(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2009-211480
(P2009-211480A)
(43)公開日　平成21年9月17日(2009.9.17)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/21	(2006.01)	G06F	17/21	530T	5B075
G06F 17/30	(2006.01)	G06F	17/30	140	5B109
		G06F	17/30	350C	
		G06F	17/21	530E	

審査請求 未請求　請求項の数 39　ＯＬ　（全 16 頁）

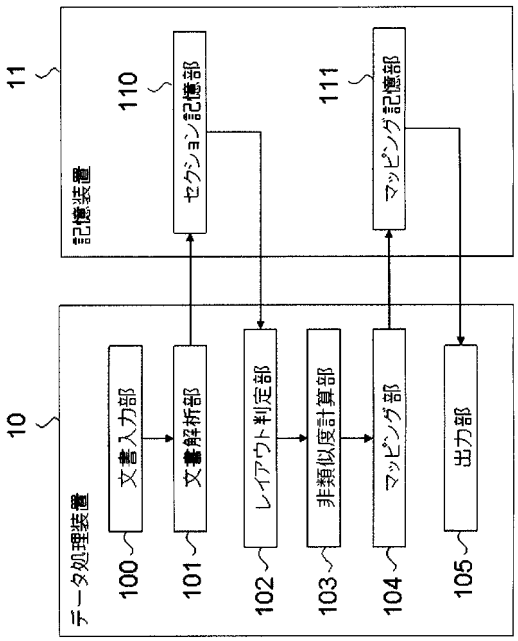
(21)出願番号　特願2008-54648(P2008-54648)
(22)出願日　平成20年3月5日(2008.3.5)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100079005
弁理士　宇高　克己
(72)発明者　森口　昌和
東京都港区芝五丁目７番１号　日本電気株式会社内
(72)発明者　辰巳　勇臣
東京都港区芝五丁目７番１号　日本電気株式会社内

Ｆターム(参考)　5B075 NR02　PR06　QM08
5B109 NH02　NH20

(54)【発明の名称】構造化文書処理システム、構造化文書処理方法及び構造化文書処理プログラム

(57)【要約】
【課題】本発明は、構造化文書をレンダリングせずにセクションのマッピングを行える構造化文書処理技術を提供することにある。また、構造化文書のレイアウト及び構造の両方を考慮したセクションのマッピングを行える構造化文書処理技術を提供することにある。
【解決手段】本発明は、構造化文書を構成している各タグに、レイアウトへの影響度に基づいて重みを割り当て、この割り当てた重み及び各構造化文書内のタグ構造に基づいて、構造化文書間の非類似度を計算する。
【選択図】図１



【特許請求の範囲】
【請求項１】
複数の構造化文書同士の非類似度を計算する構造化文書処理装置であって、
構造化文書を構成している各タグに、レイアウトへの影響度に基づいて重みを割り当てるレイアウト判定手段と、
前記割り当てた重み及び前記各構造化文書内のタグ構造に基づいて、構造化文書間の非類似度を計算する非類似度計算手段と
を有することを特徴とする構造化文書処理装置。
【請求項２】
前記計算された非類似度が小さい構造化文書同士を対応付けるマッピング手段を有することを特徴とする請求項１記載の構造化文書処理装置。
【請求項３】
マッピング手段は、複数の構造化文書を、非類似度が小さい構造化文書同士から順に対応付けていくことを特徴とする請求項２に記載の構造化文書処理装置。
【請求項４】
マッピング手段は、初めに対応付けた構造化文書を基準にして、各構造化文書を複数のグループに分割し、各グループ内の構造化文書同士を対応付けることを特徴とする請求項２又は請求項３に記載の構造化文書処理装置。
【請求項５】
前記非類似度計算手段は、前記割り当てた重み及び構造化文書を細分化したセクション内のタグ構造に基づいて、各構造化文書のセクション間の非類似度を計算することを特徴とする請求項１から請求項４のいずれかに記載の構造化文書処理装置。
【請求項６】
入力した構造化文書を解析してセクションを抽出する文書解析手段を有することを特徴とする請求項５に記載の構造化文書処理装置。
【請求項７】
レイアウト判定手段は、重みの算出基準として、構造化文書に予め定められているレイアウトの定義を用いることを特徴とする請求項１から請求項６のいずれかに記載の構造化文書処理装置。
【請求項８】
レイアウト判定手段は、重みの算出基準として、ユーザが予め定めた定義を用いることを特徴とする請求項１から請求項６のいずれかに記載の構造化文書処理装置。
【請求項９】
レイアウト判定手段は、重みの算出基準として、予めレンダリングした際に各タグのレイアウトへの影響度を測定し、それに基づいてシステムが定めた定義を用いることを特徴とする請求項１から請求項６のいずれかに記載の構造化文書処理装置。
【請求項１０】
非類似度計算手段は、非類似度の計算に、タグ構造をツリー型に変換した構造体を用いることを特徴とする請求項１から請求項９のいずれかに記載の構造化文書処理装置。
【請求項１１】
非類似度計算手段は、非類似度の計算に、タグ構造を単一バイトの文字列に変換した構造体を用いることを特徴とする請求項１から請求項９のいずれかに記載の構造化文書処理装置。
【請求項１２】
非類似度計算手段は、非類似度の計算に、構造体の編集距離を用いる請求項１０又は請求項１１に記載の構造化文書処理装置。
【請求項１３】
構造化文書のレイアウトと構造に基づいて構造化文書を対応付ける構造化文書処理システムであって、
前記構造化文書を構成する各タグにレイアウトへの影響度に基づいて重みを割り当てる

