

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第4875911号
(P4875911)

(45) 発行日 平成24年2月15日 (2012.2.15)

(24) 登録日 平成23年12月2日 (2011.12.2)

(51) Int.Cl.

F I

G 0 6 F 17/30 (2006.01)

G 0 6 F 17/30 3 4 O A

G 0 6 F 17/30 3 5 O C

請求項の数 6 (全 18 頁)

(21) 出願番号 特願2006-76501 (P2006-76501)
 (22) 出願日 平成18年3月20日 (2006.3.20)
 (65) 公開番号 特開2007-256992 (P2007-256992A)
 (43) 公開日 平成19年10月4日 (2007.10.4)
 審査請求日 平成20年12月2日 (2008.12.2)

(73) 特許権者 591117192
 ニフティ株式会社
 東京都品川区南大井6-26-1
 (73) 特許権者 000005223
 富士通株式会社
 神奈川県川崎市中原区上小田中4丁目1番
 1号
 (74) 代理人 100103528
 弁理士 原田 一男
 (72) 発明者 鎌仲 敏勝
 東京都品川区南大井六丁目26番1号 ニ
 フティ株式会社内
 (72) 発明者 松尾 亜紀
 東京都品川区南大井六丁目26番1号 ニ
 フティ株式会社内

最終頁に続く

(54) 【発明の名称】 コンテンツ特定方法及び装置

(57) 【特許請求の範囲】

【請求項1】

登録ユーザがコンテンツにアクセスした時刻であるアクセス時刻を含む、当該登録ユーザのアクセスログを格納するアクセスログ格納部と収集したコンテンツ中のキーワードに関するデータを格納するコンテンツプロファイル・データベースとに格納されているデータから、アクセス時刻及び前記キーワードに関する情報と登録ユーザとの関係を表すトランザクション・データを生成し、トランザクション・データベースに登録するステップと、

前記トランザクション・データベースに格納された未処理のトランザクション・データに係る特定の登録ユーザに関連し且つ当該未処理のトランザクション・データに含まれるキーワードの、アクセス時刻における評価値と、登録ユーザとキーワードとのこれまでの関連度を表すデータを格納するユーザプロファイル・データベースに格納されているデータから前記特定の登録ユーザに関連するキーワードにつき前記アクセス時刻における減衰された関連度とを算出して、前記特定の登録ユーザに関連するキーワードについて前記評価値及び前記減衰された関連度から前記アクセス時刻における関連度を算出し、前記ユーザプロファイル・データベースを更新する更新ステップと、

前記アクセスログ格納部に格納されているデータを用いて、所定の基準を超えてアクセスが増加したコンテンツを特定し、当該特定されたコンテンツについてのデータを前記コンテンツプロファイル・データベースから抽出し、トピック・データベースに登録する登録ステップと、

10

20

前記ユーザプロフィール・データベースに格納されている、前記特定の登録ユーザについての前記関連度が上位のキーワードと所定の類似性を有し且つ前記トピック・データベースに登録されているキーワードが出現するコンテンツを特定し、当該特定されたコンテンツの識別情報を前記特定の登録ユーザに対応して推薦トピック・データベースに登録するコンテンツ特定ステップと、

を含み、

前記トランザクション・データベースに格納された前記キーワードに関するデータが、当該キーワードの提示回数 k を含み、

前記更新ステップが、

前記未処理のトランザクション・データに含まれるキーワードの提示回数 k と所定の減衰係数 ρ によって、前記キーワードの前記評価値を $(1 - \rho^k) / (1 - \rho)$ として算出するステップ、

を含み、コンピュータにより実行されるコンテンツ特定方法。

【請求項 2】

前記コンテンツ特定ステップが、

関連語辞書から、前記ユーザプロフィール・データベースに格納されている、前記特定の登録ユーザについての前記関連度が上位のキーワードに対応して登録されている関連キーワードを抽出するステップと、

前記特定の登録ユーザについての特定のキーワードと当該特定のキーワードに対応し且つ抽出された前記関連キーワードとを含む第 1 のセットと、前記トピック・データベースに登録されている前記キーワードをコンテンツ毎にまとめた第 2 のセットとの類似度を前記コンテンツ毎に算出するステップと、

を含む請求項 1 記載のコンテンツ特定方法。

【請求項 3】

前記ユーザプロフィール・データベースには、キーワード毎に処理基準日時のデータが登録されており

前記更新ステップが、

前記処理基準日時から前記アクセス時刻までの単位時間数 t と所定減衰係数 ρ と前記これまでの関連度 g とによって、前記アクセス時刻における減衰された関連度を $\rho^t g$ として算出するステップ、

を含む請求項 1 又は 2 記載のコンテンツ特定方法。

【請求項 4】

前記登録ステップが、

各前記コンテンツにつき、処理基準時刻のアクセスユーザ数の、1 単位時間前までのアクセスユーザ数の平均からの上方乖離度を算出するステップと、

前記上方乖離度が上位所定数内のコンテンツを特定するステップと、

を含む請求項 1 乃至 3 のいずれか 1 つ記載のコンテンツ特定方法。

【請求項 5】

請求項 1 乃至 4 のいずれか 1 つ記載のコンテンツ特定方法をコンピュータに実行させるためのプログラム。

【請求項 6】

登録ユーザがコンテンツにアクセスした時刻であるアクセス時刻を含む、当該登録ユーザのアクセスログを格納するアクセスログ格納部と収集したコンテンツ中のキーワードに関するデータを格納するコンテンツプロフィール・データベースとに格納されているデータから、アクセス時刻及び前記キーワードに関する情報と登録ユーザとの関係を表すトランザクション・データを生成し、トランザクション・データベースに登録する手段と、

前記トランザクション・データベースに格納された未処理のトランザクション・データに係る特定の登録ユーザに関連し且つ当該未処理のトランザクション・データに含まれるキーワードの、アクセス時刻における評価値と、登録ユーザとキーワードとのこれまでの関連度を表すデータを格納するユーザプロフィール・データベースに格納されているデー

10

20

30

40

50

タから前記特定の登録ユーザに関連するキーワードにつき前記アクセス時刻における減衰された関連度とを算出して、前記特定の登録ユーザに関連するキーワードについて前記評価値及び前記減衰された関連度から前記アクセス時刻における関連度を算出し、前記ユーザプロフィール・データベースを更新する更新手段と、

前記アクセスログ格納部に格納されているデータを用いて、所定の基準を超えてアクセスが増加したコンテンツを特定し、当該特定されたコンテンツについてのデータを前記コンテンツプロフィール・データベースから抽出し、トピック・データベースに登録する手段と、

前記ユーザプロフィール・データベースに格納されている、前記特定の登録ユーザについての前記関連度が上位のキーワードと所定の類似性を有し且つ前記トピック・データベースに登録されているキーワードが出現するコンテンツを特定し、当該特定されたコンテンツの識別情報を前記特定の登録ユーザに対応して推薦トピック・データベースに登録する手段と、

を有し、

前記トランザクション・データベースに格納された前記キーワードに関するデータが、当該キーワードの提示回数 k を含み、

前記更新手段が、

前記未処理のトランザクション・データに含まれるキーワードの提示回数 k と所定の減衰係数 ρ によって、前記キーワードの前記評価値を $(1 - \rho^k) / (1 - \rho)$ として算出する

