(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3703080号
(P3703080)
(45)発行日　平成17年10月5日(2005.10.5)　(24)登録日　平成17年7月29日(2005.7.29)

(51)Int.Cl.[7]	F I		
G06F 12/00	G06F	12/00	546R
G06F 3/00	G06F	3/00	651B
G06F 3/14	G06F	3/14	310B
G06F 13/00	G06F	13/00	550B

請求項の数 23　(全 27 頁)

(21)出願番号　特願2000-227996(P2000-227996)
(22)出願日　平成12年7月27日(2000.7.27)
(65)公開番号　特開2002-55872(P2002-55872A)
(43)公開日　平成14年2月20日(2002.2.20)
審査請求日　平成13年6月15日(2001.6.15)

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国１０５０４　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
(74)代理人　100086243
弁理士　坂口　博
(74)代理人　100091568
弁理士　市位　嘉宏

最終頁に続く

(54)【発明の名称】ウェブコンテンツを簡略化するための方法、システムおよび媒体

(57)【特許請求の範囲】
【請求項１】
コンピュータシステムが、簡略化対象の目的ページを取得するステップと、
前記コンピュータシステムが、前記目的ページに隣接する隣接ページを取得するステップと、
前記コンピュータシステムが、前記目的ページと前記隣接ページとで共通するオブジェクトを前記目的ページから削除する差分演算を施し、簡略化ページを生成するステップと、
を含み、
前記隣接ページは、前記目的ページのＵＲＬまたは前記目的ページに含まれるリンクのＵＲＬとディレクトリが共通するＵＲＬのページ、または、
前記目的ページのＵＲＬまたは前記目的ページに含まれるリンクのＵＲＬと親ディレクトリが共通するＵＲＬのページ、または、
前記目的ページのＵＲＬを含むルートディレクトリ以下の各ディレクトリのトップページ、である
ウェブコンテンツを簡略化するための方法。
【請求項２】
コンピュータシステムが、簡略化対象の目的ページを取得するステップと、
前記コンピュータシステムが、前記目的ページに隣接する隣接ページを取得するステップと、

前記コンピュータシステムが、前記目的ページと前記隣接ページとで共通するオブジェクトを前記目的ページから削除する差分演算を施し、簡略化ページを生成するステップと、
を含み、
前記隣接ページは、前記目的ページの過去のページにおいて前記過去ページに含まれていたリンクのページ、または、
前記目的ページまたは前記隣接ページの過去のページ、である
ウェブコンテンツを簡略化するための方法。

【請求項３】
前記隣接ページの取得後に前記隣接ページのＵＲＬの優先順位付けを行うステップをさらに有する請求項１記載の方法。

【請求項４】
前記優先順位付けは、前記目的ページのＵＲＬと前記隣接ページのＵＲＬとの間のエディットディスタンス、または、前記目的ページと前記隣接ページとの間の共起回数または相互参照回数に基づくＵＲＬ間の関連性、の何れかまたは両方に基づいて決定される請求項３記載の方法。

【請求項５】
前記差分演算において、前記オブジェクトが共通するか否かの判断にはＤＰマッチングを用いる請求項１または２記載の方法。

【請求項６】
前記差分演算において、前記目的ページに含まれるオブジェクトの重要度を算出するステップを有し、
前記重要度が所定の閾値を超えている場合には前記オブジェクトが前記隣接ページのオブジェクトに共通する場合であっても前記オブジェクトを削除しない請求項１または２記載の方法。

【請求項７】
前記重要度は、重み付けされた特徴値の和で表し、
前記特徴値は、前記オブジェクトの文字サイズ、フォントその他の文字属性に割当てた数値、前記オブジェクトがバナーであることを識別判断するための数値、前記オブジェクトの画面中央位置からの変位値、前記オブジェクトに含まれるキーワード数、前記オブジェクトが追加または更新されたものであるかを示す情報に割当てた数値、前記オブジェクトの更新文字比率、前記オブジェクトが１文字であるかを示す情報に割当てた数値、前記オブジェクトのタグ種別に割当てた数値、の何れかである請求項６記載の方法。

【請求項８】
前記差分演算の後、前記簡略化ページに含まれる所定の閾値以下の重要度が低いオブジェクト、または、内容が空のテーブル要素またはリスト要素を削除するステップを有する請求項６記載の方法。

【請求項９】
前記差分演算の後、後処理ステップを有し、
前記後処理ステップでは、リストタイトルの修復、または、表頭および表側情報の修復、または、フォームのページ後部への移動、または、アノテーション情報の参照が行われる請求項１または２記載の方法。

【請求項１０】
ユーザ端末からのリクエストを受け取るステップと、
前記リクエストに応答して前記各ステップを実行し、前記簡略化ページのうち、最も小さな情報量を有する簡略化ページを選択するステップと、
前記選択された最小の情報量を有する簡略化ページを前記ユーザ端末に送信するステップと、
を有する請求項１または２記載の方法。

【請求項１１】

前記ユーザ端末は、音声ブラウザが稼動するコンピュータシステムまたは小画面の表示装置を有する情報端末である請求項１０記載の方法。
【請求項１２】
前記ユーザ端末または前記ユーザ端末に接続されたコンピュータシステムは音声認識機能および音声合成機能を備え、
音声による前記リクエストの入力ステップと、
音声による前記簡略化ページの出力ステップと、
を有する請求項１０記載の方法。
【請求項１３】
簡略化対象の目的ページを取得する手段と、
前記目的ページの比較対象となる隣接ページのＵＲＬを生成する手段と、
前記隣接ページを取得する手段と、
前記目的ページおよび隣接ページに含まれる各々のオブジェクトを比較する手段と、
前記オブジェクトの共通性を判断し、共通するオブジェクトを前記目的ページから削除して簡略化ページを生成する手段と、
を含み、
前記隣接ページのＵＲＬを生成する手段には、
前記目的ページのＵＲＬまたは前記目的ページに含まれるリンクのＵＲＬとディレクトリが共通するＵＲＬを生成する手段、または、
前記目的ページのＵＲＬまたは前記目的ページに含まれるリンクのＵＲＬと親ディレクトリが共通するＵＲＬを生成する手段、または、
前記目的ページのＵＲＬを含むルートディレクトリ以下の各ディレクトリを生成する手段、
を含むウェブコンテンツを簡略化するためのシステム。
【請求項１４】
簡略化対象の目的ページを取得する手段と、
前記目的ページの比較対象となる隣接ページのＵＲＬを生成する手段と、
前記隣接ページを取得する手段と、
前記目的ページおよび隣接ページに含まれる各々のオブジェクトを比較する手段と、
前記オブジェクトの共通性を判断し、共通するオブジェクトを前記目的ページから削除して簡略化ページを生成する手段と、
を含み、
前記隣接ページのＵＲＬを生成する手段には、
前記目的ページの過去のページにおいて前記過去ページに含まれていたリンクのＵＲＬを生成する手段、
前記目的ページまたは前記隣接ページの過去のページのＵＲＬを生成する手段、
を含むウェブコンテンツを簡略化するためのシステム。
【請求項１５】
前記隣接ページのＵＲＬの優先順位付けを行う手段をさらに有し、
前記優先順位付けは、前記目的ページのＵＲＬと前記隣接ページのＵＲＬとの間のエディットディスタンス、または、前記目的ページと前記隣接ページとの間の共起回数または相互参照回数に基づくＵＲＬ間の関連性、の何れかまたは両方に基づいて決定される請求項１３記載のシステム。
【請求項１６】
前記オブジェクトの共通性の判断手段においてＤＰマッチングを計算する手段を含む請求項１３または１４記載のシステム。
【請求項１７】
前記目的ページに含まれるオブジェクトの重要度を算出する手段を有し、
前記重要度が所定の第１閾値以上である場合には前記オブジェクトが前記隣接ページのオブジェクトに共通する場合であっても前記オブジェクトを削除しない手段と、

前記重要度が所定の第２閾値以下または前記オブジェクトの内容が空のテーブル要素またはリスト要素である場合には前記オブジェクトを削除する手段と、
を有する請求項１３または１４記載のシステム。
【請求項１８】
後処理手段をさらに有し、
前記後処理手段では、リストタイトルの修復、または、表頭および表側情報の修復、またはフォームのページ後部への移動、または、アノテーション情報の参照が行われる請求項１３または１４記載のシステム。
【請求項１９】
ユーザ端末からのリクエストを受け取る手段と、
前記簡略化ページのうち、最も小さな情報量を有する簡略化ページを選択する手段と、
前記選択された最小の情報量を有する簡略化ページを前記ユーザ端末に送信する手段と、
をさらに有する請求項１３または１４記載のシステム。
【請求項２０】
前記ユーザ端末は、音声ブラウザが稼動するコンピュータシステムまたは小画面の表示装置を有する情報端末である請求項１９記載のシステム。
【請求項２１】
前記ユーザ端末または前記ユーザ端末に接続されたコンピュータシステムは音声認識機能および音声合成機能を備え、
音声による前記リクエストの入力手段と、
音声による前記簡略化ページの出力手段と、
を有する請求項１９記載のシステム。
【請求項２２】
コンピュータ読み取り可能な記録媒体であって、コンピュータに、
簡略化対象の目的ページを取得する機能と、
前記目的ページに隣接する隣接ページを取得する機能と、
前記目的ページと前記隣接ページとで共通するオブジェクトを前記目的ページから削除する差分演算を施す機能と、
簡略化ページを生成する機能と、
を実現させるためのプログラムが記録され、
前記隣接ページは、前記目的ページのＵＲＬまたは前記目的ページに含まれるリンクのＵＲＬとディレクトリが共通するＵＲＬのページ、または、
前記目的ページのＵＲＬまたは前記目的ページに含まれるリンクのＵＲＬと親ディレクトリが共通するＵＲＬのページ、または、
前記目的ページのＵＲＬを含むルートディレクトリ以下の各ディレクトリのトップページ、である
媒体。
【請求項２３】
コンピュータ読み取り可能な記録媒体であって、コンピュータに、
簡略化対象の目的ページを取得する機能と、
前記目的ページに隣接する隣接ページを取得する機能と、
前記目的ページと前記隣接ページとで共通するオブジェクトを前記目的ページから削除する差分演算を施す機能と、
簡略化ページを生成する機能と、
を実現させるためのプログラムが記録され、
前記隣接ページは、前記目的ページの過去のページにおいて前記過去ページに含まれていたリンクのページ、または、
前記目的ページまたは前記隣接ページの過去のページ、である
媒体。

