数理モデル化と問題解決　31－9
（２０００．９．２１）

二種類の順序木より成る有向グラフの交差数減少法について

北上 始　西本 美津子
広島市立大学・情報科学部

二種類の異種木構造データベースにおける調停作業を容易にするためには，両データベースから検索された二つの部分木から構成される有向グラフの交差数を最小にすることが重要である．ここでは，データベースの質問処理において，その有向グラフの交差を最小にするための制約ソルバーの処理方式について提案する．最小交差の有向グラフは，そのグラフを構成する各部分木がゼロ交差で順序付けされる中で，部分木の葉節点列間を結ぶ辺同士が最小交差の制約を充足することで定義される．最小交差の有向グラフは，各順序木の各階層レベルにおける非葉節点順序の変更と各葉節点列における葉節点順序を変更することで達成される．ここでは，葉節点列間に結合行列を定義し，あるヒューリスティクスを用いて結合行列を対角化する方法を提案している．

A Methodology for Reducing Crossovers of A Directed Graph Constructed from Two Ordered Trees

HAJIME KITAKAMI, MITSUKO NISHIMOTO

It is very important to find two ordered trees with the same sequence of leaf nodes in order to achieve an effective reconciliation. For the reconciliation, two ordered trees that satisfy the zero-crossover constraint are useful for comparing the two heterogeneous trees. This paper proposes a new method for searching for two ordered trees that satisfy the zero-crossover constraint. This is achieved using a heuristic tree search for an interconnection matrix, which is defined by the leaf sequences (layers) of the two trees.

１．はじめに

本稿では，Goodman により分子生物学の分野で提案された調停概念(Goodman, '79)を発展させ，お互いに異なる二つの木構造データベースに対して，見通しのいい調停作業および理解しやすい調停木の表示などを支援する機能に着目する．その支援のためには，交差数の少ない二つの順序木を探索する機能が重要である．即ち，異種の木構造データベースから検索された両部分木をある有向グラフと見做し，その有向グラフから最小交差の制約を満足する有向グラフを見つけ出す機能に着目する．コンピュータ支援による調停作業は，異なる分子進化系統樹間の分析(Page, '97)，分子進化系統樹を用いた生物分類樹の見直しを初めとして，異種概念を持つ複数の人間に対する共同作業支援，画像の木構造に着目した内容検索技術などとも深い関わりがある．従って，異種木構造データベースから調停作業に必要な最小交差制約を満足する有向グラフを見つけ出す方法の研究は，応用範囲が広く大変意義深い研究であると考える．さらに，別の観点からみると，我々が扱う問題は制約プログラミングから発展してきた制約データベース(Kanellakis, '95)とも関係が深い．即ち，データベースへの問合せに最小交差制約の条件が存在する場合，その条件は制約データベースにおける有用な制約の1つとみなすことができる．以上により，お互いに異なる二つの木構造データベースを用いて最小交差制約を満足する有向グラフを見つける問題は，ある特定の限られた分野に存在する問題ではなく，応用範囲の広い制約データベース処理の問題でもある．

本論文では，ある葉節点集合を検索キーとして二つの異なる木構造データベースから検索された二つの部分木を考え，両部分木から構成される有向グラフに対して，最小交差制約を満足する有向グラフを見つけ出す方法について提案する．

２．システム構成

図1に，著者らが研究開発しているシステム(Kitakami et al., '00)の構成を示す．二つの異種木構造データベースは，どちらも有向グラフの辺（二項関係）の集合を格納しており，各辺の二次記憶上での格納順序により，節点間の順序が定義されるものとする．図1において，利用者が，解

析に必要な木構造の葉節点集合を与えると，サーチエンジンは次章で述べる上方探索により，その葉節点集合に対する部分木の頂点が二つの木構造データベースから検索される．頂点からの幅優先探索により，各部分木の辺は順序付けされるが，その順序により交差のない順序木を定義すると，二つの順序木間で葉節点列同士が一致しない．図2に，この状況の例が示されている．

