(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5462591号
(P5462591)

(45)発行日　平成26年4月2日(2014.4.2)　(24)登録日　平成26年1月24日(2014.1.24)

(51)Int.Cl.		F I		
G06Q 30/02	(2012.01)	G06Q	30/02	１５０
G06F 13/00	(2006.01)	G06F	13/00	５４０Ｐ
G06F 17/30	(2006.01)	G06F	17/30	２１０Ａ

請求項の数 11　(全 25 頁)

(21)出願番号　特願2009-250646 (P2009-250646)
(22)出願日　平成21年10月30日 (2009.10.30)
(65)公開番号　特開2011-96078 (P2011-96078A)
(43)公開日　平成23年5月12日 (2011.5.12)
審査請求日　平成23年11月7日 (2011.11.7)
審判番号　不服2012-21010 (P2012-21010/J1)
審判請求日　平成24年10月24日 (2012.10.24)

早期審査対象出願

(73)特許権者　399037405
楽天株式会社
東京都品川区東品川四丁目１２番３号
(74)代理人　110000958
特許業務法人　インテクト国際特許事務所
(72)発明者　岩淵　志学
東京都品川区東品川四丁目１２番３号　楽天株式会社内

合議体
審判長　手島　聖治
審判官　須田　勝巳
審判官　清田　健一

最終頁に続く

(54)【発明の名称】特有コンテンツ判定装置、特有コンテンツ判定方法、特有コンテンツ判定プログラム及び関連コンテンツ挿入装置

(57)【特許請求の範囲】
【請求項１】
所定のサイトに含まれる複数のＷｅｂページのうち、指定されたＷｅｂページを構成しているコンテンツを抽出する抽出手段と、
前記複数のＷｅｂページのうち、前記指定されたＷｅｂページを構成している各コンテンツが他のＷｅｂページで用いられる頻度をカウントする計算手段と、
前記指定されたＷｅｂページを構成しているコンテンツのうち、他のＷｅｂページで用いられる頻度が所定値以下のコンテンツを当該指定されたＷｅｂページに特有のコンテンツであると判定する判定手段と、
を備えることを特徴とする特有コンテンツ判定装置。
【請求項２】
請求項１に記載の特有コンテンツ判定装置において、
前記抽出手段は、１つ以上のコンテンツで構成されたコンテンツグループの単位で、Ｗｅｂページを構成しているコンテンツを抽出し、
前記計算手段は、前記指定されたＷｅｂページを構成しているコンテンツグループが他のＷｅｂページで用いられる頻度をカウントし、
前記判定手段は、前記指定されたＷｅｂページを構成しているコンテンツグループのうち、他のＷｅｂページで用いられる頻度が所定値以下のコンテンツグループを当該指定されたＷｅｂページに特有のコンテンツグループであると判定することを特徴とする特有コンテンツ判定装置。

【請求項３】
請求項２に記載の特有コンテンツ判定装置において、
前記抽出手段は、所定のマークアップ言語で記述され、Ｗｅｂページを構成するコンテンツを示すドキュメントデータに基づいて、コンテンツグループを抽出することを特徴とする特有コンテンツ判定装置。
【請求項４】
請求項３に記載の特有コンテンツ判定装置において、
前記抽出手段は、前記コンテンツを示すドキュメントデータにおいて予め定められたタグに基づいてコンテンツグループを定めることを特徴とする特有コンテンツ判定装置。
【請求項５】
請求項１乃至４の何れか１項に記載の特有コンテンツ判定装置において、
前記抽出手段は、投稿された記事が掲載されるＷｅｂページから前記記事に対して投稿されたコメントを抽出し、
前記抽出された各コメントを、コメントが示す内容別に分類する分類手段と、
出現頻度の閾値を設定する設定手段であり、前記コメントが分類された前記内容の数が多いほど前記閾値を小さくする設定手段と、
を更に備え、
前記計算手段は、前記コメントが分類された各前記内容のＷｅｂページにおける出現頻度を計算し、
前記判定手段は、前記計算手段により計算された出現頻度が前記設定された閾値以下である前記内容を前記Ｗｅｂページに特有の内容であると判定することを特徴とする特有コンテンツ判定装置。
【請求項６】
所定のサイトに含まれる複数のＷｅｂページのうち、指定されたＷｅｂページを構成しているコンテンツを抽出する抽出工程と、
前記複数のＷｅｂページのうち、前記指定されたＷｅｂページを構成している各コンテンツが他のＷｅｂページで用いられる頻度をカウントする計算工程と、
前記指定されたＷｅｂページを構成しているコンテンツのうち、他のＷｅｂページで用いられる頻度が所定値以下のコンテンツを当該指定されたＷｅｂページに特有のコンテンツであると判定する判定工程と、
を有することを特徴とする特有コンテンツ判定方法。
【請求項７】
請求項６に記載の特有コンテンツ判定方法において、
前記抽出工程は、投稿された記事が掲載されるＷｅｂページから前記記事に対して投稿されたコメントを抽出し、
前記抽出された各コメントを、コメントが示す内容別に分類する分類工程と、
出現頻度の閾値を設定する設定工程であり、前記コメントが分類された前記内容の数が多いほど前記閾値を小さくする設定工程と、
を更に含み、
前記計算工程は、前記コメントが分類された各前記内容のＷｅｂページにおける出現頻度を計算し、
前記判定工程は、前記計算工程により計算された出現頻度が前記設定された閾値以下である前記内容を前記Ｗｅｂページに特有の内容であると判定することを特徴とする特有コンテンツ判定方法。
【請求項８】
コンピュータを、
所定のサイトに含まれる複数のＷｅｂページのうち、指定されたＷｅｂページを構成しているコンテンツを抽出する抽出手段、
前記複数のＷｅｂページのうち、前記指定されたＷｅｂページを構成している各コンテンツが他のＷｅｂページで用いられる頻度をカウントする計算手段、及び、

前記指定されたＷｅｂページを構成しているコンテンツのうち、他のＷｅｂページで用いられる頻度が所定値以下のコンテンツを当該指定されたＷｅｂページに特有のコンテンツであると判定する判定手段、
として機能させることを特徴とする特有コンテンツ判定プログラム。
【請求項９】
請求項８に記載の特有コンテンツ判定プログラムにおいて、
前記抽出手段は、投稿された記事が掲載されるＷｅｂページから前記記事に対して投稿されたコメントを抽出し、
前記コンピュータを、
前記抽出された各コメントを、コメントが示す内容別に分類する分類手段、及び、
出現頻度の閾値を設定する設定手段であり、前記コメントが分類された前記内容の数が多いほど前記閾値を小さくする設定手段、
として更に機能させ、
前記計算手段は、前記コメントが分類された各前記内容のＷｅｂページにおける出現頻度を計算し、
前記判定手段は、前記計算手段により計算された出現頻度が前記設定された閾値以下である前記内容を前記Ｗｅｂページに特有の内容であると判定することを特徴とする特有コンテンツ判定プログラム。
【請求項１０】
請求項１乃至５の何れか１項に記載の特有コンテンツ判定装置と、
前記特有コンテンツ判定装置により特有のコンテンツであると判定されたコンテンツに関連する関連コンテンツを、前記指定されたＷｅｂページに挿入する挿入手段と、
を備えることを特徴とする関連コンテンツ挿入装置。
【請求項１１】
請求項１０に記載の関連コンテンツ挿入装置において、
前記特有コンテンツ判定装置は、前記指定されたＷｅｂページを構成しているコンテンツとして、投稿された記事のテキストデータが含まれている場合に、当該テキストデータを当該Ｗｅｂページに特有のコンテンツであると判定し、
前記特有コンテンツ判定装置により特有のコンテンツであると判定された記事のテキストデータから前記指定されたＷｅｂページの特徴語を抽出する特徴語抽出手段と、
それぞれ語に関連付けて記憶手段に記憶された複数のコンテンツの中から、前記抽出された特徴語に関連するコンテンツを前記関連コンテンツとして選択する選択手段と、
を更に備え、
前記挿入手段は、前記選択された関連コンテンツを、前記指定されたＷｅｂページに挿入することを特徴とする関連コンテンツ挿入装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、Ｗｅｂページを構成するコンテンツを抽出する技術分野に関する。
【背景技術】
【０００２】
従来、Ｗｅｂサイト上に公開されているＷｅｂページを構成している素材たるコンテンツを取得し、取得したコンテンツに基づいて新たなコンテンツを生成する技術が知られている。例えば、非特許文献１には、ユーザにより画像データのＵＲＬが指定されると、当該ＵＲＬに対応する画像データをＷｅｂ上から取得し、取得した画像データに基づいてバナーを自動作成する技術が開示されている。
【先行技術文献】
【非特許文献】
【０００３】
【非特許文献１】“バナー自動作成”、[online]、[平成21年10月21日検索]、インターネ

ット<ＵＲＬ：http://hyperbannermaker.com/>
【発明の概要】
【発明が解決しようとする課題】
【０００４】
Ｗｅｂサイトを構成する各Ｗｅｂページには、そのＷｅｂサイトの目的に沿った内容が掲載される。そのため、Ｗｅｂサイトを構成する各Ｗｅｂページの内容は、基本的には互いに関連性を有しているのであるが、夫々何らかの特徴を有してる場合がある。そして、そのＷｅｂページの内容を決める要因が、Ｗｅｂページを構成しているコンテンツ（例えば、テキストデータ、画像データ等）の内容である。従って、Ｗｅｂページを構成しているコンテンツの中で、そのＷｅｂページを特徴付けるコンテンツ、すなわち、そのＷｅｂページ特有のコンテンツが存在することがある。
【０００５】
非特許文献１に記載の技術は、Ｗｅｂページに特有のコンテンツを抽出するものであるが、自動的に抽出するものではなく、ユーザが手作業でコンテンツを指定しなければならず、該Ｗｅｂページに特有のコンテンツを容易に抽出することはできない。そのため、どのコンテンツがＷｅｂページ特有のコンテンツであるかをユーザが判断することができない場合、又はユーザの嗜好によって好みのコンテンツが偏ってしまう場合など、Ｗｅｂページに特有のコンテンツを適格に抽出することができない。また、対象とするＷｅｂページのページ数が多いと、ユーザの作業が甚大となってしまう問題があった。
【０００６】
また、例えば、ＨＴＭＬ（HyperText Markup Language）文書のタグの記述に基づいて、画像だけ又はテキストだけといったように、特定種類のコンテンツを全て抽出することは可能である。しかしながら、抽出されたコンテンツの中には、Ｗｅｂページ特有のものではないありふれたコンテンツも含まれるので、抽出結果としての信頼性が低く、そのため抽出結果からユーザが特有のコンテンツを探さなければならなかった。
【０００７】
本発明は以上の点に鑑みてなされたものであり、Ｗｅｂページを構成しているコンテンツの中からのそのＷｅｂページ特有のコンテンツを容易に抽出することができる特有コンテンツ判定装置、特有コンテンツ判定方法、特有コンテンツ判定プログラム等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】
【０００８】
上記課題を解決するために、請求項１に記載の発明は、所定のサイトに含まれる複数のＷｅｂページのうち、指定されたＷｅｂページを構成しているコンテンツを抽出する抽出手段と、前記複数のＷｅｂページのうち、前記指定されたＷｅｂページを構成している各コンテンツが他のＷｅｂページで用いられる頻度をカウントする計算手段と、前記指定されたＷｅｂページを構成しているコンテンツのうち、他のＷｅｂページで用いられる頻度が所定値以下のコンテンツを当該指定されたＷｅｂページに特有のコンテンツであると判定する判定手段と、を備えることを特徴とする。
【０００９】
この発明によれば、所定のサイトに含まれる複数のＷｅｂページのうち、指定されたＷｅｂページを構成している各コンテンツが複数のＷｅｂページで用いられる頻度が夫々カウントされる。他のＷｅｂページで用いられる頻度が小さいコンテンツであるほど、指定されたＷｅｂページ以外にはあまり出現しないコンテンツである。そこで、他のＷｅｂページで用いられる頻度が所定値以下であるかを判定することで、当該条件を満たす全てのコンテンツが、指定されたＷｅｂページに特有のコンテンツであると特定される。よって、Ｗｅｂページに特有のコンテンツを容易に抽出することができる。
【００１２】
請求項２に記載の発明は、請求項１に記載の特有コンテンツ判定装置において、前記抽出手段は、１つ以上のコンテンツで構成されたコンテンツグループの単位で、Ｗｅｂペー

