(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2011-123740
(P2011-123740A)
(43)公開日　平成23年6月23日(2011.6.23)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 13/00	(2006.01)	G06F 13/00	540A	5B084

審査請求 未請求　請求項の数 8　ＯＬ　（全 14 頁）

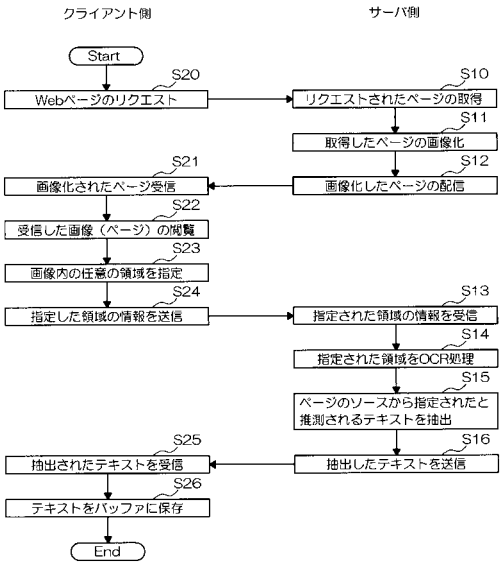
(21)出願番号　特願2009-281880 (P2009-281880)
(22)出願日　平成21年12月11日 (2009.12.11)

(71)出願人　306037311
富士フイルム株式会社
東京都港区西麻布２丁目２６番３０号
(74)代理人　100083116
弁理士　松浦　憲三
(72)発明者　福島　敏貢
東京都港区赤坂９丁目７番３号　富士フイルム株式会社内

Ｆターム(参考)　5B084 AA12　AB04　AB06　CB12　CB22
DB02　DC02　DC03

(54)【発明の名称】閲覧システム、サーバ、テキスト抽出方法及びプログラム

(57)【要約】
【課題】画像化したウェブページを端末に送信し、端末装置でウェブページを閲覧する場合において、端末装置に表示された画像内の文字を正確に抽出することができる。
【解決手段】サーバ１０は、インターネットからウェブページを取得し（ステップＳ１０）、取得したウェブページから画像を生成し（ステップＳ１１）、画像をクライアント端末２０へ送信する（ステップＳ１２）。クライアント端末２０は、画像を受信し（ステップＳ２１）、表示部２３へ表示し（ステップＳ２２）、矩形領域を指定し（ステップＳ２３）、その情報をサーバ１０へ送信する（ステップＳ２４）。サーバ１０は、画像から矩形領域の画像を切り出し、ＯＣＲ処理によりテキストを認識し（ステップＳ１４）、Ｈｔｍｌファイルのソースから認識されたテキストと最も一致度の高いテキストを抽出し（ステップＳ１５）、クライアント端末２０へ送信する（ステップＳ１６）。
【選択図】　図４



【特許請求の範囲】
【請求項１】
表示手段が設けられた端末装置と、前記端末装置と接続されたサーバとで構成された閲覧システムであって、
前記端末装置は、
前記サーバから送信された画像データを受信する端末側受信手段と、
前記受信された画像データに基づいて前記表示手段に画像を表示させる表示制御手段と、
前記表示手段に表示された画像の中の所定の領域を選択する選択手段と、
前記選択された所定の領域の情報を前記サーバへ送信する端末側送信手段と、を備え、
前記サーバは、
ウェブページのソースを取得する取得手段と、
前記取得されたウェブページのソースに基づいて当該ウェブページの画像データを生成する画像生成手段と、
前記生成された画像データを前記端末装置に送信するサーバ側送信手段と、
前記端末装置から送信された所定の領域の情報を受信するサーバ側受信手段と、
前記受信された所定の領域の情報と前記生成された画像データとに基づいて、前記所定の領域の画像からＯＣＲ処理により文字を認識する文字認識手段と、
前記ＯＣＲ処理により認識された文字と推定される文字列を前記取得されたウェブページのソースから抽出する文字列抽出手段と、を備え、
前記サーバ側送信手段は、前記抽出された文字列を前記端末装置に送信し、
前記端末側受信手段は、前記送信された文字列を受信することを特徴とする閲覧システム。
【請求項２】
前記サーバは、前記所定の領域が閾値以上であるか否かを判断する判断手段を備え、
前記所定の領域が閾値以上であると判断されなかった場合には、前記サーバ側送信手段は、前記ＯＣＲ処理により認識された文字列を送信することを特徴とする請求項１に記載の閲覧システム。
【請求項３】
前記端末側送信手段は、前記所定の領域の情報として当該所定の領域の座標の情報を前記サーバへ送信し、
前記文字認識手段は、前記生成された画像データと、前記所定の領域の座標の情報とから前記所定の領域の画像を切り出し、当該切り出された所定の領域の画像から文字を認識することを特徴とする請求項１又は２に記載の閲覧システム。
【請求項４】
前記文字列抽出手段は、前記ＯＣＲ処理により認識された文字をキーと前記取得されたソースに含まれるテキストとを比較し、前記ＯＣＲ処理により認識された文字と最も一致度の高い文字列を抽出することを特徴とする請求項１、２又は３に記載の閲覧システム。
【請求項５】
前記端末装置は、前記受信した文字列を記憶する記憶手段を備えたことを特徴とする請求項１から４のいずれかに記載の閲覧システム。
【請求項６】
請求項１から５のいずれかに記載の閲覧システムを構成するサーバ。
【請求項７】
携帯端末からウェブページの閲覧要求を受け付けるステップと、
前記受け付けられた閲覧要求に基づいてウェブページのソースを取得するステップと、
前記取得されたウェブページのソースに基づいて当該ウェブページの画像データを生成するステップと、
前記端末装置から所定の領域の情報を受信するステップと、
前記受信した所定の領域の情報と前記生成された画像データとに基づいて、前記所定の

