(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4547500号
(P4547500)
(45)発行日　平成22年9月22日(2010.9.22)　(24)登録日　平成22年7月16日(2010.7.16)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 210D
G06F 17/30 370Z
G06F 17/30 320D

請求項の数 4　(全 17 頁)

(21)出願番号　特願2006-199312(P2006-199312)
(22)出願日　平成18年7月21日(2006.7.21)
(65)公開番号　特開2008-27207(P2008-27207A)
(43)公開日　平成20年2月7日(2008.2.7)
審査請求日　平成19年6月1日(2007.6.1)

特許法第３０条第１項適用　２００６年３月２２日～２３日　社団法人　電子情報通信学会主催の「電子情報通信学会第二種研究会」において文書をもって発表

(73)特許権者　504145364
国立大学法人群馬大学
群馬県前橋市荒牧町四丁目２番地
(74)代理人　100079049
弁理士　中島　淳
(74)代理人　100084995
弁理士　加藤　和詳
(74)代理人　100085279
弁理士　西元　勝一
(74)代理人　100099025
弁理士　福田　浩志
(72)発明者　安川　美智子
群馬県桐生市相生町１丁目２４７－５　桜木ハイツＡ棟３０１号

最終頁に続く

(54)【発明の名称】検索装置及びプログラム

(57)【特許請求の範囲】
【請求項１】
複数の文書データを記憶した文書データベースから、検索語に適合する複数の文書データを取得する文書データ取得手段と、
前記文書データ取得手段によって取得された複数の文書データの各々を形態素解析することによって得られた単語に基づいて、前記文書データの各々について、前記検索語に関連する複数の関連語の各々の出現頻度を算出する頻度算出手段と、
前記頻度算出手段によって算出された前記複数の関連語の各々の出現頻度に基づいて、各関連語同士の類似度を算出する類似度算出手段と、
前記複数の関連語のクラスタリングを行って、前記類似度算出手段によって算出された類似度が高い組み合わせから前記関連語を組み合わせて、所定数の関連語クラスタを生成するクラスタリング手段と、
前記クラスタリング手段によって生成された関連語クラスタ毎に、前記関連語の出現頻度に基づいて、前記文書データ取得手段によって取得された複数の文書データのうち、該関連語クラスタの関連語によって特徴付けられる文書データを該関連語クラスタに対応付ける対応付け手段と、
前記クラスタリング手段によって生成された関連語クラスタ及び該関連語クラスタに対応付けられた文書データを示す文書データ情報を、前記検索語と同時に検索される回数が多い関連語を含む関連語クラスタから順番に、前記検索語に適合する文書データの検索結果として表示する表示手段と、

を含む検索装置。
【請求項２】
少なくとも１つの検索語からなる検索クエリを複数記憶したデータベースに基づいて、前記文書データ取得手段における検索語と同時に検索語となる単語を、前記関連語として複数取得する関連語取得手段を更に含み、
前記頻度算出手段は、前記文書データの各々について、前記関連語取得手段によって取得された複数の関連語の出現頻度を算出する請求項１記載の検索装置。
【請求項３】
少なくとも１つの検索語からなる検索クエリを複数記憶したデータベースに基づいて、前記文書データ取得手段における検索語の類義語と同時に検索語となる単語を、前記関連語として複数取得する関連語取得手段を更に含み、
前記頻度算出手段は、前記文書データの各々について、前記関連語取得手段によって取得された複数の関連語の出現頻度を算出する請求項１記載の検索装置。
【請求項４】
コンピュータを、
複数の文書データを記憶した文書データベースから、検索語に適合する複数の文書データを取得する文書データ取得手段、
前記文書データ取得手段によって取得された複数の文書データの各々を形態素解析することによって得られた単語に基づいて、前記文書データの各々について、前記検索語に関連する複数の関連語の各々の出現頻度を算出する頻度算出手段、
前記頻度算出手段によって算出された前記複数の関連語の各々の出現頻度に基づいて、各関連語同士の類似度を算出する類似度算出手段、
前記複数の関連語のクラスタリングを行って、前記類似度算出手段によって算出された類似度が高い組み合わせから前記関連語を組み合わせて、所定数の関連語クラスタを生成するクラスタリング手段、
前記クラスタリング手段によって生成された関連語クラスタ毎に、前記関連語の出現頻度に基づいて、前記文書データ取得手段によって取得された複数の文書データのうち、該関連語クラスタの関連語によって特徴付けられる文書データを該関連語クラスタに対応付ける対応付け手段、及び
前記クラスタリング手段によって生成された関連語クラスタ及び該関連語クラスタに対応付けられた文書データを示す文書データ情報を、前記検索語と同時に検索される回数が多い関連語を含む関連語クラスタから順番に、前記検索語に適合する文書データの検索結果として表示する表示手段
として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、検索装置及びプログラムにかかり、特に、検索エンジンによって文書データを検索する検索装置及びプログラムに関する。
【背景技術】
【０００２】
従来より、Ｗｅｂ検索エンジンを用いて、様々な検索が行われるようになっている。流行している物や現象、人、企業、商品、サービス、テレビ番組などについての情報を検索する際に、検索対象についてあまり詳しく知らないため、適切な関連語で検索結果を絞り込む事が容易でない場合がある。
【０００３】
また、検索対象についてある程度知っている場合であっても検索語で検索される膨大な検索結果を全て閲覧するのではなく、興味のあるページ群だけ概観したいという場合がある。
【０００４】

一般に検索対象となる文書集合の中には類似した文書が含まれることが多いことから、予め文書集合を類似度に応じてグループ化（クラスタリング）しておき、検索時にはこれらのグループ（クラスタ）と検索質問（検索クエリ）との適合度を計算するクラスタ型の検索が知られている（非特許文献１）。ある検索語で検索される検索結果Ｗｅｂページ群には、多数の類似したＷｅｂページが含まれるため、適切なクラスタリングを行うことで、検索結果を絞り込むことや、検索結果を概観することが容易になる。

【非特許文献１】徳永健伸、「情報検索と言語処理」、東京大学出版会、（１９９９）

【発明の開示】

【発明が解決しようとする課題】

【０００５】

しかしながら、上記の非特許文献１記載の技術では、検索結果Ｗｅｂページ群をＷｅｂページでクラスタリングすると、Ｗｅｂページ群の中に、ユーザの検索ニーズに合致しない雑多な情報が多数含まれているため、ユーザにとって意味が分からないクラスタや、検索対象を絞り込む上で役に立たないクラスタが生成されてしまうため、クラスタリングされた検索結果が、ユーザにとって分かりにくく、利便性が低いものとなってしまう、という問題がある。

【０００６】

本発明は、上記の問題点を解決するためになされたもので、ユーザにとって分かりやすいクラスタにより検索結果を表示することができる検索装置及びプログラム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】

