうした空間を k-次元部分空間の直交基底を固有値問題を解くことにより求めるが，本稿では 1-次元，すなわち，数直線への射影を求めるに留め，単に，クエリーを含めて対象がラベルに関して遠いか否かを判定できれば十分であるとの立場に立つ．幸いなことに，1-次元の場合は，グラフラプラシアンの正の最小固有値の近似解を二部グラフの**交差数最小化アルゴリズム** [4] により高速 (データサイズを n とした $O(n \log n)$) に求めた上で判定できるので，本稿でもこれを採用する．

元の属性空間での近接性とラベル情報に関する部分空間での近接性という二種類の距離が定まれば，意外性を有するオブジェクトはもちろん容易に同定可能であるが，本稿

[1] 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University
[a)] mh@ist.hokudai.ac.jp

Vol.2017-MPS-113 No.30
Vol.2017-BIO-50 No.30
2017/6/24

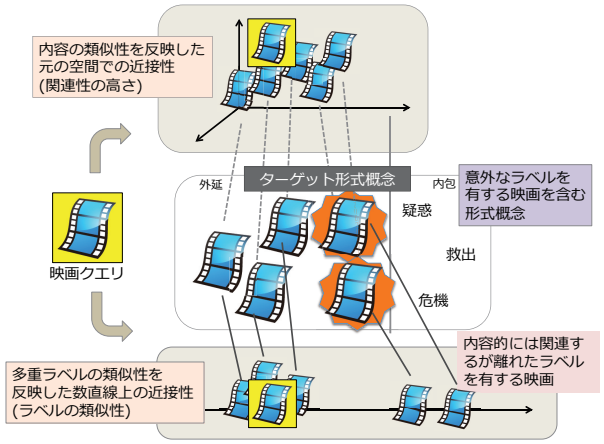

図 1 映画とその筋書きに関するデータにおける，クエリと関連性が高いが意外なラベルを有する映画を含む形式概念のイメージ

では，個々のオブジェクトを抽出対象とするのではなく，それらを構成要素の一部として含む**形式概念** [2] を抽出の対象とする．形式概念とは，オブジェクト集合 (外延) とそれらが共有する属性集合 (内包) の組であり，オブジェクト間に見出すことのできる**局所的な類似性**を陽に捉えるものである．一般に，オブジェクト間には異なる観点・視点からの類似性を見出すことが可能であり，特に推薦を目的とする検索においては，様々な観点からの類似性を扱えることは強力な武器となる．形式概念はそれを可能とする数学的な枠組みであることから，ここでは，意外なオブジェクトを含む形式概念を抽出対象とする．より具体的には，外延中の任意のオブジェクトは元の空間においてクエリと近接 (類似/関連) するが，その一部は，数直線上では離れた意外なラベルを有する形式概念を抽出ターゲットとし，それを二種類の距離制約付き形式概念探索問題として定式化する (図 1)．所与のクエリに対し，こうした形式概念を抽出することで，ユーザが新たな興味を発見するきっかけを提供したい．

2. 準備

対象 (オブジェクト) の集合 $\mathcal{O}$ と，**属性**の集合 $\mathcal{F}$ について，$R \subseteq \mathcal{O} \times \mathcal{F}$ なる二項関係 $\mathcal{R}$ を考える．ここで，$(o, f) \in R$ は，対象 o が 属性 f を有することを意味する．この時，$\mathcal{C} =< \mathcal{O}, \mathcal{F}, R >$ を**形式文脈**と呼ぶ．

対象 $o \in \mathcal{O}$ について，o が有する属性の集合を $F(o)$ と表記する．すなわち，$F(o) = \{f \in \mathcal{F} \mid (o, f) \in R\}$ である．ここで，対象集合 $O \subseteq \mathcal{O}$ と属性集合 $F \subseteq \mathcal{F}$ について，写像 $\varphi : 2^{\mathcal{O}} \to 2^{\mathcal{F}}$ と $\psi : 2^{\mathcal{F}} \to 2^{\mathcal{O}}$ を，$\varphi(O) = \bigcap_{o \in O} \mathcal{F}(o)$，および，$\psi(F) = \{o \in \mathcal{O} \mid F \subseteq \mathcal{F}(o)\}$ と定義する．φ は O 中のすべての対象が共有する属性集合を，ψ は属性集合 F を有する対象集合を同定する写像である．