領域の画像からＯＣＲ処理により文字を認識するステップと、
前記取得されたソースから前記ＯＣＲ処理により認識された文字と推定される文字列を抽出するステップと、
前記抽出された文字列を前記端末装置に送信するステップと、
を含むことを特徴とするテキスト抽出方法。
【請求項８】
請求項７に記載のテキスト抽出方法を演算装置に実行させることを特徴とするプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は閲覧システム、サーバ、テキスト抽出方法及びプログラムに係り、特に携帯端末でウェブページが閲覧可能な閲覧システム、サーバ、テキスト抽出方法及びプログラムに関する。
【背景技術】
【０００２】
近年、携帯電話にフルブラウザが搭載されることが多くなり、携帯電話からＰＣ用のウェブページを閲覧することが可能となっている。しかしながら、携帯電話でＰＣ用のウェブページを閲覧する場合には、画面が小さいため、ページのレイアウトが崩れて閲覧しづらい等といった問題が起こる場合がある。また、企業のイントラページなどは、安全性を確保するため、アクセスが制限され、携帯電話からは閲覧することができない。
【０００３】
このような問題を解決するための方法として、サーバでウェブページやイントラページを画像化して携帯電話へ配信するというシステムが考えられる。
【０００４】
引用文献１には、ウェブページをサーバ側でレンダリングし、画像に変換したページをクライアントに配信するシステムが記載されている。
【０００５】
引用文献２には、クライアント装置のウェブブラウザからＯＣＲ処理の対象とする領域を指定し、サーバでＯＣＲ処理を行うシステムが記載されている。
【０００６】
引用文献３には、画像データを文字認識（ＯＣＲ（Optical Character Reader）処理）にかけ、テキストを抽出し、さらに抽出したテキストデータを構文意味解析処理にかけることにより文章のエラーを検出し、修正を行うことで文字（文章）の認識精度を高めるシステムが記載されている。
【先行技術文献】
【特許文献】
【０００７】
【特許文献１】特開２００４－２２０２６０号公報
【特許文献２】特開２００５－３２７２５８号公報
【特許文献３】特開２００６－３５０６６３号公報
【発明の概要】
【発明が解決しようとする課題】
【０００８】
しかしながら、特許文献１に記載の発明では、クライアントに配信するウェブページは画像化されているため、テキスト領域を選択してコピーするといった操作ができなかった。
【０００９】
特許文献２に記載の発明では、ＯＣＲ処理により画像データからテキストデータを得ることはできるが、テキストデータの精度を向上させる方法については記載されていない。

【００１０】
特許文献３に記載の発明では、ＯＣＲ処理の精度が低い場合には、構文意味解析ができず、正しいテキストデータが得られないという問題がある。また、構文意味解析ができた場合であっても、得られたテキストデータが画像データに実際に含まれるテキストデータとならないという問題がある。
【００１１】
本発明はこのような事情に鑑みてなされたもので、画像化したウェブページを端末に送信し、端末装置でウェブページを閲覧する場合において、端末装置に表示された画像内の所定の領域に含まれる文字を正確に抽出することができる閲覧システム、サーバ、テキスト抽出方法及びプログラムを提供することを目的とする。
【課題を解決するための手段】
【００１２】
請求項１に記載の閲覧システムは、表示手段が設けられた端末装置と、前記端末装置と接続されたサーバとで構成された閲覧システムであって、前記端末装置は、前記サーバから送信された画像データを受信する端末側受信手段と、前記受信された画像データに基いて前記表示手段に画像を表示させる表示制御手段と、前記表示手段に表示された画像の中の所定の領域を選択する選択手段と、前記選択された所定の領域の情報を前記サーバへ送信する端末側送信手段と、を備え、前記サーバは、ウェブページのソースを取得する取得手段と、前記取得されたウェブページのソースに基づいて当該ウェブページの画像データを生成する画像生成手段と、前記生成された画像データを前記端末装置に送信するサーバ側送信手段と、前記端末装置から送信された所定の領域の情報を受信するサーバ側受信手段と、前記受信された所定の領域の情報と前記生成された画像データとに基づいて、前記所定の領域の画像からＯＣＲ処理により文字を認識する文字認識手段と、前記ＯＣＲ処理により認識された文字と推定される文字列を前記取得されたウェブページのソースから抽出する文字列抽出手段と、を備え、前記サーバ側送信手段は、前記抽出された文字列を前記端末装置に送信し、前記端末側受信手段は、前記送信された文字列を受信することを特徴とする。
【００１３】
請求項１に記載の閲覧システムによれば、サーバでは、ウェブページのソースが取得され、取得されたウェブページのソースに基づいて当該ウェブページの画像データが生成され、生成された画像データが端末装置に送信される。端末装置では、送信された画像データが受信され、受信された画像データに基づいて表示手段に画像が表示され、表示手段に表示された画像の中の所定の領域が選択され、選択された所定の領域の情報がサーバへ送信される。サーバでは、端末装置から送信された所定の領域の情報が受信され、受信された所定の領域の情報と生成された画像データとに基づいて所定の領域の画像からＯＣＲ処理により文字が認識され、ＯＣＲ処理により認識された文字と推定される文字列が取得されたソースから抽出され、抽出された文字列が端末装置に送信される。携帯端末では、サーバから送信された文字列が受信される。これにより、ＯＣＲ処理のミスにより間違ったテキストが認識された場合においても、そのミスを補完し、選択した領域に含まれる正確なテキストデータを得ることができる。例えば、下線付き文字や表の一部等ＯＣＲ処理の精度が低い場合においても、正確なテキストデータを得ることができる。
【００１４】
請求項２に記載の閲覧システムは、請求項１に記載の閲覧システムにおいて、前記サーバは、前記所定の領域が閾値以上であるか否かを判断する判断手段を備え、前記所定の領域が閾値以上であると判断されなかった場合には、前記サーバ側送信手段は、前記ＯＣＲ処理により認識された文字列を送信することを特徴とする。
【００１５】
請求項２に記載の閲覧システムによれば、サーバでは、所定の領域が閾値以上であるか否かが判断され、所定の領域が閾値以上であると判断されなかった場合には、ＯＣＲ処理により認識された文字列が端末装置へ送信される。これにより、効率よく、かつ精度よく

選択した領域に含まれるテキストデータを得ることができる。

【００１６】

請求項３に記載の閲覧システムは、請求項１又は２に記載の閲覧システムにおいて、前記端末側送信手段は、前記所定の領域の情報として当該所定の領域の座標の情報を前記サーバへ送信し、前記文字認識手段は、前記生成された画像データと、前記所定の領域の座標の情報とから前記所定の領域の画像を切り出し、当該切り出された所定の領域の画像から文字を認識することを特徴とする。

【００１７】

請求項３に記載の閲覧システムによれば、所定の領域の情報として所定の領域の座標の情報が端末装置からサーバへ送信されると、サーバでは、生成された画像データと、所定の領域の座標の情報とから所定の領域の画像が切り出され、切り出された所定の領域の画像から文字が認識される。これにより、処理能力の高いサーバで重い処理、すなわち座標に従い指定された領域の画像を抽出する処理を行い、処理能力の低い端末装置で行う処理は、処理コストの小さい矩形領域の座標の送信のみとすることができる。

【００１８】

請求項４に記載の閲覧システムは、請求項１、２又は３に記載の閲覧システムにおいて、前記文字列抽出手段は、前記ＯＣＲ処理により認識された文字をキーと前記取得されたソースに含まれるテキストとを比較し、前記ＯＣＲ処理により認識された文字と最も一致度の高い文字列を抽出することを特徴とする。

【００１９】

請求項４に記載の閲覧システムによれば、文字列抽出手段では、ＯＣＲ処理により認識された文字をキーと取得されたソースに含まれるテキストとが比較され、ＯＣＲ処理により認識された文字と最も一致度の高い文字列が抽出される。これにより、ソースから選択した領域に含まれるテキストデータを抽出することができる。

