(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 2011/105604 A1

(51) 国際特許分類:
G06F 17/30 (2006.01)

(21) 国際出願番号: PCT/JP2011/054507

(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 2010-043389 2010 年 2 月 26 日(26.02.2010) JP
特願 2010-043388 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).

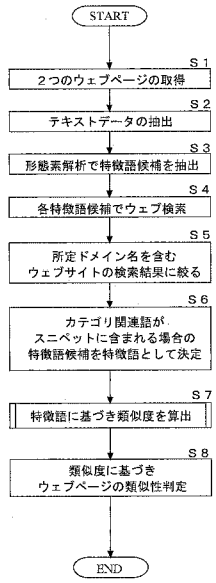
(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, 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.

[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5]



S1 Acquisition of two web pages
S2 Extraction of text data
S3 Extraction of feature term candidates using morphological analysis
S4 Web search using each feature term candidate
S5 Narrow results to search results from web sites containing a predetermined domain name
S6 Set feature term candidate as feature term when category-related term is contained in snippet
S7 Calculate degree of similarity on the basis of feature term
S8 Determine similarity between web pages on the basis of degree of similarity

(57) **Abstract**: Disclosed is an information processing device which: stores category-associated terms which are related to categories which classify objects listed in a web page, and obtains web pages (S1); extracts text data written in the web pages (S2); carries out linguist analysis on the extracted text data and extracts feature term candidates (S3); generates search queries using the feature term candidates as search keywords, and obtains search results on the basis of the search queries (S4); and determines whether or not the category-associated terms exist within snippets of the search results, and sets the characteristic word candidates for which the category-associated words do exist within the snippets as feature terms for the web page (S6).

(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に関連するカテゴリ関連語を記憶し、ウェブページを取得し（Ｓ１）、ウェブページ内に記述されているテキストデータを抽出し（Ｓ２）、抽出したテキストデータを言語解析して、ウェブページの特徴語候補を抽出し（Ｓ３）、特徴語候補を検索キーワードとする検索クエリを作成し、検索クエリに基づく検索結果を取得し（Ｓ４）、検索結果のスニペット中にカテゴリ関連語が存在するか否かを判定し、スニペット中にカテゴリ関連語が存在する場合の特徴語候補をウェブページの特徴語に決定する（Ｓ６）。

WO 2011/105604 A1

(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 (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] インターネット上には商品等を販売するためのウェブページや、個人の意見などが書き込まれたブログ（Blog）等のウェブページが存在し、さらに、取り扱われる商品等の内容やブログの内容により、様々なタイプのウェブページが存在する。このような多様なウェブページを分類したり、ウェブページの内容に適合した広告を付加したり等、ウェブページに対して様々な処理がなされている。例えば、特許文献１には、ウェブページとコンテンツカテゴリとの関連度に基づいて、ウェブページをコンテンツカテゴリに分類し、分類されたコンテンツカテゴリと広告カテゴリとをマッチングして、広告データを選択して当該ウェブページに付加するカテゴリ分類方法が開示されている。

先行技術文献

特許文献

[0003] 特許文献1：特開２００９－２６６２０４号公報

発明の概要

発明が解決しようとする課題

[0004] しかしながら、特許文献１のような従来技術では、ウェブページのテキストデータを形態素解析し、ＴＦ（Term Frequency）・ＩＤＦ（Inverse Docum

ent Frequency）値等の出現頻度に基づいて、ウェブページに含まれる特徴語を抽出して、当該特徴語と類似するキーワードに基づいて、ウェブページとコンテンツカテゴリとの関連度を算出している。このような特徴語の抽出では、流行の変化が激しいインターネット上において、精度の高い特徴語の求めることが難しかった。

[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、ウェブページの特徴語を、精度良く求めることができ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に関連するカテゴリ関連語を記憶するカテゴリ関連語記憶手段と、ウェブページを取得するウェブページ取得手段と、前記ウェブページ内に記述されているテキストデータを抽出するテキストデータ抽出手段と、抽出した前記テキストデータを言語解析して、前記ウェブページの特徴語候補を抽出する特徴語候補抽出手段と、前記特徴語候補を検索キーワードとする検索クエリを作成し、前記検索クエリに基づく検索結果を取得する検索結果取得手段と、前記カテゴリ関連語記憶手段を参照して、前記検索結果のスニペット中に前記カテゴリ関連語が存在するか否かを判定する検索結果判定手段と、前記スニペット中に前記カテゴリ関連語が存在する場合の前記特徴語候補を前記ウェブページの特徴語に決定する特徴語決定手段と、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、ノイズの形態素を記憶するノイズ形態素記憶手段を更に備え、前記特徴語候補抽出手段が、ノイズ形態素記憶手段を参照して、前記テキストデータを形態素解析によって前記言語解析して得られた形態素から、ノイズの形態素を除去するノイズ除去手段を有することを特徴とする。

[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、ウェブサイトのドメイン名を記憶するドメイン名記憶手段を更に

備え、前記検索結果判定手段が、前記カテゴリ関連語記憶手段および前記ドメイン名記憶手段を参照して、前記検索結果において、前記ドメイン名に関連したウェブサイトのスニペット中に前記カテゴリ関連語が存在するか否かを判定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、ユーザ端末からユーザ検索クエリを受信する検索クエリ受信手段と、前記ユーザ検索クエリによって得られる複数のウェブページを、前記特徴語毎にグルーピングするウェブページ・グルーピング手段と、前記特徴語毎にグルーピングした前記ウェブページに関する情報を検索結果として前記ユーザ端末に送信する送信手段を更に、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４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、前記ウェブページとして、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、を更に備え、前記ウェブページ・グルーピング手段が、前記特徴語および前記属性関連語に基づき、前記ユーザ検索クエリによって得られる複数のウェブページをグルーピングすることを特徴とする。

[0011] 請求項６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、前記ウェブページ取得手段が、前記ウェブページとして、複数のウェブページを取得し、初期データとして、前記対象の属性に関連した属性関連語、または、前記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、前記特徴語と、前記任意の２つのウェブページ間での共通の属性関連語とに基づき、前記任意の２つのウェブページが扱う対象は、互いに類似する対象であるか否かを判定する類似対象判定手段と、を更に備えたことを特徴とする。

[0012] 請求項７に記載の発明は、請求項５または請求項６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。

[0013] 請求項８に記載の発明は、請求項６または請求項７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、前記類似対象判定手段が、前記任意の２つのウェブページ間で共通の属性に係る前記属性値において等しい前記属性値がある場合に、前記任意の２つのウェブページに記載された対象は、互いに類似する対象であると判定することを特徴とする。

[0014] 請求項９に記載の発明は、請求項５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、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。

[0015] 請求項１０に記載の発明は、請求項５から請求項９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。

[0016] 請求項１１に記載の発明は、請求項１０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。

[0017] 請求項１２に記載の発明は、請求項１０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。

[0018] 請求項１３に記載の発明は、請求項５から請求項１２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。

[0019] 請求項１４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に関連するカテゴリ関連語を記憶するカテゴリ関連語記憶ステップと、ウェブページを取得するウェブページ取得ステップと、前記ウェブページ内に記述されているテキストデータを抽出するテキストデータ抽出ステップと、抽出した前記テキストデータを言語解析して、前記ウェブページの特徴語候補を抽出する特徴語候補抽出ステップと、前記特徴語候補を検索キーワードとする検索クエリを作成し、前記検索クエリに基づく検索結果を取得する検索結果取得ステップと、前記検索結果のスニペット中に前記カテゴリ関連語が存在するか否かを判定する検索結果判定ステップと、前記スニペット中に前記カテゴリ関連語が存在する場合の前記特徴語候補を前記ウェブページの特徴語に決定する特徴語決定ステップと、を有することを特徴とする。

[0020] 請求項１５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に関連するカテゴリ関連語を記憶するカテゴリ関連語記憶手段、ウェブページを取得するウェブページ取得手段、前記ウェブページ内に記述されているテキストデータを抽出するテキストデータ抽出手段、抽出した前記テキストデータを言語解析して、前記ウェブページの特徴語候補を抽出する特徴語候補抽出手段、前記特徴語候補を検索キーワードとする検索クエリを作成し、前記検索クエリに基づく検索結果を取得する検索結果取得手段、前記カテゴリ関連語記憶手段を参照して、前記検索結果のスニペット中に前記カテゴリ関連語が存在するか否かを判定する検索結果判定手段、および、前記スニペット中に前記カテゴリ関連語が存在する場合の

前記特徴語候補を前記ウェブページの特徴語に決定する特徴語決定手段として機能させることを特徴とする。

[0021] 請求項１６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に関連するカテゴリ関連語を記憶するカテゴリ関連語記憶手段、ウェブページを取得するウェブページ取得手段、前記ウェブページ内に記述されているテキストデータを抽出するテキストデータ抽出手段、抽出した前記テキストデータを言語解析して、前記ウェブページの特徴語候補を抽出する特徴語候補抽出手段、前記特徴語候補を検索キーワードとする検索クエリを作成し、前記検索クエリに基づく検索結果を取得する検索結果取得手段、前記カテゴリ関連語記憶手段を参照して、前記検索結果のスニペット中に前記カテゴリ関連語が存在するか否かを判定する検索結果判定手段、および、前記スニペット中に前記カテゴリ関連語が存在する場合の前記特徴語候補を前記ウェブページの特徴語に決定する特徴語決定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

[0022] 本発明によれば、ウェブページ内に記述されているテキストデータを抽出し、テキストデータを言語解析して、ウェブページの特徴語候補を抽出し、特徴語候補を検索キーワードとする検索クエリを作成し、検索クエリに基づく検索結果を取得し、検索結果のスニペット中にカテゴリ関連語が存在するか否かを判定し、スニペット中にカテゴリ関連語が存在する場合の特徴語候補をウェブページの特徴語に決定することにより、言語解析してウェブページの特徴語候補を抽出し、さらに、検索クエリに基づき特徴語を決定しているため、ウェブページの特徴語を、精度良く求めることができる。

図面の簡単な説明

[0023] [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のウェブページ分析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の情報提供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の検索サーバの概要構成の一例を示すブロック図である。

[図5]図１のウェブページ分析サーバにおいてウェブページを分析する動作例を示すフローチャートである。

[図6]図１の情報提供サーバのウェブページの一例を示す説明図である。

[図7]図６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8]図６のウェブページのテキストデータの分析過程の一例を示す模式図である。

[図9]図１の検索サーバの検索結果の一例を示す模式図である。

[図10]図２のドメイン名データベースに記憶されたウェブサイトのドメイン名の一例を示す模式図である。

[図11]図２のカテゴリ関連語データベースに記憶されたカテゴリ関連語の一例を示す模式図である。

[図12]図１のウェブページ分析サーバにおける類似度算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13]本発明の第２実施形態に係る情報処理システムの検索システムとしての動作においてウェブページをグルーピングする動作例を示すフローチャートである。

[図14]図１のユーザ端末に送信されるグルーピングされた検索結果の一例を示す模式図である。

[図15]図１のウェブページ分析サーバにおけるグルーピング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]本発明の第３実施形態に係る情報処理システムのウェブページ分析システムとしての動作においてウェブページ分析の全体の流れを示す説明図である。

[図17]図１６のウェブページ分析においてウェブページの特徴語を抽出してウェブページを分析する動作例を示すフローチャートである。

[図18]図１６のウェブページ分析においてウェブページから対象の属性・属

性値を抽出する動作例を示すフローチャートである。

[図19]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20]属性・属性値の抽出の様子の一例を示す模式図である。

[図21]抽出された属性・属性値の一例を示す模式図である。

[図22]図１８の属性・属性値を抽出する動作例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23]図１４のグルーピングされた検索結果の変形例を示す模式図である。

[図24]図１４のグルーピングされた検索結果の他の変形例を示す模式図である。

発明を実施するための形態

[0024] 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）

[0025] ［１．情報処理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。

[0026] 図１は、本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。

[0027] 図１に示すように、情報処理システム１は、ウェブページを分析するウェブページ分析サーバ（情報処理装置の一例）１０と、ショッピングサイトやブログを運営するための情報提供サーバ２０と、インターネット上のウェブページのウェブ検索の機能を提供する検索サーバ３０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４０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４５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。

[0028] ウェブページ分析サーバ１０と、情報提供サーバ２０と、検索サーバ３０とは、ローカルエリアネットワークやインターネット等により接続され、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４０と、ユーザ端末４５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。

[0029] 情報処理システム１は、ウェブページを分析するウェブページ分析システムとして、または、ユーザ端末４５から検索クエリに対して応答する検索システムとして機能する。

[0030] ウェブページ分析サーバ１０は、情報提供サーバ２０等に登録されたウェブページを分析して、ウェブページから特徴語を抽出したり、ウェブページ同士の類似度を算出したりする。また、ウェブページ分析サーバ１０は、分析結果に基づいて、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４０や、ユーザ端末４５等に分析結果を通知したりする。さらに、ウェブページ分析サーバ１０は、ユーザ端末４５から検索クエリを受信したり、分析結果に基づき、同じような商品に関するウェブページをグルーピングしたり、ユーザ端末４５に検索結果を送信したりする。

[0031] 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。

[0032] 検索サーバ３０は、インターネット等上の検索エンジンとして機能し、検索ＡＰＩ（Application Programming Interface）の機能を、ウェブページ分

析サーバ１０に提供したり、端末から検索キーワードを含む受信した検索クエリに基づき、ウェブページ等の検索対象を検索して、検索結果を返したりする。また、検索サーバ３０は、検索用のデータベースを構築するために、情報提供サーバ２０や、図示しないインターネット上のウェブサイトからウェブページの情報を収集し、ウェブページの情報を予め解析し、検索のためのインデックス情報や検索結果のスニペットを作成する。ここで、スニペットとは、各検索結果の説明文等のテキストであって、検索でヒットしたウェブページの要約や、そのウェブページ内に記述された検索キーワードを含んだ文章等である。

[0033] 店舗主が使用する店舗主端末４０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４０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。

[0034] ユーザが使用するユーザ端末４５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４５を使用して、商品の検索や購入等を行う。

[0035] ［２．各サーバの構成および機能］

（２．１　ウェブページ分析サーバ１０の構成および機能）

次に、ウェブページ分析サーバ１０の構成および機能について、図２を用いて説明する。

[0036] 図２は、ウェブページ分析サーバ１０の概要構成の一例を示すブロック図である。

[0037] 図２に示すように、コンピュータとして機能するウェブページ分析サーバ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。

[0038] 通信部１１は、ネットワーク３に接続してユーザ端末４５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や検索サーバ３０等の他のサーバとデータの送受信を行ったりする。

[0039] 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。

[0040] また、記憶部１２には、カテゴリ関連語記憶手段の一例としてのカテゴリ関連語データベース１２ａ（以下「カテゴリ関連語ＤＢ」とする。）や、ドメイン名記憶手段の一例としてのドメイン名データベース（以下「ドメイン名ＤＢ」とする。）１２ｂや、特徴語データベース（以下「特徴語ＤＢ」とする。）１２ｃや、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。）１２ｄ、属性・属性値データベース（以下「属性・属性値ＤＢ」とする。）１２ｅ等が構築されている。

[0041] カテゴリ関連語ＤＢ１２ａには、ウェブサイトにおける商品等や、ブログの等のカテゴリ毎に用語が記憶されている。例えば、”ワイン”のカテゴリには、”ワイン”、”お酒”等のカテゴリ関連語が、”ゴルフドライバー”のカテゴリには、”ゴルフ”、”ドライバー”、”クラブ”等のカテゴリ関連語が予め記憶されている。また、”旅行”のカテゴリには、”国内”、”海外”等のカテゴリ関連用語が、予め記憶されている。また、カテゴリは階層構造になっていて、例えば、”ワイン”のカテゴリは、”洋酒”のカテゴリに属する。これらカテゴリ関連語は、各カテゴリに対応する識別子に関連付けられて記憶されている。このように、カテゴリ関連語ＤＢ１２ａは、商品を分類するカテゴリに関連するカテゴリ関連語を記憶するカテゴリ関連語記憶手段の一例として機能する。なお、カテゴリ関連語ＤＢ１２ａには、ブログためのウェブサイトの場合、食べ物に関するカテゴリには、”レストラン”、”ランチ”等のカテゴリ関連用語が、スポーツに関するカテゴリには、”野球”、”サッカー”等のカテゴリ関連用語が記憶されている。

[0042] ドメイン名ＤＢ１２ｂには、インターネット上の様々なショッピングサイトやブログサイト等のウェブサイトのドメイン名が予め記憶されている。

[0043] 特徴語ＤＢ１２ｃには、ウェブページ分析サーバ１０による分析結果の一例として、ウェブページから、抽出したウェブページの特徴語が記憶されている。特徴語は、ウェブページのＵＲＬ（Uniform Resource Locator）等に関連付けられて記憶されている。さらに、特徴語ＤＢ１２ｃには、ノイズの形態素を除去するための情報が記憶されている。

[0044] 属性記述パターンＤＢ１２ｄ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。

[0045] 属性・属性値ＤＢ１２ｅには、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。

[0046] 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。

[0047] 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、ウェブページのテキストデータから各ウェブページの特徴語候補を抽出したり、特徴語候補から特徴語を決定したり、特徴語に基づき、ウェブページ間の類似度を算出したり、複数のウェブページを特徴語毎にグルーピングしたりする。

する。

[0048] （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。

[0049] 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、ウェブページ分析サーバ１０の構成および機能とほぼ同じであるので、ウェブページ分析サーバ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。

[0050] 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４０やユーザ端末４５やウェブページ分析サーバ１０等と通信状態を制御等するようになっている。

[0051] 記憶部２２には、情報データベース（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等が構築されている。

[0052] 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ＤＢ２２ａ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。

[0053] 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４５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。

