

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2008-27207
(P2008-27207A)

(43) 公開日 平成20年2月7日(2008.2.7)

(51) Int. Cl.	F 1	テーマコード (参考)
G 0 6 F 17/30 (2006.01)	G 0 6 F 17/30 3 7 0 Z	5 B 0 7 5
	G 0 6 F 17/30 2 1 0 D	
	G 0 6 F 17/30 3 2 0 D	

審査請求 有 請求項の数 6 O L (全 17 頁)

(21) 出願番号 (22) 出願日 特許法第30条第1項適用申請有り 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	特願2006-199312 (P2006-199312) 平成18年7月21日 (2006.7.21)	(71) 出願人 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

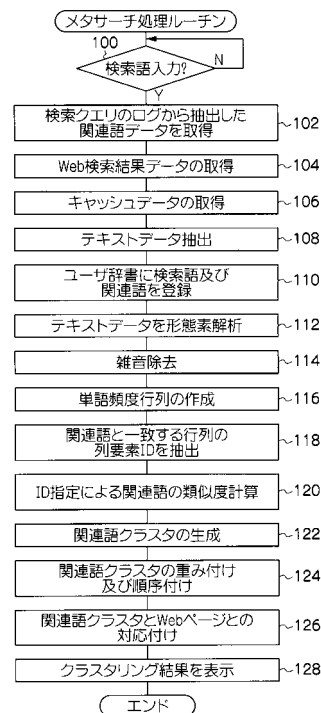
(54) 【発明の名称】 検索装置及び検索方法

(57) 【要約】

【解決課題】 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ようにする。

【解決手段】 検索クエリログから、検索語に関連する複数の関連語を取得し(102)、複数の検索エンジンによって、検索語に対応してメタサーチを行い(104)、検索されたWebページからテキストデータを抽出し(108)、抽出したテキストデータを形態素解析して複数の単語を取得し(112)、複数のWebページに対する単語頻度行列を作成し(116)、関連語にのみ注目した関連語同士の類似度を算出する。(120)そして、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、所定数の関連語クラスタを生成し(122)、関連語の検索数に基づいて、関連語クラスタの重み付けを行い、関連語クラスタを順序付けてソートし(124)、関連語クラスタのリストを検索結果として表示する(128)。

【選択図】 図3



【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、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を、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】

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を更に含み、

前記表示手段は、前記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、前記検索結果として表示する請求項 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 又は 2 記載の検索装置。

【請求項 4】

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 又は 2 記載の検索装置。

【請求項 5】

前記表示手段は、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、前記検索結果として表示する請求項 1 ~ 請求項 4 の何れか 1 項記載の検索装置。

【請求項 6】

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、

前記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し、

前記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し、

前記複数の関連語のクラスタリングを行って、前記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、所定数の関連語クラスタを生成し、

前記生成された関連語クラスタを、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ことを特徴とする検索方法。

10

20

30

40

50

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、検索装置及び検索方法にかかり、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検索方法に関する。

【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。

10

【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。

20

【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、利便性が低いものになってしまう、という問題がある。

30

【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検索方法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、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を、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。

40

【0008】

50

また、本発明に係る検索方法は、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し、前記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し、前記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し、前記複数の関連語のクラスタリングを行って、前記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、所定数の関連語クラスタを生成し、前記生成された関連語クラスタを、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ことを特徴としている。

【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。

10

【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、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。

20

【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。

30

【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。

40

【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】

50

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検索方法によれば、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。

10

【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。

20

【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。

30

【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。

40

【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。

50

【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で、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になってしまう。

【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トール(<http://inventory.jp.overture.com/>)により取得する。

10

20

30

40

50

【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.microsoft.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 Web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

10

20

30

40

50

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のように、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、単一リンク法、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。