【００２０】

請求項５に記載の閲覧システムは、請求項１から４のいずれかに記載の閲覧システムにおいて、前記端末装置は、前記受信した文字列を記憶する記憶手段を備えたことを特徴とする。

【００２１】

請求項５に記載の閲覧システムによれば、端末装置では、サーバから送信された文字列が記憶手段に記憶される。これにより、サーバから送信されたテキストを、任意のテキストフィールドへの貼り付けなどに利用することができる。すなわち、クライアント端末で選択された領域の画像に含まれるテキストのコピーと同等の効果を得ることができる。

【００２２】

請求項６に記載のサーバは、請求項１から５のいずれかに記載の閲覧システムを構成する。

【００２３】

請求項７に記載のテキスト抽出方法は、携帯端末からウェブページの閲覧要求を受け付けるステップと、前記受け付けられた閲覧要求に基づいてウェブページのソースを取得するステップと、前記取得されたウェブページのソースに基づいて当該ウェブページの画像データを生成するステップと、前記端末装置から所定の領域の情報を受信するステップと、前記受信した所定の領域の情報と前記生成された画像データとに基づいて、前記所定の領域の画像からＯＣＲ処理により文字を認識するステップと、前記取得されたソースから前記ＯＣＲ処理により認識された文字と推定される文字列を抽出するステップと、前記抽出された文字列を前記端末装置に送信するステップと、を含むことを特徴とする。

【００２４】

請求項８に記載のプログラムは、請求項７に記載のテキスト抽出方法を演算装置に実行させることを特徴とする。

【発明の効果】

【００２５】

本発明によれば、画像化したウェブページを端末に送信し、端末装置でウェブページを閲覧する場合において、端末装置に表示された画像内の所定の領域に含まれる文字を正確に抽出することができる。

【図面の簡単な説明】

【００２６】

【図１】本発明が適用された閲覧システム１の概略図である。

【図２】閲覧システム１を構成するサーバの概略図である。

【図３】閲覧システム１を構成するクライアント端末の概略図である。

【図４】閲覧システム１のクライアント端末がテキストデータをコピーする取得する処理の流れを示すフローチャートである。

【図５】クライアント端末に表示される閲覧用画像の一例である。

【図６】ＯＣＲ処理を説明するための図である。

【図７】テキスト抽出処理を説明するための図である。

【図８】一致度が最も高いテキストを抽出する方法を説明するための図である。

【図９】テキスト送信処理を説明するための図である。

【図１０】本発明が適用された閲覧システム２のクライアント端末がテキストデータをコピーする取得する処理の流れを示すフローチャートである。

【図１１】閲覧システム２のテキスト抽出処理について説明するための図である。

【発明を実施するための形態】

【００２７】

＜第１の実施の形態＞

閲覧システム１は、主として、サーバ１０と、クライアント端末２０とで構成される。サーバ１０と接続されるクライアント端末２０は１台でも良いし、複数でもよい。

【００２８】

サーバ１０は、図２に示すように、主として、ＣＰＵ１１と、データ取得部１２と、画像生成部１３と、ＯＣＲ処理部１４と、テキスト抽出部１５と、通信部１６とで構成される。

【００２９】

ＣＰＵ１１は、サーバ１０の全体の動作を統括制御する制御手段として機能するとともに、各種の演算処理を行う演算手段として機能する。ＣＰＵ１１は、制御プログラムであるファームウェア、ウェブページを表示するためのプログラムであるブラウザ、制御に必要な各種データ等を記憶するメモリ領域を有する。また、ＣＰＵ１１は、ＣＰＵ１１の作業用領域として利用されるとともに、表示用の画像データなどの一時記憶領域として利用されるメモリ領域を有する。

【００３０】

データ取得部１２は、インターネット３１と接続されており、クライアント端末２０から要求されたウェブページのコンテンツ等をインターネット３１を介して取得する。また、データ取得部１２は、文書データベース（ＤＢ）３２と接続されており、クライアント端末２０から要求された文書ファイルなどの各種データを文書ＤＢ３２から取得する。

【００３１】

画像生成部１３は、データ取得部１２が取得したコンテンツ、文書データから画像（以下、閲覧用画像という）を生成する。画像生成部１３は、生成した閲覧用画像をＣＰＵ１１のメモリ領域に記憶する。

【００３２】

ＯＣＲ処理部１４は、入力された画像に含まれる文字を識別して文書に変換する。ＯＣＲ処理自体は一般的な技術であるため、詳細な説明は省略する。

【００３３】

テキスト抽出部１５は、ＣＰＵ１１により取得されたウェブページのソースから、ＯＣＲ処理部１４が取得したテキストと最も一致度が高いテキストを抽出する。また、テキスト抽出部１５は、ＣＰＵ１１により取得された文書データから、ＯＣＲ処理部１４が取得

したテキストと最も一致度が高いテキストを抽出する。テキスト抽出部１５の処理の詳細については、後に詳述する。

【００３４】

通信部１６は、閲覧用画像等をクライアント端末２０へ送信する。また、通信部１６は、クライアント端末２０から送信されたウェブページ閲覧要求等を受信する。

【００３５】

クライアント端末２０は、例えば小型ノートパソコンや携帯電話等であり、図１に示すように、ネットワークを介してサーバ１０と接続される。クライアント端末２０は、図３に示すように、主として、ＣＰＵ２１と、入力部２２と、表示部２３と、表示制御部２４と、通信部２５とで構成される。なお、クライアント端末２０は、小型ノートパソコンや携帯電話に限定されるものではなく、ウェブブラウザを動作させ得る情報端末であればどのような端末でもよい。

【００３６】

ＣＰＵ２１は、クライアント端末２０の全体の動作を統括制御するとともに、各種の演算処理を行う演算手段として機能する。ＣＰＵ２１は、クライアント端末２０のクライアント端末情報や、各種制御に必要なプログラム等が記憶されるメモリ領域を有する。また、ＣＰＵ２１は、サーバ１０から送信された各種データを一時的に記憶するバッファを有する。

【００３７】

入力部２２は、ユーザが各種指示を入力するためのものであり、テンキー、十字キー等で構成される。

【００３８】

表示部２３は、例えば、カラー表示が可能な液晶ディスプレイである。なお、表示部２３は、カラー表示に限定されず、白黒表示でもよい。また、表示部２３は、液晶ディスプレイに限定されず、有機ＥＬ等を用いてもよい。