ジを構成しているコンテンツを抽出し、前記計算手段は、前記指定されたＷｅｂページを構成しているコンテンツグループが他のＷｅｂページで用いられる頻度をカウントし、前記判定手段は、前記指定されたＷｅｂページを構成しているコンテンツグループのうち、他のＷｅｂページで用いられる頻度が所定値以下のコンテンツグループを当該指定されたＷｅｂページに特有のコンテンツグループであると判定することを特徴とする。

【００１３】

この発明によれば、コンテンツグループの単位でＷｅｂページに特有のコンテンツが判断されるので、例えば、Ｗｅｂページ上において或るまとまりをもって表示されていたり、互いに関連性を有しているようなコンテンツをコンテンツグループとしたときに、Ｗｅｂページに特有のコンテンツとなるものを抽出することができる。

【００１４】

請求項３に記載の発明は、請求項２に記載の特有コンテンツ判定装置において、前記抽出手段は、所定のマークアップ言語で記述され、Ｗｅｂページを構成するコンテンツを示すドキュメントデータに基づいて、コンテンツグループを抽出することを特徴とする。

【００１５】

この発明によれば、Ｗｅｂページを構成するコンテンツを示すドキュメントデータに基づいてコンテンツグループが抽出されるので、適格にコンテンツグループを抽出することができる。

【００１６】

請求項４に記載の発明は、請求項３に記載の特有コンテンツ判定装置において、前記抽出手段は、前記コンテンツを示すドキュメントデータにおいて予め定められたタグに基づいてコンテンツグループを定めることを特徴とする。

【００１７】

この発明によれば、予め定められたタグに基づいてコンテンツグループが抽出されるので、Ｗｅｂページに特有のコンテンツと、特有ではないコンテンツとが夫々予め定められたタグでグループ化されている場合に、Ｗｅｂページに特有のコンテンツを判断する精度を上げることができる。

【００１８】

請求項５に記載の発明は、請求項１乃至４の何れか１項に記載の特有コンテンツ判定装置において、前記抽出手段は、投稿された記事が掲載されるＷｅｂページから前記記事に対して投稿されたコメントを抽出し、前記抽出された各コメントを、コメントが示す内容別に分類する分類手段と、出現頻度の閾値を設定する設定手段であり、前記コメントが分類された前記内容の数が多いほど前記閾値を小さくする設定手段と、を更に備え、前記計算手段は、前記コメントが分類された各前記内容のＷｅｂページにおける出現頻度を計算し、前記判定手段は、前記計算手段により計算された出現頻度が前記設定された閾値以下である前記内容を前記Ｗｅｂページに特有の内容であると判定することを特徴とする。

請求項６に記載の発明は、所定のサイトに含まれる複数のＷｅｂページのうち、指定されたＷｅｂページを構成しているコンテンツを抽出する抽出工程と、前記複数のＷｅｂページのうち、前記指定されたＷｅｂページを構成している各コンテンツが他のＷｅｂページで用いられる頻度をカウントする計算工程と、前記指定されたＷｅｂページを構成しているコンテンツのうち、他のＷｅｂページで用いられる頻度が所定値以下のコンテンツを当該指定されたＷｅｂページに特有のコンテンツであると判定する判定工程と、を有することを特徴とする。

請求項７に記載の発明は請求項６に記載の特有コンテンツ判定方法において、前記抽出工程は、投稿された記事が掲載されるＷｅｂページから前記記事に対して投稿されたコメントを抽出し、前記抽出された各コメントを、コメントが示す内容別に分類する分類工程と、出現頻度の閾値を設定する設定工程であり、前記コメントが分類された前記内容の数が多いほど前記閾値を小さくする設定工程と、を更に含み、前記計算工程は、前記コメントが分類された各前記内容のＷｅｂページにおける出現頻度を計算し、前記判定工程は、前記計算工程により計算された出現頻度が前記設定された閾値以下である前記内容を前記

Ｗｅｂページに特有の内容であると判定することを特徴とする。

【００１９】

請求項８に記載の発明は、コンピュータを、所定のサイトに含まれる複数のＷｅｂページのうち、指定されたＷｅｂページを構成しているコンテンツを抽出する抽出手段、前記複数のＷｅｂページのうち、前記指定されたＷｅｂページを構成している各コンテンツが他のＷｅｂページで用いられる頻度をカウントする計算手段、及び、前記指定されたＷｅｂページを構成しているコンテンツのうち、他のＷｅｂページで用いられる頻度が所定値以下のコンテンツを当該指定されたＷｅｂページに特有のコンテンツであると判定する判定手段、として機能させることを特徴とする。

請求項９に記載の発明は、請求項８に記載の特有コンテンツ判定プログラムにおいて、前記抽出手段は、投稿された記事が掲載されるＷｅｂページから前記記事に対して投稿されたコメントを抽出し、前記コンピュータを、前記抽出された各コメントを、コメントが示す内容別に分類する分類手段、及び、出現頻度の閾値を設定する設定手段であり、前記コメントが分類された前記内容の数が多いほど前記閾値を小さくする設定手段、として更に機能させ、前記計算手段は、前記コメントが分類された各前記内容のＷｅｂページにおける出現頻度を計算し、前記判定手段は、前記計算手段により計算された出現頻度が前記設定された閾値以下である前記内容を前記Ｗｅｂページに特有の内容であると判定することを特徴とする。

【００２０】

請求項１１に記載の発明は、請求項１乃至６の何れか１項に記載の特有コンテンツ判定装置と、前記特有コンテンツ判定装置により特有のコンテンツであると判定されたコンテンツに関連する関連コンテンツを、前記指定されたＷｅｂページに挿入する挿入手段と、を備えることを特徴とする。

【００２１】

この発明によれば、特有のコンテンツであると判定されたコンテンツに関連するコンテンツが、指定されたＷｅｂページに挿入されるので、Ｗｅｂページの特徴と関連する情報を当該Ｗｅｂページに追加することができる。

【００２２】

請求項１２に記載の発明は、請求項１１に記載の関連コンテンツ挿入装置において、前記特有コンテンツ判定装置は、前記指定されたＷｅｂページを構成しているコンテンツとして、投稿された記事のテキストデータが含まれている場合に、当該テキストデータを当該Ｗｅｂページに特有のコンテンツであると判定し、前記特有コンテンツ判定装置により特有のコンテンツであると判定された記事のテキストデータから前記指定されたＷｅｂページの特徴語を抽出する特徴語抽出手段と、それぞれ語に関連付けて記憶手段に記憶された複数のコンテンツの中から、前記抽出された特徴語に関連するコンテンツを前記関連コンテンツとして選択する選択手段と、を更に備え、前記挿入手段は、前記選択された関連コンテンツを、前記指定されたＷｅｂページに挿入することを特徴とする。

【００２３】

この発明によれば、各記事のテキストデータにその記事特有の内容が含まれているのであれば、特有コンテンツ判定装置により各記事のテキストデータを抽出することができる。これにより、Ｗｅｂページに掲載されている記事の内容に関連する情報を当該Ｗｅｂページに追加することができる。

【発明の効果】

【００２４】

本発明によれば、他のＷｅｂページで用いられる頻度が小さいコンテンツであるほど、指定されたＷｅｂページ以外に前記複数のＷｅｂページ上にあまり出現しないコンテンツであるので、他のＷｅｂページで用いられる頻度が所定値以下であるかを判定することで、当該条件を満たす全てのコンテンツが、指定されたＷｅｂページに特有のコンテンツであると特定される。よって、Ｗｅｂページに特有のコンテンツを容易に抽出することができる。

【図面の簡単な説明】
【００２５】
【図１】一実施形態に係るブログシステムＳの概要構成の一例を示す図である。
【図２】一実施形態に係るブログサーバ１の概要構成の一例を示すブロック図である。
【図３】ブロガーが指定されてからブログページに広告コンテンツが挿入されるまでの処理の概要を示す図である。
【図４】Ｗｅｂページの構成例を示す図である。
【図５】ＨＴＭＬ文書から生成されたＤＯＭツリーの一例を示す図である。
【図６】記憶部１５に記憶されたコンテンツブロック対応情報の内容の一例を示す図である。
【図７】一実施形態に係るブログサーバ１のシステム制御部２０の広告コンテンツ挿入処理における処理例を示すフローチャートである。
【図８】一実施形態に係るブログサーバ１のシステム制御部２０の１ページ対応抽出処理における処理例を示すフローチャートである。
【図９】一実施形態に係るブログサーバ１のシステム制御部２０のツリー探索処理における処理例を示すフローチャートである。
【図１０】一実施形態に係るブログサーバ１のシステム制御部２０の特有コンテンツブロック判定処理における処理例を示すフローチャートである。
【図１１】一実施形態の変形例に係るブログサーバ１のシステム制御部２０のブログ更新時処理における処理例を示すフローチャートである。
【発明を実施するための形態】
【００２６】
以下、図面を参照して本発明の実施形態について詳細に説明する。なお、以下に説明する実施の形態は、ブログサービスを提供するブログシステムにおいて、ブログページ送信するサーバ装置に対して本発明を適用した場合の実施形態である。
【００２７】
［１．ブログシステムの構成及び機能概要］
先ず、本実施形態に係るブログシステムＳの構成及び概要機能について、図１を用いて説明する。
【００２８】
図１は、本実施形態に係るブログシステムＳの概要構成の一例を示す図である。
【００２９】
図１に示すように、ブログシステムＳは、特有コンテンツ判定装置及び関連コンテンツ挿入装置の一例としてのブログサーバ１と、管理端末２と、複数のユーザ端末３と、を含んで構成されている。そして、ブログサーバ１と、各ユーザ端末３とは、ネットワークＮＷを介して、例えば、通信プロトコルにＴＣＰ／ＩＰ等を用いて相互にデータの送受信が可能になっている。なお、ネットワークＮＷは、例えば、インターネット、専用通信回線（例えば、ＣＡＴＶ（Community Antenna Television）回線）、移動体通信網（基地局等を含む）、及びゲートウェイ等により構築されている。また、ブログサーバ１と管理端末２とは、ＬＡＮ（Local Area Network）等のネットワークを介して接続されている。
【００３０】
このような構成のブログシステムＳにおいて、ブログサーバ１は、ユーザ端末３からのリクエストに応じて、ブログサービスサイトを構成するＷｅｂページを送信するＷｅｂサーバである。ユーザ端末３を利用するユーザがブログサービスサイトのユーザ登録をすると、そのユーザは、当該ブログサービスサイトにおいてユーザ自身のブログを運営することができるようになっている。そして、ユーザ登録されたユーザ（ブロガー）は、ブログサービスサイトにアクセスして、自己のブログを更新（ブログ記事（ブログ１件毎の記録）を追加）することができるようになっている。そのため、ブログサーバ１は、ブログの更新に応じて、ブログのＷｅｂページとして、１又は複数のブログ記事が掲載されるブログページを生成又は更新する。そして、ブログサーバ１は、ブログページＤＢ１０１を備