レイアウト判定手段と、
前記割り当てられた重み及び前記構造化文書のタグ構造に基づいて、構造化文書を細分化したセクション間の非類似度を計算する非類似度計算手段と、
前記計算された非類似度に基づいて構造化文書同士を対応付け、この構造化文書同士の対応付けを示す表示情報を生成するマッピング手段と
前記生成した表示情報を、通信ネットワークを介して情報端末に送信する情報配信手段と
を有することを特徴とする構造化文書処理システム。
【請求項１４】
複数の構造化文書同士の非類似度を計算する構造化文書処理方法であって、
構造化文書を構成している各タグに、レイアウトへの影響度に基づいて重みを割り当てるレイアウト判定ステップと、
前記割り当てた重み及び前記各構造化文書内のタグ構造に基づいて、構造化文書間の非類似度を計算する非類似度計算ステップと
を有することを特徴とする構造化文書処理方法。
【請求項１５】
前記計算された非類似度が小さい構造化文書同士を対応付けるマッピングステップを有することを特徴とする請求項１４に記載の構造化文書処理方法。
【請求項１６】
マッピングステップは、複数の構造化文書を、非類似度が小さい構造化文書同士から順に対応付けていくことを特徴とする請求項１５に記載の構造化文書処理方法。
【請求項１７】
マッピングステップは、初めに対応付けた構造化文書を基準にして、各構造化文書を複数のグループに分割し、各グループ内の構造化文書同士を対応付けることを特徴とする請求項１５又は請求項１６に記載の構造化文書処理方法。
【請求項１８】
前記非類似度計算ステップは、前記割り当てた重み及び構造化文書を細分化したセクション内のタグ構造に基づいて、各構造化文書のセクション間の非類似度を計算することを特徴とする請求項１４から請求項１７のいずれかに記載の構造化文書処理方法。
【請求項１９】
入力した構造化文書を解析してセクションを抽出する文書解析ステップを有することを特徴とする請求項１８に記載の構造化文書処理方法。
【請求項２０】
レイアウト判定ステップは、重みの算出基準として、構造化文書に予め定められているレイアウトの定義を用いることを特徴とする請求項１４から請求項９のいずれかに記載の構造化文書処理方法。
【請求項２１】
レイアウト判定ステップは、重みの算出基準として、ユーザが予め定めた定義を用いることを特徴とする請求項１４から請求項１９のいずれかに記載の構造化文書処理方法。
【請求項２２】
レイアウト判定ステップは、重みの算出基準として、予めレンダリングした際に各タグのレイアウトへの影響度を測定し、それに基づいてシステムが定めた定義を用いることを特徴とする請求項１４から請求項１９のいずれかに記載の構造化文書処理方法。
【請求項２３】
非類似度計算ステップは、非類似度の計算に、タグ構造をツリー型に変換した構造体を用いることを特徴とする請求項１４から請求項２２のいずれかに記載の構造化文書処理方法。
【請求項２４】
非類似度計算ステップは、非類似度の計算に、タグ構造を単一バイトの文字列に変換した構造体を用いることを特徴とする請求項１４から請求２２のいずれかに記載の構造化文

書処理方法。
【請求項２５】
非類似度計算ステップは、非類似度の計算に、構造体の編集距離を用いる請求項２３又は請求項２４に記載の構造化文書処理方法。
【請求項２６】
マッピングステップにおいて構造化文書同士を対応付けた結果を、通信ネットワークを介して送信する情報送信ステップを有することを特徴とする請求項１５から請求項２５のいずれかに記載の構造化文書処理方法。
【請求項２７】
複数の構造化文書同士の非類似度を計算するプログラムであって、前記プログラムは、情報処理装置に、
構造化文書を構成している各タグに、レイアウトへの影響度に基づいて重みを割り当てるレイアウト判定処理と、
前記割り当てた重み及び前記各構造化文書内のタグ構造に基づいて、構造化文書間の非類似度を計算する非類似度計算処理と
を実行させることを特徴とするプログラム。
【請求項２８】
前記計算された非類似度が小さい構造化文書同士を対応付けるマッピング処理を実行させることを特徴とする請求項２７に記載のプログラム。
【請求項２９】
マッピング処理は、複数の構造化文書を、非類似度が小さい構造化文書同士から順に対応付けていくことを特徴とする請求項２８に記載のプログラム。
【請求項３０】
マッピング処理は、初めに対応付けた構造化文書を基準にして、各構造化文書を複数のグループに分割し、各グループ内の構造化文書同士を対応付けることを特徴とする請求項２８又は請求項２９に記載のプログラム。
【請求項３１】
前記非類似度計算処理は、前記割り当てた重み及び構造化文書を細分化したセクション内のタグ構造に基づいて、各構造化文書のセクション間の非類似度を計算することを特徴とする請求項２７から請求項３０のいずれかに記載のプログラム。
【請求項３２】
入力した構造化文書を解析してセクションを抽出する文書解析処理を実行させることを特徴とする請求項３１に記載のプログラム。
【請求項３３】
レイアウト判定処理は、重みの算出基準として、構造化文書に予め定められているレイアウトの定義を用いることを特徴とする請求項２７から請求項３２のいずれかに記載のプログラム。
【請求項３４】
レイアウト判定処理は、重みの算出基準として、ユーザが予め定めた定義を用いることを特徴とする請求項２７から請求項３２のいずれかに記載のプログラム。
【請求項３５】
レイアウト判定処理は、重みの算出基準として、予めレンダリングした際に各タグのレイアウトへの影響度を測定し、それに基づいてシステムが定めた定義を用いることを特徴とする請求項２７から請求項３２のいずれかに記載のプログラム。
【請求項３６】
非類似度計算処理は、非類似度の計算に、タグ構造をツリー型に変換した構造体を用いることを特徴とする請求項２７から請求項３５のいずれかに記載のプログラム。
【請求項３７】
非類似度計算処理は、非類似度の計算に、タグ構造を単一バイトの文字列に変換した構造体を用いることを特徴とする請求項２７から請求３５のいずれかに記載のプログラム。