コンテンツ特定装置。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ユーザに適切なコンテンツを特定又は抽出するための技術に関する。

【背景技術】

【0002】

従来、インターネット上のコンテンツから目的のものを探し出すためには、検索エンジンが利用されてきた。検索エンジンを利用するには具体的な検索語を与える必要があり、具体的な事物を調査するのに適している。

【0003】

また、近年RSSリーダによる情報収集も注目を浴びている。このRSSリーダは、ウェブ(Web)サイトの新着、更新情報を受信するのに適しており、予め定めたカテゴリにWebページを分類した後、カテゴリの注目度やカテゴリ中の注目ページを提示する機能を備えたものも存在する。

【0004】

また、既に検索語のバースト(急激な頻出)を検出する技術が存在しているが、これを用いれば特定の検索エンジンの利用者たちの注目トピックが分かる。しかしながら、特定の個人の関心と、この注目トピックは通常異なる。また、利用者は検索語を明示的に入力する必要がある。

【0005】

さらに、既に実施されているパーソナライズ検索では、利用者が入力した過去の検索語や参照したページを後日照会する事ができる。しかし、最近の検索クエリは、検索語の選別を試行錯誤した結果、よく似た検索クエリばかりが蓄積されるという問題がある。従って、定期的に検索される語などが、似たような検索クエリによって記憶領域から押し出されてしまうということが生じる。

【0006】

なお、特開2002-14996号公報には、インターネット上のリソースを対象として、新規ドキュメントをユーザの興味領域に沿った形で提示するための技術が開示されている。そして、各ユーザのブックマーク情報は、ブックマークサーバで一元管理される。

10

20

30

40

50

ユーザはクライアント装置からブックマーク操作部を介してブックマークの操作をユーザブックマークDBに対して行うことができる。ブックマークサーバは定期的にユーザ嗜好抽出部を用いてユーザブックマークDB中の個々のユーザのブックマーク情報に基づいて、分類フォルダ毎の嗜好情報を抽出する。新規ドキュメント提案部は、各ユーザの分類フォルダ毎の嗜好情報に応じて、インターネット上のディレクトリサーバに対する検索の結果や、他の外部から与えられたドキュメント集合から適当な新規ドキュメントをユーザブックマークの一部としてユーザブックマークDBに登録するものである。但し、新規ドキュメントはユーザの嗜好に合わせられるだけで、他の観点はない。

【特許文献1】特開2002-14996号公報

【発明の開示】

【発明が解決しようとする課題】

【0007】

しかし、ユーザ自身が検索語を明確に把握していない場合には検索エンジンから適切なコンテンツを抽出するのは不可能である。また、RSSリーダーでは特定のサイトを定点観測するには適しているが、サイトが取り扱う内容とユーザの興味とは常に一致しているわけではない。また、広く世の中で注目されているサイトとユーザ自身の興味とを重ね合わせて考慮するような仕組みは存在していない。

【0008】

本発明は以上の問題を鑑みてなされたものであり、ユーザ自身が検索語を明示することなく、時と共に移りゆくユーザの関心に合わせて注目すべきコンテンツを特定又は抽出するための技術を提供することである。

【課題を解決するための手段】

【0009】

本発明に係るコンテンツ特定方法は、アクセス時刻を含む、登録ユーザのアクセスログを格納するアクセスログ格納部と収集したコンテンツ中のキーワードに関するデータを格納するコンテンツプロファイル・データベースとに格納されているデータから、アクセス時刻及び上記キーワードに関する情報と登録ユーザとの関係を表すトランザクション・データを生成し、トランザクション・データベースに登録するステップと、トランザクション・データベースに格納された未処理のトランザクション・データに係る特定の登録ユーザに関連し且つ当該未処理のトランザクション・データに含まれるキーワードの、アクセス時刻における評価値と、登録ユーザとキーワードとのこれまでの関連度を表すデータを格納するユーザプロファイル・データベースに格納されているデータから特定の登録ユーザに関連するキーワードにつきアクセス時刻における減衰された関連度とを算出して、特定の登録ユーザに関連するキーワードについて評価値及び減衰された関連度からアクセス時刻における関連度を算出し、ユーザプロファイル・データベースを更新する更新ステップと、アクセスログ格納部に格納されているデータを用いて、所定の基準を超えてアクセスが増加したコンテンツを特定し、当該特定されたコンテンツについてのデータをコンテンツプロファイル・データベースから抽出し、トピック・データベースに登録する登録ステップと、ユーザプロファイル・データベースに格納されている、特定の登録ユーザについての関連度が上位（例えば上位所定数又は所定の閾値以上）のキーワードと所定の類似性を有し且つトピック・データベースに登録されているキーワードが出現するコンテンツを特定し、当該特定されたコンテンツの識別情報を特定の登録ユーザに対応して推薦トピック・データベースに登録するコンテンツ特定ステップとを含む。

【0010】

このように登録ユーザのアクセス履歴から当該登録ユーザの関心を時系列的な要素を加味しつつ具体的な関連度付きのキーワードとして特定し、さらに所定の基準を超えてアクセスが増加したコンテンツを注目コンテンツとして特定して、関連度の高いキーワードとの類似性が高いキーワードを含む注目コンテンツを、登録ユーザに対して推薦すべきコンテンツとして特定するものである。これによって、登録ユーザは、具体的な検索語を想起できない場合においても、自らの興味に合致しつつ話題性のあるコンテンツを効率的に知

10

20

30

40

50

得することができるようになる。

【0011】

また、上で述べたコンテンツ特定ステップが、関連語辞書から、ユーザプロフィール・データベースに格納されている、特定の登録ユーザについての関連度が上位のキーワードに対応して登録されている関連キーワードを抽出するステップと、特定の登録ユーザについての特定のキーワードと当該特定のキーワードに対応し且つ抽出された関連キーワードとを含む第1のセットと、トピック・データベースに登録されているキーワードをコンテンツ毎にまとめた第2のセットとの類似度をコンテンツ毎に算出するステップとを含むようにしてもよい。このようにすれば、完全同一だけではなく類似性のあるキーワードをも考慮した形で、適切なコンテンツを特定することができるようになる。

10

【0012】

さらに、トランザクション・データベースに格納された上記キーワードに関するデータが、当該キーワードの提示回数 k を含むようにしてもよい。その場合、上で述べた更新ステップが、未処理のトランザクション・データに含まれるキーワードの提示回数 k （例えば実施の形態における出現回数又はアクセス回数）と所定の減衰係数 ρ によって、上記キーワードの評価値を $(1 - \rho^k) / (1 - \rho)$ として算出するステップを含むようにしてもよい。毎日定期的にアクセスするコンテンツと急に多数回アクセスするようになったコンテンツとは、ユーザにとってその重要度はほぼ同じであり、このような状況を同様に評価することができるようになる。

【0013】

20

さらに、ユーザプロフィール・データベースには、キーワード毎に処理基準日時のデータが登録されるようにしてもよい。そうすれば、上で述べた更新ステップは、処理基準日時からアクセス時刻までの単位時間数 t と所定減衰係数 ρ と上で述べたこれまでの関連度 g とによって、アクセス時刻における減衰された関連度を $\rho^t g$ として算出するステップを含むようにしてもよい。このようにすれば、適切に過去の影響を減衰させることができる。

