(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2011 年 9 月 1 日(01.09.2011)

PCT

(10) 国際公開番号
WO 2011/105606 A1

(51) 国際特許分類:
G06F 17/30 (2006.01) *G06Q 30/00* (2006.01)

(21) 国際出願番号: PCT/JP2011/054510

(22) 国際出願日: 2011 年 2 月 28 日(28.02.2011)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願 2010-043392 2010 年 2 月 26 日(26.02.2010) JP
特願 2010-043391 2010 年 2 月 26 日(26.02.2010) JP
特願 2010-043390 2010 年 2 月 26 日(26.02.2010) JP

(71) 出願人 (米国を除く全ての指定国について): 楽天株式会社(Rakuten, Inc.) [JP/JP]; 〒1400002 東京都品川区東品川四丁目１２番３号 Tokyo (JP).

(72) 発明者; および
(75) 発明者/出願人 (米国についてのみ): 関根 聡 (SEKINE Satoshi) [JP/JP]; 〒1520031 東京都目黒区中根一丁目６番２２号 株式会社ランゲージ・クラフト研究所内 Tokyo (JP). 竹中 孝真 (TAKENAKA Takamasa) [JP/JP]; 〒1400002 東京都品川区東品川四丁目１２番３号 楽天株式会社内 Tokyo (JP).

(74) 代理人: 特許業務法人 インテクト国際特許事務所, 外 (INTECT INTERNATIONAL PATENT OFFICE et al.); 〒1020083 東京都千代田区麹町四丁目７番２号 サンライン第７ビル４階 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

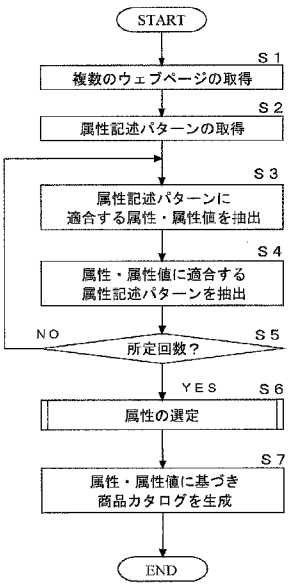
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア

[続葉有]

(54) **Title**: INFORMATION PROCESSING DEVICE, INFORMATION PROCESSING METHOD, PROGRAM FOR INFORMATION PROCESSING DEVICE, AND RECORDING MEDIUM

(54) 発明の名称 : 情報処理装置、情報処理方法、情報処理装置用のプログラム、および、記録媒体

[図4]



S1 Acquisition of plurality of web pages
S2 Acquisition of attribute description patterns
S3 Extraction of attributes/attribute values compatible with attribute description patterns
S4 Extraction of attribute description patterns compatible with attributes/attribute values
S5 Predetermined number?
S6 Selection of attributes
S7 Generation of product catalogue on basis of attributes/attribute values

(57) **Abstract**: The disclosed information processing device: obtains a plurality of web pages from the same category, said category classifying an object listed in a web page (S1); obtains attribute-related terms, which are related to the attributes of the object listed in the web pages, or attribute description patterns, which are used to describe the attributes of the object, as initial data (S2); extracts attribute-related terms for attributes which are compatible with the attribute description patterns from the plurality of web pages (S3); and extracts attribute description patterns which are compatible with the attribute-related words from the plurality of web pages (S4).

(57) 要約: ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得し（Ｓ１）、初期データとして、ウェブページに記載されている対象の属性に関連した属性関連語、または、当該対象の属性の記述に用いられる属性記述パターンを取得し（Ｓ２）、複数のウェブページから、属性記述パターンに適合する属性の属性関連語を抽出し（Ｓ３）、複数のウェブページから、属性関連語に適合する属性記述パターンを抽出する（Ｓ４）。

WO 2011/105606 A1

(AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

明　細　書

発明の名称 ：

情報処理装置、情報処理方法、情報処理装置用のプログラム、および、記録媒体

技術分野

[0001] 本発明は、インターネット上のウェブページを分析する情報処理装置、情報処理方法、情報処理装置用のプログラム、および、記録媒体の技術分野に関する。

背景技術

[0002] インターネット上のウェブサイトには商品等を扱ったウェブページが多数あり、ユーザはそのウェブページを閲覧して商品の購入を行っている。閲覧して購入する際、ユーザは通常、検索を行っている。この検索の技術において、多くのウェブページの中からユーザが欲しい商品を素早く探すために検索の技術の向上を図ることが行われている。例えば、特許文献１には、商品のカテゴリ別検索を順に大分類、中分類、小分類と検索する３層のカテゴリ別検索に限定して、６桁の整数分類コード表作成方法を考案し、この分類コード表作成方法を使用して商品分類コード表及び店舗分類コード表を作成し、これらの分類コード表をインターネットショッピングモールに設け、ショップが簡単に商品及び店舗情報の登録ができユーザが簡単に商品及び店舗を検索ができるショッピングモールにおける検索システムが開示されている。

先行技術文献

特許文献

[0003] 特許文献1：特開２００２－２３６６９４号公報

発明の概要

発明が解決しようとする課題

[0004] ところで、ワインのような商品の場合、産地、容量といった商品の属性に

関する情報や、旅行関連サービスのようなサービスの場合、料金、アクセスといったサービスの属性に関する情報が、ウェブページの中に記載されている。このような商品等の属性を抽出すれば、さまざまなサービスを提供できる可能性がある。しかし、特許文献１のような従来では、商品等の属性を抽出することは難しかった。

[0005] 本発明は、このような問題に鑑みてなされたものであり、その課題の一例は、ウェブページから商品等の属性を自動的に取得する情報処理装置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

[0006] 上記課題を解決するために、請求項１に記載の発明は、ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得するウェブページ取得手段と、初期データとして、前記ウェブページに記載されている対象の属性に関連した属性関連語、または、当該対象の属性の記述に用いられる属性記述パターンを取得する初期データ取得手段と、前記複数のウェブページから、前記属性記述パターンに適合する前記属性の属性関連語を抽出する属性抽出手段と、前記複数のウェブページから、前記属性関連語に適合する前記属性記述パターンを抽出する属性記述パターン抽出手段と、を備えたことを特徴とする。

[0007] 請求項２に記載の発明は、請求項１に記載の情報処理装置において、前記属性抽出手段および前記属性記述パターン抽出手段を交互に繰り返す繰返手段を更に備えたことを特徴とする。

[0008] 請求項３に記載の発明は、請求項１または請求項２に記載の情報処理装置において、前記属性抽出手段が、前記属性関連語として、前記属性の属性名を抽出することを特徴とする。

[0009] 請求項４に記載の発明は、請求項１から請求項３のいずれか１項に記載の情報処理装置において、抽出された前記属性関連語から属性リストを生成する属性リスト生成手段と、抽出された前記属性記述パターンのパターンリストを生成するパターンリスト生成手段と、を更に備えたことを特徴とする。

[0010] 請求項５に記載の発明は、請求項１から請求項４のいずれか１項に記載の情報処理装置において、前記属性関連語のスコア付けを行う属性スコアリング手段と、前記スコアの順に前記属性関連語のランク付けを行い、所定のランク以上の属性関連語を選択する属性選択手段と、を更に備えたことを特徴とする。

[0011] 請求項６に記載の発明は、請求項５に記載の情報処理装置において、前記属性スコアリング手段が、前記属性関連語の検索のヒット件数に基づき、前記属性関連語のスコア付けを行うことを特徴とする。

[0012] 請求項７に記載の発明は、請求項５に記載の情報処理装置において、前記属性スコアリング手段が、前記対象を販売する複数の店舗を有するウェブサイトにおいて、前記属性関連語が出現しているウェブページの前記店舗の数に基づき、前記属性関連語のスコア付けを行うことを特徴とする。

[0013] 請求項８に記載の発明は、請求項１から請求項７のいずれか１項に記載の情報処理装置において、前記カテゴリとは異なるカテゴリに属している複数のウェブページにおいて出現する前記属性関連語を取り除く属性フィルタ手段を更に備えたことを特徴とする。

[0014] 請求項９に記載の発明は、請求項１から請求項８のいずれか１項に記載の情報処理装置において、前記属性記述パターンのスコア付けを行う属性記述パターン・スコアリング手段と、前記スコアの順に前記属性記述パターンのランク付けを行い、所定のランク以上の属性記述パターンを選択する属性記述パターン選択手段と、を更に備えたことを特徴とする。

[0015] 請求項１０に記載の発明は、請求項９に記載の情報処理装置において、前記属性記述パターン・スコアリング手段が、前記属性関連語と前記属性記述パターンとが共に出現する共起数に基づき前記属性記述パターンのスコア付けを行うことを特徴とする。

[0016] 請求項１１に記載の発明は、請求項１から請求項１０のいずれか１項に記載の情報処理装置において、前記属性名同士が類似であるか否かを判定する属性名類似判定手段と、前記属性名類似判定手段により類似と判定された属

性名を集約する属性名集約手段と、を更に備えたことを特徴とする。

[0017] 請求項１２に記載の発明は、請求項１１に記載の情報処理装置において、前記属性抽出手段が、前記属性関連語として、前記属性名および前記属性名に対応する属性値を抽出し、前記属性名集約手段が、前記属性値に基づき前記属性名を集約することを特徴とする。

[0018] 請求項１３に記載の発明は、請求項１から請求項１０のいずれか１項に記載の情報処理装置において、前記ウェブページ取得手段が、前記対象の供給元のウェブページを取得し、前記初期データ取得手段、前記属性抽出手段、および、前記属性記述パターン抽出手段により、前記対象の供給元のウェブページから供給元対象属性関連語を抽出し、前記供給元対象属性関連語と前記属性関連語とを比較する属性関連語比較手段を更に備えたことを特徴とする。

[0019] 請求項１４に記載の発明は、請求項１から請求項１３のいずれか１項に記載の情報処理装置において、抽出された前記属性関連語に基づき、前記属性関連語が記載されたカタログを生成するカタログ生成手段を更に備えたことを特徴とする。

[0020] 請求項１５に記載の発明は、請求項１から請求項１４のいずれか１項に記載の情報処理装置において、前記複数のウェブページから、前記属性関連語の出現回数が所定回数以下のウェブページを抽出するウェブページ抽出手段と、

を更に備えたことを特徴とする。

[0021] 請求項１６に記載の発明は、請求項１５に記載の情報処理装置において、前記ウェブページ抽出手段が、前記属性関連語の出現回数がゼロのウェブページを抽出することを特徴とする。

[0022] 請求項１７に記載の発明は、請求項１から請求項１６のいずれか１項に記載の情報処理装置において、前記属性関連語に基づき、前記複数のウェブページをグルーピングするウェブページ・属性グルーピング手段を更に備えたことを特徴とする。

[0023] 請求項１８に記載の発明は、情報処理装置が情報処理をする情報処理方法において、ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得するウェブページ取得ステップと、前記ウェブページに記載されている対象の属性の記述に用いられる属性記述パターンを取得する属性記述パターン取得ステップと、前記複数のウェブページから、前記属性記述パターンに適合する前記属性の属性関連語を抽出する属性抽出ステップと、抽出された前記属性関連語に基づき、前記属性抽出ステップで使用する前記属性記述パターンを、前記複数のウェブページから、更に抽出する属性記述パターン抽出ステップと、を有することを特徴とする。

[0024] 請求項１９に記載の発明は、情報処理装置が情報処理をする情報処理方法において、ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得するウェブページ取得ステップと、前記ウェブページに記載されている対象の属性に関連した属性関連語を取得する属性関連語取得ステップと、前記属性の記述に用いられる属性記述パターンであって、前記複数のウェブページから、前記属性関連語に適合する前記属性記述パターンを抽出する属性記述パターン抽出ステップと、抽出された前記属性関連語に基づき、前記属性記述パターン抽出ステップで使用する属性関連語を、前記複数のウェブページから、更に抽出する属性関連語抽出ステップと、を有することを特徴とする。

[0025] 請求項２０に記載の発明は、コンピュータを、ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得するウェブページ取得手段、初期データとして、前記ウェブページに記載されている対象の属性に関連した属性関連語、または、当該対象の属性の記述に用いられる属性記述パターンを取得する初期データ取得手段、前記複数のウェブページから、前記属性記述パターンに適合する前記属性の属性関連語を抽出する属性抽出手段、および、前記複数のウェブページから、前記属性関連語に適合する前記属性記述パターンを抽出する属性記述パターン抽出手段として機能させることを特徴とする。

[0026] 請求項２１に記載の発明は、コンピュータを、ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得するウェブページ取得手段、初期データとして、前記ウェブページに記載されている対象の属性に関連した属性関連語、または、当該対象の属性の記述に用いられる属性記述パターンを取得する初期データ取得手段、前記複数のウェブページから、前記属性記述パターンに適合する前記属性の属性関連語を抽出する属性抽出手段、および、前記複数のウェブページから、前記属性関連語に適合する前記属性記述パターンを抽出する属性記述パターン抽出手段として機能させることを特徴とする情報処理装置用のプログラムを記録する。

発明の効果

[0027] 本発明によれば、ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得し、初期データとして、ウェブページに記載されている対象の属性に関連した属性関連語、または、当該対象の属性の記述に用いられる属性記述パターンを取得し、複数のウェブページから、属性記述パターンに適合する属性の属性関連語を抽出し、複数のウェブページから、属性関連語に適合する属性記述パターンを抽出することにより、同一のカテゴリに属している複数のウェブページから、属性関連語を抽出し、属性記述パターンを抽出するか、または、属性記述パターンを抽出し、属性関連語を抽出しているので、同一のカテゴリに含まれる属性を精度良く取得できる。

図面の簡単な説明

[0028] [図1]本発明の第１実施形態に係る情報処理システムの概要構成例を示す模式図である。

[図2]図１の情報処理サーバの概要構成の一例を示すブロック図である。

[図3]図１のショッピングサーバの概要構成の一例を示すブロック図である。

[図4]図１の情報処理サーバにおいてカタログを生成する動作例を示すフローチャートである。

[図5]図１のショッピングサーバのウェブページの一例を示す説明図である。

[図6]図５のウェブページのソースコードの一例を示す説明図である。

[図7]図２の属性記述パターンデータベースに記憶された属性記述パターンの一例を示す模式図である。

[図8]属性・属性値の抽出の様子の一例を示す模式図である。

[図9]抽出された属性・属性値の一例を示す模式図である。

[図10]生成された商品等カタログの一例を示す模式図である。

[図11]図１の情報処理サーバにおける属性選定のサブルーチンの一例を示すフローチャートである。

[図12]図４の商品等のカタログ生成の第１変形例の動作例を示すフローチャートである。

[図13]図１２の第１変形例の属性・属性値の抽出の様子の一例を示す模式図である。

[図14]図４の商品等のカタログ生成の第２変形例の動作例を示すフローチャートである。

[図15]図１４の属性・属性値抽出のサブルーチンの一例を示すフローチャートである。

[図16]図１４の属性記述パターン抽出のサブルーチンの一例を示すフローチャートである。

[図17]図１の情報処理サーバにおいて属性・属性値を判定する動作例を示すフローチャートである。

[図18]商品等供給元のウェブページの一例を示す説明図である。

[図19]生成された商品等カタログの一例を示す模式図である。

[図20]第２実施形態に係る情報処理システムにおいてウェブページを抽出する動作例を示すフローチャートである。

[図21]図２０のウェブページ抽出の第１変形例の動作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

[0029] 以下、図面を参照して本発明の実施形態について説明する。

（第１実施形態）

[0030] ［１．情報処理システムの構成および機能概要］

まず、本発明の第１実施形態に係る情報処理システムの構成および概要機能について、図１を用いて説明する。

[0031] 図１は、本実施形態に係る情報処理システム１の概要構成例を示す模式図である。

[0032] 図１に示すように、情報処理システム１は、ウェブページから商品等のカタログ生成したり、誤ったカテゴリに登録されたウェブページを抽出したりする情報処理サーバ（情報処理装置の一例）１０と、ショッピングサイトを運営するためや、ブログのため情報提供サーバ２０と、ショッピングサイトに出店している店舗主の店舗主端末３０と、ショッピングサイトで商品等（ショッピングサイトで提供されているサービスを含む）を購入したり、ブログを投稿するユーザのユーザ端末３５と、を備えている。なお、商品等や、ブログは、ウェブページに記載されている対象の一例である。

