(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6237334号
(P6237334)
(45)発行日　平成29年11月29日(2017.11.29)　(24)登録日　平成29年11月10日(2017.11.10)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 210D
G06F 17/30 350C

請求項の数 11　(全 26 頁)

(21)出願番号　特願2014-36700(P2014-36700)
(22)出願日　平成26年2月27日(2014.2.27)
(65)公開番号　特開2015-162076(P2015-162076A)
(43)公開日　平成27年9月7日(2015.9.7)
審査請求日　平成28年11月2日(2016.11.2)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100094525
弁理士　土井　健二
(74)代理人　100094514
弁理士　林　恒徳
(72)発明者　高橋　哲朗
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

審査官　樋口　龍弥

最終頁に続く

(54)【発明の名称】クエリ生成方法、クエリ生成プログラム、及び、クエリ生成装置

(57)【特許請求の範囲】
【請求項１】
処理ユニットが、
入力された検索語に基づいて複数の文書を検索し、前記検索された複数の文書を類似度にしたがって複数の文書集合に分類し、前記複数の文書集合を識別する情報を表示ユニットに表示する第１の工程と、
前記表示された複数の文書集合のうち排除すべき文書集合として指定された、排除文書集合内の文字列を抽出し、前記抽出した文字列毎の、前記検索された複数の文書における前記排除文書集合内での出現分布率を示すスコアを計算し、前記スコアの降順にソートされた前記文字列の文字列リストを生成する第２の工程と、
前記文字列リスト内の文字列数に比例する入力に応じて、前記文字列リストの上位数分の前記文字列を排除条件のクエリの候補として選択し、前記選択した文字列を含む文書の割合を文書集合毎に計算し前記表示ユニットに表示する第３の工程と、
を実行することを特徴とするクエリ生成方法。
【請求項２】
請求項１において、
前記第２の工程は、前記抽出した文字列毎に、前記排除文書集合内の当該文字列の頻度に応じて、前記検索された複数の文書における前記排除文書集合内での前記出現分布値を加算して前記スコアを算出するクエリ生成方法。
【請求項３】

請求項２において、
前記第２の工程は、前記抽出した文字列毎に、さらに、前記検索された複数の文書における前記排除文書集合以外の文書集合内での前記出現分布値を前記スコアから減算して、前記スコアを算出するクエリ生成方法。
【請求項４】
請求項１乃至３のいずれかにおいて、
前記第３の工程は、前記文字列リスト内の文字列数に比例する値を示すオブジェクトを前記表示ユニットに表示し、前記オブジェクトに対する前記ユーザの入力に比例する、前記文字列リストの上位数分の前記文字列を選択するクエリ生成方法。
【請求項５】
請求項４において、
前記第３の工程は、前記オブジェクトに対する入力に応じて選択される前記文字列を前記表示ユニットに更に表示するクエリ生成方法。
【請求項６】
請求項４または５において、
前記オブジェクトは、スライドバーであるクエリ生成方法。
【請求項７】
請求項１乃至６のいずれかにおいて、
前記第２の工程は、前記文書集合の特徴を示す特徴情報を前記表示ユニットに更に表示するクエリ生成方法。
【請求項８】
請求項７において、
前記特徴情報は、前記文書集合の文書に含まれる前記検索語の使用文字列、前記文書集合の文書に含まれる頻出文字列、前記文書集合の文書に含まれる文字列であって前記検索された複数の文書における前記書集合内での出現分布率が高い文字列、のうち少なくともいずれかであるクエリ生成方法。
【請求項９】
請求項１乃至８のいずれかにおいて、
前記第２の工程は、前記排除文書集合の指定を受け付けるオブジェクトを更に前記表示ユニットに表示するクエリ生成方法。
【請求項１０】
入力された検索語に基づいて複数の文書を検索し、前記検索された複数の文書を類似度にしたがって複数の文書集合に分類し、前記複数の文書集合を識別する情報を表示ユニットに表示し、
前記表示された複数の文書集合のうち排除すべき文書集合として指定された、排除文書集合内の文字列を抽出し、前記抽出した文字列毎の、前記検索された複数の文書における前記排除文書集合内での出現分布率を示すスコアを計算し、前記スコアの降順にソートされた前記文字列の文字列リストを生成し、
前記文字列リスト内の文字列数に比例する入力に応じて、前記文字列リストの上位数分の前記文字列を排除条件のクエリの候補として選択し、前記選択した文字列を含む文書の割合を文書集合毎に計算し前記表示ユニットに表示する、
処理をコンピュータに実行させるクエリ生成プログラム。
【請求項１１】
処理ユニットと、
複数の文書を記憶する記憶装置と、
表示装置と、を有し、
入力された検索語に基づいて前記複数の文書を検索し、前記検索された複数の文書を類似度にしたがって複数の文書集合に分類し、前記複数の文書集合を識別する情報を表示ユニットに表示し、前記表示された複数の文書集合のうち排除すべき文書集合として指定された、排除文書集合内の文字列を抽出し、前記抽出した文字列毎の、前記検索された複数

の文書における前記排除文書集合内での出現分布率を示すスコアを計算し、前記スコアの降順にソートされた前記文字列の文字列リストを生成し、前記文字列リスト内の文字列数に比例する入力に応じて、前記文字列リストの上位数分の前記文字列を排除条件のクエリの候補として選択し、前記選択した文字列を含む文書の割合を文書集合毎に計算し前記表示ユニットに表示するクエリ生成装置。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、クエリ生成方法、クエリ生成プログラム、及び、クエリ生成装置に関する。

【背景技術】

【０００２】

マーケティング等の目的のために、ソーシャルメディアから情報を得るというニーズが増えている。例えば、大量にあるソーシャルメディアの文書（記事）のうち、一部の文書の集合がマーケティング等の分析の対象となる。

【０００３】

分析の対象となる文書集合を選択するために、文書検索技術が使われる。例えば、マーケティング担当者は、所定の文書を検索するためクエリ（検索語、検索条件）をソーシャルメディアの文書を格納するデータベースに指定することによって、特定の条件に合致する文書を検索する。これにより、マーケティング担当者は、分析の対象となる文書集合を抽出することができる。マーケティング担当者は、分析の対象となる文書集合を選択するために適切なクエリを設定して発行する。

【先行技術文献】

【特許文献】

【０００４】

【特許文献１】特開２００８－７７１３７号公報
【特許文献２】特開２００６－２５１９３５号公報
【特許文献３】特開２０１２－８４０２９号公報
【特許文献４】特開平１１－２７２７０９号公報

【発明の概要】

【発明が解決しようとする課題】

【０００５】

しかしながら、マーケティング担当者の所望の文書集合を選択するためのクエリの生成は容易ではない。例えば、「セブンイレブン（登録商標）」の省略形である「セブン」をクエリとする場合、「ウルトラセブン（登録商標）」や「セブンスター（登録商標）」、映画の「セブン」に関する文書も合わせて抽出される。抽出された文書集合に基づいて分析を行う場合、「セブンイレブン」に関する文書以外の文書が含まれることにより、分析の精度が下がる。したがって、抽出された文書集合内の「セブンイレブン」以外に関する文書は、少ないことが望ましい。