【００３９】

表示制御部２４は、サーバ１０から送信された閲覧用画像を表示部２３に表示させる。

【００４０】

通信部２５は、サーバ１０から送信された閲覧用画像、テキストデータ等を受信する。また、通信部２５は、ウェブページ閲覧要求、領域の情報等をサーバ１０へ送信する。

【００４１】

上記のように構成された閲覧システム１の作用について説明する。閲覧システム１では、クライアント端末２０にウェブページ（又は文書データ）の画像が表示され、クライアント端末２０により所定の領域が選択されると、その領域内のテキストをコピーすることができる。図４は、クライアント端末２０が表示部２３に表示されたウェブページ内のテキストをコピーする処理の流れを示すフローチャートである。

【００４２】

クライアント端末２０のＣＰＵ２１は、メモリ領域に記憶されたウェブブラウザを起動する。入力部２２により閲覧したいウェブページの情報（URL等）が入力されると、ＣＰＵ２１は、これを受け付けてサーバ１０へリクエストを送信する（ステップＳ２０）。

【００４３】

サーバ１０のＣＰＵ１１は、リクエストを受信するとデータ取得部１２に指示を出し、データ取得部１２はインターネットからリクエストされたウェブページを取得する（ステップＳ１０）。この場合には、サーバ１０はプロキシとして動作し、外部のサーバからコンテンツ（例えば、ウェブページのＨｔｍｌファイル）を取得する。ＣＰＵ１１は、取得したコンテンツをバッファに記憶する。なお、サーバ１０は、ウェブサーバとしても機能しても良く、この場合にはサーバ１０の図示しないメモリに記憶されているコンテンツを取得する。

【００４４】

データ取得部１２は取得したコンテンツを画像生成部１３に出力し、画像生成部１３は

コンテンツから閲覧用画像を生成する（ステップＳ１１）。ウェブページのＨｔｍｌファイルを取得した場合には、画像生成部１３は、Ｈｔｍｌファイルを解析し、解析結果に基づいて文字や画像を適切に配置した結果を画像化（レンダリング）し、gif、jpeg等の画像ファイルとして保存する。

【００４５】

画像生成部１３は生成した閲覧用画像をＣＰＵ１１へ出力し、ＣＰＵ１１は閲覧用画像をクライアント端末２０へ送信する（ステップＳ１２）。

【００４６】

クライアント端末２０のＣＰＵ２１は、サーバ１０から送信された閲覧用画像を受信し（ステップＳ２１）、表示制御部２４へ出力する。表示制御部２４は、受信した画像を表示部２３へ表示させる（ステップＳ２２）。これにより、図５に示すように、クライアント端末２０にリクエストしたウェブページの画像が表示され、ユーザがウェブページを閲覧可能となる。

【００４７】

表示部２３に閲覧用画像が表示された状態下で、入力部２２によりテキストを抽出（コピー）したい領域の指定が行われる（ステップＳ２３）。領域の指定は、例えば、ユーザが入力部２２の十字キー等でカーソルを移動させ、領域の始点及び終点の位置を選択入力することにより行われる。入力部２２による入力結果がＣＰＵ２１で検出されると、ＣＰＵ２１は、図５に示すように、始点と終点とにより形成される矩形領域が指定されたと認識する。なお、領域の指定は、この形態に限らず、始点と終点の座標の値を直接入力する等の様々な方法により行うことができる。

【００４８】

ＣＰＵ２１は、認識した矩形領域の情報をサーバ１０へ送信する（ステップＳ２４）。矩形領域の情報としては、矩形領域の始点及び終点の座標が考えられる。図５に示す場合には、閲覧用画像の左上を原点（Ｘ座標、Ｙ座標共に０）とし、右方向を＋Ｘ方向、下方向を＋Ｙ方向として座標が指定される。ただし、座標の指定方法はこれに限定されるものではない。ＣＰＵ２１は、矩形領域の情報として、閲覧用画像から矩形領域を切り出し、切り出された画像を矩形領域の情報として送信するようにしてもよい。

【００４９】

サーバ１０のＣＰＵ１１は、クライアント端末２０から送信された矩形領域の情報を受信する（ステップＳ１３）。ＣＰＵ１１は、矩形領域の情報をＯＣＲ処理部１４へ出力する。

【００５０】

ＯＣＲ処理部１４は、矩形領域の情報に基づいて矩形領域に含まれる文字を認識する（ステップＳ１４）。矩形領域の情報として矩形領域の始点及び終点の座標が入力された場合には、ＯＣＲ処理部１４は、画像生成部１３から閲覧用画像を取得し、閲覧用画像と座標とから矩形領域の画像を切り出す。本実施の形態では、ＯＣＲ処理部１４は、図５の点線で囲まれた領域の画像を矩形領域の画像として切り出す。

【００５１】

そして、ＯＣＲ処理部１４は、切り出した画像をＯＣＲ処理することにより、矩形領域に含まれる文字を認識する。図６に示すように、ＯＣＲ処理部１４は、矩形領域に含まれる「ベルリンで開催された世界陸上をはじめ、週末のスポーツイベント結果ほか、今注目すべき選手についてご紹介」という文字をＯＣＲ処理し、「ペルリンで閘催された世界陸上をばじ助、週末のスポーツイ伴ント結果ほか、いま注目ずぺ舌選手1ごついてこ紹弁。」という認識結果を得る。

【００５２】

矩形領域の情報として閲覧用画像から切り出された画像が入力された場合には、ＯＣＲ処理部１４は、座標情報から画像を抽出する処理は不要であり、入力された画像を直接ＯＣＲ処理し、文字を認識する。閲覧システムの実施形態としては、一般的にクライアント端末とサーバではサーバの処理能力のほうが高いため、クライアント端末では処理コスト

の小さい矩形領域の座標の送信のみを行い、サーバで座標に従い指定された領域の画像を抽出する処理を行う方が好ましい。

【００５３】

ＯＣＲ処理部１４は、得られた認識結果をテキストデータとしてテキスト抽出部１５に出力する。テキスト抽出部１５は、バッファに記憶されたＨｔｍｌファイルを取得し、図７に示すように、Ｈｔｍｌファイルのソースに含まれるテキストの中から入力されたテキストデータと推定されるテキストを抽出する（ステップＳ１５）。ステップＳ１５の処理は、例えば、入力されたテキストデータをキーとして、ソース内から最も一致度の高いテキストを抽出することにより行われる。本実施の形態では、ページのソースとしてＨｔｍｌファイルを用いたが、Ｈｔｍｌファイルに限られるものではなく、クライアント端末２０に送信した閲覧用画像の基となるウェブページをレンダリングするために必要な情報であればどのようなものでもよい。