え、ブログページを当該ブログページＤＢ１０１に登録する。

【００３１】

また、ブログサーバ１は、システム管理者により指定されたブロガーのブログページに対して、広告コンテンツ（関連コンテンツの一例）を挿入する。広告コンテンツとしては、例えば、広告文章のテキストデータ、バナーの画像データ、動画データ、Adobe Flash（商標）やSilverlight（商標）等により生成されたリッチインターネットアプリケーション（ＲＩＡ）等がある。各ブログページに挿入される広告コンテンツは、対象のブログページに掲載されているブログ記事に関連する商品やサービスに関する広告を示すコンテンツである。そのため、ブログサーバ１は、複数の広告コンテンツが登録されている広告ＤＢ１０２を備える。そして、ブログサーバ１は、ブログページからブログ記事を抽出し、更にブログ記事から特徴語を抽出して、抽出した特徴語に関連する広告コンテンツを選択する。

【００３２】

ユーザ端末３は、ブロガーとしてのユーザや、ブログを閲覧するユーザにより利用される端末装置である。ユーザ端末３としては、例えば、パーソナルコンピュータ、ＰＤＡ、携帯電話機等が用いられる。

【００３３】

管理端末２は、ブログシステムＳのシステム管理者により使用される端末装置である。管理端末２としては、例えば、パーソナルコンピュータ等が用いられる。

【００３４】

［２．ブログサーバの構成及び機能］

次に、ブログサーバ１の構成及び機能について、図２を用いて説明する。

【００３５】

図２は、本実施形態に係るブログサーバ１の概要構成の一例を示すブロック図である。また、図３は、ブロガーが指定されてからブログページに広告コンテンツが挿入されるまでの処理の概要を示す図である。また、図４は、Ｗｅｂページの構成例を示す図である。また、図５は、ＨＴＭＬ文書から生成されたＤＯＭツリーの一例を示す図である。また、図６は、記憶部１５に記憶されたコンテンツブロック対応情報の内容の一例を示す図である。

【００３６】

図２に示すように、ブログサーバ１は、操作部１１と、表示部１２と、通信部１３と、ドライブ部１４と、記憶手段の一例としての記憶部１５と、入出力インタフェース部１６と、システム制御部２０と、を備えている。そして、システム制御部２０と入出力インタフェース部１６とは、システムバス２１を介して接続されている。

【００３７】

操作部１１は、例えば、キーボード、マウス等により構成されており、システム管理者等からの操作指示を受け付け、その指示内容を指示信号としてシステム制御部２０に出力するようになっている。表示部１２は、例えば、ＣＲＴ（Cathode Ray Tube）ディスプレイ、液晶ディスプレイ等により構成されており、文字や画像等の情報を表示するようになっている。通信部１３は、ネットワークＮＷ等に接続して、管理端末２、ユーザ端末３等との通信状態を制御するようになっている。ドライブ部１４は、例えば、フレキシブルディスク、ＣＤ（Compact Disc）、ＤＶＤ（Digital Versatile Disc）等のディスクＤＫからデータ等を読み出す一方、当該ディスクＤＫに対してデータ等を記録するようになっている。

【００３８】

記憶部１５は、例えば、ハードディスクドライブ等により構成されており、各種プログラム及びデータ等を記憶するようになっている。また、記憶部１５には、ブログページＤＢ１０１及び広告ＤＢ１０２が構築されている。ブログページＤＢ１０１には、ブログサービスサイトを構成する各ブログページ（ブログページのＨＴＭＬ文書（ドキュメントデータの一例）、ブログページの素材である画像データ等）が、例えば、そのページのＵＲ

Ｌ及びブロガーの識別情報であるユーザＩＤに対応付けて登録されている。広告ＤＢ１０２には、複数の広告コンテンツが、その広告コンテンツによる広告対象の商品やサービスに関連するキーワードに対応付けて登録されている。ここで、広告コンテンツにテキストデータ以外のコンテンツが含まれている場合には、当該コンテンツのＵＲＬも対応付けて登録されている。また、ブログページ上に表示された広告コンテンツをユーザが選択したときに広告対象の商品やサービスに関するＷｅｂページに移動するため、当該ＷｅｂページのＵＲＬも対応付けて登録されている。

【００３９】

入出力インタフェース部１６は、操作部１１～記憶部１５とシステム制御部２０との間のインタフェース処理を行うようになっている。システム制御部２０は、ＣＰＵ（Central Processing Unit）１７、ＲＯＭ（Read Only Memory）１８、ＲＡＭ（Random Access Memory）１９等により構成されている。

【００４０】

システム制御部２０は、ＣＰＵ１７が、ＲＯＭ１８や記憶部１５に記憶された各種プログラムを読み出し実行することによりブログサーバ１の各部を制御する。また、システム制御部２０は、広告コンテンツ挿入ソフトウェア（特有コンテンツ判定プログラムの一例）を実行することにより、抽出手段、計算手段、判定手段及び挿入手段として機能する。なお、広告コンテンツ挿入ソフトウェア等は、例えば、他のサーバ装置等からネットワークＮＷを介して取得されるようにしても良いし、ＣＤ－ＲＯＭ等のディスクＤＫに記録されてドライブ部１４を介して読み込まれるようにしても良い。

【００４１】

広告コンテンツ挿入ソフトウェアは、ブログページに広告コンテンツを挿入するためのプログラムである。図３に示すように、広告コンテンツ挿入ソフトウェアは、マネージャ部、素材抽出エンジン、文章解析エンジン、広告選択部等により構成されている。マネージャ部は、素材抽出エンジン、文章解析エンジン及び広告選択部の実行を制御する。素材抽出エンジンは、ブログページのＨＴＭＬ文書からＷｅｂ素材としてのコンテンツを抽出するとともに、ブログページ特有のコンテンツを判定するためのソフトウェアである。コンテンツの抽出は、後述のコンテンツブロック（コンテンツグループの一例）の単位で行われる。本実施形態においては、例えば、その記事特有の内容を含むブログ記事が、ブログページに特有のコンテンツブロックに相当する。

【００４２】

文章解析エンジンは、ブログページ特有のコンテンツとして抽出されたブログ記事から当該ブログページの特徴語を抽出するためのソフトウェアである。広告選択部は、抽出された特徴語をキーワードとして、ブログページに関連する広告コンテンツを選択するためのソフトウェアである。

【００４３】

以下に、広告コンテンツの挿入の概要について説明する。図３に示すように、システム管理者により対象のブロガーのユーザＩＤが指定される（１）。すると、システム制御部２０は、指定されたユーザＩＤに対応する全ブログページのＨＴＭＬ文書をブログページＤＢ１０１から取得して解析し、Ｗｅｂ素材たるコンテンツをコンテンツブロック単位で抽出する。そして、その抽出結果として、抽出したコンテンツブロック毎にコンテンツブロック対応情報（コンテンツ情報の一例）を生成する（２）。次いで、システム制御部２０は、抽出した各コンテンツブロックの、指定されたユーザＩＤに対応する全ブログページにおける出現頻度を計算する。本実施形態において計算される出現頻度は、例えば、出現回数（度数）である。そして、システム制御部２０は、各ブログページにおいて、出現頻度が所定の閾値以下であるコンテンツブロックを、そのブログページ特有のコンテンツブロックであると判定する（３）。

【００４４】

システム制御部２０は、特有のコンテンツブロックであると判定したコンテンツブロック、すなわち、ブログ記事に対して形態素解析等の分析を行って、ブロックページ毎の特

徴語を抽出する（４）。特徴語の抽出方法としては、種々の方法があり、また公知であるので、詳細な説明は省略する。一例としては、出現頻度の最も高い単語を特徴語とする。

【００４５】

次いで、システム制御部２０は、広告ＤＢ１０２を参照し、抽出した特徴語に関連する広告コンテンツを選択する（５）。そして、システム制御部２０は、選択した広告コンテンツをブログページに挿入して表示させるための規定（タグやデータそのものの記述等）を、ブログページのＨＴＭＬ文書に挿入する（６）。

【００４６】

次に、コンテンツブロックの抽出方法について説明する。なお、本実施形態においては、テキストデータ及び画像データがＷｅｂ素材として抽出されるものとする。

【００４７】

例えば、ブログページの構成（レイアウト）が図４に示すようなものであるとする。Ｗｅｂ素材としての各コンテンツは、ブログページ上において、それぞれ或るまとまり（かたまり）毎に表示されている。その各まとまりがコンテンツブロックに相当する。各コンテンツは、ＨＴＭＬ文書に記述されているＤＩＶタグ及びＴＡＢＬＥタグ（予め定められたタグの一例）により夫々コンテンツブロックに分けられる。つまり、各コンテンツは、ＤＩＶタグ及びＴＡＢＬＥタグ（以下、「ブロック化タグ」という）によりブロック化（グループ化）される。

【００４８】

図４には、コンテンツブロック３０１～３０６が表示されている。コンテンツブロック３０１は、例えば、ページのヘッダ部分のコンテンツブロックであり、テキストＡ及び画像ａにより構成されている。また、コンテンツブロック３０２は、例えば、各種カテゴリの商品に関するＷｅｂページに移動するためのナビゲーション部分のコンテンツブロックであり、例えば、他のＷｅｂページへのリンクを示すテキストＢ、テキストＣ及びテキストＤにより構成されている。また、コンテンツブロック３０３は、例えば、ブログの表示領域に対応するコンテンツブロックであり、ブログ等の見出しを示すテキストＥ、コンテンツブロック３０４及びコンテンツブロック３０５により構成されいている。このように、コンテンツブロックが入れ子状、つまり、階層構造になっている場合もある。この場合、コンテンツブロック３０３に含まれるコンテンツはテキストＥのみとされ、コンテンツブロック３０４及びコンテンツブロック３０５は、コンテンツブロック３０３から独立しているものとされる。コンテンツブロック３０４及び３０５は、夫々１件のブログ記事である。コンテンツブロック３０４には、ブログ記事の表題や本文を示すテキストＦ及びＧにより構成されている。コンテンツブロック３０５には、ブログ記事の表題や本文を示すテキストＨ、Ｉ及びＪと、ブログ記事に関連してブロガーにより登録された画像ｂ及びｃとにより構成されている。コンテンツブロック３０６は、例えば、コピーライト表示を示すコンテンツブロックであり、テキストＩにより構成されている。

【００４９】

これらのコンテンツブロックのうち、コンテンツブロック３０１、３０２、３０３及び３０６は、図４に示すブログページ以外のブログページ上でも比較的頻繁に現れる。一方、コンテンツブロック３０４及びコンテンツブロック３０５は、基本的に当該ブログページにだけに用いられる。よって、コンテンツブロック３０４又はコンテンツブロック３０５が、当該ブログページ特有のコンテンツブロックであると判断されることとなる。

