(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2011-248762
(P2011-248762A)
(43)公開日　平成23年12月8日(2011.12.8)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	210D	5B075
		G06F	17/30	330Z	

審査請求 未請求　請求項の数 8　ＯＬ　（全 13 頁）

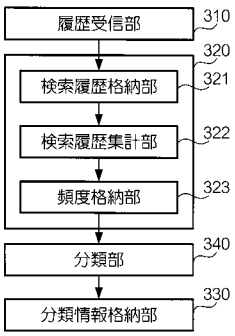
(21)出願番号　特願2010-123275 (P2010-123275)
(22)出願日　平成22年5月28日 (2010.5.28)

(71)出願人　392026693
株式会社エヌ・ティ・ティ・ドコモ
東京都千代田区永田町二丁目１１番１号
(74)代理人　110000752
特許業務法人朝日特許事務所
(72)発明者　佐々木　純
東京都千代田区永田町二丁目１１番１号
株式会社エヌ・ティ・ティ・ドコモ内

Ｆターム(参考)　5B075 NK02　NR12　PP22　PR03

(54)【発明の名称】分類装置、コンテンツ検索システム、コンテンツ分類方法、コンテンツ検索方法及びプログラム

(57)【要約】
【課題】コンテンツを、当該コンテンツの内容によらないで分類できるようにする。
【解決手段】分類情報格納部３３０は、カテゴリが既知のコンテンツに対応するカテゴリの情報を格納する。履歴受信部３１０は、コンテンツの検索履歴を受信する。検索履歴は、コンテンツを検索するために用いられた検索語を含む。コンテンツ情報格納部３２０は、検索履歴を格納するとともに、コンテンツが検索された頻度を検索語毎に集計し、その集計結果を格納する。分類部３４０は、分類情報格納部３３０に格納された情報と、コンテンツ情報格納部３２０に格納された情報とを用いて、カテゴリが未知のコンテンツを分類する。分類部３４０は、カテゴリが未知のコンテンツの検索に用いられた検索語とカテゴリが既知のコンテンツの検索に用いられた検索語（及びそのカテゴリ）とに基づいて、カテゴリが未知のコンテンツを分類する。
【選択図】図３



【特許請求の範囲】
【請求項１】
所定のカテゴリに従って分類された第１のコンテンツと、当該コンテンツを検索するために用いられた検索語とを対応付けた検索履歴を取得する第１の取得部と、
前記カテゴリが未知である第２のコンテンツを検索するために用いられた検索語を取得する第２の取得部と、
前記第１の取得部により取得された検索履歴と前記第２の取得部により取得された検索語とを比較することによって、前記第２のコンテンツを前記カテゴリに従って分類する分類部と
を備える分類装置。
【請求項２】
前記分類部は、前記検索語毎の検索された頻度に基づいて前記第２のコンテンツを分類する
ことを特徴とする請求項１に記載の分類装置。
【請求項３】
前記分類部は、検索された頻度に応じたスコアを前記検索語毎に算出し、当該スコアを用いて前記第２のコンテンツを分類する
ことを特徴とする請求項２に記載の分類装置。
【請求項４】
ＵＲＬ（Uniform Resource Locator）の少なくとも一部が共通する前記第１のコンテンツ又は前記第２のコンテンツどうしが同一のカテゴリに分類されることを特徴とする請求項１ないし３のいずれかに記載の分類装置。
【請求項５】
請求項１ないし４のいずれかに記載の分類装置と、ユーザにより入力された検索語に対応するコンテンツを当該ユーザに提示する検索装置とを有し、
前記検索装置が、前記入力された検索語に対応するコンテンツを前記カテゴリ毎に分類した態様で提示するための提示部を備える
ことを特徴とするコンテンツ検索システム。
【請求項６】
所定のカテゴリに従って分類された第１のコンテンツと、当該コンテンツを検索するために用いられた検索語とを対応付けた検索履歴を取得するステップと、
前記カテゴリが未知である第２のコンテンツを検索するために用いられた検索語を取得するステップと、
前記取得された検索履歴と前記取得された検索語とを比較することによって、前記第２のコンテンツを前記カテゴリに従って分類するステップと
を有することを特徴とするコンテンツ分類方法。
【請求項７】
所定のカテゴリに従って分類された第１のコンテンツと、当該コンテンツを検索するために用いられた検索語とを対応付けた検索履歴を取得するステップと、
前記カテゴリが未知である第２のコンテンツを検索するために用いられた検索語を取得するステップと、
前記取得された検索履歴と前記取得された検索語とを比較することによって、前記第２のコンテンツを前記カテゴリに従って分類するステップと、
入力された検索語に対応する前記第１のコンテンツ又は前記第２のコンテンツを前記カテゴリ毎に分類した態様でユーザに提示するステップと
を有することを特徴とするコンテンツ検索方法。
【請求項８】
コンピュータに、
所定のカテゴリに従って分類された第１のコンテンツと、当該コンテンツを検索するために用いられた検索語とを対応付けた検索履歴を取得するステップと、

