

文書構造を利用した自動図表配置手法における バックトラック処理の解消

獅々堀正幹 青江順一
徳島大学

1. まえがき

近年、ワードプロセッサ等の普及により、画面上で最終割り付けの確認をしながら文書を作成することが可能になった⁽¹⁾。しかし、図表に関しては、ユーザー自身が配置しなければならず、また、その配置位置が適切か否かの評価も難しい。従って、あらかじめ与えられた配置基準に準拠した位置に図表を自動配置する手法は、文書処理の重要な課題である。

本稿では、参照箇所の位置、配置位置の安定性、文書全体における配置の均等性、文章と図表の関連性等に関係する図表配置基準を定義し、文書構造⁽²⁾と図表データを用いてこれらの基準を満たす図表配置位置決定アルゴリズムを提案する。

本アルゴリズムは、参照箇所の後方近くに配置位置を定める後方処理；後方処理で決定された配置位置をより読み易い位置にするために、関連のある文章領域内に配置位置を変更する関連処理；後方処理のみでは配置位置が偏ってしまうので、文書全体的に均等に図表を配置する均等処理；配置すべきでない場所を避けて図表を配置する制約処理より成る。均等、制約処理では既配置の文章や図表をページを越えて変更する（バックトラックと呼ぶ）が必要が生じるので、処理時間が長くなるが、本手法では、このバックトラックを解消するために文章と図表領域の残量を考慮した高速なアルゴリズムを提案する。

2. 図表配置の基準とデータの定義

2. 1 基準の定義

より適切な位置に図表を配置するための基準を、情報処理学会と電子情報通信学会の論文を対象に調査し、図表配置に関する次の基準を決定する。

順序基準：番号順に配置する。

安定基準：ページの上部または下部に配置する。

後方基準：参照箇所の後方で、最も近い配置可能な領域に配置する。

関連基準：関連のある文章内に配置する。

均等基準：一定の距離内（本手法では前後1ページ内とする）に図表を配置する。

制約基準：始章、終章等（以下、特殊見出しと略する）の領域内には配置しない。

2. 2 データ要素の定義

文書は段組、図表は1フレーム⁽³⁾幅に整合する図表（片幅図表）と2フレーム幅の図表（両幅図表）を対象とし、以下のデータ要素を定義する。

(1) p 番目の段落に対するデータ要素

段落行数を $P_LINE(p)$ ；見出し $HEAD$ 、特殊見出し S_HEAD 、本文 $TEXT$ の何れかの論理構造によって示される段落の属性を $P_ATTRI(p)$ ；段落内で参照されている図表番号 d とその参照位置までの行数 h の組 (d, h) を要素とする集合を $P_REF(p)$ で表す。尚、段落の属性 P_ATTRI が $TEXT$ または $ITEM$ の場合は、未処理の段落と文章の割合によって後方処理または均等処理； S_HEAD 場合は関連処理と制約処理； $HEAD$ の場合は関連処理が行われる。

(2) d 番目の図表に対するデータ要素

図表行数を $D_LINE(d)$ ；片幅図表は $SINGLE$ 、両幅図表は $DOUBLE$ なる値をもつ図表の幅属性を $D_WIDTH(d)$ ；フレーム番号とフレーム内の上下位置を示す値（上部は TOP 、下部は $BOTTOM$ ）の組を値とする図表配置位置を $D_POS(d)$ で表す。

(3) f 番目のフレームに対するデータ要素

フレームが属するページ番号を $F_PAGE(f)$ ；フレーム内に存在する段落行数を $F_PARA(f)$ ；フレーム内の利用可能な空き行数を $F_AVAIL(f)$ ；フレーム内に含むことが可能な最大行数を $F_MAX(f)$ で表す。

図1の例では、 $P_LINE(p)=30$ 、 $P_ATTRI(p)=TEXT$ 、 $P_REF(p)={{(d, 15)}, {(d+1, 25)}}$ 、 $D_LINE(d)=10$ 、 $D_WIDTH(d)=DOUBLE$ 、 $D_POS(d)=(1, BOTTOM)$ 、 $D_LINE(d+1)=15$ 、 $D_WIDTH(d+1)=SINGLE$ 、 $D_POS(d+1)=(2, TOP)$ 、 $F_PAGE(1)=1$ 、 $F_PARA(1)=30$ 、 $F_AVAIL(1)=10$ 、 $F_MAX(1)=50$ となる。尚、図中の $r_p i$ は i 番目の図表の参照位置を示す。

また、処理中の段落 p と図表 d を配置するフレーム番号をそれぞれ $PARA_F(p)$ 、 $DIA_F(d)$ で表し、図表 d の参照箇所が存在するフレームを $REF_F(d)$ とする。

A Method for the Solution of Backtrack Process on Allocating Diagrams Automatically.