【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、ウェブコンテンツの簡略化方法およびシステムに関する。特に履歴情報を持たないウェブページや日々ＵＲＬ(Uniform Resource Locator)が変化するウェブページであっても、その内容をオンザフライで簡略化できる技術に関する。
【０００２】
【従来の技術】
近年のネットワーク技術の進展と情報機器の機能の向上、低価格化とを反映して、インターネットの利用が盛んになっている。企業、個人を問わず低コストで、また国境を意識することなく詳細な情報発信を行えるため、情報発信源のウェブページは日々爆発的に増大している。また、ウェブページ管理者の管理の下に、日々膨大な情報がアップデートされている。このような背景の下にインターネットおよびこれを用いたウェブページは従来のブロードキャスト、マスメディアに代わる、あるいはこれを補う重要な情報収集媒体になりつつある。
【０００３】
ところでウェブページの役割は多様化しつつある。たとえば単なる情報発信に止まらず、ウェブページを介した商取引（電子商取引）、ウェブページを用いた共同作業（コラボレーション）等が行われつつある。このような多様化された機能を実現するため、より利便性の高いウェブページが提供される。また、目的の情報により迅速にアクセスするために検索画面等のユーザの操作性を向上する機能がウェブページに組み込まれる。たとえば、サイト内で共通に使われるリンクリストや、イメージマップ、フォーム等である。これらはどのページにも含まれており、一般ユーザとって利便性の高い機能を提供する。
【０００４】
しかし、これら一般のウェブページは、デスクトップ型のコンピュータ画面を前提として設計されている。つまりデスクトップ型コンピュータ画面のサイズを考慮してレイアウトされている。このため、ＰＤＡ（Personal Digital Assistants）や携帯電話などの画面の小さなデバイス（以下、小画面デバイスと称する）や、ウェブページ読み上げソフトウェア（以下、音声ブラウザと称する）の場合には、素早く必要な情報にたどりつくことができないという問題がある。つまり、一般的なウェブページの場合、ページの上部にフォームやイメージマップがレイアウトされるため、小画面デバイスではこれらフォーム等の表示を何度も繰り返さなければ必要な情報に到達できない。また、音声ブラウザの場合にはこれらフォーム等の読み上げが行われた後に必要な情報の読み上げが行われる。一般に小画面デバイスの場合にはデスクトップ型コンピュータのようなビジュアルな多機能性は必要なく、また音声ブラウザの場合には操作性を向上させるビジュアルな機能は必要ない。これらビジュアル機能は小画面デバイスあるいは音声ブラウザには逆に邪魔になる。
【０００５】
そこで、ウェブページの一部を省略する簡略化（simplification）の手法が試みられる。たとえば、「http://www9.org/w9cdrom/169/169.html」に記載されているような"Dharma Transcoding"技術、あるいは、「http://www.diffweb.com/」に記載されているような"DiffWeb"（差分）技術がある。
【０００６】
"Dharma Transcoding"技術は、すでに存在するウェブページを元のレイアウトに近い状態でいくつかのページに分割し、小画面デバイスに表示しやすいページを作る技術である。この手法はページの構造や各部位の重要度などを詳細に記述した外部アノテーション情報を必要とする。
【０００７】
"DiffWeb"技術は、あらかじめ登録し、保存されたウェブページと現在のウェブページの差分を計算、提示する技術である。ユーザ毎にページのリストを登録することができ、これらのページの差分を計算することができる。この差分技術では、ページの登録、保存、

差分計算などすべての処理がユーザからの指示で行われる。同様の差分手法として、「http://www-db.stanford.edu/c3/c3.html」に記載された"ＨＴＭＬ Diff"、「http://mindit.netmind.com/mindit.shtml」に記載された"MindIt"などもある。

【０００８】

【発明が解決しようとする課題】

しかし、"Dharma Transcoding"技術では前記の通りアノテーション情報を必要とする。アノテーション情報の付与にはボランティア等の介在が必要であり、完全に自動化することは困難である。

【０００９】

"DiffWeb"技術では、ページの登録、保存、差分計算が前記の通りユーザからの指示で処理される。このため、進行中（On The Fly）の処理として差分計算をすることができない。また、プルダウンメニューなどでは、中身の文字列が削除され簡略化後のフォームが正常に動作しない恐れがある。

【００１０】

さらに、従来技術ではあらかじめ保存しておいた比較対象ページとの差分を計算することにより、シンプリフィケーションが実現されている。したがって、以下のような問題がある。

【００１１】

第１に比較対象ページがあらかじめ保存されていない場合には簡略化できないという問題がある。つまり比較ページを記録しているページのみが簡略化対象となり、初出ページの簡略化を実行できない問題がある。

【００１２】

第２に、比較対象のページが保存されていた場合でも、ＵＲＬが日々変化するページでは、簡略化できないという問題がある。たとえば朝日新聞(www.asahi.com)の記事ページは次のようにＵＲＬの中に日付が含まれている。

「http://www.asahi.com/0530/news/business30010.html」

この場合、同一ＵＲＬの過去のページは存在せず、ページ簡略化を行うことができない。

【００１３】

第３に必要な情報までも削除されてしまう問題がある。例えば、リンクリストのタイトルや、フォームといった重要な情報が削除されてしまう問題である。逆に不必要な微妙な文字列の変化などが保存されてしまう問題がある。

【００１４】

本発明の目的は、小画面デバイスや音声ブラウザを用いたウェブページの表示あるいは出力の際に、必要な情報に迅速にアクセスするためのウェブページ簡略化の手法を提供することにある。

【００１５】

また、本発明の目的は、ウェブページの簡略化を同一ＵＲＬの過去のページが存在しない場合においても実行できる手法を提供することにある。

【００１６】

また、本発明の目的は、ウェブページの簡略化をオンザフライで実行できる手法を提供することにある。

【００１７】

また、本発明の目的は、ウェブページの簡略化の際に重要な情報を欠落することなく、不必要な情報を精度良く簡略化する手法を提供することにある。

【００１８】

【課題を解決するための手段】

本願の発明の概略を説明すれば、以下の通りである。本発明は、差分計算に基づいて音声ブラウザや小画面デバイスで読みやすいページにオンザフライで変換する手法を提供する。本発明のページ簡略化で用いる差分計算では、簡略化対象のページと比較する比較ページとして、同一ＵＲＬの過去ページのみならず、隣接ページを用いる。比較により更新さ

れた情報のみを取り出し、サイト内の各ページで共通に使用されているテンプレート情報を取り除く。これにより簡略化対象ページの主なコンテンツのみを取り出す。

【００１９】

本発明により、比較のための近隣ページを自動的に取得でき、過去ページの蓄積情報がなくともオンザフライでウェブページの簡略化が実行できる。これにより小画面デバイスあるいは音声ブラウザを用いた場合の必要な情報への迅速なアクセスが可能になる。

【００２０】

具体的には本発明の方法は、簡略化対象の目的ページを取得するステップと、目的ページに隣接する隣接ページを取得するステップと、目的ページと隣接ページとで共通するオブジェクトを目的ページから削除する差分演算を施し、簡略化ページを生成するステップと、含む。

【００２１】

前記隣接ページには、目的ページのＵＲＬまたは目的ページに含まれるリンクのＵＲＬとディレクトリまたは親ディレクトリが共通するＵＲＬのページを含み、目的ページのルートディレクトリ以下の各ディレクトリのトップページ、または、目的ページの過去のページ、過去ページに含まれていたリンクのページ、さらに隣接ページの過去のページを含む。

【００２２】

前記隣接ページの取得後に前記隣接ページのＵＲＬの優先順位付けを行うことができ、優先順位付けは、目的ページのＵＲＬと隣接ページのＵＲＬとの間のエディットディスタンス、または、目的ページと隣接ページとの間の共起回数または相互参照回数に基づくＵＲＬ間の関連性に基づいて決定できる。

【００２３】

前記差分演算において、オブジェクトが共通するか否かの判断にはＤＰマッチングを用いることができ、また、目的ページに含まれるオブジェクトの重要度を算出できる。重要度が所定の閾値を超えている場合にはオブジェクトが隣接ページのオブジェクトに共通する場合であってもオブジェクトを削除しない。逆に重要度の低いオブジェクトは削除できる。

【００２４】

前記重要度は、重み付けされた特徴値の和で表し、特徴値には、オブジェクトの文字サイズ、フォントその他の文字属性に割当てた数値、オブジェクトがバナーであることを識別判断するための数値、オブジェクトの画面中央位置からの変位値、オブジェクトに含まれるキーワード数、オブジェクトが追加または更新されたものであるかを示す情報に割当てた数値、オブジェクトの更新文字比率、オブジェクトが１文字であるかを示す情報に割当てた数値、オブジェクトのタグ種別に割当てた数値、等を採用できる。

【００２５】

また、差分演算の後、後処理を行うことができ、後処理では、リストタイトルの修復、または、表頭および表側情報の修復、または、フォームのページ後部への移動、または、アノテーション情報の参照が行える。

【００２６】

また、ユーザ端末からのリクエストを受け取り、リクエストに応答して前記各ステップを実行し、簡略化ページのうち、最も小さな情報量を有する簡略化ページを選択してユーザ端末に送信することができる。ユーザ端末は、音声ブラウザが稼動するコンピュータシステムまたは小画面の表示装置を有する情報端末とすることができる。あるいはユーザ端末またはユーザ端末に接続されたコンピュータシステムに音声認識機能および音声合成機能を備え、音声によるリクエストの入力、音声による簡略化ページの出力を行うことができる。

【００２７】

【発明の実施の形態】

以下、本発明の実施の形態を図面に基づいて詳細に説明する。ただし、本発明は多くの異

なる態様で実施することが可能であり、本実施の形態の記載内容に限定して解釈すべきではない。なお、実施の形態の全体を通して同じ要素には同じ番号を付するものとする。

【００２８】

以下の実施の形態では、主に方法またはシステムについて説明するが、当業者であれば明らかなとおり、本発明は方法、システムの他、コンピュータで使用可能なプログラムが記録された媒体としても実施できる。したがって、本発明は、ハードウェアとしての実施形態、ソフトウェアとしての実施形態またはソフトウェアとハードウェアとの組合せの実施形態をとることができる。プログラムが記録された媒体としては、ハードディスク、ＣＤ－ＲＯＭ、光記憶装置または磁気記憶装置を含む任意のコンピュータ可読媒体を例示できる。

【００２９】

また以下の実施の形態では、一般的なコンピュータシステムを用いることができる。実施の形態で用いるコンピュータシステムには、中央演算処理装置（ＣＰＵ）、主記憶装置（メインメモリ：ＲＡＭ）、不揮発性記憶装置（ＲＯＭ）等を有し、バスで相互に接続される。バスには、その他コプロセッサ、画像アクセラレータ、キャッシュメモリ、入出力制御装置（Ｉ／Ｏ）等が接続されてもよい。バスには、適当なインターフェイスを介して外部記憶装置、データ入力デバイス、表示デバイス、通信制御装置等が接続される。その他、一般的にコンピュータシステムに備えられるハードウェア資源を備えることが可能なことは言うまでもない。外部記憶装置は代表的にはハードディスク装置が例示できるが、これに限られず、光磁気記憶装置、光記憶装置、フラッシュメモリ等半導体記憶装置も含まれる。データ入力デバイスには、キーボード等の入力装置、マウス等ポインティングデバイスを備えることができる。データ入力デバイスにはスキャナ等の画像読み取り装置、音声入力装置も含む。表示装置としては、ＣＲＴ、液晶表示装置、プラズマ表示装置が例示できる。また、コンピュータシステムには、パーソナルコンピュータ、ワークステーション、メインフレームコンピュータ等各種のコンピュータが含まれる。

【００３０】

（実施の形態１）

図１は、本発明の実施の形態１のシステムを示した構成図である。本実施の形態のシステムは、ユーザ端末１、プロキシサーバ２、目的のウェブページが存在するウェブサーバ３を含む。

【００３１】

ユーザ端末１は音声ブラウザが稼動するコンピュータシステムまたは小画面の表示デバイスを有する小画面デバイスである。音声ブラウザはＨＴＭＬ（Hypertext Markup Language）文書またはＸＭＬ（eXtensible Markup Language）文書等ＨＴＴＰ（HyperText Transfer Protocol）を用いて取得される情報（文書）を音声によりブラウジングする機能を有するソフトウェアである。また、小画面デバイスは、たとえばｉモード携帯電話やＰＤＡ等デスクトップ型コンピュータの表示装置よりも格段に小さい表示画面を有する情報端末である。これら音声ブラウザを用いて取得した文書を読み上げる際、あるいは小画面デバイスに表示する際には、デスクトップ型コンピュータ用にレイアウトされたウェブコンテンツをそのまま適用したのでは読み上げあるいは表示の障害になる。本実施の形態のシステムはこのようなユーザ端末１においてもスムーズに取得情報の読み上げあるいは表示を行うことを目的としている。特に音声ブラウザのユーザとなるであろう視覚障害者にとっては、視覚障害をもたない者と同様な情報取得手段を得ることになる。本発明は情報アクセスへの平等な機会を確保できる手段を提供でき、社会的な意義も大きい。