前記カテゴリが未知である第２のコンテンツを検索するために用いられた検索語を取得するステップと、
前記取得された検索履歴と前記取得された検索語とを比較することによって、前記第２のコンテンツを前記カテゴリに従って分類するステップと
を実行させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、コンテンツを分類し、又は検索するための技術に関する。
【背景技術】
【０００２】
インターネットを介して所望のコンテンツを得るために、サーチエンジン（検索エンジン）が利用されている。Ｗｅｂページの検索は、例えば、ユーザが検索語（キーワード）を入力し、その検索語を文字列として含んでいるＷｅｂページをユーザに検索結果として提示する、といった手順で行われる。なお、検索語に対応するＷｅｂページが複数ある場合には、サーチエンジン毎の規則やアルゴリズムに従って順位付けが行われる。このような検索手法は、「キーワード検索」と呼ばれている。また、キーワード検索のほかにも、あらかじめ設定されたカテゴリに従ってＷｅｂサイトやＷｅｂページを検索する、いわゆる「カテゴリ検索」も知られている。
【０００３】
特許文献１には、検索に必要なキーワードを抽出するための技術が記載されている。また、特許文献２、３には、形態素解析等の自然言語処理を用いて文書を分類するための技術が記載されている。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】特開２００２－１４９６８３号公報
【特許文献２】特開平２－１５８８７１号公報
【特許文献３】特開平１１－３２８２１１号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
ところで、コンテンツを検索する場合には、１語や２語の検索語で検索することも多い。この傾向は、ユーザが文字入力に不慣れな場合や、文字入力に特化していないデバイス（例えば、携帯電話機など）で文字を入力する場合に、より顕著である。検索語そのものの情報量が少ない場合には、ユーザが意図していないコンテンツまでもが検索結果に含まれてしまうことも多い。
【０００６】
また、カテゴリ検索には、あらかじめ分類されているコンテンツでなければ検索結果として利用できないという問題があり、コンテンツの頻繁な追加や更新に対応することが困難である。さらに、自然言語処理のように、コンテンツの内容そのものを解析する場合には、その処理に時間を要するだけでなく、一つ一つのコンテンツの情報量（文字数）が増えるほど処理時間も増加してしまう。
【０００７】
そこで、本発明は、コンテンツを、当該コンテンツの内容によらないで分類できるようにすることを目的とする。
【課題を解決するための手段】
【０００８】
本発明の一態様に係る分類装置は、所定のカテゴリに従って分類された第１のコンテンツと、当該コンテンツを検索するために用いられた検索語とを対応付けた検索履歴を取得

する第１の取得部と、前記カテゴリが未知である第２のコンテンツを検索するために用いられた検索語を取得する第２の取得部と、前記第１の取得部により取得された検索履歴と前記第２の取得部により取得された検索語とを比較することによって、前記第２のコンテンツを前記カテゴリに従って分類する分類部とを備える。

【０００９】

好ましい態様において、前記分類部は、前記検索語毎の検索された頻度に基づいて前記第２のコンテンツを分類する。

さらに好ましい態様において、前記分類部は、検索された頻度に応じたスコアを前記検索語毎に算出し、当該スコアを用いて前記第２のコンテンツを分類する。

他の好ましい態様において、前記分類装置は、ＵＲＬ（Uniform Resource Locator）の少なくとも一部が共通する前記第１のコンテンツ又は前記第２のコンテンツどうしが同一のカテゴリに分類されることを特徴とする。

【００１０】

本発明の他の態様に係るコンテンツ検索システムは、前記分類装置と、ユーザにより入力された検索語に対応するコンテンツを当該ユーザに提示する検索装置とを有し、前記検索装置が、前記入力された検索語に対応するコンテンツを前記カテゴリ毎に分類した態様で提示するための提示部を備える。

【００１１】

本発明の他の態様に係るコンテンツ分類方法は、所定のカテゴリに従って分類された第１のコンテンツと、当該コンテンツを検索するために用いられた検索語とを対応付けた検索履歴を取得するステップと、前記カテゴリが未知である第２のコンテンツを検索するために用いられた検索語を取得するステップと、前記取得された検索履歴と前記取得された検索語とを比較することによって、前記第２のコンテンツを前記カテゴリに従って分類するステップとを有する。

【００１２】

本発明の他の態様に係るコンテンツ検索方法は、所定のカテゴリに従って分類された第１のコンテンツと、当該コンテンツを検索するために用いられた検索語とを対応付けた検索履歴を取得するステップと、前記カテゴリが未知である第２のコンテンツを検索するために用いられた検索語を取得するステップと、前記取得された検索履歴と前記取得された検索語とを比較することによって、前記第２のコンテンツを前記カテゴリに従って分類するステップと、入力された検索語に対応する前記第１のコンテンツ又は前記第２のコンテンツを前記カテゴリ毎に分類した態様でユーザに提示するステップとを有する。

【００１３】

本発明の他の態様に係るプログラムは、コンピュータに、所定のカテゴリに従って分類された第１のコンテンツと、当該コンテンツを検索するために用いられた検索語とを対応付けた検索履歴を取得するステップと、前記カテゴリが未知である第２のコンテンツを検索するために用いられた検索語を取得するステップと、前記取得された検索履歴と前記取得された検索語とを比較することによって、前記第２のコンテンツを前記カテゴリに従って分類するステップとを実行させるためのものである。