【０００７】

上記の目的を達成するために本発明に係る検索装置は、複数の文書データを記憶した文書データベースから、検索語に適合する複数の文書データを取得する文書データ取得手段と、前記文書データ取得手段によって取得された複数の文書データの各々を形態素解析することによって得られた単語に基づいて、前記文書データの各々について、前記検索語に関連する複数の関連語の各々の出現頻度を算出する頻度算出手段と、前記頻度算出手段によって算出された前記複数の関連語の各々の出現頻度に基づいて、各関連語同士の類似度を算出する類似度算出手段と、前記複数の関連語のクラスタリングを行って、前記類似度算出手段によって算出された類似度が高い組み合わせから前記関連語を組み合わせて、所定数の関連語クラスタを生成するクラスタリング手段と、前記クラスタリング手段によって生成された関連語クラスタ毎に、前記関連語の出現頻度に基づいて、前記文書データ取得手段によって取得された複数の文書データのうち、該関連語クラスタの関連語によって特徴付けられる文書データを該関連語クラスタに対応付ける対応付け手段と、前記クラスタリング手段によって生成された関連語クラスタ及び該関連語クラスタに対応付けられた文書データを示す文書データ情報を、前記検索語と同時に検索される回数が多い関連語を含む関連語クラスタから順番に、前記検索語に適合する文書データの検索結果として表示する表示手段とを含んで構成されている。

【０００８】

また、本発明に係るプログラムは、コンピュータを、複数の文書データを記憶した文書データベースから、検索語に適合する複数の文書データを取得する文書データ取得手段、前記文書データ取得手段によって取得された複数の文書データの各々を形態素解析することによって得られた単語に基づいて、前記文書データの各々について、前記検索語に関連する複数の関連語の各々の出現頻度を算出する頻度算出手段、前記頻度算出手段によって算出された前記複数の関連語の各々の出現頻度に基づいて、各関連語同士の類似度を算出する類似度算出手段、前記複数の関連語のクラスタリングを行って、前記類似度算出手段によって算出された類似度が高い組み合わせから前記関連語を組み合わせて、所定数の関連語クラスタを生成するクラスタリング手段、前記クラスタリング手段によって生成された関連語クラスタ毎に、前記関連語の出現頻度に基づいて、前記文書データ取得手段によって取得された複数の文書データのうち、該関連語クラスタの関連語によって特徴付けら

れる文書データを該関連語クラスタに対応付ける対応付け手段、及び前記クラスタリング手段によって生成された関連語クラスタ及び該関連語クラスタに対応付けられた文書データを示す文書データ情報を、前記検索語と同時に検索される回数が多い関連語を含む関連語クラスタから順番に、前記検索語に適合する文書データの検索結果として表示する表示手段として機能させるためのプログラムである。

【０００９】

本発明によれば、複数の文書データを記憶した文書データベースから、検索語に適合する複数の文書データを取得し、取得された複数の文書データの各々を形態素解析して、文書データの単語を得る。そして、得られた単語に基づいて、文書データの各々について、検索語に関連する複数の関連語の各々の出現頻度を算出し、算出された複数の関連語の各々の出現頻度に基づいて、各関連語同士の類似度を算出する。

【００１０】

そして、複数の関連語のクラスタリングを行って、算出された類似度が高い組み合わせから関連語を組み合わせて、所定数の関連語クラスタを生成し、生成された関連語クラスタを、検索語に適合する文書データの検索結果として表示する。

【００１１】

従って、検索語に適合する各文書データにおける検索語に関連する複数の関連語の各々の出現頻度に基づいて関連語をクラスタリングした結果を、検索結果として表示することにより、ユーザによって入力される検索語に関係のない単語を除外して生成した関連語クラスタを検索結果として表示するため、ユーザにとって分かりやすいクラスタにより検索結果を表示することができる。

【００１２】

ここで、検索語に関連する関連語とは、検索エンジンにユーザが検索語と同時に入力した単語である。

【００１３】

また、本発明に係る検索装置は、クラスタリング手段によって生成された関連語クラスタ毎に、関連語の出現頻度に基づいて、文書データ取得手段によって取得された複数の文書データのうち、関連語クラスタの関連語によって特徴付けられる文書データを関連語クラスタに対応付ける対応付け手段を更に含み、表示手段は、関連語クラスタ及び関連語クラスタに対応付けられた文書データを示す文書データ情報を、検索結果として表示することができる。これにより、検索語に適合する文書データを関連語クラスタに対応付けて表示するため、検索結果の表示におけるユーザの利便性を向上することができる。

【００１４】

また、本発明に係る検索装置は、少なくとも１つの検索語からなる検索クエリを複数記憶したデータベースに基づいて、文書データ取得手段における検索語と同時に検索語となる単語を、関連語として複数取得する関連語取得手段を更に含み、頻度算出手段は、文書データの各々について、関連語取得手段によって取得された複数の関連語の出現頻度を算出することができる。これにより、検索クエリのログを記憶したデータベースから、検索語に関連する関連語を複数取得することができる。

【００１５】

また、本発明に係る検索装置は、少なくとも１つの検索語からなる検索クエリを複数記憶したデータベースに基づいて、文書データ取得手段における検索語の類義語と同時に検索語となる単語を、関連語として複数取得する関連語取得手段を更に含み、頻度算出手段は、文書データの各々について、関連語取得手段によって取得された複数の関連語の出現頻度を算出することができる。これにより、検索クエリのログを記憶したデータベースから、検索語の類義語に関連する関連語を複数取得することができる。

【００１６】

また、本発明に係る表示手段は、検索語と同時に検索される回数が多い関連語を含む関連語クラスタから順番に、検索結果として表示することができる。これにより、検索語との関連が強い関連語を含む関連語クラスタを先に表示することにより、ユーザの検索ニー

ズに合致することができ、検索結果の表示におけるユーザの利便性を向上させることができる。

【発明の効果】

【００１７】

以上説明したように、本発明の検索装置及びプログラムによれば、検索語に適合する各文書データにおける検索語に関連する複数の関連語の各々の出現頻度に基づいて関連語をクラスタリングした結果を、検索結果として表示することにより、ユーザによって入力される検索語に関係のない単語を除外して生成した関連語クラスタを検索結果として表示するため、ユーザにとって分かりやすいクラスタにより検索結果を表示することができる、という効果が得られる。

【発明を実施するための最良の形態】

【００１８】

以下、図面を参照して本実施の形態を詳細に説明する。なお、本実施の形態では、複数の検索エンジンを一括検索（メタサーチ）する検索装置に本発明を適用した場合について説明する。

【００１９】

図１に示すように、第１の実施の形態に係る検索システム１０は、複数の検索クエリから構成される検索クエリのログを記憶した検索クエリログデータベース１２と、少なくとも１つの検索語からなる検索クエリに対応して、Ｗｅｂページを検索する複数のＷｅｂ検索エンジン１４、及び複数のＷｅｂ検索エンジン１４によって検索されたＷｅｂページのキャッシュデータを一時的に記憶するキャッシュデータデータベース１６に接続され、かつ複数のＷｅｂ検索エンジン１４を一括検索（メタサーチ）するメタサーチエンジンを実現するメタサーチエンジンプログラムを記憶したコンピュータ１８とを備えている。