【請求項３８】
非類似度計算処理は、非類似度の計算に、構造体の編集距離を用いる請求項３６又は請求項３７に記載のプログラム。
【請求項３９】
マッピング処理における対応付けの結果を、通信ネットワークを介して送信する情報送信処理を実行させることを特徴とする請求項２８から請求項３８のいずれかに記載のプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、構造化文書を処理する構造化文書処理システム、構造化文書処理方法及び構造化文書処理プログラムに関し、特にＨＴＭＬ（Ｈｙｐｅｒ　Ｔｅｘｔ　Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ）のような特定の文書型定義（ＤＴＤ：Ｄｏｃｕｍｅｎｔ　Ｔｙｐｅ　Ｄｅｆｉｎｉｔｉｏｎ）に基づいてレイアウトを形成する構造化文書において、特定の条件で分割された領域を、レイアウトと構造の両方が類似した別の領域とマッピングする構造化文書処理システム、構造化文書処理方法及び構造化文書処理プログラムに関する。
【背景技術】
【０００２】
近年、大量の情報を含むＷｅｂコンテンツなどの構造化文書から、自動的に文書構造を解析して複数のセクションを抽出して利用するシステムが研究されている。例えば、セクションの中からユーザが必要な情報を、自動あるいは手動で選択して利用する情報提供システムがある。なお、セクションとは、構造化文書を細分化した領域のことで、構造化文書と同様にマークアップ言語（以下、タグ）で構成される。
【０００３】
しかし、その情報提供システムなどのように、常に内容が変化していくＷｅｂコンテンツのセクションを利用したアプリケーションでは、時間経過に応じてそのセクションの表示位置が変わったり、消滅したりするため、ユーザに間違ったセクションの情報を提供してしまう場合がある。
【０００４】
そこで、特許文献１では、構造化文書のレイアウトの特徴に注目し、セクションの抽出順序やセクションの表示座標、および見出しなどのレイアウト情報を用いることで、セクションのレイアウトが変わっても、セクションの位置を推定する方法が提案されている。
【０００５】
また、特許文献２で提案されている構造化文書同士の類似度を検出する技術をセクションに応用し、変化前と変化後とのセクションの文書構造を比較して、より類似した構造を持つセクション同士をマッピングさせる方法が考えられる。特に、構造化文書の文書構造がツリー型で表現できるため、そのツリーの編集距離を用いて構造化文書の類似性を判定する。
【特許文献１】特開２００７‐２９３５４３号
【特許文献２】特開２００７－０５２５５６号
【発明の開示】
【発明が解決しようとする課題】
【０００６】
第１の問題点は、特許文献１のように座標などのレイアウトの特徴を用いて、より類似したレイアウトのセクション同士をマッピングする場合、構造化文書を少なくとも１度はレンダリングして、レイアウト情報を取得しなければならないということである。
【０００７】
第２の問題点は、レンダリングせずに、特許文献２のように文書構造の類似性を判定して、より類似した構造のセクション同士をマッピングする場合、レイアウトを考慮していないために適切なマッピングができないということである。

【０００８】
本発明が解決しようとする課題は、構造化文書をレンダリングせずにセクションのマッピングを行える構造化文書処理システム、構造化文書処理方法および構造化文書処理プログラムを提供することにある。
【０００９】
また、構造化文書のレイアウト及び構造の両方を考慮したセクションのマッピングを行える構造化文書処理システム、構造化文書処理方法および構造化文書処理プログラムを提供することにある。
【課題を解決するための手段】
【００１０】
上記課題を解決するための本発明は、複数の構造化文書同士の非類似度を計算する構造化文書処理装置であって、構造化文書を構成している各タグに、レイアウトへの影響度に基づいて重みを割り当てるレイアウト判定手段と、前記割り当てた重み及び前記各構造化文書内のタグ構造に基づいて、構造化文書間の非類似度を計算する非類似度計算手段とを有することを特徴とする。
【００１１】
上記課題を解決するための本発明は、構造化文書のレイアウトと構造に基づいて構造化文書を対応付ける構造化文書処理システムであって、前記構造化文書を構成する各タグにレイアウトへの影響度に基づいて重みを割り当てるレイアウト判定手段と、前記割り当てられた重み及び前記構造化文書のタグ構造に基づいて、構造化文書を細分化したセクション間の非類似度を計算する非類似度計算手段と、前記計算された非類似度に基づいて構造化文書同士を対応付け、この構造化文書同士の対応付けを示す表示情報を生成するマッピング手段と前記生成した表示情報を、通信ネットワークを介して情報端末に送信する情報配信手段とを有することを特徴とする。
【００１２】
上記課題を解決するための本発明は、複数の構造化文書同士の非類似度を計算する構造化文書処理方法であって、構造化文書を構成している各タグに、レイアウトへの影響度に基づいて重みを割り当てるレイアウト判定ステップと、前記割り当てた重み及び前記各構造化文書内のタグ構造に基づいて、構造化文書間の非類似度を計算する非類似度計算ステップとを有することを特徴とする。
【００１３】
上記課題を解決するための本発明は、複数の構造化文書同士の非類似度を計算するプログラムであって、前記プログラムは、情報処理装置に、構造化文書を構成している各タグに、レイアウトへの影響度に基づいて重みを割り当てるレイアウト判定処理と、前記割り当てた重み及び前記各構造化文書内のタグ構造に基づいて、構造化文書間の非類似度を計算する非類似度計算処理とを実行させることを特徴とする。
【発明の効果】
【００１４】
本発明によると、レンダリングしなくてもセクションのマッピングを行えることにある。その理由は、レイアウトではなくセクションの構成タグを比較対象とするためである。
【００１５】
また、本発明によると、セクションのレイアウト及び構造の両方に基づいたマッピングを行えることにある。その理由は、各タグのレイアウトへの影響度、およびセクションのタグ構造を判定要素とするためである。
【発明を実施するための最良の形態】
【００１６】
本発明の特徴を説明するために、以下において、図面を参照して具体的に述べる。
【００１７】
本発明による構造化文書処理システムの特徴は、レイアウト判定部１０２と、非類似度計算部１０３と、マッピング部１０４とを有する点である。