【００３２】

プロキシサーバ２は、一般的なプロキシサーバの機能に加えて、ウェブページを簡略化する差分処理４の機能を有する。差分処理４の機能については後に説明する。差分処理４においてはキャッシュデータベース５を参照する。キャッシュデータベース５には、過去においてプロキシサーバ２がアクセスしたウェブページが格納される。なお、本実施の形態ではプロキシサーバ２を用いて簡略化されたウェブページを表示または音声出力する例を

説明するが、簡略化機能（差分処理４の機能）を各クライアント（ユーザ端末１）あるいはウェブサーバ３に有しても良い。また、キャッシュデータベース５はプロキシサーバ２のシステム内に存在するように図示しているが、キャッシュデータベース５は必ずしもプロキシサーバ２内に存在する必要はない。たとえばＵＲＬ、ＩＰアドレス等でその所在が特定でき、サーバの稼動状態においてアクセスできる限り他のシステムに存在しても良い。他のシステムにはＬＡＮ又はＷＡＮ接続される他のシステム、インターネットを介して接続される他のシステムが含まれる。また、キャッシュデータベース５は概念的なものであり、物理的には異なる場所（アドレス）に存在する記憶装置によって分散的に処理されても良い。以後説明する他のデータベースについて同様である。

【００３３】

ウェブサーバ３には目的のウェブページ６を有する。一般にウェブサーバ３はインターネットに接続されていることが想定されるが、必ずしもインターネットに接続されている必要はなく、ＨＴＴＰリクエストに対するレスポンスを行うサーバであれば良い。たとえば企業等団体内のイントラネット、エクストラネット等各種ネット上のサーバでも良い。

【００３４】

また、ウェブサーバ３には目的ページ６に隣接する隣接ページ７が存在する。隣接ページ７については後に詳述する。通常隣接ページ７には目的ページ６と共通なフォーム、リンクリスト、イメージマップ等のオブジェクトを含み、これら共通のオブジェクトを目的ページから削除することによりウェブページの簡略化を行う。なお、本実施の形態において隣接ページ７とともに比較対照ページの候補としてキャッシュデータベース５に記録されている過去ページも用いるが、過去ページについては本発明に必須の構成要件ではない。本発明では目的ページ６の過去ページがなくとも現在の隣接ページ７を用いて簡略化を実行できる。

【００３５】

本実施の形態のシステムでは、ユーザはユーザ端末１のブラウザのプロキシサーバの設定においてプロキシサーバ２のアドレスを指定する。そしてウェブサーバ３にアクセスすることにより簡略化されたウェブページ８を表示あるいは出力できる。

【００３６】

本システムの処理の概要は以下の通りである。ユーザ端末１は、ウェブサーバ３のアドレスを指定してＨＴＴＰリクエスト９を出す。ユーザ端末１からのリクエスト９を受けてプロキシサーバ２がウェブサーバ３にＨＴＴＰリクエスト１０を出す。プロキシサーバ２は目的ページ６、隣接ページ７およびキャッシュデータベース５内の必要に応じた過去ページの参照により差分処理４を行い、簡略化ページ８をリクエスト９の応答として返送する。ユーザ端末１は簡略化ページ８を表示し、あるいは音声出力を行う。このような簡略化方法により、目的ページ６の共通のオブジェクトを削除し、ユーザは迅速に必要な情報を得ることができる。簡略化されたページを読み上げるためユーザはよりスムーズに、ストレスを感じることなく必要な（重要な）情報を聴くことができる。特に視覚障害者には視覚的にウェブページを把握することができないので、本発明を用いたウェブページの簡略化により、スムーズに情報を取得できる。以下、差分処理の詳細を説明する。

【００３７】

図２は差分処理４の詳細を示したブロック図である。本実施の形態の差分処理モジュールには、隣接ＵＲＬ列挙１１、ディレクトリ列挙１２、ＵＲＬキャッシュ１３、更新前目的ページ/隣接ページ取得１４の各モジュールと、ウェブページを取得しＤＯＭ（Document Object Model）によるオブジェクトを出力するフェッチモジュール、ＵＲＬ優先順位演算モジュール１６、差分演算モジュール１７、最小差分選択モジュール１８を含む。本実施の形態の差分処理は、目的のページに対して、差分計算の比較対象として適切なページを自動的に検出し、差分を算出する。本実施の形態では、隣接ページという概念を導入することで、過去のページが保存されていない場合、あるいは日々ＵＲＬが変化する場合でも差分の抽出を可能とする。同時に、サイト内で共通に使用されるテンプレート（ヘッダー、フォーム、リンクリストなど）を任意のページに対して取り除くことができるようにな

る。隣接ページとは、サーバ上で同じディレクトリに存在するページ、親ディレクトリが同じページ、ディレクトリのトップページ、サイトのトップページなど目的ページに含まれるオブジェクトと共通なオブジェクトを含む可能性のあるページをいう。本実施の形態では、これらの隣接ページとの差分も計算し、その中でもっとも文字列の短くなる差分を選択することで、オンザフライ（進行中）でのウェブページ簡略化を実現する。

【００３８】

隣接ＵＲＬ列挙モジュール

図３は、隣接ＵＲＬ列挙モジュール１１の詳細を示したブロック図である。隣接ＵＲＬ列挙モジュール１１は、サイト内の近い場所にあるページ（隣接ページ７）のリストを、目的ページ６を解析して得るモジュールである。これらのページ（隣接ページ７）には目的ページ６と同じリンクリスト、イメージマップを持っていることが多いため、これら隣接ページの共通オブジェクトを削除して目的ページ６の簡略化を実施できる。

【００３９】

隣接ＵＲＬ列挙モジュール１１には、リンク列挙モジュール１９、ＵＲＬリスト２０、ディレクトリ共通ＵＲＬ選択モジュール２１、親ディレクトリ共通ＵＲＬ選択モジュール２２、リストマージモジュール２３を含む。

【００４０】

まず、プロキシサーバ２からのリクエストによって、目的ページ６からコンテンツをフェッチする（図２のフェッチモジュール１５）。フェッチによって取得された目的ページのオブジェクト（ＤＯＭ）から、目的ページ６内に含まれるＵＲＬを列挙する（リンク列挙モジュール１９）。ＵＲＬの列挙には、たとえば<a>タグのhrefアトリビュートを参照する。そしてリンク列挙モジュール１９の出力としてＵＲＬリスト２０を得る。ＵＲＬリスト２０の中から、目的ページ６のＵＲＬと同一のディレクトリをさしているＵＲＬを選択し、新たなリストを作成する(ディレクトリ共通ＵＲＬ選択モジュール２１)。さらに、親ディレクトリが共通なＵＲＬを選択してＵＲＬのリストを作成する(親ディレクトリ共通ＵＲＬ選択モジュール２２)。最後にこれらのリストをマージモジュール２３によりマージして、隣接ＵＲＬリスト２４を得る。後に説明するようにそれぞれのＵＲＬに対してフェッチモジュール１５を起動して差分演算に用いる。

【００４１】

２１、２２のいずれのモジュールによるリストがより有効かは各サイトの構造により異なる。例えば、"www.asahi.com"では、ディレクトリ共通ＵＲＬ選択モジュール２１が有効であり、"www.cnn.com"では親ディレクトリ共通ＵＲＬ選択モジュール２２が有効である。このようにサイトの構造によって有効な方法が変化するため、両方のリストをリストアップして、より有効なページ簡略化の比較ページを得ることができる。

【００４２】

以下にＵＲＬの選択例を示す。

（１）ディレクトリ共通ＵＲＬ選択例：

目的ページのＵＲＬ:

http://www.asahi.com/0606/news/national06015.html

リストアップされるＵＲＬ（一部）：

http://www.asahi.com/0606/news/national06012.html

http://www.asahi.com/0606/news/national06013.html

http://www.asahi.com/0606/news/national06014.html

（２）親ディレクトリ共通ＵＲＬ選択例：

目的ページのＵＲＬ:

http://www.cnn.com/2000/US/06/05/sea.based.defense/index.html

選択されるＵＲＬ（一部）：

http://www.cnn.com/2000/US/06/05/dday.remembrance/index.html

http://www.cnn.com/2000/US/06/05/helicopter.escape.03/index.html

http://www.cnn.com/2000/US/06/05/curbing.terrorism.02/index.html

【００４３】
ディレクトリ列挙モジュール
ディレクトリ列挙モジュール１２は、目的ページ６のＵＲＬから各ディレクトリのトップページのリストを作成する。大規模なサイトにおいて、いくつかのトップページに分かれることがあり、目的ページと共通のリンクリストを含む場合の比較対照ページの取得に有効である。以下に具体例を例示する。
目的ページのＵＲＬ:
http://www.cnn.com/2000/US/06/05/helicopter.escape.03/index.html
選択されるＵＲＬ：
http://www.cnn.com/2000/US/06/05/
http://www.cnn.com/2000/US/06/
http://www.cnn.com/2000/US/
http://www.cnn.com/2000/
http://www.cnn.com/
【００４４】
ＵＲＬキャッシュモジュール
図４はＵＲＬキャッシュモジュール１３の一例を示した構成図である。ＵＲＬキャッシュモジュール１３は、ＵＲＬキャッシュデータベース２５、検索手段２６を含む。ＵＲＬキャッシュデータベース２５は図１のキャッシュデータベース５の一部として構成されても良い。ＵＲＬキャッシュデータベース２５には、過去にプロキシサーバ２のユーザがアクセスしたページのＵＲＬとそのページ内に含まれているＵＲＬのリストが記録されている。検索手段２６は目的ページ６のＵＲＬに応じてキャッシュ内のＵＲＬを検索し、目的ページ６と同一サイトのＵＲＬリスト２７をリストアップする。
ＵＲＬキャッシュモジュール１３によりＵＲＬキャッシュから目的ページのＵＲＬに近いＵＲＬを取得することができ、適切なページの候補を増やすことが可能になる。
【００４５】
ＵＲＬ優先順位演算モジュール
ＵＲＬ優先順位演算モジュール１６は、前記した隣接ＵＲＬ列挙モジュール１１、ディレクトリ列挙モジュール１２、ＵＲＬキャッシュモジュール１３によってリストアップされたＵＲＬの類似度を算出し、ＵＲＬの優先順位付けを行うモジュールである。優先順位の順に後に説明するページのフェッチが行われる。ＵＲＬ優先順位演算モジュール１６は、より確実にレイアウトの似たページをリストアップするために、ＵＲＬのエディットディスタンス（Edit distance）、共起回数、相互参照回数を総合して決定する。
図５は、ＵＲＬ優先順位演算モジュール１６の一例を示したブロック図である。ＵＲＬ優先順位演算モジュール１６には、同一ページ除外モジュール２８、ＵＲＬエディットディスタンス算出モジュール２９、ＵＲＬ関連性算出モジュール３０、ソートモジュール３１、リダイレクションＵＲＬテーブル３２、ＵＲＬ共起回数テーブル３３、ＵＲＬ相互参照テーブル３４を含む。
【００４６】
次に、ＵＲＬ優先順位演算モジュール１６の動作を説明する。まず、隣接ＵＲＬ列挙モジュール１１、ディレクトリ列挙モジュール１２、ＵＲＬキャッシュモジュール１３でリストアップされたＵＲＬを統合し、一つのＵＲＬリストを作る。次にリダイレクションＵＲＬテーブル３２を用いて、同一ページを指し示しているＵＲＬを同一ページ除外モジュール２８で除外する。リダイレクションＵＲＬテーブル３２は、後に説明するように各フェッチモジュール１５によってメンテナンスされる。
【００４７】
次に、目的ページ６のＵＲＬとリスト中の各ＵＲＬとの類似度を算出する。類似度はＵＲＬエディットディスタンス算出モジュール２９を用いてエディットディスタンス（edit distance）を算出することにより算出できる。例えば、ＤＰマッチング（DP Matching）を用いて最長共通文字列（Longest common string）を算出し、そこから編集操作回数を算

