

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5223293号
(P5223293)

(45) 発行日 平成25年6月26日(2013.6.26)

(24) 登録日 平成25年3月22日(2013.3.22)

(51) Int.Cl. F I
G 0 6 F 17/30 (2006.01) G O 6 F 17/30 2 1 O A
 G O 6 F 17/30 1 7 O C

請求項の数 7 (全 12 頁)

(21) 出願番号	特願2007-277705 (P2007-277705)	(73) 特許権者	000000295 沖電気工業株式会社 東京都港区虎ノ門一丁目7番12号
(22) 出願日	平成19年10月25日(2007.10.25)	(74) 代理人	100090620 弁理士 工藤 宣幸
(65) 公開番号	特開2009-104528 (P2009-104528A)	(72) 発明者	福居 毅至 東京都港区虎ノ門1丁目7番12号 沖電 気工業株式会社内
(43) 公開日	平成21年5月14日(2009.5.14)	審査官	打出 義尚
審査請求日	平成22年8月11日(2010.8.11)		

最終頁に続く

(54) 【発明の名称】 位置表現抽出装置、方法及びプログラム

(57) 【特許請求の範囲】

【請求項1】

少なくとも、複数の位置表現と、これら各位置表現の位置を特定する位置座標情報とを対応付けて格納する位置表現・位置座標情報対応テーブルと、

入力された文書から1又は複数の位置表現を抽出し、上記位置表現・位置座標情報対応テーブルを参照して、抽出した上記各位置表現の位置座標情報を求める位置表現抽出手段と、

上記文書から1又は複数の非テキストコンテンツを抽出し、上記各非テキストコンテンツに含まれている1又は複数の位置情報を抽出する非テキストコンテンツ位置情報抽出手段と、

上記非テキストコンテンツ位置情報抽出手段により抽出された上記各位置情報が示す位置と、上記位置表現抽出手段により抽出された上記各位置表現の位置座標情報が示す位置との間の距離を求め、上記各非テキストコンテンツの各位置情報が示す位置から距離が最小となる位置表現を代表位置表現と判定して出力する代表位置表現判定手段と

を備えることを特徴とする位置表現抽出装置。

【請求項2】

上記代表位置表現判定手段は、上記各非テキストコンテンツの上記各位置情報が示す位置と、抽出された上記各位置表現の位置座標情報が示す位置との距離に応じて、上記各位置表現に対して優先順位を付与することを特徴とする請求項1に記載の位置表現抽出装置

。

【請求項 3】

上記代表位置表現判定手段は、当該非テキストコンテンツの上記位置情報が示す位置と、抽出されたすべての上記位置表現の位置座標情報が示す位置との距離が、予め定めた閾値より大きい場合、当該非テキストコンテンツに対する代表位置表現を選択しないことを特徴とする請求項 1 又は 2 に記載の位置表現抽出装置。

【請求項 4】

上記代表位置表現判定手段は、上記各非テキストコンテンツの上記各位置情報が示す位置と、抽出された上記各位置表現の位置座標情報が示す位置との距離が最小となる位置表現を出力することに代えて、上記各非テキストコンテンツの上記各位置情報が示す位置と、抽出された上記各位置表現の位置座標情報が示す位置との距離が小さい順に複数の位置表現を並べ替えて、予め定められた個数の上記各位置表現を出力することを特徴とする請求項 1 ~ 3 のいずれかに記載の位置表現抽出装置。

10

【請求項 5】

上記非テキストコンテンツ位置情報抽出手段は、上記各非テキストコンテンツに含まれる上記位置情報が位置座標情報でない場合、上記位置表現・位置座標情報対応テーブルを参照して、上記各非テキストコンテンツの各位置情報から位置座標情報を取得することを特徴とする請求項 1 に記載の位置表現抽出装置。

【請求項 6】

コンピュータが、少なくとも、複数の位置表現と、これら各位置表現の位置を特定する位置座標情報とを対応付けて格納する位置表現・位置座標情報対応テーブルと、位置表現抽出手段と、非テキストコンテンツ位置情報抽出手段と、代表位置表現判定手段とを備え、

20

上記位置表現抽出手段が、入力された文書から 1 又は複数の位置表現を抽出し、上記位置表現・位置座標情報対応テーブルを参照して、抽出した上記各位置表現の位置座標情報を求める位置表現抽出工程と、

上記非テキストコンテンツ位置情報抽出手段が、上記文書から 1 又は複数の非テキストコンテンツを抽出し、上記各非テキストコンテンツに含まれている 1 又は複数の位置情報を抽出する非テキストコンテンツ位置情報抽出工程と、