【発明の効果】

【００１４】

本発明によれば、コンテンツを、当該コンテンツの内容によらないで分類することが可能である。

【図面の簡単な説明】

【００１５】

【図１】本発明の一実施形態の全体構成を示す図

【図２】コンテンツ検索システムのハードウェア構成を示すブロック図

【図３】分類装置の機能的構成を示す機能ブロック図

【図４】検索履歴情報のデータ構造を例示する図

【図５】頻度情報のデータ構造を例示する図

【図６】分類情報のデータ構造を例示する図
【図７】分類装置によるコンテンツの分類方法を示すフローチャート
【図８】検索装置によるコンテンツの検索方法を示すフローチャート
【図９】分類装置及び記憶装置の構成（変形例）を示すブロック図
【発明を実施するための形態】
【００１６】
［実施形態］
図１は、本発明の一実施形態の全体構成を示す図である。本実施形態のコンテンツ検索システム１１０は、ユーザがコンテンツを検索するために用いられるものであり、複数のユーザのクライアント端末１２０によってアクセスされる。クライアント端末１２０は、通信ネットワーク１３０を介してコンテンツ検索システム１１０にアクセスするコンピュータ装置であり、例えば、パーソナルコンピュータ、携帯電話機、スマートフォンなどである。クライアント端末１２０は、通信ネットワーク１３０と通信を行う手段と、文字入力等の操作を受け付ける手段と、検索結果を表示する手段とを少なくとも備える。通信ネットワーク１３０は、インターネット、イントラネット、移動体通信網などであり、また、これらを組み合わせた複合的なネットワークであってもよい。
【００１７】
本実施形態において、コンテンツとは、ユーザが視覚的に閲覧可能な情報を含むひとまとまりのデータをいい、ここでは、文字コード（ASCII、Unicode、Shift_JIS等）によって記述された文字列を少なくとも含むものとする。コンテンツは、例えば、ＨＴＭＬ（HyperText Markup Language）形式のＷｅｂページであるが、ＰＤＦ（Portable Document Format）データをはじめ、通信ネットワーク１３０を介してやりとりが可能なさまざまなデータを含み得る。ここでいうＷｅｂページは、あらかじめ記憶されているものに限らず、ＣＧＩ（Common Gateway Interface）などによって動的に生成されたものであってもよい。コンテンツは、通信ネットワーク１３０に接続された図示せぬＷｅｂサーバに記憶されている。また、コンテンツは、音楽データのような視覚的なデータ以外のデータを含んでいてもよい。
【００１８】
コンテンツ検索システム１１０は、検索装置１１１と、分類装置１１２とを備える。検索装置１１１は、サーチエンジンの機能を有し、クライアント端末１２０から受け付けた検索クエリに応じた検索結果をクライアント端末１２０に提示する。検索クエリは、コンテンツを検索するために用いられる１又は複数の検索語を少なくとも含む。また、検索クエリは、クライアント端末１２０（又はそのユーザ）を識別する情報や、検索語以外の検索条件（検索の態様を指定する条件。前方一致検索、アンド検索など。）を含んでもよい。検索結果は、検索語に対応するコンテンツにリンクするＵＲＬを含んだＷｅｂページによってクライアント端末１２０に提供される。
【００１９】
分類装置１１２は、クライアント端末１２０のユーザによって検索されたコンテンツを当該コンテンツの特徴によって分類する機能を有する。分類装置１１２による分類は、あらかじめ決められたカテゴリを用いて行われる。本実施形態におけるカテゴリは、例えば、「スポーツ」、「映画」、「ギャンブル」、「音楽」、「占い」、「グルメ」といったものである。なお、カテゴリは、コンテンツ検索システム１１０の利用目的や利用するユーザに応じて適宜定められればよい。
【００２０】
図２は、コンテンツ検索システム１１０のハードウェア構成を示すブロック図である。検索装置１１１は、図２に示すように、制御部２１１と、記憶部２１２と、第１通信部２１３と、第２通信部２１４とを備える。制御部２１１は、検索装置１１１の各部の動作を制御する手段である。制御部２１１は、ＣＰＵ（Central Processing Unit）等の演算処理装置や主記憶装置に相当する記憶手段（メインメモリ）を備え、プログラムを実行することによって検索機能を実現する。ここにおいて、検索機能とは、検索語に応じた検索結

果を提示する機能をいい、本発明に係る提示部に相当するものである。記憶部２１２は、ＨＤＤ（Hard Disk Drive）等の補助記憶装置に相当する記憶手段を備え、検索機能に必要なデータ（例えば、いわゆるインデクス等）を記憶する。第１通信部２１３は、分類装置１１２と通信するためのインターフェースを備える。第２通信部２１４は、通信ネットワーク１３０と通信するためのインターフェースである。