【０００６】

１つの側面は、本発明は、ユーザの目的に適合する文書を抽出するクエリを生成するクエリ生成方法、クエリ生成プログラム、及び、クエリ生成装置を提供する。

【課題を解決するための手段】

【０００７】

第１の側面は、処理ユニットが、入力された検索語に基づいて複数の文書を検索し、前記検索された複数の文書を類似度にしたがって複数の文書集合に分類し、前記複数の文書集合を識別する情報を表示ユニットに表示する第１の工程と、前記表示された複数の文書集合のうち排除すべき文書集合として指定された、排除文書集合内の文字列を抽出し、前記抽出した文字列毎の、前記検索された複数の文書における前記排除文書集合内での出現分布率を示すスコアを計算し、前記スコアの降順にソートされた前記文字列の文字列リストを生成する第２の工程と、前記文字列リスト内の文字列数に比例する入力に応じて、前

記文字列リストの上位数分の前記文字列を排除条件のクエリの候補として選択し、前記選択した文字列を含む文書の割合を文書集合毎に計算し前記表示ユニットに表示する第３の工程と、を実行する。

【発明の効果】

【０００８】

第１の側面によれば、ユーザの選択により、ユーザの目的に適合する文書を抽出するクエリを生成する。

【図面の簡単な説明】

【０００９】

【図１】本実施形態における文書検索システムの構成を示す図である。

【図２】図１に示す本実施の形態例における検索サーバのハードウェア構成を説明する図である。

【図３】図１、図２に示す検索サーバのソフトウェアブロック図である。

【図４】クラスタを用いるクエリの基本的な生成処理を説明する第１の図である。

【図５】クラスタを用いたクエリの基本的な生成処理を説明する第２の図である。

【図６】本実施の形態例におけるクエリ生成処理を説明するフローチャート図である。

【図７】本実施の形態例におけるクラスタの生成処理を説明する図である。

【図８】クラスタの指定を受け付けるクライアント装置の画面の一例を示す図である。

【図９】キーワードリストの生成処理を説明する図である。

【図１０】キーワードリストの生成処理を説明するフローチャート図である。

【図１１】キーワード毎のｔｆｉｄｆ値を例示する図である。

【図１２】キーワード生成処理の具体例を説明する第１の図である。

【図１３】キーワード生成処理の具体例を説明する第２の図である。

【図１４】キーワード生成処理の具体例を説明する第３の図である。

【図１５】キーワード生成処理の具体例を説明する第４の図である。

【図１６】具体例におけるキーワードリストを示す図である。

【図１７】本実施の形態例におけるキーワードの選択処理を説明する図である。

【図１８】スライドバーを表示するクライアント装置の画面の一例を示す図である。

【図１９】排除される文書の割合を表示するクライアント装置の画面の一例を示す図である。

【発明を実施するための形態】

【００１０】

以下、図面にしたがって本発明の実施の形態について説明する。ただし、本発明の技術的範囲はこれらの実施の形態に限定されず、特許請求の範囲に記載された事項とその均等物まで及ぶものである。

【００１１】

［文書検索システム］

図１は、本実施形態における文書検索システムの構成を示す図である。図１に示すように、本実施形態における文書検索システムは、クライアント装置８０と検索サーバ（クエリ生成装置）１０とを有する。クライアント装置８０は、検索サーバ１０と通信ネットワーク５０を介して接続する。クライアント装置８０は、例えば、パーソナルコンピュータ等である。なお、図１の例では、１台のクライアント装置８０を図示しているが、検索サーバ１０は、複数のクライアント装置８０と接続してもよい。

【００１２】

クライアント装置８０は、Ｗｅｂページを閲覧するブラウザを介して、検索サーバ１０に、検索語等の検索条件を入力する。検索サーバ１０は、クライアント装置８０から入力された検索条件に基づいてクエリを生成し、文書集合からクエリに対応する文書を抽出する。そして、検索サーバ１０は、抽出した文書をクライアント装置８０に送信する。本実施の形態例における検索サーバ１０は、クライアント装置８０からの検索条件や入力情報に基づいて、適切なクエリを生成する。クライアント装置８０は、例えば、ＣＰＵ（Cent

ral Processing Unit）、メモリ、表示ユニット、入力ユニット等を有する（不図示）。

【００１３】

［検索サーバの構成］

図２は、図１に示す本実施の形態例における検索サーバ１０のハードウェア構成を説明する図である。図２に示す検索サーバ１０は、例えば、ＣＰＵ（Central Processing Unit）１０１、ＲＡＭ（Random Access Memory）２０１や不揮発性メモリ２０２等を備えるメモリ１０２、通信インタフェース部１０３を有する。各部は、バス１０４を介して相互に接続する。

【００１４】

ＣＰＵ１０１は、バス１０４を介してメモリ１０２等と接続すると共に、検索サーバ１０全体の制御を行う。メモリ１０２のＲＡＭ２０１は、ＣＰＵ１０１が処理を行うデータ等を記憶する。メモリ１０２の不揮発性メモリ２０２は、ＣＰＵ１０１が実行するＯＳ（Operating System）のプログラムを格納する領域（図示せず）や、本実施の形態例におけるクエリ生成プログラムを格納する領域２１０を備える。また、不揮発性メモリ２０２は、文書集合を格納する領域２０を有する。文書集合を格納する領域（以下、文書集合と称する）２０は、例えば、ソーシャルネットワークの記事の集合である。記事は、例えば、ブログの記事やコメント等である。不揮発性メモリ２０２は、ＨＤＤ（Hard disk drive）、不揮発性半導体メモリ等によって構成される。

【００１５】

クエリ生成プログラム領域２１０のクエリ生成プログラム（以下、クエリ生成プログラム２１０と称する）は、ＣＰＵ１０１の実行によって、本実施の形態例におけるクエリ生成処理を実現する。また、通信インタフェース部１０３は、ネットワーク５０を介して、クライアント装置８０等の通信機器との間でデータの送受信を制御する。

【００１６】

図３は、図１、図２に示す検索サーバ１０のソフトウェアブロック図である。検索サーバ１０のクエリ生成プログラム２１０（図２）は、例えば、文書検索モジュール（以下、文書検索部と称する）１１、クラスタリングモジュール（以下、クラスタリング部と称する）１２、キーワード生成モジュール（以下、キーワード生成部と称する）１３を有する。

【００１７】

文書検索部１１は、ユーザが入力した検索語をクライアント装置８０から受信し（ａ１）、検索語に基づいて検索サーバ１０が格納する文書集合２０を検索して、複数の文書を抽出する。なお、図２、図３の例では、検索対象となる文書集合２０は、検索サーバ１０に格納される。しかしながら、この例に限定されるものではない。文書集合２０は、ネットワーク５０を介して接続する１つまたは複数の他のサーバに格納されていてもよい。クラスタリング部１２は、文書検索部１１が検索して抽出した複数の文書を、文書の類似度に基づいて複数の文書集合（以下、クラスタと称する）に分類する。

【００１８】

キーワード生成部１３は、クラスタ指定受付モジュール（以下、クラスタ指定受付部と称する）３１と、キーワードリスト生成モジュール（以下、キーワードリスト生成部と称する）３２と、キーワード選択モジュール（以下、キーワード選択部と称する）３３を有する。

【００１９】

クラスタ指定受付部３１は、複数のクラスタを識別する情報をクライアント装置８０の表示ユニットに表示させるとともに、複数のクラスタのうち、検索結果から排除すべき排除対象のクラスタ（以下、排除クラスタと称する）、及び、検索結果として選択すべき選択対象のクラスタ（以下、選択クラスタと称する）の指定をユーザから受け付ける（ａ２）。キーワードリスト生成部３２は、排除クラスタ、選択クラスタに基づいて、クエリ２３における検索語の排除条件となるキーワードのリスト２２を作成する。