[0033] 情報処理サーバ１０と、情報提供サーバ２０とは、ローカルエリアネットワーク等により接続され、相互にデータの送受信が可能になっていて、サーバシステム５を構成している。そして、サーバシステム５と、店舗主端末３０と、ユーザ端末３５とは、ネットワーク３により接続され、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３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）、およびゲートウェイ等により構築されている。

[0034] 情報処理システム１は、ウェブページからカタログを生成するカタログ生成システムとして、または、誤ったカテゴリに登録されたウェブページを抽出するウェブページ抽出システムとして機能する。

[0035] 情報処理サーバ１０は、情報提供サーバ２０等に登録されたウェブページから商品等のカタログを生成したり、当該カタログをユーザ端末３５等から閲覧できるようにしたりする。また、情報処理サーバ１０は、情報提供サー

バ２０等に登録されたウェブページから誤ったカテゴリに登録されたウェブページを抽出したり、抽出結果に基づき、情報提供サーバ２０上のウェブページの整理を行ったり、ウェブページを登録した店舗主等の店舗主端末３０等に抽出結果を通知したりする。

[0036] 情報提供サーバ２０は、ショッピングサイトで商品等を販売するためのウェブサーバや、データベースサーバ等として機能し、ウェブページの登録の受け付けや、ユーザ登録や、商品等の購入手続き等の各種処理を行う。また、情報提供サーバ２０は、商品等のカテゴリ毎に分類されたウェブページを有している。また、情報提供サーバ２０は、ユーザからのブログの投稿を受け付け、ブログの内容等の基づきカテゴリ毎に分類しインターネット上に公開する。

[0037] 店舗主が使用する店舗主端末３０は、パーソナルコンピュータや携帯型無線電話機やＰＤＡ（Personal Digital Assistant）等の携帯端末である。店舗主は店舗主端末３０を使用して、ウェブページを情報提供サーバ２０に登録したり、更新したりする。

[0038] ユーザが使用するユーザ端末３５は、パーソナルコンピュータや携帯型無線電話機やＰＤＡ等の携帯端末である。ユーザはユーザ端末３５を使用して、商品等の検索や商品等の購入等を行う。

[0039] ［２．各サーバの構成および機能］

（２．１　情報処理サーバ１０の構成および機能）

次に、情報処理サーバ１０の構成および機能について、図２を用いて説明する。

[0040] 図２は、情報処理サーバ１０の概要構成の一例を示すブロック図である。

[0041] 図２に示すように、コンピュータとして機能する情報処理サーバ１０は、通信部１１と、記憶部１２と、入出力インターフェース部１３と、システム制御部１４と、を備えている。そして、システム制御部１４と入出力インターフェース部１３とは、システムバス１５を介して接続されている。

[0042] 通信部１１は、ネットワーク３に接続してユーザ端末３５等との通信状態

を制御したり、ローカルエリアネットワークに接続して、情報提供サーバ２０等の他のサーバとデータの送受信を行ったりする。

[0043] 記憶部１２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、オペレーティングシステムおよびサーバプログラム等の各種プログラムや、データ等を記憶する。なお、各種プログラムは、例えば、他のサーバ装置等からネットワーク３を介して取得されるようにしてもよいし、記録媒体に記録されてドライブ装置を介して読み込まれるようにしてもよい。

[0044] また、記憶部１２には、属性記述パターンデータベース（以下「属性記述パターンＤＢ」とする。）１２ａ、属性・属性値データベース（以下「属性・属性値ＤＢ」とする。）１２ｂ等が構築されている。

[0045] 属性記述パターンＤＢ１２ａには、商品等やブログの属性の記述に用いられる属性記述パターンの初期データや、ウェブページから抽出した属性記述パターンが記憶されている。なおブログの属性としてブログのカテゴリが挙げられる。

[0046] 属性・属性値ＤＢ１２ｂには、情報処理サーバ１０による処理の一例として、ウェブページから抽出した商品等の属性に関する属性名と属性値とが記憶される。ここで、属性関連語の一例として、属性名のみや、属性名を含む語句や、属性名と属性値との組等が挙げられる。また、属性・属性値という表記は、属性と属性値とが対になっている場合で、具体的に属性名と属性値とが組になった場合も含む。

[0047] 次に、入出力インターフェース部１３は、通信部１１および記憶部１２とシステム制御部１４との間のインターフェース処理を行うようになっている。

[0048] システム制御部１４は、ＣＰＵ（Central Processing Unit）１４ａ、ＲＯＭ（Read Only Memory）１４ｂ、ＲＡＭ（Random Access Memory）１４ｃ等により構成されている。そして、システム制御部１４は、ＣＰＵ１４ａが、ＲＯＭ１４ｂや記憶部１２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、複数のウェブページから、属性記述パターンに適合する属性名

や属性値を抽出したり、抽出した属性名や属性値から商品等のカタログを生成したりする。また、システム制御部１４は、複数のウェブページから、属性の属性名の出現回数が所定回数以下のウェブページを、誤ったカテゴリに登録されたウェブページとして抽出したりする。

[0049] （２．２　情報提供サーバ２０の構成および機能）

次に、情報提供サーバ２０の構成および機能について、図３を用いて説明する。

図３は、情報提供サーバ２０の概要構成の一例を示すブロック図である。

[0050] 図３に示すように、情報提供サーバ２０は、通信部２１と、記憶部２２と、入出力インターフェース部２３と、システム制御部２４と、を備え、システム制御部２４と入出力インターフェース部２３とは、システムバス２５を介して接続されている。なお、情報提供サーバ２０の構成および機能は、情報処理サーバ１０の構成および機能とほぼ同じであるので、情報処理サーバ１０の各構成や各機能において、異なるところを中心に説明する。

[0051] 通信部２１は、ネットワーク３やローカルエリアネットワーク等を通して、店舗主端末３０やユーザ端末３５や情報処理サーバ１０等と通信状態を制御等するようになっている。

[0052] 記憶部２２には、商品データベース（以下「情報ＤＢ」とする。）２２ａや、会員データベース（以下「会員ＤＢ」とする。）２２ｂや商品等カタログデータベース（以下「商品等カタログＤＢ」とする。）２２ｃ等が構築されている。

[0053] 情報ＤＢ２２ａには、ウェブページに記載されている対象の一例である商品、サービス、および、ブログ等に関する情報が記憶されている。例えば、情報ＤＢ２２ａには、商品等を識別するための識別子である商品ＩＤに関連付けられ、商品名（サービス名を含む）、種類、商品の画像、サービスに関連した画像、スペック、および、商品等の紹介の要約文等の商品情報や、広告情報等が記憶されている。また、情報ＤＢ２２ａには、ユーザが投稿してきたブログの記事がカテゴリ分けされて記憶されている。また、情報ＤＢ２

２ａには、ＨＴＭＬ（HyperText Markup Language）、ＸＭＬ（Extensible Markup Language）等のマークアップ言語等により記述されたウェブページのファイル等が記憶されている。また、情報ＤＢ２２ａには、製造元情報（製造元ドメインを含む）および販売元情報（販売元ドメインを含む）等の商品供給元の情報が記憶されていて、各商品の商品ＩＤに、各商品の公式の情報が記載されている商品供給元のウェブページのＵＲＬ（Uniform Resource Locator）等が関連づけられている。

[0054] 会員ＤＢ２２ｂには、会員登録されたユーザ（インターネットショップの利用者）のユーザＩＤ、名称、住所、電話番号、メールアドレス等のユーザ情報が登録されている。このようなユーザ情報は、ユーザＩＤによってユーザ毎に判別可能になっている。また、会員ＤＢ２２ｂには、ユーザがユーザ端末３５からインターネットショップのサイトにログインする際に必要な、ユーザＩＤ、ログインＩＤ、および、パスワードが登録されている。ここで、ログインＩＤおよびパスワードは、ログイン処理（ユーザの認証処理）に使用されるログイン情報である。

[0055] 商品等カタログＤＢ２２ｃには、情報処理サーバ１０により生成された商品等カタログが商品カテゴリ毎、商品毎に記憶される。

[0056] システム制御部２４は、ＣＰＵ２４ａ、ＲＯＭ２４ｂ、ＲＡＭ２４ｃ等により構成されている。そして、システム制御部２４は、ＣＰＵ２４ａが、ＲＯＭ２４ｂや記憶部２２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、店舗主によるウェブページの登録や更新や、ユーザによる商品購入処理や、商品の購買履歴をユーザＩＤ毎に記録させたりする。またユーザ端末３５からの要求により、商品等カタログの情報を送信したりする。

[0057] ［３．第１実施形態の商品等のカタログ生成システムの動作］

次に、本発明の一実施形態に係る情報処理システム１のカタログ生成システムとしての動作について図４～図１１を用いて説明する。

[0058] 図４は、情報処理サーバ１０においてウェブページを抽出する動作例を示すフローチャートである。図５は、情報提供サーバ２０のウェブページの一

例を示す説明図である。図６は、ウェブページのソースコードの一例を示す説明図である。図７は、属性記述パターンＤＢに記憶された属性記述パターンの一例を示す模式図である。図８は、属性・属性値の抽出の様子の一例を示す模式図である。図９は、抽出された属性・属性値の一例を示す模式図である。図１０は、生成された商品等カタログの一例を示す模式図である。図１１は、情報処理サーバ１０における属性選定のサブルーチンの一例を示すフローチャートである。

[0059] （３．１．商品等のカタログ生成の流れ）

まず、商品等のカタログ生成の流れについて、図４を用いて説明する。

[0060] 図４に示すように、情報処理サーバ１０は、複数のウェブページを取得する（ステップＳ１）。具体的には、情報処理サーバ１０のシステム制御部１４は、情報提供サーバ２０により運営されるショッピングサイトの同一のカテゴリに所属している商品に関して、このカテゴリの全ウェブページを、通信部１１を通して情報ＤＢ２２ａから取得する。さらに具体的には、図５に示すように、テキスト部５１、５２、５３、５４のテキストデータを含むウェブページ５０等が取得される。また、ウェブページ５０のソースコードは、図６に示すように、ＨＴＭＬ等のマークアップ言語等で記述されている。このように、情報処理サーバ１０のシステム制御部１４および通信部１１は、ウェブページに記載されている対象を分類するカテゴリにおいて、同一のカテゴリに属している複数のウェブページを取得するウェブページ取得手段の一例として機能する。

[0061] 次に、情報処理サーバ１０は、属性記述パターンを取得する（ステップＳ２）。具体的には、情報処理サーバ１０のシステム制御部１４は、下記のステップＳ３～Ｓ５におけるブートストラップ法の初期データとして、図７に示すように、属性記述パターンＤＢ１２ａの属性記述パターンリストから、初期の属性記述パターンを取得する。ここで、属性記述パターンは、図７に示すように、前部、中部、および、後部に分かれていて、例えば、属性記述パターン”［　：　］”の場合、前部”［”、中部”：”、および、後部”

］”である。前部と中部との間の語句が属性名で、中部と後部との間の語句が属性値である。また、属性記述パターンには、ＨＴＭＬタグの要素が含まれる場合がある。このように情報処理サーバ１０のシステム制御部１４は、初期データとして、ウェブページに記載されている対象の属性の記述に用いられる属性記述パターンを取得する初期データ取得手段の一例として機能する。

[0062] 次に、情報処理サーバ１０は、属性記述パターンに適合する属性・属性値を抽出する（ステップＳ３）。具体的には、情報処理サーバ１０のシステム制御部１４は、ウェブページ５０等の複数のウェブページの中から、図８に示すように、属性記述パターン６１等に適合する語句の部分（例えば”［品種：〇〇〇］”）を取り出し、属性名”品種”や、属性名”品種”に対応した属性値”〇〇〇”等を抽出する。そして、抽出した属性名および属性値は、属性リストとして属性・属性値ＤＢ１２ｂに記憶される。ここで、どんなパターンにもマッチする特殊文字、すなわち、”＊”や”？”等のワイルドカードと属性記述パターンとが用いられて、属性・属性値が抽出される。なお、属性・属性値の例として、旅行関連サービスの場合、［宿泊料金：〇〇〇］、ブログであるイベント紹介をしている場合、［会場：〇〇〇］等が挙げられる。

[0063] このように情報処理サーバ１０のシステム制御部１４は、複数のウェブページから、属性記述パターンに適合する属性の属性関連語を抽出する属性抽出手段の一例として機能する。また、情報処理サーバ１０のシステム制御部１４は、属性関連語として、属性の属性名を抽出する属性抽出手段の一例として機能する。また、情報処理サーバ１０のシステム制御部１４は、抽出された属性関連語から属性リストを生成する属性リスト生成手段として機能する。

[0064] 次に、情報処理サーバ１０は、属性・属性値に適合する属性記述パターンを抽出する（ステップＳ４）。具体的には、情報処理サーバ１０のシステム制御部１４は、図８に示すように、属性・属性値６２（例えば、属性名”品

種”および属性値”○○○”）に適合する（例えば、<td> 品種</td><td>○○○<td>）を取り出し、属性記述パターンをウェブページ５０等の複数のウェブページの中から抽出する。そして、抽出した属性記述パターンは、図７に示すように、属性記述パターンリストに追加され、属性記述パターンＤＢ１２ａに記憶される。なお、例えば、”容量　*ｍｌ”のように、属性値に関しては、抽出された属性値自体でなく、ワイルドカードが用いられてもよい。

[0065] このように情報処理サーバ１０のシステム制御部１４は、複数のウェブページから、属性関連語に適合する属性記述パターンを抽出する属性記述パターン抽出手段の一例として機能する。また、情報処理サーバ１０のシステム制御部１４は、抽出された属性記述パターンのパターンリストを生成するパターンリスト生成手段として機能する。