【００２０】

検索クエリログデータベース１２には、１つ以上の検索語からなる検索クエリが複数記憶されている。また、複数のＷｅｂ検索エンジン１４は、例えば、インターネットにおいて主要な複数の検索エンジン（ｈｔｔｐ：／／ｗｗｗ．ｙａｈｏｏ．ｃｏ．ｊｐ／、ｈｔｔｐ：／／ｓｅａｒｃｈ．ｍｓｎ．ｃｏ．ｊｐ／）を用いている。

【００２１】

また、コンピュータ１８には、テキストデータを形態素解析するための形態素解析器２０と、行列計算を行うための行列計算ライブラリ２２とが接続されている。

【００２２】

メタサーチエンジンプログラムは、後述するメタサーチ処理ルーチンを実行するためのプログラムであり、検索クエリログデータベース１２から取得した検索クエリログデータに基づいて、入力された検索語に関連する関連語を取得する関連語データ取得モジュール、検索クエリの検索語に適合するＷｅｂページを、複数のＷｅｂ検索エンジン１４によって検索し、検索されたＷｅｂページのキャッシュデータをキャッシュデータデータベース１６に一時的に記憶させる検索データ取得モジュール、キャッシュデータデータベース１６のキャッシュデータに対して、形態素解析器２０によって形態素解析を行って単語を取得し、名詞及び未知語の出現頻度を示す単語頻度行列を検索されたＷｅｂページ毎に作成する行列作成モジュール、単語頻度行列に対して行列計算ライブラリ２２によって行列計算を行い、関連語をクラスタリングして、関連語クラスタを生成するクラスタ生成モジュール、及び生成された関連語クラスタの順序付けを行うクラスタ順序付けモジュールを含んで構成されている。

【００２３】

なお、検索語に関連する関連語とは、検索エンジンへの検索クエリとして、ユーザが検索語と同時に入力した単語である。

【００２４】

次に、従来のＷｅｂ検索結果のクラスタリングの問題点について説明する。既存の文書、Ｗｅｂページ、Ｗｅｂ検索結果のクラスタリングの手法において、クラスタリング対象

となる文書群やＷｅｂページ群の全体がユーザの検索ニーズを分離可能な状態で包含している場合や、ユーザが興味を持つ文書群のみが文書群全体から明確に分離可能な場合には、効果的なクラスタ検索を行うことができる。

【００２５】

しかし、一般的に、Ｗｅｂ検索エンジンが返す検索結果は、効果的なクラスタ検索を行う上で理想的なＷｅｂページ群ではなく、ユーザにとって意味のない雑多な情報を多数含んでいる場合が多い。

【００２６】

例えば、検索語「英会話」に対してＷｅｂ検索エンジンが返す検索結果から、図２に示すようなＷｅｂページの単語頻度行列が作成された場合を考える。Ｗｅｂページの単語頻度行列に対して、Ｗｅｂページ方向に類似度計算することでＷｅｂページのクラスタリングが可能となり、また単語方向に類似度計算することで単語のクラスタリングが可能になる。図２の行列要素全てをクラスタリングに用いると、「英会話」という検索語の観点からはあまり関係のない「件」「月」「日」などの語が高い頻度で出現していたり、あるいは、逆に出現頻度が低く希少性が高かったり、また、他の語と共起していたりすることによって、語の持つ特徴量が大きくなることが、クラスタリング結果を悪化させる。例えば、類似度計算により、Ｗｅｂページのクラスタリングで｛英会話学習、英語の日記、ジオス｝｛英会話ＢＢＳ、イーオン｝のような分け方がされる場合、また、単語方向のクラスタリングで｛件、日｝｛月、無料、スクール、教材｝のような分け方がされる場合のどちらの場合も、ユーザにとってクラスタリング結果が理解しにくいものとなってしまう。

【００２７】

このように、Ｗｅｂ検索結果から構築される単語頻度行列の全体を用いると、ユーザの検索ニーズに合致しない雑多な情報が影響するために、Ｗｅｂページのクラスタリング及び単語のクラスタリングのどちらの場合も、クラスタ検索を効果的に行うことができない。

【００２８】

以下、上記のコンピュータ１８で実行される検索語による検索結果として、関連語のクラスタリング結果を表示するためのメタサーチ処理ルーチンについて図３を用いて説明する。

【００２９】

まず、ステップ１００において、ユーザが検索語を入力したか否かを判定し、ユーザがキーボードやマウス（図示省略）を操作して、検索語を入力すると、ステップ１０２へ進み、検索クエリログデータベース１２から、検索語に関連する複数の関連語を示す関連語データを取得する。

【００３０】

ここで、関連語は、検索に役立つ語を推薦するＹａｈｏｏ！（Ｒ）やＹａｈｏｏ！ＪＡＰＡＮ（Ｒ）の関連語検索の機能や検索広告のキーワード分析に用いられるものであり、検索広告では、キーワード分析を行うために、ユーザが検索語と同時に検索エンジンに入力した関連語の情報が提供されている。例えば、キーワード分析ツールにおいて、検索語「英会話」についての検索を行うユーザの検索ニーズを表す情報であって、図４のような関連語のデータを用いることにより、検索語「英会話」で得られる検索結果を、ユーザの検索ニーズに合致した情報によって絞り込む事ができるようになる。

【００３１】

なお、第１の実施の形態では、検索語の関連語のデータは、１００件を上限とする検索語の関連語と、月間検索数の予測値が得られるＯｖｅｒｔｕｒｅ（Ｒ）のキーワードアドバイスツール（ｈｔｔｐ：／／ｉｎｖｅｎｔｏｒｙ．ｊｐ．ｏｖｅｒｔｕｒｅ．ｃｏｍ／）により取得する。

【００３２】

そして、ステップ１０４では、複数のＷｅｂ検索エンジン１４を用いて、ステップ１００で入力された検索語に対応してメタサーチを行い、Ｗｅｂ検索エンジン１４の各々から

、Ｗｅｂ検索結果データとして、検索結果ＵＲＬ、Ｔｉｔｌｅ、ｓｕｍｍａｒｙ／ｓｎｉｐｐｅｔ、及びキャッシュＵＲＬを取得する。

【００３３】

なお、Ｙａｈｏｏ！（Ｒ）、Ｙａｈｏｏ！ＪＡＰＡＮ（Ｒ）、Ｇｏｏｇｌｅ（Ｒ）、ＭＳＮサーチ（Ｒ）などの主要なＷｅｂ検索エンジンでは、ライセンスを持たないメタ検索エンジンからのアクセスを禁止し、一般ユーザ向けに提供された検索サイトへの自動クエリの送信を禁止しているが、その代わりに、プログラムで検索エンジン資源にアクセスするための検索ＡＰＩやＳＤＫを提供している。例えば、Ｇｏｏｇｌｅ Ｗｅｂ ＡＰＩｓ（ｈｔｔｐ：／／ｗｗｗ．ｇｏｏｇｌｅ．ｃｏｍ／ａｐｉｓ）や、Ｙａｈｏｏ！Ｓｅａｒｃｈ Ｗｅｂ Ｓｅｒｖｉｃｅｓ ＳＤＫ（ｈｔｔｐ：／／ｄｅｖｅｌｏｐｅｒ．ｙａｈｏｏ．ｎｅｔ／ｓｅａｒｃｈ／）、ＭＳＮ Ｓｅａｒｃｈ Ｗｅｂ Ｓｅｒｖｉｃｅ ＳＤＫ（ｈｔｔｐ：／／ｍｓｄｎ．ｍｉｃｒｏｓｏｆｅ．ｃｏｍ／ｍｓｎ／ｍｓｎｓｅａｒｃｈ／）、Ｙａｈｏｏ！ＪＡＰＡＮ Ｗｅｂサービス ＳＤＫ（ｈｔｔｐ：／／ｄｅｖｅｌｏｐｅｒ．ｙａｈｏｏ．ｃｏ．ｊｐ／）があり、第１の実施の形態では、１０００件を上限とする日本語の検索結果が得られるＧｏｏｇｌｅ Ｗｅｂ ＡＰＩｓとＹａｈｏｏ！ＪＡＰＡＮ Ｗｅｅｂサービス ＳＤＫを用いてメタサーチを行っている。