上記代表位置表現判定手段が、上記非テキストコンテンツ位置情報抽出手段により抽出された上記各位置情報が示す位置と、上記位置表現抽出手段により抽出された上記各位置表現の位置座標情報が示す位置との間の距離を求め、上記各非テキストコンテンツの各位置情報が示す位置から距離が最小となる位置表現を代表位置表現と判定して出力する代表位置表現判定工程と

30

を有することを特徴とする位置表現抽出方法。

【請求項 7】

コンピュータが、少なくとも、複数の位置表現と、これら各位置表現の位置を特定する位置座標情報とを対応付けて格納する位置表現・位置座標情報対応テーブルを有し、

コンピュータを、

入力された文書から 1 又は複数の位置表現を抽出し、上記位置表現・位置座標情報対応テーブルを参照して、抽出した上記各位置表現の位置座標情報を求める位置表現抽出手段

40

、上記文書から 1 又は複数の非テキストコンテンツを抽出し、上記各非テキストコンテンツに含まれている 1 又は複数の位置情報を抽出する非テキストコンテンツ位置情報抽出手段、

上記非テキストコンテンツ位置情報抽出手段により抽出された上記各位置情報が示す位置と、上記位置表現抽出手段により抽出された上記各位置表現の位置座標情報が示す位置との間の距離を求め、上記各非テキストコンテンツの各位置情報が示す位置から距離が最小となる位置表現を代表位置表現と判定して出力する代表位置表現判定手段

として機能させる位置表現抽出プログラム。

【発明の詳細な説明】

50

【技術分野】**【0001】**

本発明は、位置表現抽出装置、方法及びプログラムに関し、例えば、ネットワーク上に存在する様々なコンテンツが有する位置表現を抽出する位置表現抽出装置、方法及びプログラムに適用し得るものである。

【背景技術】**【0002】**

近年、例えばインターネット等のネットワーク上に存在する様々なコンテンツに付与された位置情報（例えば、GPS情報など）を抽出し、この位置情報を用いた位置情報サービスが提供されている。

10

【0003】

また、例えばGPS機能付きデジタルカメラを用いた位置情報サービスも提供されている。このサービスは、位置情報（GPS情報）を取得して、撮影時に位置情報を画像に付与する。そして、地図ソフトなどを用いて、位置情報をキーとして地図上の位置を検索するというものである。

【0004】

しかしながら、ネットワーク上のコンテンツは位置情報が明示的に付与されていないものがほとんどである。また、上記GPS機能付きデジタルカメラを用いたサービスの場合も、位置情報が画像に自動的に付与される技術が定着しておらず、ほとんどの画像は位置情報が付与されないままネットワーク上に存在している。

20

【0005】

従って、上記のような位置情報サービスを充実させるためには、インターネット上のコンテンツから位置情報を自動的に抽出する技術が重要になる。

【0006】

従来、この種の技術としては、特許文献1に記載の技術がある。特許文献1には、ウェブテキストについて住所情報に関する抽出範囲を決定し、この決定した抽出範囲内のウェブテキストから複数の住所情報に該当する文字列を抽出する。そして、少なくとも抽出した文字列が特定のキーワードを含むかどうかを調べることにより各住所情報の重要度を決定し、重要度の低い住所情報に関する抽出した文字列を削除して必要な住所情報に関する文字列のみを残す、という技術である。

30

【0007】

【特許文献1】特開2004-280659号公報

【発明の開示】**【発明が解決しようとする課題】****【0008】**

しかしながら、上述した特許文献1に記載の技術は、特定のキーワードを含まない文字列の重要度に差をつけないため、一定の主題をもつ文書（例えば、ニュース記事やウェブログ（いわゆるブログ）記事など）においては主題と無関係な文字列も抽出してしまう。

【0009】

つまり、対象の文書の主題とは何ら関連しない住所情報も抽出してしまうため、最適な位置情報の抽出ができない。

40

【0010】

そのため、処理対象である入力文書の主題に強く関連する位置表現を判定することができ、その主題に対応した最適な位置表現を抽出することができる位置表現抽出装置、方法及びプログラムが求められている。

【課題を解決するための手段】**【0011】**

かかる課題を解決するために第1の本発明の位置表現抽出装置は、(1)少なくとも、複数の位置表現と、これら各位置表現の位置を特定する位置座標情報とを対応付けて格納する位置表現・位置座標情報対応テーブルと、(2)入力された文書から1又は複数の位