[0066] 次に、情報処理サーバ１０は、所定回数を判定する（ステップＳ５）。具体的には、情報処理サーバ１０のシステム制御部１４は、ステップＳ３およびステップＳ４を反復実行した回数が所定回数に達しているか否かを判定する。そして、所定回数に達していない場合（ステップＳ５；ＮＯ）、情報処理サーバ１０のシステム制御部１４は、ステップＳ３に戻り、抽出した新たな属性記述パターンにより、新たな属性・属性値を抽出する。情報処理サーバ１０のシステム制御部１４は、所定回数に達するまで、ステップＳ３およびステップＳ４を繰り返す。

[0067] このように情報処理サーバ１０のシステム制御部１４は、ステップＳ２からステップＳ４において、商品の属性の記述に用いられる属性記述パターンを取得する属性記述パターン取得ステップと、複数のウェブページから、属性記述パターンに適合する属性の属性関連語を抽出する属性抽出ステップと、抽出された属性関連語に基づき、属性抽出ステップで使用する属性記述パターンを、複数のウェブページから更に抽出する属性記述パターン抽出ステップとを実行する。情報処理サーバ１０のシステム制御部１４は、属性抽出手段および属性記述パターン抽出手段を交互に繰り返す繰返手段の一例とし

て機能する。

[0068] 所定回数に達した場合（ステップＳ５；ＹＥＳ）、情報処理サーバ１０は、属性の選定を行う（ステップＳ６）。具体的には、情報処理サーバ１０のシステム制御部１４は、ステップＳ３で抽出した属性名および属性値から、属性選定のサブルーチンにより属性を選定する。属性選定のサブルーチンでは、情報処理サーバ１０のシステム制御部１４は、属性にスコアを付けてランク付けしたり、ノイズの属性を除去したり、同義語の属性を集約する（詳細は後述）。図９に示すように、カテゴリ”ワインにおいては、属性名”品種”、”生産者”等に対して、それぞれの属性値を得る。

[0069] 次に、情報処理サーバ１０は、属性・属性値に基づき商品等カタログを生成する（ステップＳ７）。具体的には、情報処理サーバ１０のシステム制御部１４は、図１０に示すように、商品毎に属性名を並べ、属性名と属性値を組にして商品等カタログを生成する。なお、図１０に示すように、商品の画像を商品等カタログに加えてもよい。属性名の順番は、後述する属性のスコアに基づき決定してもよい。

[0070] このように情報処理サーバ１０のシステム制御部１４は、抽出された属性関連語に基づき、属性関連語が記載された商品等カタログを生成するカタログ生成手段の一例として機能する。

[0071] 次に、情報処理サーバ１０のシステム制御部１４は、他のカテゴリのウェブページに対しても、ステップＳ１～ステップＳ７を適用して、商品等カタログを生成する。そして、情報処理サーバ１０のシステム制御部１４は、生成した商品等カタログの情報を、情報提供サーバ２０に送信し、商品等カタログＤＢ２２ｃに記憶させる。

[0072] （３．２．属性の選定）

次に、属性の選定のサブルーチンについて、図１１を用いて説明する。

[0073] 図１１に示すように、情報処理サーバ１０は、属性へのスコア付けを行う（ステップＳ１０）。具体的には、ショッピングサイトが商品を販売する複数の店舗を有する場合、すなわち、サイバーモールを構成する場合、情報処

理サーバ１０のシステム制御部１４は、属性名が出現したウェブページを有する店舗の数を求め、属性のスコアとする。

[0074] 多種の店舗のウェブページに出現した属性関連語の一例の属性名は、属性として適切であるという仮定に基づいている。例えば、ワインのウェブページにおいて、適切な属性である”品種”という属性は多種の店舗のウェブページに出現する。それに対して、いずれかの属性記述ターンにマッチした不適切な属性は、１店舗のウェブページからしか獲得されないことが多く、属性のスコアが低くなる傾向がある。このように情報処理サーバ１０のシステム制御部１４は、属性関連語のスコア付けを行う属性スコアリング手段の一例として機能する。また、情報処理サーバ１０のシステム制御部１４は、ウェブページに記載されている対象を販売する複数の店舗を有するウェブサイトにおいて、属性関連語が出現しているウェブページの店舗の数に基づき、属性関連語のスコア付けを行う属性スコアリング手段の一例として機能する。

[0075] 次に、情報処理サーバ１０は、上位ランクの属性を選択する（ステップＳ１１）。具体的には、情報処理サーバ１０のシステム制御部１４は、属性のスコアの高い順に属性名をランク付けし、所定のランク以上の属性名を選択する。このように情報処理サーバ１０のシステム制御部１４は、スコアの順に属性関連語のランク付けを行い、所定のランク以上の属性関連語を選択する属性選択手段の一例として機能する。

[0076] 次に、情報処理サーバ１０は、属性のフィルタリングを行う（ステップＳ１２）。具体的には、情報処理サーバ１０のシステム制御部１４は、各カテゴリにおける属性名の出現確率を用いて、属性のフィルタリングを行う。他のカテゴリにおいても出現する属性名は、属性として不向きであるという仮定に基づいて、属性のフィルタリングが行われている。例えば、属性として不向きな”送料無料”のような語句は、多数のウェブページに出現するため、各カテゴリにおける出現確率が、似通った値になる。一方、”品種”という属性名はワインのカテゴリのウェブページにはよく出現するが、ゴルフド

ライバーや靴等のカテゴリには出現しないため、ワインのカテゴリにおける出現確率が、ワイン以外のカテゴリにおける出現確率よりも高くなる。このように情報処理サーバ１０のシステム制御部１４は、カテゴリとは異なるカテゴリに属している複数のウェブページにおいて出現する属性関連語を取り除く属性フィルタ手段の一例として機能する。

[0077] 次に、情報処理サーバ１０は、同義の属性を集約する（ステップＳ１３）。属性の中には同じ概念を持つものが存在している。例えば、ワインのカテゴリにおいて、”品種”、”ぶどう品種”、”ブドウ品種”、”セパージュ”、”葡萄品種”は同義の属性名である。情報処理サーバ１０のシステム制御部１４は、同義語辞書を用いたり、属性名同士の類似の度合いを算出したり、属性名に対応する属性値を用いたりして、同義の属性の属性名を集約する。なお、同義の属性の属性名を集約するのではなく、類似概念の属性の属性名を集約してもよい。

[0078] 具体的には、属性名”Ａ”（属性Ａ）と属性名”Ｂ”（属性Ｂ）との類似の度合いを算出する場合、属性Ａの属性値の中で、属性Ｂが持っている属性値と共通なものの割合と、属性Ｂの属性値の中で属性Ａの属性値が持っている属性値と共通なものの割合を掛け合わせた値を類似の度合いとしたり、これらの割合を元にエントロピーを計算して掛け合わせた値を類似の度合いとしたり、ジャッカード係数を類似の度合いとしたり、属性Ａと属性Ｂの属性値中で共通なものの種類の数を類似の度合いとする。

[0079] このように情報処理サーバ１０のシステム制御部１４は、属性名同士が類似であるか否かを判定する属性名類似判定手段の一例として機能する。また、情報処理サーバ１０のシステム制御部１４は、属性名類似判定手段により類似と判定された属性名を集約する属性名集約手段の一例として機能する。また、情報処理サーバ１０のシステム制御部１４は、属性関連語として、属性名および属性名に対応する属性値を抽出する属性抽出手段、および、属性値に基づき属性名を集約する属性名集約手段の一例として機能する。

[0080] 本実施形態によれば、ウェブページに記載されている対象を分類するカテ

ゴリが同一である複数のウェブページを取得し、初期データとして、ウェブページに記載されている対象の属性に関連した属性関連語、または、当該対象の属性の記述に用いられる属性記述パターンを取得し、複数のウェブページから、属性記述パターンに適合する属性の属性関連語を抽出し、複数のウェブページから、属性関連語に適合する属性記述パターンを抽出することにより、同一の前記カテゴリに属している複数のウェブページから、属性関連語を抽出し、属性記述パターンを抽出するか、または、属性記述パターンを抽出し、属性関連語を抽出しているので、同一のカテゴリに含まれる属性を精度良く取得できる。例えば、属性関連語および属性記述パターンを相互に繰り返し抽出すると、同一のカテゴリに含まれる属性を精度良く取得できる。

[0081] 情報処理サーバ１０のシステム制御部１４が、属性抽出手段および属性記述パターン抽出手段を交互に繰り返す場合、属性リストやパターンリストをブートストラップによって拡張して、初期値として与えた属性以外の属性を抽出することができる。また、この抽出された属性により、ウェブページの類似度が判定できる。また、ユーザがウェブページに関する商品等カタログを使用して、所望の商品に到達しやすくなり、ユーザの利便性の向上を図ることができる。

[0082] また、情報処理サーバ１０のシステム制御部１４が、抽出された属性関連語から属性リストを生成し、抽出された属性記述パターンのパターンリストを生成する場合、カテゴリ毎に、属性名や属性値等の属性関連語や属性記述パターンの情報を蓄積できる。

[0083] また、情報処理サーバ１０のシステム制御部１４が、属性関連語のスコア付けを行い、上位のランクの属性関連語を選択する場合、選択された属性関連語において、商品等を表す属性やブログの属性の精度が高くなる。

[0084] また、情報処理サーバ１０のシステム制御部１４が、対象を販売する複数の店舗を有するウェブサイトにおいて、属性関連語が出現しているウェブページの店舗の数に基づいて属性関連語のスコア付けを行う場合、属性関連語

を選択する際、商品等を表す属性の精度が高くなる。例えば、店舗により扱う商品等の数や、ウェブページの数が大きく異なる場合、多くの商品等を扱う店舗の影響を受けやすくなるが、店舗の数に基づき属性関連語のスコア付けを行うことにより、ある特定の店舗の影響を解消できる。

[0085] また、情報処理サーバ１０のシステム制御部１４が、他のカテゴリに属している複数のウェブページにおいて出現する属性関連語を取り除く場合、対象のカテゴリ固有の属性関連語に絞ることにより、商品等を表す属性やブログの属性の精度が高くなる。

[0086] また、情報処理サーバ１０のシステム制御部１４が、属性関連語として属性の属性名を抽出する場合、同一のカテゴリに含まれる属性・属性名を精度良く取得できる。また、属性名により、誤ったカテゴリに登録されたウェブページを抽出できる。

[0087] また、情報処理サーバ１０のシステム制御部１４が、属性名同士が類似であるか否かを判定し、類似と判定された属性名を集約する場合、重複した属性名を取り除かれ、属性名が利用しやすくなる。

[0088] また、情報処理サーバ１０のシステム制御部１４が、属性関連語として、属性名および属性名に対応する属性値を抽出し、属性値に基づき属性名を集約する場合、属性名に直結した属性値により、属性名が集約しやすくなる。

[0089] また、情報処理サーバ１０のシステム制御部１４が、対象の供給元のウェブページを取得し、初期データ取得手段、属性抽出手段、および、属性記述パターン抽出手段により、対象の供給元のウェブページから供給元対象属性関連語を抽出し、供給元対象属性関連語と属性関連語とを比較する場合、同一のカテゴリに含まれる属性をより精度良く取得できる。また、商品等の対象に関する公式な対象情報を取り入れ、生成されたカタログの精度を判定することにより、カタログの信頼性を向上させることができる。

[0090] また、情報処理サーバ１０のシステム制御部１４が、抽出された属性関連語に基づき、属性関連語が記載されたカタログを生成する場合、ユーザがウェブページに関するカタログを使用して、所望の商品等の対象に到達しやす

くなり、ユーザの利便性の向上を図ることができる。

[0091] ［４．商品等のカタログ生成システムの動作の第１変形例］

次に、商品等のカタログ生成システムの動作の第１変形例について図１２および図１３に基づきについて説明する。

[0092] なお、上記実施形態と同一または対応する部分には、同一の符号を用いて動作等を説明する。その他の変形例も同様とする。

[0093] 図１２は、商品等のカタログ生成の第１変形例の動作例を示すフローチャートである。図１３は、第１変形例の属性・属性値の抽出の様子の一例を示す模式図である。図１２に示すように、本変形例は、ブートストラップ法における初期データを、属性記述パターンではなく、属性・属性値とした点である。ステップＳ２２からステップＳ２４が、上記実施形態と異なるステップである。なお、属性・属性値ＤＢ１２ｂには、属性・属性値の初期データが記憶されている。

[0094] まず、情報処理サーバ１０は、ステップＳ１と同様に、複数のウェブページを取得する（ステップＳ２１）。

[0095] 次に、情報処理サーバ１０は、属性・属性値を取得する（ステップＳ２２）。具体的には、情報処理サーバ１０のシステム制御部１４は、下記のステップＳ２３～Ｓ２５におけるブートストラップ法の初期データとして、属性・属性値ＤＢ１２ｂの属性・属性値リストから、図１３に示すように、初期の属性・属性値６６を取得する。このように情報処理サーバ１０のシステム制御部１４は、初期データとして、商品の属性に関連した属性関連語を取得する初期データ取得手段の一例として機能する。

[0096] 次に、情報処理サーバ１０は、属性・属性値に適合する属性記述パターンを抽出する（ステップＳ２３）。具体的には、情報処理サーバ１０のシステム制御部１４は、ウェブページ５０等の複数のウェブページの中から、図１３に示すように、属性・属性値６６等に適合する語句の部分（例えば”［品種：〇〇〇］”）を取り出し、属性記述パターン” ［ ： ］”等を抽出する。そして、抽出した属性記述パターンは、属性記述パターンリストとして

属性記述パターンＤＢ１２ａに記憶される。ここで、ワイルドカードと属性・属性値とが用いられて、属性記述パターンが抽出される。

[0097] 次に、情報処理サーバ１０は、属性記述パターンに適合する属性・属性値を抽出する（ステップＳ２４）。具体的には、情報処理サーバ１０のシステム制御部１４は、図１３に示すように、属性記述パターン６７（例えば、属性記述パターンの前部”［”、中部”：”、後部”］”）に適合する、例えば、”［アルコール度数：１２．５％”］を取り出し、属性・属性値をウェブページ５０等の複数のウェブページの中から抽出する。そして、抽出した属性・属性値は、属性・属性値リストに追加され、属性・属性値ＤＢ１２ｂに記憶される。

[0098] 以下のステップＳ２５からステップＳ２８は、ステップＳ５からステップＳ８と同様である。

[0099] 以上のように、情報処理サーバ１０のシステム制御部１４は、ステップＳ２２からステップＳ２４において、商品を分類するカテゴリにおいて、同一のカテゴリに属している複数のウェブページを取得するウェブページ取得ステップと、商品の属性に関連した属性関連語を取得する属性関連語取得ステップと、属性の記述に用いられる属性記述パターンであって、複数のウェブページから、属性関連語に適合する属性記述パターンを抽出する属性記述パターン抽出ステップと、抽出された属性関連語に基づき、属性記述パターン抽出手段に使用する属性関連語を、複数のウェブページから更に抽出する属性関連語抽出ステップとを実行する。

[0100] 本変形例によれば、商品を分類するカテゴリにおいて、同一のカテゴリに属している複数のウェブページを取得し、属性・属性値ＤＢ１２ｂから商品の属性に関連した属性関連語を取得し、属性の記述に用いられる属性記述パターンであって、複数のウェブページから属性関連語に適合する属性記述パターンを抽出し、抽出された属性関連語に基づき、属性記述パターンの抽出に使用する属性関連語を、複数のウェブページから更に抽出し、抽出された属性関連語に基づき、属性関連語が記載された商品等カタログを生成するこ