【００５０】

本実施形態においては、その記事特有の内容を含むブログ記事に相当するコンテンツブロックを特有のコンテンツブロックと判定されるようにする必要がある。こうした特有の内容を含むブログ記事は、１ページ内に複数含まれている場合がある。そのため、出現頻度が所定の閾値以下であるコンテンツブロックは、全て特有のコンテンツブロックとされる。例えば、閾値の値を１回に設定する。そうすると、特有の内容を含むブログ記事は、特有のコンテンツブロックと判定され、他のブログ記事と同じような内容のみを含むブログ記事は、特有のコンテンツブロックとは判定されない。また、ヘッダ部分、ナビゲーシ

ョン部分、コピーライト表示部分等の各ブログページに共通するようなコンテンツブロックの出現頻度は夫々２回以上となるので、これらも特有のコンテンツブロックとは判定されない。なお、閾値は記憶部１５に予め記憶される。

【００５１】

図４に示すブログページのＨＴＭＬ文書をＤＯＭ（Document Object Model）ツリー、すなわち、木構造で表したものが図５である。なお、図５に示すＤＯＭツリーにおいて、本実施形態の説明に必要のないタグのノードの図示は省略している。

【００５２】

ＤＯＭツリーにおいては、ＤＩＶタグを示すＤＩＶノード及びＴＡＢＬＥタグを示すＴＡＢＬＥノードが、各コンテンツをコンテンツブロックにブロック化するノードとされる（以下、「ブロック化ノード」という）。システム制御部２０は、例えば、深さ優先探索によりＤＯＭツリーを探索し、コンテンツブロックを確定していく。具体的に、システム制御部２０は、ブロック化ノードを発見すると、当該ノードを頂点とする部分木の各ノードに規定されているコンテンツを一まとめにしてコンテンツブロックとする。ただし、コンテンツブロックが階層的に規定された結果、或るブロック化ノード（以下、「上位のブロック化ノード」という）を発見した後、当該ブロック化ノードの子孫のノードの中から更にブロック化ノード（以下、「下位のブロック化ノード」という）を発見すると、コンテンツブロックが分割される。例えば、ルートノードからの距離が近いノードほど階層的により上位のノードであるとすると、上位のブロック化ノードを頂点とする部分木（以下、「上位部分木」という）に相当するコンテンツブロックを、下位のブロック化ノードを頂点とする部分木（以下、「下位部分木」という）に相当するコンテンツブロックと、上位部分木のうち下位部分木を除く部分に相当するコンテンツブロックとに分ける（例えば、コンテンツブロック３０４とコンテンツブロック３０３）。この場合においては、前者のコンテンツブロックを、後者のコンテンツブロックよりも階層的に下位にあるものとする。例えば、コンテンツブロック３０１、３０２、３０３及び３０６の階層は１であり、コンテンツブロック３０４及び３０５の階層は２である。つまり、階層の値が小さいほど階層的により上位となる。

【００５３】

これを、ＨＴＭＬ文書のタグの記述で説明すると、基本的にはブロック化タグが記述されていると、ブロック化タグに挟まれた範囲内に規定が記述されているコンテンツが、まとめて当該ブロック化タグに対応するコンテンツブロックとされる。ただし、ブロック化タグが階層的に記述されている場合、或るブロック化タグに挟まれた範囲内に規定が記述されているコンテンツのうち、当該ブロック化タグよりも下位に記述されているブロック化タグに挟まれた範囲に規定が記述されているコンテンツを除いたものが、まとめて当該ブロック化タグに対応するコンテンツブロックとされる。

【００５４】

このようにしてコンテンツブロックを抽出すると、システム制御部２０は、抽出結果を示すコンテンツブロック対応情報を一時的に記憶部１５に保存する。図６に示すように、コンテンツブロック対応情報（符号４０１）は、コンテンツブロック毎に保存される。コンテンツブロック対応情報には、抽出元のＨＴＭＬ文書のＵＲＬ設定部分（符号４０２）とブロック構成情報（符号４０３）とにより構成されている。ブロック構成情報には、抽出された各コンテンツが設定される。ここで、テキストデータについては、ＤＯＭツリーにおけるテキストノードの内容が設定される。一方、画像データについては、画像データそのものの代わりとして、ＤＯＭツリーにおいてＩＭＧタグを示すＩＭＧノードのｓｒｃ属性としての画像データのＵＲＬが設定される。なお、本実施形態においては、ブログページ特有のコンテンツブロックであると判定されたコンテンツブロック、すなわち、ブログ記事から特徴語を抽出するので、テキストデータが抽出されれば良く、画像データについては抽出しなくても良い。

【００５５】

［３．ブログシステムの動作］

次に、ブログシステムＳの動作について、図７乃至図１０を用いて説明する。

【００５６】

図７は、本実施形態に係るブログサーバ１のシステム制御部２０の広告コンテンツ挿入処理における処理例を示すフローチャートである。

【００５７】

広告コンテンツ挿入処理は、例えば、システム管理者の操作に基づいて管理端末２から広告コンテンツ挿入処理の実行のリクエストが送信されてきたときに開始される。

【００５８】

そして、システム管理者が広告コンテンツの挿入対象とするブログを運営するブロガーのユーザＩＤを指定すると、図７に示すように、システム制御部２０は、指定されたユーザＩＤを管理端末２から受信する（ステップＳ１）。

【００５９】

次いで、システム制御部２０は、ブロック数ＮＵＭに０を設定する（ステップＳ２）。ブロック数ＮＵＭは、現時点で発見済みのコンテンツブロックの個数である。また、ＮＵＭはグローバル変数であり、後述する１ページ対応抽出処理及びツリー探索処理からアクセスが可能である。

【００６０】

次いで、システム制御部２０は、受信したユーザＩＤに対応する最初のブログページのＨＴＭＬ文書をブログページＤＢ１０１から取得する（ステップＳ３）。次いで、システム制御部２０は、取得したＨＴＭＬ文書を指定して、後述する１ページ対応抽出処理を実行する（ステップＳ４）。この１ページ対応抽出処理では、取得したＨＴＭＬ文書からコンテンツブロックが抽出され、コンテンツブロック対応情報が保存される。

【００６１】

次いで、システム制御部２０は、受信したユーザＩＤに対応する全てのブログページのコンテンツブロックを抽出したか否かを判定する（ステップＳ５）。このとき、システム制御部２０は、コンテンツブロックを抽出していないブログページが存在する場合には（ステップＳ５：ＮＯ）、次のブログページのＨＴＭＬ文書をブログページＤＢ１０１から取得して（ステップＳ６）、ステップＳ４に移行する。そして、システム制御部２０は、ステップＳ４～Ｓ６の処理を繰り返して全てのブログページのコンテンツブロックを抽出すると（ステップＳ５：ＹＥＳ）、ステップＳ７に移行する。

【００６２】

ステップＳ７において、システム制御部２０は、受信したユーザＩＤに対応する最初のブログページのＨＴＭＬ文書を特定する。

【００６３】

次いで、システム制御部２０は、取得したＨＴＭＬ文書を指定して、後述する特有コンテンツブロック判定処理を実行する（ステップＳ８）。この特有コンテンツブロック判定処理では、特定したＨＴＭＬ文書からコンテンツブロックが抽出され、ブログページ特有のコンテンツブロックが判定される。

【００６４】

次いで、システム制御部２０は、特有と判定されたコンテンツブロックを構成する各テキストデータからブログページの特徴語を抽出する（ステップＳ９）。次いで、システム制御部２０は、抽出した特徴語に基づいて、ブログページに関連する広告ページを当該ブログページに挿入する（ステップＳ１０）。具体的に、システム制御部２０は、抽出した特徴語をキーワードとし、広告ＤＢ１０２に参照して当該キーワードに対応する広告コンテンツを選択する。次いで、システム制御部２０は、特定したＨＴＭＬ文書上の所定の位置に、選択した広告コンテンツの規定を挿入する。例えば、システム制御部２０は、広告コンテンツにテキストデータが含まれている場合には、当該テキストデータの内容をＨＴＭＬ文書に追加する。また、例えば、システム制御部２０は、広告コンテンツに画像データが含まれている場合には、当該画像データを表示するためのＩＭＧタグをＨＴＭＬ文書に追加する。また、例えば、システム制御部２０は、広告対象の商品やサービスに関する

Ｗｅｂページへのリンク情報をＨＴＭＬ文書に追加する。

【００６５】

システム制御部２０は、特定したＨＴＭＬ文書に広告コンテンツの規定を挿入すると、当該ＨＴＭＬ文書で、ブログページＤＢ１０１に登録されているＨＴＭＬ文書を更新する（ステップＳ１１）。

【００６６】

次いで、システム制御部２０は、受信したユーザＩＤに対応する全てのブログページに広告コンテンツを挿入したか否かを判定する（ステップＳ１２）。このとき、システム制御部２０は、広告コンテンツを挿入していないブログページが存在する場合には（ステップＳ１２：ＮＯ）、次のブログページのＨＴＭＬ文書を特定して（ステップＳ１３）、ステップＳ８に移行する。そして、システム制御部２０は、ステップＳ８～Ｓ１３の処理を繰り返して全てのブログページに広告コンテンツを挿入すると（ステップＳ１２：ＹＥＳ）、記憶部１５に保存させておいた全てのコンテンツブロック対応情報を、記憶部１５から削除する（ステップＳ１４）。システム制御部２０は、この処理を終えると、広告コンテンツ挿入処理を終了させる。

【００６７】

図８は、本実施形態に係るブログサーバ１のシステム制御部２０の１ページ対応抽出処理における処理例を示すフローチャートである。

【００６８】

図８に示すように、システム制御部２０は、先ず、取得したＨＴＭＬ文書のＤＯＭツリーをＲＡＭ１９上に生成する（ステップＳ２１）。

【００６９】

次いで、システム制御部２０は、階層ＬＶに０を設定する（ステップＳ２２）。階層ＬＶは、ＤＯＭツリーにおいて現在探索中のノードが属するコンテンツブロックの階層である。ＬＶはグローバル変数であり、１ページ対応抽出処理及び後述するツリー探索処理からアクセスが可能である。

【００７０】

次いで、システム制御部２０は、ＤＯＭツリーのルートノードを指定して（ステップＳ２３）、ツリー探索処理を実行する（ステップＳ２４）。ツリー探索処理は再帰呼び出しが可能であり、このツリー探索処理により、Ｗｅｂページから全てのコンテンツブロックが抽出され、コンテンツブロック対応情報が生成される。

【００７１】

次いで、システム制御部２０は、ツリー探索処理により生成された各コンテンツブロック対応情報を記憶部１５に保存する（ステップＳ２５）。システム制御部２０は、この処理を終えると、１ページ対応抽出処理を終了させる。

【００７２】

図９は、本実施形態に係るブログサーバ１のシステム制御部２０のツリー探索処理における処理例を示すフローチャートである。

【００７３】

図９に示すように、システム制御部２０は、先ず、指定されたノードの種類を判定する（ステップＳ３１）。このとき、システム制御部２０は、指定されたノードの種類が、ＤＩＶノード又はＴＡＢＬＥノード（ブロック化ノード）である場合には、すなわち、コンテンツブロックが発見された場合には（ステップＳ３１：ＤＩＶ又はＴＡＢＬＥ）、ステップＳ３２に移行する。

【００７４】

ステップＳ３２において、システム制御部２０は、ブロック数ＮＵＭに１を加算するとともに、階層ＬＶに１を加算する。次いで、システム制御部２０は、ブロック番号ＢＮ［ＬＶ］にＮＵＭを設定する（ステップＳ３３）。ブロック番号ＢＮ［ＬＶ］は、現在探索中のノードが属する階層ＬＶで示されるコンテンツブロックのブロック番号である。このブロック番号は、コンテンツブロックの発見順に付与される。また、ＢＮ［ＬＶ］は、グ