【００２１】

分類装置１１２は、図２に示すように、制御部２２１と、記憶部２２２と、第１通信部２２３とを備える。これらの各部は、検索装置１１１の同名の構成要素と同様のハードウェア構成を有する。ただし、記憶部２２２は、その記憶するデータが記憶部２１２とは相違し、コンテンツの分類に必要なデータを記憶している。

【００２２】

図３は、分類装置１１２の機能的構成を示す機能ブロック図である。分類装置１１２は、プログラムを実行することにより、図３に示す履歴受信部３１０、コンテンツ情報格納部３２０、分類情報格納部３３０及び分類部３４０の各部に相当する機能を実現する。また、コンテンツ情報格納部３２０は、より詳細には、検索履歴格納部３２１、検索履歴集計部３２２及び頻度格納部３２３に機能的に分類される。

【００２３】

履歴受信部３１０は、検索装置１１１から検索履歴を受信する。ここにおいて、検索履歴とは、検索語と、その検索語を用いて検索され、ユーザが閲覧するためにクライアント端末１２０で選択したコンテンツとを対応付けたものをいう。ここでいう「選択」とは、ユーザの操作（クリック等）に基づくものである。履歴受信部３１０により受信されるコンテンツは、分類の対象であるコンテンツ、すなわち、まだ分類されておらず、カテゴリが未知であるコンテンツと、カテゴリが既知であるコンテンツの双方が含まれ得る。

【００２４】

コンテンツ情報格納部３２０は、記憶部２２２にコンテンツ情報を格納する。コンテンツ情報は、検索履歴情報と頻度情報の総称である。検索履歴情報は、検索履歴を表すデータである。検索履歴格納部３２１は、履歴受信部３１０により受信された検索履歴に基づき、記憶部２２２に検索履歴情報を格納する。

【００２５】

頻度情報は、あるコンテンツがどの検索語によって何回検索されたかを表すデータである。本実施形態において、頻度情報は、コンテンツを表すＵＲＬと、当該ＵＲＬに対応する検索語と、その検索された頻度とを対応付けたデータである。なお、ここでいう「頻度」は、検索クエリとして得られた延べ回数ではなく、例えば、同一のユーザが同一の検索語で同一のコンテンツを繰り返し何回も検索した場合には、これらを１回の検索とみなした値であってもよい。この場合、頻度の値は、単位時間当たりにあるコンテンツをある検索語で検索したユーザの人数に相当する。

【００２６】

頻度情報は、検索履歴集計部３２２によって集計され、頻度格納部３２３によって記憶部２２２に格納される。検索履歴集計部３２２は、検索履歴格納部３２１によって格納された検索履歴情報を読み出し、その検索履歴情報の中で重複しているコンテンツを検索語毎に集計し、頻度を算出する。頻度格納部３２３は、検索履歴集計部３２２による集計結果を記憶部２２２に格納する。

【００２７】

図４は、検索履歴情報のデータ構造を例示する図である。この例において、検索履歴情報は、ユーザＩＤと、タイムスタンプと、検索語と、ＵＲＬとを対応付けて記述したデータである。ユーザＩＤは、ユーザを一意的に識別するためのデータである。ユーザの識別は、パスワード等による認証やCookieの利用など、周知の適当な技術によって実現されればよい。タイムスタンプは、ユーザが検索を行った時刻（以下「検索時刻」という。）を表すデータである。検索時刻は、クライアント端末１２０で検索クエリの送信時に計測されてもよいし、検索クエリの受信時に検索装置１１１で計測されてもよい。