出してエディットディスタンスを求める。エディットディスタンスが短いほど類似度は高い。

【００４８】

次に、ＵＲＬ関連性算出モジュール３０を用いてＵＲＬの関連性を算出する。ＵＲＬ関連性算出モジュール３０は、エディットディスタンスの値と、ＵＲＬの相互参照回数、ＵＲＬの共起回数とを総合して、目的ページのＵＲＬとの関連性を数値化する。ＵＲＬの共起回数および相互参照回数は、各々ＵＲＬ共起回数テーブル３３、ＵＲＬ相互参照回数テーブル３４に記録され、フェッチモジュール１５内のＨＴＭＬパーザ（後に説明する）において常に算出、更新される。ＵＲＬ関連性算出モジュール３０は、ＵＲＬの相互参照回数、共起回数、エディットディスタンスを重みつき加算し、各ＵＲＬの関連性を算出する。最後に、ソートモジュール３１において降順でリストを並び替え、優先順位を決定する。決定された優先順位は優先度順ＵＲＬリスト３５として出力される。

【００４９】

更新前目的ページ／隣接ページ取得モジュール

更新前目的ページ／隣接ページ取得モジュール１４は、目的ページ６またはそれに隣接するページ７の更新前ページがキャッシュデータベース５に存在する場合にはこれらをも差分（比較）対象として選択するためのモジュールである。

図６は、更新前目的ページ／隣接ページ取得モジュールの一例を示した構成図である。

更新前目的ページ／隣接ページ取得モジュール１４には、検索キー作成モジュール３６、検索モジュール３７、ＨＴＭＬパーザ３９を含み、前記キャッシュデータベース５が参照される。

【００５０】

ＵＲＬ優先順位演算モジュール１６によって、類似度順にソートされたＵＲＬリスト３５の過去のページがキャッシュデータベース５内に存在するかを検索モジュール３７を用いて検索する。存在していた場合、キャッシュデータベース５から検索条件を満足するウェブページのリスト３８を抽出し、ＨＴＭＬパーザ３９を用いてＨＴＭＬ文書からウェブページリスト４０（ＤＯＭツリーからなる）を生成する。 キャッシュデータベース５にはＵＲＬリストだけでなく、ウェブページのコンテンツも記録されるので、更新前目的ページ／隣接ページリスト４１はＤＯＭツリーの状態で得られる。得られた更新前目的ページ／隣接ページリスト４１は目的ページ６の差分対象として選択する。

【００５１】

キャッシュデータベース５に対する検索キーは検索キー作成モジュール３６で生成される。検索キーには、目的ページ６のＵＲＬに加えて、前述のディレクトリ列挙モジュール１２によって生成されるＵＲＬ列を用いる。

また、目的ページ６が検索エンジン（ホームページ検索を主な機能とするウェブページ）の検索結果など「Query文字列」を含む場合、他のキーワードによる検索結果（ＵＲＬ）も検索キーに加えることができる。

【００５２】

フェッチモジュール

図７は、フェッチモジュールの一例を示す構成図である。フェッチモジュール１５には、ダウンロードモジュール４２およびＨＴＭＬパーザ４３を含む。フェッチすべきＵＲＬ４４の入力を受けて、ダウンロードモジュール４２は入力されたＵＲＬのウェブサーバ３にＨＴＴＰリクエストを発行する。ウェブサーバ３はリクエストを受けてＨＴＭＬファイル４５を返送する。ダウンロードモジュール４２は受け取ったＨＴＭＬファイルのＵＲＬをリダイレクションＵＲＬテーブル３２に記録する。一方、ＨＴＭＬファイル４５はＨＴＭＬパーザ４３に送付され、ＤＯＭツリー４６に変換される。ＨＴＭＬパーザ４３は、ＵＲＬ毎の共起回数、相互参照回数を算出し、ＵＲＬ共起回数テーブル３３およびＵＲＬ相互参照テーブル３４に記録する。これらリダイレクションＵＲＬテーブル３２、ＵＲＬ共起回数テーブル３３およびＵＲＬ相互参照テーブル３４はフェッチモジュール１５によってメンテナンスされ、前記ＵＲＬ優先順位演算モジュール１６で利用される。

【００５３】
差分演算モジュール
差分演算モジュール１７は、目的ページ６のＤＯＭツリーと前記手法で選択された比較対象ページのＤＯＭツリーとから差分のＤＯＭツリーを生成する。
図８は差分演算モジュールの一例を示したブロック図である。差分演算モジュール１７には、線形化モジュール４７、ＤＰマッチングモジュール５０、重要度算出モジュール５２、共通ノード削除モジュール５３を含む。ＵＲＬ優先順位演算モジュール１６によって並び替えられた各ＵＲＬは、前記したフェッチモジュール１５によってＤＯＭツリーに変換される。一方、更新前目的ページ／隣接ページ取得モジュール１４によって取得されたリスト４１はＤＯＭツリー状態で得られるためフェッチの必要はない。
【００５４】
これら隣接ＵＲＬ列挙（１１）、ディレクトリ列挙（１２）、ＵＲＬキャッシュ（１３）、更新前目的ページ／隣接ページ取得（１４）の各モジュールで選択されたリストに対応するページのＤＯＭツリーを前記の通り生成し、これを差分演算の一方の入力として目的ページ６の比較対象とする。目的ページ６のＤＯＭツリーは差分演算の他の入力となる。
【００５５】
本実施の形態の差分過程で重要なノードが削除されるのを防ぐために、あらかじめノードの重要度を重要度算出モジュール５２によって算出し、ある閾値を越えた重要度を持つノードは共通ノードであっても削除しない。また、次のクリーンアップモジュール５４で重要度の低いノードが取り除かれることでそのページに独自でかつ、重要度の高い情報のみを取得できる。
【００５６】
以下、ＤＰマッチングを用いた手法を図８に従って説明する。まず、線形化モジュール４７が、目的ページと比較対象ページそれぞれのＤＯＭツリーからノードリスト４８，４９を作成する。線形化モジュール４７はＤＯＭツリーを巡回して、テキストノード、イメージ（画像）ノードを選択する。このとき、ページ上のフォームが削除されないように、フォームノードの下のノードを選択せず、フォームの処理を後に他の実施の形態として説明する「後処理」モジュールにおいて行うようにすることもできる。また、クライアントサイドスクリプト(JavaScript, VBScript)を保存するためにコメントノードも選択しないにようにすることができる。
【００５７】
次に、２つのノードリストに対してＤＰマッチング５０を行い、共通に含まれるノードのリストを算出する。ＤＰマッチングは２つの記号列から、共通するもっとも長い記号列（Longest Common String）を算出するためのアルゴリズムである。例えば、"abcdefgh"と"bcdlgh"であれば"bcdgh"が出力される。本実施の形態ではこのアルゴリズムをノードのリストに対して適応することで共通に含まれるノードのリスト(いわばLongest Common "Node" String)を得る。
【００５８】
重要度算出モジュール
差分演算において、重要ノード（タイトルを示す文字列など）が削除されるのを防ぐため、あらかじめ目的のページの各テキストノードおよびイメージノードに対し、重み付けを行う。共通ノード削除モジュール５３では、重みが閾値を越えていた場合、削除しない。ノードの重要度を算出する手法を以下に例示する。なお以下に例示する以外にその他の重要度算出手法を適用できることは勿論である。ここではいくつかの特徴値の重みつき総和により重要度を決定する手法を説明する。各ノードの重要度ｓは次の式により算出される。
$s = \Sigma Wi*Pi$
ただし、Pi = 各特徴値
Wi = 各特徴値に対する重み付け
【００５９】

以下に特徴値の例を述べる。
[文字サイズ]　レンダリングされたときの文字サイズ（size）とデフォルトフォントサイズ（default size）の差を特徴値Piとする。
Pi = size - default size
大きな文字サイズを有するノードほど重要度が高いという経験則に基づく。文字属性を特徴値Piに加味することもできる。その場合、各属性値に応じてPiに加算される。たとえばボールド、イタリック等のフォント、赤等の色が指定されている場合、下線、二重下線等の属性が指定されている場合には一般に重要度が高いと判断でき、このような属性に応じてPiに加算する。
【００６０】
[テンプレートによるバナーの除去]　バナーである可能性が高いイメージリンクの場合には重要度を下げる。バナーテンプレートは、画像サイズ、リンク先文字列(/doubleclick/, /ads/など)、直後のリンク文字列(Click Here)などを判断基準とすることができる。テンプレートからの距離を特徴値Piとすることができる。
【００６１】
[ノードのポジション]　レンダリングされたときにノードが表示されるポジションにより、重み付けを行う。図９に示すように中心部のものほど位置重要度を上げ、周辺のものほど重要度を下げる。図９ではウィンドウ内の色の濃い領域ほど重要度が高いことを示している。特徴値Piは、各ノードの各ピクセル値の位置重要度の総和で算出できる。
【００６２】
[キーワード検出による重要度上げ]　目的ページのキーワード解析を行い、キーワードを含むノードの重要度を上げることができる。システムが持っている重要キーワードと、ページ内を解析して決定されたキーワードがノード内に含まれている個数をそのノードの特徴値とすることができる。
【００６３】
[追加されたノードと更新されたノード]　追加されたノード（比較対象のページには含まれないノード）の重要度を上げるため、追加されたノードに対して特徴値を1、それ以外は0とすることができる。Wiは正の値とする。
【００６４】
[更新されたノードの場合更新文字列比率]　追加ではなく、更新されたノードの場合、ノード内の文字数に対する更新文字数の割合を特徴値とすることができる。Wiは正の値とする。
【００６５】
[1文字の時は下げる]　一字のみのノードの重要度を下げるため、特徴値1を割りあてることができる。この時Wiは負の値とする。
【００６６】
[タグ種別]　ノードによっては、タグの種類によって明確に重要度が判別できるものがある。そのようなタグに特徴値を割り当てる。デフォルトは0とする。たとえば、フォームノードを保存するためにFORMノードに正の特徴値を割り当てることができる。
【００６７】
共通ノード削除モジュール
共通ノード削除モジュール５３は、共通ノードリスト５１に含まれるノードを目的ページ６のＤＯＭツリーから削除することで差分のＤＯＭツリーを生成する。ただし、重要度算出モジュール５２で重要度が高いと判断されたノードは削除しない。保存するノードは、一定の閾値によって決定される。閾値はシステムの持つデフォルト値以外に、ユーザによって指定される閾値も利用可能である。出力は差分ＤＯＭツリーとなる。
【００６８】
クリーンアップモジュール
クリーンアップモジュール５４は、差分処理の最後で、重要度の低いノードと空白ノードを削除する。まず、重要度算出モジュール５２においてきわめて低い重要度であると判断

されたノードを削除する。削除されるノードはある閾値によって判別される。この閾値はシステムの持つデフォルト値以外に、ユーザによって指定される閾値も利用可能である。次に、空白のテーブルセル<TD>、リストアイテム<LI>などを削除する。テーブルセルは、列または行のすべてが空白の場合にのみ削除を行う。