【００３４】

次のステップ１０６では、キャッシュＵＲＬに基づいて、キャッシュデータを取得し、キャッシュデータをキャッシュデータデータベース１６に格納し、ステップ１０８において、キャッシュデータのＨＴＭＬソースファイルから、ＥＵＣ－ＪＰテキストであるテキストデータを抽出する。

【００３５】

そして、ステップ１１０で、形態素解析器２０のユーザ辞書に対して、入力された検索語及びステップ１０２で取得された関連語を登録し、ステップ１１２において、形態素解析器２０によって、ステップ１０８で抽出したテキストデータを形態素解析して、形態素解析結果として複数の単語を取得し、ステップ１１４で、形態素解析結果から、雑音を除去し、検索語の周辺の名詞及び未知語のみを抽出する。なお、形態素解析には、ＣｈａＳｅｎ（ｈｔｔｐ：／／ｃｈａｓｅｎ．ｎａｉｓｔ．ｊｐ／ｈｉｋｉ／Ｃｈａｓｅｎ／）を使用し、検索語や関連語をＣｈａＳｅｎのユーザ辞書に登録することにより、１つの語が複数の語に分割されていないようにしている。

【００３６】

そして、ステップ１１６において、抽出された名詞及び未知語で、図２に示すような複数のＷｅｂページに対する単語頻度行列を作成し、ステップ１１８で、作成された単語頻度行列における関連語と一致する単語の列要素ＩＤを抽出し、ステップ１２０において、抽出した列要素ＩＤを指定して、行列計算ライブラリ２２によって、関連語にのみ注目した関連語同士の類似度を算出する。

【００３７】

ここで、上述したように、検索結果から作成されるＷｅｂページの単語頻度行列全体に対して、単語方向の類似度計算を行うと、検索結果の中の雑多な情報がクラスタリングに悪影響を及ぼしてしまう。これに対して、例えば、図２に示すような検索結果に対して、図４に含まれる関連語で絞り込みを行い、関連語「スクール」「無料」「教材」のみの出現頻度で類似度を算出すると、図５に示すように、「英会話」に興味を持つユーザにとって重要でない語「件」「月」「日」を、類似度計算の対象から除外することができる。

【００３８】

そして、ステップ１２２において、ステップ１２０で算出された関連語同士の類似度に基づいて、関連語のクラスタリングを行い、類似度が高い組み合わせから関連語を組み合わせ、所定数の関連語クラスタになるまで、類似度が高い組み合わせから関連語の組み合わせを行い、所定数の関連語クラスタを生成する。例えば、図５にように、関連語「スクール」「無料」「教材」に限定して関連語のクラスタリングを行うことで、｛無料、教材｝｛スクール｝のような関連語クラスタを生成し、関連語クラスタ｛無料、教材｝を特徴

付けるページとして｛英会話学習、英語の日記、英会話ＢＢＳ｝を関連語クラスタ｛無料、教材｝に対応付け、また、関連語クラスタ｛スクール｝を特徴付けるページとして｛ジオス、イーオン｝を関連語クラスタ｛スクール｝に対応付ける。

【００３９】

なお、関連語のクラスタリングを行うために、第１の実施の形態では、連想計算のライブラリとして汎用連想計算エンジンＧＥＴＡ（ｈｔｔｐ：／／ｇｅｔａ．ｅｘ．ｎｉｉ．ａｃ．ｊｐ／）を利用している。ＧＥＴＡでは、単一リンク方、完全リンク法、群平均法、ＷＡＲＤ法、階層的ベイズクラスタリング（ＨＢＣ）などの代表的なクラスタリングの距離計算のアルゴリズムを指定できる。

【００４０】

また、検索数（月間検索数の予測値）が多い関連語で限定した関連語のクラスタリングにより、多くのユーザの検索ニーズに合致する関連語クラスタを生成することができる。

【００４１】

次のステップ１２４では、関連語の検索数に基づいて、ステップ１２２で生成された関連語クラスタの重み付けを行い、重みに基づいて関連語クラスタを順序付けて、関連語クラスタをソートする。関連語クラスタC_iの重みは以下の数式によって算出する。

【００４２】

【数１】

$$C_i = \sum_{t=1}^{T} f_t$$

ここで、f_tは関連語クラスタC_iに含まれる関連語w_tの検索数の総和であり、Ｔは関連語クラスタC_iに含まれる関連語の数である。

【００４３】

例えば、関連語「子供」が、「英会話　子供」「子供　英会話　教室」のような複数の検索で用いられている場合は関連語「子供」の検索数の総和は、「英会話　子供」「子供　英会話　教室」の検索数の和となる。図４の例では、「スクール」の検索数が２２７９６件、「無料」と「教材」の検索数がそれぞれ６６４７件、２２８５件となっている。従って、図５の関連語クラスタ｛無料、教材｝｛スクール｝の重みはそれぞれ８９３２、２２７９６と計算される。

【００４４】

そして、ステップ１２６において、関連語クラスタとＷｅｂ検索データが示すＷｅｂページとの対応付けを行い、ステップ１２８で、図６に示すように、ソートされた関連語クラスタのリストを検索結果として表示して、メタサーチ処理ルーチンを終了する。図５に示したような関連語クラスタが生成された場合には、検索結果において関連語クラスタが｛スクール｝｛無料、教材｝の順で表示される。このように、第１の実施の形態では、検索語の関連語のデータを用いて、検索で頻繁に用いられる関連語のみを用いた関連語のクラスタリングを行い、更に、生成された関連語クラスタを関連語の検索数で重み付けし、関連語クラスタをソートして検索結果を表示する。

【００４５】

また、検索結果の表示では、図６に示すように、関連語クラスタのリスト表示の下に、関連語クラスタの詳細表示として、関連語クラスタに対応付けられたＷｅｂページの文書データ情報としてのタイトルや概要、ＵＲＬも表示されるようになっている。