図５は、頻度情報のデータ構造を例示する図である。この例において、頻度情報は、ＵＲＬと、検索語と、頻度とを対応付けて記述したデータである。
なお、検索履歴情報又は頻度情報は、コンテンツが分類済みであるか否か（すなわち、カテゴリが既知か未知か）を表すフラグを含んでいてもよい。
【００２８】
分類情報格納部３３０は、コンテンツの分類結果を表す分類情報を記憶部２２２に格納する。分類情報は、記憶部２２２にあらかじめ記憶されているものと、分類部３４０による分類結果として事後的に格納されるものとがある。本実施形態の分類情報は、ＵＲＬと、カテゴリＩＤと、スコアとを対応付けて記述したデータである。カテゴリＩＤは、既知のカテゴリを一意的に識別するために割り当てられるデータである。また、スコアは、ＵＲＬによって表されるコンテンツがそれぞれのカテゴリに適合する度合いを示す値であり、本実施形態においては、「１」が最大（最も適合する）で「０」が最小であるとする。
【００２９】
図６は、分類情報のデータ構造を例示する図である。この例においては、ＵＲＬ「http://aaa.com/aaa.cgi?a1=1&a2=2」で表されるコンテンツは、カテゴリＩＤ「２４」で表されるカテゴリに最も適合している（相応しい）ということになる。なお、カテゴリは、一つのコンテンツにつき１種類でなくてもよい。例えば、コンテンツは、所定のスコア以上のカテゴリのすべてに属するとしてもよいし、スコアの大きい順に所定数のカテゴリに属するとしてもよい。また、図６においては、スコアが「０」である分類情報の表示を省略しているが、実際のデータは、各ＵＲＬと各カテゴリによって考えられる組み合わせのそれぞれについて、「０」かそれ以外のスコアが付与されているものとする。
【００３０】
分類部３４０は、分類情報格納部３３０に格納された分類情報を用いて、カテゴリが未知であるコンテンツを分類し、その分類結果を分類情報格納部３３０に供給する。分類部３４０は、カテゴリが既知のコンテンツの検索履歴とカテゴリが未知のコンテンツを検索するために用いられた検索語とを比較することによって、カテゴリが未知のコンテンツを分類する。
【００３１】
本実施形態において、分類部３４０は、コンテンツ情報格納部３２０に格納されたコンテンツ情報からカテゴリが未知であるコンテンツに関するデータを取得し、これを分類の対象とする。次に、分類部３４０は、分類情報格納部３３０を介して分類情報を取得し、分類モデルを生成する。ここにおいて、分類モデルとは、コンテンツがどのカテゴリに属するのが適当であるかを数学的にモデル化したものである。分類部３４０は、分類モデルを用いて、分類対象のコンテンツのそれぞれにどのカテゴリが適しているかを算出し、その算出結果を分類情報として分類情報格納部３３０に供給する。分類モデルは、例えば、ＳＶＭ（Support Vector Machine）やニューラルネットワークを用いて実現可能である。
【００３２】
以上の構成のもと、コンテンツ検索システム１１０は、クライアント端末１２０からの要求（すなわち検索クエリ）に応じて検索を実行し、検索結果をクライアント端末１２０に提示する。また、コンテンツ検索システム１１０は、適当なタイミングで、カテゴリが未知のコンテンツに関する情報を取得し、コンテンツの分類を繰り返し実行する。
また、コンテンツ検索システム１１０は、周知のキーワード検索による検索結果と、コンテンツをカテゴリ毎に分類して表示する検索結果とをユーザに提供することができる。コンテンツ検索システム１１０は、これらの検索結果を両方提示してもよいし、ユーザによる事前の設定に応じていずれかの検索結果を択一的に提示するようにしてもよい。
【００３３】
図７は、分類装置１１２によるコンテンツの分類方法を示すフローチャートである。図７に示すように、分類装置１１２の制御部２２１は、最初に既知の分類情報を格納する（ステップＳ１１）。ステップＳ１１において、制御部２２１は、カテゴリに関する既存の情報源を利用して分類情報を得る。かかる情報源としては、例えば、ＯＤＰ（Open Direc

tory Project）に登録されているＷｅｂサイトを利用可能である。あるいは、周知の代表的なサーチエンジンを用いてカテゴリ名によって検索を実行し、その検索結果の上位に提示されるＷｅｂサイトを分類情報に利用したり、既存のポータルサイトによって分類されているカテゴリを分類情報に利用したりすることも可能である。なお、分類装置１１２が用いるカテゴリと既存の情報源によるカテゴリとに相違がある場合には、あらかじめ対応表を作成しておき、そのときに用いるカテゴリに置き換えられるようにすることが望ましい。また、分類情報のうち、スコアは、適当な値があらかじめ機械的に（あるいは手作業で）設定される。

【００３４】

次に、制御部２２１は、検索履歴を検索装置１１１から受信し、これを検索履歴情報として格納する（ステップＳ１２）。制御部２２１は、検索クエリが発生する毎に検索履歴を受信してもよいし、検索履歴を所定時間毎又は所定数毎に一括して受信する動作を繰り返してもよい。このような動作が繰り返されることにより、検索履歴情報が徐々に蓄積される。一定量の検索履歴情報が蓄積されると、制御部２２１は、頻度情報をコンテンツ毎に集計し、格納する（ステップＳ１３）。

【００３５】

続いて、制御部２２１は、このようにして格納された分類情報と頻度情報とに基づき、分類モデルを生成する（ステップＳ１４）。分類モデルの具体的な生成方法の一例は、ＳＶＭによる回帰分析を行う場合であれば、以下のとおりである。

【００３６】

制御部２２１は、カテゴリが既知であるコンテンツの特徴量として、頻度情報をベクトル形式で表現したX_iを用いる。X_iは、具体的には、以下の（１）式で表される。ここにおいて、x_{ij}は、各コンテンツ及び検索語を一意的に識別するためのＩＤをそれぞれｉ、ｊとした場合の頻度の値であり、コンテンツの総数をＮ、検索語の総数をＶとすると、$1 \leqq i \leqq N$、$1 \leqq j \leqq V$をそれぞれ満たす。つまり、X_iは、コンテンツの検索語毎の頻度をコンテンツ毎に表すものである。

$$X_i = (x_{i1}, x_{i2}, x_{i3}, \cdots, x_{iV}) \quad \cdots (1)$$

【００３７】

なお、制御部２２１は、必要に応じて、X_iに対数化や正規化を行ってもよい。対数化や正規化は、x_{ij}のそれぞれの差が大きすぎ、大小関係の特徴が強く現れすぎる場合に有効である。対数化は、例えば、x_{ij}を$\log_{10}(x_{ij}+1)$、すなわち１０を底とする$(x_{ij}+1)$の対数に置き換えることで実現可能である。また、正規化は、X_iの絶対値が１となるように、各要素（x_{ij}）を要素の二乗和で除算することで実現可能である。