これら写像をもとに，$\varphi(O) = F$ かつ $\psi(F) = O$ なる対象集合 O と属性集合 F の組 $C = (O, F)$ を，形式文脈 $\mathcal{C}$ における**形式概念** (*FC:Formal Concept*) と定める [2]．ここで，O と F それぞれを，形式概念 C の**外延**，および，**内包**と呼ぶ．なお，写像 φ と φ が定義される形式文脈 $\mathcal{C}$ を明確にする場合は，$\varphi_{\mathcal{C}}$，$\psi_{\mathcal{C}}$ と表記する．

互いに素な頂点集合 V_0 と V_1 を部集合とする二部グラフを，$G = (V_0, V_1, E)$ と表記する．ここで，$E \subseteq V_0 \times V_1$ である．

行列 A の転置行列を A^T と表記する．

3. 多重ラベル分類

M の属性に関する (実) ベクトルで表現される N のデータ $\boldsymbol{x}_i$ $(1 \leq i \leq N)$ から成る $M \times N$ データ行列を $X = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_N)$ とする．いま，Ls を (クラス) ラベルの集合とした時，各データ $\boldsymbol{x}_1$ には，Ls 中のいくつかのラベル (多重ラベル) が付与されているものとし，それを $L(\boldsymbol{x}_i)$ $(\subseteq Ls)$ で参照する．

ラベルが未知のデータ $\boldsymbol{q} \in \mathbb{R}^M$ について，X 中のデータとの類似性をもとに，$\boldsymbol{q}$ に付与されるであろう多重ラベルを推定するタスクを，**多重ラベル分類問題**と呼ぶ．

3.1 低次元空間への埋め込みに基づくラベル推定

Sun 等は文献 [3] において，X 中の各データを頂点，各ラベル $\ell_i \in Ls$ が付与されたデータ集合 (頂点集合) をハイパーエッジとするハイパーグラフ G により，データ群とそれら多重ラベルの関係を表現し，G のスター展開によって得られる二部グラフの正規化ラプラシアン $\mathcal{L}$ をスペクトル分解することで，多重ラベル間の類似性を反映した属性部分空間を (近似的に) 同定し，そこでのデータ間の近接性をもとに，ラベルが未知のデータのラベル推定を行なう手法を提案した．その詳細は文献 [3] に譲るが，簡単に述べると，正規化ラプラシアン $\mathcal{L}$ の固有ベクトルを正の固有値が小さな方から k $(< M)$ 並べた $N \times k$ 行列 H_M の第 i-行ベクトル $\tilde{\boldsymbol{x}}_i$ は，X 中のデータ $\boldsymbol{x}_i$ の，多重ラベルの類似性が反映された k-次元部分空間への埋め込み (embedding) を与える [*1]．よって，

$$H_M^T \simeq W^T X \qquad (1)$$

なる $M \times k$ 行列 $W = (\boldsymbol{w}_1 \cdots \boldsymbol{w}_k)$ がわかれば，ラベルが未知のデータ $\boldsymbol{q} \in \mathbb{R}^M$ の k-次元部分空間での (列ベクトル) 表現が $W^T\boldsymbol{q}$ で与えられることから，H_M^T の列ベクトル $\tilde{\boldsymbol{x}}_i^T$ と $W^T\boldsymbol{q}$ の類似性に基いて，$\boldsymbol{q}$ のラベル推定を行なうことができる．

ここで，$H_M = (\boldsymbol{h}_1 \cdots \boldsymbol{h}_k)$ とすると，式 (1) は

$$(\boldsymbol{h}_1 \cdots \boldsymbol{h}_k) \simeq (X^T\boldsymbol{w}_1 \cdots X^T\boldsymbol{w}_k)$$

と書けるから，こうした W は，$\boldsymbol{h}_i \simeq X^T\boldsymbol{w}_i$ の最小二乗解

[*1] 正確には，各頂点の次数による補正操作等も必要となる．

$\boldsymbol{w}_i$ を計算することで得られる．

データ行列 X の k-次元部分空間への埋め込みである行列 H_M は，正規化ラプラシアン $\mathcal{L}$ の固有ベクトルにより与えられるが，行列の固有値計算は一般に負荷が大きく，昨今の大規模データを扱う際には何らかの工夫が必要となる．