Masami Shishibori, Jun-ichi Aoe

The University of Tokushima

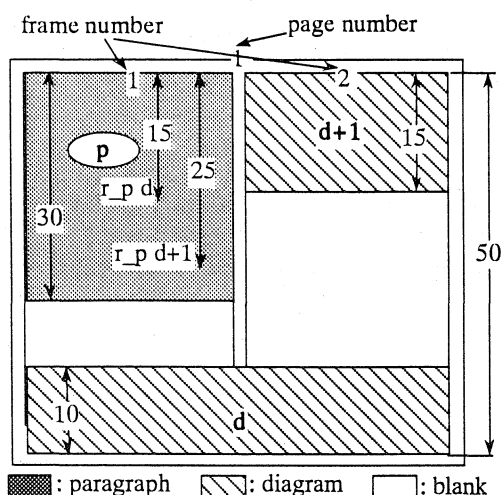


図1 データ要素の例

Fig. 1 Examples of data elements.

3. 図表配置決定処理

3.1 後方処理

3.1.1 片幅図表配置処理

片幅図表の処理は、REF_F(d)とDIA_F(d-1)の距離によって、図2に示すような2種類に分けられる。まず、図2(a)はREF_F(d)がDIA_F(d-1)以前にある場合、DIA_F(d-1)以前のフレームにD_LINE(d)以上の空き領域があったとしても、順序基準を満たすため、DIA_F(d-1)以前のフレームには配置できない。即ち、図表dはDIA_F(d-1)またはDIA_F(d-1)+1のフレームに配置される。この場合の配置手順を以下に示す。

Step1-1: $F_AVAIL(DIA_F(d-1)) \geq D_LINE(d)$ が成立すればStep 1-2に、不成立ならばStep 1-3に進む。

Step1-2: DIA_F(d)をDIA_F(d-1)とし、dをDIA_F(d)の上部に配置する。

Step1-3: DIA_F(d)をDIA_F(d-1)+1とし、dをDIA_F(d)の上部に配置する。

図2(b)のようにREF_F(d)がDIA_F(d-1)以後にある場合、REF_F(d)以前のフレームには空き領域がないので、図表dはREF_F(d)またはREF_F(d)+1のフレームに配置される。この配置手順を以下に示す。

Step3-1: $F_AVAIL(REF_F(d)) \geq D_LINE(d)$ が成立すればStep 3-2に、不成立ならばStep 3-3に進む。

Step3-2: DIA_F(d)をREF_F(d)とし、dをDIA_F(d)の下部に配置する。

Step3-3: DIA_F(d)をREF_F(d)+1とし、dをDIA_F(d)の上部に配置する。

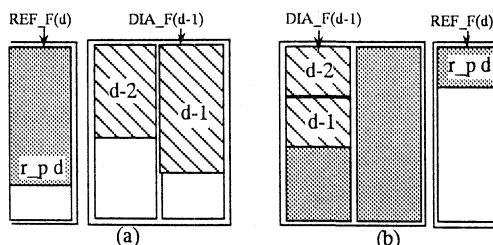


図2 片幅図表配置の説明図

Fig. 2 Illustration of the single column diagram allocation.

3.1.2 両幅図表配置処理

両幅図表処理では、ページ内に両幅図表分 ($D_LINE(d)*2$) の未使用行数が残っていたとしても、両フレームにそれぞれD_LINE(d)の未使用行数がなければ配置できない。従って、ページ内に既配置の図表や文章の再配置処理が必要となるが、2つの制約が生じる。まず、既配置の図表を次ページに移動すれば、配置対象の図表が既配置の図表より前に配置されることになり、順序基準に反する。また、文章を次ページに移動すれば、配置対象の図表は対応する参照箇所より以前に配置され、後方基準に反する。即ち、後方への再配置はページを越えてはいけないという一つの制約が生じる。次に、本手法は基準を満たしながら逐次的に配置決定するので、既配置の図表や文章を前方に再配置すれば、今まで満たしていた基準が破棄される。即ち、前方への再配置はできないという制約も生じる。これらの制約から、ページの最終フレームに含まれる図表や文章は再配置できないことになり、段組み文書では、再配置可能な図表や文章はページの先頭フレームのものに限られる。以上より、両幅図表処理は図3のように2種類に分類される。

ページの最終フレームに図表や文章が含まれる場合(図3(a))、それらの図表や文章は再配置できないため、まず、最終フレームの未使用行数を調べる。以下に、この場合の手順を示す。

Step4-1: $F_AVAIL(DIA_F(d-1)) \geq D_LINE(d)$ ならばStep 4-2、不成立ならばStep 4-5に進む。

Step4-2: $F_AVAIL(DIA_F(d-1)) + F_AVAIL(DIA_F(d-1)-1) \geq D_LINE(d)*2$ ならばStep4-3、不成立ならばStep4-5に進む。