【００５４】

最も一致度の高いテキストを抽出する方法について、図８を用いて説明する。ＯＣＲ処理部１４により「ＡＢＣ」というテキストが認識された場合には、テキスト抽出部１５は、「ＡＢＣ」というテキストとソースとを順番に比較し、一致度を算出する。例えば、「ＡＢＣ」というテキストとソース内のテキスト「ＡＶＡ」との一致度は３３％であり、「ＡＢＣ」というテキストとソース内のテキスト「ＶＡＢ」との一致度は０％であり、「ＡＢＣ」というテキストとソース内のテキスト「ＡＢＡ」との一致度は６６％であり、「ＡＢＣ」というテキストとソース内のテキスト「ＥＡＣ」との一致度は３３％である。一致度が最も高いのは、「ＡＢＣ」というテキストとソース内のテキスト「ＡＢＡ」とを比較した場合であるため、テキスト抽出部１５は、ソース内のテキスト「ＡＢＡ」を抽出する。

【００５５】

図７に示す場合には、テキスト抽出部１５は、ステップＳ１４で認識されたテキスト「ペルリンで閘催された世界陸上をばじ助、週末のスポーツイ伴ント結果ほか、いま注目ずぺ舌選手1ごついてこ紹弁。」をキーとして、ソース内から最も一致度の高いテキストの抽出を行う。その結果、テキスト抽出部１５は、「ベルリンで開催された世界陸上をはじめ、週末のスポーツイベント結果ほか、いま注目すべき選手についてご紹介。」というテキストを抽出する。

【００５６】

そして、テキスト抽出部１５は、抽出されたテキストをクライアント端末２０で指定された矩形領域に含まれるテキストと判定する。クライアント端末２０で指定された矩形領域に含まれるテキストは、必ずソース内に含まれるテキストである。したがって、ソース内に含まれるテキストからＯＣＲ処理の結果得られたテキストを推測して抽出することにより、ＯＣＲ処理のミスにより間違ったテキストが認識された場合においても、そのミスを補完し、正しいテキストを抽出することができる。

【００５７】

なお、本実施の形態では、ステップＳ１５において、ステップＳ１０で取得され、バッファに記憶されたＨｔｍｌファイルを用いたが、ステップＳ１５の処理の前に改めてＨｔｍｌファイルを取得してもよい。また、ステップＳ１５においては、ソースに含まれるテキスト全てを抽出対象としても良いし、ソースがＨｔｍｌファイルでメタ情報（タグ）が含まれている場合等であれば、タグを除いたレンダリングの対象となるテキストのみを抽出対象としても良い。

【００５８】

テキスト抽出部１５は、抽出したテキストをＣＰＵ１１に出力し、図９に示すように、ＣＰＵ１１はテキストをクライアント端末２０へ送信する（ステップＳ１６）。クライアント端末２０のＣＰＵ２１は、サーバ１０から送信されたテキストを受信し（ステップＳ２５）、受信したテキストをＣＰＵ２１内のバッファに記憶する（ステップＳ２６）。バッファに保存したテキストは、例えば任意のテキストフィールドへの貼り付けなどに利用

することなどが考えられる。

【００５９】

本実施の形態によれば、ウェブページや文書データを画像化してクライアント端末に表示させる場合に、クライアント端末に表示された画像の一部を選択することにより、選択した領域に含まれる正確なテキストデータを得ることができる。そして、得られたテキストデータを記憶することにより、クライアント端末で選択された領域の画像に含まれるテキストをコピーすることと同等の効果を得ることができる。

【００６０】

従来のシンクライアント型ブラウザでは、クライアント端末で閲覧されるウェブページは画像化されているため、ウェブページに含まれるテキストをコピーすることはできなかった。しかしながら、ＯＣＲ処理とソートからのテキスト抽出とを組み合わせることにより、シンクライアント型ブラウザを用いる場合においても所望のテキストのコピーアンドペーストが可能となる。

【００６１】

また、本実施の形態によれば、下線付き文字や表の一部等ＯＣＲ処理の精度が低い場合においても、正確なテキストデータをコピーすることができる。例えば、ステップＳ２３で図５の一点鎖線で囲んだ領域が矩形領域として選択された場合には、ステップＳ１４のＯＣＲ処理において、行間の線が原因で上段のテキストは正確な認識結果は得られない。しかしながら、図７に示すようにソースと比較することにより、「各党の政権公約比較「安全保障」」及び「候補者情報」「マニフェスト」「選挙ニュース」というテキストを抽出することができる。

【００６２】

なお、本実施の形態では、図４に示すようにウェブページを閲覧する場合を例に作用を説明したが、ウェブページの閲覧のみでなく、文書データを閲覧する場合においても同様の方法により、選択した矩形領域内のテキストを抽出することができる。

【００６３】

＜第２の実施の形態＞

第１の実施の形態は、ＯＣＲ処理のミスにより間違ったテキストが認識された場合においても、そのミスを補完し、正しいテキストを抽出するため、ソースに含まれるテキストの中からテキストを抽出する処理を行なったが、必ずしもソースからのテキスト抽出処理が必要とは限らない。例えば、単語等テキストの長さが短い場合には、ＯＣＲ処理の制度が高いため、処理結果が正しい場合も多い。

【００６４】

第２の実施の形態は、クライアント端末で選択された矩形領域の大きさ、即ちテキストの長さに応じてテキスト抽出処理をするかしないかを異ならせる形態である。以下、第２の実施の形態に係る閲覧システム２について説明する。なお、閲覧システム２の構成は閲覧システム１と同様であるため、説明を省略する。また、第１の実施の形態と同一の部分については、同一の符号を付し、詳細な説明を省略する。

【００６５】

図１０は、閲覧システム２において、クライアント端末２０により選択された領域内のテキストをコピーする処理の流れを示すフローチャートである。

【００６６】

クライアント端末２０のＣＰＵ２１は、メモリ領域に記憶されたウェブブラウザを起動する。入力部２２により閲覧したいウェブページの情報（URL等）が入力されると、ＣＰＵ２１は、これを受け付けてサーバ１０へリクエストを送信する（ステップＳ２０）。

【００６７】

サーバ１０のＣＰＵ１１は、リクエストを受信するとデータ取得部１２に指示を出し、データ取得部１２はインターネットからリクエストされたウェブページを取得する（ステップＳ１０）。データ取得部１２は取得したコンテンツを画像生成部１３に出力し、画像生成部１３はコンテンツから閲覧用画像を生成する（ステップＳ１１）。画像生成部１３

は生成した閲覧用画像をＣＰＵ１１へ出力し、ＣＰＵ１１は閲覧用画像をクライアント端末２０へ送信する（ステップＳ１２）。

【００６８】

クライアント端末２０のＣＰＵ２１は、サーバ１０から送信された閲覧用画像を受信し（ステップＳ２１）、表示制御部２４へ出力する。表示制御部２４は、受信した画像を表示部２３へ表示させる（ステップＳ２２）。これにより、クライアント端末２０にリクエストしたウェブページの画像が表示され、ユーザがウェブページを閲覧可能となる。