【００６９】

最小差分選択モジュール

最小差分選択モジュール１８は、各比較ページに対応する差分演算モジュール１７の出力から最小サイズの差分を選択する。最も効率的に簡略化された差分ページをプロキシサーバ２の出力（簡略化ページ８）としてユーザ端末１に返送し、ユーザ端末１は簡略化ページ８をブラウジングする。

【００７０】

本実施の形態のシステムおよび簡略化方法によれば、過去ページが存在しない場合であっても、比較対象のページを得ることができ、目的ページの簡略化を行うことができる。また、各種の隣接ページ（比較対象ページ）を網羅的に取得するため、より適切な精度の高い簡略化を行える。また、差分処理においてノードの重要度をチェックするため必要な情報を欠落させる確率が低くできる。また、クリーンアップモジュールを用いるため無用なノードを削除でき、より重要な情報のみを残して簡略化の精度を向上することができる。

【００７１】

実際のウェブページに本実施の形態のシステムおよび簡略化方法を適用した例を示す。図１０は通常のブラウザによりニュースページを表示した表示図である。つまり図１０は本実施の形態の処理を施す前の表示に対応する。図１１は、本実施の形態のシステムを用いて図１０のページを表示した表示図である。ページ上部および左側に存在していたリンクリストが削除され、ページ固有の情報であるニュース本文が残っていることがわかる。音声ブラウザ等でブラウジングした時により速やかにニュース本文に到達できることがわかる。

【００７２】

図１２は検索画面の一例を通常のブラウザにより表示した例を示す表示図である。検索結果が画面中央部に表示されている。図１３は、本実施の形態のシステムを用いて図１２のページを表示した表示図である。前記図１１の場合と同様にリンクリストが削除され、検索結果が残されている。なお、画面左に位置する検索機能をもつフォームはその位置にリンクだけが残され、フォーム自体はページの後部に移動される。これにより音声ブラウザによる読み上げが検索結果に素早く到達できる。フォームの移動については後に説明する。

【００７３】

前記した本実施の形態のシステムおよび簡略化方法により、ページの文字数、リンク数、ページ内エレメント数ともに半分程度に削減できることが本発明者らの検討により判明した。表１は、CNN、朝日新聞、SUNTIMESの任意のページにおいて、本実施の形態のシステムおよび方法を適用した結果を示す表である。ばらつきがあるのもの、おおむねオリジナルページに対して40% から60% までに情報を削減できていることがわかる。

【００７４】

【表１】

Site	Number of Characters			Number of Links			Number of Elements		
	original	transcoded		original	transcoded		original	transcoded	
CNN	4294	2557	60%	167	75	45%	228	116	51%
Suntimes	3446	2770	80%	59	17	29%	93	41	44%
Asahi	1880	1020	54%	40	4	10%	65	13	20%

また、表２は各種検索ページの開始から検索結果を表示するまでの情報量の比較を示した表である。情報が大幅に削減されており、音声ブラウザなどでよりすばやく検索結果に到達できるようになっている事がわかる。

【００７５】
【表２】

Page	Original	Transcoded
Yahoo	14 links 1 image map	0 link
Lycos	15 links 1 form	7 links
Infoseek	16 links and 1 form	2 links

（実施の形態２）
図１４は、本発明の実施の形態２のシステムを示した構成図である。本実施の形態のシステムは、実施の形態１のシステムを構成する各構成要件に加えてプロキシサーバ６０に後処理モジュール６１、ＤＯＭからＨＴＭＬへの変換モジュール６２、ユーザプロファイル６３を有する。本発明の実施の形態のシステムは、特に視覚障害者のユーザに適する。視覚障害者のユーザがアクセスすることによって自動的にページを読みやすく変換して出力するプロキシサーバ６０を提示できる。このようなプロキシサーバ６０は、小画面デバイスや、電話によるWEBアクセスのサーバへの応用も勿論可能である。
【００７６】
図１４に示すように、ユーザはユーザ端末１のブラウザのプロキシサーバとして本システムのプロキシサーバ６０のアドレスを指定する。ユーザプロファイル６３により、ユーザの要求に合わせた簡略化を行える。たとえば、ユーザがあるページに対して、より絞り込んだ情報のみを要求した場合、本システムは共通ノード削除モジュール５３（図８）およびクリーンアップモジュール５４（図８）における、閾値の値を上げることで情報をより重要度の高いものだけに絞り込んで提示することができる。ユーザによる閾値コントロールは、各ページの最下部に「情報量減らす」と「情報量増やす」という二つのリンクを追加し、これらのリンクが選択されたときに閾値を変更するという方法で実現できる。同様に他のパラメータをユーザがコントロールすることも考えられる。このようなカスタマイゼーション機能は、本発明に容易に統合することが可能である。
【００７７】
また、差分ＤＯＭツリー５５に対していくつかのヒューリスティクスを用いて一部の情報をさらに修復することでさらに精度を上げることが可能である。後処理モジュール６１では、このようなヒューリスティクスに基づいた処理を行うことができる。
【００７８】
図１５は後処理モジュール６１の一例を示したブロック図である。ここでは、タグ構造の分析に基づいた差分結果（差分処理の出力）の自動修正の例を示す。後処理モジュール６１には、リストタイトル修復モジュール６４、表頭・表側修復モジュール６５、フォーム移動モジュール６６を有する。
【００７９】
リストタイトル修復モジュール６４では、リスト(ＨＴＭＬのオーダードリスト<OL>とアンオーダードリスト<UL>)のタイトルが削除されてしまっていた場合にこれを目的ページのＤＯＭツリーを参照して修復する。図１６はリストタイトル修復モジュール６４による修復の例を示す表示図である。オリジナル(同図（ａ）)の「同ジャンルの他のニュース」というリンクリストのタイトルを示す文字列が差分ページ(同図（ｂ）)では削除されている。このようなリンクリストのタイトルは各ページに共通に含まれる場合が多いため、このような現象が生じる可能性がある。リンクリストのタイトルはそのリストの意味を的確に表しており、残されるべきである。そこで、リストタイトル修復モジュール６４では、次のような条件でタイトルを識別、修復する。
1)リスト内の項目が一つでも残っている。
2)リストの直前の文字列がヘッダー、ボールド、拡大フォントのいずれかでかつ、50文字

以内である。
このような場合、直前の文字列をタイトルと判定し、同図（ｃ）に示すように、修復する。
【００８０】
表頭・表側修復モジュール６５においては、同様にテーブル内のセルが残っていた場合にテーブルの先頭のセルを修復する。
【００８１】
フォーム移動モジュール６６では、図１７に示すように、フォームをページの下部に移動する。図１７（ａ）はフォームの存在した位置に移動したフォームへのリンクを残した図であり、図１７（ｂ）はページ下部に移動されたフォームを示す図である。ページ上部にあるフォームは音声ブラウザを用いた場合大きな障害となる。一方、検索フォームのように重要なフォームはページに残すことが好ましい。そこでフォーム移動モジュール６６では、フォームをページ下部へ移動するとともに、そのフォームへのリンクをフォームが存在した場所に残すことにより、簡略化とフォームの保存を両立させることができる。
【００８２】
なお、本実施の形態では実施の形態１とは相違して差分ページ（ＤＯＭ）をＤＯＭからＨＴＭＬへの変換モジュール６２によりＨＴＭＬに変換する。得られる差分ページはＨＴＭＬファイルである（６７）。このような場合、ダイナミックＨＴＭＬの対応しないブラウザでの表示が可能になる。
【００８３】
（実施の形態３）
図１８は、本発明の実施の形態３のシステムを示した構成図である。本実施の形態のシステムは、アノテーション情報による差分ページの修復、修正を可能にする例である。
【００８４】
従来技術の項で前記した通り、詳細なアノテーション情報に基づいて小画面デバイス用の画面出力を得る手法が提案・開発されている。本実施の形態のシステムでは、このようなアノテーション情報と組みあわせることで、より精度の高い出力を得ることができる。ここではアノテーション情報が後処理において用いられる例を示す。図１８に示すように、本実施の形態のプロキシサーバ７０には実施の形態２のプロキシサーバ６０の構成（ユーザプロファイル６３を除く）に加えてアノテーション用の後処理モジュール７１を有する。また本実施の形態のシステムはアノテーションデータベース７２を有する。
【００８５】
目的ページ６に対して、ボランティア７４が詳細なアノテーション情報を入力する。アノテーション情報はアノテーションサーバ７３に入力され、アノテーションデータベース７２に記録される。アノテーション情報がアノテーションデータベース７２に存在する場合、その情報を加味した後処理を行うことができる。アノテーション情報がない場合には実施の形態２と同様な処理が行われる。
【００８６】
図１９はアノテーション用後処理モジュール７１の一例を示したブロック図である。アノテーション用後処理モジュール７１には差分部分マーキングモジュール７５、グループ分割およびグループ選択モジュール７６、グループ並び替えモジュール７７を有する。
【００８７】
ここでは、ページ上のビジュアルなブロックがアノテーションによって指定されているとする。まず、目的ページ６のＤＯＭに対して、差分出力に含まれているノードにマーキングを行う（７５）。次にアノテーション情報に基づいて、ページを分割し、「差分を含まないグループ」を削除する（７６）。残ったグループに対しては、コンテンツをすべて復活させる（７７）。これにより、グループ単位での差分を得ることができ、ビジュアルなブロックを加味した出力を得ることができる。
【００８８】
（実施の形態４）

図２０は、本発明の実施の形態４のシステムを示した構成図である。本実施の形態のシステムは、音声入力に対し音声でブラウジングする形態を例示する。本実施の形態のプロキシサーバ８０における差分処理４は実施の形態２と同様である。また、実施の形態３のアノテーション情報（アノテーション用後処理７１）との組み合わせを適用しても良い。

【００８９】

本実施の形態においては、ユーザ端末はボイスＸＭＬブラウザ８１、電話８２、インターネット電話８３、簡易音声ブラウザ８４である。また、電話８２およびインターネット電話８３をユーザ端末とする場合には音声認識ブラウジングサーバ８５を、ボタン操作による電話８２および簡易音声ブラウザ８４をユーザ端末とする場合にはボタン操作音声ブラウジングサーバ８６を必要とする。さらに、プロキシサーバ８０にはＤＯＭからボイスＸＭＬへの変換モジュール８７、ＤＯＭからＨＴＭＬへの変換モジュール８８を有する。

【００９０】

ボイスＸＭＬブラウザ８１はクライアントサイドで稼動する音声応答を実現するソフトウェア機能である。電話８２では音声入力またはボタン操作入力に対する音声応答を得、インターネット電話８３では音声入力に対する音声応答を得る。簡易音声ブラウザ８４は、キー操作入力に対し音声出力を得る。

【００９１】

音声認識ブラウジングサーバ８５は、音声合成エンジンと音声認識エンジンとを有する。電話８２またはインターネット電話８３から入力された音声を音声認識エンジンで認識し、入力内容を解析する。入力内容はコマンドまたはデータとしてプロキシサーバ８０に送付される。プロキシサーバ８０から受け取ったボイスＸＭＬデータを音声合成エンジンによって音声に変換し、ユーザサイドの端末（電話またはインターネット電話）で再生する。

【００９２】

ボタン操作音声ブラウジングサーバ８６は、電話８２からのボタン操作入力をＤＴＭＦ信号処理部で処理しコマンドまたはデータに変換する。また、簡易音声ブラウザ８４からのキー操作をキー操作処理部で解析しコマンドまたはデータに変換する。コマンドまたはデータはプロキシサーバ８０に送付される。プロキシサーバ８０からの応答はＨＴＭＬファイルとして受け取り、音声合成エンジンがこれを音声に変換してユーザ端末である電話８２または簡易音声ブラウザ８４で再生する。