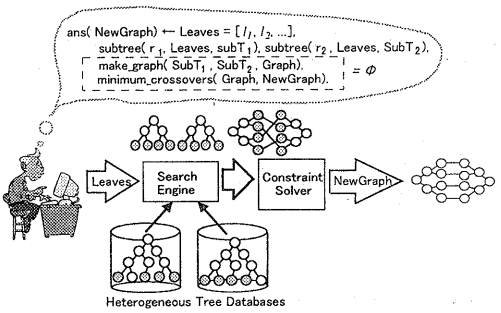

図１．システム構成図

図1の制約ソルバーでは，サーチエンジンにより検索された両順序木から構成されるグラフを初期値とし，節点の順序を変化させながら最小交差をもつ二つの順序木を見つける．即ち，制約ソルバーでは，図2の初期グラフから出発し，木の枝同士の交差がゼロでかつ葉節点列間の交差が最小になるような二つの順序木が見つけ出される．見つけ出された二つの順序木は，適当な画面表示ソフトウェアにより，画面上に表示される．

3．データモデル

本章では，二つの異種系統樹データベースおよびそれらから検索される二つの部分木はいずれも同じデータモデルで表現されるとする．また，異種木構造データベースを順序木とみなしたときの順序は，データの格納順によって決められているとする．別の見方をすると，この順序により，検索時にデータが取り出される順番が決定される．さらに，本章では，二つの部分木から構成される有向グラフはゼロ交差にできると仮定する．ゼロ交差にできない問題については，本章のデータモデルを応用することで対処する．以下，順序木の構造を定義した後，交差のない順序木を解くためのゼロ交差制約充足問題を定義し，その制約充足問題を解くために重要な基本操作について述べる．

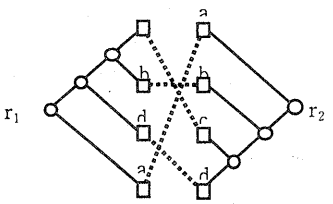

図２．初期グラフの例

3.1 順序木の構造

本質的な部分を浮き彫りにするために，高さn—1の木は正規化されているものとする．すなわち，ある葉の深さがn—1よりも小さいとき，深さがn—1になるように途中にダミーの節点が追加されているものとする．そのような木を定義するために，あるnレベル階層（$n \geqq 2$）の有向グラフ$T = g(V,R,n)$を考える．この有向グラフが下記の条件を満足するとき，その有向グラフTを順序木と呼ぶ．ただし，Vはある節点の集合であり，RはV上の2項関係$V \times V$の部分集合として定義される辺の集合である．

【条件1】Vは次のようにn個の部分集合に分割される．

$V = N_1 \cup N_2 \cup \cdots \cup N_n$ （$N_i \cap N_j = \phi$，$i \neq j, 1 \leqq i \leqq n, 1 \leqq j \leqq n$）．

以下，N_iをレベルiの節点集合と呼び，nを木の高さと呼ぶ．

【条件2】N_1の要素数$|N_1|$は1である．以下，この節点要素をTの根と呼ぶ．

【条件3】Rは，次のように$n-1$の部分集合に分割される．

$R = B_1 \cup B_2 \cup \cdots \cup B_{n-1}$ ，（$B_i \cap B_j = \phi$，$i \neq j$），

$B_i \subset N_i \times N_{i-1}$，$1 \leqq i \leqq n-1$．

このとき，同じ終点を有する任意の2つの辺(d_1, e)，$(d_2, e) \in R$に対して，$d_1 = d_2$を満足する．以下では辺(d, e)において，始点 d を親節点と呼び，終点 e を子節点と呼ぶことにする．