【００４６】

次に、第１の実施の形態のクラスタリングと従来のクラスタリングとの比較実験について説明する。ここでは、検索語として、Ｃｌｕｓｔｙ　ｔｈｅ　Ｃｌｕｓｔｅｒｉｎｇ

Ｅｎｇｉｎｅ（ｈｔｔｐ：／／ｃｌｕｓｔｙ．ｊｐ／）のトップページで例示されているクラスタ検索の検索語の例６語（英会話、介護、携帯電話、胃がん、悪質商法、受験）を用いて、関連語のクラスタリングとＷｅｂページのクラスタリングのとの結果を比較した。

【００４７】

関連語のクラスタリングに利用する関連語の数、生成するクラスタの数、クラスタリングの距離計算のアルゴリズムなど条件を変えることで、生成される関連語クラスタが変化する。異なる条件の下で関連語のクラスタリングとＷｅｂページのクラスタリングとをそれぞれ行い、クラスタリング結果を比較した。

【００４８】

関連語のクラスタリングを図７に示す条件で行い、Ｗｅｂページのクラスタリングを図８に示す条件で行い、検索語を「英会話」とした場合のＷｅｂページのクラスタリング結果を図９に示す。また、検索語を「英会話」とした場合の関連語のクラスタリング結果では、図１０に示すように、「無料、教材、上達法」は、無料の英会話教材を使って英会話の勉強をする場合をイメージすることができ、「マンツーマン、個人、プライベート、レッスン、講師」は、個人的に英会話のレッスンを受けたい場合をイメージすることができ、「ビジネス、ラジオ、日常、旅行」は、ラジオ番組を聴いて、英会話を習得したい場合をイメージすることができる。

【００４９】

また、検索語を「英会話」とした場合のＷｅｂページのクラスタリング結果を図１１に示し、また、検索語を「英会話」とした場合の関連語のクラスタリング結果を図１２に示す。

【００５０】

上記の比較結果では、関連語のクラスタリングの結果とＷｅｂページのクラスタリングの結果とには、ほとんど共通点がなく、Ｗｅｂページのクラスタリングでは、ユーザの検索意図とは無関係な意味の分からないクラスタが生成される傾向が見られた。

【００５１】

これに対して関連語のクラスタリングでは、ユーザにとって馴染みがあると思われる関連語がクラスタリング結果に現れ、ユーザ層や検索目的ごとの関連語クラスタが生成される傾向が見られた。

【００５２】

次に、実際に検索を行うユーザの立場で、関連語のクラスタリングの結果とＷｅｂページのクラスタリングの結果とを比較する評価実験について説明する。まず、評価者は、大学院生及び大学学部生（男性、２０代前半）１０名であり、検索語は図１３に示す２０語を用いた。なお、クラスタリングは、図７、８に示す条件で行った。

【００５３】

関連語のクラスタリングの結果とＷｅｂページのクラスタリングの結果とを左右並べて表示し、「どちらのクラスタリング結果が見やすいか」を評価者に質問して、回答を得た。評価者１０人が２０語のクラスタリング結果の比較を行い、合計２００件の回答が得られた。２００件のうち、１６１件が「関連語のクラスタリングの結果が見やすい」、３９件が「Ｗｅｂページのクラスタリングの結果が見やすい」という結果であった。

【００５４】

また、検索語別及び評価者別の回答結果を図１４及び図１５のそれぞれに示す。また、評価者別のクラスタリング結果１件当たりの平均閲覧時間を図１６に示す。検索語によって、また、評価者によって評価が分かれているが、Ｗｅｂページのクラスタリングと比較して、関連語のクラスタリングの方がユーザにとって分かりやすく見やすい結果を表示できていると推察される。

【００５５】

第１の実施の形態における関連語のクラスタリングでは、類義語（例えば、「試験」と「模試」）、共起語（例えば、「航空券」と「空席」と「予約」）、集合（例えば、「レ

クサス」と「ハリアー」と「アイシス」と「ウィッシュ」）、表記の揺れ（例えば、「プレーヤー」と「プレイヤー」）、複合語（例えば、「機種」と「変更」）がそれぞれ１つのクラスタにまとまる傾向が見受けられた。この傾向により、関連語のクラスタリングは、検索結果ページ群をコーパスとした関連語のシソーラス構築に相当するものといえる。

【００５６】

以上説明したように、第１の実施の形態に係る検索システムによれば、検索語に適合する各Ｗｅｂページにおける検索語に関連する複数の関連語の各々の出現頻度に基づいて関連語をクラスタリングした結果を、検索結果として表示することにより、ユーザによって入力される検索語に関係のない単語を除外して生成した関連語クラスタを検索結果として表示するため、ユーザにとって分かりやすいクラスタにより検索結果を表示することができる。

【００５７】

また、検索されたＷｅｂページを関連語クラスタに対応付けて表示するため、検索結果の表示におけるユーザの利便性を向上することができる。

【００５８】

また、検索クエリのログを記憶したデータベースから、自動的に検索語に関連する関連語を複数取得することができる。

【００５９】

また、検索語と同時に検索される回数が多い関連語を含む関連語クラスタから順に表示することにより、検索語との関連が強い関連語を含む関連語クラスタを先に表示するため、ユーザの検索ニーズに合致することができ、検索結果の表示におけるユーザの利便性を向上させることができる。

【００６０】

また、ユーザが頻繁に利用する検索語の関連語を用いた関連語のクラスタリングにより、ユーザにとって分かりやすい見やすいクラスタリング結果の表示を行うことができる。

【００６１】

また、複数の検索エンジンを一括検索することにより、質の良い多数の検索結果を得ることができる。

【００６２】

また、得られた多数の結果をクラスタリングして表示することでユーザにとって概観しやすい検索結果表示を行うことができる。

【００６３】

また、関連語の検索数で関連語クラスタを重み付けすることで、頻繁に参照される関連語クラスタを検索結果の上位に表示することができる。

【００６４】

なお、上記の実施の形態では、コンピュータが既存の複数の検索エンジンを利用して、Ｗｅｂ検索結果データを取得する場合を例に説明したが、コンピュータに検索エンジンの機能が搭載されており、Ｗｅｂページを複数記憶したデータベースから検索語に適合するＷｅｂページを取得するようにしてもよい。この場合には、関連語の取得や関連語クラスタリングの機能が、検索エンジンの一つの機能となる。

【００６５】

また、メタサーチ処理ルーチンなどのプログラムをコンピュータで実行する場合を例に説明したが、これに限定されるものではなく、検索システムが携帯情報端末を含んで構成されており、携帯情報端末で、メタサーチ処理ルーチンを含むプログラムを実行するように構成してもよい。

【００６６】

次に第２の実施の形態について説明する。なお、第１の実施の形態と同様の構成部分については、同一符号を付して説明を省略する。

【００６７】

第２の実施の形態では、関連語を取得するための検索語や、Ｗｅｂ検索データを取得す

るための検索語を修正して、再度クラスタリングすることができる点が第１の実施の形態と異なっている。

【００６８】

図１７に示すように、第２の実施の形態に係る検索システム２１０は、検索クエリログデータベース１２、Ｗｅｂ検索エンジン１４、キャッシュデータデータベース１６、及び検索語を類義語に修正するために、複数の単語の各々に対する類義語を記憶した検索語修正用シソーラスデータベース２１２に接続されたコンピュータ２１８を備えている。なお、類義語とは、一般的な意味の類義語の他に、分割した単語や、表記の揺れとなる単語を含む。