【００２０】

キーワード選択部３３は、キーワードリスト２２内のキーワード数に比例する値を示すスライドバーをクライアント装置８０の表示ユニットに表示する。また、キーワード選択部３３は、ユーザによるスライドバーの操作を受け付けるとともに（ａ３）、スライドバーの値に対応するキーワードを含む文書の割合をクラスタ毎に表示する。スライドバーの値に対応するキーワードがクエリ２３の排除条件となる。

【００２１】

次に、本実施の形態例におけるクエリ生成処理を説明する前に、クラスタを用いたクエリ２３の基本的な生成処理を図に基づいて説明する。

【００２２】

［クラスタを用いたクエリの生成］

図４は、クラスタを用いるクエリ２３の基本的な生成処理を説明する第１の図である。本実施の形態例では、ユーザが、コンビニエンスストアの「セブンイレブン」に関する文書を検索して抽出する場合を前提とする。「セブンイレブン」は、「セブン」と省略して用いられることが多い。したがって、ユーザは、例えば、「セブン」をクエリ２３の検索語として入力する。この結果、文書検索部１１は、文書内に検索語「セブン」を含む複数の文書２０ａを取得する。

【００２３】

ただし、検索語「セブン」にしたがって検索された複数の文書２０ａは、「セブンイレブン」に関する記事に加えて、「ウルトラセブン」や「セブンスター」、映画の「セブン」に関する記事も含む。検索された複数の文書２０ａを対象として、マーケティング等の分析処理を行う場合、「セブンイレブン」に関する記事以外の記事が含まることにより、分析の精度が下がる。したがって、検索結果から、「セブンイレブン」に関する記事以外の記事が排除されることが望ましい。そこで、クラスタリング部１２は、検索語「セブン」にしたがって検索された複数の文書２０ａを、複数のクラスタに分類する。

【００２４】

図５は、クラスタを用いたクエリ２３の基本的な生成処理を説明する第２の図である。クラスタリング部１２は、検索語「セブン」に基づいて抽出された複数の文書２０ａを、類似性のある文書同士をまとめて、複数のクラスタを生成する。図５の例では、クラスタリング部１２は、例えば、複数の文書２０ａ内の各文書を、いずれかのクラスタに分類する。

【００２５】

図５の例において、クラスタリング部１２は、複数の文書２０ａを、複数のクラスタＣ１１～Ｃ１４に分類する。クラスタＣ１１は、例えば、セブンイレブンの話題に関する文書を多く含む。また、クラスタＣ１２は、ウルトラマンのセブンの話題に関する文書を、クラスタＣ１３は煙草のマイルドセブンの話題に関する文書を多く含む。クラスタＣ１４は、セブンイレブン、ウルトラマン、煙草以外の話題に関する文書を含む。

【００２６】

次に、クラスタリング部１２は、例えば、クラスタＣ１１～Ｃ１４に基づいて、各クラスタを代表する語句（以下、代表語と称する）Ｃｋを抽出する。代表語Ｃｋとは、例えば、対象のクラスタにより多く頻出し、対象のクラスタ以外のクラスタにはほとんど含まれない単語である。例えば、クラスタリング部１２は、対象のクラスタへの出現頻度が高く、かつ、対象外のクラスタへの出現頻度が少ない単語を代表語として抽出する。図５の例において、クラスタＣ１１の代表語は、「コンビニ」「アイス」等である。また、クラスタＣ１２の代表語は、「ウルトラマン」「フィギュア」等であって、クラスタＣ１３の代表語は、「マイルドセブン」煙草」等である。ユーザは、代表語を、検索語と組み合わせる排除条件（ＮＯＴ）の候補として使用する。

【００２７】

例えば、ユーザは、代表語「フィギュア」を用いて、クエリ２３「セブン ａｎｄ ＮＯＴフィギュア」を生成する。これにより、ユーザは、検索語「セブン」を含み、かつ、代表語「フィギュア」を含まない文書の集合を抽出できる。つまり、ユーザは、ウルトラマ

ンの話題を示すクラスタＣ１２が含む文書の多くを検索結果から排除することができる。しかしながら、クエリ２３「セブン ａｎｄ ＮＯＴフィギュア」に基づいて文書集合を検索すると、例えば、「セブンイレブンでフィギュアを買った」等の文章を含む文書が検索結果から排除されてしまう。つまり、検索によって抽出したい文書についても、検索結果から排除されてしまう。

【００２８】

このように、クラスタの代表語を用いることによって、検索語による検索結果から不要な文書を排除するキーワード候補が取得可能になるものの、検索によって抽出したいユーザ所望の文書についても検索結果から排除されてしまうことがある。したがって、代表語を用いてクエリ２３を生成する場合であっても、所望の文書を抽出できるとは限らない。また、検索結果を分類した複数のクラスタから所望のクラスタを選択して文書を抽出する場合であっても、クラスタリングの精度が完全ではないことから、不要な文書が抽出されてしまう場合がある。

【００２９】

また、クエリ２３による文書検索では、根本的には、ユーザが所望する完全な文書集合を取得することは困難である。つまり、クエリ２３による文書検索によると、不要な文書を完全に排除することや、抽出したい文書を完全に選択することは困難である。したがって、ユーザは、所望の文書集合とできるだけ近い文書集合を抽出可能にするクエリ２３を生成する。ただし、ユーザは、所望する完全な文書の集合の内容を、予め検知していない場合がある。したがって、ユーザは、クラスタの代表語を用いて試行錯誤を重ねながら、所望の文書集合と近いと思われる文書集合を抽出可能なクエリ２３を生成する。

【００３０】

しかしながら、各クラスタの代表語を組み合わせたとしても、所望の文書集合と近い文書集合を抽出するクエリ２３を生成することは容易ではない。また、ユーザは所望の文書集合の内容を予め検知しているわけではないため、試行錯誤を重ねたとしても最適な文書集合が抽出されるとは限らない。

【００３１】

本実施の形態例における検索サーバ１０は、複数のクラスタを識別する情報を表示して、ユーザに、検索結果から排除すべき文書を多く有する排除クラスタを指定させる。そして、検索サーバ１０は、排除クラスタ内のキーワードを抽出し、抽出したキーワード毎の、検索された複数の文書における排除クラスタ内での出現分布率を示すスコアを計算し、スコアの降順にソートされたキーワードのキーワードリスト２２を生成する。そして、検索サーバ１０は、キーワードリスト２２内のキーワード数に比例する値を示すスライドバー等の操作オブジェクトを表示して、ユーザのスライドバー等の入力に応じて、キーワードリスト２２の上位数分のキーワードを排除条件のクエリ２３の候補として選択し、選択したキーワードを含む文書の割合をクラスタ毎に計算して表示する。

【００３２】

即ち、本実施の形態例における検索サーバ１０は、検索語による検索結果を分類した複数のクラスタから、排除すべき文書を多く有する排除クラスタをユーザに指定させる。検索サーバ１０は、排除クラスタの指定に基づいて、排除クラスタ内の文書をより多く排除可能であって、排除クラスタ以外のクラスタから排除される文書量を抑えるキーワードをその有効性の順に有するリストを生成する。したがって、ユーザは、排除クラスタを指定するだけで、検索サーバ１０に、クエリ２３の排除条件のキーワード候補（キーワードリスト２２）を生成させることができる。