【条件4】入次数がゼロなる節点の集合はN_1だけである．また，出次数がゼロなる節点集合はN_nだけであり，N_nの各要素を葉節点と呼ぶことにする．

【条件5】レベルiの節点集合N_iに存在する全ての節点に対して，ある順序列が与えられている．すなわち，節点を$d^{(i)}_p \in N_i$で表現すると，順序列は$\sigma_i = d^{(i)}_1\, d^{(i)}_2 \cdots d^{(i)}_\alpha$で表現される．ただし，$1 \leqq p \leqq \alpha$，$\alpha$は$N_i$の節点数を意味する．以後，この順序列の順序関係を2項関係 $d^{(i)}_1 <_B d^{(i)}_2$，$d^{(i)}_2 <_B d^{(i)}_3$，$\cdots$，$d^{(i)}_{p-1} <_B d^{(i)}_p$で表現し，$n$レベル階層グラフを$g(V,R,n,\Sigma)$で表現する（$\Sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$）．

3.2 ゼロ交差制約

葉節点数が同じ二つの順序木$T_1 = g(V_1, R_1, n_1, \Sigma_1)$，$T_2 = g(V_2, R_2, n_2, \Sigma_2)$において，一方の木の葉節点集合から他方の木の葉節点集合への全単射（1対1）が与えられているものとする．ここでは，木T_jの枝同士が交差しな

い条件をT_jのゼロ交差制約(CS_1)と呼ぶ．また，二つの木T_1,T_2の葉節点レベル間を対応させることにより定義される辺集合を考えたとき，その中の辺同士が交差しない条件をT_1,T_2間のゼロ交差制約(CS_2)と呼ぶ．以下，これらの二種類の制約を総称して，T_1とT_2により構成される有向グラフのゼロ交差制約と呼ぶ．

3.3 結合行列

図2の葉節点列間の結合関係を表わす結合行列を図3に示す．図の左側順序木及び右側順序木に定められている葉の順序は，各々，OL_1=[c, b, d, a]及びOL_2=[a, b, c, d]で表現されている．左側順序木の葉cと右側順序木の葉節点列[a, b, c, d]との間の結合関係は，結合行列 Mat の1行目の行ベクトル(0　0　1　0)の中に示されている．左側順序木のcは右側順序木のa, b, dと接続関係をもたないので，行ベクトルの第一，第二，第四の要素は0で表現されているが，第三要素は両部分木間の葉c同士で接続関係をもつので1で表現されている．結合行列の要素を m_{ij} で表現すると，葉節点列間の交差数 C(Mat)は次のように定義される．

$$C(Mat) = \Sigma_{[1\leq j<k\leq n]} \Sigma_{[1\leq \alpha<\beta\leq n]} m_{j,\beta} m_{k,\alpha} \qquad (1)$$

3.4 基本操作

ゼロ交差制約の充足には，二つの基本操作がある．第一は，木構造データベースから検索された部分木に対して，ゼロ交差の順序木を見つけるための幅優先探索であり，第二は，葉節点列において，ある葉節点を含む部分木を考え，与えられた二つの葉節点を区別するのに有用な二つの最大部分木の探索である．以下では，その最大部分木が有する葉節点集合をクラスタと呼ぶ．

次に，図4を用いて，二つの葉節点 $d^{(n)}_p$, $d^{(n)}_q \in N_n$ を区別する二つのクラスタを探索する方法について述べる．葉節点の集合N_nを有する順序木$T=g(V,R,n,\Sigma)$は，ゼロ交差制約(CS_1)を満足しているとしよう．また，探索される二つのクラスタは，二つの葉節点 $d^{(n)}_p$, $d^{(n)}_q \in N_n$を区別する最大クラスタC_1,C_2であるとする．ただし，$d^{(n)}_p \in C_1$, $d^{(n)}_q \in C_2$であり，$C_1 \cap C_2 = \phi$である．葉節点 $d^{(n)}_p$のクラスタとは，$d^{(n)}_p$を含む部分木の葉節点集合を指す．以上の性質を有する二つの最大クラスタ　C_1,C_2は，以下の手順により探索することができる．