【0014】

また、上で述べた登録ステップは、各コンテンツにつき、特定時刻のアクセスユーザ数の、1単位時間前までのアクセスユーザ数の平均からの上方乖離度（例えば実施の形態における $A_t(p)$ ）を算出するステップと、上方乖離度が上位所定数内のコンテンツを特定するステップとを含むようにしてもよい。注目が集まっていることをこの上方乖離度によって特定できる。

30

【0015】

本発明に係る方法は、コンピュータ・ハードウェアとプログラムとの組み合わせにより実施される場合があり、このプログラムは、例えばフレキシブルディスク、CD-ROM、光磁気ディスク、半導体メモリ、ハードディスク等の記憶媒体又は記憶装置に格納される。また、ネットワークなどを介してデジタル信号として配信される場合もある。尚、中間的な処理結果はメインメモリ等の記憶装置に一時保管される。

【発明の効果】

【0016】

40

本発明によれば、ユーザ自身が検索語を明示することなく、時と共に移りゆくユーザの関心に合わせて注目すべきコンテンツを特定又は抽出することができるようになる。

【発明を実施するための最良の形態】

【0017】

本発明の一実施の形態に係るシステム概要図を図1に示す。例えばインターネットであるネットワーク101には、複数のユーザ端末103と、複数のウェブ(Web)サーバ107と、本実施の形態において主要な処理を実施するコンテンツ推薦サーバ105とが接続されている。ユーザ端末103には、例えばウェブブラウザのプラグインとして専用のアプリケーションがインストールされ、当該アプリケーションによってアクセス先のデータを含むアクセスログのデータがコンテンツ推薦サーバ105に送信されるようになっ

50

ている。そのような構成でない場合には、コンテンツ推薦サーバ１０５がインターネット・サービス・プロバイダ（ＩＳＰ：Internet Service Provider）内に設置されており、ユーザ端末１０３は、コンテンツ推薦サーバ１０５を経由してＷｅｂサーバ１０７にアクセスするため、コンテンツ推薦サーバ１０５がアクセス先のデータを含むアクセスログを取得できるようになっている場合もある。どのような構成であっても、コンテンツ推薦サーバ１０５は、登録ユーザのアクセスログを取得できるようになっている。そして、コンテンツ推薦サーバ１０５は、当該アクセスログなどから、登録ユーザに対して、当該登録ユーザが関心を有し且つ最近注目されているＷｅｂページ（すなわちコンテンツ）の推薦を行うための処理を実施する。

【００１８】

このようなコンテンツ推薦サーバ１０５の機能ブロック図を図２乃至図４を用いて説明する。図２は、コンテンツ推薦サーバ１０５の前処理及びユーザプロファイル生成処理を実施する部分の機能ブロック図を示す。コンテンツ推薦サーバ１０５は、ユーザ端末１０３のアクセス先のデータを含むアクセスログを取得するための処理を実施するアクセスログ取得部１と、アクセスログ取得部１が取得したアクセスログを格納するアクセスログ格納部３と、ネットワーク１０１に接続されているＷｅｂサーバ１０７からＷｅｂページ・データを収集するＷｅｂページ収集部７と、Ｗｅｂページ収集部７が収集したＷｅｂページ・データを格納するＷｅｂページＤＢ９と、ＷｅｂページＤＢ９から周知の手法を用いてキーワードを抽出するキーワード抽出部１１と、キーワード抽出部１１によって抽出されたキーワードのデータをＵＲＬ（Uniform Resource Locator）と共に格納するコンテンツプロファイルＤＢ１３と、アクセスログ格納部３に格納されているデータとコンテンツプロファイルＤＢ１３に格納されているデータとを連結したデータを生成するログ連結部５と、ログ連結部５によって生成されたデータを格納するトランザクションＤＢ１５と、ユーザが興味のあるキーワード等のデータを格納するユーザプロファイルＤＢ１９と、トランザクションＤＢ１５に新たに格納されたデータとユーザプロファイルＤＢ１９に格納されている過去のユーザプロファイルとを用いてユーザプロファイルを更新するための処理を実施するユーザプロファイル生成部１７とを有する。

【００１９】

図３は、コンテンツ推薦サーバ１０５のコンテンツ選別処理及びマッチング処理を実施する部分の機能ブロック図を示す。コンテンツ推薦サーバ１０５は、アクセスログ格納部３に格納されているデータを用いて処理を行い、該当するデータをコンテンツプロファイルＤＢ１３から抽出する処理を行うコンテンツ選別部２１と、コンテンツ選別部２１によって抽出されたデータを格納するトピックＤＢ２３と、特定の語に関連する語が登録されている関連語辞書２７と、関連語辞書２７に格納されているデータを用いてユーザプロファイルＤＢ１９に含まれるキーワードを処理して該当するデータをトピックＤＢ２３から抽出するマッチング部２５と、マッチング部２５の処理結果である各ユーザへの推薦ＵＲＬなどのデータを格納する推薦トピックＤＢ２９と、ユーザ端末１０３に推薦ＵＲＬなどを出力する推薦出力部３１とを有する。

【００２０】

図４は、関連語辞書２７の生成処理を実施する部分の機能ブロック図を示す。コンテンツ推薦サーバ１０５は、登録ユーザによる操作ログを格納する操作履歴ＤＢ３５と、操作推薦出力部３１が出力した推薦ＵＲＬを受信して表示したユーザ端末１０３から登録ユーザによる推薦ＵＲＬのクリックに関するデータを取得し、推薦トピックＤＢ２９から対応するキーワードを抽出して操作履歴ＤＢ３５に格納する操作ログ取得部３３と、操作履歴ＤＢ３５から関連語辞書のデータを生成する関連語辞書生成部３７とを含む。

【００２１】

次に、図５乃至図２６を用いて図１乃至図４に示したシステムの処理を説明する。まず、コンテンツ推薦サーバ１０５は、前処理を実施する（図５：ステップＳ１）。この前処理については図６乃至図１２を用いて説明する。まず、Ｗｅｂページ収集部７は、ネットワーク１０１を介してＷｅｂサーバ１０７に対してＷｅｂページの収集処理を実施し、収

10

20

30

40

50

集した Web ページのデータを URL に対応して Web ページ DB 9 に格納する (図 6 : ステップ S 1 1) 。例えば、Web ページ DB 9 には図 7 に示すようなデータフォーマットでデータを格納する。すなわち、Web ページ・データの取得日時、Web ページ・データの URL、Web ページのタイトル、Web ページの内容を格納する。

【 0 0 2 2 】

また、キーワード抽出部 1 1 は、Web ページ DB 9 に格納された各 Web ページについて周知のキーワード抽出処理を実施し、抽出されたキーワード等を URL 等に対応してコンテンツプロファイル DB 1 3 に格納する (ステップ S 1 3) 。例えば、コンテンツプロファイル DB 1 3 には図 8 に示すようなデータフォーマットでデータを格納する。すなわち、元となる Web ページ・データの取得時刻、URL、抽出されたキーワード、本 URL の Web ページにおいて本キーワードが出現する回数、抽出処理の際に算出されたスコアなどが格納されるようになっている。キーワード毎にレコードが生成される。なお、スコアについては格納しなくとも良い。