10

【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】

20

【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。

30

【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。

40

【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。

50

【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

10

【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

20

【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。

30

【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。

40

【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。

50

【 0 0 5 5 】

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。

【 0 0 5 6 】

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。

10

【 0 0 5 7 】

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。

【 0 0 5 8 】

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。

20

【 0 0 5 9 】

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。

【 0 0 6 0 】

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。

【 0 0 6 1 】

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。

30

【 0 0 6 2 】

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。

【 0 0 6 3 】

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。

【 0 0 6 4 】

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。

40

【 0 0 6 5 】

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。

【 0 0 6 6 】

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

50

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

10

【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。

20

【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。

30

【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。

40

【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

50

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。

10

20

30

40

50

。

【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。

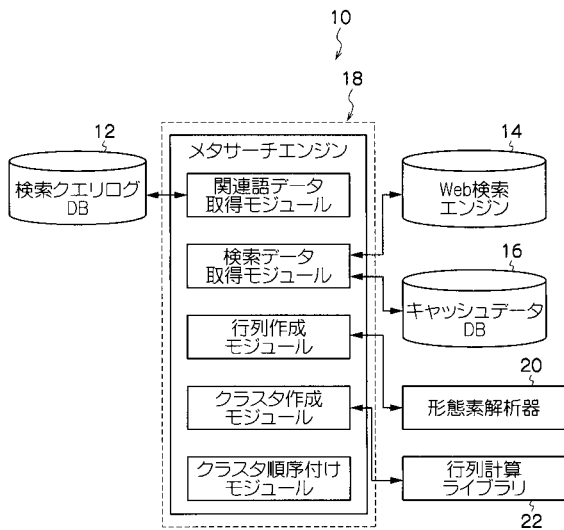
【図18】第2の実施の形態に係るコンピュータのメタサーチ処理ルーチンの内容を示すフローチャートである。

【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】

		単語					
		生	師	無料	口	スクール	教材
Web ページ	英会話学習			1	4		1
	英語の日記	1		1	1		1
	英会話BBS		2	2		2	2
	ジオス	1			4	1	
	イーオン		1			1	

【 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,ウイレッジ,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,ダイアログ)
- 9.(英会話,∞,英語,件)
- 10.(賃貸,殿堂,ヘルス,海老名)

【 図 1 0 】

- 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,シエーン)
- 10.(オンライン,電話)

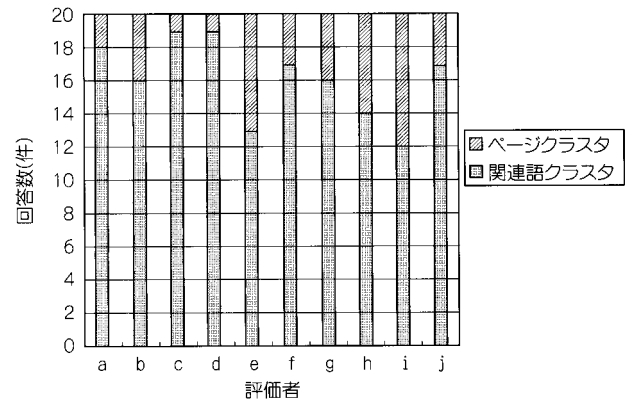
【 図 1 1 】

- 1.(看護,受験,コース,東進)
- 2.(CAD,クラブ,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 1 2 】

- 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)