【００９３】

図２１は、代表的な音声ブラウザの構成を示した図である。ＨＴＴＰリクエストに対する応答として受け取ったＨＴＭＬ文書（ファイル）をパージング・解析モジュール９０で解析し、読み上げ情報９１に変換する。読み上げ情報９１は読み上げ制御モジュール９２で制御されて、音声合成エンジン９３によって音声信号に合成される。合成された音声はスピーカ９４から出力される。なお、読み上げ制御モジュール９２はキーボード等の入力デバイス９５からの入力を受け付けて読み上げ情報を制御できる。

【００９４】

図２２は、ボイスＸＭＬ用の後処理モジュール８９およびＤＯＭからボイスＸＭＬへの変換モジュール８７の一例を示したブロック図である。リストタイトル修復モジュール６４、表頭・表側修復モジュール６５、フォーム移動モジュール６６については実施の形態２と同様である。但し本実施の形態では、各モジュールにおいて修復・移動を行うと同時にリストタイトル９６、テーブルタイトル９７、フォームタイトル９８をリストアップしタイトルリスト９９を生成する。このタイトルはそれぞれのタイトルの内容を「かたまり」としてとらえ、このかたまりの指標であると考えることができる。たとえばリストタイトルはリンクリストを１つのかたまりとして、フォームであれば１つのフォームを１つのかたまりとして考えることができる。これら「かたまり」を認識することによりＨＴＭＬファイルからの音声応答を生成できる。なお、この際に実施の形態３で説明したアノテーション情報により音声合成精度を高めることができる。

【００９５】

得られたタイトルリストに対してキーワード解析１００を施し、音声認識用のボキャブラリを生成する（１０２）。一方ＤＯＭツリーとして得られている差分ページからリンク（アンカータグ）をリストアップし（１０１）、キーワード解析１００を施して音声認識用のボキャブラリを生成する（１０２）。得られたボキャブラリから応答文および文法を生成する（１０３）。

【００９６】

本実施の形態のシステムおよび方法によれば、音声入力あるいは簡単なキー操作入力（ボタン入力）によって、ウェブページの閲覧を音声出力により得ることができる。視覚障害者がウェブコンテンツにアクセスする場合にバリアフリーを実現する有効な手段を提供できる。なお、コンテンツの内容が簡略化されるので音声応答による読み上げがスムーズに進むのは勿論である。また、コンピュータの操作に不慣れなユーザに対しても同様に簡便なウェブコンテンツへのアクセスを可能にする手法を提供できる。

【００９７】

以上、本発明者によってなされた発明を発明の実施の形態に基づき具体的に説明したが、本発明は前記実施の形態に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることは言うまでもない。

【００９８】

【発明の効果】

本願で開示される発明のうち、代表的なものによって得られる効果は、以下の通りである。すなわち、小画面デバイスや音声ブラウザを用いたウェブページの表示あるいは出力の際に、必要な情報に迅速にアクセスするためのウェブページ簡略化の手法を提供できる。また、ウェブページの簡略化を同一ＵＲＬの過去のページが存在しない場合においても実行できる。また、ウェブページの簡略化をオンザフライで実行できる。また、ウェブページの簡略化の際に重要な情報を欠落することなく、不必要な情報を精度良く簡略化する手法を提供できる。

【図面の簡単な説明】

【図１】本発明の実施の形態１のシステムを示した構成図である。

【図２】差分処理の詳細を示したブロック図である。

【図３】隣接ＵＲＬ列挙モジュールの詳細を示したブロック図である。

【図４】ＵＲＬキャッシュモジュールの一例を示した構成図である。

【図５】ＵＲＬ優先順位演算モジュールの一例を示したブロック図である。

【図６】更新前目的ページ／隣接ページ取得モジュールの一例を示した構成図である。

【図７】フェッチモジュールの一例を示す構成図である。

【図８】差分演算モジュールの一例を示したブロック図である。

【図９】ノード位置による重要度の相違を示す図である。

【図１０】通常のブラウザによりニュースページを表示した表示図である。

【図１１】本実施の形態のシステムを用いて図１０のページを表示した表示図である。

【図１２】検索画面の一例を通常のブラウザにより表示した例を示す表示図である。

【図１３】本実施の形態のシステムを用いて図１２のページを表示した表示図である。

【図１４】本発明の実施の形態２のシステムを示した構成図である。

【図１５】後処理モジュールの一例を示したブロック図である。

【図１６】リストタイトル修復モジュールによる修復の例を示す表示図である。

【図１７】（ａ）はフォームの存在した位置に移動したフォームへのリンクを残した図であり、図１７（ｂ）はページ下部に移動されたフォームを示す図である。

【図１８】本発明の実施の形態３のシステムを示した構成図である。

【図１９】アノテーション用後処理モジュールの一例を示したブロック図である。

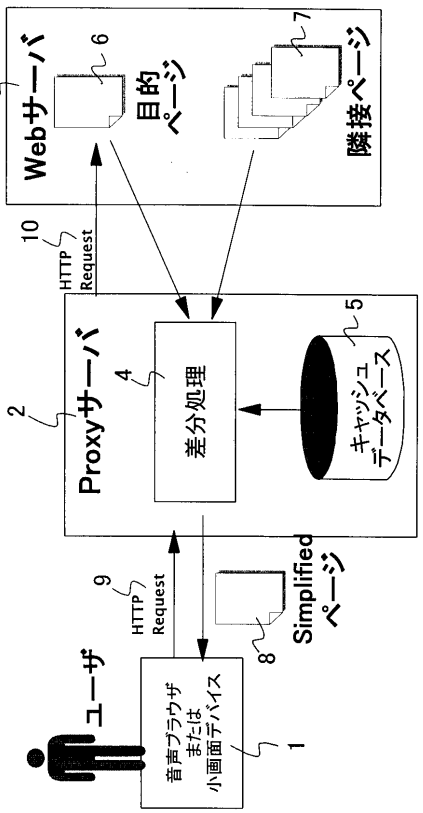
【図２０】本発明の実施の形態４のシステムを示した構成図である。

【図２１】代表的な音声ブラウザの構成を示した図である。

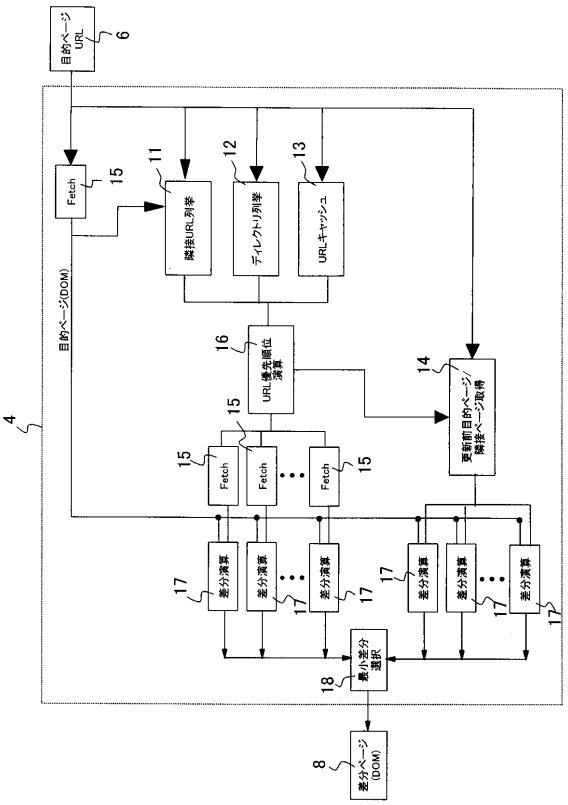
【図２２】ボイスＸＭＬ用の後処理モジュールおよびＤＯＭからボイスＸＭＬへの変換モジュールの一例を示したブロック図である。

【符号の説明】
１…ユーザ端末、２…プロキシサーバ、３…ウェブサーバ、４…差分処理、５…キャッシュデータベース、６…目的ページ、７…隣接ページ、８…簡略化ページ、９…ＨＴＴＰリクエスト、１０…ＨＴＴＰリクエスト、１１…隣接ＵＲＬ列挙モジュール、１２…ディレクトリ列挙モジュール、１３…ＵＲＬキャッシュモジュール、１４…更新前目的ページ／隣接ページ取得モジュール、１５…フェッチモジュール、１６…ＵＲＬ優先順位演算モジュール、１７…差分演算モジュール、１８…最小差分選択モジュール、１９…リンク列挙モジュール、２０…ＵＲＬリスト、２１…ディレクトリ共通ＵＲＬ選択モジュール、２２…親ディレクトリ共通ＵＲＬ選択モジュール、２３…マージモジュール、２４…隣接ＵＲＬリスト、２５…ＵＲＬキャッシュデータベース、２６…検索手段、２７…ＵＲＬリスト、２８…同一ページ除外モジュール、２９…ＵＲＬエディットディスタンス算出モジュール、３０…ＵＲＬ関連性算出モジュール、３１…ソートモジュール、３２…リダイレクションＵＲＬテーブル、３３…ＵＲＬ共起回数テーブル、３４…ＵＲＬ相互参照テーブル、３５…優先度順ＵＲＬリスト、３６…検索キー作成モジュール、３７…検索モジュール、３８…リスト、３９…ＴＨＭＬパーザ、４０…ウェブページ、４１…更新前目的ページ／隣接ページリスト、４２…ダウンロードモジュール、４３…ＨＴＭＬパーザ、４５…ＨＴＭＬファイル、４６…ＤＯＭツリー、４７…線形化モジュール、４８，４９…ノードリスト、５０…ＤＰマッチングモジュール、５１…共通ノードリスト、５２…重要度算出モジュール、５３…共通ノード削除モジュール、５４…クリーンアップモジュール、５５…差分ＤＯＭツリー、６０…プロキシサーバ、６１…後処理モジュール、６２…変換モジュール、６３…ユーザプロファイル、６４…リストタイトル修復モジュール、６５…表頭・表側修復モジュール、６６…フォーム移動モジュール、７０…プロキシサーバ、７１…アノテーション用後処理モジュール、７２…アノテーションデータベース、７３…アノテーションサーバ、７４…ボランティア、７５…差分部分マーキングモジュール、７６…グループ選択モジュール、７７…グループ並び替えモジュール、８０…プロキシサーバ、８１…ボイスＸＭＬブラウザ、８２…電話、８３…インターネット電話、８４…簡易音声ブラウザ、８５…音声認識ブラウジングサーバ、８６…ボタン操作音声ブラウジングサーバ、８７…変換モジュール、８８…変換モジュール、８９…後処理モジュール、９０…パージング・解析モジュール、９１…読み上げ情報、９２…読み上げ制御モジュール、９３…音声合成エンジン、９４…スピーカ、９５…入力デバイス、９６…リストタイトル、９７…テーブルタイトル、９８…フォームタイトル、９９…リストアップしタイトルリスト、１００…キーワード解析。

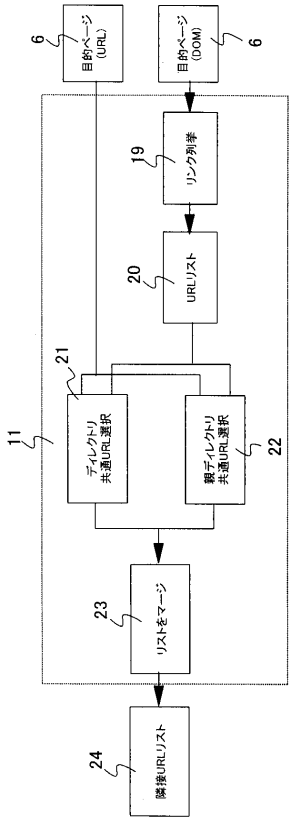
【図１】


Webサーバ
目的ページ
隣接ページ
HTTP Request
Proxyサーバ
差分処理
キャッシュデータベース
HTTP Request
Simplified ページ
ユーザ
音声ブラウザまたは小画面デバイス