50

置表現を抽出し、位置表現・位置座標情報対応テーブルを参照して、抽出した各位置表現の位置座標情報を求める位置表現抽出手段と、(3)文書から1又は複数の非テキストコンテンツを抽出し、各非テキストコンテンツに含まれている1又は複数の位置情報を抽出する非テキストコンテンツ位置情報抽出手段と、(4)非テキストコンテンツ位置情報抽出手段により抽出された各位置情報が示す位置と、位置表現抽出手段により抽出された上記各位置表現の位置座標情報が示す位置との間の距離を求め、各非テキストコンテンツの各位置情報が示す位置から距離が最小となる位置表現を代表位置表現と判定して出力する代表位置表現判定手段とを備えることを特徴とする。

【0012】

第2の本発明の位置表現抽出方法は、コンピュータが、少なくとも、複数の位置表現と、これら各位置表現の位置を特定する位置座標情報とを対応付けて格納する位置表現・位置座標情報対応テーブルと、位置表現抽出手段と、非テキストコンテンツ位置情報抽出手段と、代表位置表現判定手段とを備え、(1)位置表現抽出手段が、入力された文書から1又は複数の位置表現を抽出し、位置表現・位置座標情報対応テーブルを参照して、抽出した各位置表現の位置座標情報を求める位置表現抽出工程と、(2)非テキストコンテンツ位置情報抽出手段が、文書から1又は複数の非テキストコンテンツを抽出し、各非テキストコンテンツに含まれている1又は複数の位置情報を抽出する非テキストコンテンツ位置情報抽出工程と、(3)代表位置表現判定手段が、非テキストコンテンツ位置情報抽出手段により抽出された各位置情報が示す位置と、位置表現抽出手段により抽出された各位置表現の位置座標情報が示す位置との間の距離を求め、各非テキストコンテンツの各位置情報が示す位置から距離が最小となる位置表現を代表位置表現と判定して出力する代表位置表現判定工程とを有することを特徴とする。

【0013】

第3の本発明の位置表現抽出プログラムは、コンピュータが、少なくとも、複数の位置表現と、これら各位置表現の位置を特定する位置座標情報とを対応付けて格納する位置表現・位置座標情報対応テーブルを有し、コンピュータを、(1)入力された文書から1又は複数の位置表現を抽出し、位置表現・位置座標情報対応テーブルを参照して、抽出した各位置表現の位置座標情報を求める位置表現抽出手段、(2)文書から1又は複数の非テキストコンテンツを抽出し、各非テキストコンテンツに含まれている1又は複数の位置情報を抽出する非テキストコンテンツ位置情報抽出手段、(3)非テキストコンテンツ位置情報抽出手段により抽出された各位置情報が示す位置と、位置表現抽出手段により抽出された各位置表現の位置座標情報が示す位置との間の距離を求め、各非テキストコンテンツの各位置情報が示す位置から距離が最小となる位置表現を代表位置表現と判定して出力する代表位置表現判定手段として機能させるものである。

【発明の効果】

【0014】

本発明によれば、処理対象である入力文書の主題に強く関連する位置表現を判定することで、当該文書の主題に最適な位置表現を抽出することができる。

【発明を実施するための最良の形態】

【0015】

(A)第1の実施形態

以下、本発明の位置表現抽出装置、方法及びプログラムの第1の実施形態を図面を参照しながら詳細に説明する。

【0016】

第1の実施形態は、本発明の位置表現抽出装置、方法及びプログラムを利用して、入力されたウェブ文書(例えば、HTML(HyperText Markup Language)言語で記述されたウェブページなど)から主題に関連する位置表現を抽出する実施形態を例示して説明する。

【0017】

(A-1)第1の実施形態の構成

10

20

30

40

50

図 1 は、第 1 の実施形態の位置表現抽出装置 10 の内部構成を示す内部構成図である。図 1 において、第 1 の実施形態の位置表現抽出装置 10 は、位置表現抽出部 11、非テキストコンテンツ位置情報抽出部 12、代表位置表現判定部 13、位置表現・位置座標対応情報テーブル 14 を少なくとも有して構成される。

【0018】

なお、第 1 の実施形態の位置表現抽出装置 10 の機能はソフトウェア処理により実現されるものである。つまり、例えば、位置表現抽出部 11 と非テキストコンテンツ位置情報抽出部 12 と位置表現判定部 13 は、専用の制御デバイスあるいは、プログラムを実行するプロセッサと、そのプロセッサで実行するプログラムとデータを記憶する RAM (Random Access Memory)、ROM (Read Only Memory)、HDD (Hard Disk Drive) などの記憶デバイスを含むハードウェア資源により実現される。