【００６９】

表示部２３に閲覧用画像が表示された状態下で、テキストを抽出（コピー）したい矩形領域の指定が行われる（ステップＳ２３）。指定された矩形領域の情報はＣＰＵ２１で検出され、ＣＰＵ２１は、認識した矩形領域の情報をサーバ１０へ送信する（ステップＳ２４）。

【００７０】

サーバ１０のＣＰＵ１１は、クライアント端末２０から送信された矩形領域の情報を受信する。ＣＰＵ１１は、受信された矩形領域の情報に基づいて、矩形領域の大きさ（面積）を算出する（ステップＳ１７）。

【００７１】

ＣＰＵ１１は、矩形領域の情報をＯＣＲ処理部１４へ出力する。ＯＣＲ処理部１４は、矩形領域の情報に基づいて矩形領域に含まれる文字を認識する（ステップＳ１４）。

【００７２】

ＣＰＵ１１はステップＳ１３で受信された矩形領域の大きさが閾値以上であるか否かを判断する（ステップＳ１８）。なお、閾値は、予め設定された任意の値であり、ＣＰＵ１１のメモリ領域に記憶されている。閾値は、必要に応じてクライアント端末２０等から変更することもできる。閾値としては、ＯＣＲ処理により正しい結果が得られる最大の長さ（単語レベルの長さ）のテキストが含まれるような面積とすることが望ましい。

【００７３】

矩形領域の大きさが閾値以上である場合（ステップＳ１８でＹＥＳ）は、クライアント端末２０により指定された領域に含まれるテキストは文章等の長いテキストであると推定される。テキストが長い場合には、ＯＣＲ処理の精度は低く、正確に文字が認識できない場合が多い。したがって、ＯＣＲ処理部１４は得られた認識結果をテキストデータとしてテキスト抽出部１５に出力し、テキスト抽出部１５はバッファに記憶されたＨｔｍｌファイルのソースに含まれるテキストの中から入力されたテキストデータと推定されるテキストを抽出する（ステップＳ１５）。テキスト抽出部１５は抽出されたテキストをＣＰＵ１１に出力し、ＣＰＵ１１はテキストをクライアント端末２０へ送信する（ステップＳ１９）。これにより、ＯＣＲ処理のミスにより間違ったテキストが認識される可能性が高い場合においても、そのミスを補完し、正しいテキストを抽出することができる。

【００７４】

矩形領域の大きさが閾値以上でない場合（ステップＳ１７でＮＯ）は、クライアント端末２０により指定された領域に含まれるテキストは単語レベルであると推定される。単語であれば、ＯＣＲ処理の精度がある程度期待できる。また、短いテキストをソースから抽出することで、間違ったテキストを抽出しやすくなり、精度が低下するおそれもある。したがって、この場合には、ＯＣＲ処理部１４は得られた認識結果をＣＰＵ１１に出力し、ＣＰＵ１１はテキストをクライアント端末２０へ送信する（ステップＳ１９）。

【００７５】

ステップＳ１８～Ｓ１９の処理について、図１１を用いて具体的に説明する。閾値が「５０」である場合に、ステップＳ１７で算出された面積が「２００」である場合には、算出された面積「２００」は閾値「５０」より大きいため、Ｈｔｍｌファイルのソースに含まれるテキストの中から正しいと推定されるテキストを抽出し、その結果をクライアント端末２０で指定された矩形領域に含まれるテキストと判定する。それに対し、ステップＳ１７で算出された面積が「１０」である場合には、算出された面積「１０」は閾値「５０

」より小さいため、テキスト抽出は行わず、ＯＣＲ処理により得られた結果をクライアント端末２０で指定された矩形領域に含まれるテキストと判定する。

【００７６】

クライアント端末２０のＣＰＵ２１は、サーバ１０から送信されたテキストを受信し（ステップＳ２５）、受信したテキストをＣＰＵ２１内のバッファに記憶する（ステップＳ２６）。バッファに保存したテキストは、例えば任意のテキストフィールドへの貼り付けなどに利用することなどが考えられる。

【００７７】

本実施の形態によれば、矩形領域の大きさに応じて送信するテキストの抽出方法を変えることにより、効率、精度の良い処理を行うことができる。

【００７８】

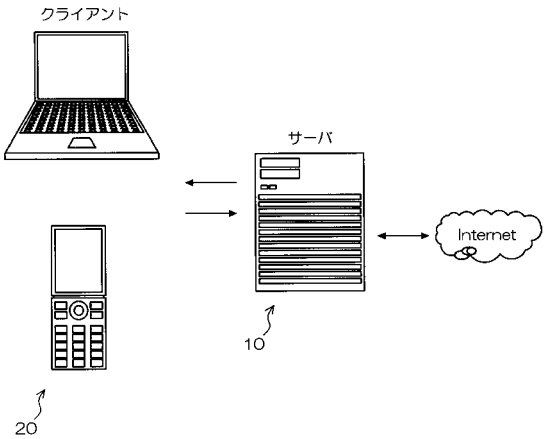
なお、上記第１、第２の実施の形態では、サーバとクライアント端末とを有するシステムを例に説明したが、本発明は、システムに限らず、外部の装置へ画像を配信するサーバとして提供することもできる。また、サーバ、クライアント端末に適用するプログラムとして提供することもできる。

【符号の説明】

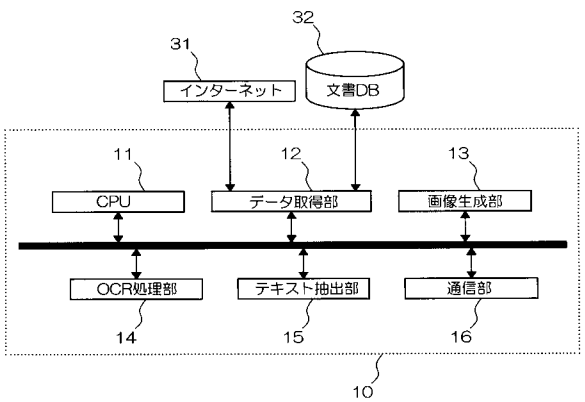
【００７９】

１、２：閲覧システム、１０：サーバ、１１：ＣＰＵ、１２：データ取得部、１３：画像生成部、１４：ＯＣＲ処理部、１５：テキスト抽出部、１６：通信部、２０：クライアント端末、２１：ＣＰＵ、２２：入力部、２３：表示部、２４：表示制御部、２５：通信部