【００３３】

そして、検索サーバ１０は、キーワードのリスト内のキーワード数に比例する値を示すスライドバーとともに、スライドバーの値に対応する上位数分のキーワードを含む文書の割合をクラスタ毎に表示する。これにより、ユーザは、スライドバーの値に応じて、排除クラスタ内で排除される文書の割合と、排除クラスタ以外のクラスタ内で排除される文書の割合とのバランスを確認しながら、スライドバーの値を指定することができる。つまり

、ユーザは、排除すべきクラスタから排除される文書の割合と、選択したいクラスタから排除される文書の割合とのバランスを確認できることにより、最適なバランスを選択することができる。

【００３４】

検索サーバ１０は、スライドバーの値が示す上位数分のキーワードを、クエリ２３の排除条件として選択しクエリ２３を生成する。したがって、ユーザは、キーワード自体を意識することなく、最適なクエリ２３を取得することができる。

【００３５】

このように、ユーザは、所望する完全な文書の集合の内容を予め検知していなくても、排除クラスタの指定とスライドバーによる指定とを行うだけで、ユーザが所望する文書集合を抽出可能にするクエリ２３を生成させることが可能になる。つまり、本実施の形態例における検索サーバ１０は、試行錯誤を重ねることなく、簡易な操作にしたがって、ユーザの意図を反映させた排除条件のキーワードの絞込みを可能にする。また、ユーザは、所望の文書集合に近い文書集合を確実に抽出可能になる。

【００３６】

次に、本実施の形態例におけるクエリ生成処理をフローチャート図に基づいて説明する。

【００３７】

［フローチャート］

図６は、本実施の形態例におけるクエリ生成処理を説明するフローチャート図である。初めに、検索サーバ１０の文書検索部１１は、クライアント装置８０から文書集合２０の検索を行うための検索語を受信する（Ｓ１１）。なお、検索語は、複数のキーワードによる組み合わせ（キーワード集合）であってもよい。次に、文書検索部１１は、検索語に基づいてクエリ２３を生成し、文書集合２０の検索処理を行う（Ｓ１２）。文書検索部１１は、検索処理の結果、クエリ２３が示す条件に合致する複数の文書２０ａを取得する。

【００３８】

次に、検索サーバ１０のクラスタリング部１２は、検索によって取得した複数の文書２０ａを、文書の類似性に基づいて複数のクラスタに分類する（Ｓ１２）。クラスタリング部１２は、例えば、ｋ－ｍｅａｎｓ、ワン・パスクラスタリング等の公知の技術を用いて、クラスタを生成する。例えば、本実施の形態例におけるクラスタリング部１２は、論文「Criterion Functions for Document Clustering Experiments and Analysis （文献名：Technical Report CS Dept. 01-40, Univ. Minnesota, Ying Zhao and George Karypis 2001年）」に記述されるクラスタリングの技術に基づいて、クラスタを生成する。

【００３９】

次に、キーワード生成部１３のクラスタ指定受付部３１は、クライアント装置８０の表示ユニットに、生成した複数のクラスタを識別する情報を表示させる（Ｓ１４）。複数のクラスタを識別する情報は、例えば、クラスタＩＤやアイコン等である。クラスタ指定受付部３１は、例えば、複数のクラスタを識別する情報をクライアント装置８０にダウンロードさせ、クライアント装置８０で動作するウェブブラウザを介して表示させる。また、クラスタ指定受付部３１は、さらに、複数のクラスタのうち、排除クラスタ、選択クラスタに対する指定を受け付けるオブジェクトを表示する（Ｓ１５）。オブジェクトは、例えば、ラジオボタンである。

【００４０】

また、クラスタ指定受付部３１は、選択クラスタ及び排除クラスタのユーザによる指定を可能にするために、各クラスタの特徴を表す特徴情報を表示する（Ｓ１６）。ユーザは、特徴情報を参照することによって、クラスタが有する話題を識別することができる。ユーザは、クラスタが有する話題に基づいて、クラスタが有する文書が検索結果から排除されることが望ましいか否か、クラスタが有する文書が検索結果として選択されることが望ましいか否かを判断可能になる。

【００４１】

特徴情報とは、例えば、クラスタの文書に含まれる検索語を使用する文字列の一部、クラスタの文書に含まれる頻出語、クラスタの文書に含まれる単語であって検索された複数の文書に対する文書集合への出現比率が高い単語（代表語）等である。ただし、特徴情報は、この例に限定されるものではなく、クラスタが有する文書の主題が識別可能になる情報であればいずれの情報であってもよい。

【００４２】

次に、クラスタ指定受付部３１は、ユーザによる選択クラスタ、排除クラスタの指定を受け付ける（Ｓ１７）。ユーザは、少なくとも、排除クラスタを１つ指定する。また、選択クラスタは、必ずしも指定されなくてもよい。ユーザは、排除すべき文書を多く有するクラスタを排除クラスタに指定する。また、ユーザは、検索結果として選択されることが望ましい文書を多く有するクラスタを選択クラスタに指定する。

【００４３】

ユーザによるクラスタの指定を受け付けると（Ｓ１８のＹＥＳ）、キーワード生成部１３のキーワードリスト生成部３２は、キーワードリスト２２を生成する（Ｓ１９）。キーワードリスト２２は、クエリ２３の排除条件の候補となる複数のキーワードを有する。また、キーワードリスト２２は、検索された文書に対する排除クラスタへの出現分布率が高いときにより大きい値を有するスコアの降順に、キーワードを有する。キーワードリスト２２の詳細については、別の図にしたがって後述する。

【００４４】

このように、ユーザは、複数のクラスタから排除クラスタを指定するだけで、クラスタに対する指定項目（排除、選択）を反映させた、クエリ２３の排除条件のキーワード候補を検索サーバ１０に生成させることができる。したがって、ユーザは、クエリ２３の排除条件のキーワードの候補を考える必要がない。

【００４５】

次に、キーワード生成部１３のキーワード選択部３３は、クライアント装置８０にスライドバーを表示させ（Ｓ２０）、ユーザによるスライドバーに対する操作を受け付ける（Ｓ２１）。スライドバーの値は、キーワードリスト２２内のキーワード数に比例する。つまり、ユーザは、スライドバーの値を変更させることによって、クエリ２３の排除条件となるキーワードの数を変動させることができる。

【００４６】

そして、キーワード選択部３３は、スライドバーの値に応じた、キーワードリスト２２の上位数分のキーワードをクエリ２３の排除条件の候補として選択し、選択したキーワードを含む文書の割合をクラスタ毎に計算する（Ｓ２２）。選択したキーワードを含む文書の割合とは、選択したキーワードが排除条件として適用されることによって、排除される文書の割合を表す。そして、キーワード選択部３３は、クラスタ毎に、選択したキーワードを含む文書の割合を表示する（Ｓ２３）。

【００４７】

キーワード選択部３３は、スライドバーが示す値が更新される度に、クラスタ毎の排除される文書の割合を計算し直して、表示する（Ｓ２２、Ｓ２３）。スライドバーの値が確定すると（Ｓ２４のＹＥＳ）、キーワード選択部３３は、検索語と、排除条件とする上位数分のキーワードとの組み合わせによって、クエリ２３を生成する（Ｓ２５）。そして、例えば、キーワード選択部３３は、検索語と排除条件である上位数分のキーワードとに基づいたクエリ２３をクライアント装置８０の表示ユニットや、検索サーバ１０のメモリ１０２に出力する。または、キーワード選択部３３は、検索語と排除条件である上位数分のキーワードとに基づいたクエリ２３にしたがって文書集合２０を検索し直し、文書集合を抽出する。