そこで本稿では，二部グラフの**交差数最小化**により，ラプラシアンの第二固有ベクトル (Fiedler ベクトル) [*2] の良い近似が得られることに注目する [4]．文献 [4] では，データサイズを n として，交差数最小化問題の近似解を $O(n \log n)$ で高速に計算可能なアルゴリズムが与えられており，これを用いることで，先の議論における k-次元部分空間として，$k = 1$，すなわち，数直線を考え，そこでのデータ間の近接性に基づいて多重ラベル間の類似性を考えるものとする．

3.2 二部グラフの交差数最小化

二部グラフ $G = (V_0, V_1, E)$ は，各部集合 V_i の頂点を，それぞれ並行する直線 L_i 上に並べ，隣接する頂点間を直線で結ぶことで図示できる．ここで，各頂点 $v \in V_0 \cup V_1$ に対して，対応する直線上での座標を割り当てる関数を $h : V_0 \cup V_1 \to \mathbb{R}$ とすると，h は G のひとつの**描画** (drawing) を与える．

G のある描画 h において，辺 $e \in E$ と交差する辺の総数を $cr_h(e)$ で参照すると，描画 h における辺の交差数は $cr(h) = \frac{1}{2}\sum_{e \in E} cr_h(e)$ で与えられる．G の**交差数最小化問題**とは，$h^* = \arg\min_h \ cr(h)$ なる描画 h^* を求める問題である．

一方，描画関数 h により定義される G の隣接頂点間の各直線上での座標の差分総和 $L_h = \sum_{u \in V_0, v \in V_1} |h(u) - h(v)|$ が最小となる頂点の並び順 (序数) を決める問題は，**線形配置問題**と呼ばれる．

文献 [4] では，ラプラシアンの二次形式最小化問題として定式化される (正規化) カット最小化の意味での**二部グラフの最適分割問題**と線形配置問題が，ともにラプラシアンの第二固有ベクトル (Fiedler Vector) を求めることで (近似的に) 解けること，および，交差数最小化問題の解が線形配置問題のよい近似解を与えることが示された．このことから，交差数最小化に基づく二部グラフの分割 (クラスタリング) 手法が提案され，その有効性が理論的・実験的に示されている [4]．

本稿では，そこで示された交差数最小化問題の近似解を高速に求めるアルゴリズムを用いてラプラシアンの第二固有ベクトルを近似的に求め，それをデータ行列 X の直線上 (1-次元部分空間) への埋め込みと考える．

[*2] 正の最小固有値に属する固有ベクトル．

4. 意外性を有する検索・推薦のための距離制約付き形式概念探索

本節では，ラベルが未知のクエリオブジェクト q に対して，意外性のある対象を検索・推薦するタスクを，ふたつの距離制約付きの形式概念探索問題として定式化する．具体的には，データベース中のオブジェクトに付与された多重ラベルの類似性を反映した空間での距離と，それらオブジェクトが属する元の空間での距離のもとで，元の空間では q と類似度が高い (近い) という意味で関連性を有するが，q の推定ラベルとは異なるラベルを有するという意味で意外性を有するオブジェクトを外延の一部として含む形式概念を高速に抽出することで，意外性を考慮した検索・推薦を実現する．

いま，M-次元ベクトル $\boldsymbol{x}_i \in \mathbb{R}$ で表現される N のオブジェクト群から成るデータベースを考え，これを行列 $X = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_N)$ で表す．ここで，各オブジェクト $\boldsymbol{x}_i$ には，Ls 中のいくつかのラベル (多重ラベル) が付与されているものとし，それを $L(\boldsymbol{x}_i)$ $(\subseteq Ls)$ で参照する．また，クエリオブジェクトを $\boldsymbol{q} \in \mathbb{R}$ とし，特にそのラベルは未知であると仮定する．

4.1 クエリに関する意外性

本稿で提案する，クエリ $\boldsymbol{q}$ に関する意外性は，$\boldsymbol{q}$ の推定ラベルと，各オブジェクトに付与されたラベルとの遠さ (非類似性) をひとつの根拠とするものである．ここではまず，Sun 等の多重ラベル推定法 [3] での考え方を利用して，ラベルの遠さを認識する方法を与える．