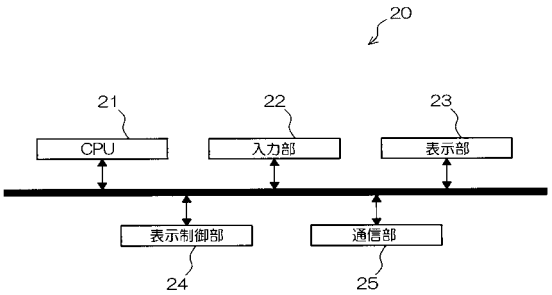
【図１】



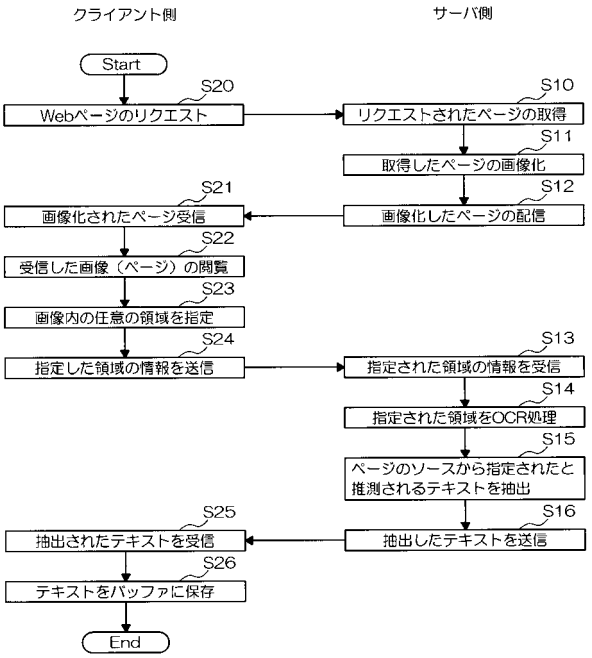
【図２】



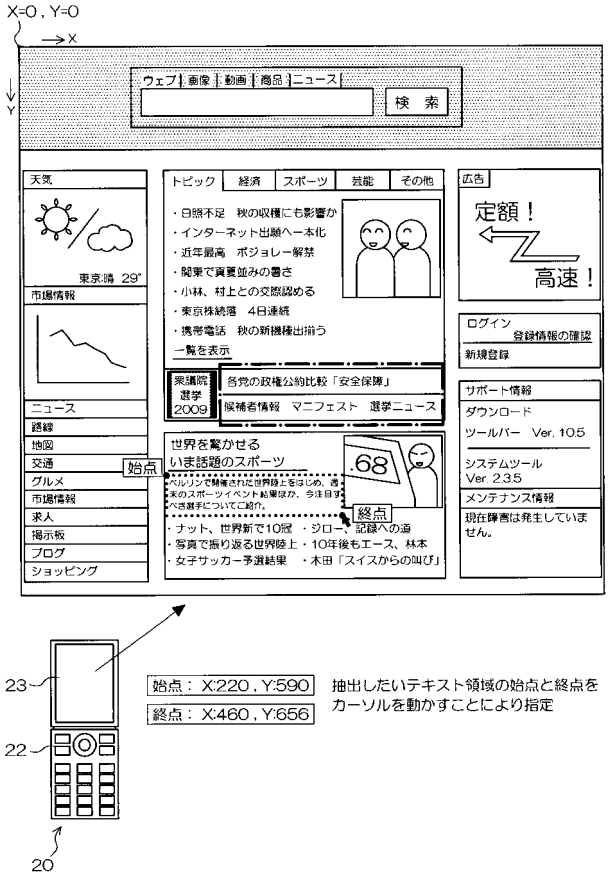
【図３】



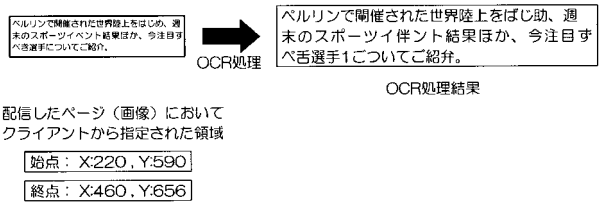
【図４】


クライアント側
サーバ側
Start
S20
Webページのリクエスト
S10
リクエストされたページの取得
S11
取得したページの画像化
S12
画像化したページの配信
S21
画像化されたページ受信
S22
受信した画像（ページ）の閲覧
S23
画像内の任意の領域を指定
S24
指定した領域の情報を送信
S13
指定された領域の情報を受信
S14
指定された領域をOCR処理
S15
ページのソースから指定されたと推測されるテキストを抽出
S16
抽出したテキストを送信
S25
抽出されたテキストを受信
S26
テキストをバッファに保存
End

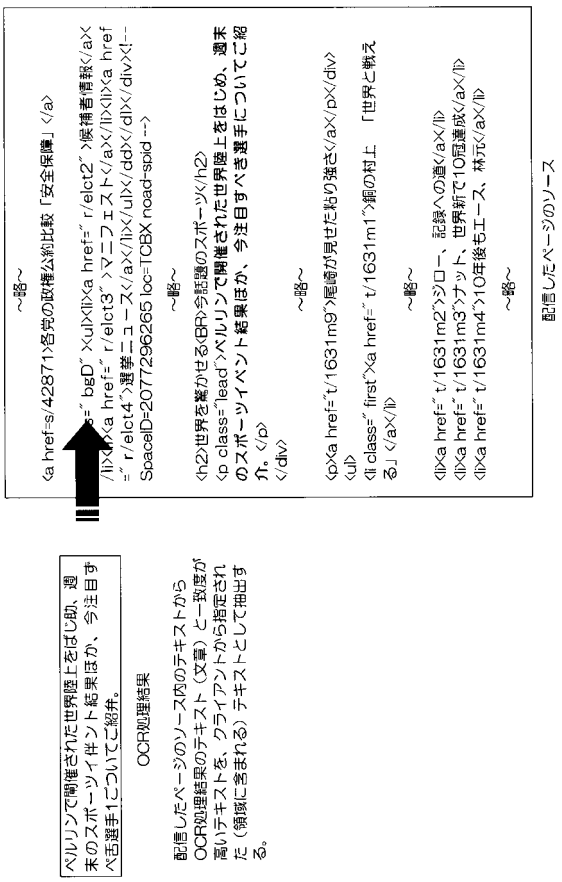
【図５】


X=0，Y=0
X
Y
ウェブ 画像 動画 商品 ニュース
検 索
天気
東京晴 29°
市場情報
ニュース
路線
地図
交通
グルメ
市場情報
求人
掲示板
ブログ
ショッピング
トピック 経済 スポーツ 芸能 その他
・日照不足 秋の収穫にも影響か
・インターネット出願へ一本化
・近年最高 ポジョレー解禁
・関東で真夏並みの暑さ
・小林、村上との交際認める
・東京株続落 4日連続
・携帯電話 秋の新機種出揃う
一覧を表示
衆議院選挙 2009
各党の政権公約比較「安全保障」
候補者情報 マニフェスト 選挙ニュース
始点
世界を驚かせる いま話題のスポーツ
・ナット、世界新で10冠 ・ジロー、記録への道
・写真で振り返る世界陸上・10年後もエース、林本
・女子サッカー予選結果 ・木田「スイスからの叫び」
終点
広告
定額！
高速！
ログイン
登録情報の確認
新規登録
サポート情報
ダウンロード
ツールバー Ver. 105
システムツール Ver 2.35
メンテナンス情報
現在障害は発生していません。
23
22
20
始点： X:220，Y:590
終点： X:460，Y:656
抽出したいテキスト領域の始点と終点をカーソルを動かすことにより指定