【 0 0 2 3 】

一方、アクセスログ取得部 1 は、ユーザ端末 1 0 3 から Web ページへのアクセスに関するデータを受信し、アクセス先 URL 及びユーザ ID を含むアクセスログを生成してアクセスログ格納部 3 に格納する (ステップ S 1 5) 。例えば、アクセスログ格納部 3 には、図 9 に示すようなデータフォーマットでデータを格納する。すなわち、アクセス日時、ユーザ ID、アクセス先の URL である参照 URL とが格納されるようになっている。

【 0 0 2 4 】

さらに、ログ連結部 5 は、コンテンツプロファイルとアクセスログとを URL で連結する処理を実施し、処理結果をトランザクション DB 1 5 に格納する (ステップ S 1 7) 。具体的には、アクセス時刻、ユーザ ID 及び URL については、アクセスログ格納部 3 から抽出され、当該 URL に対応してコンテンツプロファイル DB 1 3 に格納されているキーワード及び回数が抽出され、トランザクション DB 1 5 に格納される。例えば、トランザクション DB 1 5 には図 1 0 に示すようなデータフォーマットでデータを格納する。すなわち、アクセス時刻、ユーザ ID、キーワード及び回数が格納されるようになっている。なお、アクセス時刻が所定の単位時間 (例えば 1 日) 毎であれば、アクセス時刻、ユーザ ID 及びキーワードで、レコードをマージして回数は合計される。このような場合、この「回数」については、アクセス回数とも呼ぶものとする。

【 0 0 2 5 】

図 5 の説明に戻って、次にユーザプロファイル生成部 1 7 は、ユーザプロファイル DB 1 9 とトランザクション DB 1 5 とを用いて、ユーザプロファイル生成処理を実施する (ステップ S 3) 。ユーザプロファイル生成処理については図 1 1 乃至図 1 7 を用いて説明する。本実施の形態では、ユーザとキーワードとの関係をユーザプロファイル DB 1 9 に格納するが、その際キーワードにつきユーザとの関連性を表すスコアを、図 1 1 (a) 及び (b) に示すように時間に応じて減衰させる。すなわち、図 1 1 (a) に示すように、 t_0 で特定のキーワードのスコアが X であった場合、図 1 1 (b) に示すように、1 単位時間後の t_1 になると ρ ($0 < \rho < 1$) 倍になり、さらに 1 単位時間後の t_2 になるとさらに ρ 倍になり、そしてさらに 1 単位時間後の t_3 になるものとする。すなわち、 t_3 のスコアは t_0 のスコアの $\rho^{t_0-t_3}$ 倍になる。一般的に、時刻 t_i におけるアクセス回数 (キーワードの出現回数) が n_i の場合の現在のスコア $g(t_0)$ は、以下のように表される。

【 数 1 】

$$\begin{aligned} g(t_0) &= f(n_0)\rho^0 + f(n_1)\rho^1 + f(n_2)\rho^2 + \cdots \\ &= \sum_{i=0} f(n_i)\rho^i \end{aligned}$$

なお、 $f(n)$ は、 n 回アクセスしたときのスコアである。

【 0 0 2 6 】

また、本実施の形態では、アクセス回数（キーワードの出現回数）とスコアの関係については、以下の事項を前提とする。すなわち、図 12 に示すように、最近 n 日間、毎日 1 回アクセスした場合のスコアの合計値（ $= 1 + \rho + \dots + \rho^{n-1}$ ）と、今日 1 日に n 回アクセスした場合のスコア（ $= f(n)$ ）とが同じであるとする。そうすると、 $f(n)$ は以下のように表される。

【数 2】

$$f(n) = \frac{1 - \rho^n}{1 - \rho}$$

ここで $0 < \rho < 1$ である。

【0027】

なお、ある時点 τ でのスコア $g(\tau)$ が分かっている場合には、現時刻 t において n 回のアクセスがあった場合のスコア $g(t)$ は、 $g(\tau)$ を用いて以下の式で算出される。

$$g(t) = f(n) + \rho^{t-\tau} g(\tau)$$

【0028】

このような前提の下、図 13 に示すような処理をユーザプロフィール生成部 17 が実施する。まず、ユーザプロフィール生成部 17 は、トランザクション DB 15 から未処理所定単位（例えば毎日に処理を行う場合には本日分）のトランザクション・データを抽出してユーザ ID でグループ化し、各グループのデータを例えばメインメモリなどの記憶装置に格納する（ステップ S21）。例えば図 14 に示すようなデータがトランザクション DB 15 に格納されている場合には、グループ A、グループ B、グループ C にまとめられる。次に、未処理のユーザ ID を 1 つ特定する（ステップ S23）。そして、特定されたユーザ ID の過去のユーザプロフィールを、ユーザプロフィール DB 19 から読み出す（ステップ S25）。例えば図 15 に示すようなデータがユーザプロフィール DB 19 から読み出されるものとする。図 15 に示すように、ユーザプロフィール DB 19 には、最終訪問日時（日単位で処理をする場合には最終訪問日。但し、アクセス日時が存在しないがユーザ ID が処理対象として抽出されると、処理日時又は処理日となる。）、ユーザ ID、キーワード及び当該キーワードのスコアが登録されるようになっている。図 15 の例では、ユーザ ID が「1000」のユーザと、ユーザ ID が「3388」であるユーザと、ユーザ ID が「2621」であるユーザとが登録されている。

【0029】

さらに、現時刻（アクセス日時又はアクセス日） t と過去ユーザプロフィールの最終訪問日時を取得し、スコアの減衰処理を実施する（ステップ S27）。具体的には、現時刻 t と最終訪問日時の差（例えば日単位）を t とすると、 ρ^t 倍して、例えばメインメモリ等の記憶装置に格納する。例えば、現時刻 t は図 14 から 2006 年 2 月 14 日であり、最終訪問日時が 2006 年 2 月 10 日であるので 4 日経っており、 ρ^4 を乗ずる。 $\rho = 0.9$ であるとする、図 16 に示したようなスコアが算出される。なお、所定の閾値（例えば 0.1）より小さいスコアのレコードについては削除するようにする。これによって処理量を削減できる。

【0030】

そして、特定されたユーザ ID の読み出されたトランザクション・データに含まれるアクセス回数（キーワードの出現回数）に応じたスコアを算出し、例えばメインメモリ等の記憶装置に格納する（ステップ S29）。回数を k とすると $(1 - \rho^k) / (1 - \rho)$ を算出する。例えば、ユーザ ID 「1000」のキーワード「トリノ」については、 $(1 - 0.9^3) / (1 - 0.9) = 2.71$ となる。同様に、ユーザ ID 「1000」のキーワード「下村・子」については、 $(1 - 0.9^1) / (1 - 0.9) = 1$ となる。さらに、ユーザ ID 「1000」のキーワード「モーグル」については、 $(1 - 0.9^2) / (1 - 0.9) = 1.9$ となる。

【0031】

最後に、ステップ S27 の減衰処理の結果とステップ S29 で算出されたスコアを加算

10

20

30

40

50

し、ユーザプロフィールDB 19を更新する(ステップS 31)。ユーザID「1000」のキーワード「トリノ」については、 $0.51 + 2.71 = 3.22$ で更新される。さらに、ユーザID「1000」のキーワード「下村・子」については、 $0 + 1.0 = 1.0$ で更新される。「下村・子」についてはレコードが存在していなかったので追加される。また、ユーザID「1000」のキーワード「モーグル」についても、 $0 + 1.9 = 1.9$ で更新される。「モーグル」についてもレコードが存在していなかったので追加される。このような処理を実施すれば、図17に示すようなデータがユーザプロフィールDB 19に登録される。