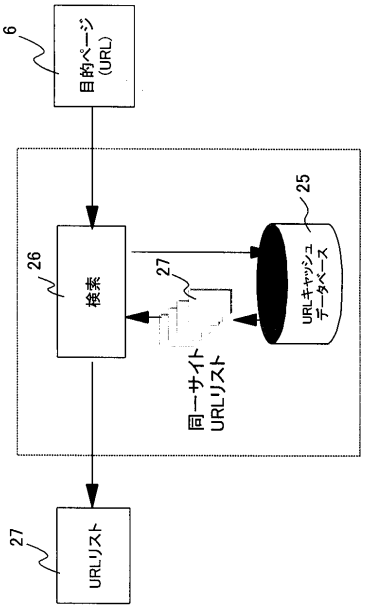
【図２】


目的ページURL
Fetch
隣接URL列挙
ディレクトリ列挙
URLキャッシュ
目的ページ(DOM)
URL優先順位決定
更新前目的ページ/隣接ページ取得
差分演算
最小差分選択
差分ページ(DOM)

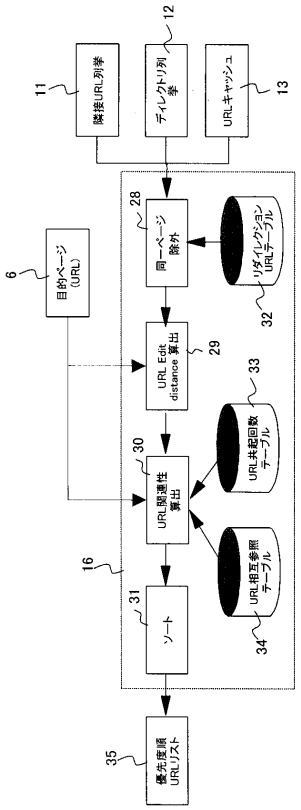
【図３】


目的ページ(URL)
目的ページ(DOM)
リンク列挙
URLリスト
ディレクトリ共通URL選択
親ディレクトリ共通URL選択
リストをマージ
隣接URLリスト

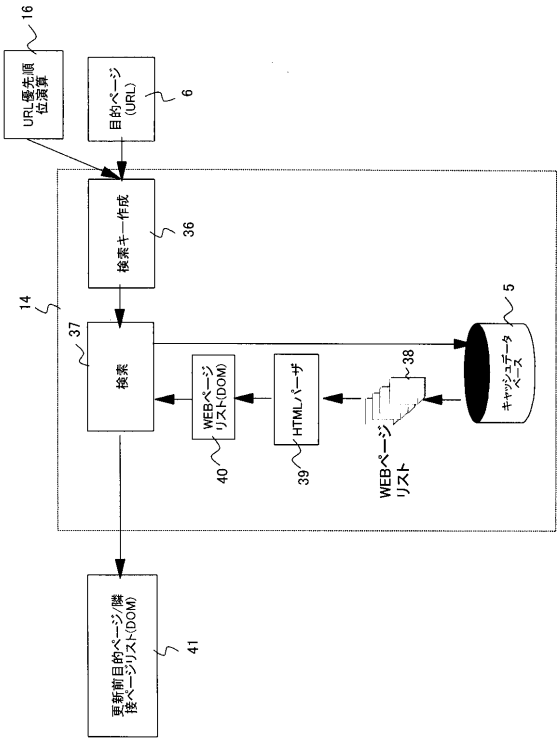
【図４】


目的ページ(URL)
検索
URLキャッシュデータベース
同一サイトURLリスト
URLリスト

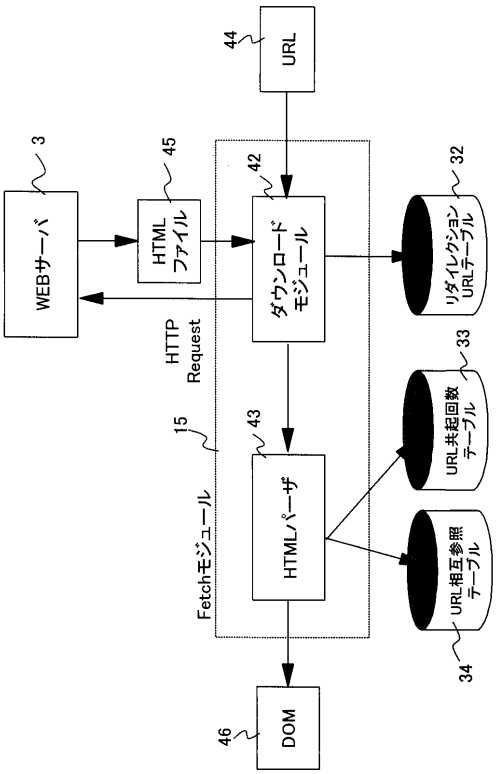
【図５】


11
隣接URL列挙
12
ディレクトリ列挙
13
URLキャッシュ
6
目的ページ(URL)
28
同一ページ除外
32
リダイレクションURLテーブル
29
URL Edit distance 算出
33
URL共起回数テーブル
30
URL関連性算出
34
URL相互参照テーブル
16
31
ソート
35
優先度順URLリスト

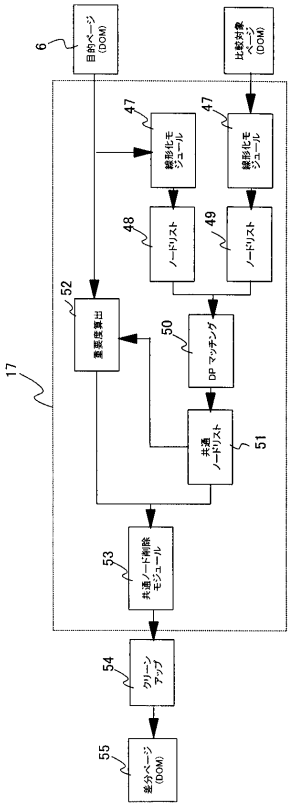
【図６】


16
URL優先順位演算
6
目的ページ(URL)
36
検索キー作成
14
37
検索
40
WEBページリスト(DOM)
39
HTML/パーザ
38
WEBページリスト
5
キャッシュデータベース
41
更新前目的ページ/隣接ページリスト(DOM)

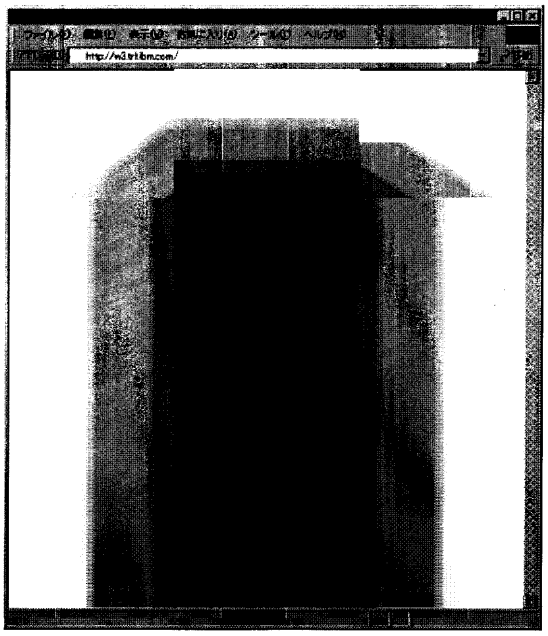
【図７】


44
URL
3
WEBサーバ
45
HTMLファイル
42
ダウンロードモジュール
32
リダイレクションURLテーブル
HTTP Request
15
Fetchモジュール
43
HTMLパーザ
33
URL共起回数テーブル
34
URL相互参照テーブル
46
DOM

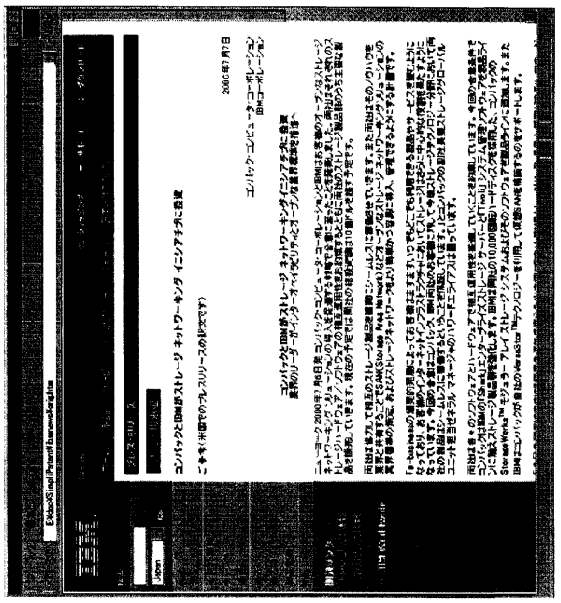
【図８】


6
目的ページ(DOM)
比較対象ページ(DOM)
47
線形化モジュール
48
ノードリスト
49
ノードリスト
52
重要度算出
50
DP マッチング
51
共通ノードリスト
17
53
共通ノード削除モジュール
54
クリーンアップ
55
差分ページ(DOM)

