6F-06

イベント指向データ管理手法を用いた系図表示
– 複数の系譜情報のデータ構造 –¶

平塚聡§, 杉山正治*, 横澤大典*, 生田敦司*, 柴田みゆき*, 松浦亨**
四條畷学園短期大学§, 大谷大学*, 北海道大学病院**

1. はじめに

一般に系図作成時には，多くの婚姻を結んだような家系を整理するために，目的に応じて細分化された系図を作成する場合がある．一方，細分化された系図を統合して大きな一つの系図を作成することもある．このときに，複数の系図を再利用し統合されたひとつの系図を生成することが可能であれば，作成労力を軽減することができる．統合対象の複数の系図には同一人物を示す個性が存在することが多いが，そのような個性を自動的に同定することができれば，それを起点として系図を統合することが容易となる．その為には，個性や系図上の関係の類似度を自動的に算定することができれば助けとなる．しかし，人の目視や手作業によってしても，個性の同定は容易な作業ではない．

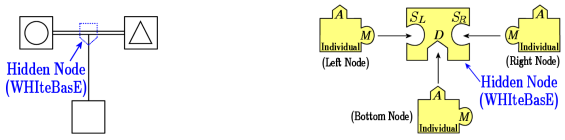
この種の問題については，不特定多数のユーザが家系図情報をそれぞれ登録することにより系図を作成するサービスが既に存在する [3]．そこでは新たに登録しようとする一件の個性と，データベース上の既にある個性との重複登録を避けるための検索が必要である．この操作は，全レコードに対して一回の走査で済む．一方，複数の系図を統合する際に求められるのは系図同士の照合であるため，効率の良い検索手法が必要である．

我々はこれまでに，新しい系図データ管理手法である WHIteBasE (Widespread Hand to InTErconnect BASic Elements) を用いた系図表示手法を提案し，プロトタイプソフトウェアを開発してきた [1]．WHIteBasE モデルでは，人物を示す個性（individual）と，個性同士の婚姻・親子関係を示す不可視結節点 WHIteBasE, および個性と WHIteBasE を結ぶひとつまたは複数の線分によって系図が表示される．

本研究では，WHIteBasE を用いた複数の系図データが与えられた場合に，それらにまたがる同一個性を同定するためのデータセットを生成するアルゴリズムを提案し，その有効性を検討する．

2. WHIteBasE

WHIteBasE の概要を示す．1 つの親子関係は 1 つのイベントとして不可視結節点 WHIteBasE (図 1(a)) を用いて管理される．WHIteBasE による結合モデルを



(a) 系図表示スタイル　(b) WHIteBasE モデル

図 1: 婚姻関係と子の発生を表す基本結合

図 1(b) に示す．WHIteBasE は婚姻を示す線分と親子を示す線分の交点に置く．

WHIteBasE は集合 W_i を，個性は集合 I_j を用いて

$$W_i = \{S_L, S_R, D_j, \mathbf{Q}\} \quad I_j = \{A, M_k\} \quad \begin{cases} i = 0, 1, \cdots, i_{max} \\ j = 0, 1, \cdots, j_{max} \\ k = 0, 1, \cdots, k_{max} \end{cases} \quad (1)$$

で表される．ここで i, j, k は各要素の ID を，$i_{max}, j_{max}, k_{max}$ は各要素の最大値を，S_L, S_R は左右の個性 (両親) の ID を，D_j は下位世代 (子) の ID を，A は上位世代 (親) の WHIteBasE の ID を，M_k は婚姻相手の WHIteBasE の ID を，それぞれ表す．I_j は個性名称や付帯情報が格納されたデータテーブルで管理される．W_i は個性情報とは異なるデータテーブルで管理される．再婚は複数の WHIteBasE により管理される．$\mathbf{Q}$ は WHIteBasE が管理する座標値の集合である．