ローバル変数である。

【００７５】

次いで、システム制御部２０は、ブロック番号ＢＮ［ＬＶ］のコンテンツブロックに対応するコンテンツブロック対応情報を初期化する（ステップＳ３４）。具体的に、システム制御部２０は、コンテンツブロック対応情報を格納する領域をＲＡＭ１９上に設定し、取得したＨＴＭＬ文書のＵＲＬを、当該領域に設定する。

【００７６】

次いで、システム制御部２０は、指定されたノードの子ノードのうち、未だ探索されていない子ノードが存在するか否かを判定する（ステップＳ３５）。このとき、システム制御部２０は、未だ探索されていない子ノードが存在する場合には（ステップＳ３５：ＹＥＳ）、ステップＳ３６に移行する。

【００７７】

ステップＳ３６において、システム制御部２０は、探索されていない子ノードのうちの１つの子ノードを指定して、ツリー探索処理を実行する（ステップＳ３７）。システム制御部２０は、ツリー探索処理を終えると、ステップＳ３５に移行する。

【００７８】

そして、システム制御部２０は、ステップＳ３５～Ｓ３７の処理を繰り返して全ての子ノードのツリー探索処理を終えると（ステップＳ３５：ＮＯ）、ステップＳ３８に移行する。なお、システム制御部２０は、指定されたノードの子ノードが１つも存在しない場合にも、ステップＳ３８に移行する。ステップＳ３８において、システム制御部２０は、階層ＬＶから１を減算して、ツリー探索処理を終了させる。

【００７９】

ステップＳ３１において、システム制御部２０は、指定されたノードの種類がテキストノードである場合には（ステップＳ３１：テキスト）、指定されたノードの内容（テキストデータ）を、ブロック番号ＢＮ［ＬＶ］のコンテンツブロックに対応するコンテンツブロック対応情報中のブロック構成情報に追加設定する（ステップＳ３９）。システム制御部２０は、この処理を終えると、ツリー探索処理を終了させる。

【００８０】

ステップＳ３１において、システム制御部２０は、指定されたノードの種類がＩＭＧノードである場合には（ステップＳ３１：ＩＭＧ）、指定されたノードのｓｒｃ属性として設定されている画像データのＵＲＬを取得し、取得したＵＲＬを、ブロック番号ＢＮ［ＬＶ］のコンテンツブロックに対応するコンテンツブロック対応情報中のブロック構成情報に追加設定する（ステップＳ４０）。システム制御部２０は、この処理を終えると、ツリー探索処理を終了させる。

【００８１】

ステップＳ３１において、システム制御部２０は、指定されたノードの種類が、ＤＩＶノード、ＴＡＢＬＥノード、テキストノード、及びＩＭＧノードの何れでもない場合には（ステップＳ３１：その他）、指定されたノードの子ノードのうち、未だ探索されていない子ノードが存在するか否かを判定する（ステップＳ４１）。このとき、システム制御部２０は、未だ探索されていない子ノードが存在する場合には（ステップＳ４１：ＹＥＳ）、探索されていない子ノードのうちの１つの子ノードを指定して（ステップＳ４２）、ツリー探索処理を実行する（ステップＳ４３）。システム制御部２０は、ツリー探索処理を終えると、ステップＳ４１に移行する。

【００８２】

一方、システム制御部２０は、指定されたノードの全ての子ノードのツリー探索処理を終えた場合、又は、指定されたノードの子ノードが１つも存在しない場合には（ステップＳ４１：ＮＯ）、ツリー探索処理を終了させる。

【００８３】

図１０は、本実施形態に係るコンテンツ生成サーバ１のシステム制御部２０の特有コンテンツブロック判定処理における処理例を示すフローチャートである。

【００８４】
図１０に示すように、システム制御部２０は、先ず、１ページ対応抽出処理と同様に、指定されたＨＴＭＬ文書のＤＯＭツリー生成（ステップＳ６１）、ブロック数ＮＵＭ及び階層ＬＶに対して０の設定を行い（ステップＳ６２）、ＤＯＭツリーのルートノードを指定して（ステップＳ６３）、ツリー探索処理を実行する（ステップＳ６４）。
【００８５】
次いで、システム制御部２０は、ブロック番号ｉに１を設定する（ステップＳ６５）。次いで、システム制御部２０は、ブロック番号ｉのコンテンツブロックの出現頻度を計算する（ステップＳ６６）。
【００８６】
具体的に、システム制御部２０は、ステップＳ６４のツリー探索処理において生成されたコンテンツブロック対応情報ｉ（ブロック番号ｉのコンテンツブロックに対応するコンテンツブロック対応情報）のブロック構成情報と、記憶部１５に保存されている各コンテンツブロック対応情報のブロック構成情報とを比較する。このとき、システム制御部２０は、ブロック構成情報の内容が一致する場合には、出現回数１回としてカウントする。このとき、システム制御部２０は、ブロック構成情報中におけるコンテンツの規定順は無視してかまわない。また、システム制御部２０は、記憶部１５に保存されているコンテンツブロック対応情報のブロック構成情報に規定されている一部のコンテンツがコンテンツブロック対応情報ｉのブロック構成情報に規定されている全部のコンテンツに一致する場合も、出現回数１回としてカウントしても良い。更に、システム制御部２０は、コンテンツブロック対応情報のブロック構成情報中に規定されているテキストデータ同士を比較する場合には、テキストデータが示す文章等そのものが一致するか否かを判定するのではなく、その文章等により表現されている実質的な内容を比較しても良い。例えば、システム制御部２０は、夫々のテキストデータの形態素解析等を行うことによりテキストデータから単語を抽出し、抽出した単語同士を比較しても良い。そして、システム制御部２０は、全ての単語が一致した場合にテキストデータ同士が一致したと判断しても良いし、所定の割合以上で単語が一致した場合にテキストデータ同士が一致したと判断しても良い。システム制御部２０は、このようにしてコンテンツブロック対応情報ｉのブロック構成情報と、記憶部１５に保存されている全てのコンテンツブロック対応情報のブロック構成情報とを比較して、出現頻度を計算する。
【００８７】
システム制御部２０は、出現頻度を計算すると、計算した出現頻度が、記憶部１５に記憶されている閾値以下であるか否かを判定する（ステップＳ６７）。このとき、システム制御部２０は、出現頻度が閾値以下である場合には（ステップＳ６７：ＹＥＳ）、ブロック番号ｉのコンテンツブロックを、特有のコンテンツブロックの１つであると判定する（ステップＳ６８）。つまり、システム制御部２０は、ブロック番号ｉのコンテンツブロックを、指定されたＨＴＭＬ文書が対応するブログページに特有のコンテンツブロックに加える。
【００８８】
システム制御部２０は、出現頻度が閾値よりも大きい場合（ステップＳ６７：ＮＯ）、又は、ステップＳ６８の処理を終えた場合には、ブロック番号ｉに１を加算して（ステップＳ６９）、ブロック番号ｉがブロック数ＮＵＭの値より大きいか否かを判定する（ステップＳ７０）。このとき、システム制御部２０は、ブロック番号ｉがブロック数ＮＵＭの値以下である場合には（ステップＳ７０：ＮＯ）、ステップＳ６６に移行する。そして、システム制御部２０は、ツリー探索処理において抽出された全てのコンテンツブロックの出現頻度を計算すると（ステップＳ７０：ＹＥＳ）、特有コンテンツブロック判定処理を終了させる。
【００８９】
なお、システム制御部２０は、ステップＳ６４のツリー探索処理によりコンテンツブロックを抽出していたが、広告コンテンツ挿入処理から実行された１ページ対応抽出処理（

図７ステップＳ４）において、受信したブロガーのユーザＩＤに対応する全てのブログページについてコンテンツブロックが抽出され、その結果としてコンテンツブロック対応情報が記憶部１５に記憶されているので、再度コンテンツブロックを抽出しなくても良い。その場合には、指定されたＨＴＭＬ文書のＵＲＬに基づいて、当該ＨＴＭＬ文書が対応するブログページを構成する各コンテンツブロックのコンテンツブロック対応情報を記憶部１５から取得することができる。

【００９０】

［４．変形例１］

次に、本実施形態の変形例について、図１１を用いて説明する。

【００９１】

これまでの説明においては、システム管理者によりブロガーが指定されたときに、指定されたブロガーのブログページに広告コンテンツを挿入していたが、ブログが更新されたタイミングで広告コンテンツを挿入しても良い。

【００９２】

図１１は、本実施形態の変形例に係るブログサーバ１のシステム制御部２０のブログ更新時処理における処理例を示すフローチャートである。なお、図１１において、図７と同様の処理については同様のステップ番号を付してある。

【００９３】

先ず、ブログの更新に先立ち、ブロガーは、ユーザ端末３を操作してブログサービスサイトにアクセスし、自身のユーザＩＤとパスワードとを入力することによりブログサービスサイトにログインする。このログインにより、ブログサーバ１はユーザ端末３に対してセッションＩＤを発行し、セッションＩＤとユーザＩＤとを対応付けて管理する。ユーザ端末３からブログサーバ１へのリクエストにはセッションＩＤが含まれているので、ブログサーバ１は、どのブロガーからのリクエストであるかを特定することができる。

【００９４】

そして、ブロガーが新しいブログ記事の登録操作を行うと、ユーザ端末３は、ブログ記事のデータ（表題や本文等のテキストデータ、画像データ等）をブログサーバ１に送信し、図１１に示すように、ブログサーバ１のシステム制御部２０は、ブログ記事のデータを受信する（ステップＳ７１）。次いで、システム制御部２０は、ブロガーのユーザＩＤに対応するブログページの中から、更新すべきブログページのＨＴＭＬ文書をブログページＤＢ１０１から取得する（ステップＳ７２）。次いで、システム制御部２０は、受信したブログ記事のデータに基づいて、取得したＨＴＭＬ文書を更新する（ステップＳ７３）。例えば、システム制御部２０は、取得したＨＴＭＬ文書に、ブログ記事用のＴＡＢＬＥタグ又はＤＩＶタグを追加し、当該タグに挟まれた形で、受信したブログ記事の表題や本文のテキストデータ等を追加する。次いで、システム制御部２０は、ブログ記事のデータを追加したＨＴＭＬ文書で、ブログページＤＢ１０１に登録されているＨＴＭＬ文書を更新する（ステップＳ７４）。

【００９５】

次いで、システム制御部２０は、ブロガーのユーザＩＤに対応する全てのブログページからコンテンツブロックを抽出する（ステップＳ３～Ｓ６）。

【００９６】

次いで、システム制御部２０は、ステップＳ７３において更新したＨＴＭＬ文書を指定して、特有コンテンツブロック判定処理を実行し（ステップＳ８）、特有と判定されたコンテンツブロックを構成する各テキストデータからブログページの特徴語を抽出する（ステップＳ９）。

【００９７】

次いで、システム制御部２０は、指定されたＨＴＭＬ文書から、既存の広告コンテンツの規定を削除し（ステップＳ７５）、抽出した特徴語をキーワードとして、関連する広告コンテンツの規定を挿入する（ステップＳ１０）。つまり、システム制御部２０は、ブログページ上に表示される広告コンテンツを変更する。