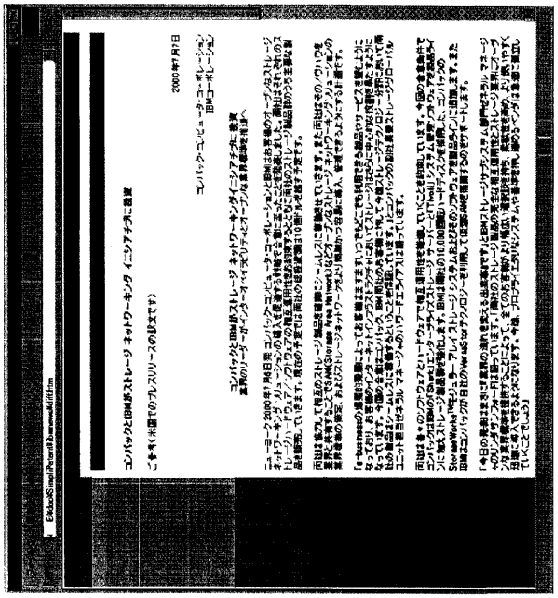
【図９】



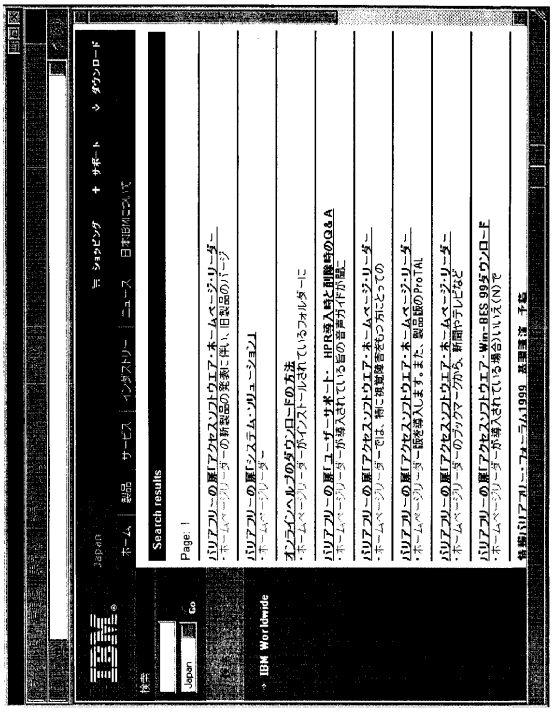
【図１０】



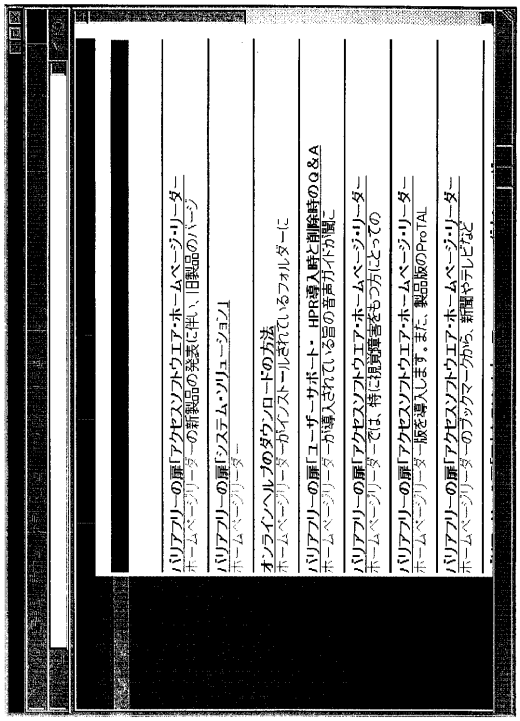
【図１１】



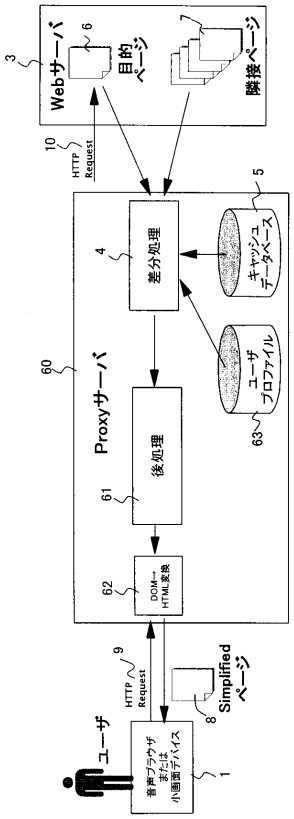
【図１２】



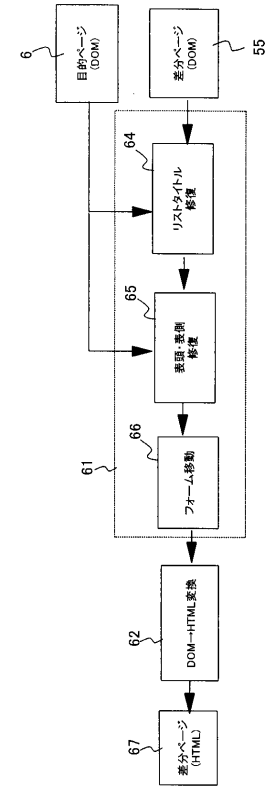
【図１３】



【図１４】



【図１５】



【図１６】

(a)

同ジャンルの他のニュース

- 半導体メーカー、デジタル需要増見込み設備投資を積極化(14:15)
- 太陽電池の生産 本家・米国抜いて日本が世界一に(12:50)
- 企業会計ルールの統一、欧米主導に日本危機感(11:12)
- 税金や公共料金の支払い、電話・ネットでも可能に(23:37)
- TTNetが学校を対象に割安サービス開始 5月から(20:24)
- 伝ルノーが三菱自動車を買収(13:33)

(b)

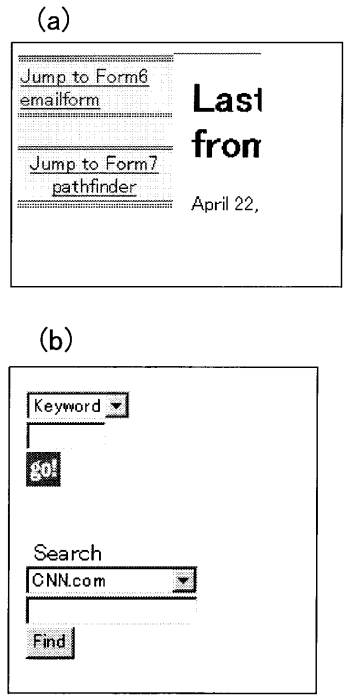
- 半導体メーカー、デジタル需要増見込み設備投資を積極化(14:15)
- 企業会計ルールの統一、欧米主導に日本危機感(11:12)

(c)

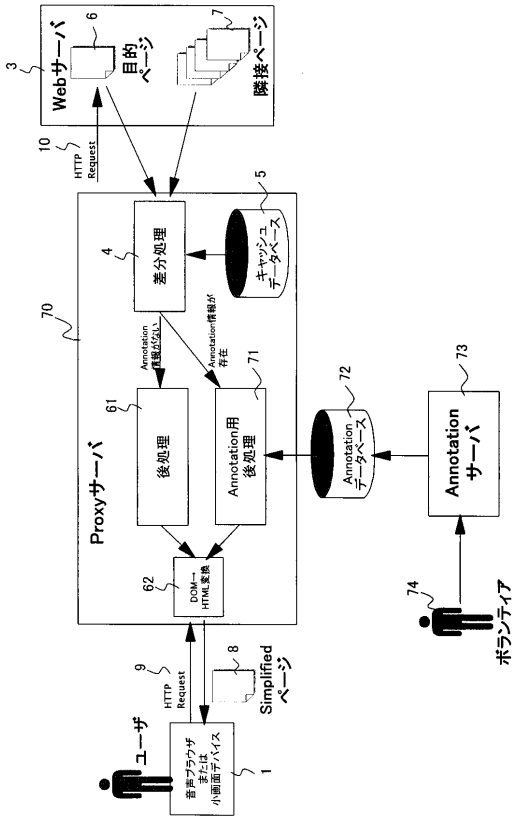
同ジャンルの他のニュース

- 半導体メーカー、デジタル需要増見込み設備投資を積極化(14:15)
- 企業会計ルールの統一、欧米主導に日本危機感(11:12)

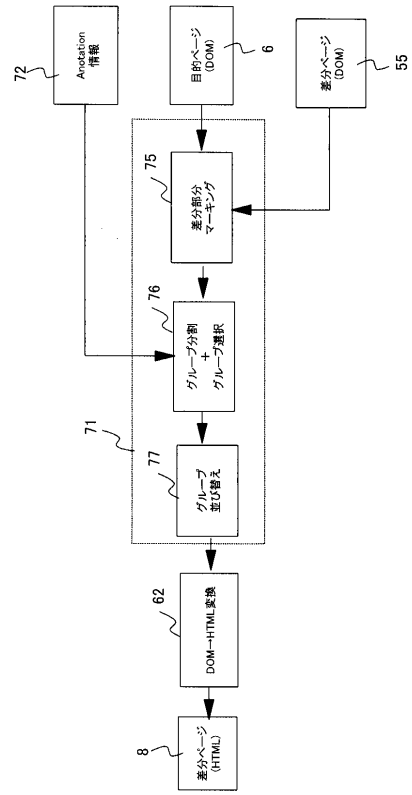
【図１７】



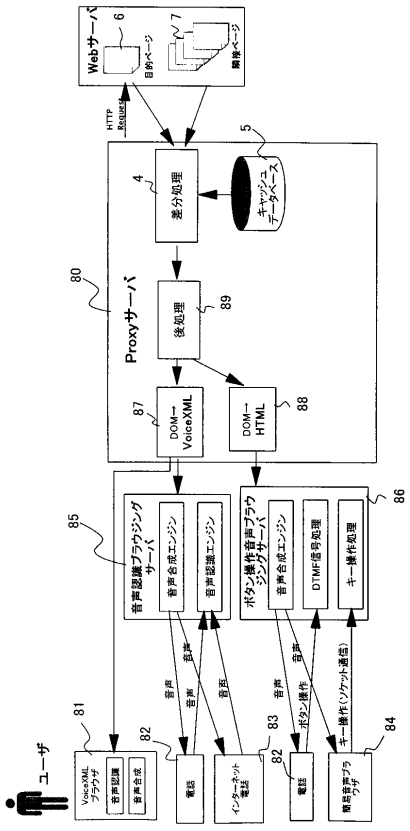
【図１８】



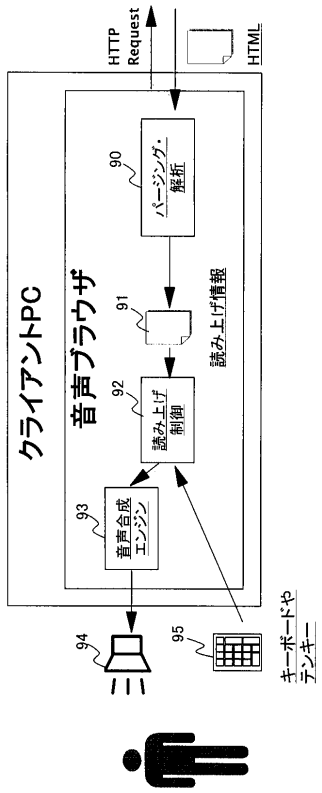
【図１９】



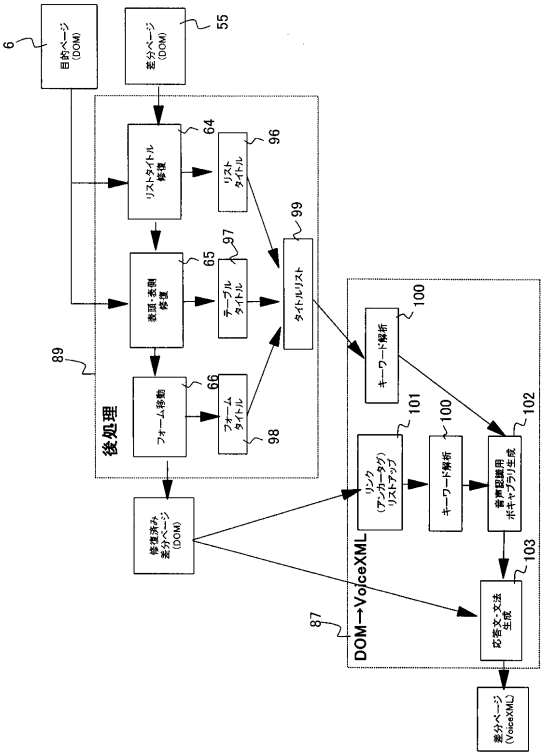
【図２０】



【図２１】


HTTP Request
HTML
クライアントPC
音声ブラウザ
パージング・解析
90
91
読み上げ情報
読み上げ制御
92
音声合成エンジン
93
94
95
キーボードやテンキー

【図２２】


6
目的ページ (DOM)
差分ページ (DOM)
55
後処理
89
リストタイトル修復
64
リストタイトル
96
表頭・表側修復
65
テーブルタイトル
97
フォーム移動
66
フォームタイトル
98
タイトルリスト
99
修復済み差分ページ (DOM)
DOM→VoiceXML
87
キーワード解析
100
リンク(アンカータグ)リストアップ
101
キーワード解析
100
音声認識用ボキャブラリ生成
102
応答文・文法生成
103
差分ページ (VoiceXML)

フロントページの続き

(72)発明者　▲高▼木　啓伸
神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内
(72)発明者　浅川　智恵子
神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

審査官　平井　誠

(56)参考文献　特開２０００－１４８６４２（ＪＰ，Ａ）
特表２００２－５１８７２６（ＪＰ，Ａ）
特開平１０－２４０６０４（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F 12/00
G06F　3/00
G06F　3/14
G06F 13/00