上方探索により，二つの葉節点 $d^{(n)}_p$, $d^{(n)}_q \in N_n$ の分岐点 $d^{(i)}$を見つけ，それに連結される複数の部分木から葉節点 $d^{(n)}_p$を含む部分木 $SubT_1$及び葉節点 $d^{(n)}_q$を含む部分木 $SubT_2$ を探索する．ただし，$i<n$であり，$SubT_1$ と $SubT_2$ の節点には共通節点がないとする．また，$SubT_1$ と $SubT_2$ は，各々，考えられる部分木の中で最大の部分木であるとする．

下方探索により，二つの部分木 $SubT_1$ と $SubT_2$ の各々に対する葉節点の集合C_1,C_2を計算する．ただし，$d^{(n)}_p \in C_1$, $d^{(n)}_q \in C_2$, $C_1 \cap C_2 = \phi$を満足する．

4．制約ソルバー

本章では，図1の制約ソルバーにおいて，二つの部分木により構成される有向グラフの交差数を最小にする方法を提案する．二つの順序木より構成される有向グラフを最小交差の状態にするには，各順序木に対する制約(CS_1)を満足させる中で，葉節点列同士の接続において制約(CS_2)を最大の近似度で満足させることが重要である．最大の近似度とは，辺の交差数が最小の状態を意味する．もし，この状況で制約(CS_2)を完全に満足させることができれば，最小交差数はゼロを意味する．制約(CS_2)を最大の近似度で充足させる前に，(CS_1)を満足するような順序木の1つ(初期値)を取り敢えず求めておくことは，前章の幅優先探索により達成可能である．しかし，この段階では，図2の点線で示されたように，双方の部分木の葉節点間を結ぶ辺同士に交差が生じてしまう．

この状態から辺の交差を最小にするために，葉節点の順序列において葉節点の順序を交換しなければならない．しかし，旨く交換しなければ，木の枝の間に交差ができてしまう．ここでは，両木構造の葉節点間にだけ結合行列を考え，ゼロ交差制約(CS_1)に違反しないような行列変換(葉節点間の順序交換)を行うことにより，ゼロ交差制約

$$Mat = \begin{array}{c} \\ c \\ b \\ d \\ a \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{array}{l} OL_1 = [c, b, d, a] \\ \\ OL_2 = [a, b, c, d] \end{array}$$

図3．結合行列の例

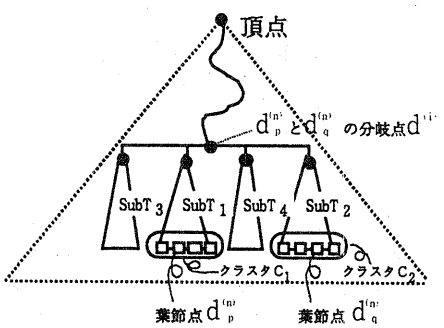

図4．　葉節点に関するクラスタの探索

(CS_2)に対する準最適解を得る方法を提案する．ゼロ交差制約(CS_1)に違反しないような行列変換は，前章で探索されるクラスタ同士の交換により保証される．この行列変換により単位行列にできるだけ近づけることにより，ゼロ交差制約(CS_2)を近似的に充足させることが可能である．紙面の都合上，アルゴリズムの大枠だけを図5に示しておく．

次に，時間計算量Ωについて考察する．ただし，簡単のため両順序木の構造は同一とし，各木の階層数をnとする．i 番目の階層レベルに存在する節点数をM_iで表わすと，各木の葉節点数はM_nである．したがって，二つの木を葉節列間で接続した有向グラフの階層数は$2n+1$である．結合行列を対角化する際にバックトラックがないとすれば，計算量は次のとおりである．

$$O(h_1 \times M_n^2) \qquad (1)$$

また，可能なバックトラックを全て実施したとすれば，計算量は次のとおりである．

$$O(h_2 \times M_n^2 \times M_n\,!) \qquad (2)$$

ここで，$h_1=h_2=n-1$である．これを式(1)～(2)に代入すると，次の時間計算量が得られる．

$$\Omega = O(n \times M_n^2) \sim O(n \times M_n^2 \times M_n\,!) \qquad (3)$$

ここでは，結合行列上で1の値を対角線上の左上から右下へと順に乗せていくというヒューリスティクスを使っている．したがって，クラスタ交換が無効になることが原因で生ずるバックトラックの回数が多くなければ，計算量は$O(n \times M_n^2$程度と考えられる．