とにより、ユーザがウェブページに関する商品等カタログを使用して、所望の商品に到達しやすくなり、ユーザの利便性の向上を図ることができる。

[0101] ［５．商品等のカタログ生成システムの動作の第２変形例］

次に、商品等のカタログ生成システムの動作の第２変形例について図１４～図１６に基づきについて説明する。本変形例では、ブートストラップ法のステップにおいて、属性の選定を行ったり、属性記述パターンの選定を行ったりしている。

[0102] 図１４は、商品等のカタログ生成の第２変形例の動作例を示すフローチャートである。図１５は、属性・属性値抽出のサブルーチンの一例を示すフローチャートである。図１６は、属性記述パターン抽出のサブルーチンの一例を示すフローチャートである。

[0103] （５．１．ウェブページの抽出の流れ）

まず、図１４に示すように、情報処理サーバ１０は、ステップＳ１およびステップＳ２と同様に、複数のウェブページを取得し（ステップＳ３１）、属性記述パターンを取得する（ステップＳ３２）。

[0104] 次に、情報処理サーバ１０は、属性記述パターンに基づき属性・属性値を抽出する（ステップＳ３３）。具体的には、情報処理サーバ１０のシステム制御部１４は、属性・属性値抽出のサブルーチンにより属性・属性値を抽出する。属性・属性値抽出のサブルーチンでは、情報処理サーバ１０のシステム制御部１４は、属性記述パターンに適合する属性・属性値を抽出したり、属性へのスコア付けを行ったり、上位のランクの属性を選択したり、属性のフィルタリングを行ったりする。

[0105] 次に、情報処理サーバ１０は、属性・属性値に基づき属性記述パターンを抽出する（ステップＳ３４）。具体的には、情報処理サーバ１０のシステム制御部１４は、属性記述パターン抽出のサブルーチンにより属性記述パターンを抽出する。属性記述パターン抽出のサブルーチンでは、情報処理サーバ１０のシステム制御部１４は、属性・属性値に適合する属性記述パターンを抽出したり、属性記述パターンと属性・属性値との共起確率を算出したり、

スコアを算出したり、上位のランクの属性記述パターンを選択したりする。

[0106] 次に、情報処理サーバ１０は、ステップＳ５と同様に、所定回数を判定する（ステップＳ３５）。

[0107] 次に、情報処理サーバ１０は、同義の属性を集約する（ステップＳ３６）。具体的には、情報処理サーバ１０のシステム制御部１４は、ステップＳ３３～ステップＳ３５のブートストラップ法により求めた属性名に対して、ステップＳ１３と同様に、同義の属性の属性名の集約を行う。

[0108] 次に、情報処理サーバ１０は、ステップＳ７と同様に、属性、属性値に基づき、商品等カタログを生成する（ステップＳ３７）。

[0109] （５．２．属性・属性値の抽出）

次に、属性・属性値抽出のサブルーチンについて、図１５を用いて説明する。

[0110] 図１５に示すように、情報処理サーバ１０は、属性記述パターンに適合する属性・属性値を抽出する（ステップＳ４０）。具体的には、情報処理サーバ１０のシステム制御部１４は、ステップＳ３と同様に、属性記述パターンに適合する属性・属性値を抽出する。

[0111] 次に、情報処理サーバ１０は、属性選定のサブルーチンにおけるステップＳ１０～ステップＳ１２と同様に、属性へのスコア付けを行い（ステップＳ４１）、上位のランクの属性を選択し（ステップＳ４２）、属性のフィルタリングを行う（ステップＳ４３）。

[0112] （５．３．属性記述パターンの抽出）

次に、属性記述パターン抽出のサブルーチンについて、図１６を用いて説明する。

[0113] 図１６に示すように、情報処理サーバ１０は、ステップＳ４と同様に属性・属性値に適合する属性記述パターンを抽出する（ステップＳ４５）。

[0114] 次に、情報処理サーバ１０は、属性記述パターンと属性・属性値との共起確率を算出する（ステップＳ４６）。具体的には、情報処理サーバ１０のシステム制御部１４は、属性関連語と属性記述パターンとが共に出現する共起

数の一例として、属性記述パターンｔと、対象となっている同一カテゴリの複数のウェブページにおいて、属性・属性値の対ｉとの共起数ｆ（ｉ，ｔ）を算出する。そして、情報処理サーバ１０のシステム制御部１４は、属性記述パターンｔと属性・属性値の組ｉが共起する確率、すなわち、式（１）の共起確率Ｐｔ（ｉ）を算出する。

Ｐｔ（ｉ）＝ｆ（ｉ，ｔ）／Ｎｔ　　・・・（１）

ここで、Ｎｔは、抽出した属性記述パターンｔの数である。

[0115] 次に、情報処理サーバ１０は、エントロピー（スコア）を算出する（ステップＳ４７）。様々な属性・属性値と共起する属性記述パターンは、適切な属性記述パターンであるという仮定に基づいて、情報処理サーバ１０のシステム制御部１４は、属性記述パターンに対するエントロピーＨ（ｔ）を式（２）により算出する。

$$H(t) = -\sum_{i \in I} Pt(i) \times \log_2 Pt(i) \quad \cdots (2)$$

[0116] ここで、Ｉは、属性・属性値の組ｉを要素とする属性・属性値の組の集合、属性記述パターンｔを要素とするＴは属性記述パターン集合である。

[0117] 次に、情報処理サーバ１０は、上位のスコアの属性記述パターンを選択する（ステップＳ４８）。具体的には、情報処理サーバ１０のシステム制御部１４は、スコアとしてエントロピーＨ（ｔ）の高い属性記述パターンからランク付けをして、所定のランク以上の属性記述パターンを選択する。このように情報処理サーバ１０のシステム制御部１４は、属性記述パターンのスコア付けを行う属性記述パターン・スコアリング手段の一例として機能する。また、情報処理サーバ１０のシステム制御部１４は、スコアの順に属性記述パターンのランク付けを行い、所定のランク以上の属性記述パターンを選択する属性記述パターン選択手段の一例として機能する。また、情報処理サーバ１０のシステム制御部１４は、属性関連語と属性記述パターンとが共に出現する共起数に基づき属性記述パターンのスコア付けを行う属性記述パターン・スコアリング手段の一例として機能する。

[0118] なお、ステップＳ４６～ステップＳ４８は、ステップＳ５までに得られた

属性記述パターンリストに対して、属性記述パターンの選定のステップとして使用されて、無駄な属性記述パターンを省くことができる。

[0119] 以上のように、本変形例では、特に、ブートストラップの回数を増やした場合に、抽出される属性・属性値や、属性記述パターンが増大することを防止することができる。

[0120] また、情報処理サーバ１０のシステム制御部１４が、属性記述パターンのスコア付けを行い、上位のランクの属性記述パターンを選択する場合、属性・属性値を抽出するための属性記述パターンの精度が高くなる。

[0121] また、情報処理サーバ１０のシステム制御部１４が、属性関連語と属性記述パターンとが共に出現する共起数に基づき属性記述パターンのスコア付けを行う場合、属性記述パターンを選択する上でのスコアの精度が高くなる。

[0122] なお、本実施形態やその変形例において、属性のスコアとして、店舗数でなく、属性名が出現したウェブページの数でもよい。情報処理サーバ１０のシステム制御部１４が、属性スコアリング手段として、属性関連語の検索のヒット件数に基づき、属性関連語のスコア付けを行う。この場合、店舗が多く集まるサイバーモール以外にも適用できる。

[0123] また、商品等の対象の供給元のウェブページから、各商品等の対象の属性・属性値を求め、この属性・属性値により、商品等カタログの属性・属性値の精度の判定を行ってもよい。この場合、商品等に関する公式な商品等情報を取り入れ、生成された商品等カタログの精度を判定することにより、商品等カタログの信頼性を向上させることができる。

[0124] 例えば、図１７に示すように、情報処理サーバ１０のが、商品等の供給元対象の製造元や輸入販売元等の対象供給元のウェブページを通信部を通して取得する（ステップＳ５１）。具体的には、情報処理サーバ１０のシステム制御部１４が、情報ＤＢ２２ａに記憶されたＵＲＬ等を参照して、図１８に示すような、各商品等の商品ＩＤに対応した供給元のウェブページを取得する。なお、商品ＩＤは、ショッピングサイトにおいて、ウェブページに予め付されている商品ＩＤや、ウェブページのテキストデータから抽出した商品

ＩＤでもよい。このように、情報処理サーバ１０のシステム制御部１４および通信部１１が、商品等の対象の供給元のウェブページを取得するウェブページ取得手段の一例として機能する。

[0125] 次に、情報処理サーバ１０のが、供給元商品の属性・属性値を抽出する（ステップＳ５２）。具体的には、情報処理サーバ１０のシステム制御部１４が、ステップＳ２～ステップＳ６や、ステップＳ２２～ステップＳ２６や、ステップＳ３２～ステップＳ３６のようにして、商品供給元のウェブページから、商品供給元のウェブページに関する属性・属性値を抽出する。

[0126] 次に、情報処理サーバ１０のが、供給元商品等の供給元対象の属性・属性値と商品等カタログの属性・属性値とを比較して、商品等カタログの精度を判定する。具体的には、情報処理サーバ１０のシステム制御部１４が、商品等カタログの属性名に、供給元商品等の供給元対象の属性名があるか否か、また、属性名がある場合、その属性値が一致しているか否かを比較する。そして、同一の属性名が少ない場合や属性値が一致しない場合は、生成された商品等カタログの精度が低いと判断する。また、属性値が一致しない場合は、店舗側での入力ミスと考えることができる。そして、属性名が一致している数や、属性値が一致している数に所定値を設け、所定値以上の場合の商品等カタログの情報が、ユーザ端末３５から閲覧されるようにする。このように、情報処理サーバ１０のシステム制御部１４および通信部１１が、初期データ取得手段、属性抽出手段、および、属性記述パターン抽出手段により、商品等の対象の供給元のウェブページから供給元対象属性関連語を抽出し、供給元対象属性関連語と、属性関連語とを比較する属性関連語比較手段の一例として機能する。

[0127] 次に、生成された商品等カタログの変形例について図１９を用いて説明する。

図１９は、生成された商品等カタログの一例を示す模式図である。

[0128] 図１９に示すように、属性名”製造年”に対して、属性値”１９９５年”および属性値”１９９６年”によりグルーピングされている。情報処理サー

バ１０のシステム制御部１４が、ステップＳ６や、ステップＳ２６や、ステップＳ３６等で、属性関連語を求めた後、属性名”製造年”対して属性値”１９９５年”を有するウェブページを収集し、また、属性名”製造年”対して属性値”１９９６年”を有するウェブページを収集して、属性関連語に基づき、複数のウェブページの情報をグルーピングする。図１９に示すように、商品の商品名と、他の属性とが、属性名”製造年”対する属性値によりグルーピングされる。

[0129] 情報処理サーバ１０のシステム制御部１４が、属性関連語に基づき、複数のウェブページをグルーピングする場合、共通する属性によりグルーピングされたウェブページにまとめることができる。また、ユーザが見やすい検索結果に反映させる等、利用価値が高まる。

[0130] （第２実施形態）

次に、本発明の第２実施形態に係る情報処理システム１の動作について、図を用いて説明する。なお、前記第１実施形態と同一または対応する部分には、同一の符号を用いて異なる構成および作用のみを説明する。その他の実施形態および変形例も同様とする。

[0131] ［６．第２実施形態に係るウェブページ抽出システムの動作］

次に、本発明の第２実施形態に係る情報処理システム１のウェブページ抽出システムとしての動作について図２０を用いて説明する。

[0132] まず、ウェブページの抽出の流れについて、図２０を用いて説明する。

図２０は、第２実施形態に係る情報処理システムにおいてウェブページを抽出する動作例を示すフローチャートである。

[0133] 図２０に示すように、情報処理サーバ１０は、複数のウェブページを取得し、属性を求める（ステップＳ６０）。具体的には、情報処理サーバ１０のシステム制御部１４は、ステップＳ１～ステップＳ６と同様に複数のウェブページを取得し、選定した属性の属性名および属性値を求める。または、情報処理サーバ１０のシステム制御部１４は、ステップＳ２１～ステップＳ２６と同様に複数のウェブページを取得し、選定した属性の属性名および属性

値を求める。

[0134] 次に、情報処理サーバ１０は、各ウェブページの属性の出現回数を算出する（ステップＳ６１）。具体的には、情報処理サーバ１０のシステム制御部１４は、ステップＳ６０で取得した各ウェブページで、選定した属性の属性名の出現回数を算出する。なお、集約された属性名の同義語も考慮して、出現回数が算出される。

[0135] 次に、情報処理サーバ１０は、出現回数が所定回数以下のウェブページを抽出する（ステップＳ６２）。具体的には、情報処理サーバ１０のシステム制御部１４は、ステップＳ６０で取得したウェブページで、選定した属性の属性名の出現回数がゼロ、すなわち、選定した属性の属性名が出現しないウェブページを抽出する。複数の属性名がある場合は、どの属性名も出現しないウェブページを抽出する。情報処理サーバ１０のシステム制御部１４は、この抽出されたウェブページは誤ったカテゴリに登録されたウェブページである、と判定する。

[0136] このように情報処理サーバ１０のシステム制御部１４は、複数のウェブページから、属性関連語の出現回数が所定回数以下のウェブページを抽出するウェブページ抽出手段の一例として機能する。また、情報処理サーバ１０のシステム制御部１４は、属性関連語の出現回数がゼロのウェブページを抽出するウェブページ抽出手段の一例として機能する。

[0137] 次に、情報処理サーバ１０のシステム制御部１４は、他のカテゴリのウェブページに対しても、ステップＳ６０～ステップＳ６２を適用して、誤ったカテゴリに登録されたウェブページを抽出する。そして、情報処理サーバ１０のシステム制御部１４は、抽出したウェブページに関する情報を、店舗主に知らせるために店舗主端末３０に送信したり、抽出したウェブページを正しいカテゴリに移動させたりする。

[0138] 本実施形態によれば、ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得し、初期データとして、ウェブページに記載されている対象の属性に関連した属性関連語、または、当該対

象の属性の記述に用いられる属性記述パターンを取得し、複数のウェブページから、属性記述パターンに適合する属性の属性関連語を抽出し、複数のウェブページから、属性関連語に適合する前記属性記述パターンを抽出し、複数のウェブページから、属性関連語の出現回数が所定回数以下のウェブページを抽出することにより、誤ったカテゴリに登録されたウェブページを抽出することができる。例えば、ワインセラーの商品がワインのカテゴリに登録された場合、ワインセラーに関するウェブページには、ワインの属性のひとつ”品種”等の属性名が出てこない確率が高い。また、属性の属性名を抽出する場合、属性名により、誤ったカテゴリに登録された商品ウェブページを抽出できる。

[0139] また、情報処理サーバ１０のシステム制御部１４が、属性関連語の出現回数がゼロのウェブページを抽出する場合、誤ったカテゴリに登録されたウェブページには、属性関連語が出現する確率が低いので、誤ったカテゴリに登録されたウェブページを容易に抽出することができる。