【００９８】
そして、システム制御部２０は、広告コンテンツの規定が挿入されたＨＴＭＬ文書で、ブログページＤＢ１０１に登録されているＨＴＭＬ文書を更新し（ステップＳ１１）、記憶部１５から全てのコンテンツブロック対応情報を削除する（ステップＳ１４）。
【００９９】
なお、ブログの更新に伴ってブログページを新規に生成しなければならない場合の処理も、基本的に上述した処理と同様で良い。ただし、新規に生成されたブログページには、広告コンテンツは未だ挿入されていないので、ステップＳ７５における広告コンテンツの規定の削除は行われない。
【０１００】
［５．変形例２］
これまでの説明においては、ブログページに特有のコンテンツの判定に用いられる閾値として１回を設定していたが、２回以上の値を閾値として設定しても良い。
【０１０１】
例えば、閾値を１回とした場合には、出現頻度が１回であるコンテンツブロック（ブログ記事）がブログページに特有のコンテンツとして抽出され、抽出されたブログ記事のテキストデータから特徴語が抽出される。このとき、抽出された各ブログ記事のテキストデータのデータ量が少ないと、そこから抽出される単語の数は少なくなる。そして、十分な数の単語を抽出することができないと、どの単語が特徴語であるかを全く判断することができない場合や、的確に判断することができない場合がある。そこで、閾値の値を上げて、ブログページに特有のコンテンツと判定される条件をゆるめることで、特徴語を抽出する対象となるブログ記事を増やしていく。これにより、特徴語を抽出することが可能となる。
【０１０２】
具体的には、ブログサーバ１のシステム制御部２０が、最初は閾値を１回に設定して、ブログページに特有のコンテンツブロックを判定することにより、出現回数が１回のブログ記事を抽出して特徴語を抽出する。このとき、システム制御部２０は、特徴語を抽出することができないと判定した場合には、閾値を２回に変更して、ブログ記事の抽出及び特徴語の抽出を行う。システム制御部２０は、それでも特徴語を抽出することができないと判定した場合には、閾値を３回に変更して、ブログ記事の抽出及び特徴語の抽出を行う。システム制御部２０は、こうした処理を、特徴語が抽出することができるまで継続する。つまり、特有のコンテンツブロックの抽出結果に基づく処理を正常に行うことができなかった場合に、閾値を上げるのである。
【０１０３】
ただし、閾値を無制限に上げていくと、ブログ記事ではないものも抽出されてしまうので、閾値がある程度まで上がると処理を中断するものとする。例えば、閾値が、指定されたブロガーに対応するブログページのページ数の値にまで上がると、各ブログページで共通して用いられるコンテンツブロックを抽出してしまうので、閾値がブログページのページ数の値になったら処理を中断しても良い。
【０１０４】
また例えば、ブログページの所定ページ数あたり１回のみ出現するコンテンツブロックをブログページ特有のコンテンツブロックであると、システム管理者側で予め定めても良い。この場合、指定されたブロガーに対応するブログページのページ数に比例して、閾値としての出現回数を変えても良い。
【０１０５】
［６．変形例３］
これまでの説明においては、ブログページに特有のコンテンツの判定に用いられる出現頻度として、出現回数（度数）を用いていたが、指定されたブロガーに対応するブログページの全コンテンツブロックに対する出現回数の割合（相対度数）を用いても良い。
【０１０６】

例えば、ブロガーが登録したブログ記事に対して、他のユーザからコメントを登録することができ、ブログ記事とともにコメントが閲覧可能になっているとする。このコメントのテキストデータもブログページを構成するコンテンツの１つとなる。ブログサーバ１のシステム制御部２０は、コメントのテキストデータをブログページに追加する場合、当該ブログページのＨＴＭＬ文書に、ブロック化タグの記述を追加した上で当該テキストデータを追加することにより、コメントのテキストデータを、ブログ記事や他のコメントのテキストデータとは独立したコンテンツブロックとする。そして、システム制御部２０は、コンテンツブロックとしてコメントのテキストデータを抽出し、抽出したコメントのテキストデータが特有の内容を有している場合には、そのコメントに関連する広告コンテンツをブログページに挿入するようにする。

【０１０７】

ところで、或るブログ記事に対して複数のコメントが登録された場合において、複数のコメントの各内容が、例えば、多数派の意見と少数派の意見といったように、頻繁に出現する内容と、あまり頻繁には出現しない内容とに分かれる場合がある。このとき、多数派の意見は、一般的な意見であり、あまり特徴的な内容ではないと考えることができる。一方、少数派の意見は、特異な意見であり、ブログページに特有の内容と考えることができる。そうした場合に、少数派の意見を示すコメントをブログページ特有のコンテンツとして抽出したい。

【０１０８】

しかしながら、多数派の意見の数と少数派の意見の数は、相対的なものであり、コメントの総数によって変化する。こうした場合において、出現頻度として度数を用い、閾値を例えば１回とすると、頻繁に現れない内容（少数派の意見）を適切に抽出することができない場合がある。そこで、出現頻度として相対度数を用い、閾値を所定の割合に設定するのである。このときの閾値は任意に設定することができる。例えば、抽出されたコンテンツブロックの内容がＮ個のパターン（Ｎは２以上の整数）に分けられる場合、少数派の意見を区別するために、閾値には１÷Ｎ未満の範囲で閾値を設定しても良い。このように、システム制御部２０が、その時々の状況に応じて閾値を変更しても良い。

【０１０９】

なお、ブログ等のような記事に対してコメント等を登録することができるシステムとして、例えば、或るユーザが登録したつぶやきに対して、これをフォローするつぶやきを他のユーザが登録することができるTwitter（商標）や、電子掲示板等がある。

【０１１０】

以上説明したように、本実施形態によれば、ブログサーバ１のシステム制御部２０が、ＨＴＭＬ文書が指定されることによって順次指定されたブログページを構成しているコンテンツを抽出し、指定されたブログページを構成している各コンテンツの出現頻度を計算し、指定されたブログページを構成するコンテンツのうち、出現頻度が所定の閾値以下のコンテンツを当該ブログページに特有のコンテンツであると判断する。

【０１１１】

従って、出現頻度が小さいコンテンツであるほど、指定されたブログページ以外にはあまり出現しないコンテンツであるので、出現頻度が閾値以下であるかを判定することで、当該条件を満たす全てのコンテンツが、指定されたブログページに特有のコンテンツであると特定される。よって、ブログページに特有のコンテンツを容易に抽出することができる。

【０１１２】

また、ブログサーバ１のシステム制御部２０が、指定されたブログページに特有のコンテンツに関連する広告コンテンツを当該ブログページに挿入する。

【０１１３】

従って、ブログページの特徴と関連する情報をＷｅｂページに追加することができる。

【０１１４】

また、ブログサーバ１のシステム制御部２０が、指定されたブログページを構成してい

るコンテンツとして、ブログの記事のテキストデータが含まれている場合に、当該テキストデータを、当該ブログページに特有のコンテンツであると判定し、ブログの記事のテキストデータから当該ブログページの特徴語を抽出し、当該特徴語をキーワードとして、予め関連付けられている広告コンテンツを、当該ブログページに挿入する。

【０１１５】

従って、ブログページに掲載されているブログの内容に関連する広告を当該ブログページに追加することができる。

【０１１６】

また、ブログサーバ１のシステム制御部２０が、ブログサービスサイトに含まれる複数のブログページ上における各コンテンツの出現頻度を計算する。

【０１１７】

従って、ブログサービスサイトに含まれる複数のＷｅｂページ（例えば、指定されたブロガーのユーザＩＤに対応する複数のブログページ）上において、指定されたブログページを構成している各コンテンツの出現頻度が計算されるので、ブログサービスサイト内で共通して用いられるコンテンツは、特有のコンテンツではないと判定することが可能となり、判断精度を上げることができる。

【０１１８】

また、ブログサーバ１のシステム制御部２０が、１つ以上のコンテンツで構成されるコンテンツブロックの単位で、ブログページを構成しているコンテンツを抽出し、指定されたブログページを構成している各コンテンツブロック出現頻度を計算し、指定されたブログページを構成するコンテンツブロックのうち、出現頻度が閾値以下のコンテンツブロックを当該ブログページに特有のコンテンツブロックであると判断する。

【０１１９】

従って、ブログページ上において、例えば、ヘッダ部分、ナビゲーション部分、ブログが表示される部分、コピーライト表示の部分等のように、１つ以上のコンテンツがまとまりをもってコンテンツブロックとして表示されている場合に、ブログページに特有のコンテンツブロックを抽出することができる。

【０１２０】

また、ブログサーバ１のシステム制御部２０が、ブログページを構成しているコンテンツを当該ブログページのＨＴＭＬ文書に基づいて抽出し、ＨＴＭＬ文書においてＤＩＶタグ又はＴＡＢＬＥタグに基づいて、コンテンツブロックを定める。

【０１２１】

従って、ＤＩＶタグにより、ＨＴＭＬ文書の作成の際に明示的にブロック化された１つ以上のコンテンツを特定することができ、また、ＴＡＢＬＥタグにより、表形式でブロック化されて表示される１つ以上のコンテンツを特定することができるので、例えば、これらのタグにより、ブログページに特有のコンテンツと、特有ではないコンテンツとがブロック化されている場合に、Ｗｅｂページに特有のコンテンツを判断する精度を上げることができる。

【０１２２】

なお、上記実施形態においては、Ｗｅｂページを構成しているコンテンツとして、テキストデータ及び画像データを抽出していたが、抽出対象のコンテンツはこれらに限られるものではない。例えば、Ｗｅｂページ上に表示されるコンテンツ、又は、Ｗｅｂページが表示されている際に再生されるコンテンツ（例えば、動画データ、音声データ、電子文書等）であれば良い。また、所定の種類のコンテンツのみを抽出しても良い。

【０１２３】

また、上記実施形態においては、指定されたブログページを構成する各コンテンツブロックに対応するコンテンツブロック対応情報を、指定されたブロガーのユーザＩＤに対応する全てのブログページを構成する各コンテンツブロック対応情報と比較することによって各出現頻度が計算されていた。つまり、指定されたブログページを構成する各コンテンツブロックの出現頻度を計算する場合に、指定されたブロガーに対応する全てのブログペ

ージを対象とした範囲に出現する頻度を計算するようになっていたが、対象とする範囲はこれだけに限られるものではない。例えば、予め定められたページ数分のブログページを対象としても良いし、ブログサービスサイトを構成する全てのブログページを対象としても良い。

【０１２４】

また、上記実施形態においては、ＤＩＶタグに挟まれているコンテンツ、及び、ＴＡＢＬＥタグに挟まれているコンテンツを、コンテンツブロックとしてグループ化して抽出していたが、コンテンツをグループ化するタグとしては、これらのみに限られるものではない。

【０１２５】

また、上記実施形態においては、Ｗｅｂページに特有のコンテンツをコンテンツブロックの単位で抽出していたが、各コンテンツをそのまま一つずつ抽出しても良い。

【０１２６】

また、上記実施形態として、Ｗｅｂページに特有のコンテンツに関連するコンテンツとして、商品やサービスに関する広告を示す広告コンテンツを、当該Ｗｅｂページに挿入していたが、関連するコンテンツであれば広告コンテンツに限られるものではない。例えば、特有のコンテンツと判定されたブログ記事等のコンテンツに関連する画像データ（静止画や動画像）を、背景画像や挿入画像（挿絵等）として挿入しても良い。具体的には、例えば、画像データ用のデータベースを構築し、当該データベースに、画像データとキーワードとを対応付けて登録しておく。画像データに対応付けられるキーワードは、その画像データによって表される画像を示す単語や当該画像に関連する単語である。そして、特有のコンテンツと判定されたコンテンツから特徴語を抽出し、抽出した特徴語をキーワードとして、関連する画像データをデータベースから選択する。そして、対象のＨＴＭＬ文書のＢＯＤＹタグに、選択した画像データのＵＲＬをｂａｃｋｇｒｏｕｎｄ属性として挿入したり、対象のＨＴＭＬ文書の所定位置に、選択した画像データを表示するＩＭＧタグを挿入したりする。これにより、特有のコンテンツと判定されたブログ記事等のコンテンツの内容に適した画像をＷｅｂページに挿入することができる。