【００６９】

また、コンピュータ２１８には、形態素解析器２０と行列計算ライブラリ２２とが接続されている。

【００７０】

次に、第２の実施の形態におけるメタサーチ処理ルーチンについて図１８を用いて説明する。なお、第１の実施の形態と同様の処理については、同一符号を付して詳細な説明を省略する。

【００７１】

まず、ステップ１００において、ユーザが検索語を入力したか否かを判定し、検索語が入力されると、ステップ１０２で、検索クエリログデータベース１２から、検索語に関連する複数の関連語を示す関連語データを取得する。そして、ステップ１０４では、入力された検索語に対応してメタサーチを行い、Ｗｅｂ検索エンジン１４の各々から、Ｗｅｂ検索結果データを取得し、次のステップ１０６では、キャッシュデータを取得し、キャッシュデータデータベース１６に格納する。

【００７２】

そして、ステップ１０８において、キャッシュデータからテキストデータを抽出し、ステップ１１０で、形態素解析器２０のユーザ辞書に対して、入力された検索語及び関連語を登録し、ステップ１１２において、抽出したテキストデータを形態素解析して、形態素解析結果として複数の単語を取得し、ステップ１１４で、形態素解析結果から、雑音を除去し、検索語の周辺の名詞及び未知語のみを抽出する。

【００７３】

そして、ステップ１１６において、抽出された名詞及び未知語で、複数のＷｅｂページに対する単語頻度行列を作成し、ステップ１１８で、作成された単語頻度行列における関連語と一致する単語の列要素ＩＤを抽出し、ステップ１２０において、抽出した列要素ＩＤを指定して、関連語にのみ注目した関連語同士の類似度を算出する。

【００７４】

そして、ステップ１２２において、関連語のクラスタリングを行い、所定数の関連語クラスタを生成し、次のステップ１２４では、生成された関連語クラスタの重み付けを行い、重みに基づいて関連語クラスタを順序付けて、関連語クラスタをソートする。

【００７５】

そして、ステップ１２６において、関連語クラスタとＷｅｂ検索データが示すＷｅｂページとの対応付けを行い、ステップ１２８で、ソートされた関連語クラスタのリストを検索結果として表示する。

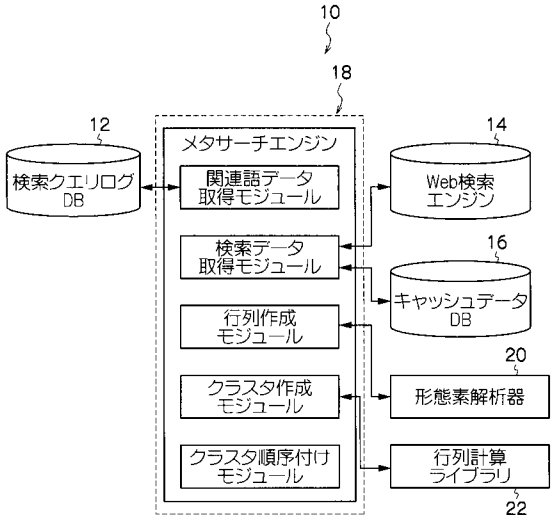
【００７６】

次のステップ２３０では、検索結果として表示された関連語クラスタを修正するか否かを判定し、ユーザから関連語クラスタの修正が指示されない場合には、メタサーチ処理ルーチンを終了するが、ユーザがキーボードやマウスを操作して、関連語クラスタの修正を指示すると、ステップ２３２で、関連語データを修正するか否かを判定し、ユーザが関連語データの修正を指示しない場合には、ステップ２３８へ移行するが、一方、ユーザがキーボードやマウスを操作して、関連語データの修正を指示した場合には、ステップ２３４へ移行する。

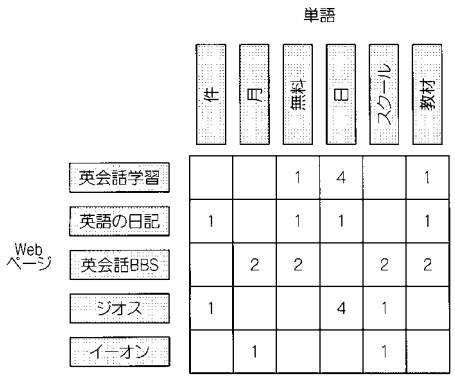
【００７７】
ステップ２３４では、関連語データ取得用に、修正した検索語を作成する。例えば、ユーザの入力により、修正した検索語を作成するか、または、検索語修正用シソーラスデータベース２１２から検索語の類似語を自動的に取得して、修正した検索語を作成する。次のステップ２３６では、修正済みの検索語と同時に検索される関連語を、検索クエリログデータベース１２から抽出して、関連語データを取得し、ステップ２３８へ移行する。
【００７８】
ステップ２３８において、Ｗｅｂ検索結果データを修正するか否かを判定し、ユーザがＷｅｂ検索結果データの修正を指示しない場合には、ステップ１０６へ戻り、新たに取得された関連語データに基づいて、再び関連語クラスタを生成するが、一方、ユーザがキーボードやマウスを操作して、Ｗｅｂ検索結果データの修正を指示した場合には、ステップ２４０へ移行する。
【００７９】
ステップ２４０では、Ｗｅｂ検索結果データ取得用に、修正した検索語を作成する。例えば、ユーザの入力により、修正した検索語を作成するか、または、検索語修正用シソーラスデータベース２１２から検索語の類似語を自動的に取得して、修正した検索語を作成する。次のステップ２４２では、修正済みの検索語に対応してメタサーチを行い、Ｗｅｂ検索エンジン１４の各々から、Ｗｅｂ検索結果データを取得して、ステップ１０６へ戻り、新たに取得された関連語データ及びＷｅｂ検索結果データに基づいて、再び関連語クラスタを生成する。
【００８０】
以上説明したように、第２の実施の形態に係る検索システムによれば、関連語クラスタを検索結果として表示した後に、検索語を修正して、新たに取得した関連語データ及びＷｅｂ検索結果データを用いて、検索結果となる関連語クラスタを生成することができるため、ユーザにとって更に分かりやすいクラスタにより検索結果を表示することができる。
【図面の簡単な説明】
【００８１】
【図１】第１の実施の形態に係る検索システムを示すブロック図である。
【図２】複数のＷｅｂページにおける単語頻度行列を示すイメージ図である。
【図３】第１の実施の形態に係るコンピュータのメタサーチ処理ルーチンの内容を示すフローチャートである。
【図４】検索語と関連語との組み合わせに対する検索数を示す表である。
【図５】関連語の列要素に限定した単語頻度行列を示すイメージ図である。
【図６】第１の実施の形態に係る検索結果表示のイメージ図である。
【図７】関連語のクラスタリングの条件を示す表である。
【図８】Ｗｅｂページのクラスタリングの条件を示す表である。
【図９】検索語を「英会話」とした場合のＷｅｂページのクラスタリング結果を示す図である。
【図１０】検索語を「英会話」とした場合の関連語のクラスタリング結果を示す図である。
【図１１】検索語を「受験」とした場合のＷｅｂページのクラスタリング結果を示す図である。
【図１２】検索語を「受験」とした場合の関連語のクラスタリング結果を示す図である。
【図１３】ユーザ評価に用いた検索語を示す図である。
【図１４】複数の検索語各々におけるクラスタリング結果の見やすさを示すグラフである。
【図１５】複数の評価者各々におけるクラスタリング結果の見やすさを示すグラフである。
【図１６】複数の評価者各々におけるクラスタリング結果の平均閲覧時間を示すグラフである。