[0140] ［７．ウェブページ抽出システムの動作の第１変形例］

次に、ウェブページ抽出システムの動作の第１変形例について図２１を用いて説明する。本変形例では、ブートストラップ法のステップにおいて、属性の選定を行ったり、属性記述パターンの選定を行ったりしている。

[0141] 図２１は、ウェブページ抽出の第１変形例の動作例を示すフローチャートである。

[0142] （５．１．ウェブページの抽出の流れ）

まず、図２１に示すように、情報処理サーバ１０は、ステップＳ３１～ステップＳ３６と同様に、複数のウェブページを取得し、属性を求め、同義の属性を集約する（ステップＳ６５）。

[0143] 次に、情報処理サーバ１０は、ステップＳ６１およびステップＳ６２と同様に、各ウェブページの属性の属性名の出現回数を算出し（ステップＳ６６）、出現回数が所定回数以下のウェブページを抽出する（ステップＳ６７）。

[0144] 以上のように、本変形例では、特に、ブートストラップの回数を増やした場合に、抽出される属性・属性値や、属性記述パターンが増大することを防止することができる。

[0145] また、情報処理サーバ１０のシステム制御部１４が、属性記述パターンのスコア付けを行い、上位のランクの属性記述パターンを選択する場合、属性・属性値を抽出するための属性記述パターンの精度が高くなる。

[0146] また、情報処理サーバ１０のシステム制御部１４が、属性関連語と属性記述パターンとの共起数に基づき属性記述パターンのスコア付けを行う場合、属性記述パターンを選択する上でのスコアの精度が高くなる。

[0147] なお、本実施形態やその変形例において、属性のスコアとして、店舗数でなく、属性名が出現したウェブページの数でもよい。情報処理サーバ１０のシステム制御部１４が、属性スコアリング手段として、属性関連語の検索のヒット件数に基づき、属性関連語のスコア付けを行う。この場合、店舗が多く集まるサイバーモール以外にも適用できる。

[0148] また、情報処理サーバ１０のシステム制御部１４が、ウェブページ抽出手段として、複数の属性関連語において、属性関連語の出現回数がゼロの割合が、所定以上のウェブページを抽出してもよい。また、属性関連語の出現回数がゼロでなくても、少数出現回数に基づいてもよい。

[0149] さらに、本発明は、上記各実施形態に限定されるものではない。上記各実施形態は、例示であり、本発明の特許請求の範囲に記載された技術的思想と実質的に同一な構成を有し、同様な作用効果を奏するものは、いかなるものであっても本発明の技術的範囲に包含される。

符号の説明

[0150] ３：ネットワーク

５：サーバシステム

１０：情報処理サーバ（情報処理装置）

１２ａ：属性記述パターンＤＢ

１２ｂ：属性・属性名ＤＢ

２０：情報提供サーバ

２２ｃ：商品等カタログＤＢ

請求の範囲

[請求項1] ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得するウェブページ取得手段と、

初期データとして、前記ウェブページに記載されている対象の属性に関連した属性関連語、または、当該対象の属性の記述に用いられる属性記述パターンを取得する初期データ取得手段と、

前記複数のウェブページから、前記属性記述パターンに適合する前記属性の属性関連語を抽出する属性抽出手段と、

前記複数のウェブページから、前記属性関連語に適合する前記属性記述パターンを抽出する属性記述パターン抽出手段と、

を備えたことを特徴とする情報処理装置。

[請求項2] 請求項１に記載の情報処理装置において、

前記属性抽出手段および前記属性記述パターン抽出手段を交互に繰り返す繰返手段を更に備えたことを特徴とする情報処理装置。

[請求項3] 請求項１または請求項２に記載の情報処理装置において、

前記属性抽出手段が、前記属性関連語として、前記属性の属性名を抽出することを特徴とする情報処理装置。

[請求項4] 請求項１から請求項３のいずれか１項に記載の情報処理装置において、

抽出された前記属性関連語から属性リストを生成する属性リスト生成手段と、

抽出された前記属性記述パターンのパターンリストを生成するパターンリスト生成手段と、

を更に備えたことを特徴とする情報処理装置。

[請求項5] 請求項１から請求項４のいずれか１項に記載の情報処理装置において、

前記属性関連語のスコア付けを行う属性スコアリング手段と、

前記スコアの順に前記属性関連語のランク付けを行い、所定のランク

以上の属性関連語を選択する属性選択手段と、

を更に備えたことを特徴とする情報処理装置。

[請求項6] 請求項５に記載の情報処理装置において、

前記属性スコアリング手段が、前記属性関連語の検索のヒット件数に基づき、前記属性関連語のスコア付けを行うことを特徴とする情報処理装置。

[請求項7] 請求項５に記載の情報処理装置において、

前記属性スコアリング手段が、前記対象を販売する複数の店舗を有するウェブサイトにおいて、前記属性関連語が出現しているウェブページの前記店舗の数に基づき、前記属性関連語のスコア付けを行うことを特徴とする情報処理装置。

[請求項8] 請求項１から請求項７のいずれか１項に記載の情報処理装置において、

前記カテゴリとは異なるカテゴリに属している複数のウェブページにおいて出現する前記属性関連語を取り除く属性フィルタ手段を更に備えたことを特徴とする情報処理装置。

[請求項9] 請求項１から請求項８のいずれか１項に記載の情報処理装置において、

前記属性記述パターンのスコア付けを行う属性記述パターン・スコアリング手段と、

前記スコアの順に前記属性記述パターンのランク付けを行い、所定のランク以上の属性記述パターンを選択する属性記述パターン選択手段と、

を更に備えたことを特徴とする情報処理装置。

[請求項10] 請求項９に記載の情報処理装置において、

前記属性記述パターン・スコアリング手段が、前記属性関連語と前記属性記述パターンとが共に出現する共起数に基づき前記属性記述パターンのスコア付けを行うことを特徴とする情報処理装置。

[請求項11] 請求項１から請求項１０のいずれか１項に記載の情報処理装置において、

前記属性名同士が類似であるか否かを判定する属性名類似判定手段と、

前記属性名類似判定手段により類似と判定された属性名を集約する属性名集約手段と、

を更に備えたことを特徴とする情報処理装置。

[請求項12] 請求項１１に記載の情報処理装置において、

前記属性抽出手段が、前記属性関連語として、前記属性名および前記属性名に対応する属性値を抽出し、

前記属性名集約手段が、前記属性値に基づき前記属性名を集約することを特徴とする情報処理装置。

[請求項13] 請求項１から請求項１０のいずれか１項に記載の情報処理装置において、

前記ウェブページ取得手段が、前記対象の供給元のウェブページを取得し、

前記初期データ取得手段、前記属性抽出手段、および、前記属性記述パターン抽出手段により、前記対象の供給元のウェブページから供給元対象属性関連語を抽出し、前記供給元対象属性関連語と前記属性関連語とを比較する属性関連語比較手段を更に備えたことを特徴とする情報処理装置。

[請求項14] 請求項１から請求項１３のいずれか１項に記載の情報処理装置において、

抽出された前記属性関連語に基づき、前記属性関連語が記載されたカタログを生成するカタログ生成手段を更に備えたことを特徴とする情報処理装置。

[請求項15] 請求項１から請求項１４のいずれか１項に記載の情報処理装置において、

前記複数のウェブページから、前記属性関連語の出現回数が所定回数以下のウェブページを抽出するウェブページ抽出手段と、

を更に備えたことを特徴とする情報処理装置。

[請求項16] 請求項１５に記載の情報処理装置において、

前記ウェブページ抽出手段が、前記属性関連語の出現回数がゼロのウェブページを抽出することを特徴とする情報処理装置。

[請求項17] 請求項１から請求項１６のいずれか１項に記載の情報処理装置において、

前記属性関連語に基づき、前記複数のウェブページをグルーピングするウェブページ・属性グルーピング手段を更に備えたことを特徴とする情報処理装置。

[請求項18] 情報処理装置が情報処理をする情報処理方法において、

ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得するウェブページ取得ステップと、

前記ウェブページに記載されている対象の属性の記述に用いられる属性記述パターンを取得する属性記述パターン取得ステップと、

前記複数のウェブページから、前記属性記述パターンに適合する前記属性の属性関連語を抽出する属性抽出ステップと、

抽出された前記属性関連語に基づき、前記属性抽出ステップで使用する前記属性記述パターンを、前記複数のウェブページから、更に抽出する属性記述パターン抽出ステップと、

を有することを特徴とする情報処理方法。

[請求項19] 情報処理装置が情報処理をする情報処理方法において、

ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得するウェブページ取得ステップと、

前記ウェブページに記載されている対象の属性に関連した属性関連語を取得する属性関連語取得ステップと、

前記属性の記述に用いられる属性記述パターンであって、前記複数

のウェブページから、前記属性関連語に適合する前記属性記述パターンを抽出する属性記述パターン抽出ステップと、

抽出された前記属性関連語に基づき、前記属性記述パターン抽出ステップで使用する属性関連語を、前記複数のウェブページから、更に抽出する属性関連語抽出ステップと、

を有することを特徴とする情報処理方法。

[請求項20] コンピュータを、

ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得するウェブページ取得手段、

初期データとして、前記ウェブページに記載されている対象の属性に関連した属性関連語、または、当該対象の属性の記述に用いられる属性記述パターンを取得する初期データ取得手段、

前記複数のウェブページから、前記属性記述パターンに適合する前記属性の属性関連語を抽出する属性抽出手段、および、

前記複数のウェブページから、前記属性関連語に適合する前記属性記述パターンを抽出する属性記述パターン抽出手段として機能させることを特徴とする情報処理装置用のプログラム。

[請求項21] コンピュータを、

ウェブページに記載されている対象を分類するカテゴリが同一である複数のウェブページを取得するウェブページ取得手段、

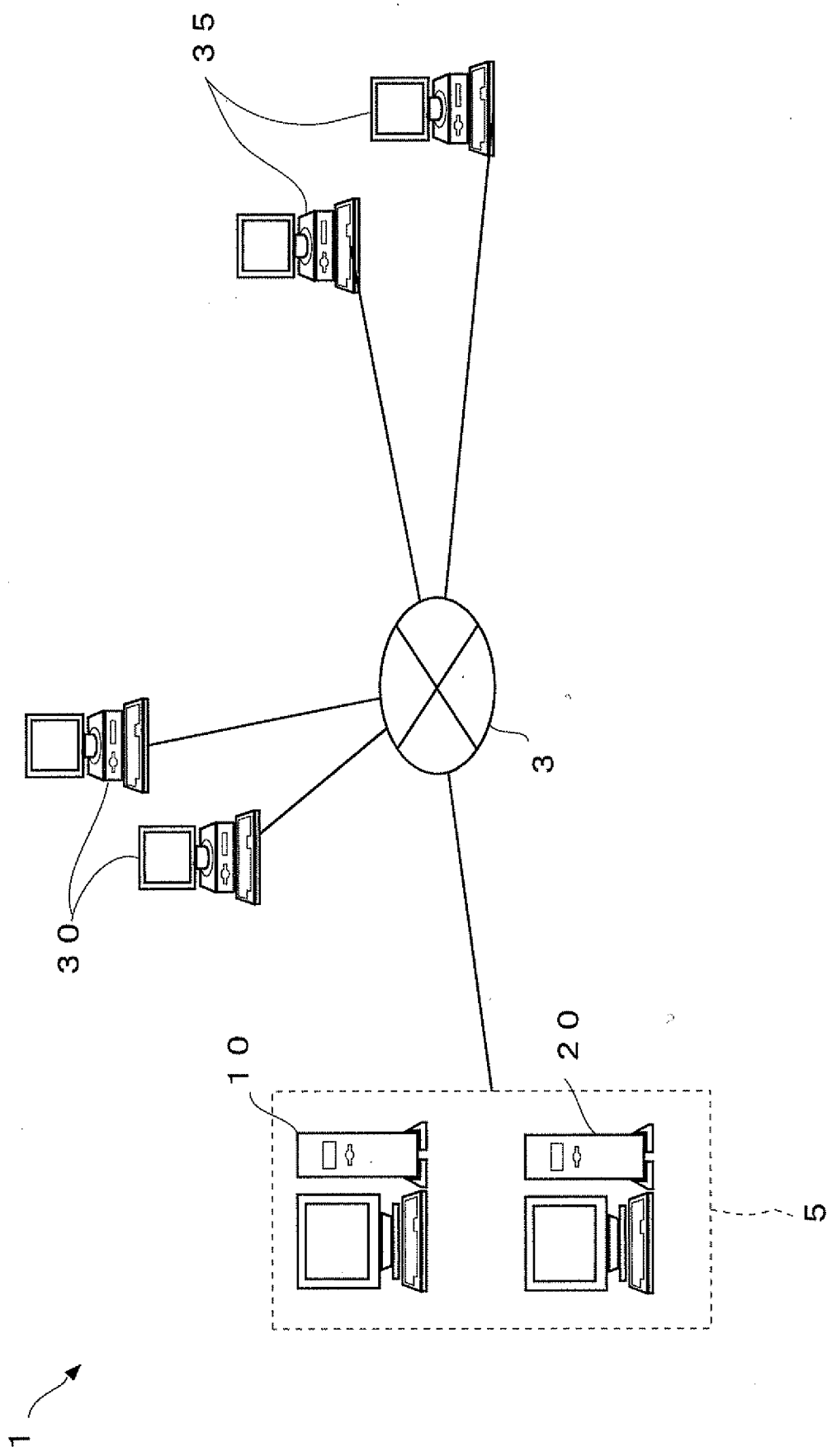
初期データとして、前記ウェブページに記載されている対象の属性に関連した属性関連語、または、当該対象の属性の記述に用いられる属性記述パターンを取得する初期データ取得手段、