10

【0019】

位置表現抽出部 11 は、入力されたウェブ文書 15 から 1 又は複数の位置表現を抽出し、位置表現・位置座標対応情報テーブル 14 を参照しながら、この抽出した各位置表現に対応する位置座標を求めるものである。また、位置表現抽出部 11 は、抽出した各位置表現の位置座標を代表位置表現判定部 13 に与えるものである。

【0020】

また、位置表現抽出部 11 による位置表現の抽出方法としては、種々の方法を適用することができるが、例えば、予め設定された複数の位置表現を格納する位置表現データベースを参照して、パターンマッチングにより、ウェブ文書 15 から位置表現を抽出する方法を適用することができる。

20

【0021】

そのため、例えばパターンマッチングにより位置表現を抽出する位置表現抽出部 11 の内部構成としては図 2 に示すような構成を備える。

【0022】

図 2 は、位置表現抽出部 11 の内部構成を示す内部構成図である。図 2 に示すように、位置表現抽出部 11 は、位置表現抽出実行部 111、位置表現データベース 112 を少なくとも有して構成される。

【0023】

位置表現抽出実行部 111 は、位置表現データベース 112 を参照しながら、入力されたウェブ文書 15 から、パターンマッチングにより位置表現を抽出するものである。パターンマッチング技術は種々の方法を適用することができ、例えば、位置表現データベース 112 に格納される位置表現と一致する文字列を抽出する。なお、完全一致しない場合でも、一部に位置表現を含む文字列を抽出するようにしてもよい。

30

【0024】

位置表現データベース 112 は、予め設定された複数の位置表現を格納するデータベースである。

【0025】

なお、図 2 では、位置表現データベース 112 を有する場合を示したが、位置表現・位置座標対応情報テーブル 14 を用いて、位置表現・位置座標対応情報テーブル 14 に格納されている複数の位置表現とパターンマッチングにより抽出するようにしてもよい。

40

【0026】

非テキストコンテンツ位置情報抽出部 12 は、入力されたウェブ文書 15 を構成する非テキストコンテンツ (例えば、静止画像、動画像、音声など) に含まれる 1 又は複数の位置情報を抽出し、この抽出した各位置情報の位置座標を求めるものである。また、非テキストコンテンツ位置情報抽出部 12 は、抽出した各位置表現の位置情報を代表位置表現判定部 13 に与えるものである。

【0027】

ここで、非テキストコンテンツとしては、例えば、ウェブ文書に含まれる静止画像、動画像、音声などが該当する。また、これら非テキストコンテンツから位置情報を抽出する

50

方法としては、その非テキストコンテンツの種類に応じて、種々の方法を適用することができる。

【 0 0 2 8 】

図 3 は、非テキストコンテンツが静止画像である場合、この静止画像に含まれる位置情報を抽出するときの非テキストコンテンツ位置情報抽出部 1 2 の内部構成を示す内部構成図である。

【 0 0 2 9 】

図 3 に示すように、非テキストコンテンツ位置情報抽出部 1 2 は、メタデータ検出部 1 2 1、位置情報抽出部 1 2 2 を少なくとも有するものである。

【 0 0 3 0 】

メタデータ検出部 1 2 1 は、入力されたウェブ文書 1 5 に含まれている静止画像データから、当該静止画像データに含まれているメタデータを検出するものである。

【 0 0 3 1 】

ここで、メタ情報について説明する。例えば、デジタルカメラなどで撮影された画像データには、当該静止画像に関連する情報が例えば E x i f (Exchangeable Image File Format) 形式のメタデータとして付与されている。そこで、メタデータ検出部 1 2 1 は、静止画像データから、このメタデータを抽出するようにする。

【 0 0 3 2 】

なお、非テキストコンテンツの種類は静止画像に限定されるものではなく、また静止画像のメタデータの規格形式も特に限定されるものではなく、種々のものを広く適用することができる。

【 0 0 3 3 】

位置情報抽出部 1 2 2 は、メタデータ検出部 1 2 1 により検出されたメタデータを受け取り、このメタデータに含まれる 1 又は複数の位置情報を抽出するものである。

【 0 0 3 4 】

代表位置表現判定部 1 3 は、位置表現抽出部 1 1 により抽出された位置表現の位置情報と、非テキストコンテンツ位置情報抽出部 1 2 により抽出された位置表現の位置情報とに基づいて、ウェブ文書 1 5 の主題に関連する位置表現を求め、この位置表現を代表位置表現リスト 1 6 として出力するものである。