【図１７】第２の実施の形態に係る検索システムを示すブロック図である。
【図１８】第２の実施の形態に係るコンピュータのメタサーチ処理ルーチンの内容を示すフローチャートである。
【符号の説明】
【００８２】
１０、２１０　検索システム
１２　検索クエリログデータベース
１４　検索エンジン
１６　キャッシュデータデータベース
１８、２１８　コンピュータ
２０　形態素解析器
２２　行列計算ライブラリ
２１２　検索語修正用シソーラスデータベース

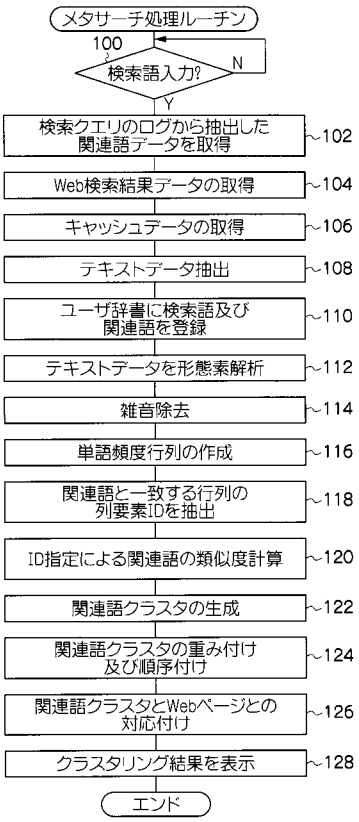
【図１】



【図２】



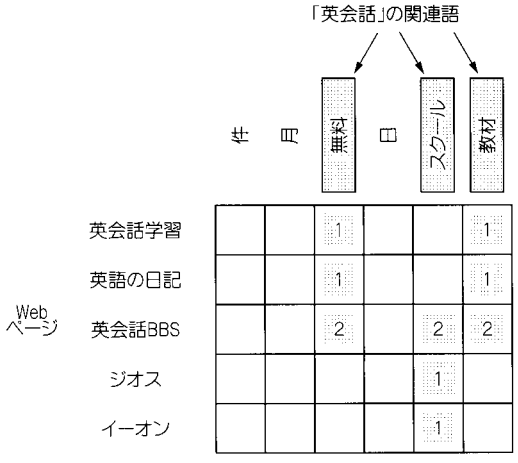
【図３】



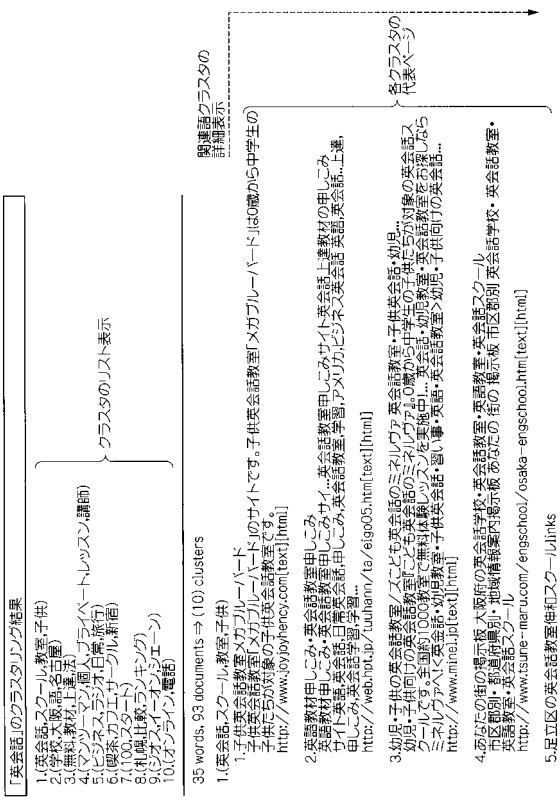
【図４】

検索数	検索語と関連語
86044	英会話
22796	英会話 **スクール**
8281	英会話 教室
7847	子供 英会話
6647	英会話 **無料**
5724	英会話 学校
4039	nhk 英会話
3191	英会話 喫茶
2901	ビジネス 英会話
2684	英会話 大阪
2674	マンツーマン 英会話
2312	100 語 スタート 英会話
2285	英会話 **教材**
2228	日常 英会話
2181	英会話 ジオス

【図５】



【図６】



【図７】

生成する関連語クラスタの数	10個
表示する代表ページ数	10件
利用する関連語数	上位40語
距離計算のアルゴリズム	HBC
検索語周辺の有効距離語数	10語

【図８】

生成するページクラスタの数	10個
表示する代表単語数	上位4語
距離計算のアルゴリズム	HBC
検索語周辺の有効距離語数	10語

【図９】

1.(英会話,通信,英語,教材)
2.(英会話,中途,サンドイッチ,解約)
3.(保育園,9月,ヴィレッジ,エディ)
4.(英会話,英語,教材,学習)
5.(スミス,案内,英会話,コース)
6.(ミキハウス,ACC,溝の口,エポック)
7.(英会話,スクール,レッスン,英語)
8.(ゴルフ,動画,夙川,ダイアローグ)
9.(英会話,∞,英語,件)
10.(賃貸,殿堂,ヘルス,海老名)

【図１０】

1.(英会話,スクール,教室,子供)
2.(学校,大阪,語,名古屋)
3.(無料,教材,上達,法)
4.(マンツーマン,個人,プライベート,レッスン,講師)
5.(ビジネス,ラジオ,日常,旅行)
6.(喫茶,カフェ,サークル,新宿)
7.(100,スタート)
8.(札幌,比較,ランキング)
9.(ジオス,イーオン,シェーン)
10.(オンライン,電話)

【図１１】

1.(看護,受験,コース,東進)
2.(CAD,クリップ,整体,受験)
3.(士,受験,会計士,心理)
4.(受験,大学,勉強,英語)
5.(東京音楽大学,周期,バック,権)
6.(試験,受験,申込,者)
7.(小学校,受験,幼稚園,幼児)
8.(格言,医学部,歯学部,購入)
9.(受験,中学,家庭教師,高校)
10.(美,芸大,美術,デッサン)