前記複数のウェブページから、前記属性記述パターンに適合する前記属性の属性関連語を抽出する属性抽出手段、および、

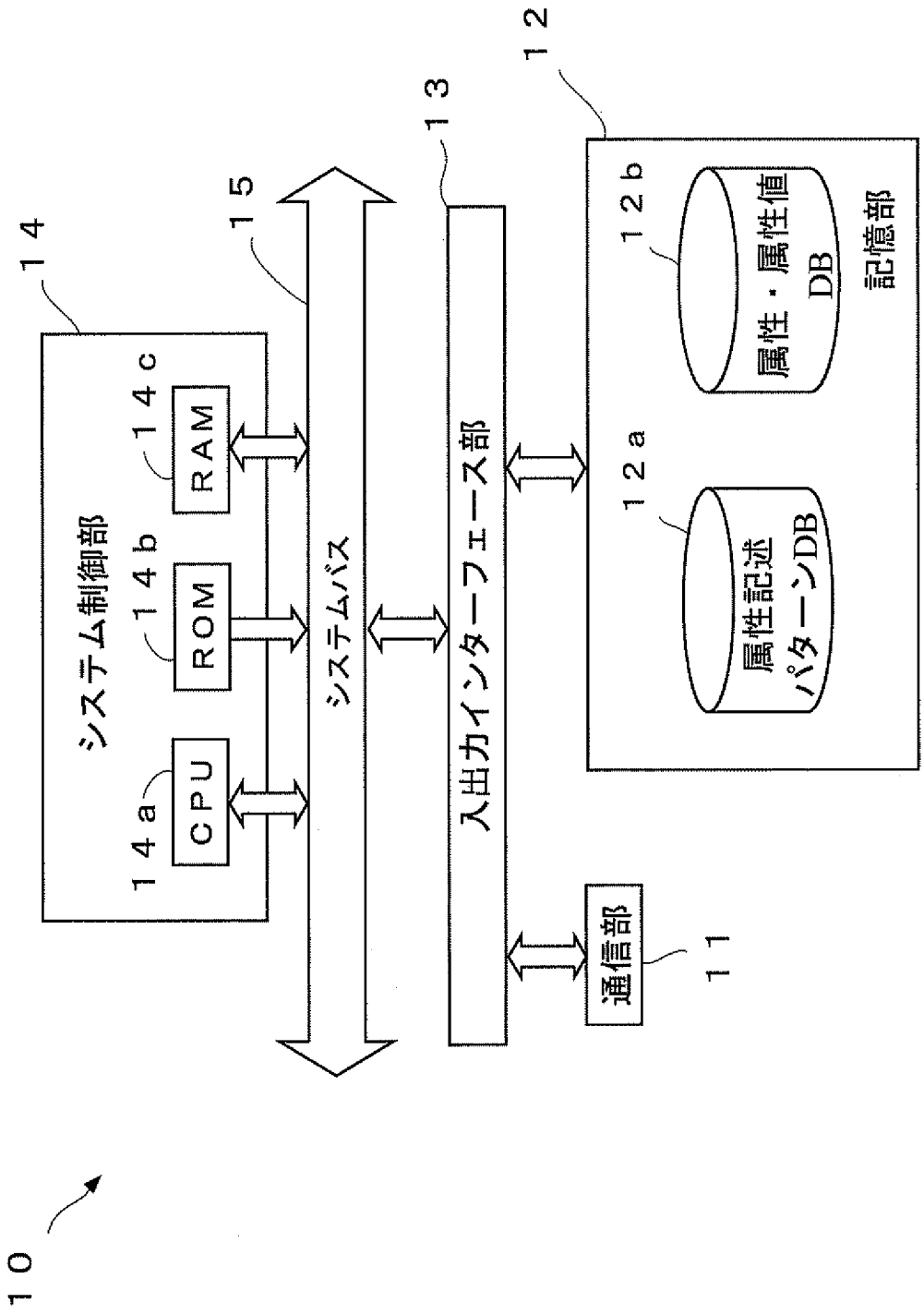
前記複数のウェブページから、前記属性関連語に適合する前記属性記述パターンを抽出する属性記述パターン抽出手段として機能させることを特徴とする情報処理装置用のプログラムを記録したコンピュー

タ読み取り可能な記録媒体。

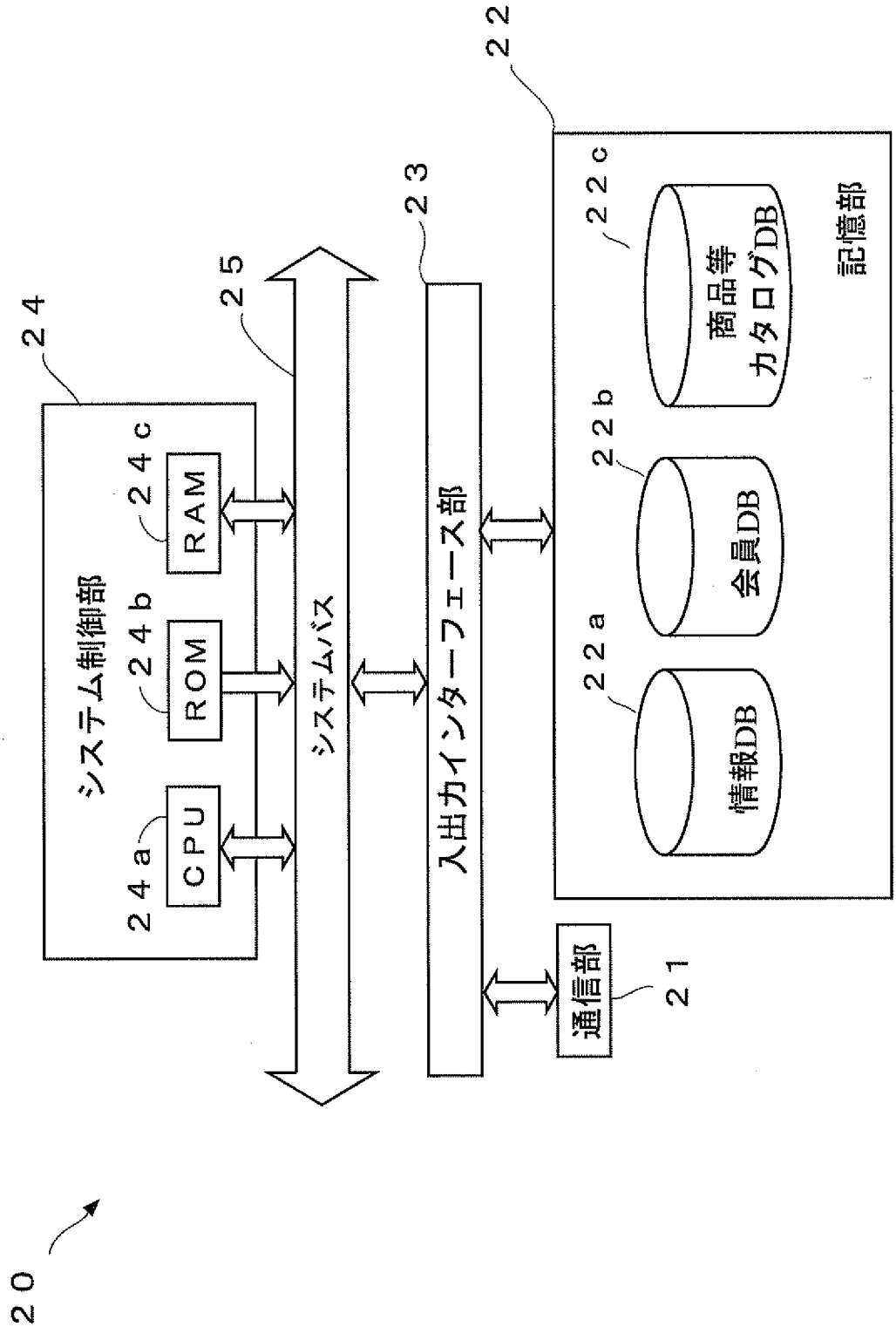
[図1]



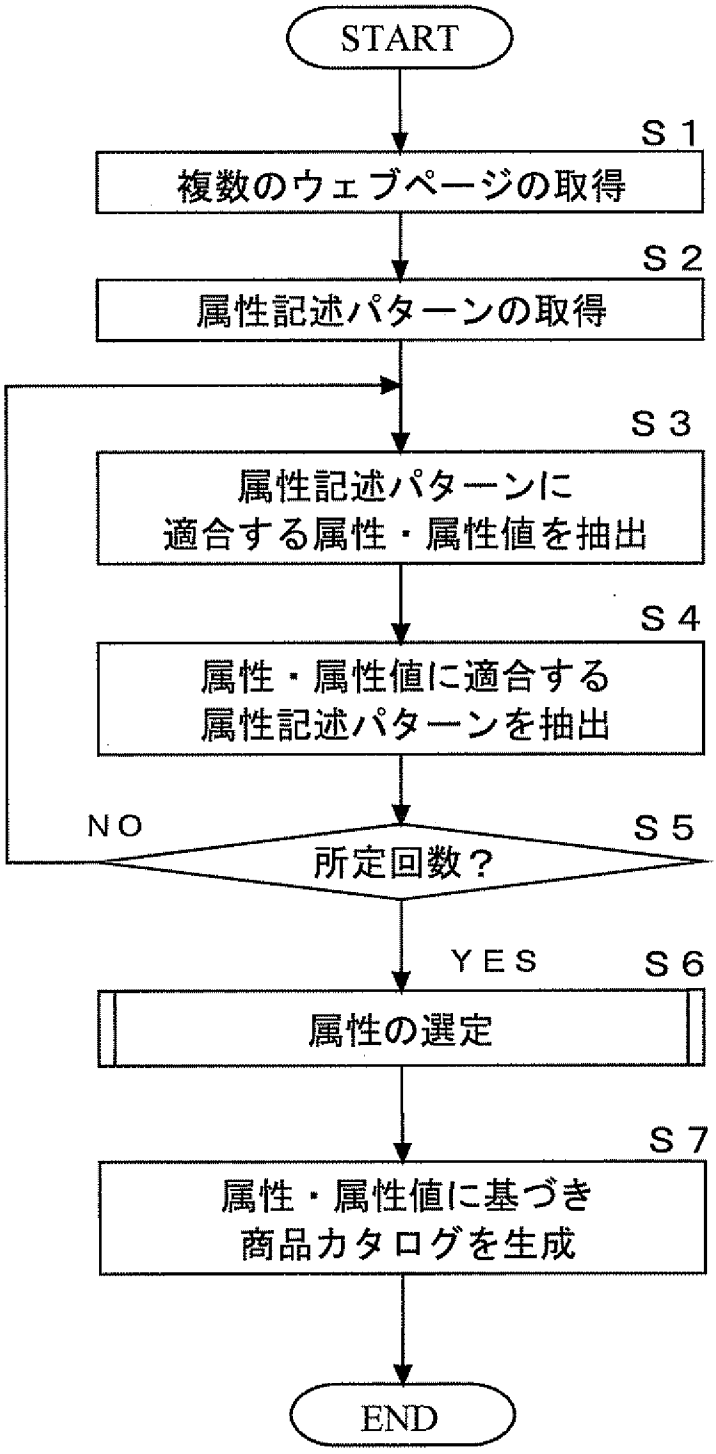
[図2]


10
14
システム制御部
14a
CPU
14b
ROM
14c
RAM
15
システムバス
13
入出力インターフェース部
12
記憶部
12a
属性記述
パターンDB
12b
属性・属性値
DB
11
通信部

[図3]


20
システム制御部
24
CPU
24a
ROM
24b
RAM
24c
システムバス
25
入出力インターフェース部
23
記憶部
22
情報DB
22a
会員DB
22b
商品等カタログDB
22c
通信部
21

[図4]


START
S１
複数のウェブページの取得
S２
属性記述パターンの取得
S３
属性記述パターンに
適合する属性・属性値を抽出
S４
属性・属性値に適合する
属性記述パターンを抽出
NO
S５
所定回数？
YES
S６
属性の選定
S７
属性・属性値に基づき
商品カタログを生成
END

[図5]


５０
SHOP
５１
ツェ〇・〇〇・△△
[送料無料]
当店通常価格 〇△円（税込）
＜商品詳細＞
[品種：〇〇〇]
[生産者：・・・・・・・]
[アルコール度数：１２．５%]
[容量：７５０ｍｌ]
・・・・・・・・
・・・・・・・
５２
商品番号 0000000
５３
・・・・・・・・簡単なラッピングで素敵なプレゼントになります
！ツェ〇・〇〇・△△・ＱＢＡ・・・・・・・・・・・
・・・・・・・・・・・・・白ワイン・・・・・・・・・・・・・
・・・・・
５４
〇〇
◇◇
・・・
・・・

[図6]

```
<html>
<head>
・・・・・・・・・・・・・・・・・・・・・・・・・・・・
<title>［〇〇市場］・・・。ツェ〇・〇〇・△△・・
・・・・・
</title>
・・・・・・・・・・・・・・・・・・・・・・・・・・・・
</head>
<body>
・・・・・
・・・・・ツェ〇・〇〇・△△・・・・・・
・・・・・・・・・・・・・・・・・
・・・・・・・［送料無料］・・・・・・・・・
・・・・・・・・・・・・・・［品種：〇〇〇］・・・・
・・・・・・・・・・・・・
・・・・・・・・・・・・・・・・・・・・・・・・・・・・
・・・・・・・・・・・・・・・・
<table・・・・・>
・・・・・
  <td>〇〇</td><td>◇◇</td>
・・・・・
</table>
・・・・・・・・
</body>
</html>
```

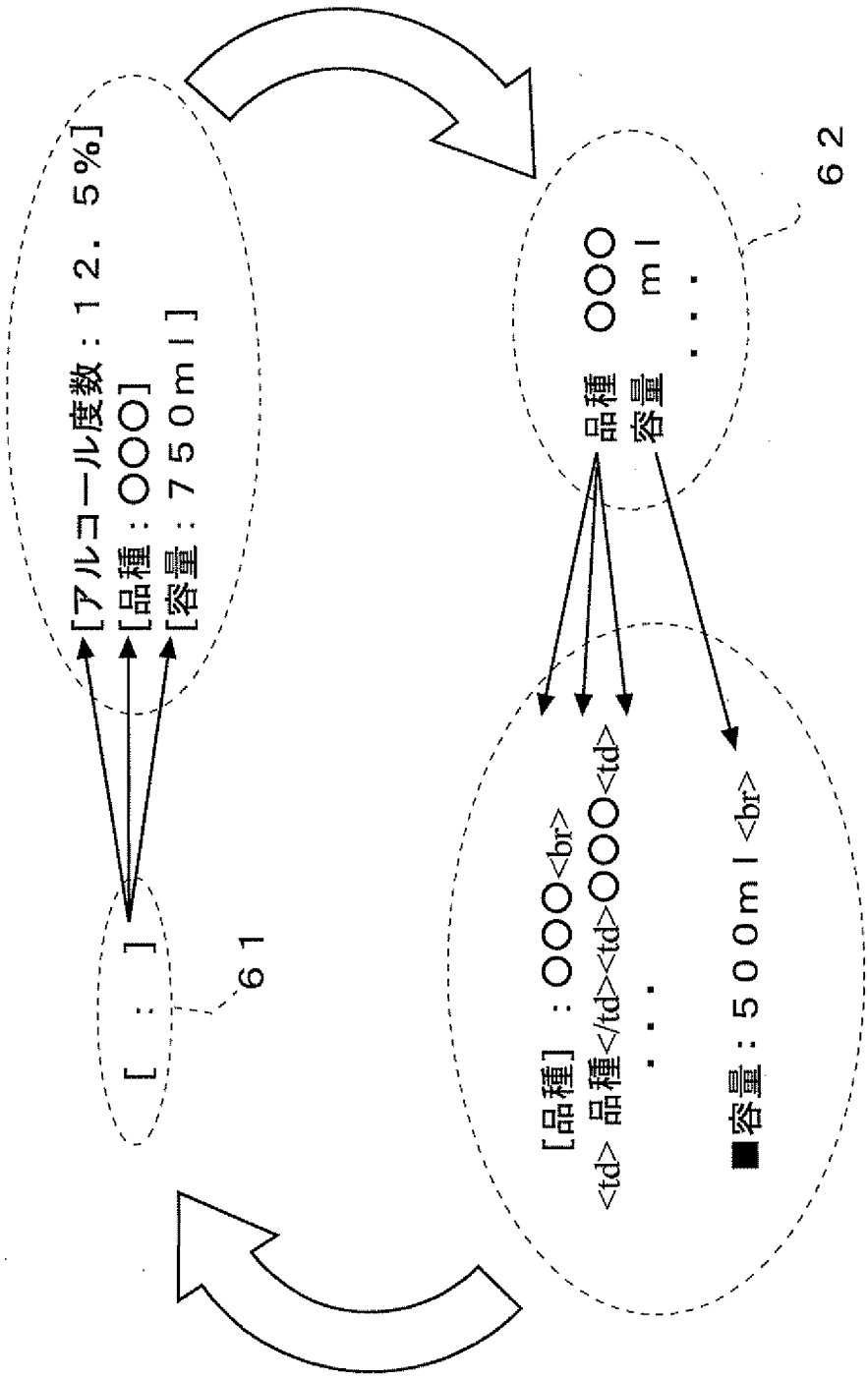
[図7]