【００３８】

また、制御部２２１は、カテゴリの特徴量として、分類情報をベクトル形式で表現したY_iを用いる。Y_iは、具体的には、以下の（２）式で表される。ここにおいて、y_{ik}は、コンテンツのＩＤをｉ、カテゴリＩＤをｋとした場合のスコアである。カテゴリＩＤは、カテゴリの総数をＬとすると、$1 \leqq k \leqq L$を満たす。つまり、Y_iは、コンテンツのカテゴリ毎のスコアをコンテンツ毎に表すものである。

$$Y_i = (y_{i1}, y_{i2}, y_{i3}, \cdots, y_{iL}) \quad \cdots (2)$$

【００３９】

制御部２２１は、これらの特徴量から分類モデルを生成する。分類モデルは、X_iを説明変数、Y_iを目的変数とするものであり、以下においては$F_k(X_i)$と表現する。$F_k(X_i)$は、カテゴリ毎（すなわちカテゴリＩＤ毎）に生成される。制御部２２１は、このようにして生成される分類モデル（$F_k(X_i)$）に対して、カテゴリが未知であるコンテンツの特徴量であるX_i'を入力する（なお、X_iとX_i'の相違点は、カテゴリが既知であるか未知であるかの１点のみである。）。そうすると、制御部２２１は、カテゴリが未知であるコンテンツの目的変数として、上述したY_iに相当する特徴量を得ることができる。

【００４０】

制御部２２１は、このようにして生成された分類モデルを用いて、カテゴリが未知であるコンテンツについてカテゴリ毎のスコアを算出し、その算出結果を分類情報として格納する（ステップＳ１５）。なお、制御部２２１は、スコアに対して閾値を設定し、閾値を下回るスコアをすべて「０」であるとみなしてもよい。このようにすることで、実態に即していない分類がされることを防ぐことが可能である。また、制御部２２１は、タイムスタンプに基づいて頻度又はスコアに対して重み付けを行い、より新しい検索履歴ほど分類に強い影響を与えるようにしてもよい。

【００４１】

制御部２２１は、一定量の検索履歴が新たに蓄積される毎に、あるいは一定間隔で、図７に示す分類処理を繰り返す。ただし、制御部２２１は、いったん分類情報が格納された後には、ステップＳ１１の処理をスキップし、ステップＳ１２の処理から実行すればよい。また、ステップＳ１５の処理によって分類情報が新たに格納されたコンテンツは、その後はカテゴリが既知のコンテンツとして扱われる。

【００４２】

図８は、検索装置１１１によるコンテンツの検索方法を示すフローチャートである。図８に示すように、検索装置１１１の制御部２１１は、ユーザから検索クエリを受信することにより、検索語を取得する（ステップＳ２１）。制御部２１１は、取得した検索語に基づき、周知のキーワード検索によってコンテンツのリストを生成する（ステップＳ２２）。ステップＳ２２において生成されるリストは、検索語を含むコンテンツを適当に順位付けしたものである。

【００４３】

次に、制御部２１１は、分類装置１１２に格納されている分類情報を参照することにより、ステップＳ２２において生成されたリストに含まれるコンテンツをカテゴリ毎に分類する（ステップＳ２３）。そして、制御部２１１は、ユーザが入力した検索語に対応するコンテンツをカテゴリ毎に分類した態様で表示するための検索結果情報を当該ユーザのクライアント端末１２０に送信する（ステップＳ２４）。検索結果情報の表示態様は、例えば、各カテゴリのコンテンツのリストをスコアが高い順に表示するものであってもよいし、各カテゴリのコンテンツのリストをタブで切り替えて表示するものであってもよい。なお、制御部２１１は、コンテンツに対応する要約文を検索結果情報に含めてもよい。このようにすれば、ユーザによる各コンテンツの取捨選択を容易にすることができる。また、制御部２１１は、カテゴリが未知のコンテンツが検索結果に含まれる場合などには、必要に応じて、キーワード検索を行っただけの（未分類の）リストを検索結果情報に含めてもよい。

【００４４】

検索結果情報には、目的のコンテンツに直接リンクするＵＲＬではなく、いったんコンテンツ検索システム１１０を経由して目的のコンテンツにリンクするＵＲＬ（いわゆるリダイレクトＵＲＬ）が記述されている。このＵＲＬには、パラメータとして、目的のコンテンツのＵＲＬに加え、ユーザが入力した検索語が含まれる。このようにすることで、コンテンツ検索システム１１０は、コンテンツを検索するために用いられた検索語を特定し、検索履歴を取得することが可能である。

【００４５】

以上のとおり、本実施形態によれば、コンテンツを検索するために用いられた検索語を利用することで、コンテンツの内容によらない分類を行うことが可能となる。よって、本実施形態によれば、コンテンツに対して自然言語処理を実行することなくコンテンツを分類することが可能となる。ただし、コンテンツの内容によらない分類が有効に機能するためには、当該コンテンツの検索が既にある程度行われていることが条件となる。