一般的な系図表示ソフトウェアでは個性同士がリンクされているが，WHIteBasE モデルでは全ての個性を WHIteBasE が管理し，個性同士は直接リンクしない．

3. 系図を照合する際の問題

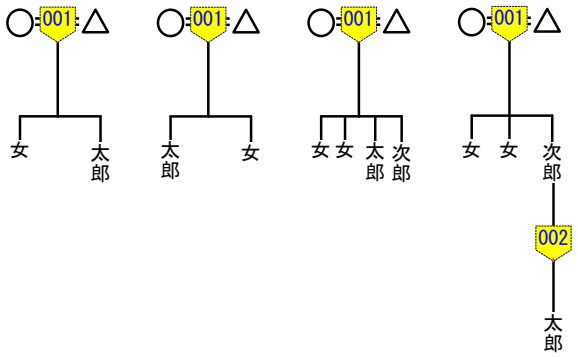
2 つの系図データを統合する場合，双方で同一人物とされる個性を同定する必要がある．個性の同一性は個性の名称，付帯情報，および他のノードとの関係によって識別することができる．

ある系図のある名称を持つ個性は，別の系図の同じ名称の個性と同一人物とみなすべき蓋然性が大きいと考えられる．ただし，同姓同名の人物が付帯情報のみで区別されて存在することもありうるため，他のノードとの関係の評価を含めて総合的に同定を行う必要がある．

ノード同士の関係を見る場合，ある個性を中心としてそこからの関係を評価するよりも，婚姻関係と親子関係を一単位として，その単位ごとに照合する方が効率的である．WHIteBasE のデータ構造はこのような処理に適している．

付帯情報は自由記述に用いられる領域であり，今回は識別要素に含めない．

¶Data Structure of Genealogy with Several Possible Origins
§Satoshi Hiratsuka: Shijonawate Gakuen Junior College
*Seiji Sugiyama, Atsushi Ikuta, Daisuke Yokozawa, and Miyuki Shibata: Otani University
**Tohru Matsuura: Hokkaido University Hospital



(a) 系図 1　(b) 系図 1'　(c) 系図 2　(d) 系図 3

図 2: WHIteBasE を用いた系図表示の例

4. WHIteBasE 一致度検出手法

比較対象の系図 A と B に含まれる WHIteBasE のベクトル $\boldsymbol{b}^A$ と $\boldsymbol{b}^B$ を,

$$\boldsymbol{b}^A = (b_1^A, b_2^A, \cdots, b_m^A)^T, \quad \boldsymbol{b}^B = (b_1^B, b_2^B \cdots b_n^B)^T \quad (2)$$

で表すとき，同一の婚姻･親子関係を表す一致度を $b_i^A b_j^B$ とすると，一致度行列$\boldsymbol{C}$ は,

$$\boldsymbol{C} = \begin{pmatrix} b_1^A b_1^B & b_1^A b_2^B & \cdots & b_1^A b_n^B \\ b_2^A b_1^B & b_2^A b_2^B & \cdots & b_2^A b_n^B \\ \vdots & \vdots & \ddots & \vdots \\ b_m^A b_1^B & b_m^A b_2^B & \cdots & b_m^A b_n^B \end{pmatrix} \quad (3)$$

$$b_i^A b_j^B = w_\alpha (w_s c_s + w_d c_d) + w_\beta c_f \quad (4)$$

$$c_s = \frac{U_s}{U_s + V_s}, \quad c_d = \frac{U_d}{U_d + V_d}, \quad c_f = 1 - \frac{|x-y|}{x+y} \quad (5)$$

で表現できる．ここで，m, n は系図 A, B の WHIteBasE の個数を表す．一致度は，WHIteBasE に連なる両親と子に，個性の名称が一致するものがどの割合で存在するかにより算出する．