【０１２７】

また、Ｗｅｂページに特有のコンテンツの用途としては、関連するコンテンツをＷｅｂページに挿入することのみに限られるものではない。例えば、Ｗｅｂページに特有のコンテンツに基づいて、新たなコンテンツを生成しても良い。

【０１２８】

また、上記実施形態においては、サーバ装置に対して本発明の特有コンテンツ判定装置を適用していたが、記憶手段やネットワーク上からＨＴＭＬ文書を取得することができれば、端末装置等に対して特有コンテンツ判定装置を適用しても良い。

【０１２９】

また、上記実施形態においては、ＨＴＭＬ文書に対して本発明のドキュメントデータを適用していたが、マークアップ言語で記述され、Ｗｅｂページを構成するコンテンツを示すデータ（例えば、ＸＨＴＭＬ（Extensible HyperText Markup Language）文書等）に対してドキュメントデータを適用しても良い。

【０１３０】

また、上記実施形態においては、ブログサービスサイトにおけるブログページを構成するコンテンツを抽出していたが、対象とするサイト及びページの種類はこれらのみに限られるものではない。

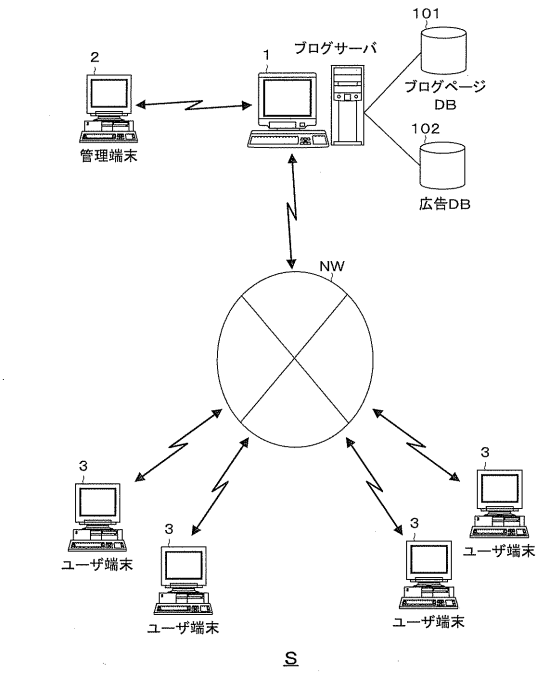
【符号の説明】

【０１３１】

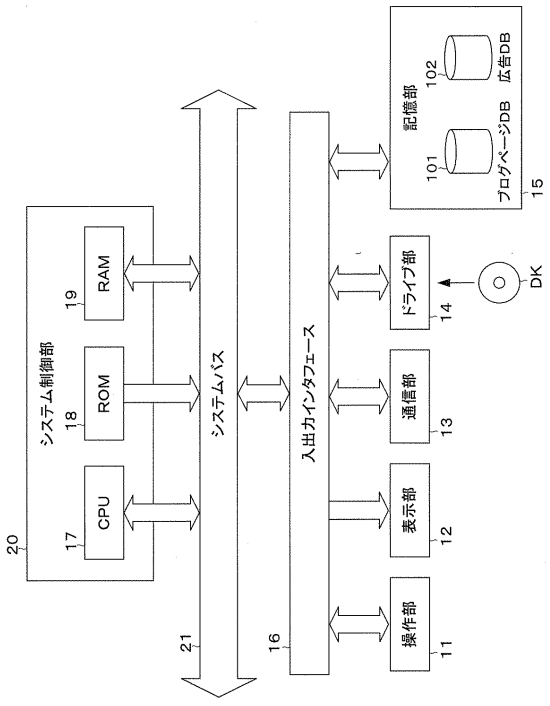
１　ブログサーバ
２　管理端末
３　ユーザ端末
１１　操作部

１２　表示部
１３　通信部
１４　ドライブ部
１５　記憶部
１６　入出力インタフェース部
１７　ＣＰＵ
１８　ＲＯＭ
１９　ＲＡＭ
２０　システム制御部
２１　システムバス
１０１　ブログページＤＢ
１０２　広告ＤＢ
ＮＷ　ネットワーク
Ｓ　ブログシステム

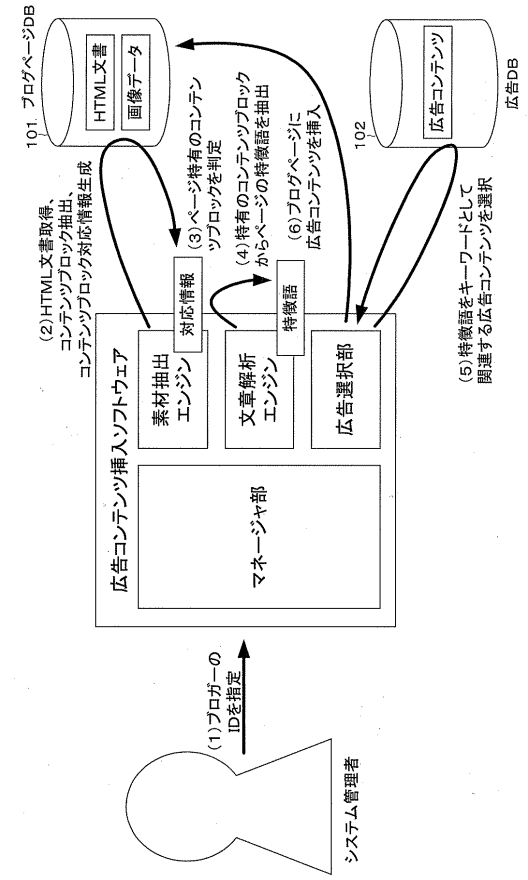
【図１】



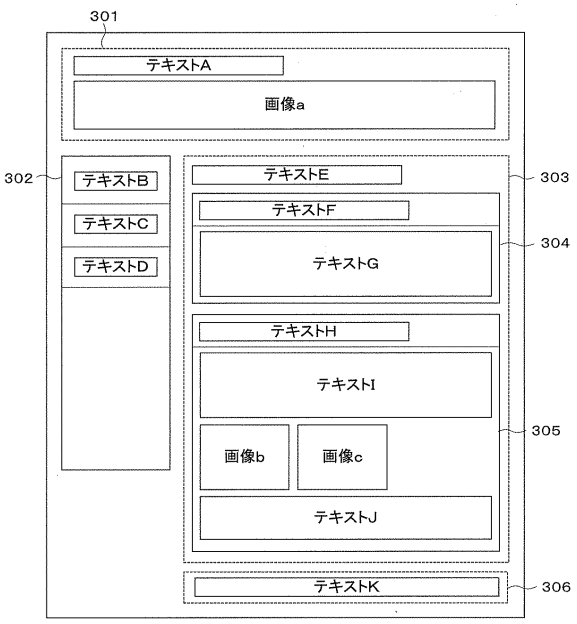
【図２】



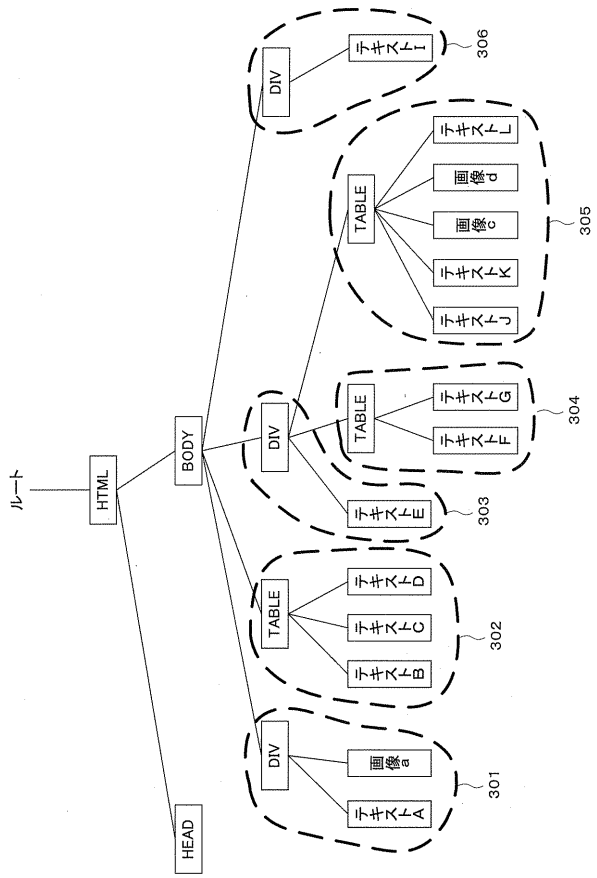
【図３】


ブログページDB
101
HTML文書
画像データ
広告コンテンツ
広告DB
102
(1)ブロガーのIDを指定
システム管理者
(2)HTML文書取得、コンテンツブロック抽出、コンテンツブロック対応情報生成
(3)ページ特有のコンテンツブロックを判定
(4)特有のコンテンツブロックからページの特徴語を抽出
(5)特徴語をキーワードとして関連する広告コンテンツを選択
(6)ブログページに広告コンテンツを挿入
広告コンテンツ挿入ソフトウェア
素材抽出エンジン
対応情報
文章解析エンジン
特徴語
広告選択部
マネージャ部

【図４】


301
テキストA
画像a
302
テキストB
テキストC
テキストD
テキストE
303
テキストF
テキストG
304
テキストH
テキストI
画像b
画像c
305
テキストJ
テキストK
306

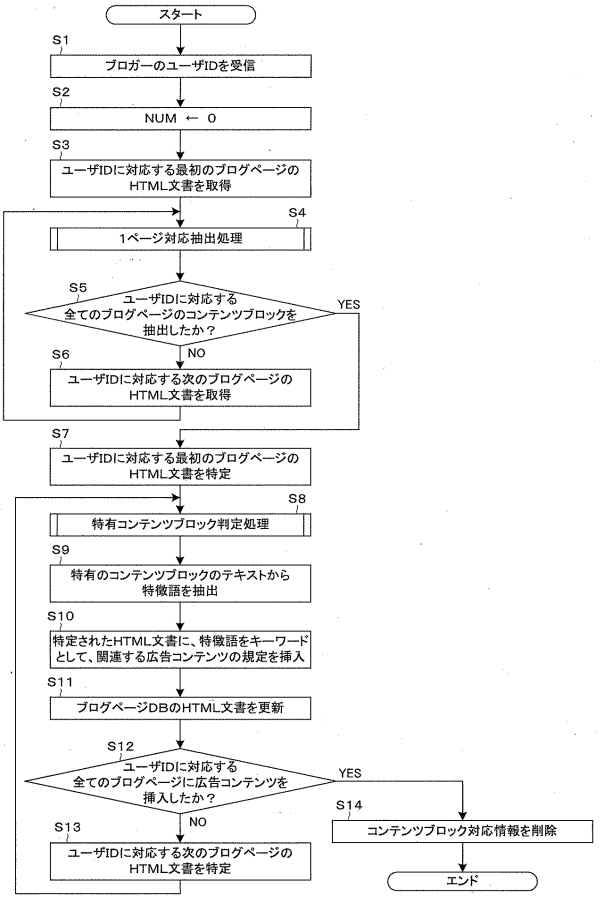
【図５】


ルート
HTML
BODY
HEAD
DIV
TABLE
テキストA
画像a
テキストB
テキストC
テキストD
テキストE
テキストF
テキストG
テキストH
テキストI
テキストJ
テキストK
テキストL
画像c
画像d
301
302
303
304
305
306