【 0 0 3 5 】

ここで、代表位置表現判定部 1 3 による代表位置表現の求める方法としては、非テキストコンテンツ位置情報抽出部 1 2 により抽出された位置表現の位置情報を基準とする。そして、位置表現抽出部 1 1 からの位置表現の位置情報の中から、上記基準点に最も近いものを代表位置表現とする方法を適用することができる。これは、ウェブ文書 1 5 に含まれている非テキストコンテンツの位置情報がウェブ文書 1 5 の主題に強く関連する場合が多いからである。

【 0 0 3 6 】

位置表現・位置座標対応情報テーブル 1 4 は、複数の位置表現と各位置表現の位置座標情報とを対応付けたテーブルであり、例えば、HDD、ROM等の記憶デバイスに記憶されるものである。

【 0 0 3 7 】

図 4 は、位置表現・位置座標対応情報テーブル 1 4 の構成例を示す構成図である。図 4 に示すように、位置表現・位置座標対応情報テーブル 1 4 は、「位置表現 1 4 1」、「緯度 1 4 2」、「経度 1 4 3」を管理項目として有する。

【 0 0 3 8 】

「位置表現 1 4 1」は、位置を表現するものを格納し、例えば、住所、施設名、建物名、スポット名などが該当する。「緯度 1 4 2」、「経度 1 4 3」は、対応する位置表現の位置座標情報とするものである。

【 0 0 3 9 】

なお、ここで説明した第 1 の実施形態の位置表現抽出装置 1 0 の構成は、機能別に説明

10

20

30

40

50

したものであり、実現するハードウェアの物理構成が各部分毎に明確に分離され、独立に用意される必要はない。

【 0 0 4 0 】

例えば位置表現抽出部 1 1 と非テキストコンテンツ位置情報抽出部 1 2 と位置表現判定部 1 3 を実現するプロセッサと R A M、R O M は共通のものであってもよく、さらにプログラムを格納する H D D は位置表現・位置座標対応テーブル 1 4 の H D D と共通のものであってもよく、さらには他の機能を実現する装置の一部を利用してよい。また、本装置の構成する部分の一部はネットワークで接続された他の場所に分散配置してもよい。

【 0 0 4 1 】

(A - 2) 第 1 の実施形態の動作

10

次に、第 1 の実施形態の位置表現抽出装置 1 0 による位置表現抽出処理について図面を参照しながら説明する。

【 0 0 4 2 】

図 5 は、この実施形態の動作説明で用いる、処理対象のウェブ文書 1 5 の内容例を示すものである。

【 0 0 4 3 】

図 5 において、ウェブ文書 1 5 は、テキスト文書 1 5 1、非テキストコンテンツ 1 5 2、広告データ 1 5 3、を有して構成されるものとする。

【 0 0 4 4 】

非テキストコンテンツ 1 5 2 は、例えばデジタルカメラ等により撮像された静止画像であり、この静止画像には位置情報を含む E x i f 形式のメタデータが付与されているものとする。

20

【 0 0 4 5 】

広告データ 1 5 3 は、例えば、ユーザが作成したウェブ文書 1 5 に付与された広告データであり、広告主の住所が付与されているものとする。

【 0 0 4 6 】

この広告データ 1 5 3 は、テキスト文書、静止画像・動画像などの非テキストコンテンツなどを組み合わせて構成されるものであってよい。また、第 1 の実施形態の位置表現抽出装置 1 0 は、この広告データについても、テキスト文書と非テキストコンテンツとに応じた位置情報の抽出が可能である。

30

【 0 0 4 7 】

図 5 に示す例では、広告データ 1 5 3 は、テキスト文書として、「A 電気工業」、「大阪府大阪市中央区本町 X - Y - Z」を含むものとする。

【 0 0 4 8 】

図 6 は、第 1 の実施形態の位置表現抽出処理の動作を示すフローチャートである。

【 0 0 4 9 】

まず、処理対象のウェブ文書 1 5 が位置表現抽出装置 1 0 に入力されると、位置表現抽出部 1 1 により、ウェブ文書 1 5 からすべての位置表現が抽出される (ステップ S 1 0 1)。

【 0 0 5 0 】

40

例えば、図 5 に示すウェブ文書 1 5 が入力されると、位置表現抽出部 1 1 は、ウェブ文書 1 5 から、「大阪府大阪市中央区 X - Y - Z」、「大阪城」、「大阪ドーム」を抽出する。

【 0 0 5 1 】

ステップ 1 0 2 では、ステップ 1 0 1 において抽出した位置表現の数を調べ、1 個も取得できなかった場合は結果なしとして処理を終了する (ステップ S 1 0 2)。

【 0 0 5 2 】

一方、1 個以上の位置表現を取得できた場合には、ステップ S 1 0 3 に移行する。この実施形態の動作説明の例では、3 個の位置表現を取得できたので、ステップ S 1 0 3 に移行する。