ただし，U は双方の WHIteBasE で名前が現れる個性の数，V は片方の WHIteBasE にのみ名前が現れる個性の数，下添字 s は WHIteBasE に連なる親，d は子を表す．x, y はそれぞれの WHIteBasE に連なる全ての子の数を表す．w は重み係数で，w_α と w_β は名前の一致による得点と子の人数の一致による得点の比重を調整し，w_s と w_d は名前一致による得点のうち親と子の比重を調整する．

5. 一致度検出手法の適用例

図 2 を用いて，(a) の WHIteBasE と，(b)–(d) の各 WHIteBasE の一致度算出の例を示す．図中の「女」(むすめ) は，日本における歴史的な系図において，名前が不明な女性を示す際によく用いられる記法である．なお，婚姻を示す線分と親子を示す線分の交点にある五角形の記号は WHIteBasE を表す．

(a) と (b) にそれぞれ現れる WHIteBasE001 の一致度を求める場合，まずどちらの両親も「△」と「○」なので，U_s は親の総数の 4，V_s は 0 であるため，親に関する c_s は最大の 1 となる．また，系図 1 側に子として現われる「太郎」「女」は系図 1' 側にも現れ，また系図 1' 側の個性もすべて系図 1 側に現れるので，U_d は子の総数の 4，V_d は 0 となるので，子の名前に関する c_d も最大の 1 である．更に，WHIteBasE が管理する子の数に関する c_f も最大値の 1 となる．重み係数の値によらずすべての得点が最大となるため，この 2 つの WHIteBasE は同一として扱うことができる．

(a) と (c) の場合，親の構成は一致するため，親に関する c_s は最大の 1 だが，(c) の「次郎」が (a) に現れず，その他の名前の個性は全て双方に現れるため，U_d は 5，V_d は 1 となり，子の名前に関する c_d は 5/6 となる．また，子の数に関する c_f は，$1 - |2-4|/(2+4)$ すなわち 2/3 となる．子の名前に関しては (c) の 2 件の「女」は双方とも (a) に存在するとしてカウントされるが，2 件の「女」の存在によって子の数が増えることから，子の数に関する得点が減点されている．したがって「女」などの同一名称の子の数が異なる場合は最高得点は得られない．

(d) は WHIteBasE001 と WHIteBasE002 があるが，まず (a) の WHIteBasE001 と (d) の WHIteBasE001 を比較する．親に関する c_s は最大の 1 だが，子の名前に関する c_d は 3/5，子の数に関する c_f は 4/5 となる．対して，WHIteBasE002 を比較対象とする場合は，親に関しては一致する個性が皆無なので c_s は 0，子の名前に関する c_d は 2/3，子の数に関する c_f は 2/3 となる．

仮に全ての重み係数を 0.5 とした場合の各一致度行列は，(a) と (b) の場合は $\boldsymbol{C}$=(1)，(a) と (c) の場合は $\boldsymbol{C}$=$(\frac{19}{24})$，(a) と (d) の場合は $\boldsymbol{C}$=$(\frac{4}{5} \quad \frac{1}{2})$ となる．

6. おわりに

本研究では複数の系図で同一の個性を同定するための照合アルゴリズムを提案した．系譜・系図情報を扱う上では，史料や情報の異説を照合する処理が必要となる [2] が，本手法は高速な処理を行うことに資する．

重み係数の具体的な値は，ユーザが直接設定するほか，実際にある系図資料の例から学習により生成することを予定している．

参考文献

[1] S. Sugiyama, A. Ikuta, D. Yokozawa, M. Shibata, T. Matsuura: “Displaying Genealogy with A-doptions and Multiple Remarriages Using the WHIteBasE”, Lecture Notes in Computer Science (LNCS) 8104, pp. 325-336, 2013

[2] 横澤, 生田, 杉山, 平塚, 柴田, 松浦 :『イベント指向データ管理手法を用いた系図表示 -複数の系譜情報の併記手法-』第 78 回情報処理学会全国大会講演論文集, 4F-02, pp. 4-517〜4-518, 2016

[3] MyHeritage, https://www.myheritage.jp/