【0032】

その後、全てのユーザIDについて処理が完了したか判断し(ステップS 33)、未処理のユーザIDが存在していればステップS 23に戻り、全てのユーザIDについて処理が完了していれば元の処理に戻る。

【0033】

このようにして、時間軸方向で適切に減衰され且つ定量化されたスコアが、ユーザの興味に関連するキーワード毎にユーザプロフィールDB 19に登録されるようになる。なお、この段階で、スコアでソートして、キーワードを絞り込んでも良い。例えば、上位所定数のキーワードを特定したり、閾値以上のスコアを有するキーワードを特定するようにしても良い。

【0034】

図5の説明に戻って、次に、コンテンツ選別部21は、コンテンツプロフィールDB 13とアクセスログ格納部3とを用いて、コンテンツ選別処理を実施する(ステップS 5)。このコンテンツ選別処理については図18乃至図21を用いて説明する。

【0035】

コンテンツ選別部21は、アクセスログ格納部3に格納されたアクセスログのデータから所定単位時間(例えば日単位)のアクセスユーザ数をURL毎にカウントし、カウント結果を例えばメインメモリ等の記憶装置に格納する(図18:ステップS 41)。例えば、図19に示すようなデータがアクセスログ格納部3に格納されているとすると、例えば図20に示すようなデータが生成される。すなわち、2006年2月14日にURL 1にアクセスしたユーザの数は「3」であり、2006年2月13日にURL 1にアクセスしたユーザの数は「1」であり、2006年2月12日にURL 1にアクセスしたユーザの数は「2」である。なお、時刻 τ にWebページ p にアクセスしたユーザ数を、 $U_{\tau}(p)$ と表すものとする。

【0036】

そして、未処理のURLを1つ特定し(ステップS 43)、 $U_{\tau}(p)$ の平均を以下のとおり算出し、例えばメインメモリ等の記憶装置に格納する(ステップS 45)。

【数3】

$$\overline{U}_{t-1}(p) = \frac{\sum_{\tau=1}^{t-1} U_{\tau}(p)}{t-1}$$

このように、現時点 t を含まない直前の $t-1$ の段階までのWebページ p の平均ユーザ数が算出される。

【0037】

この $U_{\tau}(p)$ の平均を用いて以下の式に従ってスコア $A_t(p)$ を算出し、例えばメインメモリ等の記憶装置に格納する(ステップS 47)。

【数4】

$$A_t(p) = \frac{U_t(p) - C_t \overline{U}_{t-1}(p)}{C_t \overline{U}_{t-1}(p)} U_t(p)$$

10

20

30

40

50

この式は、仮に $C_t = 1$ だとすると、 $U_t(p)$ の平均ユーザ数からのずれに対して、時刻 t におけるユーザ数を乗じた値となる。すなわち、平均ユーザ数からのずれ（上方乖離度）が大きいほど $A_t(p)$ が大きな値となって出てくる。より具体的には、より多くのユーザから注目をあびるようになると、 $A_t(p)$ が大きな値になるので、バーストを検出することができる。

【0038】

但し、 C_t は時間帯 t による補正係数であり、例えば1時間毎に設定する場合もある。この場合、0時台には $C_t = 0.9$ 、1時台 = 0.8 、・・・23時台 = 1.0 のようにする。これは、夜間のアクセスが多く、早朝のアクセスが少ないなど、アクセスが集中する時間帯にアクセスされたページのスコアが不当に高く評価される問題を解消するためである。1日を単位時間とする場合には、日毎に設定するようにする。曜日毎に設定するようにしても良い。また、 C_t については固定しても良い。

10

【0039】

そして、未処理のURLが存在するか判断し、未処理のURLが存在する場合にはステップS43に戻る。一方、未処理のURLが存在しない場合には、 $A_t(p)$ の値でURLをソートし、上位所定数のURLのデータをコンテンツプロファイルDB13から抽出して、トピックDB23に登録する（ステップS51）。そして元の処理に戻る。トピックDB23に格納されるデータのフォーマット例を図21に示す。図21の例では、本URLを検出した時刻であるバースト時刻と、URLと、当該URLに関連するキーワードと、スコアとが登録されるようになっている。

20

【0040】

図5の説明に戻って、次に、マッチング部25は、ユーザプロファイルDB19、関連語辞書27及びトピックDB23を用いてマッチング処理を実施し、ユーザに推薦すべきURLのリストをユーザ毎に推薦トピックDB29に格納する（ステップS7）。マッチング処理については図22乃至図24を用いて説明する。まず、マッチング部25は、各ユーザのユーザプロファイルに含まれるキーワード（例えばスコア上位3位までのキーワード）をユーザプロファイルDB19から抽出し、当該キーワードを関連語辞書27によってグループ化し、当該グループのデータを例えばメインメモリ等の記憶装置に格納する（ステップS61）。グループ化については、図23及び図24を用いて説明する。例えば、関連語辞書27には図23に示すようなフォーマットでデータが格納される。すなわち、キーワード1と、キーワード1に関連するキーワード2と、それらの関連度とが格納されるようになっている。

30

【0041】

図17の例では、ユーザID「1000」のユーザプロファイル中には、「トリノ」、「下村・子」、「モーグル」、「フィギュア」、「代表選考」が登録されているが、スコアの値で上位3つに限定すると、「トリノ」「モーグル」「フィギュア」が特定される。一方、関連語辞書27には、「トリノ」と「スケルトン」の組、「トリノ」と「ハーフパイプ」の組、「トリノ」と「フィギュア」の組、「トリノ」と「モーグル」の組、「トリノ」と「大谷多●」の組、「トリノ」と「下村・子」の組、「トリノ」と「村上・枝」の組、「トリノ」と「安川静●」の組、「モーグル」と「大谷多●」の組、「モーグル」と「下村・子」の組、「代表選考」と「深田真●」の組、「代表選考」と「伊藤美●」の組、「代表選考」と「安川静●」の組と、「代表選考」と「村上・枝」の組とが登録されているとする。

40

【0042】

そうすると、図24に示すようなグラフが描ける。但し、キーワードに対応する四角は、大きいものほどユーザプロファイル中でスコアが大きい、又は関連語辞書27において関連度が大きいことを表している。これによって「トリノ」に関連するキーワードのグループであるグループ1 = {トリノ, フィギュア, モーグル, スケルトン, ハーフパイプ} が構成される。「安川静●」「村上・枝」「下村・子」「大谷多●」については相対的に関連度が低いのでグループに登録されていない。また、「モーグル」に関連するキーワー

50

ドのグループであるグループ2 = { モーグル, トリノ, 下村・子, 大谷多・ } が構成される。さらに、「フィギュア」に関連するキーワードのグループであるグループ3 = { フィギュア, トリノ } が構成される。

【0043】

次に、未処理のユーザを1人特定し(ステップS63)、未処理のキーワードグループを1つ特定する(ステップS65)。そして、特定されたキーワードグループと、トピックDB23に格納されているキーワードとの類似度をトピックDB23のURL毎に算出し、例えばメインメモリ等の記憶装置に格納する(ステップS67)。類似度は、例えば以下の式で算出される。

$$J(W, V) = (W \cap V) / (W \cup V)$$

10

なお、 $J(W, V)$ は周知の J a c c a r d C o e f f i c i e n t である。Wは、ステップS61で生成され且つステップS65で特定されたキーワードグループであり、Vは、トピックDB23内の特定のURLのキーワードグループである。従って、分母はW ∪ Vのキーワード数、分子はW ∩ Vのキーワード数である。