【００４８】

クエリ２３の排除条件の候補として選択するキーワードの数が変動することによって、各クラスタから排除される文書量も変動する。このとき、排除クラスタから排除される文書の量と、選択クラスタから抽出される文書の量とは、トレードオフの関係にある。具体

的に、キーワードの数を増加させて排除クラスタから排除される文書量を増加させるようとする、選択クラスタから抽出される文書量も低減する傾向にある。一方、キーワードの数を低減させて選択クラスタから抽出される文書量を増加させようとすると、排除クラスタから排除される文書量も低減してしまう傾向にある。

【００４９】

排除クラスタから排除される文書量と、選択クラスタから抽出される文書量との望ましいバランスは、検索ケースによって異なる。例えば、排除クラスタの文書を可能な限り検索結果から排除したい場合、選択クラスタから抽出される文書量が低下したとても、排除クラスタから排除される文書量が多い方が望ましい。したがって、排除クラスタの文書を可能な限り検索結果から排除したい場合、キーワードの数が多い方が、ユーザが所望する文書集合に近い文書集合を抽出可能になり易い。

【００５０】

一方、選択クラスタの文書を可能な限り検索結果として抽出したい場合、排除クラスタから排除される文書量が少なかったとしても、選択クラスタから抽出される文書量が多い方が望ましい。したがって、選択クラスタの文書を可能な限り検索結果として抽出したい場合、キーワードの数が少ない方が、ユーザが所望する文書集合に近い文書集合を抽出可能になり易い。

【００５１】

このように、排除クラスタから排除される文書量と選択クラスタから抽出される文書量との望ましいバランス数が検索ケースによって異なるところ、ユーザは、排除クラスタから排除される文書量と選択クラスタから抽出される文書量とのバランスを確認しながら、スライドバーの値を選択することができる。これにより、ユーザは、意図に沿った文書集合を抽出可能にするクエリ２３の排除条件のキーワード数を選択可能になる。また、ユーザは、スライドバー等のオブジェクトを操作するだけで、キーワード自体を意識することなくクエリ２３を生成させることが可能になる。

【００５２】

次に、図６で説明したフローチャート図の処理を具体例に対応させて説明する。

【００５３】

図７は、本実施の形態例におけるクラスタの生成処理を説明する図である。本実施の形態例における検索サーバ１０の文書検索部１１は、図４で説明した処理と同様にして、受け付けた検索語「セブン」に基づいて、文書集合２０を検索する（図６のＳ１１、Ｓ１２）。この結果、文書検索部１１は、文書内に検索語「セブン」を含む複数の文書２０ａを取得する。次に、クラスタリング部１２は、検索語「セブン」にしたがって検索された複数の文書２０ａを、複数のクラスタＣ１１～Ｃ１４に分類する（Ｓ１３）。クラスタＣ１１～Ｃ１４は、図４で説明したとおりである。

【００５４】

次に、クラスタ指定受付部３１は、クライアント装置８０の表示ユニットに、複数のクラスタを識別する情報を表示させるとともに（図６のＳ１４）、ラジオボタン等のオブジェクトを表示して、排除クラスタ及び選択クラスタへの指定を受け付ける（Ｓ１５）。このとき、クラスタ指定受付部３１は、選択クラスタ及び排除クラスタのユーザによる指定を可能にするために、各クラスタの特徴を表す特徴情報を表示する（Ｓ１６）。

【００５５】

図８は、クラスタの指定を受け付けるクライアント装置８０の表示ユニットが表示する画面の一例を示す図である。図８は、クラスタ毎に、クラスタを識別する文書マークで示されるアイコンに加えて、クラスタの指定を受け付けるボタンメニューＲ１～Ｒ４と、クラスタの特徴情報であるクラスタ内の検索語の使用例Ｔ１～Ｔ４とを有する。図８の例において、各クラスタＣ１１～Ｃ１４を識別する情報は、文書マークで示されるアイコンである。

【００５６】

また、ボタンメニューＲ１～Ｒ４は、「選択」「排除」「その他」のいずれかを指定さ

せるラジオボタンである。例えば、初め、全てのクラスタは、「その他」に指定される。「その他」に指定されるクラスタの文書は、キーワードリスト２２の生成処理に使用されない。処理の詳細については後述するが、キーワード生成部１３は、排除クラスタ、選択クラスタに基づいて、キーワードリスト２２を生成する。クラスタの話題が識別できない場合、ユーザは、例えば、クラスタを「その他」のクラスタとする。

【００５７】

図８の例において、クラスタ内の検索語「セブン」の使用例Ｔ１～Ｔ４は、クラスタ内の文書のうち、一部の文書における検索語の使用部分の文字列である。例えば、クラスタＣ１１の検索語の使用部分の文字列Ｔ１は、「この夏６、７、８、９月、私がよく購入した商品を発表したと思います。第１位　昆布とかつおのうま味・おでん［セブンイレブン］」である。また、図８に示すように、検索語の使用部分の文字列のうち検索語「セブン」は、例えば、太字等によって強調され表示される。なお、検索語は、例えば、斜体、下線等によって強調されて表示されてもよい。ユーザは、クラスタ内の検索語の使用例を参照することによって、クラスタ内の文書で検索語「セブン」がどのように参照されているかを確認可能になり、クラスタが有する話題を識別することができる。

【００５８】

なお、クラスタ内の検索語の使用例Ｔ１～Ｔ４は、ユーザがクラスタＣ１１～Ｃ１４それぞれからランダムに選択した文書内での検索語の使用例であってもよい。また、図６のフローチャート図で前述したとおり、特徴情報は、検索語の使用例Ｔ１～Ｔ４の他に、例えば、クラスタ内の頻出語や代表語等であってもよい。

【００５９】

図８の例において、クラスタＣ１２の検索語の使用例Ｔ２によるとクラスタＣ１２が「ウルトラマン」に関する話題を有することが識別可能になる。本実施の形態例では、「セブンイレブン」に関する文書を抽出することを目的とする。したがって、「ウルトラマン」の話題を有する文書は、検索結果として抽出したい文書に当たらない可能性があることから、ユーザは、例えば、クラスタＣ１２を排除クラスタとしてボタンメニューＲ２に指定する。また、クラスタＣ１３の検索語の使用例Ｔ３によると、クラスタＣ１３が「タバコ」に関する話題を有することが識別可能になる。同様にして、「タバコ」の話題を有する文書は、検索結果として抽出したい文書に当たらない可能性があることから、ユーザは、例えば、クラスタＣ１３を排除クラスタとしてボタンメニューＲ３に指定する。

【００６０】

図８の例では、排除クラスタに加えて、選択クラスタが指定される。クラスタＣ１１の検索語の使用例Ｔ１によると、クラスタＣ１１が「セブンイレブン」に関する話題を有することが識別可能になる。したがって、ユーザは、例えば、クラスタＣ１１を選択クラスタとしてボタンメニューＲ１に指定する。また、クラスタＣ１４は、特定の話題を有していない。したがって、ユーザは、例えば、クラスタＣ１４を、その他のクラスタとする（Ｒ４）。