従来の方式にしたがい，全ての階層間に結合行列を

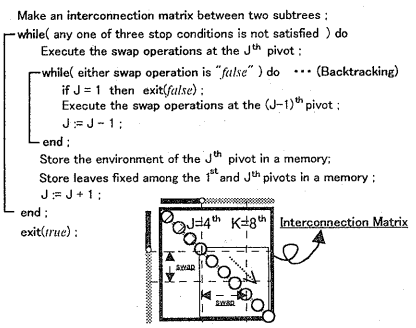

図5．アルゴリズムの大枠

作成する場合について考えてみよう．$i-1$レベル階層とiレベル階層との間の結合行列に着目すると，解の候補数は，$M_i\,! \times {}_lC_m$個あり，1つの解を計算するための計算量は，$O(M_i \times M_{i-1})$である($l=M_i=M_{i-1}$, ${}_lC_m$は二項係数)．これにより，上記のh_1,h_2に相当する部分は，以下のようになる．

$$h_1 = O(\pi_{2 \le i \le n} M_{i-1} \times M_i) > O(n) \qquad (4)$$

$$h_2 = O(\pi_{2 \le i \le n} M_{i-1} \times M_i \times M_{i-1}\,! \times {}_lC_m) > O(n) \qquad (5)$$

ただし，πはiに関する乗積を表わす．式(4)～(5)を式(1)～(2)に代入すると，次式が得られる．

$$\Omega = O((\pi_{[2 \le i \le n]} M_{i-1} \times M_i) \times M_n^2) \sim$$

$$O((\pi_{[2 \le i \le n]} M_{i-1} \times M_i \times M_{i-1}\,! \times {}_lC_m) \times M_n^2 \times M_n\,!) \qquad (6)$$

以上により，従来の方式では，各結合行列の計算量の積をとった大きな計算量をもつことになり，本提案方式はそのような悪影響を回避することができる．

5．おわりに

本論文では，二種類の異種木構造データベースから検索された二つの部分木に対して，最小交差制約を満たす二つの順序木の計算方法を提案した．提案方式では，両系統樹の葉の階層間に唯一の結合行列を定義し，クラスタ交換により木の枝に交差が生じないような行列変換を行った．これにより，単位行列を効果的に見つけることができた．計算量の評価や実装による測定により，本提案方式の有効性を確認することができた．

今後は，遺伝的プログラミングや並列処理について検討を進めていく予定である．そこでは，致死遺伝子の発生をできるだけ防ぐ方法，両親の遺伝的形質が不必要に失われないようにする方法，アニーリング手法等の導入により局所探索能力を追加する方法などの検討が重要であると思われる．

参 考 文 献

1) Goodman, M., Czelusniak, J., Romero-Herrera, A. E., and Matsuda, G.: Fitting the Gene Lineage into its Species Lineage: A parsimony strategu illustrated by Cladograms Constructed from Globin Sequences, Systematic Zoology, Vol. 28, pp.132-168 (1979).

2) Page, R.D.M., and Charleston, M.A.: From Gene to Organismal Phylogeny: Reconciled Trees and the Gene/Species Tree Problem", Molecular Phylogenetics and Evolution, Vol.7, pp231-240 (1997).

3) Kanellakis P. Constraint Programming and Database Languages: A Tutorial, Proc. of ACM POS'95, pp.46-53, 1995.

4) Hajime Kitakami and Mitsuko Nishimoto: A Constraint Solver for Reconciling Heterogeneous Trees, IC-AI'2000, CSREA Press, Vol. III, June 26-29, 2000, pp.1419-1425.