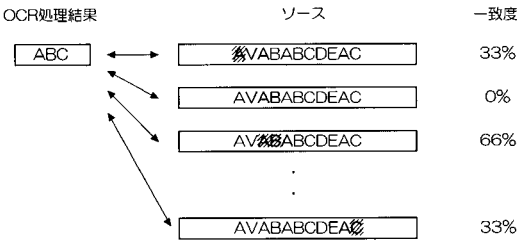
【図６】


ペルリンで開催された世界陸上をはじめ、週末のスポーツイベント結果ほか、今注目すべき選手についてご紹介。
OCR処理
ペルリンで開催された世界陸上をばじ助、週末のスポーツイ伴ン結果ほか、今注目すべ舌選手1こついてご紹弁。
OCR処理結果
配信したページ（画像）において
クライアントから指定された領域
始点： X:220，Y:590
終点： X:460，Y:656

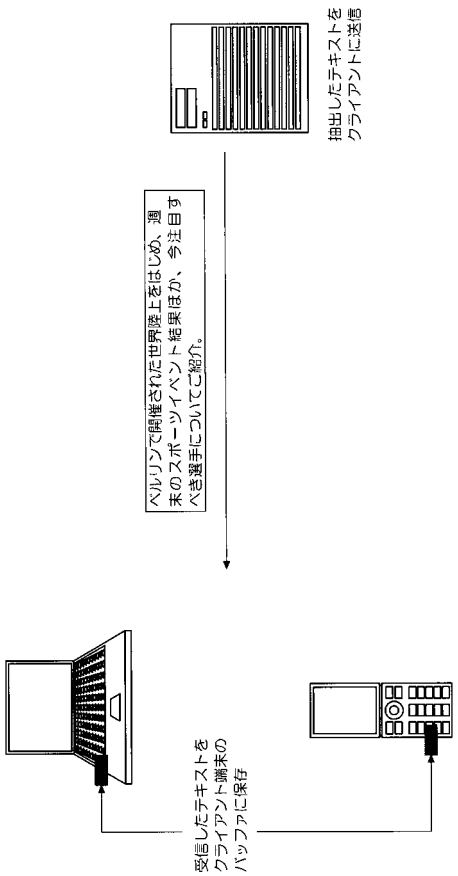
【図７】


～略～
各党の政権公約比較「安全保障」
～略～
<h2>世界を驚かせるくBR>今話題のスポーツ</h2>
ペルリンで開催された世界陸上をはじめ、週末のスポーツイベント結果ほか、今注目すべき選手についてご紹介。
～略～
尾崎が見せた粘り強さ
～略～
ジロー、記録への道
ナット、世界新で10冠達成
10年後もエース、林元
～略～
配信したページのソース
ペルリンで開催された世界陸上をばじ助、週末のスポーツイ伴ン結果ほか、今注目すべ舌選手1こついてご紹弁。
OCR処理結果
配信したページのソースからOCR処理結果のテキスト（文章）と一致度が高いテキストを、クライアントから指定された（領域に含まれる）テキストとして抽出する。

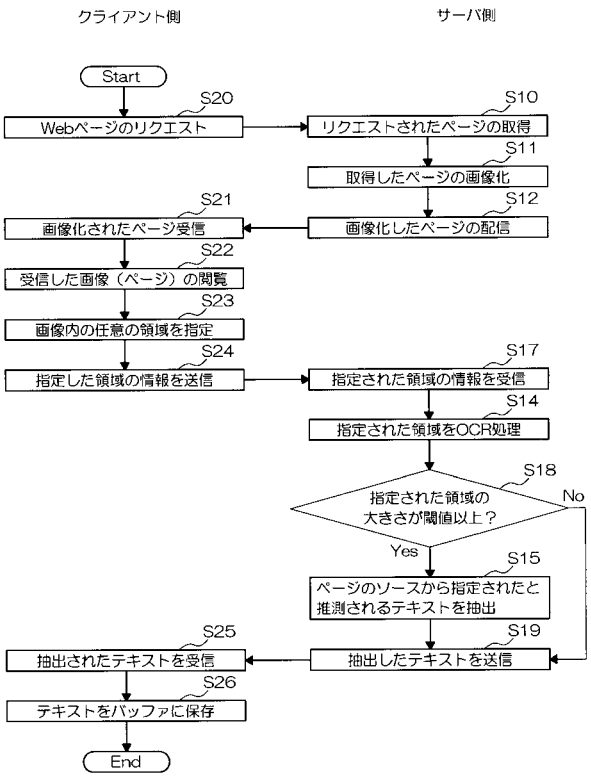
【図８】


OCR処理結果
ソース
一致度
ABC
AVABABCDEAC
33%
AVABABCDEAC
0%
AVABABCDEAC
66%
AVABABCDEAC
33%

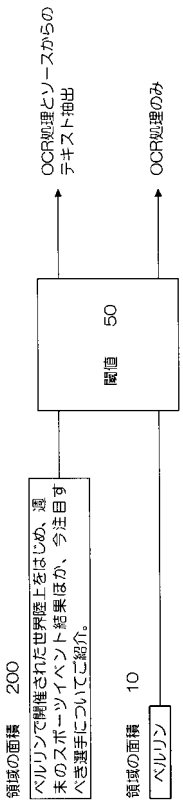
【図９】


抽出したテキストを
クライアントに送信
ベルリンで開催された世界陸上をはじめ、週末のスポーツイベント結果ほか、今注目すべき選手についてご紹介。
受信したテキストを
クライアント端末の
バッファに保存

【図１０】


クライアント側
サーバ側
Start
S20
Webページのリクエスト
S10
リクエストされたページの取得
S11
取得したページの画像化
S12
画像化したページの配信
S21
画像化されたページ受信
S22
受信した画像（ページ）の閲覧
S23
画像内の任意の領域を指定
S24
指定した領域の情報を送信
S17
指定された領域の情報を受信
S14
指定された領域をOCR処理
S18
指定された領域の
大きさが閾値以上？
No
Yes
S15
ページのソースから指定されたと
推測されるテキストを抽出
S19
抽出したテキストを送信
S25
抽出されたテキストを受信
S26
テキストをバッファに保存
End

【図１１】


OCR処理とソースからの
テキスト抽出
OCR処理のみ
閾値 50
領域の面積 200
ベルリンで開催された世界陸上をはじめ、週末のスポーツイベント結果ほか、今注目すべき選手についてご紹介。
領域の面積 10
ベルリン