【００１８】
レイアウト判定部１０２は、構造化文書の各タグのレイアウトへの影響度に基づいて、重みをタグに割り当てる。
【００１９】
非類似度計算部１０３は、タグ構造、およびレイアウト判定部１０２で各タグに割り当てられた重みに基づいて、比較するセクション同士の非類似度を計算する。
【００２０】
マッピング部１０４は、非類似度計算部１０３で算出した非類似度に基づいて、適切に構造化文書をマッピングする。
【００２１】
図１は、本発明による構造化文書処理システムの構成の一例を示すブロック図である。本実施の形態では、構造化文書処理システムは、ハードウェアで構成することも可能であるが、以下ではプログラムに従って動作するパーソナルコンピュータなどの情報処理端末によって実現する場合を用いて説明する。尚、構造化文書処理システムは、構造化文書を複数のセクションに分割して配信するシステム等のビジネスモデルに適用されてもよい。この場合、構造化文書処理システムは、例えば、構造化文書をレンダリングするソフトウェアを搭載した携帯電話やＰＤＡ、パーソナルコンピュータ等のユーザ端末と、構造化文書を処理する構造化文書処理サーバとを含んでもよい。
【００２２】
図１に示すように、本実施の形態では、構造化文書処理システムは、プログラム制御により動作するデータ処理装置１０と、情報を記憶する記憶装置１１とを有する。
【００２３】
データ処理装置１０は、具体的には、プログラムに従って動作するパーソナルコンピュータやサーバ等によって実現される。
【００２４】
データ処理装置１０は、文書入力部１００と、文書解析部１０１と、レイアウト判定部１０２と、非類似度計算部１０３と、マッピング部１０４と、出力部１０５とを有する。また、記憶装置１１は、具体的には、メモリやハードディスク装置等によって実現される。記憶装置１１は、セクション記憶部１１０と、マッピング記憶部１１１とを有する。
【００２５】
文書入力部１００は、外部から構造化文書を取得し、文書解析部１０１に出力する機能を備える。例えば、文書入力部１００は、ユーザの操作に従って、記憶装置１１から構造化文書を読み出し、文書解析部１０１に出力する。また、例えば、文書入力部１００は、インターネット等の通信ネットワークを介して構造化文書（例えば、Ｗｅｂコンテンツなど）を受信し、文書解析部１０１に出力する。
【００２６】
文書解析部１０１は、文書入力部１００から取得した構造化文書を解析して、複数のセクションを抽出し、セクション記憶部１１０に記憶させる機能を備える。なお、取得した構造化文書をそのまま単一セクションとしてセクション記憶部１１０に記憶させてもよい。
【００２７】
レイアウト判定部１０２は、セクション記憶部１１０からセクションを取得し、各セクションを構成するタグのレイアウトへの影響度に基づいて、各タグに重みを割り当てる機能を備える。また、レイアウト判定部１０２は、タグに重みを割り当てたセクションを、非類似度計算部に出力する機能を備える。なお、ここで説明するセクションには、セクションの集合であるセクショングループ（構造化文書そのものを含む）を含んでもよい。
【００２８】
例えば、レイアウト判定部１０２は、タグのレイアウト定義をＤＴＤ（Ｄｏｃｕｍｅｎｔ　Ｔｙｐｅ　Ｄｅｆｉｎｉｔｉｏｎ）から取得し、各タグのレイアウトへの影響度をブロック要素（見出しや段落など、レイアウトを構成する基本要素）およびインライン要素

（強調やリンクなど、表示情報に役割や機能を与える要素）の２種類に分類する。そして、レイアウト判定部１０２は、ブロック要素に重みを大きく与え、一方インライン要素には重みを小さく与えることによって、レイアウトへの影響度に基づいた重み付けをする。

【００２９】

非類似度計算部１０３は、レイアウト判定部１０２から比較するセクション（少なくとも２以上のセクション）を取得して、セクションのタグ構造に基づいて、セクション間の構造の類似性を示す非類似度を計算する機能を備える。また、非類似度計算部１０３は、セクション間の非類似度を、マッピング部１０４に出力する機能を備える。

【００３０】

例えば、非類似度計算部１０３は、各セクションのタグ構造をツリー型に変換して、ツリー同士の編集距離を計算する。その際、レイアウト判定部１０２で計算した重みをツリーの各ノードの編集コストとすることによって、レイアウトへの影響度およびセクションの構造の両方に基づいた非類似度を計算する。

【００３１】

マッピング部１０４は、非類似度計算部１０３から各セクションの非類似度を取得し、最も非類似度が小さい、即ち最も類似しているセクションの組み合わせをマッピングする（対応付ける）機能を備える。また、マッピング部１０４は、セクションのマッピング結果を、マッピング記憶部１１１に記憶させる機能を備える。例えば、マッピング部１０４は、構造化文書Ｄａと構造化文書Ｄｂとの比較において、最も非類似度が小さいセクションの組み合わせから順に、ＤａとＤｂとのセクションをすべてマッピングする。また、セクション数の違いからマッピングできずに残ってしまったセクションは、空セクションφとマッピングする。なお、マッピング部１０４は、セクションの相対位置を考慮し、マッピングしたセクションを基準に、構造化文書内のセクションの集合を２つのグループに分けて、その各グループ内でマッピングを行ってもよい。また、マッピング部１０４は、セクションの絶対位置を考慮し、構造化文書内のセクションの階層構造に基づいて、セクションの集合を複数のグループに分け、その各グループ内でマッピングを行ってもよい。

【００３２】

出力部１０５は、マッピング記憶部１１１が記憶しているセクションのマッピング結果を表示情報として外部に出力する機能を備える。例えば、出力部１０５は、マッピング記憶部１１１から、ユーザが指定したセクションと、そのセクションとマッピングされているセクションとを抽出し、液晶表示部やディプレイ装置等の表示装置に表示させたり、通信ネットワークを介して情報端末に送信したりする。

【００３３】

記憶装置１１は、セクション記憶部１１０と、マッピング記憶部１１１とを含む。

【００３４】

セクション記憶部１１０は、文書解析部１０１が解析した構造化文書のセクションを、構造化文書毎に記憶する。

【００３５】

マッピング記憶部１１１は、マッピング部１０４が計算したセクションのマッピング結果を記憶する。

【００３６】

次に、動作について説明する。図２は、構造化文書処理システムがレイアウトと文書構造に基づいてセクションをマッピングする処理の一例を示す流れ図である。

【００３７】

まず、文書入力部１００は、構造化文書を取得する。例えば、文書入力部１００は、記憶装置１１に格納されている構造化文書を読み出す。また、例えば、文書入力部１００は、通信ネットワークを介して構造化文書（例えば、Ｗｅｂコンテンツなど）を受信する。