【0044】

そして、全てのキーワードグループについて処理したか判断する(ステップS69)。未処理のキーワードグループが存在していれば、ステップS65に戻る。一方、未処理のキーワードグループが存在しない場合には、類似度Jでソートし、類似度Jが大きい順に所定数のURLを特定して、トピックDB23内の当該URLの対応データを推薦トピックDB29に格納する(ステップS71)。

20

【0045】

さらに、全てのユーザについて処理したか判断し(ステップS73)、未処理のユーザが存在している場合にはステップS63に戻る。一方、全てのユーザを処理した場合には、元の処理に戻る。

【0046】

図5の処理に戻って、最後に推薦出力部31は、例えばユーザ端末103からの要求に応じて当該ユーザ端末103の登録ユーザについての推薦URLを推薦トピックDB29から読み出し、当該推薦URLのリストをユーザ端末103に出力する(ステップS9)。ユーザ端末103は、コンテンツ推薦サーバ105から、登録ユーザが興味を有しており且つ最近注目されている推薦URLを受信し、表示装置に表示する。例えば、Webブラウザのプラグインとして提供されているアプリケーションによってリンクの形で登録ユーザに提示される。

30

【0047】

このようにすれば、ユーザが明確に把握していないようなキーワードであっても上で述べたような処理によって抽出され、さらに当該キーワードに関連し且つ最近注目されているサイトのURLが、自動的に提示されるようになるため、効率的にWebページを閲覧することができるようになる。

【0048】

なお、関連語辞書27については、例えば図25及び図26に示すような処理にて構成される場合がある。例えば、推薦出力部31は、上で述べたようにユーザ端末103に推薦URLのリストを送信し、ユーザ端末103は、コンテンツ推薦サーバ105から推薦URLのリストを受信し、表示装置に表示することによって、登録ユーザに推薦URLのリストを提示する(ステップS81)。これに対して、登録ユーザが、推薦URLのうちいずれかを選択してクリックすると、ユーザ端末103は、当該推薦URLの選択を受け付け、当該推薦URLの選択データをコンテンツ推薦サーバ105に送信する。コンテンツ推薦サーバ105の操作取得部33は、ユーザ端末103から推薦URLの選択データを受信すると、推薦トピックDB29から当該選択に係る推薦URLに対応して登録されたキーワードを読み出し、操作履歴DB35に登録する(ステップS83)。例えば操作履歴DB35には、図26に示すようなデータフォーマットでデータが蓄積される。すなわち、アクセス時刻と、ユーザIDと、キーワードと、参照URLとが格納されるように

40

50

なっている。

【0049】

次に、関連語辞書生成部37は、周知の関連度算出処理を実施する（ステップS85）。これによって、例えば同じURLを参照URLとするキーワードにつき関連度が算出される。そして、関連語辞書生成部37は、算出された関連度に従って、例えば所定の閾値以上の関連度を有するキーワードの組及びその関連度を含む関連語辞書データを生成し、関連語辞書27に登録する（ステップS87）。

【0050】

このような処理を実施することによって、登録ユーザによる実際の操作履歴に基づき、適切な関連語が関連語辞書に蓄積されるようになる。従って、推薦URLを選択する際にも適切なキーワードグループが構成されるようになり、適切な類似度が算出され、最終的に適切な推薦URLが特定されるようになる。

【0051】

以上本発明の一実施の形態を説明したが、本発明はこれに限定されるものではない。例えば、図2乃至図4に示した機能ブロックは必ずしも実際のプログラム構成に対応しない場合もある。また、処理フローについても、処理結果が変わらない限りにおいて順番の入れ替えや並列処理が可能である。

【0052】

なお、ユーザ端末103、コンテンツ推薦サーバ105、Webサーバ107は、図27のようなコンピュータ装置であって、メモリ2501（記憶装置）とCPU2503（処理装置）とハードディスク・ドライブ（HDD）2505と表示装置2509に接続される表示制御部2507とリムーバブル・ディスク2511用のドライブ装置2513と入力装置2515とネットワークに接続するための通信制御部2517とがバス2519で接続されている。オペレーティング・システム（OS：Operating System）及び本実施の形態における処理を実施するためのアプリケーション・プログラムは、HDD2505に格納されており、CPU2503により実行される際にはHDD2505からメモリ2501に読み出される。必要に応じてCPU2503は、表示制御部2507、通信制御部2517、ドライブ装置2513を制御して、必要な動作を行わせる。また、処理途中のデータについては、メモリ2501に格納され、必要があればHDD2505に格納される。本発明の実施の形態では、上で述べた処理を実施するためのアプリケーション・プログラムはリムーバブル・ディスク2511に格納されて頒布され、ドライブ装置2513からHDD2505にインストールされる。インターネットなどのネットワーク及び通信制御部2517を経由して、HDD2505にインストールされる場合もある。このようなコンピュータ装置は、上で述べたCPU2503、メモリ2501などのハードウェアとOS及び必要なアプリケーション・プログラムとが有機的に協働することにより、上で述べたような各種機能を実現する。

【図面の簡単な説明】

【0053】

【図1】本発明の実施の形態に係るシステム概要を説明するための図である。

【図2】コンテンツ推薦サーバの第1の機能ブロック図である。

【図3】コンテンツ推薦サーバの第2の機能ブロック図である。

【図4】コンテンツ推薦サーバの第3の機能ブロック図である。

【図5】本発明の実施の形態に係るメイン処理フローを示す図である。

【図6】前処理の処理フローを示す図である。

【図7】WebページDBのデータフォーマット例を示す図である。

【図8】コンテンツプロファイルDBのデータフォーマット例を示す図である。

【図9】アクセス履歴DBのデータフォーマット例を示す図である。

【図10】トランザクションDBのデータフォーマット例を示す図である。

【図11】（a）及び（b）は、ユーザプロファイルにおけるスコアの時間減衰を説明するための図である。

10

20

30

40

50

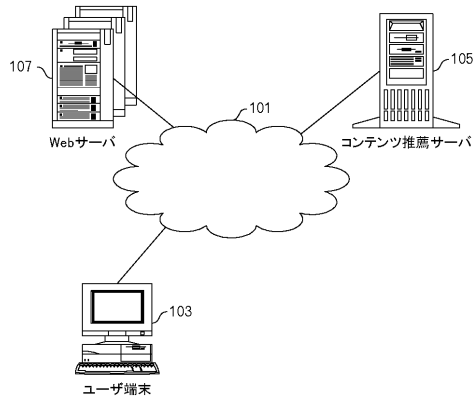
- 【図 1 2】アクセス頻度と減衰の調整モデルを説明するための図である。
 【図 1 3】ユーザプロファイル生成処理の処理フローを示す図である。
 【図 1 4】ユーザプロファイル生成処理を説明するためのデータ例を示す図である。
 【図 1 5】ユーザプロファイル生成処理を説明するためのデータ例を示す図である。
 【図 1 6】ユーザプロファイル生成処理を説明するためのデータ例を示す図である。
 【図 1 7】ユーザプロファイル生成処理を説明するためのデータ例を示す図である。
 【図 1 8】コンテンツ選別処理の処理フローを示す図である。
 【図 1 9】コンテンツ選別処理を説明するためのデータ例を示す図である。
 【図 2 0】コンテンツ選別処理を説明するためのデータ例を示す図である。
 【図 2 1】トピック D B のデータフォーマット例を示す図である。
 【図 2 2】マッチング処理の処理フローを示す図である。
 【図 2 3】関連語辞書のデータフォーマット例を示す図である。
 【図 2 4】キーワードのグループ化を説明するための図である。
 【図 2 5】関連語辞書作成処理の処理フローを示す図である。
 【図 2 6】操作履歴 D B のデータフォーマット例を示す図である。
 【図 2 7】コンピュータの機能ブロック図である。