前	中	後ろ
[	:	]
	:	

初期パターン

■	:	
<td>	</td><td>	</td>

[図8]


[アルコール度数：１２．５%]
[品種：○○○]
[容量：７５０ｍｌ]
[　：　]
６１
品種　○○○
容量　ｍｌ
・・・
６２
[品種]：○○○
<td>品種</td><td>○○○<td>
・・・
■容量：５００ｍｌ

[図9]

ワイン

属性名	属性値
品種	○○○、○▼
生産者	△△、・・・
アルコール度数	１２．５％
：	：

ゴルフドライバー

属性名	属性値
重量	５０ｇ、・・・
ロフト角	１０．５°、・・・
シャフト	・・・
：	：

[図10]

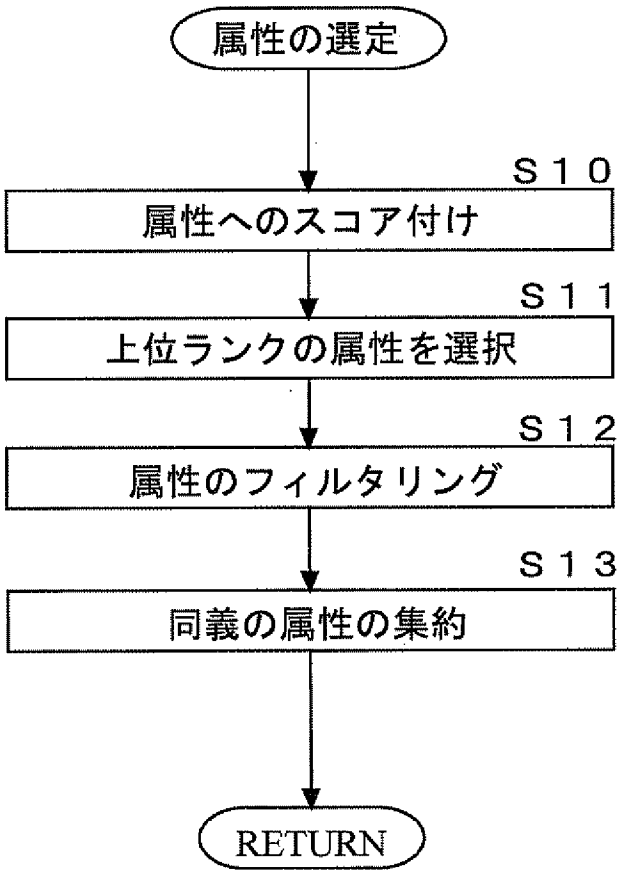
●●ワイン　商品番号１２３４５

品種	○○
生産者	△△
容量	７５０ｍｌ
製造年	２０００年
価格	￥９１２

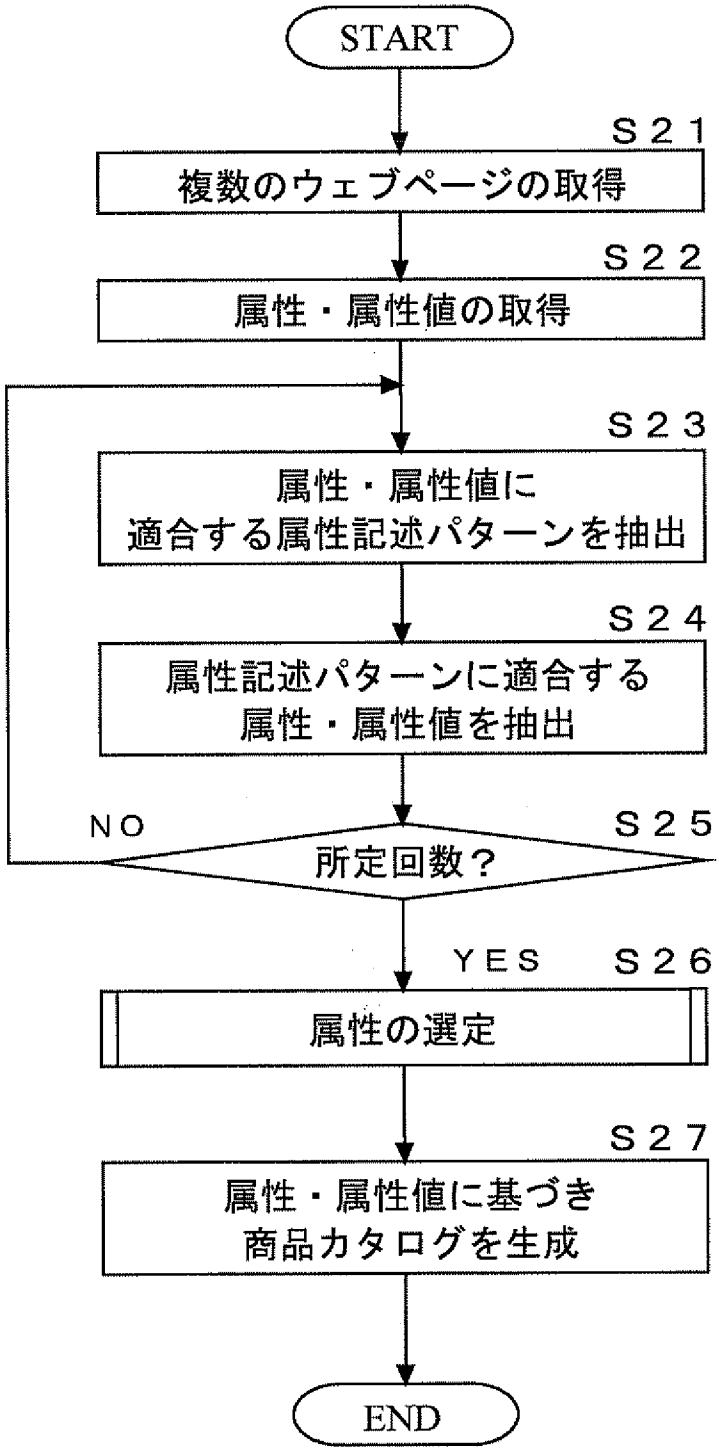
▽▲ワイン　商品番号１２３４６

品種	○●
生産者	▲△
容量	７５０ｍｌ
製造年	１９９５年
価格	￥２，９００

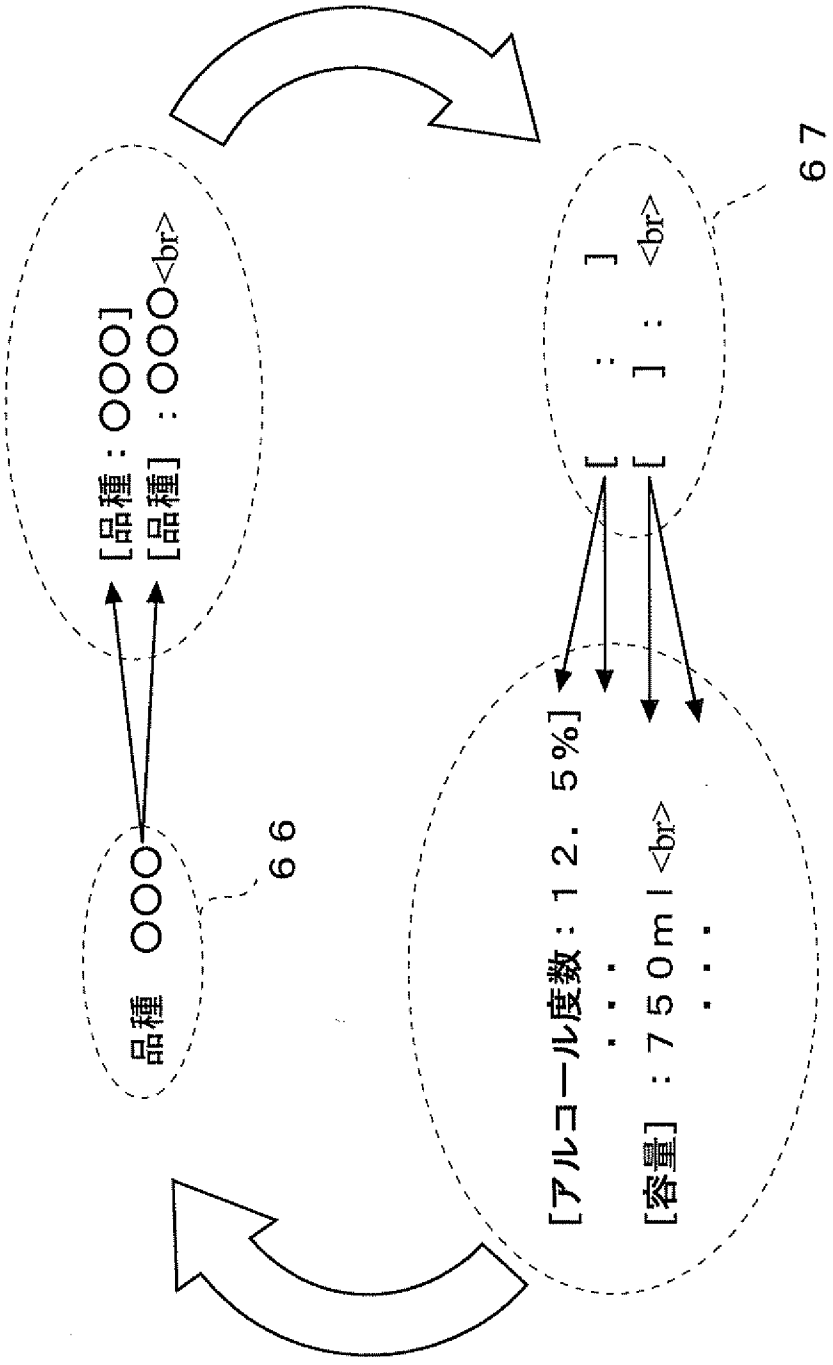
[図11]


属性の選定
S10
属性へのスコア付け
S11
上位ランクの属性を選択
S12
属性のフィルタリング
S13
同義の属性の集約
RETURN

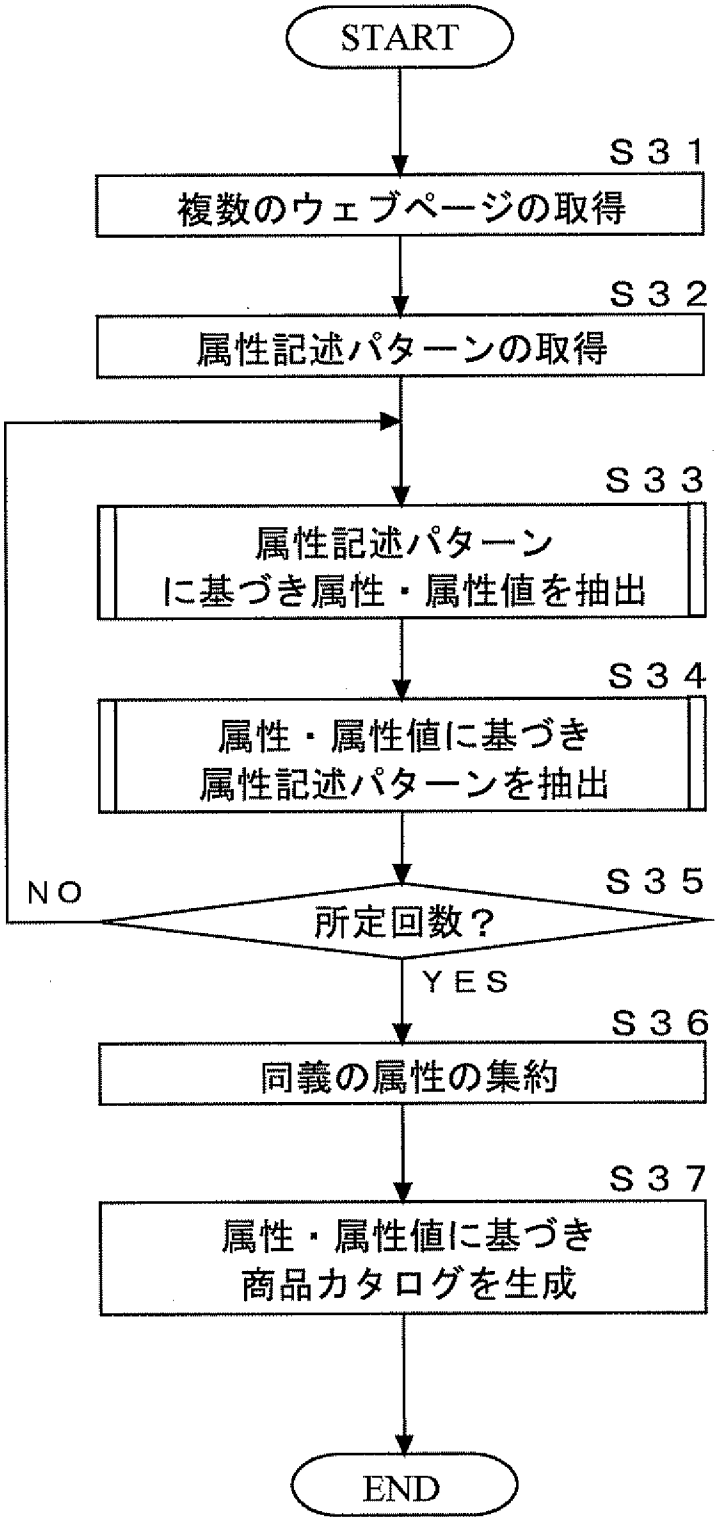
[図12]


START
S21
複数のウェブページの取得
S22
属性・属性値の取得
S23
属性・属性値に
適合する属性記述パターンを抽出
S24
属性記述パターンに適合する
属性・属性値を抽出
NO
S25
所定回数？
YES
S26
属性の選定
S27
属性・属性値に基づき
商品カタログを生成
END