4.1.1 ラベルの遠さ

先に述べた通り，Sun 等の手法では，X 中の各オブジェクト $\boldsymbol{x}_i$ と，それに付与された多重ラベル $L(\boldsymbol{x}_i)$ の関係を二部グラフ G_{label} として表現し，その正規化ラプラシアンのスペクトル分解により，多重ラベルの類似性が反映された k-次元部分空間への $\boldsymbol{x}_i$ の埋め込みを与えた．ここで，G_{label} は，各オブジェクト $\boldsymbol{x}_i$ に対応する頂点を v_i とする頂点集合 $V_X = \{v_1, \ldots, v_N\}$ と，ラベル集合 Ls をそれぞれ部集合とし，辺集合 $E \subseteq V_X \times Ls$ は次の通り定義される：

$$E = \{(v_i, \ell) \mid v_i \in V_X, \ell \in L(\boldsymbol{x}_i)\}.$$

すなわち，$G_{label} = (V_X, Ls, E)$ は，各オブジェクトとそれに付与されたそれぞれのラベルを隣接関係によって表現した二部グラフである．こうした二部グラフのラプラシアンの固有ベクトルを正の固有値が小さな方から k $(< M)$ 並べた行列の第 i 行が，X 中のオブジェクト $\boldsymbol{x}_i$ の，多重ラベル間の類似性が反映された k-次元部分空間への埋め込みとなる．

$G_{label} = (V_X, Ls, E)$ のラプラシアンの固有値計算はコストの高い処理であり，特に大規模データに対しては避け

ることが望ましい．そこで本稿では，文献 [4] で提案された二部グラフの交差数最小化アルゴリズムにより，G_{label} のラプラシアンの第二固有ベクトルを近似した N-次元ベクトル $\tilde{\boldsymbol{h}} = (h_1, \ldots, h_N)^T$ を求め，$\tilde{\boldsymbol{h}}$ を，多重ラベル間の類似性が反映された 1-次元部分空間，すなわち，数直線 (L_{label} とする) 上への X の埋め込みと考える．

より正確に述べると，G_{label} の交差数最小化の結果得られる部集合 V_X 上の序数を，全単射 $\pi : V_X \to \{1, \ldots, N\}$ で表すと，$\tilde{\boldsymbol{h}}$ は，

$$\tilde{\boldsymbol{h}} = (\pi(v_1), \ldots, \pi(v_N))^T$$

で与えられ，その各成分 $\pi(v_i)$ は，データ行列 X 中のオブジェクト $\boldsymbol{x}_i$ の L_{label} 上の座標を表す．

よって，式 (1) より，$\tilde{\boldsymbol{h}}^T \simeq \boldsymbol{w}^T X$ なるベクトル $\boldsymbol{w}$ がわかれば，クエリ $\boldsymbol{q}$ の L_{label} 上の座標は，$\boldsymbol{w}^T\boldsymbol{q}$ で与えられる．こうした $\boldsymbol{w}$ は，$\tilde{\boldsymbol{h}} \simeq X^T\boldsymbol{w}$ の最小二乗解を求めることで同定できる．

L_{label} はデータベース中のオブジェクトに付与された多重ラベル間の類似性を反映した数直線であるから，クエリ $\boldsymbol{q}$ のラベルは，L_{label} 上での近傍オブジェクトのそれと類似すると推定される．逆に，L_{label} 上で $\boldsymbol{q}$ と離れて位置するオブジェクトは，$\boldsymbol{q}$ とは類似していないラベルを有すると推測できるだろう．ここでは，数直線 L_{label} 上での距離に基づいて，ラベル情報の遠さを認識するものとする．

具体的には，近傍パラメータ δ_L (> 0) を導入し，L_{label} 上で δ_L より離れて位置するオブジェクトのラベルは，$\boldsymbol{q}$ のそれとは離れている (非類似である) と考える．すなわち，$\boldsymbol{q}$ とオブジェクト $\boldsymbol{x}_i$ について，$|\boldsymbol{w}^T\boldsymbol{q} - \pi(v_i)| > \delta_L$ が成り立つ時，$\boldsymbol{q}$ と $\boldsymbol{x}_i$ のラベルは離れているとする．

4.1.2 ラベルの意外性

上述した通り，オブジェクトが有するラベルの非類似性は，数直線 L_{label} 上での離れ具合により認識可能であるが，非類似性は必ずしも意外性を含意するものではない．クエリ $\boldsymbol{q}$ との関連性が低いオブジェクトのラベルが，$\boldsymbol{q}$ のそれと類似しないことはある意味自然であり，そこに意外性を見出すことは困難である．様々な意外性の定義が考えられるが，ここでは，データ行列 X が表現された元の M-次元空間での近接性が示唆するオブジェクト間の関連性をもとに，関連性の高いオブジェクト同士は多くの場合，類似するラベルを有するであろうとの仮説のもと，$\boldsymbol{q}$ との関連性が高いオブジェクトが，$\boldsymbol{q}$ の (推定) ラベルとは離れたラベルを有する時，それは意外なものであると考え，検索結果として積極的に取り込むことを提案する．