Step4-3: DIA_F(d-1)内の図表や文章を移動し、両フレームにD_LINE(d)分の空き領域を確保出来ればStep4-4、出来なければStep4-5に進む。

Step4-4: DIA_F(d)をDIA_F(d-1)とした後、F_PAGE(

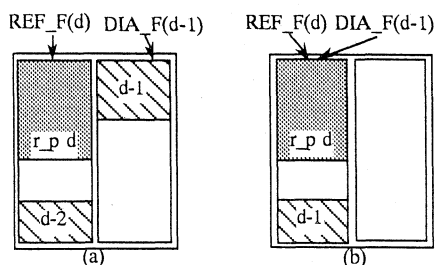


図3 両幅図表配置の説明図

Fig. 3 Illustration of the double column diagram allocation.

$DIA_F(d) = F_PAGE(REF_F(d))$ ならば、 $D_POS(d)$ を $(DIA_F(d), BOTTOM)$ とし、 $F_PAGE(DIA_F(d)) \neq F_PAGE(REF_F(d))$ ならば、 $D_POS(d)$ を $(DIA_F(d), TOP)$ とする。

Step4-5: $DIA_F(d)$ を $DIA_F(d-1)+1$ とし、 $D_POS(d)$ を $(DIA_F(d), TOP)$ とする。

次に、ページの最終フレームが空フレームの場合(図3(b))、先頭フレームの未使用行数を調べ、もし、その行数が $D_LINE(d)$ 以上ならば、図表dをこのページに配置する。以下に、この手順を示す。

Step5-1: $F_AVAIL(DIA_F(d-1)) \geq D_LINE(d)$ が成立すればStep5-4、不成立ならばStep5-2に進む。

Step5-2: $F_AVAIL(DIA_F(d-1)) + F_AVAIL(DIA_F(d-1) + 1) \geq D_LINE(d) * 2$ が成立すればStep5-3、不成立ならばStep5-5に進む。

Step5-3: $DIA_F(d-1)$ 内の図表や文章を移動し、両フレームに $D_LINE(d)$ 分の空き領域を確保出来ればStep5-4、出来なければStep5-5に進む。

Step5-4: Step4-4と同様。

Step5-5: $DIA_F(d)$ を $DIA_F(d-1)+2$ とし、 $D_POS(d)$ を $(DIA_F(d), TOP)$ とする。

尚、図表・文章の移動処理は順序基準を守るために、移動すべき図表や文章領域(移動領域と呼ぶ)を前方向に拡張し、後方基準を守るために、参照箇所手前までの文章→参照箇所に対応する図表→参照箇所の順番で移動領域を拡張する。

3.2 均等処理

多くの図表が集中した箇所では参照された場合(図4(a))、図表が小さければ後方処理で対応できる(図4(b))が、大きい場合は図表が文書の後方に追いやられて後方フレームに図表だけが固り、図表と参照箇所との距離が大きくなる(図4(c))。均等処理では、これを次のように解決する。

段落及び図表の全行数をそれぞれ ALL_P_LINE , ALL_D_LINE ; 段落p及び図表dが配置されたときの

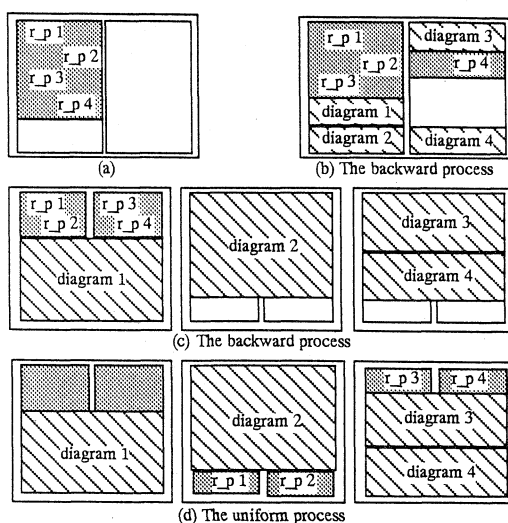


図4 均等処理の例

Fig. 4 Examples of the uniform process.