【００３８】

次に、文書解析部１０１は、文書入力部１００から取得した構造化文書を解析して、複数のセクションを抽出し、構造化文書毎にセクション記憶部１１０に記憶させる（ステッ

プＳ１１）。
【００３９】
続いて、レイアウト判定部１０２は、セクション記憶部１１０からセクションを取得し、タグのレイアウトへの影響度に基づいて、各タグに重みを割り当てる（ステップＳ１２）。
【００４０】
次に、非類似度計算部１０３は、レイアウト判定部１０２から各構造化文書のセクションを取得し、タグに割り当てられた重みとタグの構造とに基づいて、比較元のセクションとこれ以外のセクションとの間の各非類似度を計算する（ステップＳ１３）。
【００４１】
続いて、マッピング部１０４は、非類似度計算部１０３からセクションの非類似度を取得して、その非類似度に基づいて複数の構造化文書のセクション同士をマッピングし、マッピング記憶部１１１に記憶させる（ステップＳ１４）。
【００４２】
また、出力部１０５は、マッピング記憶部１１１が記憶するセクションのマッピング結果を出力する。
【００４３】
以上のように、本実施の形態によれば、レイアウトへの影響度に基づいてタグに重みを割り当て、その重みとセクションのタグ構造に基づいて非類似度を計算し、非類似度に基づいてセクションをマッピングすることにより、レンダリングせずにレイアウトや文書構造に基づいた適切なセクションのマッピングが可能となる。
【００４４】
例えば、ある２つの構造化文書の比較において、微妙にレイアウトが異なっているセクションがある場合、あるセクションを構成するタグの種類や構造が、他のセクションと比べて最も類似していれば、それをマッピングすることができる。また、定性的にしか把握できなかった類似性を、非類似度という形で定量的に把握することができる。
【００４５】
次に、本発明による構造化文書処理システムの具体的な実施例について説明する。
【００４６】
まず、構造化文書処理システムの第１の実施例について説明する。なお、本実施例における構造化文書処理システムは、第１の実施の形態で示した構造化文書処理システムに相当する。また、本実施例では、データ処理装置がパーソナルコンピュータであり、データ記憶装置が磁気ディスク装置であるものとする。
【００４７】
パーソナルコンピュータ（データ処理装置）は、文書入力手段、文書解析手段、レイアウト判定手段、非類似度計算手段、マッピング手段、及び出力手段として機能する中央演算装置を含む。また、磁気ディスク装置（記憶装置）は、パーソナルコンピュータによって解析または計算されたセクション情報や非類似度情報を記憶する。なお、データ処理装置は、サーバや携帯電話等でもよく、端末の種類によらない。また、本実施例では、構造化文書の例として、Ｗｅｂコンテンツを対象とする。例えば、パーソナルコンピュータは、インターネットを介してＷｅｂコンテンツを受信する。
【００４８】
本実施例では、まず、中央演算装置は、Ｗｅｂコンテンツを受信して、受信したＷｅｂコンテンツを解析してセクションを抽出する。そして、中央演算装置は、抽出したセクション情報を磁気ディスク装置に記憶させる。なお、受信したＷｅｂコンテンツをそのまま単一セクションとしてもよい。また、構造化文書は、レンダリングするために作成されたものであれば、ＨＴＭＬやＸＭＬなどの種類に寄らない。本実施例では、構造化文書としてＨＴＭＬを扱う。
【００４９】
図３は、１つの構造化文書の文書構造例を示す図であり、図４は、２つの構造化文書が

解析されて複数のセクションが抽出された後のセクションのレイアウト構成を示す図である。本実施例では、中央演算装置は、図３のように、構造化文書Ｄ１から複数のセクションを抽出し、図４のように、セクションの抽出順序に基づいてそれぞれセクション１～８のように番号を割り振る。もう一方の構造化文書Ｄ２でも同様にセクションを抽出し、そのセクション情報を磁気ディスク装置に記憶させる。

【００５０】

次に、中央演算装置は、磁気ディスク装置からセクション情報を取得し、各セクションのタグにレイアウトへの影響度に基づいて重みを割り当てる。本実施例では、中央演算装置は、各タグに割り当てる重みを、ＤＴＤに基づいて、ブロック要素（見出しや段落など、レイアウトを構成する基本要素）およびインライン要素（強調やリンクなど、表示情報に役割や機能を与える要素）の２つの種類に分けて計算する。なお、磁気ディスク装置から取得するセクションは、構造化文書を解析したセクションに限らず、他のセクションや、構造化文書そのものでもよい。また、取得するセクションあるいは構造化文書は複数でもよい。

【００５１】

図５は、図３の構造化文書Ｄ１のタグをブロック要素およびインライン要素別に分け、重みとしてそれぞれ１００と１を割り当てた例を示す説明図である。なお、図５の重みの値は、レイアウトへの影響度を強く評価したいため、ブロック要素を示す“ｄｉｖ”等に対して「１００」、インライン要素を示す“ｈ”、“ａ”、“ｉｍｇ”等に対して「１」のように、重みの差を大きくしたが、例えば、レイアウトへの影響度を緩く評価するならば、ブロック要素を「３」、インライン要素を「１」のように、重みの差を小さくしてもよい。また、レイアウトではなく、タグの構造の類似性を強く評価したいならば、ブロック要素およびインライン要素を共に１にして、重みの差を無くしてもよい。また、レイアウトへの影響度を決めるタグの種類は、ブロック要素およびインライン要素という定義を使用せずに、ＤＴＤの別の定義や、ＣＳＳ（Cascading Style Sheets）などのＤＴＤ以外の構造化文書内の要素の表示を定義したレイアウト定義に従ってもよい。また、ユーザが予め定義したものでもよい。また、予めレンダリングした際に各タグのレイアウトへの影響度を計算し、その記憶している結果に基づいてもよい。

【００５２】

次に、中央演算装置は、セクションのタグ構造に基づいて、比較するセクションとの編集距離を計算する。本実施例では、タグの構造をツリー型に変換し、ツリー構造の編集距離を計算する。さらに、ノード一つの編集コストを各タグの重みとし、それに基づいて編集距離から非類似度を計算する。

【００５３】

図６は、図５のセクション１をツリー型に変換した例を示す説明図である。なお、図６のＲはツリーの根（Ｒｏｏｔ）を示している。また、図６では、重みの値が大きいノードは表示サイズを大きく、逆に重みの値が小さいノードは小さく表現している。

【００５４】

図７は、図６のセクション１のツリーを比較元として、別のセクションであるセクション２、セクション３、セクション４のそれぞれとの編集距離および非類似度の計算例を示す説明図である。本実施例では、ツリーの編集距離の計算において、レイアウトへの影響度を中心に計算するため、レイアウト要素（ブロック要素やインライン要素など）が同じならば、タグ名に関係しない。また、本実施例では、置換を使用せずに、置換＝削除＋挿入と見なす。さらに、本実施例では、レイアウトにツリーの兄弟順序も影響すると考え、ツリーの兄弟要素を入れ替えての構造一致は許可しない。

【００５５】

例えば、セクション２のツリー構造を示した（ａ）のツリーとの編集距離の計算において、比較元のツリーからインライン要素“ａ”と“ｉｍｇ”との２つを削除（ｄｅｌ）した構造と同等になるため、編集距離が２となり、重みに基づいた非類似度は２となる。

【００５６】

同様に、セクション３のツリー構造を示した（ｂ）のツリーは、比較元のツリーから、インライン要素を１つ削除、１つ挿入（ｉｎｓ）、ブロック要素を１つ削除した構造と同等になるため、編集距離は３となり、重みに基づいた非類似度は１０２となる。