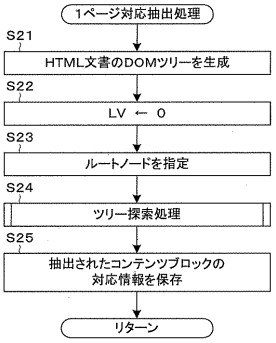
【図６】


15
401
402
403
HTML文書XのURL
テキストA
画像a
テキストB
テキストC
テキストD
テキストE
テキストF
テキストG
テキストH
テキストI
画像b
画像c
テキストJ
テキストK
HTML文書YのURL
テキストA
画像a
テキストB
テキストC
テキストD
テキストM
画像e
画像f
テキストN
テキストO

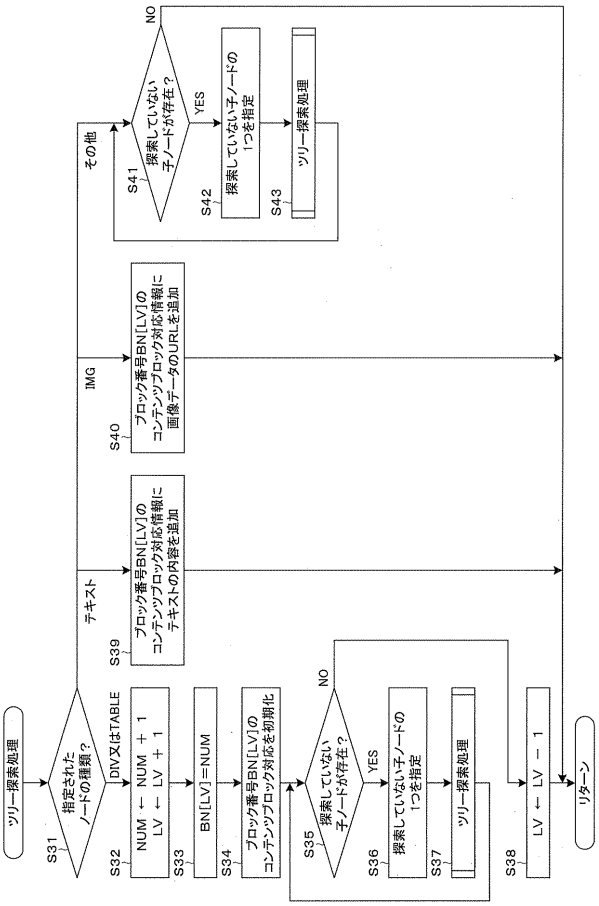
【図７】


スタート
S1
ブロガーのユーザIDを受信
S2
NUM ← 0
S3
ユーザIDに対応する最初のブログページの
HTML文書を取得
S4
1ページ対応抽出処理
S5
ユーザIDに対応する
全てのブログページのコンテンツブロックを
抽出したか？
YES
NO
S6
ユーザIDに対応する次のブログページの
HTML文書を取得
S7
ユーザIDに対応する最初のブログページの
HTML文書を特定
S8
特有コンテンツブロック判定処理
S9
特有のコンテンツブロックのテキストから
特徴語を抽出
S10
特定されたHTML文書に、特徴語をキーワードと
として、関連する広告コンテンツの規定を挿入
S11
ブログページDBのHTML文書を更新
S12
ユーザIDに対応する
全てのブログページに広告コンテンツを
挿入したか？
YES
NO
S13
ユーザIDに対応する次のブログページの
HTML文書を特定
S14
コンテンツブロック対応情報を削除
エンド

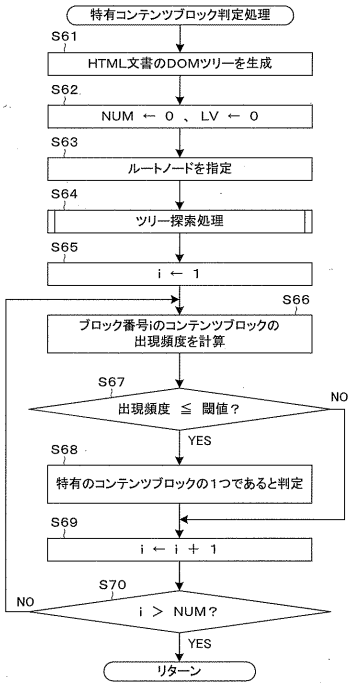
【図８】


1ページ対応抽出処理
S21
HTML文書のDOMツリーを生成
S22
LV ← 0
S23
ルートノードを指定
S24
ツリー探索処理
S25
抽出されたコンテンツブロックの
対応情報を保存
リターン

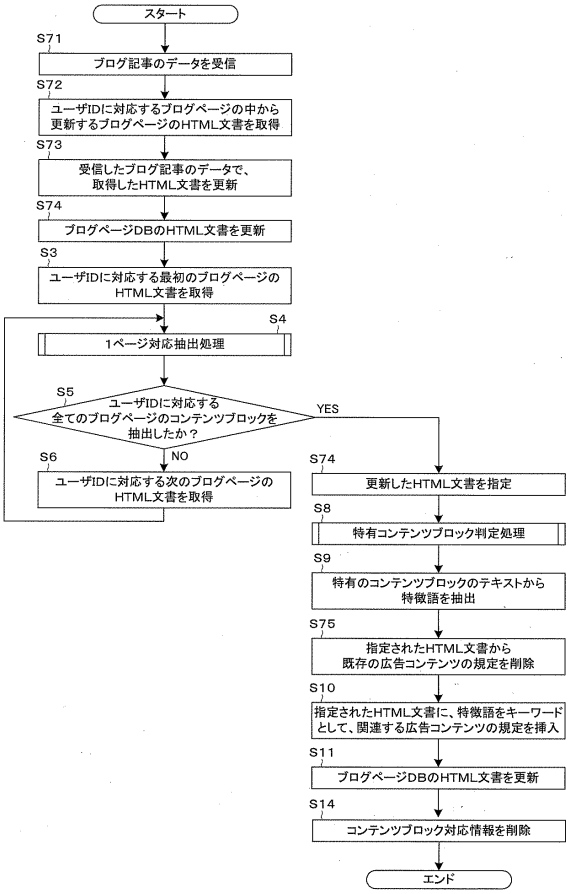
【図９】


ツリー探索処理
指定された
ノードの種類？
DIV又はTABLE
S31
NUM ← NUM + 1
LV ← LV + 1
S32
BN[LV]=NUM
S33
ブロック番号BN[LV]の
コンテンツブロック対応を初期化
S34
探索していない
子ノードが存在？
NO
YES
S35
探索していない子ノードの
1つを指定
S36
ツリー探索処理
S37
LV ← LV − 1
S38
リターン
テキスト
S39
ブロック番号BN[LV]の
コンテンツブロック対応情報に
テキストの内容を追加
IMG
S40
ブロック番号BN[LV]の
コンテンツブロック対応情報に
画像データのURLを追加
その他
S41
探索していない
子ノードが存在？
YES
S42
探索していない子ノードの
1つを指定
S43
ツリー探索処理
NO

【図１０】


特有コンテンツブロック判定処理
S61
HTML文書のDOMツリーを生成
S62
NUM ← 0 、 LV ← 0
S63
ルートノードを指定
S64
ツリー探索処理
S65
i ← 1
S66
ブロック番号iのコンテンツブロックの
出現頻度を計算
S67
出現頻度 ≦ 閾値？
NO
YES
S68
特有のコンテンツブロックの1つであると判定
S69
i ← i + 1
S70
i > NUM?
NO
YES
リターン

【図１１】


スタート
S71
ブログ記事のデータを受信
S72
ユーザIDに対応するブログページの中から
更新するブログページのHTML文書を取得
S73
受信したブログ記事のデータで、
取得したHTML文書を更新
S74
ブログページDBのHTML文書を更新
S3
ユーザIDに対応する最初のブログページの
HTML文書を取得
S4
1ページ対応抽出処理
S5
ユーザIDに対応する
全てのブログページのコンテンツブロックを
抽出したか？
YES
NO
S6
ユーザIDに対応する次のブログページの
HTML文書を取得
S74
更新したHTML文書を指定
S8
特有コンテンツブロック判定処理
S9
特有のコンテンツブロックのテキストから
特徴語を抽出
S75
指定されたHTML文書から
既存の広告コンテンツの規定を削除
S10
指定されたHTML文書に、特徴語をキーワード
として、関連する広告コンテンツの規定を挿入
S11
ブログページDBのHTML文書を更新
S14
コンテンツブロック対応情報を削除
エンド

フロントページの続き

(56)参考文献 特開２００５－３０２０４１（ＪＰ，Ａ）
国際公開第２００８／１０８５１５（WO，Ａ1）
特表２０１０－５３３８９７（ＪＰ，Ａ）
特開２００５－１２２６９０（ＪＰ，Ａ）
特開２００８－１３９９２８（ＪＰ，Ａ）
特開２００９－０５３９８３（ＪＰ，Ａ）
特開２００９－２０５４９９（ＪＰ，Ａ）
特開２００６－１４６５０６（ＪＰ，Ａ）
Ｓａｎｄｉｐ Ｄｅｂｎａｔｈ，Ｐｒａｓｅｎｊｉｔ Ｍｉｔｒａ，Ｎｉｒｍａｌ Ｐａｌ，ａｎｄ Ｃ．Ｌｅｅ Ｇｉｌｅｓ，“Ａｕｔｏｍａｔｉｃ Ｉｄｅｎｔｉｆｉｃａｔｉｏｎ ｏｆ Ｉｎｆｏｒｍａｔｉｖｅ Ｓｅｃｔｉｏｎｓ ｏｆ Ｗｅｂ Ｐａｇｅｓ”，ＩＥＥＥ ＴＲＡＮＳＡＣＴＩＯＮＳ ＯＮ ＫＮＯＷＬＥＤＧＥ ＡＮＤ ＤＡＴＡ ＥＮＧＩＮＥＥＲＩＮＧ，ＶＯＬ．１７，ＮＯ．９，ＳＥＰＴＥＭＢＥＲ ２００５，ｐｐ．１２３３－１２４６
吉田光男，山本幹雄，“教師情報を必要としないＷｅｂページ群のコンテンツ自動抽出ツールの提案”，Ｉｎｔｅｒｎｅｔ＜URL：ｈｔｔｐ：／／ｄｂ－ｅｖｅｎｔ．ｊｐｎ．ｏｒｇ／ｄｅｉｍ２００９／ｐｒｏｃｅｅｄｉｎｇｓ／ｆｉｌｅｓ／Ａ８－４．ｐｄｆ＞，第1回データ工学と情報マネジメントに関するフォーラム－ＤＥＩＭフォーラム－論文集，ＤＥＩＭフォーラム２００９，電子情報通信学会データ工学研究専門委員会，２００９年０５月０９日

(58)調査した分野(Int.Cl.，ＤＢ名)
G06Q10/00-50/00