[0054] 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。

[0055] （２．３　検索サーバ３０の構成および機能）

次に、検索サーバ３０の構成および機能について、図４を用いて説明する。

図４は、検索サーバ３０の概要構成の一例を示すブロック図である。

[0056] 図４に示すように、検索サーバ３０は、通信部３１と、記憶部３２と、入出力インターフェース部３３と、システム制御部３４と、を備え、システム制御部３４と入出力インターフェース部３３とは、システムバス３５を介して接続されている。なお、検索サーバ３０の構成および機能は、ウェブページ分析サーバ１０の構成および機能とほぼ同じであるので、ウェブページ分析サーバ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。

[0057] 通信部３１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、ユーザ端末４５やウェブページ分析サーバ１０等と通信状態を制御等するようになっている。

[0058] 記憶部３２には、検索データベース（以下「検索ＤＢ」とする。）３２ａ等が構築されている。

[0059] 検索ＤＢ３２ａには、インデクサーにより生成されたインデックス情報や、スニペット等が記憶されている。

[0060] システム制御部３４は、ＣＰＵ３４ａ、ＲＯＭ３４ｂ、ＲＡＭ３４ｃ等により構成されている。そして、システム制御部３４は、ＣＰＵ３４ａが、ＲＯＭ３４ｂや記憶部３２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、インターネット上のウェブページの情報を収集し、ウェブページの情報を予め解析し、インデックス情報や検索結果のスニペットを作成したり、ウェブページ分析サーバ１０やユーザ端末４５からの検索クエリに基づき、検索をしたり、検索結果を送信したりする。

[0061] ［３．第１実施形態のウェブページ分析システムの動作］

次に、本発明の第１実施形態に係る情報処理システム１のウェブページ分析システムとしての動作について図５～図１２を用いて説明する。

[0062] 図５は、ウェブページ分析サーバ１０においてウェブページを分析する動作例を示すフローチャートである。図６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７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８は、ウェブページのテキストデータの分析過程の一例を示す模式図である。図９は、検索サーバ３０の検索結果の一例を示す模式図である。図１０は、ドメイン名ＤＢ１２ｂに記憶されたウェブサイトのドメイン名の一例を示す模式図である。図１１は、カテゴリ関連語ＤＢ１２ａに記憶されたカテゴリ関連語の一例を示す模式図である。図１２は、ウェブページ分析サーバ１０における類似度算出のサブルーチンの一例を示すフローチャートである。

[0063] （３．１．ウェブページの分析の流れ）

まず、ウェブページの分析の流れについて、図５を用いて説明する。

[0064] 図５に示すように、ウェブページ分析サーバ１０は、２つのウェブページを取得する（ステップＳ１）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、情報ＤＢ２２ａにおいて、同一のカテゴリ（例えば、カテゴリ”ワイン”）に所属している商品に関して、第１のウェブページ

および第２のウェブページを、通信部１１を通して情報ＤＢ２２ａから取得する。これらウェブページは、店舗主端末４０を通して、店舗主等により登録されたウェブページや、ユーザ端末４５から投稿されたブログのウェブページである。このように、ウェブページ分析サーバ１０のシステム制御部１４および通信部１１は、ウェブページを取得する、第１のウェブページおよび第２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。

[0065] 次に、ウェブページ分析サーバ１０は、各ウェブページからテキストデータを抽出する（ステップＳ２）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、各ウェブページ内に記述されているテキストデータを抽出する。さらに具体的には、図６に示すように、ウェブページ５０において、テキスト部５１、５２、５３、５４のテキストデータが抽出される。また、図７に示すように、ＨＴＭＬ等のマークアップ言語等で記述されたソースコードの中のデータも利用する。例えば、ＨＴＭＬタグのtitle要素の部分のテキストデータも抽出される。

[0066] このように、ウェブページ分析サーバ１０のシステム制御部１４は、各ウェブページ内に記述されているテキストデータを抽出するテキストデータ抽出手段の一例として機能する。

[0067] 次に、ウェブページ分析サーバ１０は、抽出したテキストデータから言語解析の一例である形態素解析によって特徴語候補を抽出する（ステップＳ３）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、形態素解析のプログラムを用い、抽出したテキストデータを形態素に分解し、各文節の品詞等を求める。図８に示すように、例えば、語句群５３ａが得られる。そして、ウェブページ分析サーバ１０のシステム制御部１４は、語句群５３ａから、ノイズを除去して、特徴語候補群５３ｂを抽出する。例えば、”［　　］”のように括弧や記号で括られた部分や、動詞句や、動詞句に係る語句等が、語句群５３ａから取り除かれ、特徴語候補群５３ｂが得られる。ここで、括弧や記号で括られた部分はノイズである場合が多く、また

、動詞句や動詞句に係る語句は商品やサービスやブログに関連する語句でないことが多い。

[0068] なお、形態素解析として、一般的な形態素解析のプログラムを用いればよく、形態素解析する際、複合語を形成する過程があってもよい。また、他方のウェブページについても、特徴語候補が抽出される。ウェブページ分析サーバ１０のシステム制御部１４は、このようなノイズの形態素に関する情報を、予め特徴語ＤＢ１２ｃに記憶しておく。また、ウェブページ分析サーバ１０のシステム制御部１４は、特徴語候補としてではなく、特徴語として抽出してもよい。

[0069] このようにウェブページ分析サーバ１０のシステム制御部１４は、抽出したテキストデータを言語解析して、各ウェブページの特徴語候補を抽出する特徴語候補抽出手段の一例として機能する。また、ウェブページ分析サーバ１０のシステム制御部１４は、抽出した前記テキストデータを言語解析して、各ウェブページの特徴語を抽出する特徴語抽出手段の一例として機能する。また、ウェブページ分析サーバ１０のシステム制御部１４は、ノイズの形態素を記憶するノイズ形態素記憶手段の一例として機能する。また、ウェブページ分析サーバ１０のシステム制御部１４は、ノイズ形態素記憶手段を参照して、テキストデータを形態素解析によって言語解析して得られた形態素から、ノイズの形態素を除去するノイズ除去手段の一例として機能する。

[0070] 次に、ウェブページ分析サーバ１０は、各特徴語候補によりウェブ検索を行う（ステップＳ４）。具体的には、まず、ウェブページ分析サーバ１０のシステム制御部１４は、検索サーバ３０が提供する検索ＡＰＩの機能を利用するために、各特徴語候補を検索キーワードとする各検索クエリを作成する。そして、ウェブページ分析サーバ１０のシステム制御部１４は、通信部１１を通して、検索サーバ３０に送信し、検索サーバ３０から特徴語候補毎の検索結果を受信する。図９に示すように、検索結果ページ６０の検索結果毎に、検索結果のタイトル部６１、スニペット部６２が含まれる。なお、使用する検索エンジンによって、タイトル部とスニペット部とに明確に分けられ

ていない場合は、検索結果のテキストデータの一部をスニペットとする。このように、ウェブページ分析サーバ１０のシステム制御部１４は、特徴語候補を検索キーワードとする検索クエリを作成し、検索クエリに基づく検索結果を取得する検索結果取得手段の一例として機能する。

[0071] 次に、ウェブページ分析サーバ１０は、ショッピングサイト等のウェブサイトの検索結果に絞る（ステップＳ５）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、検索結果のリンク先のＵＲＬ（Uniform Resource Locator）において、ショッピングサイト等のウェブサイトのドメイン名を含む検索結果に絞る。さらに具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、図１０のように、ドメイン名ＤＢ１２ｂを参照して、上位（例えば３０位）の検索結果のうち、ＨＴＭＬのタグ”<a href=・・・>”の情報やスニペット部６２のＵＲＬ６２ａの中にショッピングサイト等のウェブサイトのドメイン名を含む検索結果に絞る。なお、ウェブページ分析サーバ１０は、ステップＳ４の検索前に、ドメイン名ＤＢ１２ｂを参照してもよい。この場合、検索するドメインを検索ＡＰＩのパラメータとして、ウェブページ分析サーバ１０は、検索を行う。

[0072] 次に、ウェブページ分析サーバ１０は、カテゴリ関連語がスニペットに含まれる場合の特徴語候補を特徴語として決定する（ステップＳ６）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ５で絞られた検索結果のスニペットの中に、カテゴリ関連語が含まれる場合の特徴語候補を選択し、この特徴語候補を特徴語として決定する。さらに具体的には、図１１に示すように、ウェブページ分析サーバ１０のシステム制御部１４は、カテゴリが”ワイン”の場合、検索結果のスニペットに、カテゴリ関連語”ワイン”または”酒”が含まれる特徴語候補を選択し、特徴語候補を特徴語として決定する。そして、図８に示すように、例えば、語句群５３ｃが得られ、特徴語ＤＢ１２ｃに記憶される。

[0073] このようにウェブページ分析サーバ１０のシステム制御部１４は、カテゴリ関連語記憶手段の一例のカテゴリ関連語ＤＢ１２ａを参照して、検索結果

のスニペット中にカテゴリ関連語が存在するか否かを判定する検索結果判定手段の一例として機能する。また、ウェブページ分析サーバ１０のシステム制御部１４は、スニペット中にカテゴリ関連語が存在する場合の特徴語候補を各ウェブページの特徴語に決定する特徴語決定手段の一例として機能する。また、ウェブページ分析サーバ１０のシステム制御部１４は、カテゴリ関連語記憶手段およびドメイン名記憶手段を参照して、検索結果において、ドメイン名に関連したウェブサイトのスニペット中にカテゴリ関連語が存在するか否かを判定する検索結果判定手段の一例として機能する。

[0074] 次に、ウェブページ分析サーバ１０は、特徴語に基づき類似度を算出する（ステップＳ７）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ウェブページ間の類似度を算出するサブルーチンにより、類似度を算出する。なお、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳで抽出した特徴語候補に基づき類似度を算出してもよい。

[0075] このようにウェブページ分析サーバ１０のシステム制御部１４は、特徴語に基づき、第１のウェブページと第２のウェブページとの間の類似度を算出する類似度算出手段の一例として機能する。

[0076] 次に、ウェブページ分析サーバ１０は、類似度に基づきウェブページの類似性の判定を行う（ステップＳ８）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、類似度算出のサブルーチンにより算出した類似度が所定の値以上である場合、ウェブページ同士は類似商品等の類似対象を扱うウェブページであると判定する。なお、類似商品等の類似対象を扱うウェブページであると判定されたウェブページのうち、さらに類似度が高い場合、同一の商品等の同一の対象を扱うウェブページであると判定してもよい。このようにウェブページ分析サーバ１０のシステム制御部１４は、類似度が所定の値以上である場合、第１のウェブページと第２のウェブページとは類似である対象を扱うウェブページであると判定する類似対象判定手段の一例として機能する。

[0077] ウェブページの類似性の判定を利用して、同一の対象や類似の対象を扱う

ウェブページを集め、ユーザ端末４５を通して、ユーザが閲覧できるように、ウェブページのデータベースを構築する。例えば、ウェブページに、ある商品に対するＧＵＩ（Graphical User Interface）のボタン等を設け、ユーザがクリックすることにより、同一の対象や類似の対象を扱うウェブページの一覧が見られるようにする。

[0078] （３．２．特徴語に基づく類似度の算出）

次に、特徴語に基づく類似度の算出のサブルーチンについて、図１２を用いて説明する。

[0079] 図１２に示すように、ウェブページ分析サーバ１０は、各特徴語のＩＤＦ値を算出する（ステップＳ１０）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、各特徴語に対して、下記の式（１）に示すようなＩＤＦ（Inverse Document Frequency（逆出現頻度））の値を、算出する。

$$IDF_w = \log(N / N_w) \quad \cdots (1)$$

[0080] ここで、IDF_wは特徴語ｗのＩＤＦ値であり、Ｎはあるショッピングサイトの全ウェブページ数であり、N_wは前記ショッピングサイトにおいて特徴語ｗが出現するウェブページ数（ウェブページ出現数の一例）である。特徴語ｗのウェブページ数N_wの値が小さいほどIDF_wの値が大きくなり、特徴語ｗは商品を扱うウェブページに特有な語句である可能性が高くなる。なお、ＩＤＦ値を算出するためのウェブページの範囲（各ウェブページが属するウェブサイトの範囲の一例）は、検索エンジンがクローリングしたページ数でもよいし、ある国のウェブページや、ある国の複数のショッピングサイトのウェブページ（各ウェブページが属するウェブサイトの一例）や、特定しているカテゴリにおける上位のカテゴリのウェブページ等でもよい。

[0081] 次に、ウェブページ分析サーバ１０は、ＩＤＦ値に基づく類似度を算出する（ステップＳ１１）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、下記の式（２）に示すように、ステップＳ１０で算出したＩＤＦ値に基づき、第１のウェブページと第２のウェブページとの類似度S_{12}

を算出する。

S_{12}＝［Σｃ∈C_{12}（ＩＤＦｃ）］／［Σｕ∈U_{12}（ＩＤＦｕ）］・・・（２）

[0082] ここで、集合U_{12}は、第１のウェブページの特徴語と第２のウェブページの特徴語とを合わせた集合である。すなわち、第１のウェブページの特徴語の集合と、第２のウェブページの特徴語の集合との和集合である。集合C_{12}は、第１のウェブページと第２のウェブページとの共通の特徴語の集合である。すなわち、第１のウェブページの特徴語の集合と、第２のウェブページの特徴語の集合との積集合である。また、式（２）の分母は、集合U_{12}の要素ｕである特徴語のＩＤＦ値の和であり、式（２）の分子は、集合C_{12}の要素ｃである共通の特徴語のＩＤＦ値の和である。式（２）に示すように、第１のウェブページと第２のウェブページとにおいて、共通の特徴語があればあるほど、類似度S_{12}が高くなり、また、その特徴語のＩＤＦ値が高いほど、類似度S_{12}が高くなる。

[0083] このようにウェブページ分析サーバ１０のシステム制御部１４は、特徴語が、各ウェブページが属するウェブサイトの範囲において出現するウェブページ出現数に基づき、第１のウェブページと第２のウェブページとの間の類似度を算出する類似度算出手段の一例として機能する。またウェブページ分析サーバ１０のシステム制御部１４は、第１のウェブページおよび第２のウェブページにおいて共通の特徴語のウェブページ出現数に基づき、類似度を算出する類似度算出手段の一例として機能する。また、ウェブページ分析サーバ１０のシステム制御部１４は、各ウェブページが属するウェブサイトの範囲に存在するウェブページの数とウェブページ出現数とから、前記特徴語の逆出現頻度を算出し、当該逆出現頻度に基づき類似度を算出する類似度算出手段の一例として機能する。

[0084] この式（２）の類似度に基づき、ステップＳ８において、ウェブページ分析サーバ１０が、ウェブページの類似性の判定を行う。

[0085] 以上、本実施形態は、ウェブページ５０内に記述されているテキスト部５

１、５２、５３等のテキストデータを抽出し、テキストデータを言語解析して、ウェブページの特徴語候補を抽出し、特徴語候補を検索キーワードとする検索クエリを作成し、検索クエリに基づく検索結果ページ６０を取得し、検索結果ページのスニペット部６２のスニペット中にカテゴリ関連語が存在するか否かを判定し、スニペット中にカテゴリ関連語が存在する場合の特徴語候補をウェブページの特徴語に決定する。従って、本実施形態によれば、言語解析してウェブページの特徴語候補を抽出し、さらに、検索クエリに基づき特徴語を決定しているため、ウェブページの特徴語を、精度良く求めることができる。

[0086] また、本実施形態によれば、特徴語候補を絞る際、ウェブ検索により求まるスニペットを使用することにより、インターネット上における語句の使用状況を反映でき、高精度の特徴語を抽出することができる。さらに、スニペットに限ることにより、高速に特徴語を求めることができる。さらにまた、ウェブ検索のスニペットを使用することにより、商品等のトレンドが変化してもトレンドに追随でき、適切な特徴語（例えば、商品名や型番）が求められる。また、カテゴリ関連語がスニペット含まれるか否かのように、ウェブページに含まれる対象の大枠を表すカテゴリ関連語を用いて判定することで、言語解析したウェブページに含まれる内容を反映させた適切な特徴語を決定する判定が精度良く容易にできる。

[0087] また、本実施形態によれば、ウェブページ５０内に記述されているテキスト部５１、５２、５３等のテキストデータを抽出し、テキストデータを言語解析して、ウェブページの特徴語を抽出し、当該特徴語が、各ウェブページが属するウェブサイトの範囲において出現するウェブページ出現数に基づき、第１のウェブページと第２のウェブページとの間の類似度を算出し、類似度が所定の値以上である場合、第１のウェブページと第２のウェブページとは類似である対象を扱うウェブページであると判定することにより、言語解析してウェブページの特徴語を抽出し、特徴語のウェブページ出現数に基づき類似度を算出しているため、ウェブページの類似性を求めることができる

。また、本実施形態によれば、ウェブページの類似度を精度良く求めることができる。さらに、本実施形態によれば、ウェブページで扱われる商品等の対象の類似性を求めて、例えば、同一の対象や類似の対象を扱うウェブページをまとめ上げることができるため、ユーザの利便性の向上を図ることができる。

[0088] また、ウェブページ分析サーバ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分類するカテゴリに関連するカテゴリ関連語をカテゴリ関連語ＤＢ１２ａに記憶し、第１のウェブページおよび第２のウェブページを取得し、各ウェブページ５０内に記述されているテキスト部５１、５２、５３等のテキストデータを言語解析して各ウェブページの特徴語候補を抽出し、特徴語候補を検索キーワードとする検索クエリを作成して検索クエリに基づく検索結果ページ６０を取得し、検索結果ページのスニペット部６２のスニペット中にカテゴリ関連語が存在するか否かを判定し、スニペット中にカテゴリ関連語が存在する場合の特徴語候補を各ウェブページの特徴語に決定し、当該特徴語が、各ウェブページが属するウェブサイトの範囲において出現するウェブページ出現数に基づき、第１のウェブページと第２のウェブページとの間の類似度を算出し、類似度が所定の値以上である場合、第１のウェブページと第２のウェブページとは類似である対象を扱うウェブページであると判定してもよい。