【００５７】

さらに、セクション４のツリー構造を示した（ｃ）のツリーは、比較元のツリーから、インライン要素を２つ削除、２つ挿入した構造と同等になるため、編集距離は４となり、重みに基づいた非類似度は４となる。

【００５８】

なお、類似判定を厳密にするため、同じ要素でもタグ名が異なっていれば編集コストに一定の重みを与えて置換を使用してもよい。また、類似判定を緩和するため、ツリーの兄弟要素を入れ替えての構造一致を許可してもよい。また、タグの構造を用いるのならば、ツリー型に変換しなくてもよい。例えば、ブロック要素を１、及びインライン要素を０として、二進法的な表現にタグ構造を変換し、変換した文字列の編集距離を求めてもよい。また、中央演算装置は、非類似度を計算した時点で処理を終えてもよい。この場合、磁気ディスク装置に記憶される情報は、非類似度となる。

【００５９】

次に、中央演算装置は、最も非類似度が小さいセクションの組み合わせから順にマッピングする。本実施例では、マッピングしたセクションを基準にセクションを２つのグループに分割し、それぞれのグループでセクションをマッピングするという処理を繰り返す。

【００６０】

図８は、図４の構造化文書Ｄ１とＤ２とのマッピング例を示す説明図である。例えば、まず、最も非類似度が小さいセクション４（Ｄ１）とセクション４（Ｄ２）との組み合わせをマッピングする。このマッピングによって、構造化文書が２つのグループに分割され、それぞれＤ１ではＧ１１とＧ１２、Ｄ２ではＧ２１とＧ２２となる。

【００６１】

次に、それぞれのグループで最も非類似度が小さいセクションの組み合わせをマッピングする。例えば、図９のように、Ｇ１１：Ｇ２１では、セクション２（Ｄ１）とセクション３（Ｄ２）、およびＧ１２：Ｇ２２では、セクション８（Ｄ１）とセクション９（Ｄ２）との組み合わせをマッピングする。

【００６２】

そして、Ｄ１ではグループＧ１１がＧ１１１とＧ１１２に、Ｇ１２がＧ１２１とＧ１２２に、一方Ｄ２ではＧ２１がＧ２１１とＧ２１２に、Ｇ２２がＧ２２１とＧ２２２に分割される。なお、Ｇ１２２、Ｇ２１２及びＧ２２２はセクションが存在しない空グループである。

【００６３】

続いて、今までと同様にそれぞれのグループで最も非類似度が小さいセクションの組み合わせをマッピングする。例えば、図１０のように、Ｇ１１１：Ｇ２１１では、セクション１（Ｄ１）とセクション１（Ｄ２）、及びＧ１２１：Ｇ２２１では、セクション５（Ｄ１）とセクション７（Ｄ２）の組み合わせをマッピングする。

【００６４】

そして、Ｄ１ではグループＧ１１１がＧ１１１１とＧ１１１２に、Ｇ１２１がＧ１２１１とＧ１２１２に、一方Ｄ２ではＧ２１１がＧ２１１１とＧ２１１２に、Ｇ２２１がＧ２２１１とＧ２２１２に分割される。ここで、Ｇ１１２：Ｇ２１２では、セクション３（Ｄ１）の組み合わせの相手がＤ２に存在しないため、存在しないセクションφとマッピングされる。