【００４６】

また、本実施形態によれば、コンテンツを検索するために実際に用いられた検索語を利用することで、各ユーザの意図に即した分類を行うことが可能であるともいえる。例えば、携帯電話機向けのコンテンツにおいては、画面サイズ等の表示上の制約から、文字の情

報量が比較的少ない場合がある。かかるコンテンツをその内容（すなわち、コンテンツに含まれる文字列）に基づいて分類した場合、その分類結果にユーザの意図を的確に反映させられず、分類の精度・確度が低下する可能性がある。よって、コンテンツ検索システム１１０は、文字列としての情報量が比較的少ないコンテンツの検索に適用するのに好適であるといえる。

【００４７】

［変形例］

上述した実施形態は、本発明の実施の一態様である。本発明は、上述した実施形態に対して以下の変形を適用した態様で実施することも可能である。なお、以下に示す変形例は、必要に応じて、各々を適当に組み合わせて実施されてもよいものである。

【００４８】

（変形例１）

上述したとおり、本発明に係るコンテンツは、その分類に際して自然言語処理が不要である。すなわち、本発明に係るコンテンツは、文字列を含まないデータであっても分類可能である。したがって、本発明は、文字列を含むか否かを問わず、ユーザが検索可能なあらゆるデータを分類の対象にすることができる。

【００４９】

（変形例２）

本発明に係る分類装置は、ＵＲＬの少なくとも一部が共通する複数のコンテンツがある場合に、これらが同一のカテゴリに分類されるように動作するものであってもよい。このようにすれば、カテゴリが未知のコンテンツをより高速に分類することが可能であるとともに、検索されたことが一度もないコンテンツであっても分類が可能になる場合がある。なお、本例において、カテゴリの共通化は、ドメイン名、ホスト名、ＦＱＤＮ（Fully Qualified Domain Name）などを単位として行われる。また、ＵＲＬがパス名（ディレクトリ名）によって階層分けされている場合には、かかる階層がカテゴリに対応付けられていてもよい。さらに、ＵＲＬにパラメータ部（図４等の例における「？」以降の文字列）が含まれる場合には、パラメータ部以外の部分が共通しているＵＲＬのカテゴリを共通化することも可能である。

【００５０】

（変形例３）

上述した実施形態においては、コンテンツの分類処理は、カテゴリが未知のコンテンツに対してのみ実行された。しかし、本発明は、カテゴリが既知のコンテンツにも分類処理を実行し、必要に応じて、コンテンツのカテゴリを変更できるようにしてもよい。このようにすれば、コンテンツの内容が時間の経過に応じて（ＵＲＬは変えずに）変更されたとしても、より適切なカテゴリに再分類することが可能となる。具体的には、制御部２２１は、カテゴリが既知のコンテンツに対して上述した分類モデルを適用し、既に算出されたスコアと新たに算出されたスコアとの間に一定以上の乖離がある場合に、既に算出されたスコアを新たに算出されたスコアに書き換えるようにしてもよい。

【００５１】

（変形例４）

本発明は、コンテンツ情報や分類情報の記憶手段を分類装置から分離した態様でも実施可能である。すなわち、本発明に係る分類装置は、コンテンツ情報や分類情報を格納する手段を別体に構成したものであってもよい。

【００５２】

図９は、本例に係る分類装置及び記憶装置の構成を示すブロック図である。図９に示すように、分類装置９１０は、第１取得部９１１と、第２取得部９１２と、分類部９１３と、供給部９１４とを備える。また、記憶装置９２０は、コンテンツ情報を記憶する手段であり、上述した実施形態のコンテンツ情報格納部３２０に相当する機能を少なくとも備える。記憶装置９３０は、分類情報を記憶する手段であり、上述した実施形態の分類情報格納部３３０に相当する機能を少なくとも備える。なお、記憶装置９２０は、分類装置９１

０を実施する事業者とは異なる事業者のサーチエンジンであってもよい。

【００５３】

第１取得部９１１は、記憶装置９２０からコンテンツ情報（検索履歴情報、頻度情報）を取得する。第２取得部９１２は、記憶装置９３０から分類情報を取得する。分類部９１３は、上述した実施形態の分類部３４０に相当する機能を備える。供給部９１４は、分類部９１３による分類結果を記憶装置９３０に供給する。

【００５４】

（変形例５）

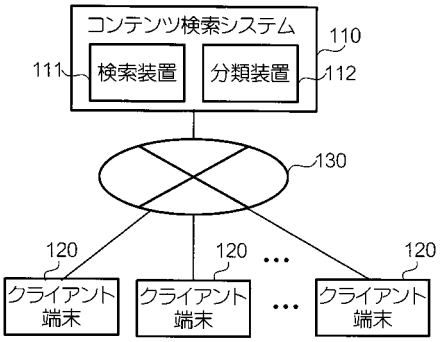
本発明は、上述した実施形態のように分類装置と検索装置とを別体にするのではなく、これらを単一のコンピュータ装置で実現することも可能である。また、本発明は、コンテンツを分類し、又は検索するための方法や、コンピュータ装置を上述した分類装置や検索装置として機能させるためのプログラムとしても提供可能である。かかるプログラムは、光ディスク等の記録媒体に記録した形態で提供されたり、インターネット等のネットワークを介して、コンピュータにダウンロードさせ、これをインストールして利用可能にするなどの形態で提供されたりすることも可能である。