具体的には，元の M-次元空間における近傍パラメータ δ_M (> 0) を導入し，適当な距離関数 $dist_M : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ のもとで，オブジェクト $\boldsymbol{x}_i$ が

関連性制約：$\boldsymbol{q}$ と $\boldsymbol{x}_i$ の関連性は高い

$dist_M(\boldsymbol{q}, \boldsymbol{x}_i) \leq \delta_M$，および，

ラベルの遠さ制約：$\boldsymbol{q}$ と $\boldsymbol{x}_i$ のラベルは離れている

$|\boldsymbol{w}^T\boldsymbol{q} - \pi(v_i)| > \delta_L$

を満たす時，$\boldsymbol{x}_i$ **はクエリ $\boldsymbol{q}$ に関して意外なオブジェクトである**と定める．

4.2 距離制約付き形式概念探索による意外性を考慮した検索

ここでは，所与のクエリに関する意外性を考慮した検索タスクを，距離制約付き形式概念探索問題として定式化する．

$M \times N$ データ行列 X のもとで形式概念を抽出するにあたり，まず X に対応した形式文脈 $\mathcal{C}_X$ を与える．

X 中の各オブジェクト $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iM})^T \in \mathbb{R}^M$ を o_i で参照し，その第 j-成分 x_{ij} が o_i の属性 f_j の値であると考えると，X におけるオブジェクト集合と属性集合は，それぞれ $\mathcal{O}_X = \{o_1, \ldots, o_N\}$，および，$\mathcal{F}_X = \{f_1, \ldots, f_M\}$ となる．また，形式文脈では，各オブジェクトはそれが有する属性集合により特徴付けられることから，X において観測されるオブジェクトと属性の関係 $\mathcal{R}_X \subseteq \mathcal{O}_X \times \mathcal{F}_X$ を次の通り定める [*3]：

$$\mathcal{R}_X = \{(o_i, f_j) \mid \boldsymbol{x}_i = (x_{i1}, \ldots, x_{iN})^T \text{ in } X, x_{ij} \neq 0\}.$$

よって，X に対応する形式文脈は $\mathcal{C}_X =< \mathcal{O}_X, \mathcal{F}_X, \mathcal{R}_X >$ となる．

データ行列 X とその対応する形式文脈 $\mathcal{C}_X$ のもとで，クエリ $\boldsymbol{q}$ に関する意外性を考慮した検索を実現するために，ここでは次の条件を満たす形式概念 FC の抽出を試みる：

> FC の外延は，$\boldsymbol{q}$ との関連性が高いオブジェクトから成り，かつ，その中には $\boldsymbol{q}$ に関して意外なものが含まれる．

形式的にはこうしたタスクは，先に議論した，X に関する多重ラベル情報を表現した二部グラフの交差数最小化により得られるオブジェクト集合上の序数を与える写像 π，および，数直線 L_{label} 上への X の埋め込み $\tilde{\boldsymbol{h}} = (\pi(v_1), \ldots, \pi(v_N))^T$ とそれを与えるベクトル $\boldsymbol{w}$ を用いて，距離制約付きの形式概念探索問題として次の通り定式化できる．

定義：距離制約付き形式概念探索によるクエリに関する意外性を考慮した検索

$M \times N$ データ行列とその対応する形式文脈をそれぞれ X，および，$\mathcal{C}_X =< \mathcal{O}_X, \mathcal{F}_X, \mathcal{R}_X >$ とする．いま，クエリオブジェクト $\boldsymbol{q}$ に関する意外性を考慮した検索を，次の条件を満たす $\mathcal{C}_X$ の形式概念 $FC = (O, F)$ を列挙するタスクと定める．

[*3] ここでは，X における成分値 0 が，対応する属性を持たないこと，あるいは，無視可能なことを意味するものと想定しており，例えば，単語の $tfidf$ 値を成分とする文書データ行列等がこれに当てはまる．成分値 0 に何らかの意味がある場合は，もちろんこの定式化は不適切である．

関連性制約： 任意の $o_i \in O$ について，X 中の対応するベクトル $\boldsymbol{x}_i$ は $dist_M(\boldsymbol{q}, \boldsymbol{x}_i) \leq \delta_M$ を満たす．