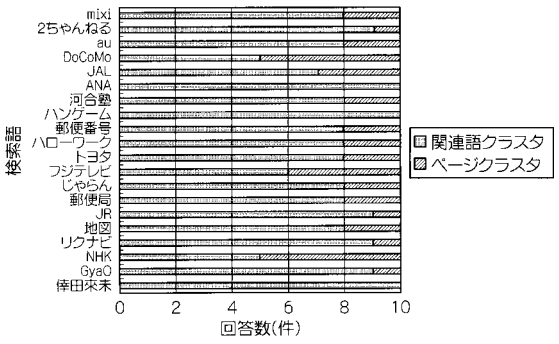
【図１２】

1.(受験,高校,大学,情報)
2.(中学,家庭教師,塾,ランキング,算数)
3.(偏差値,英語,勉強,法,医学部,予備校)
4.(小学校,幼稚園,面接)
5.(掲示板,参考書,校,東大)
6.(関西,宝塚,税理士,音楽)
7.(資格,学,士,社会)
8.(看護,一覧,福祉,ネット)
9.(2ちゃんねる,立命館,宿)
10.(マネージャー,ケア)

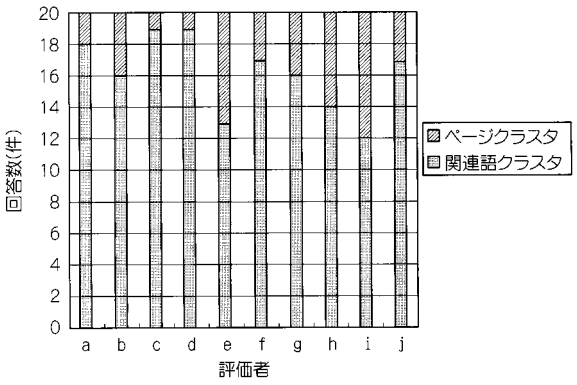
【図１３】

mixi,2ちゃんねる,au,DoCoMo,JAL,ANA,
河合塾,ハンゲーム,郵便番号,ハローワーク,
トヨタ,フジテレビ,じゃらん,郵便局,JR,地図,
リクナビ,NHK,GyaO,倖田來未

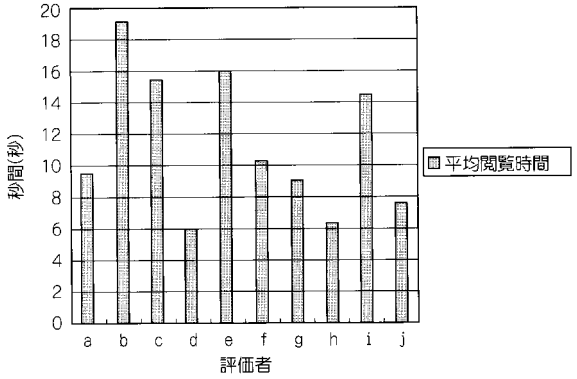
【図１４】



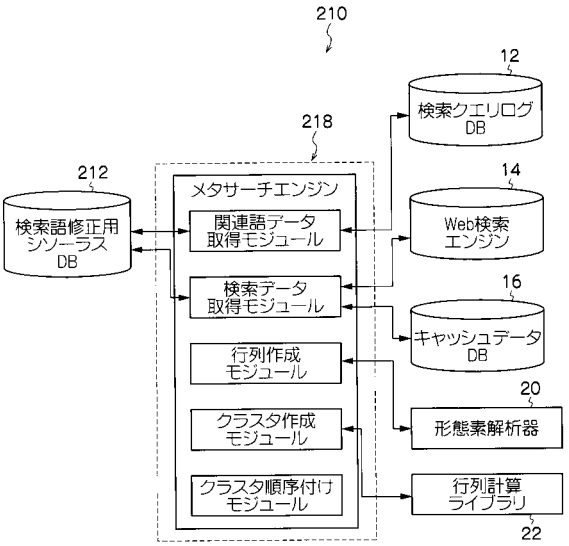
【図１５】



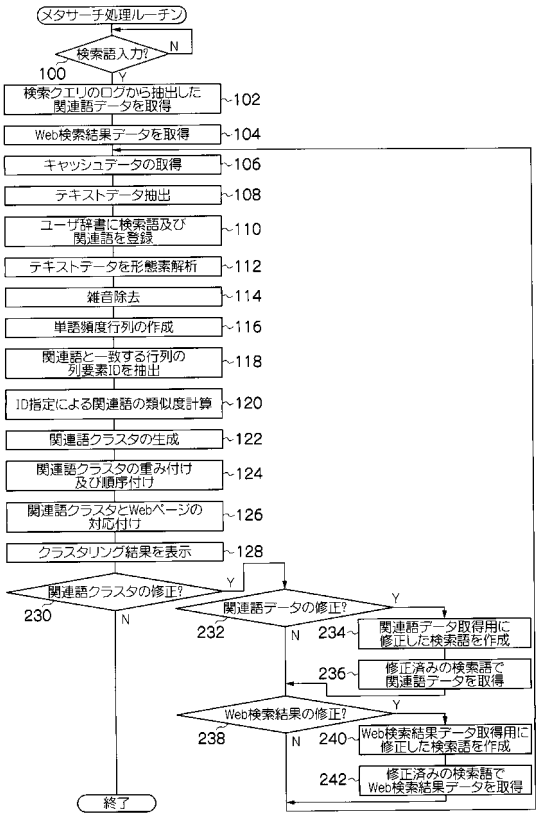
【図１６】



【図１７】


210
12
検索クエリログ DB
218
212
メタサーチエンジン
検索語修正用 シソーラス DB
関連語データ 取得モジュール
14
Web検索 エンジン
検索データ 取得モジュール
16
キャッシュデータ DB
行列作成 モジュール
20
クラスタ作成 モジュール
形態素解析器
クラスタ順序付け モジュール
行列計算 ライブラリ
22

【図１８】


メタサーチ処理ルーチン
検索語入力?
N
Y
100
検索クエリのログから抽出した関連語データを取得
102
Web検索結果データを取得
104
キャッシュデータの取得
106
テキストデータ抽出
108
ユーザ辞書に検索語及び関連語を登録
110
テキストデータを形態素解析
112
雑音除去
114
単語頻度行列の作成
116
関連語と一致する行列の列要素IDを抽出
118
ID指定による関連語の類似度計算
120
関連語クラスタの生成
122
関連語クラスタの重み付け及び順序付け
124
関連語クラスタとWebページの対応付け
126
クラスタリング結果を表示
128
関連語クラスタの修正?
230
Y
N
関連語データの修正?
232
Y
N
234
関連語データ取得用に修正した検索語を作成
236
修正済みの検索語で関連語データを取得
Web検索結果の修正?
238
Y
N
240
Web検索結果データ取得用に修正した検索語を作成
242
修正済みの検索語でWeb検索結果データを取得
終了

フロントページの続き

(72)発明者　横尾　英俊
群馬県桐生市相生町５丁目２８４－６４
(72)発明者　内山　智文
神奈川県川崎市中原区小杉町１－５１６－２　ダイホーステージ武蔵小杉２０４

審査官　波内　みさ

(56)参考文献　特開２００５－３４６５６０（ＪＰ，Ａ）
特開平１１－３２８２２０（ＪＰ，Ａ）
特開２００３－２０８４４７（ＪＰ，Ａ）
特開平１１－３２８２２１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　　１７／３０