【００６１】

図９は、キーワードリスト２２の生成処理を説明する図である。キーワードリスト生成部３２は、図８の画面において指定された排除クラスタＣ１２、Ｃ１３、及び、選択クラスタＣ１１が有する文書を入力として、クエリ２３の排除条件の候補となるスコア付きのキーワードのリストを生成する（図６のＳ１８、Ｓ１９）。排除クラスタのみが指定される場合、キーワードリスト生成部３２は、排除クラスタ内の文書に含まれるキーワードを抽出し、キーワードの複数のクラスタにおける排除クラスタ内での出現分布率に基づいてスコアを算出する。この場合、スコアは、複数のクラスタにおける排除クラスタ内での出現分布率が高いときにより大きい値となる。

【００６２】

また、排除クラスタに加えて選択クラスタが指定される場合、キーワードリスト生成部３２は、排除クラスタ及び選択クラスタ内の文書に含まれるキーワードを抽出し、キーワードの複数のクラスタにおける排除クラスタ内での出現分布率に基づいてスコアを算出す

る。この場合、スコアは、複数のクラスタにおける排除クラスタ内での出現分布率が高く、かつ、選択クラスタ内での出現分布率が低いときにより大きい値となる。

【００６３】

ここで、キーワードリスト２２の生成処理をより具体的に説明する。

【００６４】

［キーワードリストの生成］

図１０は、キーワードリスト２２の生成処理を説明するフローチャート図である。まず、キーワードリスト生成部３２は、排除クラスタ及び選択クラスタ内の文書に含まれるキーワードを複数抽出する（Ｓ３１）。そして、キーワードリスト生成部３２は、抽出したキーワードのスコアの値を０に初期化する（Ｓ３２）。

【００６５】

なお、各クラスタは、クラスタ内に含まれる単語毎にｔｆｉｄｆ（term frequency inverse document frequency）値を有する。ｔｆｉｄｆ値は、例えば、単語が、クラスタ内で特徴的である度合いを識別するための指標である。この例では、ｔｆｉｄｆ値は、対象のクラスタにより偏って出現する度合いを表す。ｔｆｉｄｆ値は、単語がクラスタ内に出現する回数「ｔｆ」と、全文書において当該単語が出現する文書数「ｄｆ」とに基づいて、計算式「ｔｆｉｄｆ＝ｔｆ／ｌｏｇ(ｄｆ／Ｎ)」にしたがって算出される。計算式内の値「Ｎ」は全文書数を表す。

【００６６】

ｔｆｉｄｆ値は、例えば、対象のクラスタへの出現頻度が高く、かつ、全クラスタ内への出現率が少ない場合により大きい値を有する。言い換えると、対象のクラスタへの出現頻度が高くても全クラスタ内での出現率が高い場合は、いずれのクラスタにも出現することを示すため、ｔｆｉｄｆ値は大きな値にはならない。単語毎のｔｆｉｄｆ値の例については、図１１で例示する。

【００６７】

そして、キーワードリスト生成部３２は、選択クラスタとして指定されたクラスタがあるか否かを判定する（Ｓ３３）。選択クラスタがある場合（Ｓ３３のＹＥＳ）、キーワードリスト生成部３２は、抽出したキーワードにしたがって、選択クラスタ内の各文書を検索する（Ｓ３４）。文書にキーワードが含まれる場合（Ｓ３５のＹＥＳ）、キーワードリスト生成部３２は、キーワードのｔｆｉｄｆ値をスコアから減算する（Ｓ３６）。つまり、キーワードが選択クラスタ内に出現する場合、排除条件のキーワードとして不適切である可能性が高いため、スコアの値は減算される。一方、キーワードが含まれない場合（Ｓ３５のＮＯ）、キーワードリスト生成部３２は、ｔｆｉｄｆ値をスコアから減算しない。キーワードリスト生成部３２は、選択クラスタ内の全ての文書について（Ｓ３７のＹＥＳ）、工程Ｓ３４～Ｓ３６の処理を行う。

【００６８】

次に、キーワードリスト生成部３２は、抽出したキーワードにしたがって、排除クラスタ内の各文書を検索する（Ｓ３８）。文書にキーワードが含まれる場合（Ｓ３９のＹＥＳ）、キーワードリスト生成部３２は、キーワードのｔｆｉｄｆ値をスコアに加算する（Ｓ４０）。つまり、キーワードが排除クラスタ内に出現する場合、排除条件のキーワードとして適切である可能性が高いため、スコアの値は加算される。一方、キーワードが含まれない場合（Ｓ３９のＮＯ）、キーワードリスト生成部３２は、ｔｆｉｄｆ値をスコアに加算しない。

【００６９】

キーワードリスト生成部３２は、排除クラスタ内の全ての文書について（Ｓ４１のＹＥＳ）、工程Ｓ３８～Ｓ４０の処理を行う。そして、キーワードリスト生成部３２は、スコアの降順にキーワードをソートする（Ｓ４２）。これにより、キーワードリスト生成部３２は、スコアの降順にキーワードを有するキーワードのリスト２２を生成する。なお、前述したように、キーワード生成部３２は、「その他」に指定されるクラスタ内の文書を加味することなく、キーワードリスト２２を生成する。

【００７０】
次に、キーワードリスト２２の生成処理を具体例に対応させて説明する。
【００７１】
図１１は、キーワード毎のｔｆｉｄｆ値を例示する図である。図１１の表Ｈ１は、クラスタＩＤ、文書ＩＤ、単語、ｔｆｉｄｆ値を有する。クラスタＩＤは、クラスタを識別する情報であって、文書ＩＤは、文書を識別する情報である。単語は、クラスタ内の文書に含まれる単語である。図１１の表Ｈ１において、例えば、ＩＤ「１」のクラスタＣ１は、文書ｄｏｃ１、ｄｏｃ２等を有する。また、例えば、ＩＤ「２」のクラスタＣ２は、文書ｄｏｃ３０３等を有する。なお、この例では、文書の一部の情報を表しているが、実際には、各クラスタは多数の文書を有する。
【００７２】
図１１の例において、文書ｄｏｃ１は、例えば、単語「コンビニ」「おにぎり」等を含む。この例では、一部の単語を表しているが、実際には、各文書は多数の単語を有する。単語「コンビニ」のｔｆｉｄｆ値は「４２．７」であって、単語「おにぎり」のｔｆｉｄｆ値は「４０．３」である。また、文書ｄｏｃ２は、例えば、単語「四国」「コンビニ」等を含み、単語「四国」のｔｆｉｄｆ値は「５８．７」であって、単語「コンビニ」のｔｆｉｄｆ値は「４２．７」である。つまり、単語「四国」は、単語「おにぎり」「コンビニ」よりも、ＩＤ「１」のクラスタＣ１により偏って出現することを意味する。
【００７３】
また、表Ｈ１において、文書ｄｏｃ３０３は、例えば、単語「コンビニ」「マイルドセブン」「煙草」等を含む。この例では、一部の単語を表しているが、実際には、各文書は多数の単語を有する。単語「コンビニ」のｔｆｉｄｆ値は「３８．１」、単語「マイルドセブン」のｔｆｉｄｆ値は「３７．８」、単語「煙草」のｔｆｉｄｆ値は「３３．６」である。ＩＤ「１」のクラスタＣ１内の文書ｄｏｃ１、ｄｏｃ２、ＩＤ「２」のクラスタＣ２内の文書ｄｏｃ３０３はいずれも、単語「コンビニ」のｔｆｉｄｆ値を有する。これは、単語「コンビニ」が、文書ｄｏｃ１、ｄｏｃ２、ｄｏｃ３０３のいずれにも含まれることを示す。また、ＩＤ「２」のクラスタの単語「コンビニ」のｔｆｉｄｆ値は、ＩＤ「１」のクラスタＣ１の単語「コンビニ」のｔｆｉｄｆ値より小さい。これは、単語「コンビニ」が、ＩＤ「２」のクラスタＣ２内の文書ｄｏｃ３０３よりも、ＩＤ「１」のクラスタＣ１内の文書ｄｏｃ１、ｄｏｃ２に、より偏って出現することを示す。
【００７４】
図１２は、キーワード生成処理の具体例を説明する第１の図である。図１２の表Ｈ１は、図１１の表Ｈ１と同一である。また、この例において、ＩＤ「１」のクラスタＣ１は選択クラスタ、ＩＤ「２」のクラスタＣ２は排除クラスタに該当する。
【００７５】
キーワードリスト生成部３２は、ＩＤ「１」の選択クラスタＣ１、及び、ＩＤ「２」の排除クラスタＣ２内の文書から、例えば、キーワード「コンビニ」「おにぎり」「四国」「マイルドセブン」「煙草」等を抽出する（図９のＳ３１）。そして、キーワードリスト生成部３２は、抽出した各キーワードのスコアを０に初期化したキーワードリスト２２を生成する（Ｓ３２）。具体例において、選択クラスタ（ＩＤ「１」）が存在することから（Ｓ３３のＹＥＳ）、キーワードリスト生成部３２は、キーワード「コンビニ」「おにぎり」「四国」「マイルドセブン」「煙草」にしたがって、ＩＤ「１」の選択クラスタＣ１内の文書を検索する（Ｓ３４）。
【００７６】
具体例において、ＩＤ「１」の選択クラスタＣ１内の文書ｄｏｃ１は、キーワード「コンビニ」「おにぎり」を含む（Ｓ３５のＹＥＳ）。したがって、キーワードリスト生成部３２は、キーワード「コンビニ」のｔｆｉｄｆ値「４２．７」（Ｙ１１）、キーワード「おにぎり」のｔｆｉｄｆ値「４０．３」（Ｙ１２）をそれぞれスコアから減算する（Ｓ３６）。したがって、図１２のキーワードリスト２２－１におけるキーワード「コンビニ」のスコアは値「－４２．７」（Ｙ１３）、キーワード「おにぎり」のスコアは値「－４０