【符号の説明】

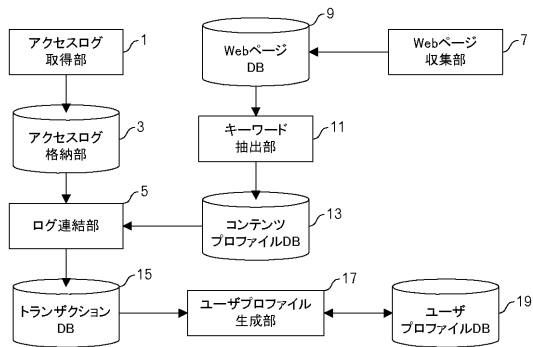
【 0 0 5 4 】

- | | | | | |
|-------|-----------------|-------|--------------|----|
| 1 | アクセスログ取得部 | 3 | アクセスログ格納部 | |
| 5 | ログ連結部 | 7 | Web ページ収集部 | 20 |
| 9 | Web ページ D B | 1 1 | キーワード抽出部 | |
| 1 3 | コンテンツプロファイル D B | | | |
| 1 5 | トランザクション D B | 1 7 | ユーザプロファイル生成部 | |
| 1 9 | ユーザプロファイル D B | | | |
| 2 1 | コンテンツ選別部 | 2 3 | トピック D B | |
| 2 5 | マッチング部 | 2 7 | 関連語辞書 | |
| 2 9 | 推薦トピック D B | 3 1 | 推薦出力部 | |
| 3 3 | 操作取得部 | 3 5 | 操作履歴 D B | |
| 3 7 | 関連語辞書生成部 | | | |
| 1 0 1 | ネットワーク | 1 0 3 | ユーザ端末 | 30 |
| 1 0 5 | コンテンツ推薦サーバ | 1 0 7 | Web サーバ | |

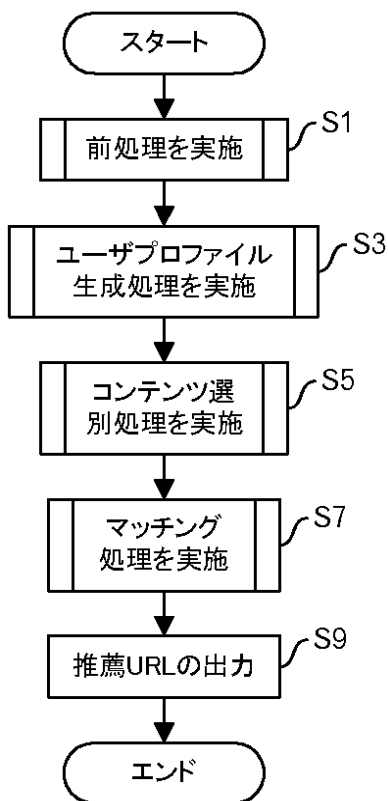
【図 1】



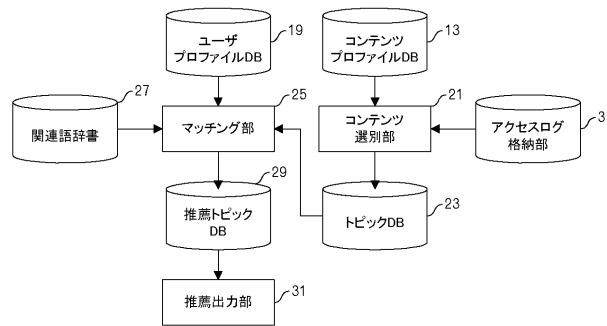
【図 2】



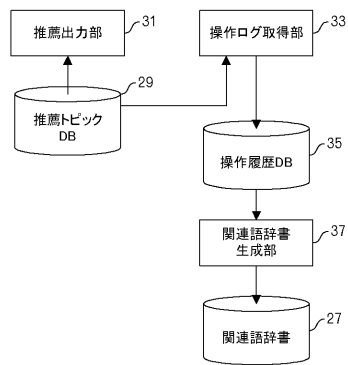
【図 5】



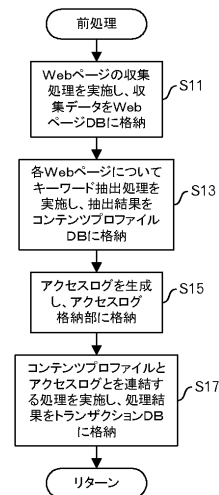
【図 3】



【図 4】



【図 6】



【図 7】

取得時刻	URL	タイトル	内容

【図 8】

取得時刻	URL	キーワード	回数	スコア

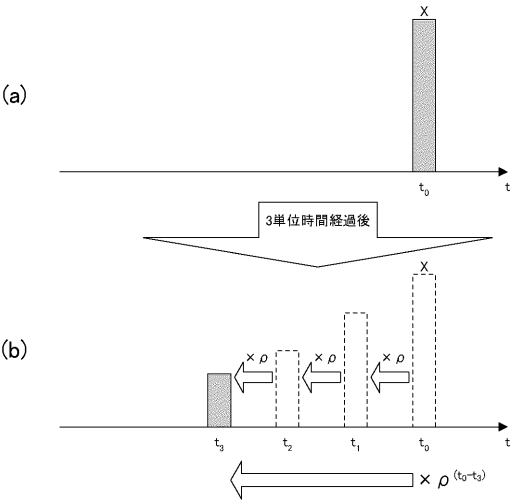
【図 9】

アクセス時刻	ユーザID	参照URL

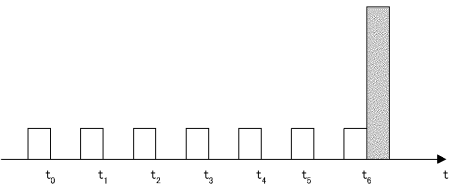
【図 10】

アクセス時刻	ユーザID	キーワード	回数

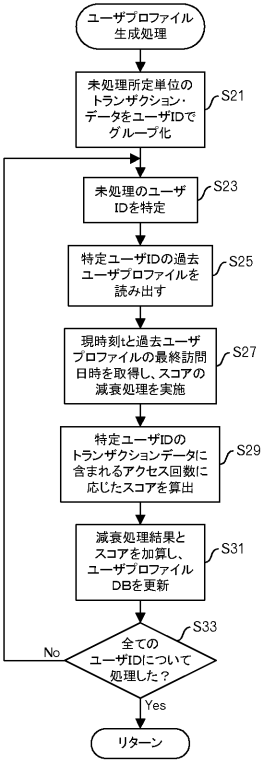
【図 11】



【図 12】



【図 13】



【図 14】

アクセス時刻	ユーザID	キーワード	回数
2006-02-14	1000	トリノ	3
2006-02-14	1000	下村●子	1
2006-02-14	1000	モーグル	2
2006-02-14	1203	パレンティン	5
2006-02-14	1203	チョコ	3
2006-02-14	2621	天気	2
2006-02-14	2621	スキー場	4
2006-02-14	2621	積雪	2
⋮	⋮	⋮	⋮