50

【 0 0 5 3 】

次に、位置表現抽出部 1 1 では、位置表現・位置座標対応情報テーブル 1 4 を参照して、各位置表現に対応する位置座標情報（緯度情報、経度情報）を取得する（ステップ S 1 0 3 ）。

【 0 0 5 4 】

例えば、「大阪府大阪市中央区本町 X - Y - Z 」の位置表現に対して「緯度 3 4 . 6 8 4 1 経度 1 3 5 . 5 0 4 0 」が取得され、「大阪ドーム」の位置表現に対して「緯度 3 4 . 6 6 9 4 経度 1 3 5 . 4 7 6 0 」が取得され、「大阪城」の位置表現に対して「緯度 3 4 . 6 8 7 4 経度 1 3 5 . 5 2 5 9 」が取得される。

【 0 0 5 5 】

また、非テキストコンテンツ位置情報抽出部 1 2 では、ウェブ文書 1 5 からすべての非テキストコンテンツが抽出される（ステップ S 1 0 4 ）。

【 0 0 5 6 】

例えば、ウェブ文書 1 5 には非テキストコンテンツ 1 5 2 として「大阪城の写真」が含まれているので、非テキストコンテンツ位置情報抽出部 1 2 はこの「大阪城の写真」を抽出する。

【 0 0 5 7 】

ステップ S 1 0 5 では、ステップ S 1 0 4 において抽出された非テキストコンテンツの数を調べ、1 個も取得できなかった場合にはステップ S 1 0 8 に移行し、代表位置表現判定部 1 3 は、位置表現抽出部 1 1 からのすべての位置表現が選択されたものとみなされ、すべての位置表現をリストとして出力し、処理を終了する（ステップ S 1 0 8 ）。

【 0 0 5 8 】

一方、1 個以上の非テキストコンテンツが抽出されると、ステップ S 1 0 6 に移行する。この実施形態の動作説明では、1 個の非テキストコンテンツ 1 5 2 が抽出されたので、ステップ S 1 0 8 に移行する。

【 0 0 5 9 】

次に、非テキストコンテンツでは、抽出された非テキストコンテンツに含まれるメタデータを抽出し、このメタデータから位置情報を抽出する（ステップ S 1 0 6 ）。

【 0 0 6 0 】

このとき、複数個の非テキストコンテンツがある場合には、複数個の非テキストコンテンツのそれぞれの位置情報を抽出し、複数の位置情報を取得するようにしてもよい。さらに、1 個の非テキストコンテンツのメタデータから複数の位置情報を取得するようにしてもよい。例えば、E x i f 形式のメタデータからは、撮影場所と撮影対象の場所の 2 個の位置情報が得られることがある。この場合には、それぞれの位置情報を抽出する。

【 0 0 6 1 】

また、メタデータから取得する位置情報が緯度情報・経度情報ではない場合もある。この場合、例えばメタデータの位置情報として「位置表現」が格納されているのであれば、非テキストコンテンツ位置情報抽出部 1 2 は、位置表現・位置座標対応情報テーブル 1 2 を用いて緯度情報・経度情報を求めるようにしてもよい。

【 0 0 6 2 】

例えば、この実施形態の動作説明では、非テキストコンテンツ 1 5 2 である「大阪城の写真」のメタデータには、「緯度 3 4 . 6 8 8 0 経度 1 3 5 . 5 3 4 4 」の位置情報が付加されているものとする。そこで、非テキストコンテンツ位置情報抽出部 1 2 は、この「緯度 3 4 . 6 8 8 0 経度 1 3 5 . 5 3 4 4 」を抽出する。

【 0 0 6 3 】

ステップ S 1 0 1 において位置表現抽出部 1 1 により抽出された位置表現は、代表位置表現判定部 1 3 に与えられる。また、非テキストコンテンツ位置情報抽出部 1 2 により取得された位置情報も、代表位置表現判定部 1 3 に与えられる。

【 0 0 6 4 】

そして、代表位置表現判定部 1 3 により、位置表現抽出部 1 1 により抽出された位置表

10

20

30

40

50

現の中から、ウェブ文書15の主題に強く関連する位置表現(代表位置表現)が選択される(ステップS107)。

【0065】

このとき、代表位置表現判定部13は、非テキストコンテンツ位置情報抽出部12からの位置情報を基準点として、代表位置表現を選択する。

【0066】

このように、非テキストコンテンツの位置情報を基準として利用することは、例えば、画像がウェブ文書15に含まれている場合、その画像がウェブ文書15の主題に強く関連している可能性が非常に高いからである。