．３」（Ｙ１４）となる。

【００７７】

図１３は、キーワード生成処理の具体例を説明する第２の図である。図１３の表Ｈ１は、図１１の表Ｈ１と同一である。具体例において、ＩＤ「１」の選択クラスタＣ１内の文書ｄｏｃ２は、キーワード「四国」「コンビニ」を含む（Ｓ３５のＹＥＳ）。したがって、キーワードリスト生成部３２は、キーワード「四国」のｔｆｉｄｆ値「５８．４」（Ｙ２１）、キーワード「コンビニ」のｔｆｉｄｆ値「４２．７」（Ｙ２２）をそれぞれスコアから減算する（Ｓ３６）。したがって、図１３のキーワードリスト２２－２におけるキーワード「コンビニ」のスコアは値「－８５．４（＝－４２．７－４２．７）」（Ｙ２３）、キーワード「四国」のスコアは値「－５８．４」（Ｙ２４）となる。

【００７８】

図１４は、キーワード生成処理の具体例を説明する第３の図である。図１４の表Ｈ１は、図１１の表Ｈ１と同一である。次に、キーワードリスト生成部３２は、抽出したキーワード「コンビニ」「おにぎり」「四国」「マイルドセブン」「煙草」にしたがって、ＩＤ「２」の排除クラスタＣ２内の各文書を検索する（Ｓ３８）。具体例において、ＩＤ「２」の排除クラスタＣ２内の文書ｄｏｃ３０３は、キーワード「コンビニ」「マイルドセブン」「煙草」を含む（Ｓ３５のＹＥＳ）。したがって、キーワードリスト生成部３２は、キーワード「コンビニ」のｔｆｉｄｆ値「３８．１」（Ｙ３１）をスコアに加算する（Ｓ３６）。したがって、図１４のキーワードリスト２２－３におけるキーワード「コンビニ」のスコアは値「（－４７．３＝－８５．４＋３８．１）」（Ｙ３２）となる。

【００７９】

図１５は、キーワード生成処理の具体例を説明する第４の図である。図１５の表Ｈ１は、図１１の表Ｈ１と同一である。次に、キーワードリスト生成部３２は、キーワード「マイルドセブン」のｔｆｉｄｆ値「３７．８」（Ｙ４１）をスコアに加算するとともに、キーワード「煙草」のｔｆｉｄｆ値「３３．６」（Ｙ４２）をスコアに加算する（Ｓ３６）。したがって、図１３のキーワードリスト２２－４におけるキーワード「マイルドセブン」のスコアは値「３７．８」（Ｙ４３）、キーワード「煙草」のスコアは値「３３．６」（Ｙ４４）となる。

【００８０】

図１６は、具体例において生成されるキーワードリスト２２－５を示す図である。図１６のキーワードリスト２２－５は、スコアの降順にキーワードを有する。図１６のキーワードリスト２２－５において、最もスコアの高いキーワードは「マイルドセブン」である。これは、キーワード「マイルドセブン」が、クエリ２３の排除条件として有効性が高いことを示す。次に、排除条件として有効性が高いキーワードは、「煙草」である。

【００８１】

このように、キーワードリスト生成部３２は、ユーザから指定された排除クラスタ、選択クラスタに基づいて、排除クラスタ内への出現分布率が高く、かつ、選択クラスタ内での出現分布率が低いときにより大きい値を有するスコアの降順にキーワードを有するキーワードリスト２２を生成することができる。なお、図１１～図１６の具体例では、排除クラスタに加えて選択クラスタが指定される場合を例示しているが、排除クラスタのみが指定される場合、キーワードリスト生成部３２は、排除クラスタへの出現分布率が高いときにより大きい値を有するスコアの降順にキーワードを有するキーワードリスト２２を生成する。

【００８２】

図１７は、本実施の形態例におけるキーワードの選択処理を説明する図である。図１６のようなキーワードリスト２２を生成すると、キーワード選択部３３は、キーワードリスト２２内のキーワード数の比例した値を示すスライドバー等のオブジェクトを、クライアント装置８０に表示させる（図６のＳ２０）。そして、キーワード選択部３３は、ユーザのスライドバーに対する操作にしたがって、キーワードリスト２２の上位数分のキーワード２２ａを排除条件のクエリ２３候補として選択する（Ｓ２２）。キーワードリスト２２

は、スコアの降順にキーワードを有する。したがって、キーワードリスト２２の上位から順に排除条件とするキーワードが選択されることによって、排除条件とするキーワードを効率的に選択することが可能になる。

【００８３】

図１８は、スライドバーＳＢを表示するクライアント装置８０の表示ユニットが表示する画面の一例を示す図である。図１８は、クラスタを識別する情報とクラスタの指定を示すボタンメニューＲ１～Ｒ４とに加えて、スライドバーＳＢを有する。なお、キーワード選択部３３は、例えば、スライドバーＳＢの代わりに、キーワードリスト２２内のキーワード数に比例する複数の項目（例えば、高、中、低等）を表示するドロップダウンリストや、ボタン等を表示して、ユーザに選択させてもよい。