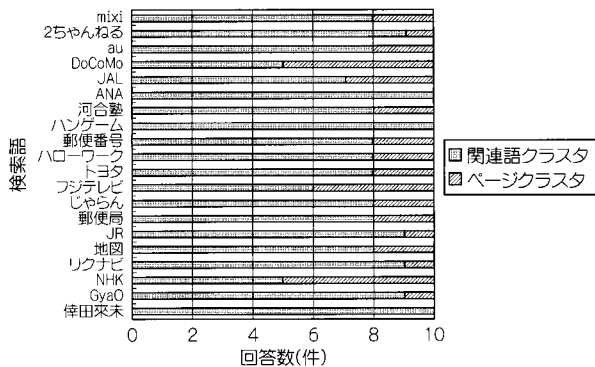
【 図 1 5 】



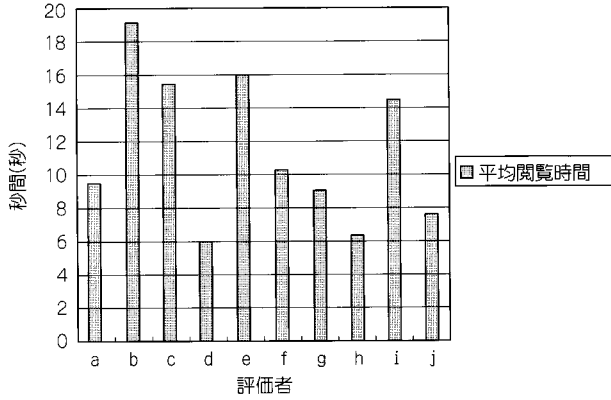
【 図 1 3 】

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,じやらん,郵便局,JR,地図,
リクナビ,NHK,GyaO,倅田來未

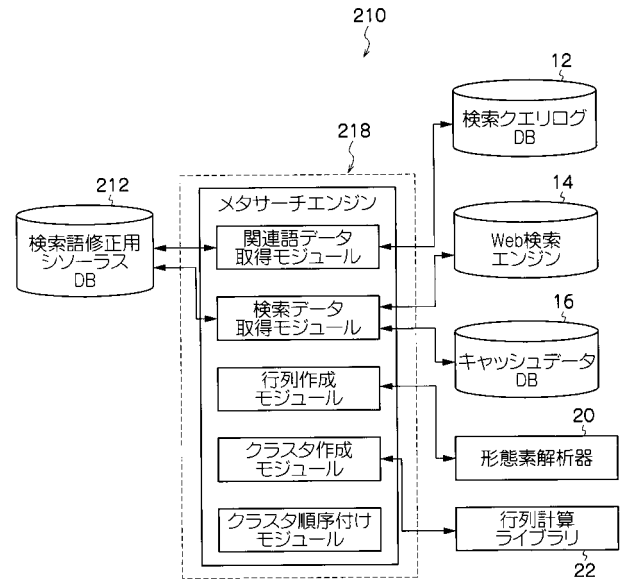
【 図 1 4 】



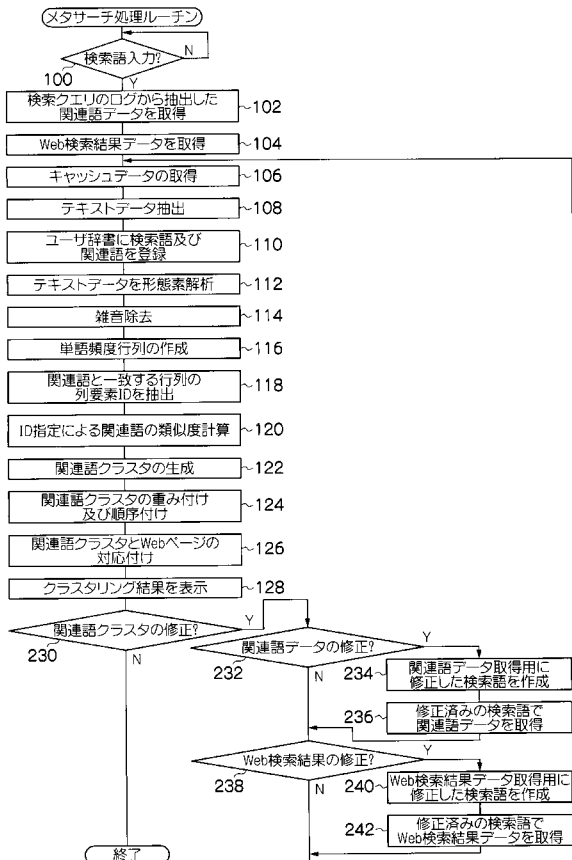
【図16】



【図17】



【図18】



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

Fターム(参考) 5B075 NK31 NR12