【0067】

これにより、ウェブ文書15の主題に最も強く関連する位置表現を判定し、この位置表現を代表位置表現として出力することができる(ステップS108)。その結果、当該ウェブ文書15に最適な位置表現を出力することができる。

【0068】

例えば、非テキストコンテンツ152の位置情報が「緯度34.6880 経度135.5344」であるから、代表位置表現判定部13は、この緯度情報・経度情報を基準点とする。

【0069】

そして、代表位置表現判定部13は、位置表現「大阪府大阪市中央区X-Y-Z」、「大阪城」、「大阪ドーム」のそれぞれの緯度情報・経度情報を用いて、基準点に最も近い距離の位置表現を選択する。

【0070】

このとき、代表位置表現判定部13は、種々の方法を適用して、各位置表現の位置座標情報と基準点との間の距離を求めることができる。例えば、球面三角法などを用いて厳密に計算しても良いし、また例えば直角三角形の公式を用いた簡単なものでも良い。

【0071】

そうすると、代表位置表現判定部13は、基準点「緯度34.6880 経度135.5344」に対して、最も距離が近いものは「大阪城」(緯度34.6874 経度135.5259)であると判定し、この位置表現「大阪城」を代表位置表現として選択し、選択された代表位置表現リスト16を出力して終了する。

【0072】

(A-3)第1の実施形態の効果

以上のように、第1の実施形態によれば、対象とするウェブ文書に複数の位置表現が存在する場合において、非テキストコンテンツから求めた位置情報を基準として、この基準から最も距離が近い位置表現を出力することで、対象のウェブ文書の主題に関連の強い位置表現を選択及び出力することが可能となる。

【0073】

(B)他の実施形態

第1の実施形態では、処理対象として「ウェブ文書」を例示して説明したが、一般的なHTML言語で記述されたウェブページに限定されるものではない。また、例えば、RDFやPDFなどいわゆるリッチメディアと呼ばれる、テキストとそれ以外のさまざまなメディア情報(非テキストコンテンツ)を併せ持つ文書に適用することができる。

【0074】

第1の実施形態において、代表位置表現判定部13は、非テキストコンテンツの位置情報が示す位置との間の距離が最も近い位置表現を選択し、この1個の位置表現を出力する場合を例示した。しかし、基準点からの距離が近い順に複数個の位置表現を並べ替えて、これら複数の位置表現を出力してもよい。これにより、文書の主題の関連度に応じた優先順位で位置表記を出力することができる。また、この場合、予め定めた個数の位置表現を出力するようにしてもよい。

【0075】

10

20

30

40

50

さらに、主題とは無関係の非テキストコンテンツによる悪影響を抑えるために、すべての位置表現又は一部の位置表現と基準点との距離が、予め設定された閾値よりも大きな距離の位置表現しか存在しない場合は、その非テキストコンテンツに対する代表位置表現をあえて出力しないようにしてもよい。

【 0 0 7 6 】

第 1 の実施形態の説明においては、ウェブ文書に含まれる非テキストコンテンツの個数は 1 個であったが、複数の非テキストコンテンツが含まれるウェブ文書にも適用できる。この場合、各非テキストコンテンツに対応した代表位置表現が選択され、複数出力される。

【 0 0 7 7 】

非テキストコンテンツの位置情報が示す位置が複数ある場合、各位置情報の位置毎に、代表位置表現を出力するようにしてもよい。

【 0 0 7 8 】

第 1 の実施形態では、非テキストコンテンツが、例えばデジタルカメラなどで撮影した静止画像の場合を例示して説明したが、第 1 の実施形態でも示したが、動画画像や音声などの場合にも適用できる。この場合、非テキストコンテンツ位置情報抽出部は、動画画像データや音声データの設定データを検出し、この設定データから位置情報を抽出することで実現することができる。