[0089] この場合、ウェブページ分析サーバ１０のシステム制御部１４は、特徴語候補を絞る際、ウェブ検索により求まるスニペットを使用することにより、インターネット上における語句の使用状況を反映でき、高精度の特徴語を抽出することができる。さらに、ウェブページ分析サーバ１０のシステム制御部１４は、スニペットに限ることにより、高速に特徴語を求めることができる。さらにまた、ウェブ検索のスニペットを使用することにより、商品等のトレンドが変化してもトレンドに追随でき、適切な特徴語が求められ、ウェブページ分析サーバ１０のシステム制御部１４は、ウェブページの類似度を精度良く求めることができる。

[0090] また、ウェブページ分析サーバ１０のシステム制御部１４が、インターネットにおける所定の範囲を、各ウェブページが属するウェブサイトとし、特徴語が出現するウェブページ出現数を算出し、当該ウェブページ出現数に基づき、類似度を算出する場合、当該ウェブサイトを反映した類似度を算出できる。

[0091] また、ウェブページ分析サーバ１０のシステム制御部１４が、各ウェブページが属するウェブサイトの範囲に存在するウェブページの数と前記ウェブページ出現数とから、前記特徴語の逆出現頻度を算出し、当該逆出現頻度に基づき前記類似度を算出する場合、特徴語の逆出現頻度により、逆出現頻度が高い順等に特徴語のランク付けができるため、類似性の判断がしやすくなる。

[0092] また、ウェブページ分析サーバ１０のシステム制御部１４が、第１のウェブページおよび第２のウェブページにおいて共通の特徴語のウェブページ出現数に基づき、類似度を算出する場合、共通の特徴語により的確に類似度を算出できる。

[0093] また、ウェブページ分析サーバ１０のシステム制御部１４が、テキストデータを形態素解析して得られた形態素から、ノイズの形態素を除去する場合、的確な特徴語候補を抽出できるため、特徴語の精度が高くなる。

[0094] また、ウェブページ分析サーバ１０のシステム制御部１４が、カテゴリ関連語ＤＢ（カテゴリ関連語記憶手段）１２ａを参照して、検索結果において、ドメイン名に関連したウェブサイトのスニペット中にカテゴリ関連語が存在するか否かを判定する場合、対象の説明がなされているウェブページを多く有するウェブサイトに絞るため、対象を適切に特徴付ける特徴語を抽出でき、特徴語の精度が高くなる。

[0095] また、ウェブページ分析サーバ１０のシステム制御部１４が、特徴語に基づき、第１のウェブページと第２のウェブページとの間の類似度を算出し、類似度が所定の値以上である場合、第１のウェブページと第２のウェブページとは類似対象を扱うウェブページであると判定し、ウェブページの商品の

類似性を求めて、例えば、同一の対象や類似の対象を扱うウェブページをまとめ上げることができ、ユーザの利便性の向上を図ることができる。

[0096] （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。

[0097] ［４．第２実施形態に係る検索システムの動作］

次に、本発明の第２実施形態に係る情報処理システム１の検索システムとしての動作について図１３～図１５を用いて説明する。

[0098] 図１３は、ウェブページ分析サーバ１０においてウェブページをグルーピングする動作例を示すフローチャートである。図１４は、ユーザ端末４５に送信されるグルーピングされた検索結果の一例を示す模式図である。図１５は、ウェブページ分析サーバ１０におけるグルーピングのサブルーチンの一例を示すフローチャートである。

[0099] （４．１．ウェブページのグルーピングの流れ）

ウェブページのグルーピングの流れについて、図１３を用いて説明する。

[0100] まず、ユーザはユーザ端末４５に、探している商品等のキーワードを入力する。例えば、”白ワイン”と入力し検索を行うと、検索キーワード”白ワイン”を含む検索クエリがウェブページ分析サーバ１０に送信される。

[0101] 次に、図１３に示すように、ウェブページ分析サーバ１０は、ユーザ端末４５から検索クエリを受信する（ステップＳ１５）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、通信部１１を通して、ユーザ端末４５から検索キーワードを含む検索クエリを受信する。このように、ウェブページ分析サーバ１０のシステム制御部１４および通信部１１は、ユーザ端末からユーザ検索クエリを受信する検索クエリ受信手段の一例として機能する。

[0102] 次に、ウェブページ分析サーバ１０は、検索クエリに基づき検索をする（

ステップＳ１６）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、情報ＤＢ２２ａに対して、検索キーワードに対応する商品等のウェブページの検索を行う。

[0103] 次に、ウェブページ分析サーバ１０は、検索結果に基づく複数のウェブページを取得する（ステップＳ１７）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、情報ＤＢ２２ａにおいて、検索キーワードによりヒットした商品等の対象を扱う複数のウェブページの情報を取得する。これらウェブページは、例えば、店舗主端末４０を通して、店舗主等により登録されたウェブページや、ユーザ端末４５から投稿されたブログのウェブページである。このように、ウェブページ分析サーバ１０のシステム制御部１４および通信部１１は、ユーザ検索クエリに基づく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。

[0104] 次に、ウェブページ分析サーバ１０は、各ウェブページからテキストデータを抽出する（ステップＳ１８）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ２と同様にテキストデータを抽出する。

[0105] 次に、ウェブページ分析サーバ１０は、抽出したテキストデータから言語解析の一例である形態素解析によって特徴語候補を抽出する（ステップＳ１９）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ３と同様に、特徴語候補を抽出する。

[0106] 次に、ウェブページ分析サーバ１０は、各特徴語候補によりウェブ検索を行う（ステップＳ２０）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ４と同様にウェブ検索を行う。

[0107] 次に、ウェブページ分析サーバ１０は、ショッピングサイト等のウェブサイトの検索結果に絞る（ステップＳ２１）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ５と同様に、ショッピングサイト等のウェブサイトの検索結果に絞る。

[0108] 次に、ウェブページ分析サーバ１０は、カテゴリ関連語がスニペットに含

まれる場合の特徴語候補を特徴語として決定する（ステップＳ２２）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ６と同様に、特徴語候補を特徴語として決定する。

[0109] 次に、ウェブページ分析サーバ１０は、特徴語に基づき複数のウェブページをグルーピングする（ステップＳ２３）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、グルーピングのサブルーチンにより、各特徴語のＩＤＦ値を算出し、任意の２つのウェブページ間において、ＩＤＦ値から類似度（任意の２つのウェブページ間の第２類似度の一例）を求め、類似度に基づき類似のウェブページ同士をグルーピングする。図１４に示すように、検索結果のウェブページ７０において、商品等毎にグループ名部７１が設けられる。グループ名部７１には、具体的な商品名やサービス名やブログからの特徴語が利用される。また、各グループ名部７１以下に、同じグルーピングに属するウェブページのタイトル部７２と、スニペット部７３とが、集められる。例えば、タイトル部７２およびスニペット部７３がリスティングされる。

[0110] このようにウェブページ分析サーバ１０のシステム制御部１４は、取得したユーザ検索クエリに基づく複数のウェブページのうち任意の２つのウェブページ間の第２類似度を、類似度として、特徴語に基づき算出する類似度算出手段の一例として機能する。

[0111] 次に、ウェブページ分析サーバ１０は、ユーザ端末に検索結果を送信する（ステップＳ２４）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、通信部１１を通して、図１４に示すような検索結果を表示させるＨＴＭＬ等のマークアップ言語等の情報を、検索クエリを送信してきたユーザ端末４５に送信する。このように、ウェブページ分析サーバ１０のシステム制御部１４および通信部１１は、特徴語毎にグルーピングしたウェブページに関する情報を検索結果としてユーザ端末４５に送信する送信手段の一例として機能する。

[0112] なお、例えば「ボルドー　ワイン」という語をユーザが検索した場合、ウ

ェブページ分析サーバ１０のシステム制御部１４は、ボルドー産ワインに関する複数のウェブページを取得し、その複数のウェブページそれぞれの特徴語を上記手法で求め、特徴語が同じまたは類似のもの同士をまとめて、検索結果としてユーザ端末４５に送信してもよい。

[0113] さらに、例えば、Ａ，Ｂ，Ｃ，Ｄの４つのウェブページがある場合に、ウェブページ分析サーバ１０のシステム制御部１４は、ＡＢの組み合わせ及びＣＤの組み合わせの類似度が、所定の値以上であれば、ＡＢを第１グループ、ＣＤを第２グループとしたグルーピングを行ってもよい。また、ウェブページ分析サーバ１０のシステム制御部１４は、ＡＢＣの各ウェブページ間の類似度が、所定の値以上ならば、ＡＢＣをグルーピングしてもよい。

[0114] （４．２．特徴語に基づくグルーピング）

次に、特徴語に基づくグルーピングのサブルーチンについて、図１５を用いて説明する。

[0115] 図１５に示すように、ウェブページ分析サーバ１０は、各特徴語のＩＤＦ値を算出する（ステップＳ２５）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ１０と同様に、各特徴語のＩＤＦ値を算出する。

[0116] 次に、ウェブページ分析サーバ１０は、ＩＤＦ値に基づく類似度を算出する（ステップＳ２６）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ１１と同様に、ＩＤＦ値に基づく類似度（第２類似度の一例）を算出する。

[0117] 次に、ウェブページ分析サーバ１０は、類似度に基づき類似のウェブページ同士をグルーピングする（ステップＳ２７）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、算出した類似度（第２類似度の一例）が所定の値以上である場合、ウェブページ同士は類似対象を扱うウェブページであると判定し、類似の対象を扱うウェブページとして同じグループにする。なお、類似対象を扱うウェブページであると判定されたウェブページのうち、さらに類似度が高い場合、同一の対象を扱うウェブページである

と判定してもよい。このようにウェブページ分析サーバ１０のシステム制御部１４は、第２類似度が所定の値以上である場合のウェブページ同士をグルーピングするウェブページ・グルーピング手段の一例として機能する。また、ウェブページ分析サーバ１０のシステム制御部１４は、第２類似度に基づき、複数のウェブページを特徴語毎にグルーピングするウェブページ・グルーピング手段の一例として機能する。

[0118] なお、検索キーワードに対応するカテゴリが複数にまたがる場合、ウェブページ分析サーバ１０のシステム制御部１４は、カテゴリ毎に以上の処理を行い、検索結果をカテゴリ毎に分けて、ウェブページのグループを表示するようにする。

[0119] また、ウェブページ分析サーバ１０のシステム制御部１４は、所定数以上の共通する特徴語を互いに有するウェブページ同士をグルーピングしてもよい。ウェブページ分析サーバ１０のシステム制御部１４は、各特徴語のＩＤＦ値を算出して各特徴語にＩＤＦ値によるランクを設け、ＩＤＦ値が所定以上の共通する特徴語を所定数以上互いに有するウェブページ同士をグルーピングしてもよい。

[0120] このようにウェブページ分析サーバ１０のシステム制御部１４は、ユーザ検索クエリによって得られる複数のウェブページを特徴語毎にグルーピングするウェブページ・グルーピング手段の一例として機能する。

[0121] 以上、本実施形態によれば、ユーザ端末４５からユーザ検索クエリを受信し、ユーザ検索クエリに基づく複数のウェブページを、ウェブページの中から取得し、各ウェブページ５０内に記述されているテキスト部５１、５２、５３等のテキストデータを言語解析して各ウェブページの特徴語候補を抽出し、特徴語候補を検索キーワードとする検索クエリを作成して検索クエリに基づく検索結果ページ６０を取得し、検索結果ページのスニペット部６２のスニペット中にカテゴリ関連語が存在するか否かを判定し、スニペット中にカテゴリ関連語が存在する場合の特徴語候補を各ウェブページの特徴語に決定し、複数のウェブページを特徴語毎にグルーピングし、特徴語毎にグルー

ピングしたウェブページに関する情報を検索結果（検索結果のウェブページ７０）としてユーザ端末４５に送信する場合、複数のウェブページを特徴語毎にグルーピングすることにより、ユーザが見やすい検索結果に反映させる等、利用価値が高まり、また、ユーザが見やすい検索結果を提示してユーザの利便性の向上を図ることができる。

[0122] また、本実施形態によれば、ユーザ端末４５からユーザ検索クエリを受信し、ユーザ検索クエリに基づく複数のウェブページを取得し、当該取得したユーザ検索クエリに基づく複数のウェブページのうち任意の２つのウェブページ間の第２類似度を類似度として、特徴語に基づき算出し、第２類似度が所定の値以上である場合のウェブページ同士をグルーピングし、グルーピングしたウェブページに関する情報を検索結果としてユーザ端末に送信することにより、複数のウェブページを特徴語毎にグルーピングすることにより、ユーザが見やすい検索結果に反映させる等、利用価値が高まり、また、ユーザが見やすい検索結果を提示してユーザの利便性の向上を図ることができる。また、ウェブページ分析サーバ１０のシステム制御部１４が、類似度に閾値を設けたことにより、類似度により一律に、ウェブページをグルーピングすることができる。

[0123] また、ウェブページ分析サーバ１０のシステム制御部１４が、第２類似度に基づき、複数のウェブページを特徴語毎にグルーピングする場合、類似度により異なる複数の特徴語でもウェブページをグルーピングしやすくなる。

[0124] また、ウェブページ分析サーバ１０のシステム制御部１４が、特徴語に基づき、複数のウェブページのうち任意の２つのウェブページ間の類似度を算出し、類似度に基づき、複数のウェブページをグルーピングする場合、類似度により異なる複数の特徴語でもウェブページをグルーピングしやすくなる。

[0125] （第３実施形態）

次に、本発明の第３実施形態に係る情報処理システム１のウェブページ分析システムとしての動作について図１６～図２２を用いて説明する。

［５．第３実施形態に係るウェブページ分析システムの動作］

[0126] （５．１．ウェブページ分析の全体の流れ）

まず、ウェブページ分析の全体の流れについて、図１６を用いて説明する。

図１６は、ウェブページ分析サーバ１０においてウェブページ分析の全体の流れを示す説明図である。なお、図中に後述するフローチャートのステップに対応する符号を示した。

[0127] 図１６に示すように、まず、ウェブページ分析サーバ１０は、ウェブページ取得手段の一例として、情報提供サーバ２０の情報ＤＢ２２ａから、複数のウェブページを取得する。

[0128] ウェブページを取得した後、ウェブページ分析サーバ１０の処理は２つに分かれる。２つの処理は、ウェブページからウェブページの特徴語を抽出してウェブページ間の類似度を求める処理と、ウェブページから商品の属性・属性値を求める処理である。なお、属性抽出手段と、属性記述パターン抽出手段とが所定の回数繰り返され、ブートストラップ法を構成する。

[0129] 最後に、ウェブページ分析サーバ１０は、類似対象判定手段の一例として、類似度と属性とに基づき、２つのウェブページが類似対象を扱うウェブページであるか否かを判定する。

[0130] （５．２．ウェブページの特徴語を抽出して分析する流れ）

次に、ウェブページ分析サーバ１０においてウェブページの特徴語を抽出してウェブページを分析する流れについて、図１７を用いて説明する。

[0131] 図１７は、ウェブページ分析サーバ１０においてウェブページの特徴語を抽出してウェブページを分析する動作例を示すフローチャートである。

[0132] 図１７に示すように、ウェブページ分析サーバ１０は、複数のウェブページを取得する（ステップＳ３０）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ１で２つのウェブページを取得するように、複数のウェブページを取得する。

[0133] 次に、ウェブページ分析サーバ１０は、各ウェブページからテキストデー

タを抽出する（ステップＳ３１）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ２と同様に各ウェブページからテキストデータを抽出する。

[0134] 次に、ウェブページ分析サーバ１０は、抽出したテキストデータから言語解析の一例として形態素解析によって特徴語候補を抽出する（ステップＳ３２）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ３と同様に特徴語候補を抽出する。

[0135] 次に、ウェブページ分析サーバ１０は、各特徴語候補によりウェブ検索を行う（ステップＳ３３）。具体的には、まず、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ４と同様に、ウェブ検索を行う。

[0136] 次に、ウェブページ分析サーバ１０は、ショッピングサイト等のウェブサイトの検索結果に絞る（ステップＳ３４）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ５と同様にウェブサイトの検索結果に絞る。

[0137] 次に、ウェブページ分析サーバ１０は、カテゴリ関連語がスニペットに含まれる場合の特徴語候補を特徴語として決定する（ステップＳ３５）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ６と同様に特徴語候補を特徴語として決定する。

[0138] 次に、ウェブページ分析サーバ１０は、特徴語に基づき類似度を算出する（ステップＳ３６）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ７同様に類似度の算出のサブルーチンにより類似度（取得した複数のウェブページのうち任意の２つのウェブページ間の第３類似度の一例）を算出する。このようにウェブページ分析サーバ１０のシステム制御部１４は、取得した複数のウェブページのうち任意の２つのウェブページ間の第３類似度を類似度として、特徴語に基づき算出する類似度算出手段の一例として機能する。