【００６５】

以下も同様に分割されたグループでのマッピングを繰り返し、最終的に図１１のようなマッピング結果になる。

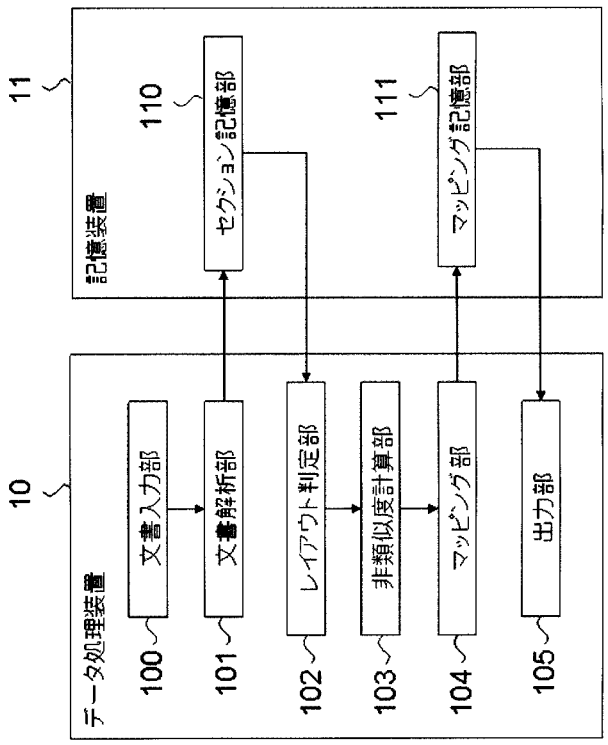
【００６６】

なお、図１２のように、セクションの文書内における階層構造を利用してグループ分割

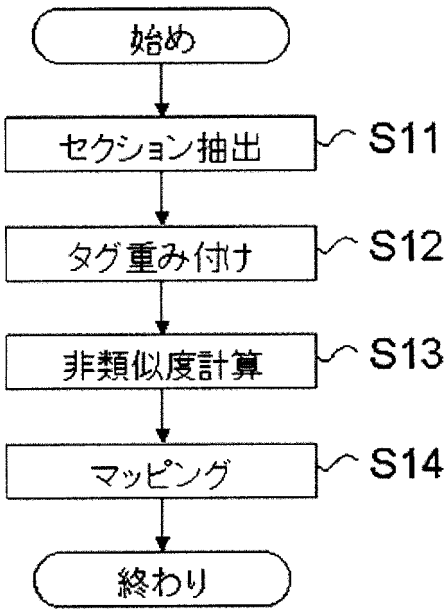
してもよい。
【００６７】
例えば、深さ２まで探索すると、まず、図１３（ａ）のように、深さ１の階層において、グループのマッピングを行う。次に、図１３（ｂ）のように、深さ２の階層において、グループのマッピングを行う。その後に、それぞれの分割されたグループ内において、セクションのマッピングを行う。
【００６８】
上述した本発明は、レイアウトを持った構造化文書を複数のセクションに分割して利用するアプリケーションにおいて、レイアウトと文書構造に基づいて類似したセクションに適切にマッピングすることができるため、構造化文書の変化に強いマッピングが可能となる。また、構造化文書の類似性を定量的に把握できるようになる。
【００６９】
例えば、ブログなどのレイアウト構成がよく変化するＷｅｂコンテンツにおいて、特定のセクションの更新情報を管理するアプリケーションでは、他のセクションが削除あるいは追加され、セクションの構成が変化したとしても、その特定のセクションを一貫して管理することができる。
【００７０】
また、Ｗｅｂコンテンツ全体の更新情報を管理するアプリケーションでも、どのセクションが削除されたり追加されたりしたかを識別することができる。さらに、全くＵＲＬが異なるＷｅｂコンテンツ同士でも、その類似性をセクション単位で把握できるため、特定の類似性をもったセクションをすべてのＷｅｂコンテンツで非表示にするなどの処理が可能となる。
【図面の簡単な説明】
【００７１】
【図１】構造化文書処理システムの他の構成例を示すブロック図である。
【図２】構造化文書処理システムがセクションをマッピングする処理の一例を示す流れ図である。
【図３】構造化文書の文書構造の例を示す説明図である。
【図４】セクションのレイアウト構成を示す説明図である。
【図５】ブロック要素およびインライン要素のタグに対して、重みとしてそれぞれ１００と１を割り当てた例を示す説明図である。
【図６】セクションをツリー型に変換する例を示す説明図である。
【図７】編集距離および非類似度の計算例を示す説明図である。
【図８】構造化文書Ｄ１とＤ２とのマッピング手順を示す説明図である。
【図９】構造化文書Ｄ１とＤ２とのマッピング手順を示す説明図である。
【図１０】構造化文書Ｄ１とＤ２とのマッピング手順を示す説明図である。
【図１１】構造化文書Ｄ１とＤ２とのマッピングが完了した例を示す説明図である。
【図１２】構造化文書Ｄ１の階層構造例を示す説明図である。
【図１３】構造化文書Ｄ１とＤ２とのの階層構造を利用したマッピング例を示す説明図である。
【符号の説明】
【００７２】
１０　データ処理装置
１１　記憶装置
１００　文書入力部
１０１　文書解析部
１０２　レイアウト判定部
１０３　非類似度計算部
１０４　マッピング部
１０５　出力部

１１０　セクション記憶部
１１１　マッピング記憶部

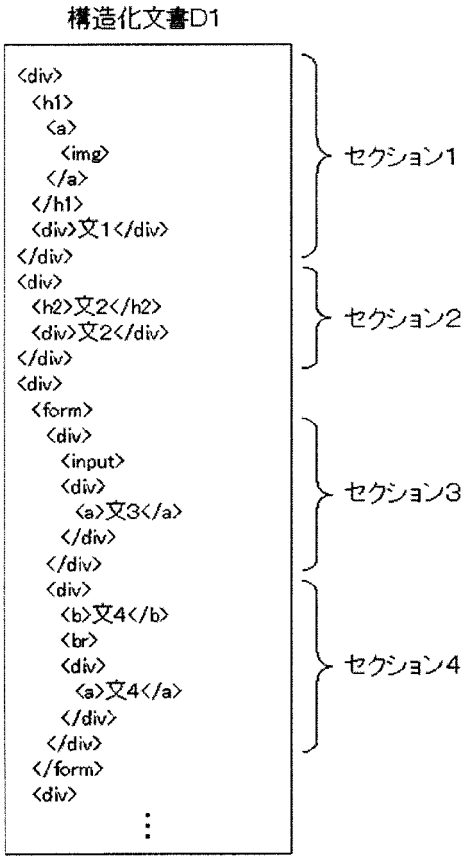
【図１】



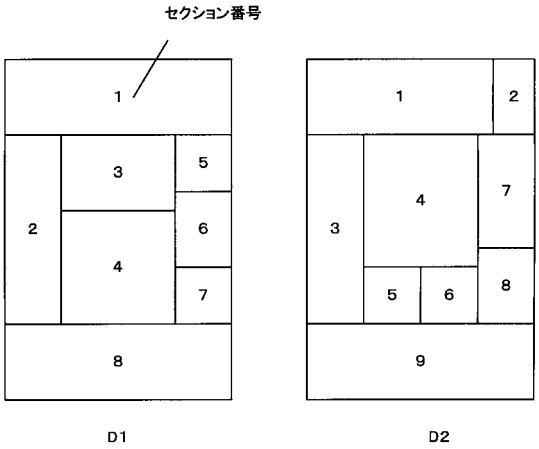
【図２】



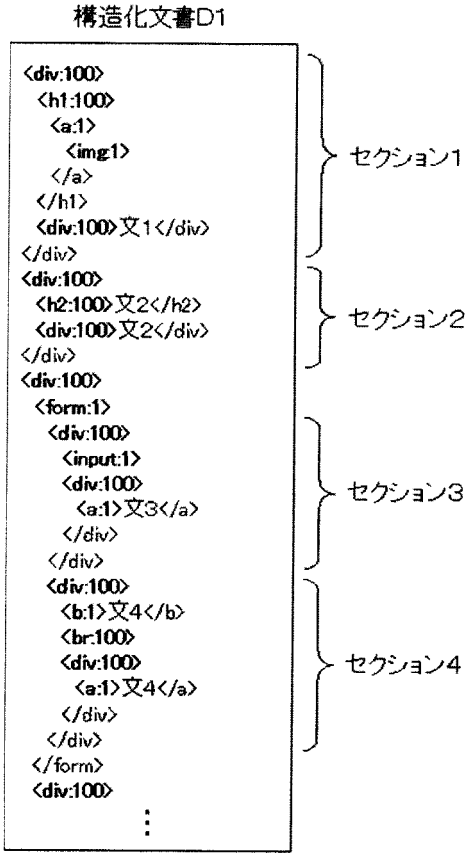
【図３】



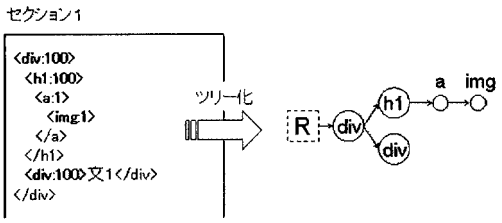
【図４】



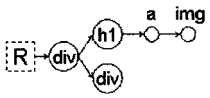
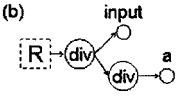
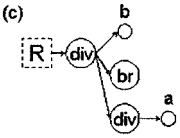
【図５】