【 図面の簡単な説明 】

【 0 0 7 9 】

【 図 1 】 第 1 の実施形態の位置表現抽出装置の内部構成を示す内部構成図である。

【 図 2 】 第 1 の実施形態の位置表現抽出部の内部構成を示す内部構成図である。

【 図 3 】 第 1 の実施形態の非テキストコンテンツ位置情報抽出部の内部構成を示す内部構成図である。

【 図 4 】 第 1 の実施形態の位置表現・位置座標対応情報テーブルの構成を示す構成図である。

【 図 5 】 第 1 の実施形態のウェブ文書の内容を示す説明図である。

【 図 6 】 第 1 の実施形態の位置表現抽出処理の動作を示すフローチャートである。

【 符号の説明 】

【 0 0 8 0 】

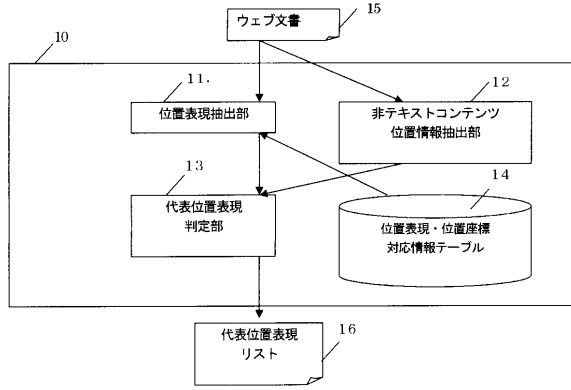
1 0 ... 位置表現抽出装置、 1 1 ... 位置表現抽出部、 1 2 ... 非テキストコンテンツ位置情報抽出部、 1 3 ... 代表位置表現判定部、 1 4 ... 位置表現・位置座標対応情報テーブル。

10

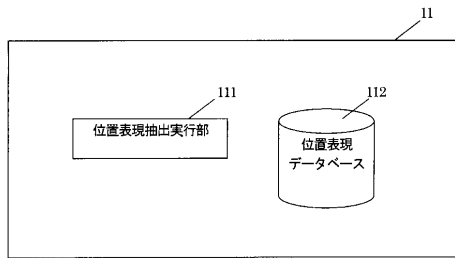
20

30

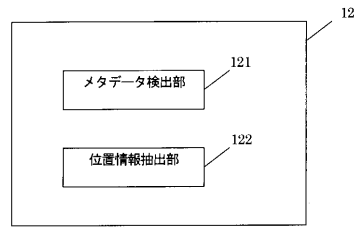
【図1】



【図2】



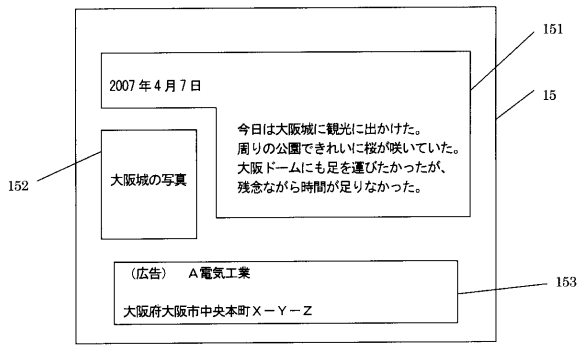
【図3】



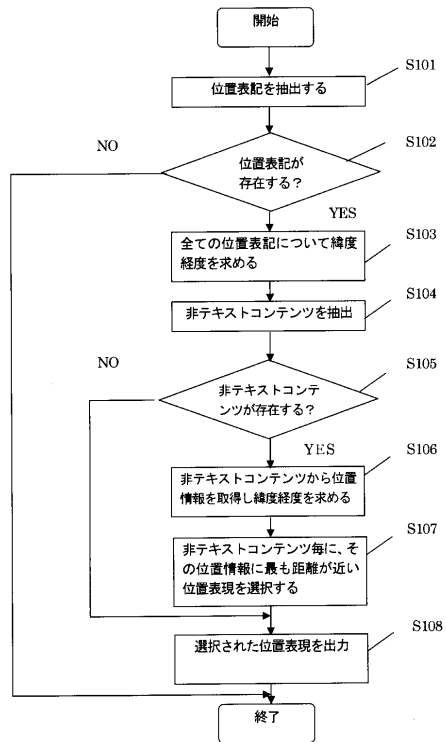
【図4】

位置表現	緯度	経度
大阪府大阪市中央区本町X-Y-Z	34.6841	135.5040
...
大阪ドーム	34.6694	135.4760
大阪城	34.6874	135.5259
...

【図5】



【図6】



フロントページの続き

- (56)参考文献 特開2004-280659(JP,A)
特開2007-129434(JP,A)
特開2007-052581(JP,A)
李龍, 地域ウェブ情報検索のための2次元領域質問処理法, 情報処理学会研究報告, 日本, 社団法人情報処理学会, 2003年 7月17日, Vol.2003, No.71, pp.413-420
山田直治, 携帯端末を考慮した詳細度と地域的スコープによるWEBページ分類, FIT2002 情報科学技術フォーラム 情報技術レターズ, 日本, 社団法人電子情報通信学会, 2002年 9月13日, Vol.1, pp.53-54

- (58)調査した分野(Int.Cl., DB名)
G06F 17/30