[0139] 次に、ウェブページ分析サーバ１０は、類似度に基づきウェブページの類似性の判定を行う（ステップＳ３７）。具体的には、ウェブページ分析サー

バ１０のシステム制御部１４は、ステップＳ８と同様にウェブページの類似性の判定を行う。

[0140] ウェブページ同士の類似度が所定の値以上と判定された場合（ステップＳ３７；ＹＥＳ）、ウェブページ分析サーバ１０は、類似度と属性とに基づき、ウェブページの類似性の判定を行う（ステップＳ３８）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、類似度と、任意の２つのウェブページ間での共通の属性関連語とに基づき、任意の２つのウェブページとは類似商品等の類似対象を扱うウェブページであるか否かを判定する。さらに具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、類似度（第３類似度）が所定の値以上である場合で、かつ、任意の２つのウェブページ間で共通の属性に係る属性値において等しい属性値がある場合に、任意の２つのウェブページとは類似商品等の類似対象を扱うウェブページであると判定する。また、共通の属性において、属性値が等しい場合の数が、全てである場合や、半数以上の等場合に、同一の商品等の同一対象を扱ったウェブページと判定してもよい。なお、属性関連語等については、図１６に示すように、ウェブページ分析サーバ１０が別の処理により求める（詳細は後述）。このようにウェブページ分析サーバ１０のシステム制御部１４は、第３類似度と、任意の２つのウェブページ間での共通の属性関連語とに基づき、任意の２つのウェブページが扱う対象は、互いに類似する対象であるか否かを判定する類似対象判定手段の一例として機能する。また、ウェブページ分析サーバ１０のシステム制御部１４は、第３類似度が所定の値以上である場合で、かつ、任意の２つのページ間で共通の属性に係る属性値において等しい属性値がある場合に、任意の２つのウェブページが扱う対象は、互いに類似する対象であると判定する類似対象判定手段の一例として機能する。

[0141] 次に、ウェブページ同士の類似度が所定の値以上と判定されない場合（ステップＳ３７；ＮＯ）、ウェブページは、類似商品等の類似対象を扱っていないとして、処理を終了する。

[0142] （５．３．商品の属性・属性値の抽出の流れ）

次に、ウェブページから商品の属性・属性値を抽出する動作の流れについて、図１８～図２２を用いて説明する。

[0143] 図１８は、ウェブページ分析サーバ１０においてウェブページから商品の属性・属性値を抽出する動作例を示すフローチャートである。図１９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２０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２１は、抽出された属性・属性値の一例を示す模式図である。図２２は、ウェブページ分析サーバ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。

[0144] 図１８に示すように、ウェブページ分析サーバ１０は、ステップＳ３０で、複数のウェブページを取得した後、属性記述パターンを取得する（ステップＳ４０）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、下記のステップＳ４１～Ｓ４３におけるブートストラップ法の初期データとして、図１９に示すように、属性記述パターンＤＢ１２ｄ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１９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。このようにウェブページ分析サーバ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。

[0145] 次に、ウェブページ分析サーバ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Ｓ４１）。具体的には、ウェブページ分析サーバ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２０に示すように、属性記述パターン８１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ＤＢ１２ｅ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。

[0146] このようにウェブページ分析サーバ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、ウェブページ分析サーバ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、ウェブページ分析サーバ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。

[0147] 次に、ウェブページ分析サーバ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Ｓ４２）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、図２０に示すように、属性・属性値８２（例えば、属性名”品種”および属性値”〇〇〇”）に適合する（例えば、〈td〉品種〈/td〉〈td〉〇〇〇〈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１９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ＤＢ１２ｄ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。

[0148] このようにウェブページ分析サーバ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、ウェブページ分析サ

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。

[0149] 次に、ウェブページ分析サーバ１０は、所定回数を判定する（ステップＳ４３）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ４１およびステップＳ４２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Ｓ４３；ＮＯ）、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ４１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。ウェブページ分析サーバ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Ｓ２１およびステップＳ２２を繰り返す。なお、ウェブページ分析サーバ１０のシステム制御部１４は、抽出された属性・属性値の数（例えば、属性名の数や属性値の数や属性名と属性値の組み合わせの数）、または、属性記述パターンの数が、所定数以上になったか否かを判定してもよい。所定数以上になった場合に、ウェブページ分析サーバ１０のシステム制御部１４は、次のステップＳ４４の処理を行う。

[0150] このようにウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ４０からステップＳ４４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。ウェブページ分析サーバ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。

[0151] 所定回数に達した場合（ステップＳ４３；ＹＥＳ）、ウェブページ分析サーバ１０は、属性の選定を行う（ステップＳ４４）。具体的には、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ４１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、ウェブページ分析サーバ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２１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。

[0152] 抽出して選定された属性名および属性値は、ステップＳ３８において、ウェブページ分析サーバ１０が、ウェブページの類似性を判定する際に利用される。

[0153] なお、ブートストラップ法の初期データとして、属性・属性値ＤＢ１２ｅの属性・属性値リストから、初期の属性・属性値を取得してもよい。このようにウェブページ分析サーバ１０のシステム制御部１４は、初期データとして、対象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。

[0154] そして、この場合、ウェブページ分析サーバ１０のシステム制御部１４は、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。

[0155] （５．４．属性の選定）

次に、属性の選定のサブルーチンについて、図２２を用いて説明する。

[0156] 図２２に示すように、ウェブページ分析サーバ１０は、属性へのスコア付けを行う（ステップＳ５０）。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、ウェブページ分析サーバ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。

[0157] 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。このようにウェブページ分析サーバ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、ウェブページ分析サーバ１０のシステム制御部１４は、ウェブページで扱われ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。

[0158] 次に、ウェブページ分析サーバ１０は、上位ランクの属性を選択する（ステップＳ５１）。具体的には、ウェブページ分析サーバ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。このようにウェブページ分析サーバ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。

[0159] 次に、ウェブページ分析サーバ１０は、属性のフィルタリングを行う（ステップＳ５２）。具体的には、ウェブページ分析サーバ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関連語の一例の属性名は、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。このようにウェブページ分析サーバ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。

[0160] 次に、ウェブページ分析サーバ１０は、同義の属性を集約する（ステップＳ５３）。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。ウェブページ分析サーバ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、類似概念の属性の属性名でもよい。

[0161] 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とを掛け合わせた値を類似の度合いとしてもよいし、これらの割合を元にエントロピーを計算し、掛け合わせた値を類似の度合いとしてもよいし、ジャッカード係数を類似の度合いとてもよいし、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してもよい。

[0162] このようにウェブページ分析サーバ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、ウェブページ分析サーバ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、ウェブページ分析サーバ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。

[0163] 以上、本実施形態によれば、各ウェブページとして、複数のウェブページを取得し、初期データとして、対象の属性に関連した属性関連語、または、

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前記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し、第２類似度および属性関連語に基づき、ウェブページをグルーピングすることにより、例えば、類似度によりグルーピングしたウェブページ７０をさらに、ウェブページ９０のように、共通する属性によりサブグルーピングできる。また、ユーザがより見やすい検索結果に反映させる等、利用価値が高まる。

[0164] また、本実施形態によれば、各ウェブページとして、複数のウェブページを取得し、当該取得した複数のウェブページのうち任意の２つのウェブページ間の第３類似度を類似度として、特徴語に基づき算出し、初期データとして、対象の属性に関連した属性関連語、または、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、抽出し、第３類似度と、任意の２つのウェブページ間での共通の属性関連語とに基づき、任意の２つのウェブページが扱う対象は、互いに類似する対象であるか否かを判定することにより、ウェブページで扱われる商品等の対象の類似性を求めることができる。また、本実施形態によれば、同一や類似の対象を扱うウェブページをまとめ上げることができる。特に、類似度の他に、更に共通の属性関連語に基づき判定することにより、判定の精度が向上する。ユーザの利便性も向上させることもできる。

[0165] また、本実施形態によれば、ウェブページとして、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し、特徴語および属性関連語に基づき、ウェブペー

ジをグルーピングすることにより、例えば、特徴語によりグルーピングしたウェブページ７０をさらに、ウェブページ９０のように、共通する属性によりサブグルーピングできる。また、ユーザがより見やすい検索結果に反映させる等、利用価値が高まる。

[0166] ウェブページ分析サーバ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。

[0167] ウェブページ分析サーバ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、任意の２つのウェブページ間で共通の属性に係る属性値において等しい属性値がある場合に、任意の２つのウェブページに記載された対象は、互いに類似する対象であると判定すると、同じ対象であれば属性値が一致しやすいので、同一や類似の対象を扱うウェブページであると精度良く判定しやすくなる。なお、属性名はカテゴリ特有の場合が多いが、属性記述パターンはカテゴリ特有のものもあるが、様々なカテゴリに適用可能である。

[0168] ウェブページ分析サーバ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。

[0169] ウェブページ分析サーバ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、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場合、選択された属性関連語において、商品等の対象を表す属性の精度が高くなる。

[0170] ウェブページ分析サーバ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き、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、商品等の対象を表す属性の精度が高くなる。例えば、店舗により扱う対象（アイテム）の数やウェブページの数が大きく異なる場合、多くの商品等の対象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。

[0171] ウェブページ分析サーバ１０のシステム制御部１４が、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場合、対象とするカテゴリ固有の属性関連語に絞ることにより、商品等を表す属性の精度が高くなる。

[0172] ウェブページ分析サーバ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、第３類似度が所定の値以上である場合で、かつ、任意の２つのページ間で共通の属性に係る属性値において等しい属性値がある場合に、任意の２つのウェブページが扱う対象は、互いに類似する対象であると判定すると、同じ対象であれば属性値が一致しやすいので、同一や類似の対象を扱うウェブページであると精度良く判定しやすくなる。

[0173] なお、検索サーバ３０は、サーバシステム５外である、他の検索サイトの検索サーバでもよい。この場合、ウェブページ分析サーバ１０は、ネットワーク３を通して、検索クエリを送信し、検索結果を受信する。また、スニペットを求める際のウェブ検索する範囲は、情報提供サーバ２０内のウェブページに限ったり、ある国や地域のウェブページに限ったりしてもよい。

[0174] また、属性の選定や属性記述パターンの選定を、ステップＳ２１やステップＳ２２のブートストラップ法の中で行ってもよい。この場合、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。

[0175] ここで、属性記述パターンの選定する場合、属性記述パターンと属性・属性値との共起確率を算出し、共起確率から算出されるエントロピー等を計算して属性記述パターンのスコア付けを行い、スコアに基づき属性記述パター

ンを選定する。

[0176] また、属性のスコアとして、店舗の数でなく、属性名が出現したウェブページの数でもよい。ウェブページ分析サーバ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。

[0177] 次に、ウェブページをグルーピングの変形例について図２３および図２４を用いて説明する。

[0178] 図２３は、図１４のグルーピングされた検索結果の一例を示す模式図である。図２４は、図１４のグルーピングされた検索結果の他の変形例を示す模式図である。

[0179] まず、図２３に示すように、ウェブページ分析サーバ１０のシステム制御部１４は、類似度に基づき類似のウェブページ同士をグルーピングしたり、所定数以上の共通する特徴語を互いに有するウェブページ同士にグルーピングしたりして、特徴語によりグループ名部９１を設定する。

[0180] さらに、ウェブページ分析サーバ１０のシステム制御部１４は、グルーピングされた各グループを、属性に基づき、サブグループ名部９４のように、サブグルーピングをする。ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ４４において求めた属性に基づき、例えば、属性名が”容量”に関して、容量が”７５０ｍｌ”と”３７５ｍｌ”とに分けてサブグルーピングする。ウェブページ分析サーバ１０のシステム制御部１４は、”容量”・”７５０ｍｌ”や”容量”・”３７５ｍｌ”のような共通する属性名・属性値を有するウェブページを、特徴語（または第２類似度）によりグルーピングされたウェブページの中から収集する。そして、ウェブページ９０において、グループ名部９１およびサブグループ名部９４によってサブグループ化された領域に、ウェブページのタイトル部９２と、スニペット部９３とが表示される。

[0181] また、図２４に示すように、ウェブページ分析サーバ１０のシステム制御

部１４は、グループ名部９５”容量：７５０ｍｌ”、”容量：３７５ｍｌ”のように、属性に基づき、グルーピングを行ってもよい。ウェブページ分析サーバ１０のシステム制御部１４は、属性名・属性値として、”容量”・”７５０ｍｌ”等を有するウェブページを収集する。さらに、グルーピングされた各グループを、特徴語や類似度に基づき、サブグループ名部９９のように、サブグルーピングをする。ウェブページ分析サーバ１０のシステム制御部１４は、属性に基づきグルーピングされたウェブページの特徴語を求めたり、類似度を求めたりする。そして、ウェブページ分析サーバ１０のシステム制御部１４は、ステップＳ２７のように、類似度に基づき類似のウェブページ同士をサブグルーピングしたり、所定数以上の共通する特徴語を互いに有するウェブページ同士をサブグルーピングしたりする。

[0182] ここで、属性を求める際、ステップＳ４０で必要な複数のウェブページの範囲は、特徴語に基づきグルーピングされた範囲でもよいし、ウェブサイト単位でもよいし、ステップＳ４０で扱うのと同じ範囲でもよい。

[0183] このように、ウェブページ分析サーバ１０のシステム制御部１４は、特徴語および属性関連語に基づき、前記ユーザ検索クエリによって得られる複数のウェブページをグルーピングするウェブページ・グルーピング手段の一例として機能する。また、ウェブページ分析サーバ１０のシステム制御部１４は、第２類似度および属性関連語に基づき、ウェブページ同士をグルーピングするウェブページ・グルーピング手段の一例として機能する。

[0184] ウェブページ分析サーバ１０のシステム制御部１４が、特徴語および属性関連語に基づき、ウェブページをグルーピングし、または、第２類似度および属性関連語に基づき、グルーピングする場合、特徴語や類似度によりグルーピングしたウェブページをさらに、共通する属性によりサブグルーピングでき、または、共通する属性によりグルーピングしたウェブページをさらに、特徴語や類似度によりサブグルーピングできる。従って、ユーザがより見やすい検索結果に反映させる等、利用価値が高まる。

[0185] なお、ウェブページ分析サーバ１０のシステム制御部１４は、抽出した、

または、決定した特徴語と、任意の２つのウェブページ間での共通の属性関連語とに基づき、任意の２つのウェブページが扱う対象は、互いに類似する対象であるか否かを判定してもよい。例えば、２つのウェブページにおいて、共通の特徴語がいくつか存在し、かつ、当該２つのウェブページ間で共通の属性に係る属性値において等しい属性値がある場合、ウェブページ分析サーバ１０のシステム制御部１４は、互いに類似する対象を扱うウェブページと判定する。

[0186] この場合、ウェブページ分析サーバ１０のシステム制御部１４は、ウェブページで扱われる商品等の対象の類似性を求めることができる。また、ウェブページ分析サーバ１０のシステム制御部１４は、同一や類似の対象を扱うウェブページをまとめ上げることができる。特に、特徴語の他に、更に共通の属性関連語に基づき判定することにより、判定の精度が向上する。

[0187] また、ウェブページ分析サーバ１０のシステム制御部１４は、特徴語を索引として、ウェブページを整理した辞書を作成してもよい。例えば、ウェブページ分析サーバ１０のシステム制御部１４は、特徴語に基づき、ウェブページをグルーピングすることで、特徴語を索引として関連したウェブページをまとめる。さらに、ウェブページ分析サーバ１０のシステム制御部１４は、属性名と属性値を索引として、ウェブページを整理した辞書を作成してもよい。例えば、ウェブページ分析サーバ１０のシステム制御部１４は、属性名・属性値に基づき、ウェブページをグルーピングすることで、特徴語を索引として関連したウェブページをまとめる。

[0188] また、ウェブページ分析サーバ１０のシステム制御部１４は、同一または極めた類似したウェブページ、名寄せのように、どちらか１つにするか、２つ併せるかして、まとめてもよい。例えば、ウェブページ分析サーバ１０のシステム制御部１４は、類似度が所定の値（例えば、ステップＳ８の場合よりさらに高い所定の値）以上の場合、ウェブページを１つにまとめる。

[0189] 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

[0190] ３：ネットワーク

５：サーバシステム

１０：ウェブページ分析サーバ（情報処理装置）

１２ａ：カテゴリ関連語ＤＢ（カテゴリ関連語記憶手段）

１２ｂ：ドメイン名ＤＢ

１２ｄ：属性記述パターンＤＢ

１２ｅ：属性・属性値ＤＢ

２０：情報提供サーバ

３０：検索サーバ

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に関連するカテゴリ関連語を記憶するカテゴリ関連語記憶手段と、

ウェブページを取得するウェブページ取得手段と、

前記ウェブページ内に記述されているテキストデータを抽出するテキストデータ抽出手段と、

抽出した前記テキストデータを言語解析して、前記ウェブページの特徴語候補を抽出する特徴語候補抽出手段と、

前記特徴語候補を検索キーワードとする検索クエリを作成し、前記検索クエリに基づく検索結果を取得する検索結果取得手段と、

前記カテゴリ関連語記憶手段を参照して、前記検索結果のスニペット中に前記カテゴリ関連語が存在するか否かを判定する検索結果判定手段と、

前記スニペット中に前記カテゴリ関連語が存在する場合の前記特徴語候補を前記ウェブページの特徴語に決定する特徴語決定手段と、

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、

ノイズの形態素を記憶するノイズ形態素記憶手段を更に備え、

前記特徴語候補抽出手段が、ノイズ形態素記憶手段を参照して、前記テキストデータを形態素解析によって前記言語解析して得られた形態素から、ノイズの形態素を除去するノイズ除去手段を有する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、

ウェブサイトのドメイン名を記憶するドメイン名記憶手段を更に備え、

前記検索結果判定手段が、前記カテゴリ関連語記憶手段および前記ドメイン名記憶手段を参照して、前記検索結果において、前記ドメイン名に関連したウェブサイトのスニペット中に前記カテゴリ関連語が

存在するか否かを判定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、

ユーザ端末からユーザ検索クエリを受信する検索クエリ受信手段と、

前記ユーザ検索クエリによって得られる複数のウェブページを、前記特徴語毎にグルーピングするウェブページ・グルーピング手段と、

前記特徴語毎にグルーピングした前記ウェブページに関する情報を検索結果として前記ユーザ端末に送信する送信手段を更に、

備えたことを特徴とする情報処理装置。

[請求項5] 請求項４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、前記ウェブページとして、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、を更に備え、