未処理行数を $REM_P_LINE(p)$, $REM_D_LINE(d)$; 未処理の段落と図表の割合を $RATIO$ とする。初期状態として、 REM_P_LINE は ALL_P_LINE , REM_D_LINE は ALL_D_LINE とすると、図表pと段落dが配置される毎に、次が計算できる。

$REM_P_LINE(p) := REM_P_LINE(p-1) - P_LINE(p);$

$REM_D_LINE(d) := REM_D_LINE(d-1) - D_LINE(d);$

$RATIO := REM_D_LINE(d) / REM_P_LINE(p);$

$RATIO$ が閾値を越えた場合、配置処理を後方処理から均等処理へ切り替える。また、後方処理では参照箇所の後方に図表を配置していたが、均等処理では参照箇所が出現したフレームに図表を配置する。本手法により、参照箇所以前の図表配置が生じるが、各図表は均等に配置され、各図表と参照箇所の距離は近できる(図4(d))。また、提案手法は配置済みの図表をバックトラックしない高速手法である。

3.3 その他の処理

関連処理は、話題の区切りに当たる見出し(区切り見出し)が出現した場合、前章節内で参照された下配置済みの図表を上部に配置し直す。尚、章節から章、項から章節への変化に対しては話の流れが変化するので、区切り見出しは、章見出し、及び項見出しの後に出現した節見出しとする。

更に、図表配置が制約される領域(制約領域)を避けて図表配置を行う制約処理では、制約領域以降に図表を配置する場合は、制約領域が終了する位置を参照箇所とみなす。また、制約領域以前に図表を

配置する場合は、制約領域の行数を予め全段落行数から差し引くことにより、均等処理が行われる。本来ならば制約領域に図表が配置されると、それを避けるためバックトラックして配置済みの図表を前方に再配置しなければならないが、本手法では文章や図表の残量を考慮することにより、これを解決している。

4. 実験結果による評価

4. 1 基準の評価

提案基準が実際の文書にどの程度適合するかを評価する(評価A)。対象文書は、論文50編(約440ページ、図表556個、片幅図表446個、両幅図表110個)であり、制約領域は始章内及び終章以降の領域とした。調査結果を表1に示す。

後方基準に不適応な図表は、大半が均等及び制約基準を満たすために参照箇所の前方に配置されたものである。即ち、後方基準と均等基準は相反しており、対象文書では参照箇所との距離を優先しているので、後方基準の適応率が低下したと思われる。均等基準に対して、図表サイズがほぼ1ページに当たるものが連続して参照されたため、参照箇所との距離が離れた図表が6個存在した。制約基準の対象となる図表は14個存在したが、すべて制約領域外に配置されていた。表1から全基準の平均適応率は93%となり、導入した基準は妥当であると言える。

表1 評価Aによる結果

基準	適応率(%)	不適応な図表数(個)
順序	99.3	2
安定	100	0
後方	70.1	81
関連	91.4	24
均等	98.9	3
制約	100	0

表2 評価Bによる結果

基準	適応率(%)	不適応な図表数(個)
順序	100	0
安定	100	0
後方	100	0
関連	93.5	18
均等	98.9	3
制約	100	0

4. 2 自動配置の実験結果による評価

評価Aと同じ文書データに対して提案手法の実験結果を評価する(評価B)。なお、均等処理における閾値は1とした。評価結果を表2に示す。但し、後方基準と均等基準は相反するため、均等処理を行った図表は後方基準の対象外とした。均等処理が行われた図表は146個あり、もし、均等処理をしなければ均等基準に反する図表は64個であった。この結果、均等処理をしなければ適応率は88%であるが、この処理を行えば99%と向上し、均等処理の有効性がわかる。表2から全基準の平均適応率は99%となり、本手法の有効性が確認できる。

更に、処理時間に関しては、文書構造抽出の計算量は文書の段落数 p に比例し $O(p)$ 、図表配置の計算量は文書の段落数と図表数 d に依存し、バックトラックは起こらないため、 $O(p+d)$ となる。実験システムは、計算機NEC PC-98AP2上に言語Cで開発され、評価データの中で段落数と図表数が最も多い文書A(段落数=112、図表数=23)において約0.6秒で文章構造を抽出し、約7ミリ秒で図表配置位置を決定し、全体でも1秒以内でレイアウトが行えた。また、各種データの記憶量も $O(p+d)$ であり、文書Aで約3キロバイトの容量で稼働できた。

5. まとめ

本稿では、図表配置に対する参照箇所の位置、関連のある文章領域、配置の均等性に対する基準を考慮し、バックトラックなしに高速に図表配置を決定する手法を提案した。また、50編の論文データに対する実験結果から、本手法による図表配置処理の有効性を実証した。

片幅と両幅に属さない図表や他の段組みに対する処理の考案、また他種の文書についても本稿で提案した基準とその基準に基づく自動配置手法の有効性を確かめることが今後の課題である。

参考文献

- (1)上林：“文書エディタの現状と将来展望”，情報学会誌，31-11，pp.1535-1542(1990-11)。
- (2)土井，福井，山口，竹林，岩井：“文書構造抽出技法の開発”，信学論(D-II)，J76-D-II，9，pp.2042-2052(1993-9)。
- (3)福井，山口，土井，岩井：“文書自動レイアウトシステムにおける図表配置候補生成方式”，情報論文誌，34-7，pp.1499-1506(1993-7)。