【符号の説明】

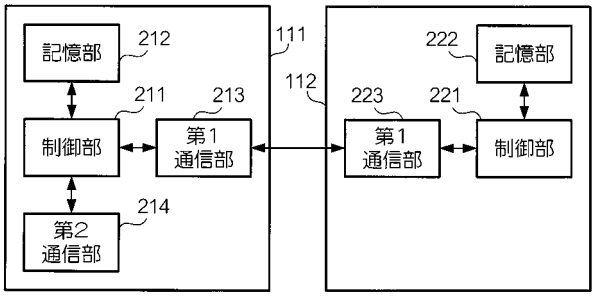
【００５５】

１１０…コンテンツ検索システム、１１１…検索装置、１１２…分類装置、１２０…クライアント端末、１３０…通信ネットワーク、２１１、２２１…制御部、２１２、２２２…記憶部、２１３、２２３…第１通信部、２１４…第２通信部、３１０…履歴受信部、３２０…コンテンツ情報格納部、３２１…検索履歴格納部、３２２…検索履歴集計部、３２３…頻度格納部、３３０…分類情報格納部、３４０…分類部、９１０…分類装置、９１１…第１取得部、９１２…第２取得部、９１３…分類部、９１４…供給部、９２０、９３０…記憶装置

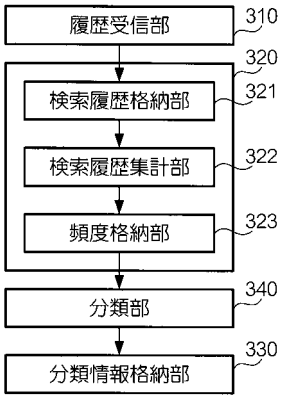
【図１】



【図２】



【図３】



【図４】

ユーザID	タイムスタンプ	検索語	URL
000	2010/5/28 07:00:00	A	http://aaa.com/aaa.cgi?a1=1&a2=2
000	2010/5/28 07:31:15	B	http://bbb.com/bbb.html
000	2010/5/28 08:02:30	C	http://bbb.com/bbb.cgi?b1=1
001	2010/5/28 09:35:10	D	http:/ddd.co.jp/ddd/ddd.gif
001	2010/5/28 12:40:02	E	http://eee.com/eee.cgi?e1=1;jsessionid=111
002	2010/5/28 08:24:10	F	http://fff.com/fff.html
002	2010/5/28 08:27:28	G	http://fff.com/fff/fff/fff/fff.cgi?id=fff
⋮	⋮	⋮	⋮

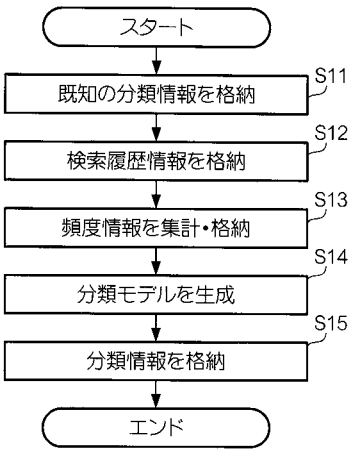
【図５】

URL	検索語	頻度
http://aaa.com/aaa.cgi?a1=1&a2=2	A	34
http://aaa.com/aaa.cgi?a1=1&a2=2	C	664
http://aaa.com/aaa.cgi?a1=1&a2=2	D	65
http://aaa.com/aaa.cgi?a1=1&a2=2	F	216
http://aaa.com/aaa.cgi?a1=1&a2=2	H	45
http://bbb.com/bbb.html	B	74
http://bbb.com/bbb.html	C	57
⋮	⋮	⋮

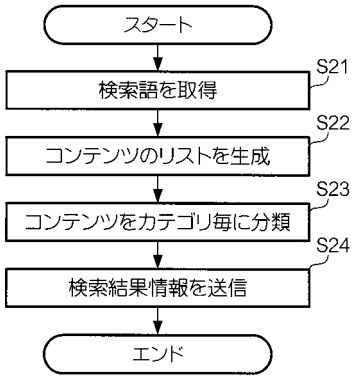
【図６】

URL	カテゴリID	スコア
http://aaa.com/aaa.cgi?a1=1&a2=2	1	0.45
http://aaa.com/aaa.cgi?a1=1&a2=2	5	0.81
http://aaa.com/aaa.cgi?a1=1&a2=2	24	1.00
http://bbb.com/bbb.html	3	0.26
http://bbb.com/bbb.html	9	0.71
http://bbb.com/bbb.html	11	0.57
http://bbb.com/bbb.html	41	0.98
⋮	⋮	⋮

【図７】



【図８】



【図９】