前記ウェブページ・グルーピング手段が、前記特徴語および前記属性関連語に基づき、前記ユーザ検索クエリによって得られる複数のウェブページをグルーピングすることを特徴とする情報処理装置。

[請求項6] 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、

前記ウェブページ取得手段が、前記ウェブページとして、複数のウェブページを取得し、

初期データとして、前記対象の属性に関連した属性関連語、または、前記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、

前記特徴語と、前記任意の２つのウェブページ間での共通の属性関連語とに基づき、前記任意の２つのウェブページが扱う対象は、互いに類似する対象であるか否かを判定する類似対象判定手段と、

を更に備えたことを特徴とする情報処理装置。

[請求項7] 請求項５または請求項６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8] 請求項６または請求項７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、

前記類似対象判定手段が、前記任意の２つのウェブページ間で共通の属性に係る前記属性値において等しい前記属性値がある場合に、前記任意の２つのウェブページに記載された対象は、互いに類似する対象であると判定することを特徴とする情報処理装置。

[請求項9] 請求項５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10] 請求項５から請求項９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11] 請求項１０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12] 請求項１０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13] 請求項５から請求項１２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14] 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に関連するカテゴリ関連語を記憶するカテゴリ関連語記憶ステップと、

ウェブページを取得するウェブページ取得ステップと、

前記ウェブページ内に記述されているテキストデータを抽出するテキストデータ抽出ステップと、

抽出した前記テキストデータを言語解析して、前記ウェブページの特徴語候補を抽出する特徴語候補抽出ステップと、

前記特徴語候補を検索キーワードとする検索クエリを作成し、前記

検索クエリに基づく検索結果を取得する検索結果取得ステップと、

前記検索結果のスニペット中に前記カテゴリ関連語が存在するか否かを判定する検索結果判定ステップと、

前記スニペット中に前記カテゴリ関連語が存在する場合の前記特徴語候補を前記ウェブページの特徴語に決定する特徴語決定ステップと、

を有することを特徴とする情報処理方法。

[請求項15] 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に関連するカテゴリ関連語を記憶するカテゴリ関連語記憶手段、

ウェブページを取得するウェブページ取得手段、

前記ウェブページ内に記述されているテキストデータを抽出するテキストデータ抽出手段、

抽出した前記テキストデータを言語解析して、前記ウェブページの特徴語候補を抽出する特徴語候補抽出手段、

前記特徴語候補を検索キーワードとする検索クエリを作成し、前記検索クエリに基づく検索結果を取得する検索結果取得手段、

前記カテゴリ関連語記憶手段を参照して、前記検索結果のスニペット中に前記カテゴリ関連語が存在するか否かを判定する検索結果判定手段、および、

前記スニペット中に前記カテゴリ関連語が存在する場合の前記特徴語候補を前記ウェブページの特徴語に決定する特徴語決定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16] 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に関連するカテゴリ関連語を記憶するカテゴリ関連語記憶手段、

ウェブページを取得するウェブページ取得手段、

前記ウェブページ内に記述されているテキストデータを抽出するテ

キストデータ抽出手段、

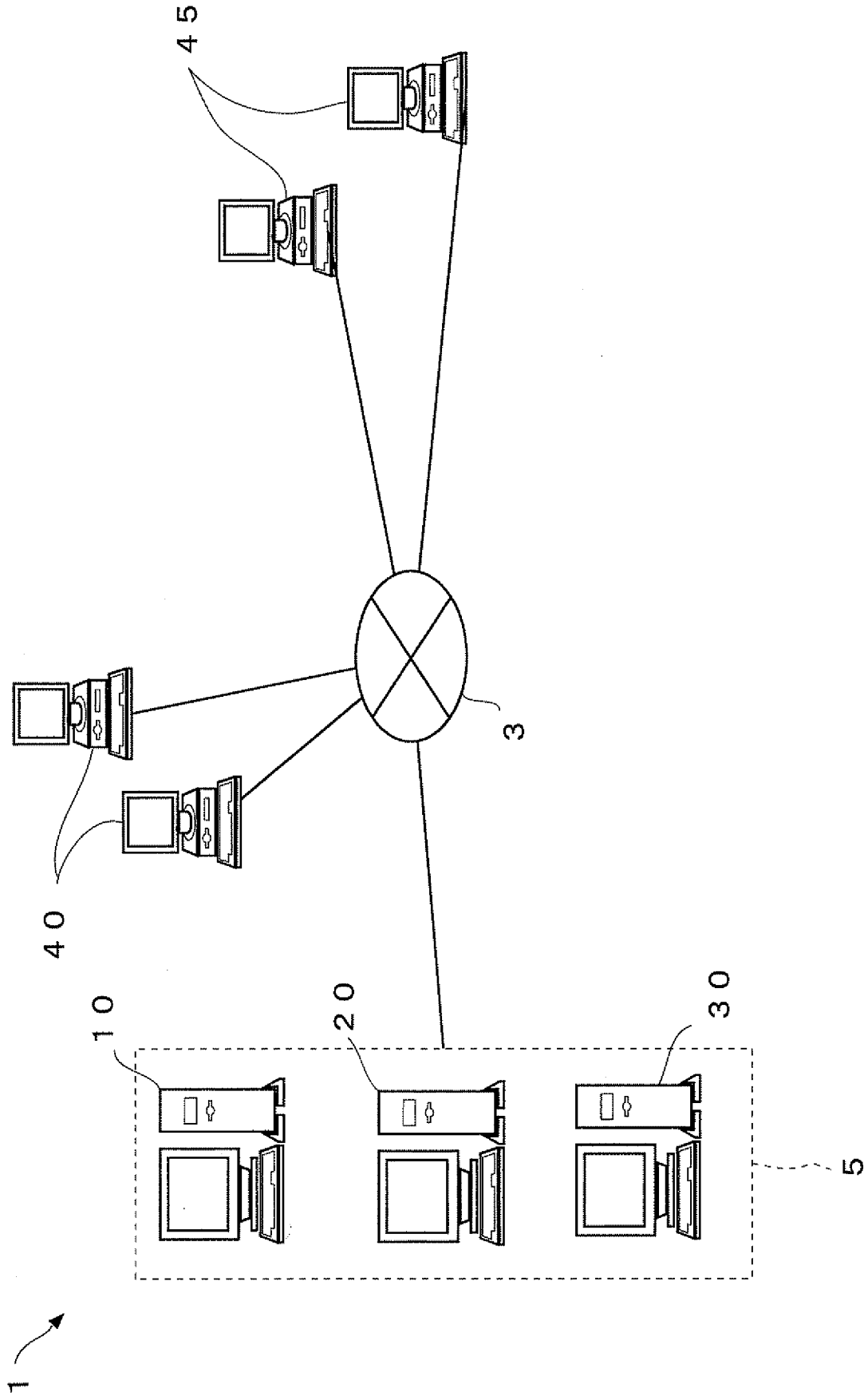
抽出した前記テキストデータを言語解析して、前記ウェブページの特徴語候補を抽出する特徴語候補抽出手段、

前記特徴語候補を検索キーワードとする検索クエリを作成し、前記検索クエリに基づく検索結果を取得する検索結果取得手段、

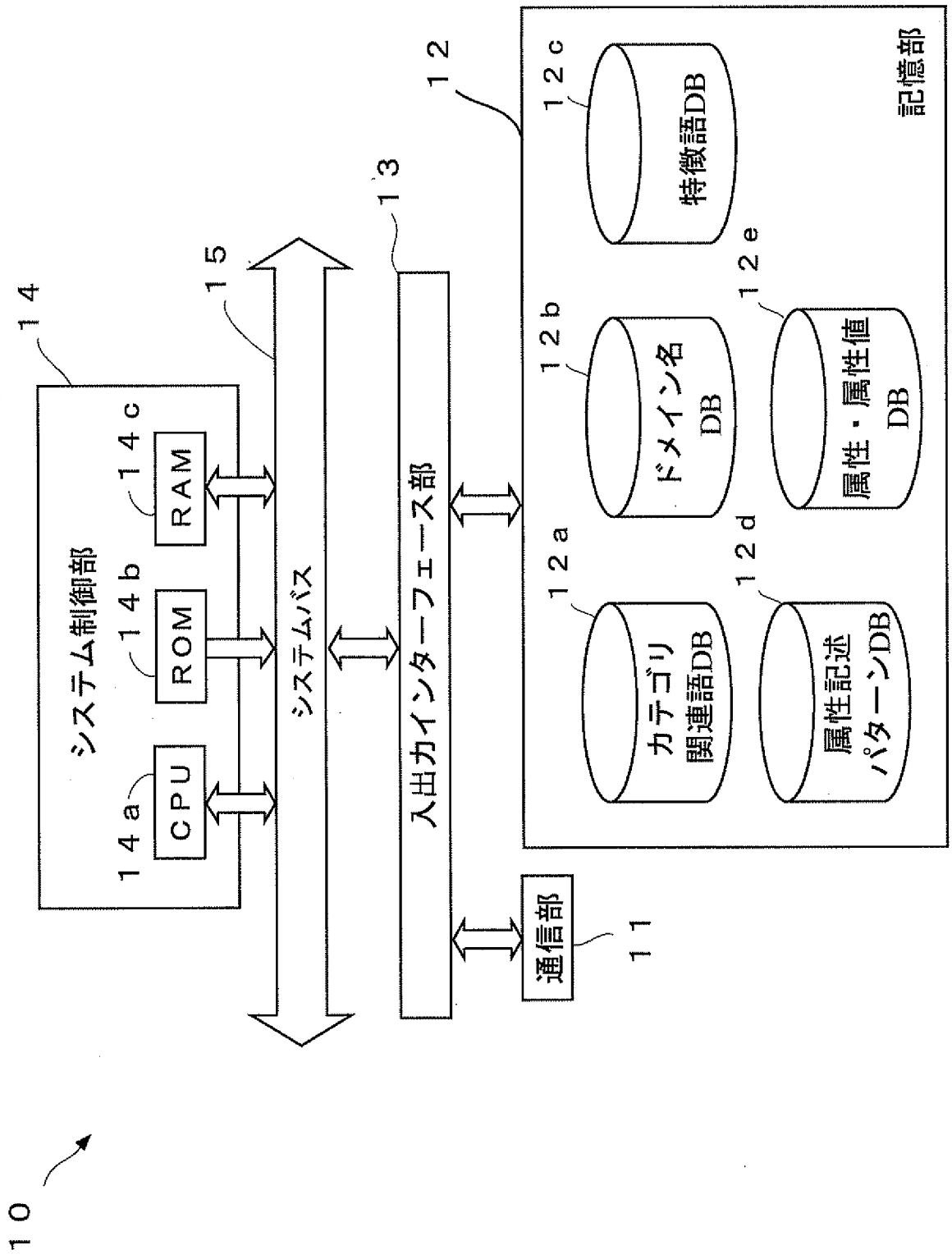
前記カテゴリ関連語記憶手段を参照して、前記検索結果のスニペット中に前記カテゴリ関連語が存在するか否かを判定する検索結果判定手段、および、

前記スニペット中に前記カテゴリ関連語が存在する場合の前記特徴語候補を前記ウェブページの特徴語に決定する特徴語決定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。

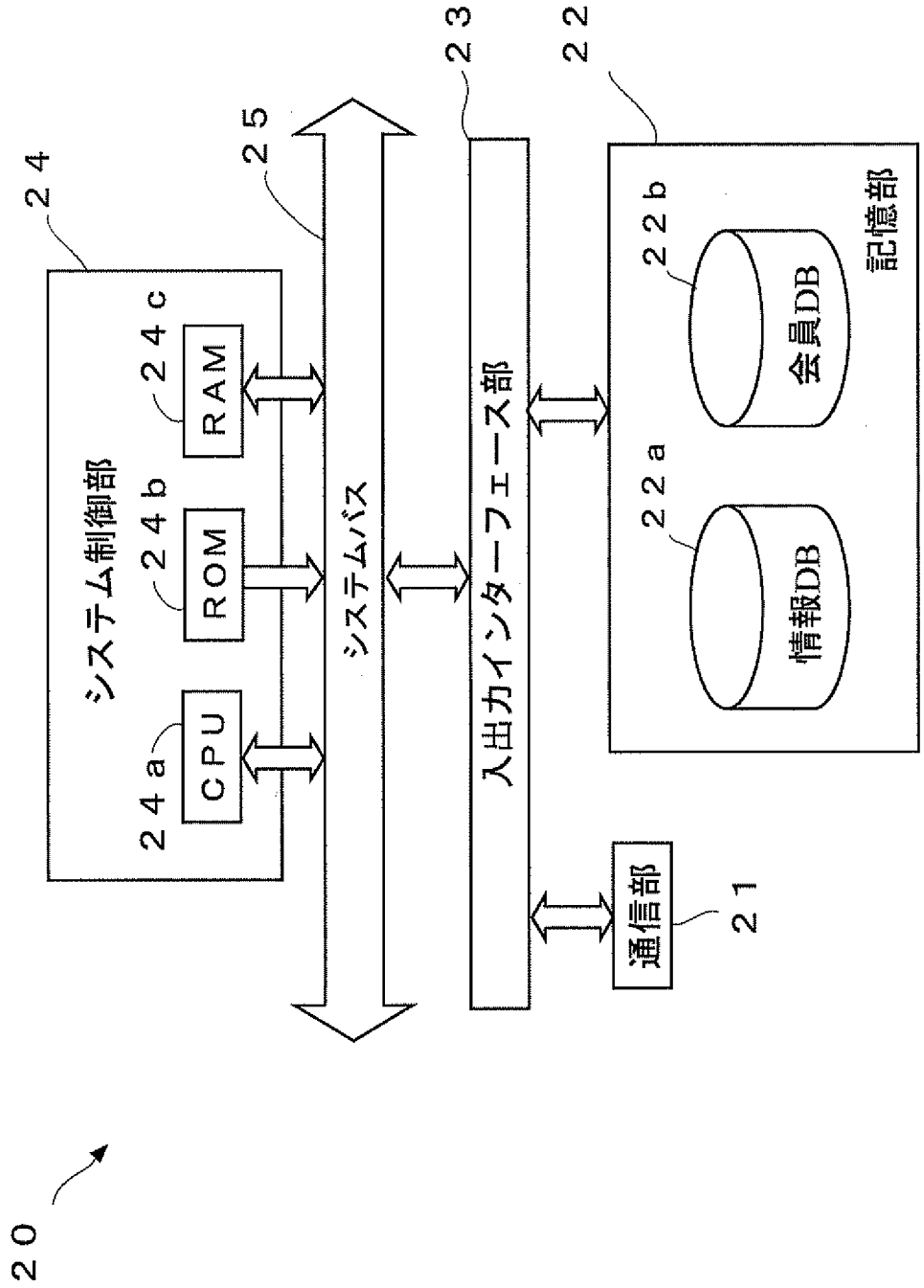
[図1]



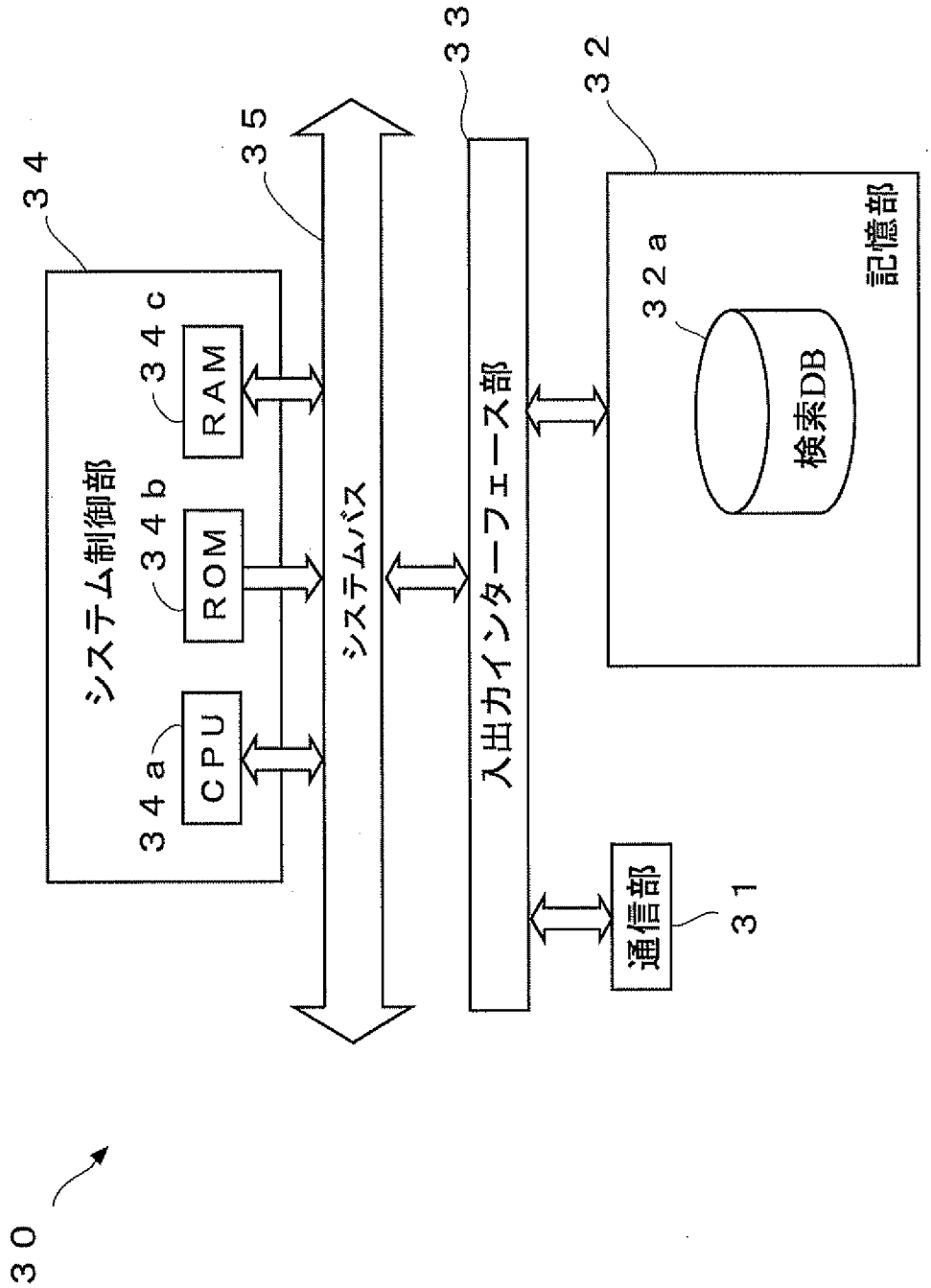
[図2]