ラベルの遠さ制約： ある $o_i \in O$ が存在し，$|\boldsymbol{w}^T\boldsymbol{q} - \pi(o_i)| > \delta_L$ である．

ここで，δ_M と δ_L は，それぞれ M-次元空間，および，数直線 L_{label} における非負実数の近傍パラメータである．■

4.3 意外性を有するオブジェクトを含む形式概念探索アルゴリズム

形式文脈を $\mathcal{C}_X =< \mathcal{O}_X, \mathcal{F}_X, \mathcal{R}_X >$ とする．ここでは，上述したふたつの距離制約を満たす $\mathcal{C}_X$ の形式概念を列挙する深さ優先探索アルゴリズムを与える．

4.3.1 形式概念探索の基本戦略

いま，$\mathcal{O}_X = \{o_1, \ldots, o_N\}$ 上の全順序 $\prec$ を仮定し，$\mathcal{O}_X$ の任意の部分集合 P の要素は，この順序に従って整列しているものとする．この時，全順序 $\prec$ のもとで，$\mathcal{O}_X$ のべき集合 $2^{\mathcal{O}_X}$ は，包含関係に基づいて，空集合を根とする集合列挙木で表現できる．ここで，集合 P の最終要素を $tail(P)$ で参照すると，親子関係にある P と P' において，$P = P' \setminus \{tail(P')\}$ が成り立つ．

定義より，$\mathcal{C}_X$ における各形式概念 (O, F) について，その外延と内包がそれぞれ $O = \psi(\varphi(P))$ および $F = \varphi(P)$ となるオブジェクト集合 $P \subseteq \mathcal{O}_X$ が存在する．よって，集合列挙木の各ノード P を深さ優先で辿りながら，$(\psi(\varphi(P)), \varphi(P))$ を計算してやれば，すべての形式概念を漏れなく抽出することができる．

より具体的には，$\emptyset$ で初期化したオブジェクト集合 P を起点として，$tail(P)$ の後続オブジェクトを追加することで P を P' へと拡張した後，概念 $(\psi(\varphi(P')), \varphi(P'))$ を計算する処理を，深さ優先で繰り返せばよい．

4.3.2 意外性を有するオブジェクトを含む形式概念探索

上述した探索基本戦略に従い，ここでは，意外性を有するオブジェクトを含む形式概念の抽出手続きを与える．

本稿で抽出ターゲットとする形式概念の外延に含まれる任意のオブジェクトは，元の空間においてクエリ $\boldsymbol{q}$ と関連 (類似) することを要請される．よって，ここでは，$\boldsymbol{q}$ と関連するオブジェクトのみを考慮した形式文脈 $\tilde{\mathcal{C}}_X$ における形式概念探索を行なうものとする．

より形式的に述べると，形式文脈 $\mathcal{C}_X =< \mathcal{O}_X, \mathcal{F}_X, \mathcal{R}_X >$ において，元の空間でクエリ $\boldsymbol{q}$ と関連するオブジェクト集合 $\tilde{\mathcal{O}}_X$ は，

$$\tilde{\mathcal{O}}_X = \{o_i \mid o_i \in \mathcal{O}_X, dist_M(\boldsymbol{q}, \boldsymbol{x}_i) \leq \delta_M\}$$

で与えられ，これらオブジェクトのみに限定した関係 $\tilde{\mathcal{R}}_X$ は

$$\tilde{\mathcal{R}}_X = \mathcal{R}_X \cap (\tilde{\mathcal{O}}_X \times \mathcal{F}_X)$$

と表すことができる．これらを用いて，不要なオブジェクトを除去した形式文脈を $\tilde{\mathcal{C}}_X =< \tilde{\mathcal{O}}_X, \mathcal{F}_X, \tilde{\mathcal{R}}_X >$ と定める．

抽出ターゲットとなる形式概念に課せられたもうひとつの重要な制約は，その外延が，数直線 L_{label} 上で $\boldsymbol{q}$ と離れたオブジェクトを少なくともひとつ含むことである．探索過程でこうした形式概念のみが抽出されることが保証できれば，効率の良い処理が可能となろう．ここでは，集合列挙木を定めるオブジェクト集合上の全順序 $\prec$ を工夫することでそれを実現する．