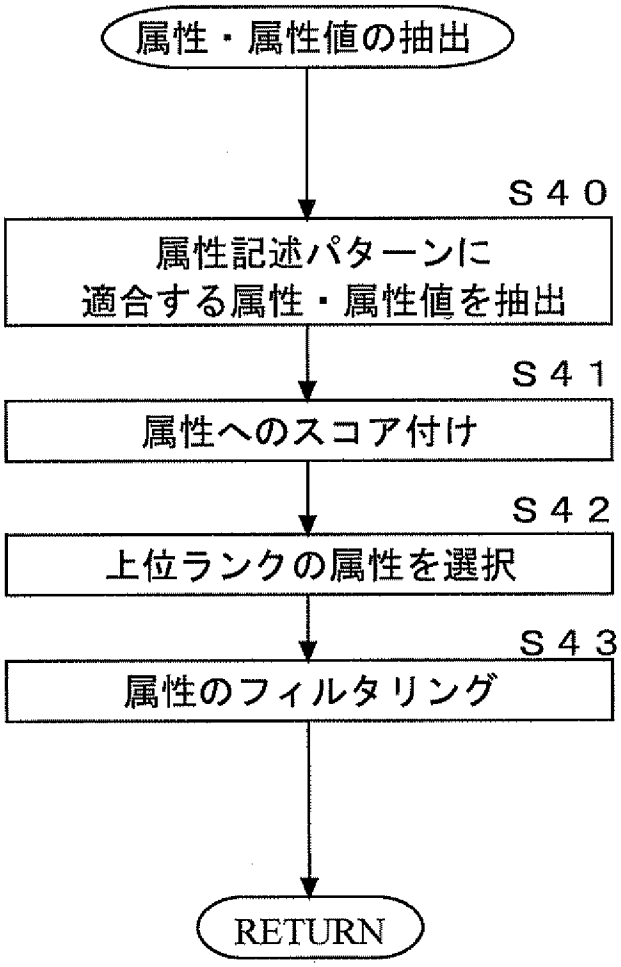
[図13]



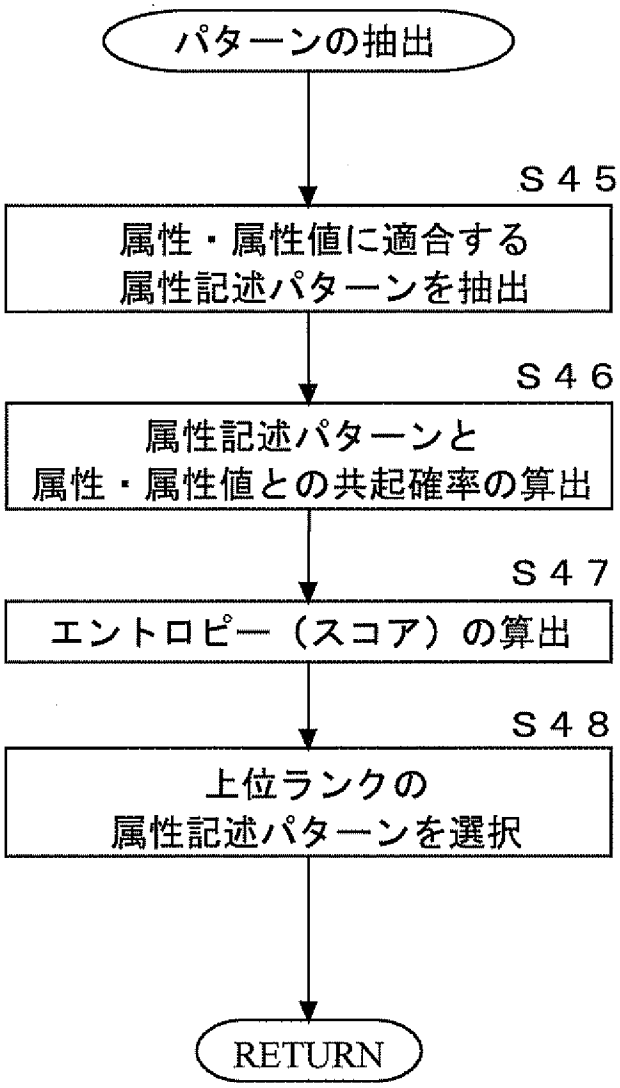
[図14]


START
S31
複数のウェブページの取得
S32
属性記述パターンの取得
S33
属性記述パターン
に基づき属性・属性値を抽出
S34
属性・属性値に基づき
属性記述パターンを抽出
S35
所定回数？
NO
YES
S36
同義の属性の集約
S37
属性・属性値に基づき
商品カタログを生成
END

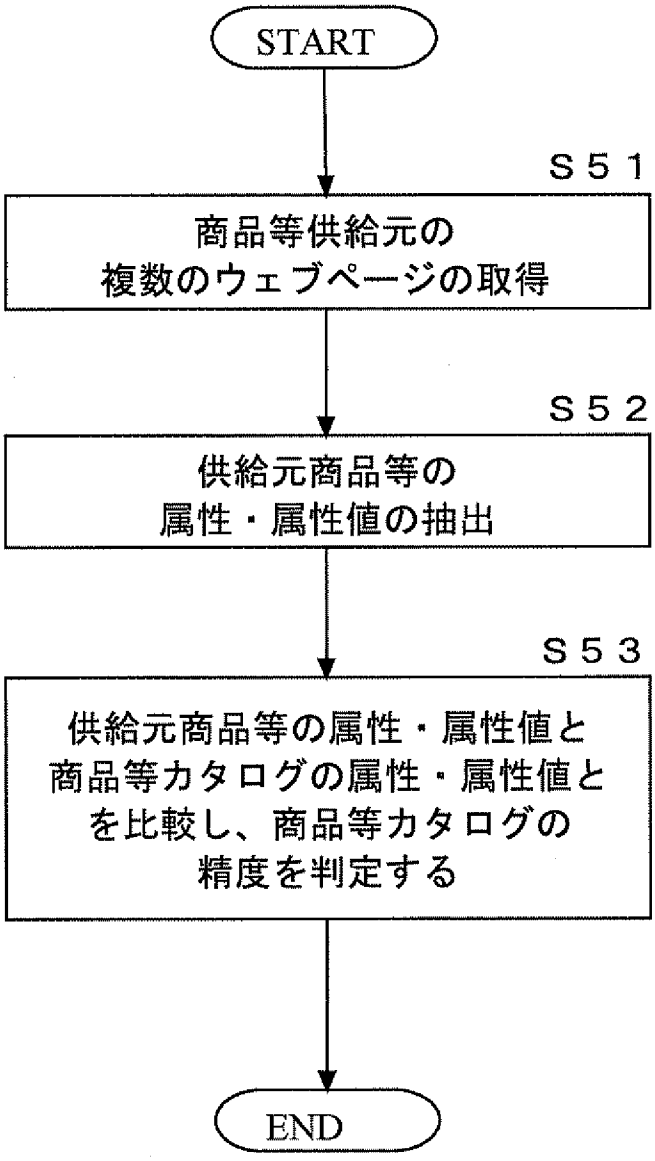
[図15]


属性・属性値の抽出
S40
属性記述パターンに
適合する属性・属性値を抽出
S41
属性へのスコア付け
S42
上位ランクの属性を選択
S43
属性のフィルタリング
RETURN

[図16]


パターンの抽出
S45
属性・属性値に適合する
属性記述パターンを抽出
S46
属性記述パターンと
属性・属性値との共起確率の算出
S47
エントロピー（スコア）の算出
S48
上位ランクの
属性記述パターンを選択
RETURN

[図17]


START
S５１
商品等供給元の
複数のウェブページの取得
S５２
供給元商品等の
属性・属性値の抽出
S５３
供給元商品等の属性・属性値と
商品等カタログの属性・属性値と
を比較し、商品等カタログの
精度を判定する
END

[図18]


◆◆◆社
メルマガ
検索
トップページ＞ワイン＞ドイツ
ツェ○・○○・△△
カタログ
No.123456788
容量 750ｍｌ
容器 びん
果実名 ：
収穫地 ：
醸造地 ：
味のタイプ：

[図19]

製造年別カタログ

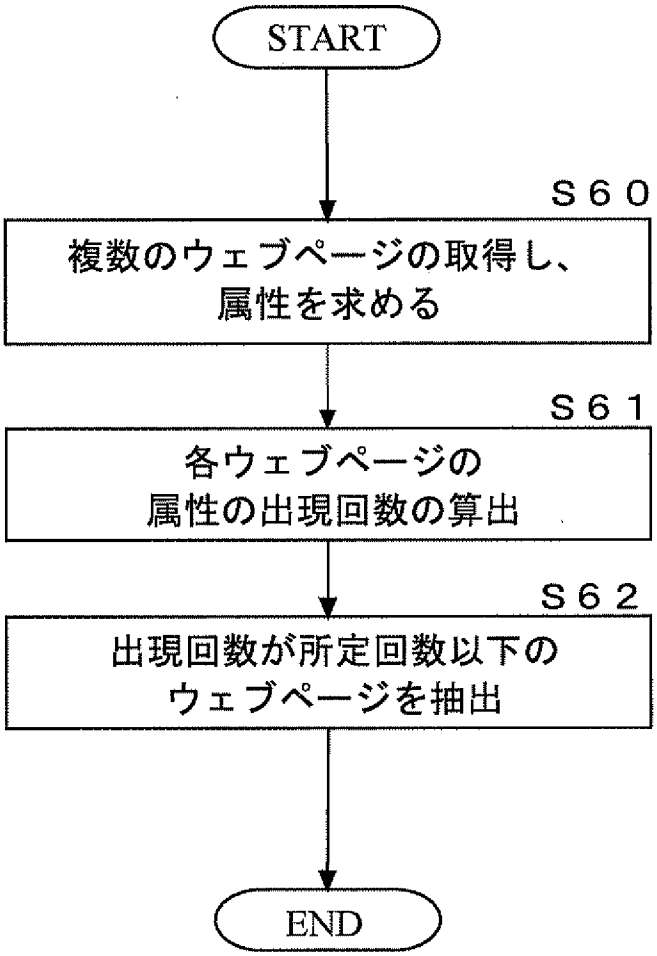
◆ 製造年：１９９５年

ワイン名	品種	生産者	価格	商品番号
●●ワイン	○○	△△	￥１，５００	１０４６１
▽▲ワイン	○●	▲△	￥２，９００	１２３４６
⋮		⋮		⋮
■▽ワイン	□○	○△	￥３，５００	１３３５４

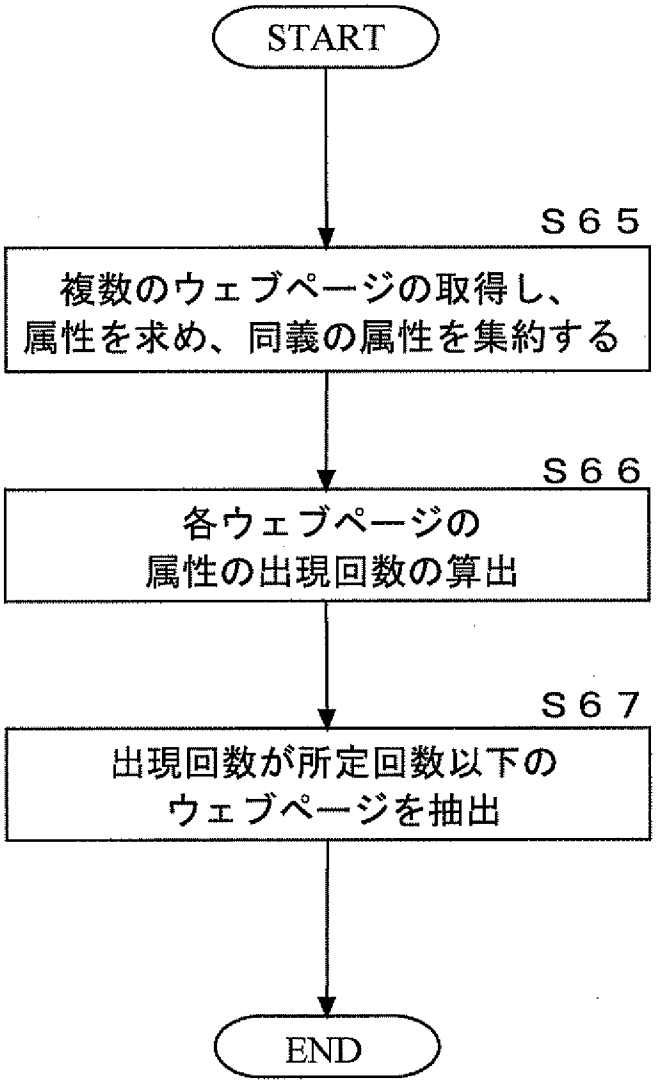
◆ 製造年：１９９６年

ワイン名	品種	生産者	価格	商品番号
●●ワイン	○○	△△	￥１，３００	１３４７８
▽▲ワイン	○●	▲△	￥２，５００	１５３５６
⋮				⋮

[図20]


START
S60
複数のウェブページの取得し、
属性を求める
S61
各ウェブページの
属性の出現回数の算出
S62
出現回数が所定回数以下の
ウェブページを抽出
END

[図21]


START
S65
複数のウェブページの取得し、
属性を求め、同義の属性を集約する
S66
各ウェブページの
属性の出現回数の算出
S67
出現回数が所定回数以下の
ウェブページを抽出
END

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2011/054510

A. CLASSIFICATION OF SUBJECT MATTER
G06F17/30(2006.01)i, *G06Q30/00*(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F17/30, G06Q30/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2011
Kokai Jitsuyo Shinan Koho	1971-2011	Toroku Jitsuyo Shinan Koho	1994-2011

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 2008-269106 A (Osaka Industrial Promotion Organization), 06 November 2008 (06.11.2008), entire text; all drawings (Family: none)	1-21
P,X	Satoshi SEKINE, "Shopping Site ni Okeru Shohin no Doitsusei, Ruijisei no Suitei Shuho", Proceedings of the 16th annual meeting of the Association for Natural Language Processing, 08 March 2010 (08.03.2010), pages 254 to 257	1-21

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 March, 2011 (15.03.11)	29 March, 2011 (29.03.11)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2011/054510

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P,X	Hiroki SAKACHI, "Shohin Page Karano Zokusei· Zokuseichi Chushutsu to Doitsu Shohin Clustering Shuho", Proceedings of the 16th annual meeting of the Association for Natural Language Processing, 08 March 2010 (08.03.2010), pages 371 to 374	1-21

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１１／０５４５１０

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30(2006.01)i, G06Q30/00(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30, G06Q30/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１１年
日本国実用新案登録公報	１９９６－２０１１年
日本国登録実用新案公報	１９９４－２０１１年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
X	JP 2008-269106 A (財団法人大阪産業振興機構) 2008.11.06, 全文, 全図 (ファミリーなし)	1-21
P, X	関根聡, ショッピングサイトにおける商品の同一性、類似性の推定手法, 言語処理学会第１６回年次大会発表論文集, 2010.03.08, ｐｐ．２５４－２５７	1-21

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの

「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「Ｏ」口頭による開示、使用、展示等に言及する文献

「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
１５．０３．２０１１	２９．０３．２０１１

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｍ　４４４０
松田　直也
電話番号　０３－３５８１－１１０１　内線　３５９９

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
P, X	坂地泰紀，商品ページからの属性・属性値抽出と同一商品クラスタリング手法，言語処理学会第１６回年次大会発表論文集，2010.03.08，ｐｐ．３７１－３７４	1-21