10
システム制御部
14
CPU
14a
ROM
14b
RAM
14c
システムバス
15
入出力インターフェース部
13
通信部
11
記憶部
12
カテゴリ関連語DB
12a
ドメイン名DB
12b
特徴語DB
12c
属性記述パターンDB
12d
属性・属性値DB
12e

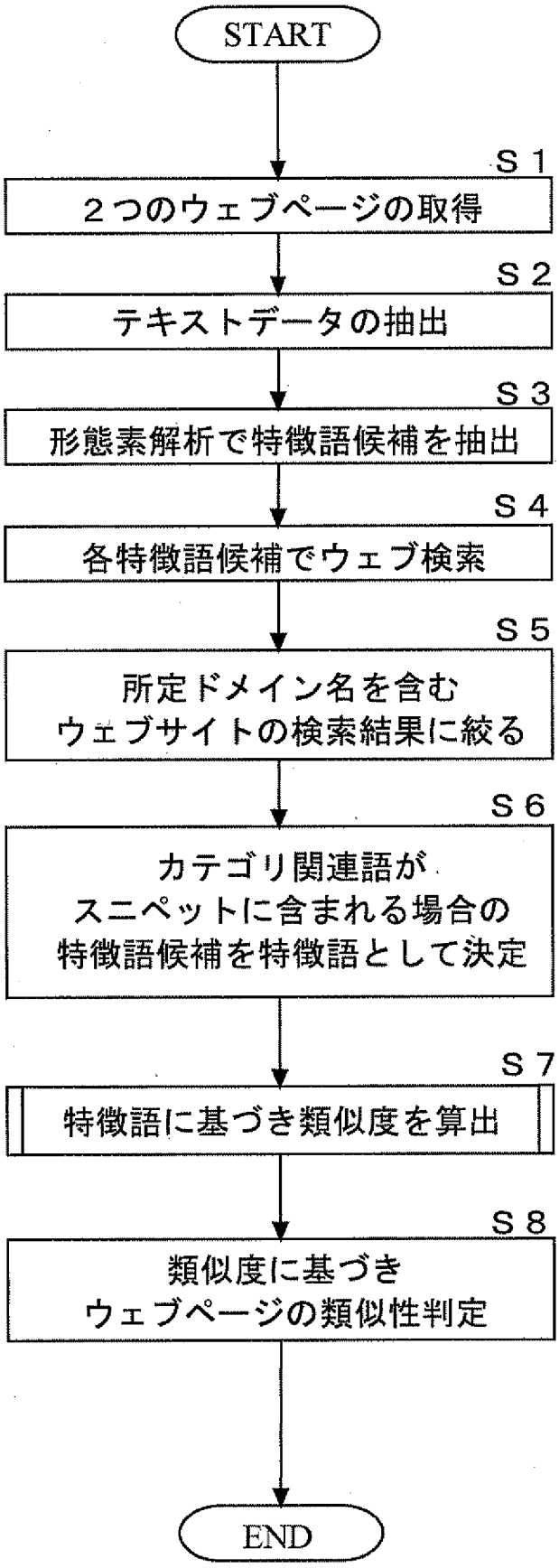
[図3]


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

[図4]


30
システム制御部
34
CPU
34a
ROM
34b
RAM
34c
システムバス
35
入出力インターフェース部
33
記憶部
32
検索DB
32a
通信部
31

[図5]


START
S 1
２つのウェブページの取得
S 2
テキストデータの抽出
S 3
形態素解析で特徴語候補を抽出
S 4
各特徴語候補でウェブ検索
S 5
所定ドメイン名を含む
ウェブサイトの検索結果に絞る
S 6
カテゴリ関連語が
スニペットに含まれる場合の
特徴語候補を特徴語として決定
S 7
特徴語に基づき類似度を算出
S 8
類似度に基づき
ウェブページの類似性判定
END

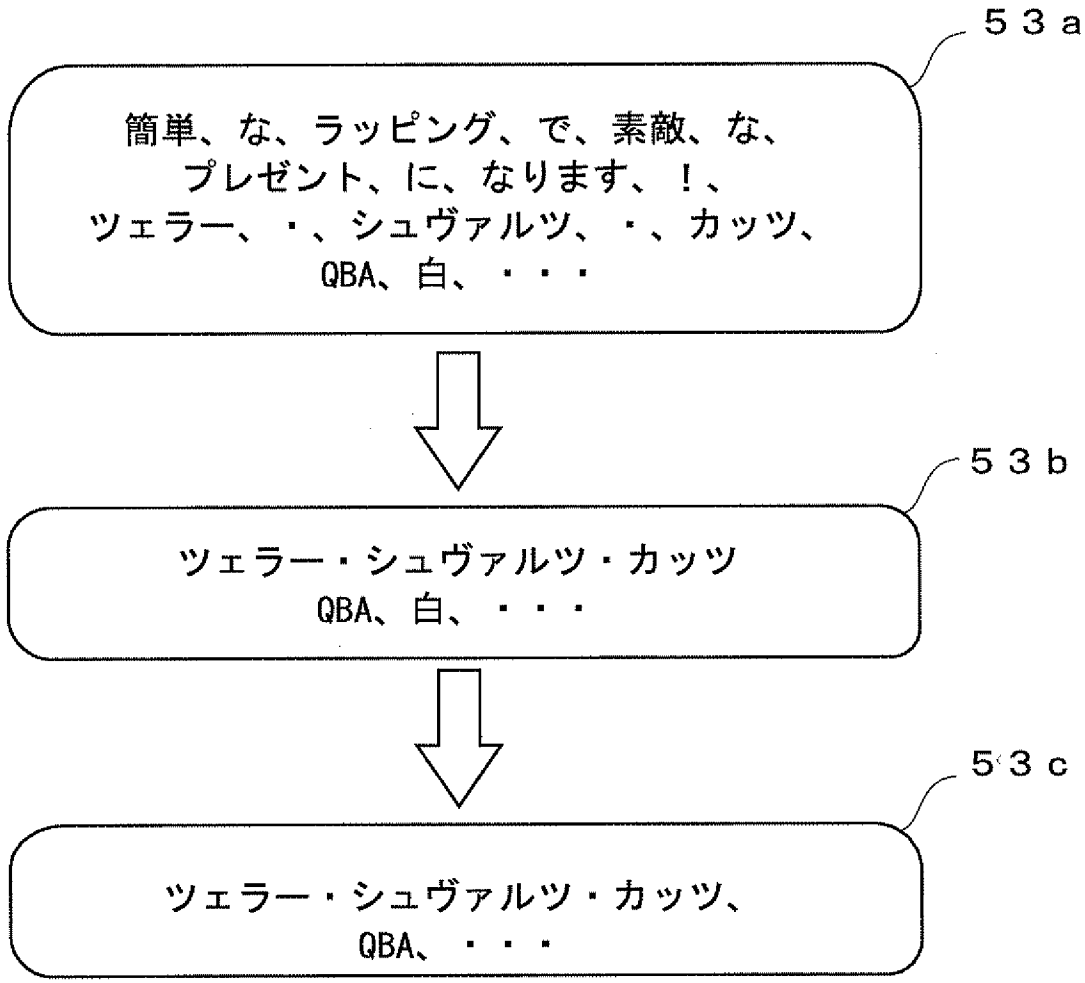
[図6]


５０
SHOP
５１
ツェ○・○○・△△
[送料無料]
当店通常価格 ○△円（税込）
＜商品詳細＞
[品種：○○○]
[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○・○○・△△・ＱＢＡ・・・・・・・・・・・
・・・・・・・・・・・・・白ワイン・・・・・・・・・・・・・
・・・・・
５４
○○
◇◇
・・・
・・・

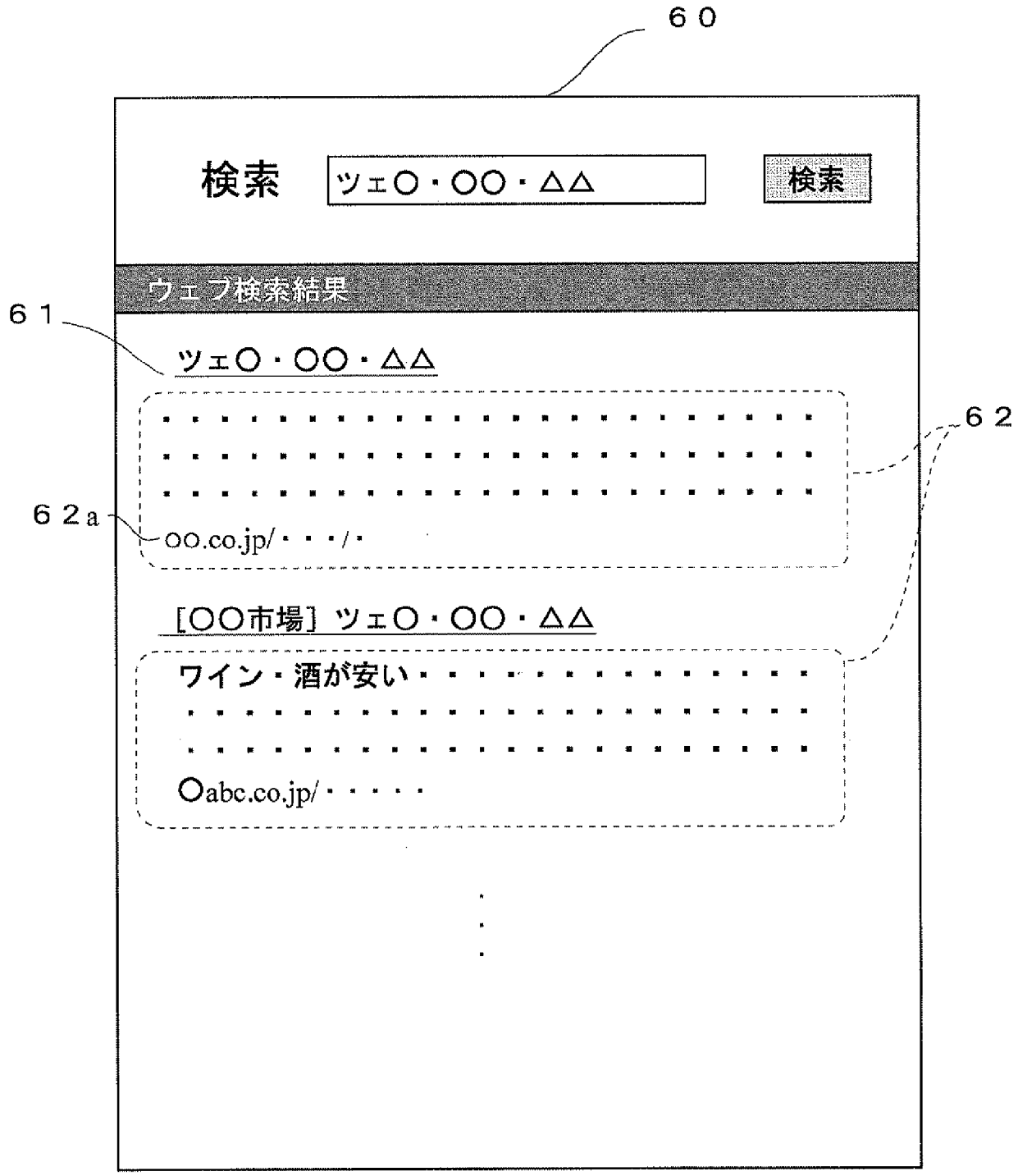
[図7]

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

[図8]


５３ａ
簡単、な、ラッピング、で、素敵、な、
プレゼント、に、なります、！、
ツェラー、・、シュヴァルツ、・、カッツ、
QBA、白、・・・
５３ｂ
ツェラー・シュヴァルツ・カッツ
QBA、白、・・・
５３ｃ
ツェラー・シュヴァルツ・カッツ、
QBA、・・・

[図9]



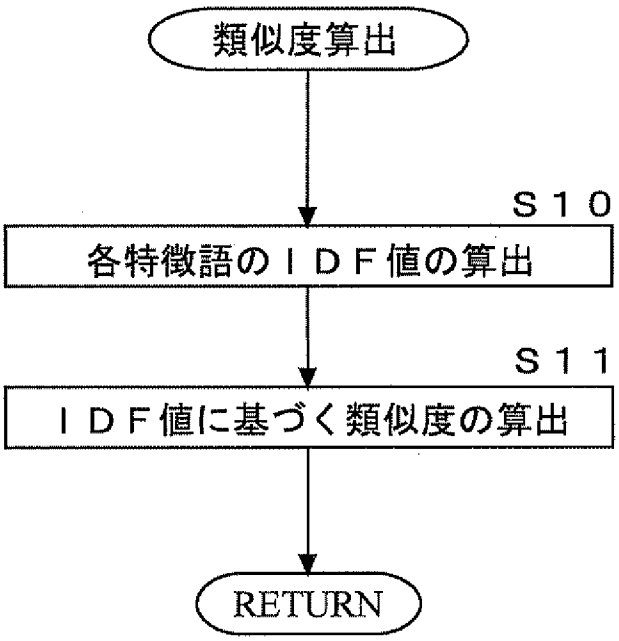
[図10]

ウェブサイト	ウェブサイトのドメイン名
○ABC	○abc.co.jp
▲▲	▲▲.com
・・・	・・・

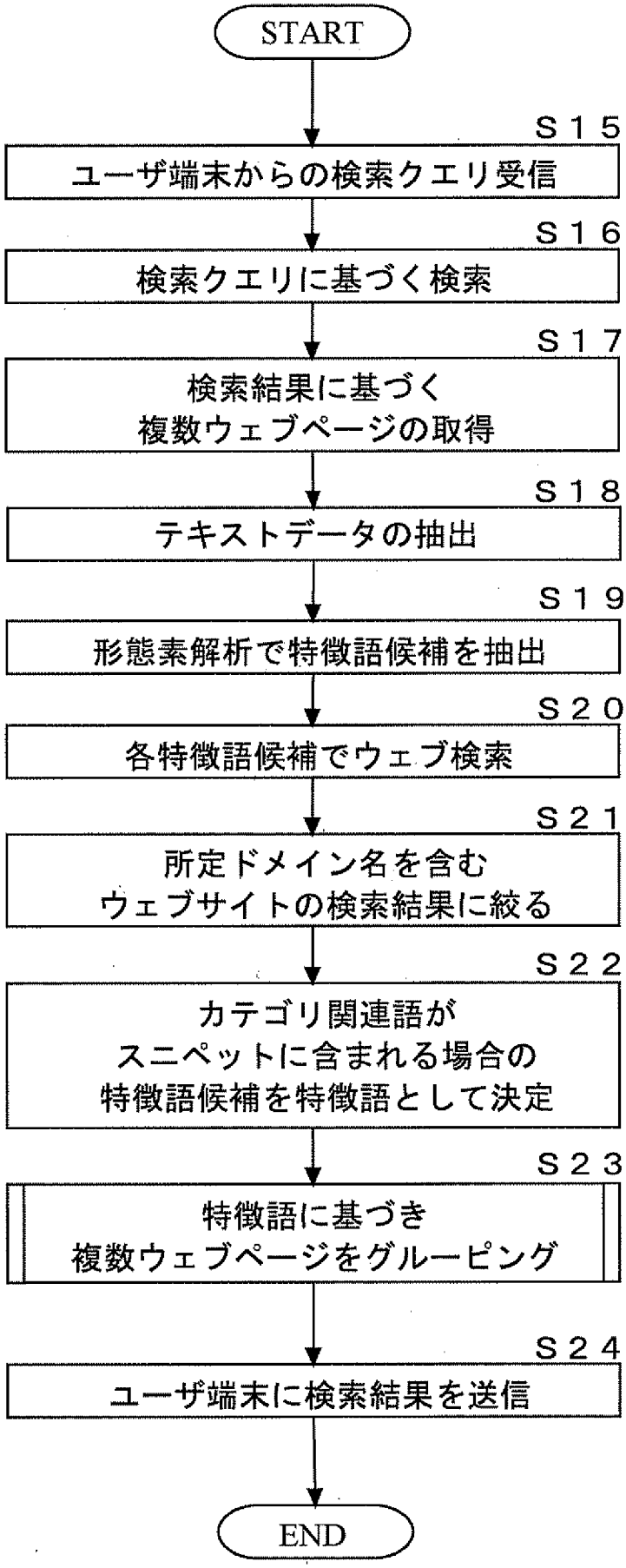
[図11]