集合列挙木の定義より，オブジェクト集合 P の拡張処理で追加されるオブジェクト o は，$tail(P) \prec o$ を満たすものである．よって，オブジェクト集合 $\tilde{\mathcal{O}}_X$ を，L_{label} 上で $\boldsymbol{q}$ の近傍に位置するものから成る $\tilde{\mathcal{O}}_X^+$ と，そうでないもの $\tilde{\mathcal{O}}_X^-$ に分割し，$\tilde{\mathcal{O}}_X^-$ を前方，$\tilde{\mathcal{O}}_X^+$ を後方とする $\tilde{\mathcal{O}}_X$ 上の全順序 $\prec$ を考えれば，$\tilde{\mathcal{O}}_X$ の任意の部分集合 P において，$\tilde{\mathcal{O}}_X^+$ 中のオブジェクトが $\tilde{\mathcal{O}}_X^-$ 中のそれに先行して出現することはない．このことから，$\tilde{\mathcal{O}}_X^-$ 中のオブジェクトを要素とする単集合 P に対して，深さ優先の拡張処理を行えば，その過程で得られる形式概念は，$\boldsymbol{q}$ の近傍にはないオブジェクトを少なくともひとつ含むものとなる．また，こうした形式概念を得るには，単集合 $P \subseteq \tilde{\mathcal{O}}_X^-$ を起点とする拡張処理だけを考えれば十分である．

以上の処理をまとめた擬似コードを図 2 に示す．図中，先の議論で与えた $\delta_M, \delta_L, dist_M, \boldsymbol{w}$, および, π を既知のものとして用いている．また，集合 S の先頭要素を $head(S)$ で参照する．なお，本稿ではここで触れるに留めるが，計算効率に悪影響を及ぼす形式概念の重複枚挙を回避するための処理も組み込んでいる (手続き FCFind 冒頭の **if** 文))．その詳細については，例えば文献 [5] を参照されたい．

5. おわりに

本稿では，意外性を考慮した検索・推薦を実現するための，形式概念探索手法を議論した．所与のクエリに対し，ラベル情報に関する意外性を有するオブジェクトを含む形式概念を抽出することで，ユーザが新たな興味を発見するきっかけを提供できるものと期待している．

現在，システムの実装作業を進めている段階であり，実験結果については口頭発表時に報告したい．実験に用いるデータとしては，種々ジャンルを表す (多重) ラベルが付与された映画の筋書き (文書) データを考えている．この場合，クエリとして与えた映画と筋書きが類似しているが，異なるジャンルに属するとの意味で意外な映画が検索・推薦される．なお，実験結果を分析する際には次の点について検討を行い，必要に応じて手法の改良・洗練化を進めたい．

- ここでは，意外性を考慮した検索タスクを，ふたつの距離制約を満たす形式概念の列挙問題として定式化したが，他にも様々な制約の代替案が考えられることは

[Input] $\mathcal{C}_X = (\mathcal{O}_X, \mathcal{F}_X, \mathcal{R}_X)$: a formal context obtained from the data matrix $X = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_N)$
$\boldsymbol{q}$: a query vector
[Output] $\mathcal{FC}$: the set of formal concepts with interesting objects for $\boldsymbol{q}$

procedure MAIN($\mathcal{C}_X$, $\boldsymbol{q}$) :
  $\mathcal{FC} \leftarrow \emptyset$;
  $\tilde{\mathcal{O}}_X \leftarrow \{o_i \mid o_i \in \mathcal{O}_X, dist_M(\boldsymbol{q}, \boldsymbol{x}_i) \leq \delta_M\}$;
  $\tilde{\mathcal{R}}_X \leftarrow \mathcal{R}_X \cap (\tilde{\mathcal{O}}_X \times \mathcal{F}_X)$;
  $\tilde{\mathcal{C}}_X \leftarrow < \tilde{\mathcal{O}}_X, \mathcal{F}_X, \tilde{\mathcal{R}}_X >$; // reduced formal context
  $\tilde{\mathcal{O}}_X^- \leftarrow \{o_i \mid o_i \in \tilde{\mathcal{O}}_X, |\boldsymbol{w}^T\boldsymbol{q} - \pi(v_i)| > \delta_L\}$;
  $\tilde{\mathcal{O}}_X^+ \leftarrow (\tilde{\mathcal{O}}_X \setminus \tilde{\mathcal{O}}_X^-)$;
  Fix a total order $\prec$ on $\tilde{\mathcal{O}}_X$ such that
    for any $o^- \in \tilde{\mathcal{O}}_X^-$ and $o^+ \in \tilde{\mathcal{O}}_X^+$, $o^- \prec o^+$;
  while $\tilde{\mathcal{O}}_X^- \neq \emptyset$ **do**
    begin
      $o \leftarrow head(\tilde{\mathcal{O}}_X^-)$;
      $\tilde{\mathcal{O}}_X^- \leftarrow (\tilde{\mathcal{O}}_X^- \setminus \{o\})$; // removing o from $\tilde{\mathcal{O}}_X^-$;
      $C \leftarrow (\tilde{\mathcal{O}}_X^- \cup \tilde{\mathcal{O}}_X^+)$; // candidate obj.
      FCFIND($\{o\}$, $\emptyset$, C) ;
    end
  return $\mathcal{FC}$;