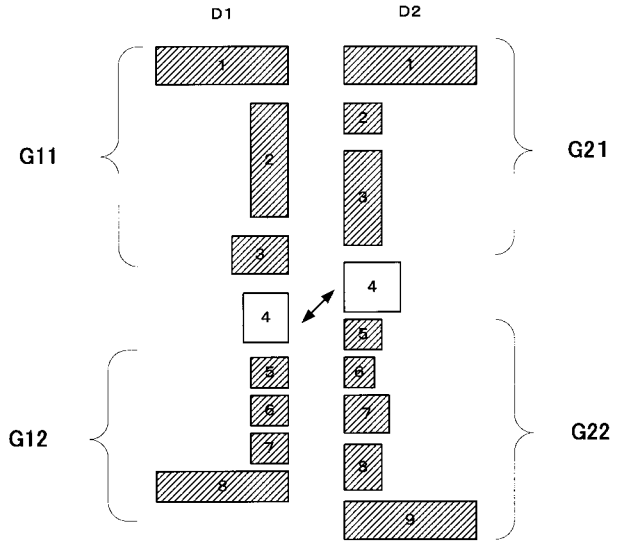
【図６】



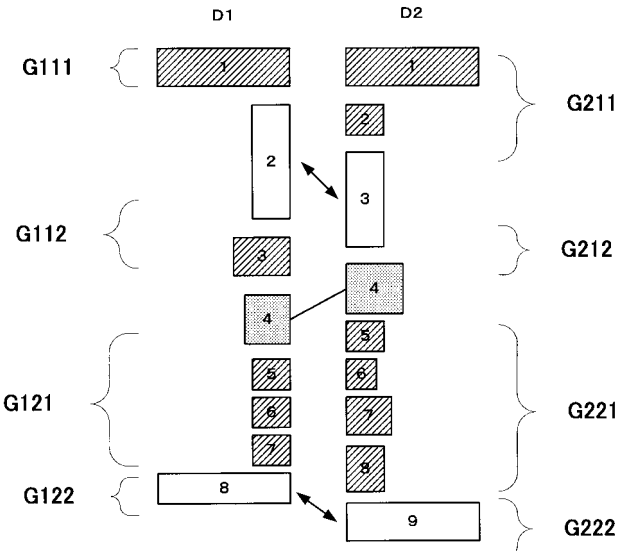
【図７】

比較元セクション	非類似度	比較先セクション
	**2** inline:del:2	(a) 
	102 inline:del:1 inline:ins:1 block:del:1	(b) 
	4 inline:del:2 inline:ins:2	(c) 

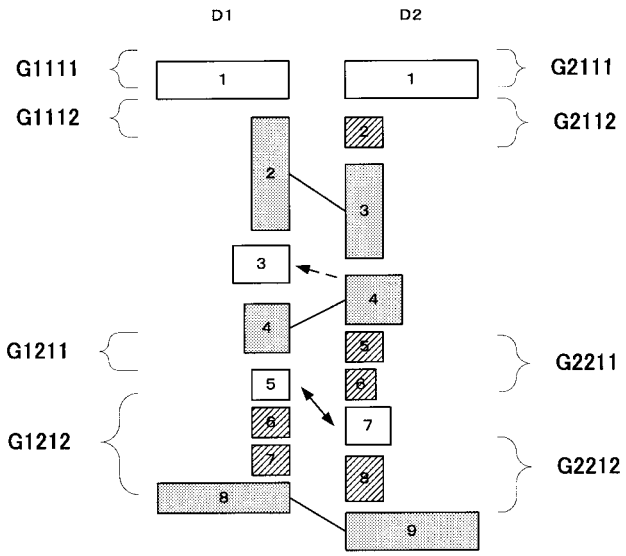
【図８】


D1
D2
G11
G12
G21
G22

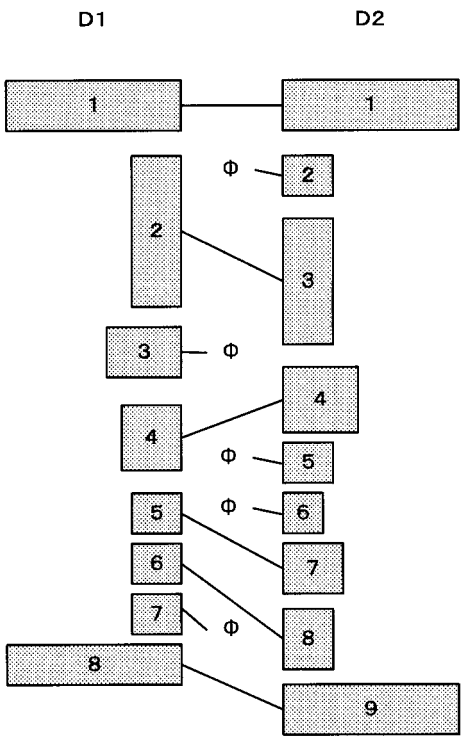
【図９】


D1
D2
G111
G112
G121
G122
G211
G212
G221
G222

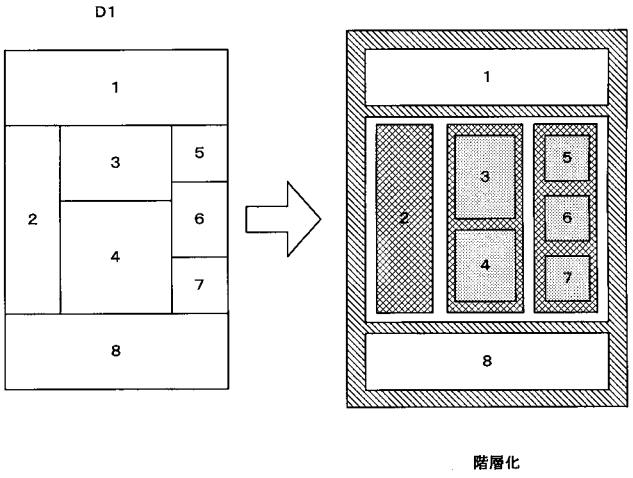
【図１０】


D1
D2
G1111
G1112
G1211
G1212
G2111
G2112
G2211
G2212

【図１１】


D1
D2

【図１２】



【図１３】