カテゴリ	カテゴリ関連語
ワイン	ワイン、酒
ゴルフドライバー	ゴルフ、ドライバー、クラブ
男性用シューズ	シューズ、メンズ、・・・

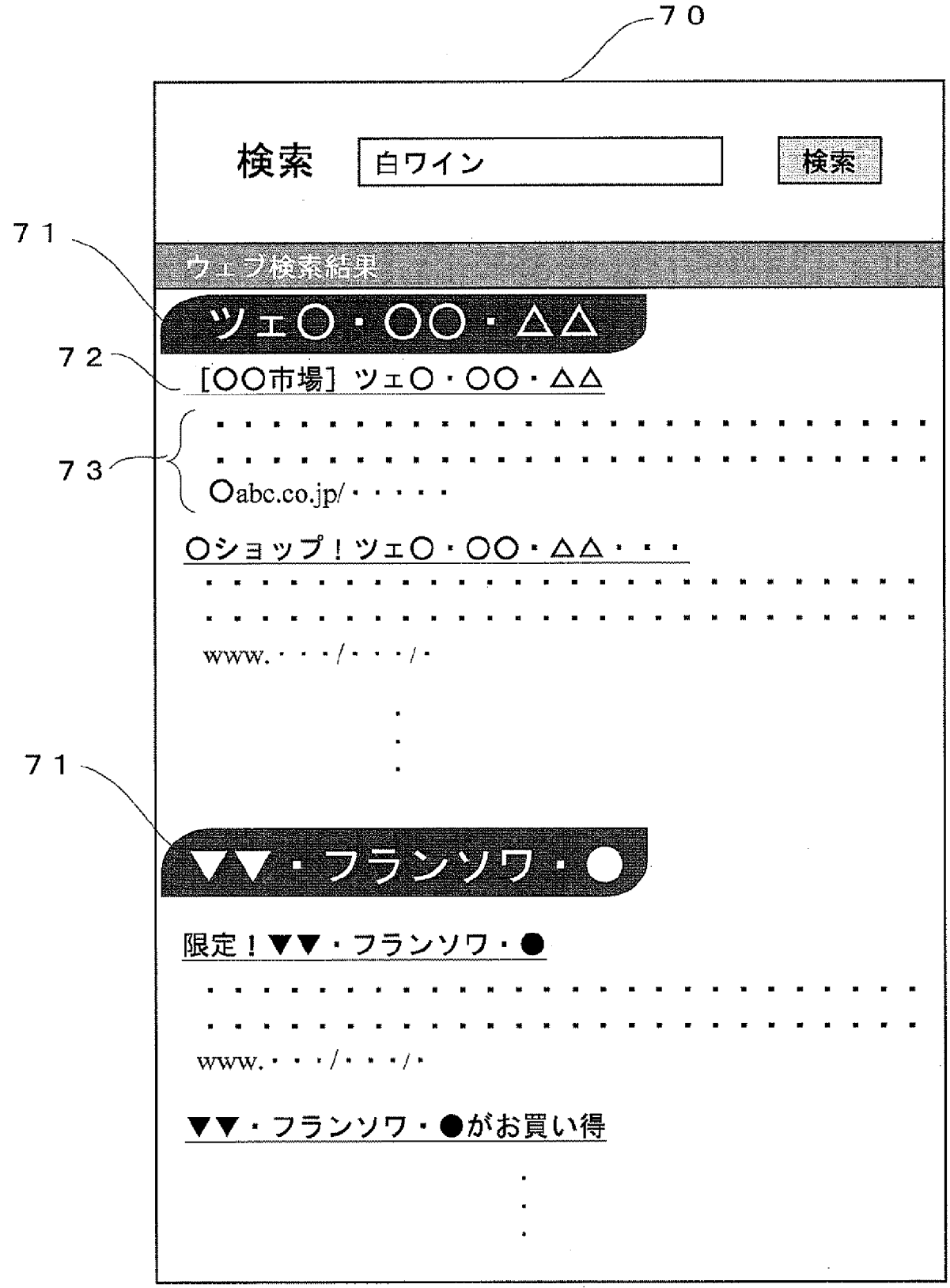
[図12]



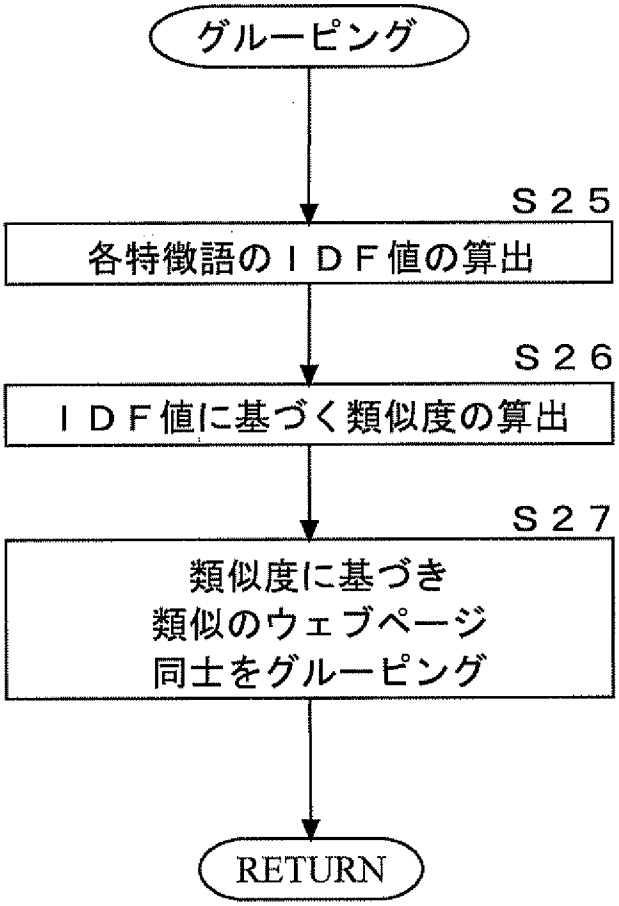
[図13]


START
S15
ユーザ端末からの検索クエリ受信
S16
検索クエリに基づく検索
S17
検索結果に基づく
複数ウェブページの取得
S18
テキストデータの抽出
S19
形態素解析で特徴語候補を抽出
S20
各特徴語候補でウェブ検索
S21
所定ドメイン名を含む
ウェブサイトの検索結果に絞る
S22
カテゴリ関連語が
スニペットに含まれる場合の
特徴語候補を特徴語として決定
S23
特徴語に基づき
複数ウェブページをグルーピング
S24
ユーザ端末に検索結果を送信
END

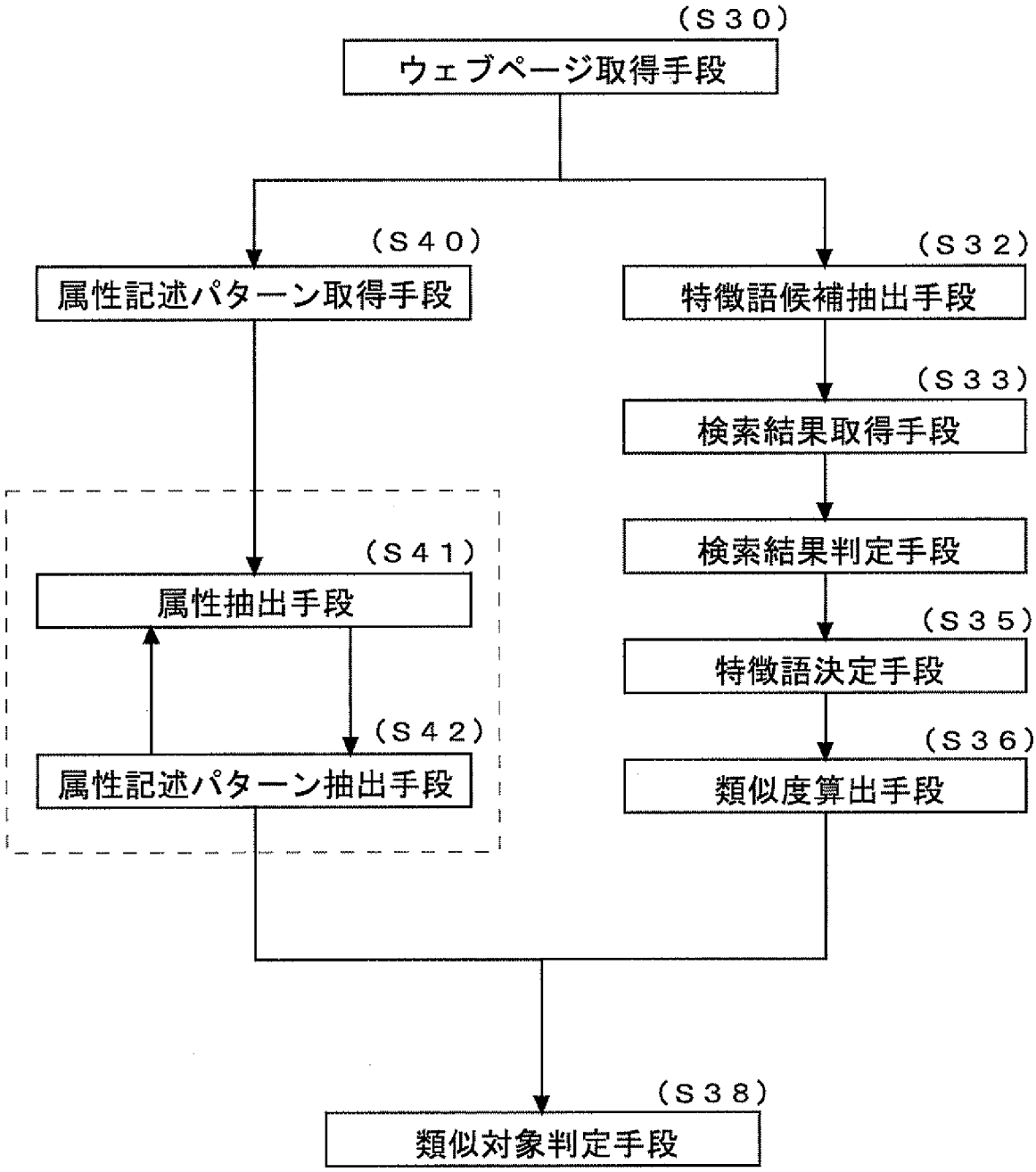
[図14]


７０
検索
白ワイン
検索
７１
ウェブ検索結果
ツェ○・○○・△△
７２
[○○市場] ツェ○・○○・△△
７３
○abc.co.jp/・・・・・
○ショップ！ツェ○・○○・△△・・・
www.・・・/・・・/・
７１
▼▼・フランソワ・●
限定！▼▼・フランソワ・●
www.・・・/・・・/・
▼▼・フランソワ・●がお買い得

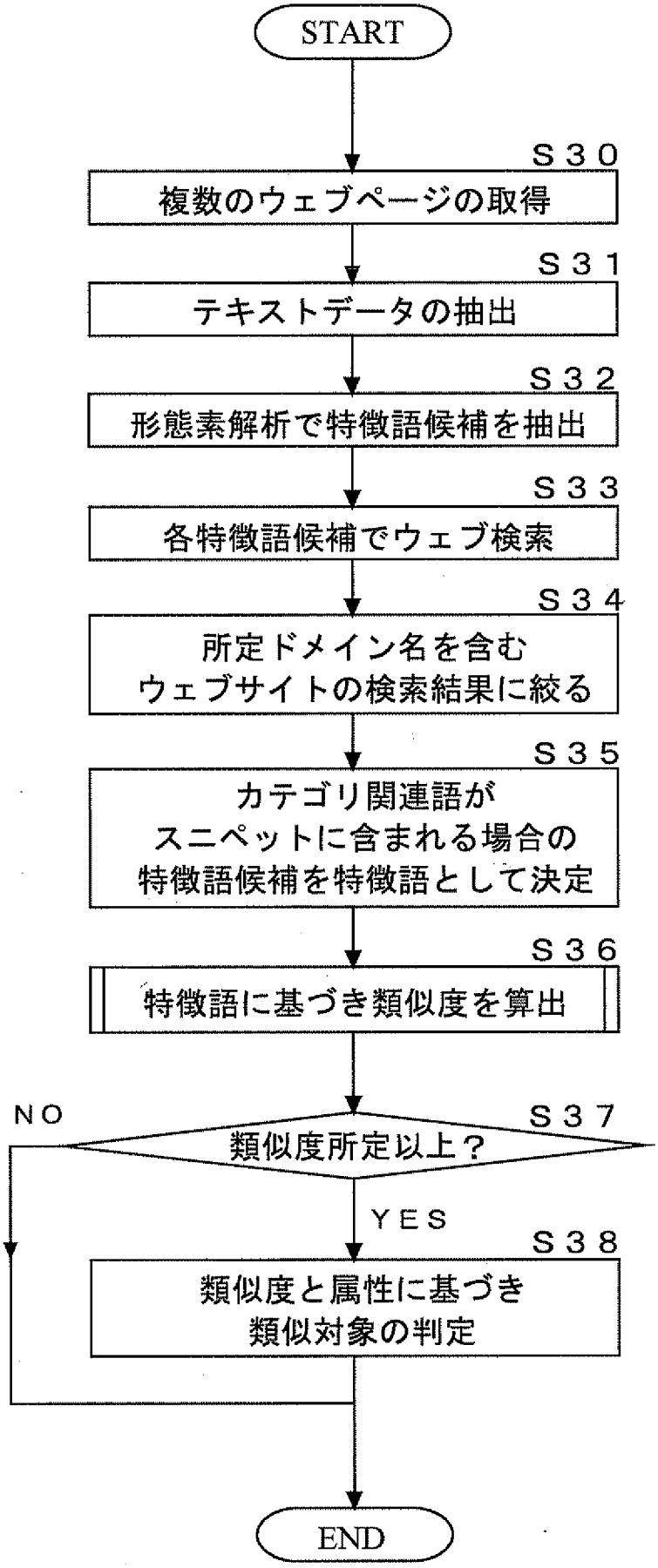
[図15]


グルーピング
S25
各特徴語のＩＤＦ値の算出
S26
ＩＤＦ値に基づく類似度の算出
S27
類似度に基づき
類似のウェブページ
同士をグルーピング
RETURN

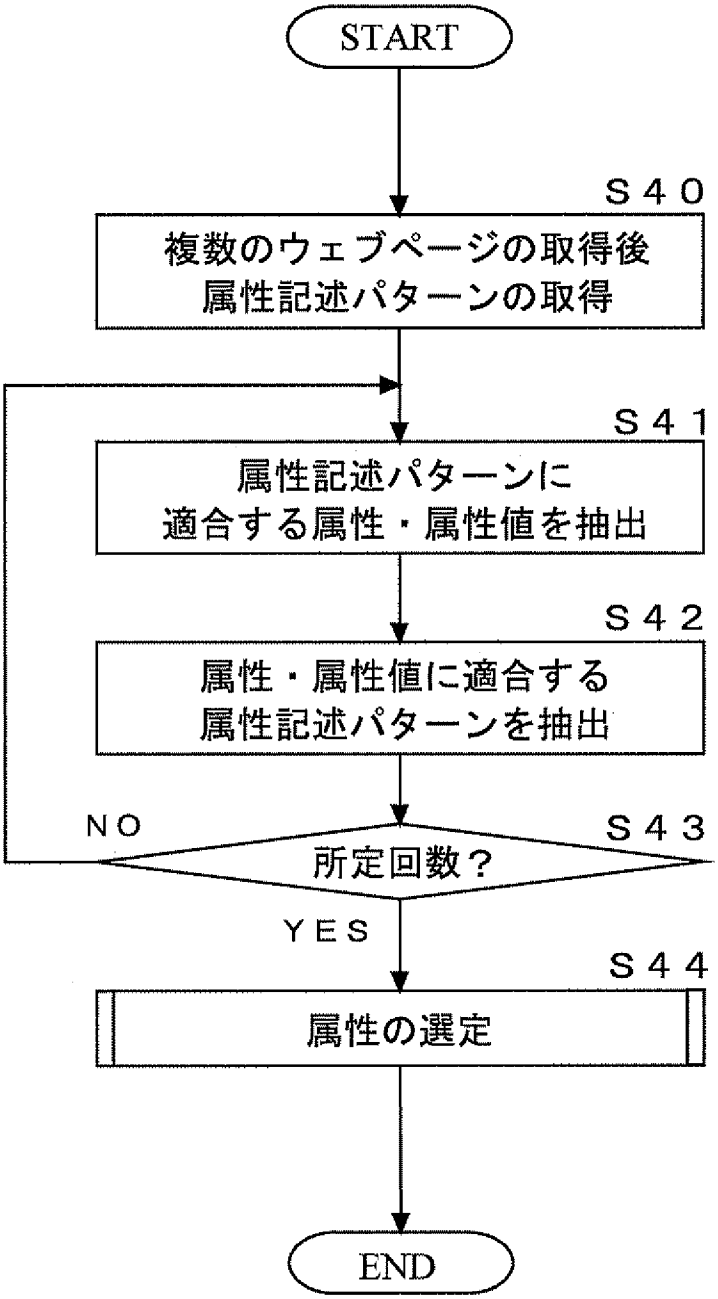
[図16]


（Ｓ３０）
ウェブページ取得手段
（Ｓ４０）
属性記述パターン取得手段
（Ｓ３２）
特徴語候補抽出手段
（Ｓ３３）
検索結果取得手段
検索結果判定手段
（Ｓ４１）
属性抽出手段
（Ｓ３５）
特徴語決定手段
（Ｓ４２）
属性記述パターン抽出手段
（Ｓ３６）
類似度算出手段
（Ｓ３８）
類似対象判定手段

[図17]


START
S30
複数のウェブページの取得
S31
テキストデータの抽出
S32
形態素解析で特徴語候補を抽出
S33
各特徴語候補でウェブ検索
S34
所定ドメイン名を含む
ウェブサイトの検索結果に絞る
S35
カテゴリ関連語が
スニペットに含まれる場合の
特徴語候補を特徴語として決定
S36
特徴語に基づき類似度を算出
S37
類似度所定以上？
NO
YES
S38
類似度と属性に基づき
類似対象の判定
END