【図 15】

最終訪問日時	ユーザID	キーワード	スコア
2006-02-10	1000	フィギュア	2.88
2006-02-10	1000	代表選考	1.82
2006-02-10	1000	トリノ	0.77
2006-02-13	3388	ブログ	1.29
2006-02-13	3388	こむごう	3.78
2006-02-07	2621	積雪	0.80
2006-02-07	2621	苗場	1.82
2006-02-07	2621	渋滞情報	1.95
⋮	⋮	⋮	⋮

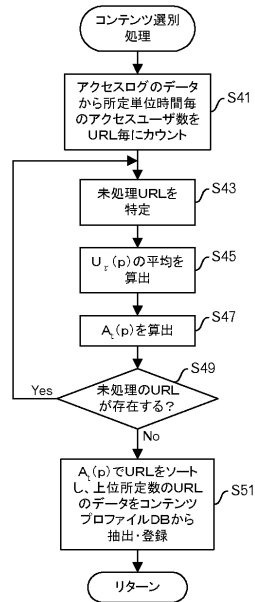
【図 16】

最終訪問日時	ユーザID	キーワード	スコア
2006-02-14	1000	フィギュア	1.89
2006-02-14	1000	代表選考	1.19
2006-02-14	1000	トリノ	0.51
2006-02-13	3388	ブログ	1.29
2006-02-13	3388	こむごう	3.78
2006-02-14	2621	積雪	0.38
2006-02-14	2621	苗場	0.09
2006-02-14	2621	渋滞情報	0.94
⋮	⋮	⋮	⋮

【図 17】

最終訪問日時	ユーザID	キーワード	スコア
2006-02-14	1203	バレンタイン	4.10
2006-02-14	1203	チョコ	2.71
2006-02-14	1000	トリノ	3.22
2006-02-14	1000	下村●子	1.00
2006-02-14	1000	モーグル	1.90
2006-02-14	1000	フィギュア	1.89
2006-02-14	1000	代表選考	1.19
2006-02-13	3388	ブログ	1.29
2006-02-13	3388	こむぎ	3.78
2006-02-14	2621	天気	1.90
2006-02-14	2621	スキー場	3.44
2006-02-14	2621	積雪	2.28
2006-02-14	2621	渋滞情報	0.94
⋮	⋮	⋮	⋮

【図 18】



【図 19】

アクセス時刻	ユーザID	参照URL
2006-02-14	1000	http://URL1
2006-02-14	1023	http://URL1
2006-02-14	http://URL3	http://URL2
2006-02-14	1024	http://URL2
2006-02-14	1024	http://URL3
2006-02-14	1024	http://URL1
⋮	⋮	⋮
2006-02-13	9999	http://URL1
2006-02-13	9989	http://URL2
2006-02-13	8879	http://URL2
⋮	⋮	⋮

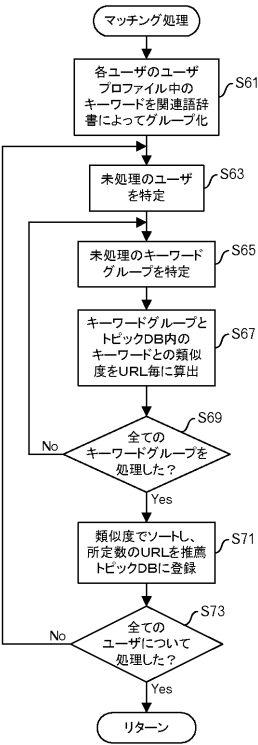
【図 21】

バースト時刻	URL	キーワード	スコア

【図 20】

アクセス時刻	参照URL	ユーザ数
2006-02-14	http://URL1	3
2006-02-14	http://URL2	2
2006-02-14	http://URL3	1
⋮	⋮	
2006-02-13	http://URL1	1
2006-02-13	http://URL2	2
⋮	⋮	
2006-02-12	http://URL1	2
2006-02-12	http://URL2	1
2006-02-12	http://URL3	1
⋮	⋮	

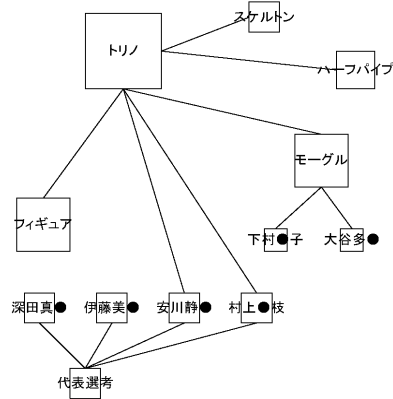
【図 2 2】



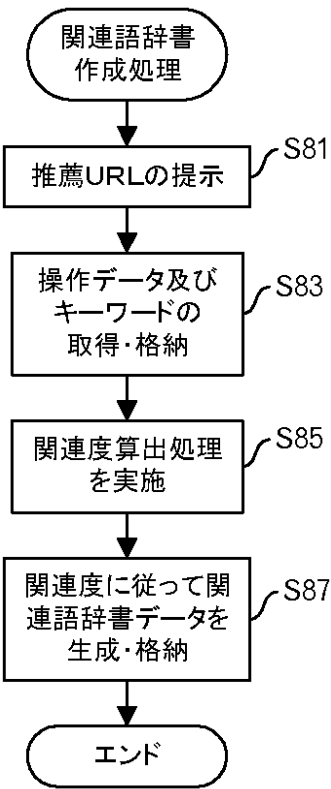
【図 2 3】

キーワード1	キーワード2	関連度

【図 2 4】



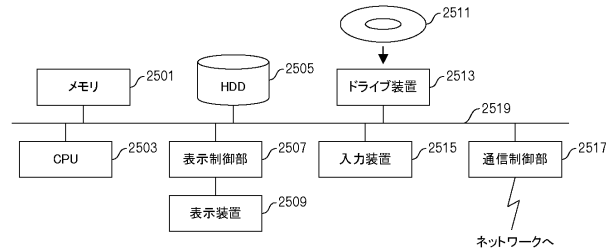
【図 2 5】



【図 2 6】

アクセス時刻	ユーザID	キーワード	参照URL

【図 2 7】



フロントページの続き

- (72)発明者 樋沼 英雄
東京都品川区南大井六丁目2番1号 ニフティ株式会社内
- (72)発明者 成田 智也
東京都品川区南大井六丁目2番1号 ニフティ株式会社内
- (72)発明者 稲越 宏弥
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
- (72)発明者 内野 寛治
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
- (72)発明者 岡本 青史
神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内

審査官 岩田 淳

- (56)参考文献 特開2003-173351(JP,A)
特開平10-326289(JP,A)
特開2002-278989(JP,A)
特開2003-173352(JP,A)
特開2005-309998(JP,A)
吉村 武, 外3名, 第68回(平成18年)全国大会講演論文集(3) データベースとメディアネットワーク, 2006年 3月 7日, P.3-217~3-218

- (58)調査した分野(Int.Cl., DB名)
G06F 17/30