procedure FCFIND(P, $PrevExt$, $Cand$) :
  $FC \leftarrow (Ext = \psi_{\tilde{\mathcal{C}}_X}(\varphi_{\tilde{\mathcal{C}}_X}(P)), \varphi_{\tilde{\mathcal{C}}_X}(P))$; // formal concept
  if $\exists o \in (Ext \setminus PrevExt)$ such that $o \prec tail(P)$ **then**
    return; // detected duplicate formal concept
  endif
  $\mathcal{FC} \leftarrow \mathcal{FC} \cup \{FC\}$;
  while $Cand \neq \emptyset$ **do**
    begin
      $o \leftarrow head(Cand)$;
      $Cand \leftarrow (Cand \setminus \{o\})$; // removing o from $Cand$;
      $NewCand \leftarrow (Cand \setminus PrevExt)$; // new candidate obj.
      if $NewCand = \emptyset$ **then continue** ;
      FCFIND($P \cup \{o\}$, Ext, $NewCand$) ;
    end

図 2 意外性を有するオブジェクトを含む形式概念探索アルゴリズム

言うまでもない．例えば，現在は，外延中のすべてのオブジェクトが元の空間でクエリと関連することを要請するが，形式概念であることによって局所的な類似性は保証されることから，少なくともひとつのオブジェクトがクエリと関連すればよいとの立場も考えられる．また，意外性の根拠となるラベルの遠さを，パラメータ δ_L を陽に導入して数直線 L_{label} 上で判断しているが，パラメータ設定の煩わしさは否定できない．それを避けるためには，例えば，L_{label} 上での，外延中のオブジェクトとクエリとの最短距離が最大となる形式概念を抽出対象とする最適化問題としての定式化が考えられる．こうした最適化，あるいは，Top-N 最適化アプローチは，列挙アプローチでは出力概念数が膨大となる場合にも有望であり，十分検討に値する．

- 提案した枠組みでは，ラベル情報に関する数直線 L_{label} 上へのデータの埋め込みを，オブジェクト集合上の序数 (順列) として与えたが，Sun 等の手法 [3] に倣い，ラベルの重要度を反映した辺重みを二部グラフ G_{label} に導入することで，より一般的な実数値座標への埋め込みを得ることができる．そこでのオブジェクト間の近接関係が検索結果に及ぼす影響を観察することは興味深い．
- 高次元データを対象とする場合には事前の次元圧縮処理が必要となるが，その方法や度合いは得られる検索・推薦結果を大きく左右する．こうした影響の観察・考察を通して，適切な次元圧縮の方針を得ることも重要である．

参考文献

[1] Salton, G. and McGill, M. J.: Introduction to Modern Information Retrieval, 440 pages, McGraw-Hill, 1983.
[2] Ganter, B. and Wille, R.: Formal Concept Analysis - Mathematical Foundations, 284 pages, Springer, 1999.
[3] Sun, L., Ji, S. and Ye, J.: Hypergraph Spectral Learning for Multi-Label Classification, Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'08, pp. 668 – 676, 2008.
[4] Ahmad, W. and Khokhar, A.: cHawk: An Efficient Biclustering Algorithm Based on Bipartite Graph Crossing Minimization, Proc. of the 2007 VLDB Workshop on Data Mining in Bioinformatics, 2007.
[5] Haraguchi, M. and Okubo, Y.: An Extended Branch-and-Bound Search Algorithm for Finding Top-N Formal Concepts of Documents, New Frontiers in Artificial Intelligence, JSAI 2006 Conference and Workshops, Tokyo, Japan, June 5-9, 2006, Revised Selected Papers, LNCS-4384, pp. 276 – 288, 2007.