[図18]


START
S40
複数のウェブページの取得後
属性記述パターンの取得
S41
属性記述パターンに
適合する属性・属性値を抽出
S42
属性・属性値に適合する
属性記述パターンを抽出
NO
S43
所定回数？
YES
S44
属性の選定
END

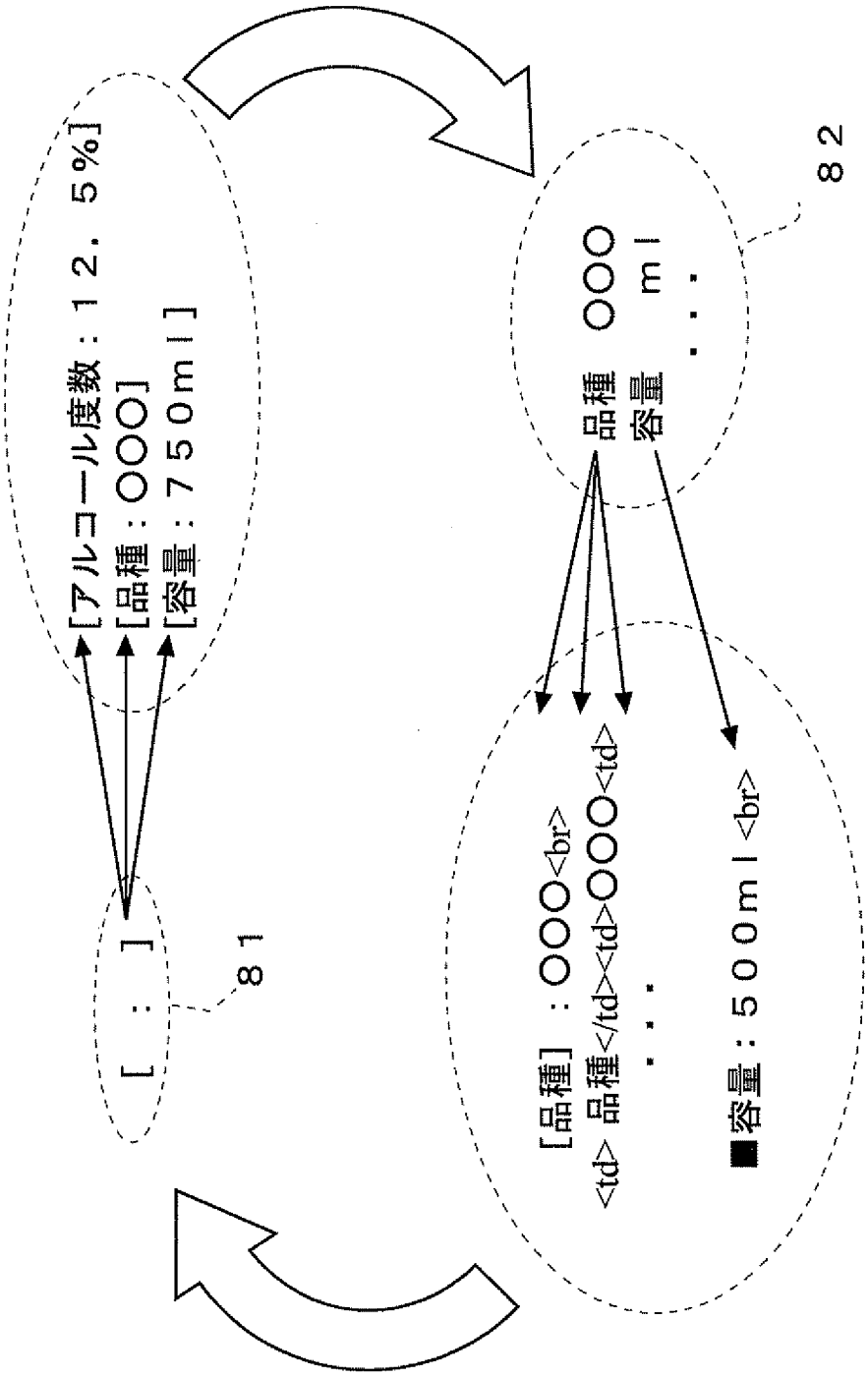
[図19]

前部	中部	後部
[	:	]
	:	

初期パターン

■	:	
<td>	</td><td>	</td>

[図20]


[アルコール度数：１２．５％]
[品種：ＯＯＯ]
[容量：７５０ｍｌ]
[　：　]
８１
８２
品種　ＯＯＯ
容量　ｍｌ
：
：
[品種]　：ＯＯＯ
<td>品種</td><td>ＯＯＯ<td>
：
：
■容量：５００ｍｌ

[図21]

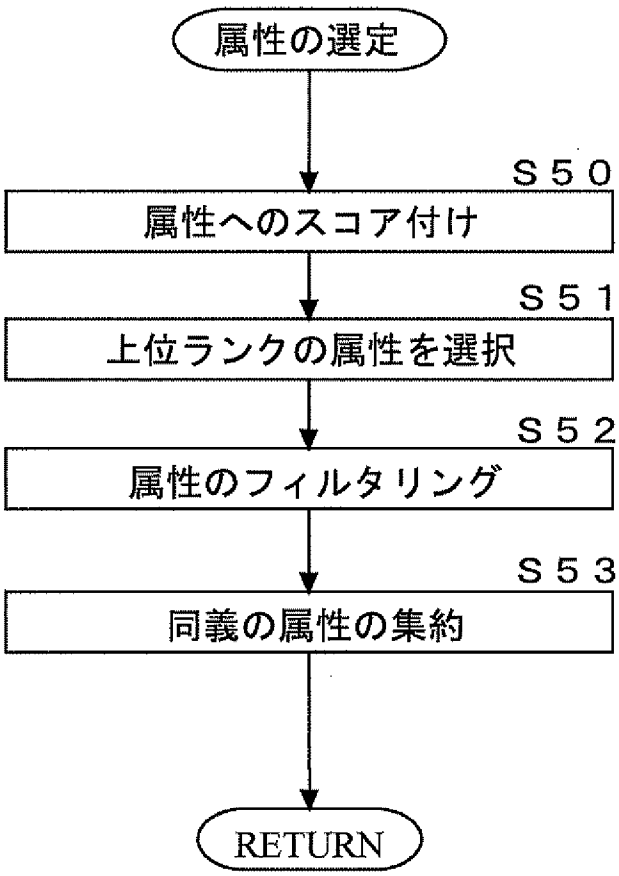
ワイン

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

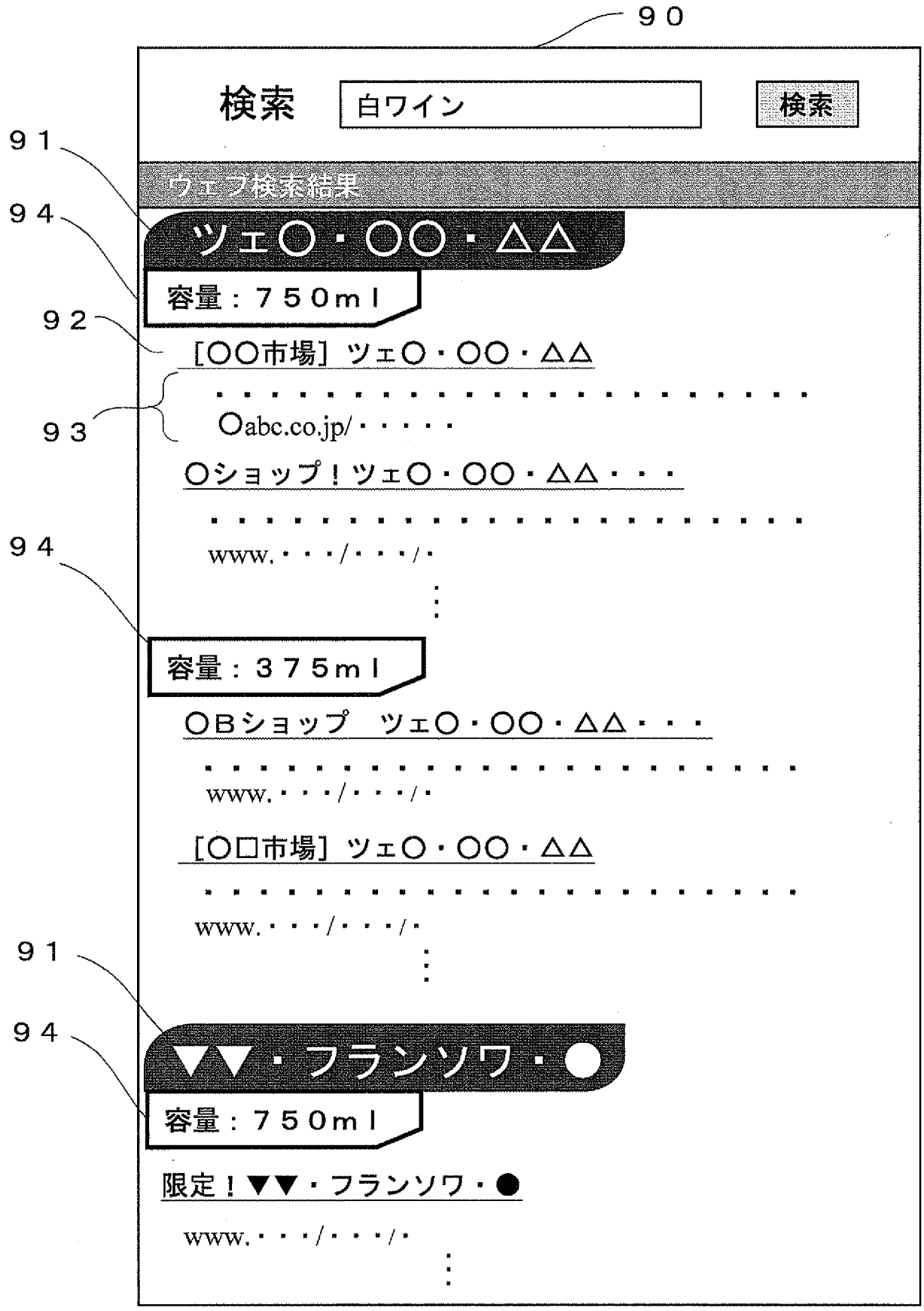
ゴルフドライバー

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

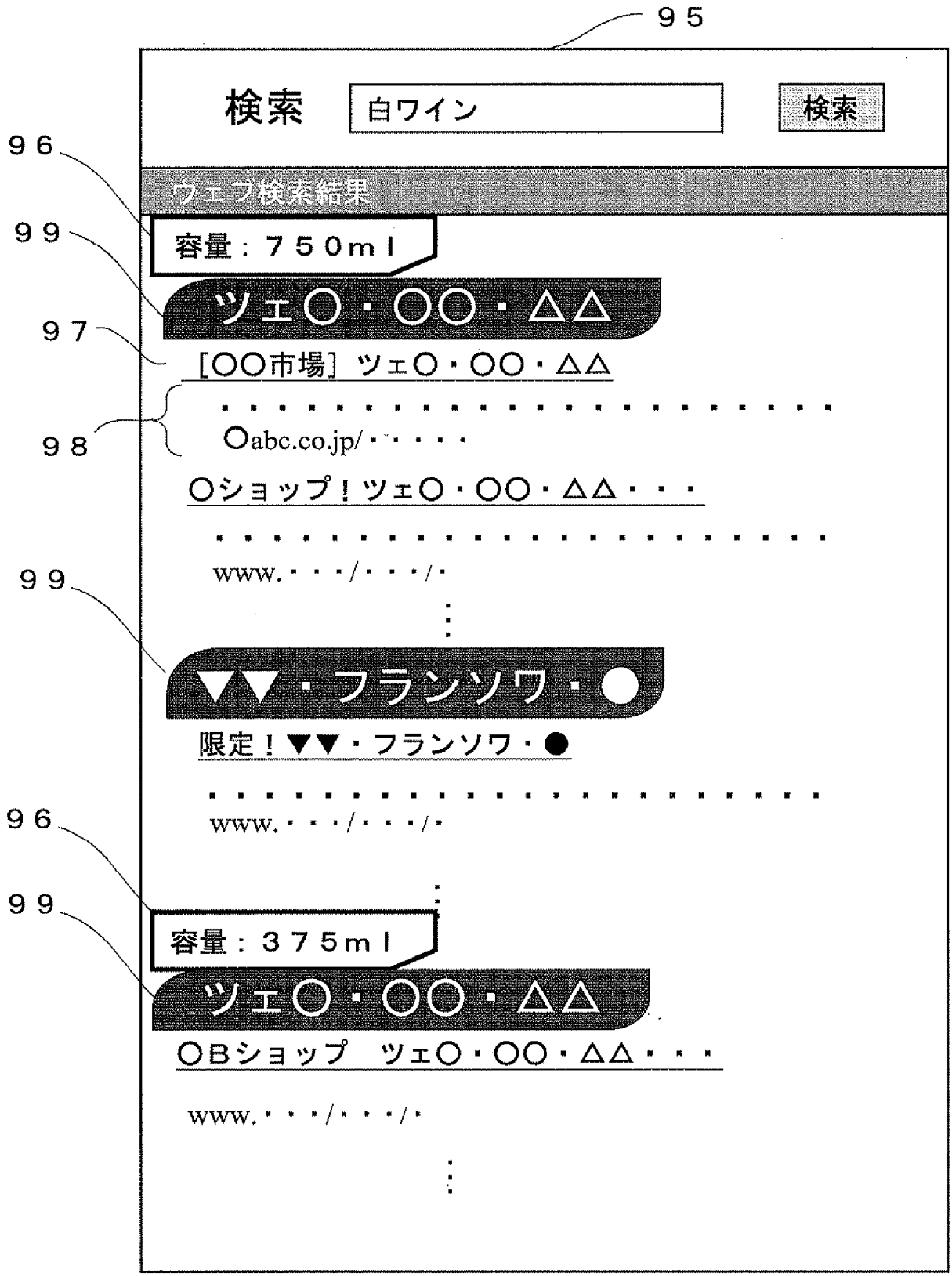
[図22]


属性の選定
S50
属性へのスコア付け
S51
上位ランクの属性を選択
S52
属性のフィルタリング
S53
同義の属性の集約
RETURN

[図23]


９０
検索
白ワイン
検索
９１
ウェブ検索結果
９４
ツェ○・○○・△△
容量：７５０ｍｌ
９２
[○○市場] ツェ○・○○・△△
９３
○abc.co.jp/・・・・・
○ショップ！ツェ○・○○・△△・・・
www.・・・/・・・/・
９４
容量：３７５ｍｌ
○Ｂショップ　ツェ○・○○・△△・・・
www.・・・/・・・/・
[○□市場] ツェ○・○○・△△
www.・・・/・・・/・
９１
９４
▼▼・フランソワ・●
容量：７５０ｍｌ
限定！▼▼・フランソワ・●
www.・・・/・・・/・

[図24]


９５
検索
白ワイン
検索
９６
ウェブ検索結果
９９
容量：７５０ｍｌ
ツェ〇・〇〇・△△
９７
［〇〇市場］ツェ〇・〇〇・△△
９８
〇abc.co.jp/・・・・・
〇ショップ！ツェ〇・〇〇・△△・・・
www.・・・/・・・/・
９９
▼▼・フランソワ・●
限定！▼▼・フランソワ・●
www.・・・/・・・/・
９６
９９
容量：３７５ｍｌ
ツェ〇・〇〇・△△
〇Ｂショップ　ツェ〇・〇〇・△△・・・
www.・・・/・・・/・

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2011/054507

A. CLASSIFICATION OF SUBJECT MATTER
G06F17/30(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

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 Y	Chikara HASHIMOTO, "Construction of Domain Dictionary for Fundamental Vocabulary and its Application to Automatic Blog Categorization with the Dynamic Estimation of Unknown Words' Domains", Journal of Natural Language Processing, 10 October 2008 (10.10.2008), vol.15, no.5, pages 73 to 97	1-3,14-16 4-13
Y	Masahiro KAKIZAKI, "Visualization of search result for improving usability", FIT2008, Dai 7 Kai Forum on Information Technology Koen Ronbunshu, separate vol.2, 20 August 2008 (20.08.2008), pages 105 to 106	4-13

☒ 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
16 March, 2011 (16.03.11)	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/JP2011/054507

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2008-269106 A (Osaka Industrial Promotion Organization), 06 November 2008 (06.11.2008), entire text; all drawings (Family: none)	5-13
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	1-16
P,X	Akio KOBAYASHI, "Shopping Site no Shohin Page Title Karano Shohin Kanren Yogo no Chushutsu to Shohin Catalogue eno Shohin Page no Himo Tsuke Shuho", Proceedings of the 16th annual meeting of the Association for Natural Language Processing, 08 March 2010 (08.03.2010), pages 367 to 370	1-3,14-16

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１１／０５４５０７

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

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

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 Y	橋本力，基本語ドメイン辞書の構築と未知語ドメイン推定を用いたブログ自動分類法への応用，自然言語処理，2008.10.10，第１５巻，第５号，ｐｐ．７３－９７	1-3，14-16 4-13
Y	垣崎正宏，ユーザビリティ向上を目指した検索結果の視覚化，ＦＩＴ２００８　第７回情報科学技術フォーラム　講演論文集　第２分冊，2008.08.20，ｐｐ．１０５－１０６	4-13

☑　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
１６．０３．２０１１	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
Y	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)	5-13
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	1-16
P, X	小林暁雄, ショッピングサイトの商品ページタイトルからの商品関連用語の抽出と商品カタログへの商品ページの紐付け手法, 言語処理学会第１６回年次大会発表論文集, 2010.03.08, ｐｐ．３６７－３７０	1-3, 14-16