【００８４】

また、図１８の例において、例えば、スライドバーＳＢの左端の値に対応するキーワードの数は０個である。一方、スライドバーＳＢの右端の値に対応するキーワードの数は、例えば、排除クラスタ内の文書がすべて検索結果から排除される上位数分のキーワード数に対応する。この場合、スライドバーＳＢの値が右端に設定される場合のキーワードの数は、検索ケースによって異なる。ただし、スライドバーＳＢの右端の値に対応するキーワードの数は、所定の値に予め設定されていてもよいし、スコアが所定値以上のキーワードの数であってもよい。

【００８５】

ユーザは、例えば、図１８に示すスライドバーＳＢのノブ（つまみ）ｐｐの位置を変化させることによって、スライドバーＳＢの示す値を変更させる。図１８の例において、スライドバーＳＢのノブｐｐを右方向に変更させた場合、スライドバーＳＢの値に応じて選択されるキーワード数が増加する。

【００８６】

図１９は、排除される文書の割合を表示するクライアント装置８０の表示ユニットが表示する画面の一例を示す図である。図１９は、スライドバーＳＢに加えて、クラスタ毎に、スライドバーＳＢによって示される排除条件のキーワードが適用された場合に排除される文書の割合を示す棒グラフＥＢ１～ＥＢ４を表示する。図１９の例において、棒グラフＥＢ１～ＥＢ４に加えて、クラスタが排除クラスタであるか、選択クラスタであるかが識別可能に表示されることにより、ユーザは、排除クラスタの文書がどの程度排除され、選択クラスタの文書がどの程度、排除されずに残るかを確認することができる。

【００８７】

したがって、ユーザは、スライドバーの値に応じて、排除すべきクラスタから排除される文書の割合と、選択したいクラスタから排除される文書の割合とのバランスを確認できる。ユーザは、排除クラスタ内で排除される文書の割合と、排除クラスタ以外のクラスタ内で排除される文書の割合とのバランスを確認しながら、最適なバランスを実現するスライドバーの値を指定することができる。したがって、ユーザは、スライドバーの値に対応する、所望の文書に近い文書集合を抽出可能な排除条件のキーワードを取得することができる。

【００８８】

なお、ユーザは、その他に指定されるクラスタの文書がどの程度、排除されるかを検知することによって、その他に指定されるクラスタとして指定されるクラスタが有する特徴を識別することが可能になる。クラスタの文書が有する特徴を識別することが可能になることによって、ユーザは、その他に指定されるクラスタを、例えば、排除クラスタや選択クラスタに指定し直すことが可能になる。これにより、ユーザは、クラスタの指定と、スライドバー等のオブジェクト操作によるキーワード数の調整とを繰り返すことによって、所望の文書により近い文書の集合を抽出することができる。

【００８９】

以上のように、本実施の形態例におけるクエリ生成方法は、処理ユニットが、入力された検索語に基づいて複数の文書を検索し、検索された複数の文書を類似度にしたがって複

数の文書集合に分類し、複数の文書集合を識別する情報を表示ユニットに表示する第１の工程を有する。また、本実施の形態例におけるクエリ生成方法は、表示された複数の文書集合のうち排除すべき文書集合として指定された、排除文書集合内の文字列を抽出し、抽出した文字列毎の、検索された複数の文書における排除文書集合内での出現分布率を示すスコアを計算し、スコアの降順にソートされた文字列の文字列リストを生成する第２の工程を有する。また、本実施の形態例におけるクエリ生成方法は、文字列リスト内の文字列数に比例する入力に応じて、文字列リストの上位数分の文字列を排除条件のクエリ２３の候補として選択し、選択した文字列を含む文書の割合を文書集合毎に計算し表示ユニットに表示する第３の工程を有する。

【００９０】

したがって、ユーザは、排除クラスタを指定するだけで、検索サーバ１０に、クエリ２３の排除条件のキーワード候補（キーワードリスト２２）を生成させることができる。また、ユーザは、スライドバーの値に応じて、排除クラスタ内で排除される文書の割合と、排除クラスタ以外のクラスタ内で排除される文書の割合とのバランスを確認しながら、スライドバーの値を指定することができる。つまり、ユーザは、排除すべきクラスタから排除される文書の割合と、選択したいクラスタから排除される文書の割合とのバランスを確認できることにより、最適なバランスを選択することができる。また、ユーザは、キーワード自体を意識することなく、最適なクエリ２３を取得することができる。

【００９１】

したがって、ユーザは、所望する完全な文書の集合の内容を予め検知していなくても、排除クラスタの指定とスライドバーによる指定とを行うだけで、ユーザが所望する文書集合を抽出可能にするクエリ２３を生成させることが可能になる。つまり、本実施の形態例における検索サーバ１０は、試行錯誤を重ねることなく、簡易な操作にしたがって、ユーザの意図を反映させた排除条件のキーワードの絞込みを可能にする。また、ユーザは、所望の文書集合に近い文書集合を確実に抽出可能になる。

【００９２】

また、本実施の形態例におけるクエリ生成方法によると、第２の工程は、抽出した文字列毎に、排除文書集合内の当該文字列の頻度に応じて、検索された複数の文書における排除文書集合内での出現分布値を加算してスコアを算出する。これにより、全てのクラスタにおける排除クラスタ内での出現分布率が高いときにより大きい値を有するスコアを算出可能になる。

【００９３】

また、本実施の形態例におけるクエリ生成方法によると、第２の工程は、抽出した文字列毎に、さらに、検索された複数の文書における排除文書集合以外の文書集合内での出現分布値をスコアから減算して、スコアを算出する。これにより、全てのクラスタにおける排除クラスタ内での出現分布率が高く、かつ、選択クラスタ内での出現分布率が低いときにより大きい値を有するスコアを算出可能になる。

【００９４】

また、本実施の形態例におけるクエリ生成方法によると、第３の工程は、文字列リスト内の文字列数に比例する値を示すオブジェクトを表示ユニットに表示し、オブジェクトに対するユーザの入力に比例する、文字列リストの上位数分の文字列を選択する。これにより、ユーザは、オブジェクトを操作することによって、クエリの排除条件として選択するキーワード自体を意識することなく、キーワードの数を指定することができる。

【００９５】

また、本実施の形態例におけるクエリ生成方法によると、第２の工程は、文書集合の特徴を示す特徴情報を表示ユニットに更に表示する。また、特徴情報は、文書集合の文書に含まれる検索語の使用文字列、文書集合の文書に含まれる頻出文字列、文書集合の文書に含まれる文字列であって検索された複数の文書における書集合内での出現分布率が高い文字列、のうち少なくともいずれかである。ユーザは、クラスタの特徴情報を参照することにより、クラスタが有する話題を識別可能になり、複数のクラスタから検索結果から排除

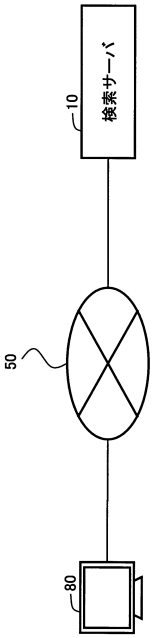
すべきクラスタ、及び、検索結果に残したいクラスタを指定することができる。
【００９６】
また、本実施の形態例におけるクエリ生成方法によると、第２の工程は、排除文書集合の指定を受け付けるオブジェクトを更に表示ユニットに表示する。これにより、ユーザは、オブジェクトを操作することによって、複数のクラスタのうち、検索結果から排除すべきクラスタを簡易に指定することができる。
【００９７】
［他の実施の形態例］
なお、本実施の形態例における検索サーバ１０は、図１８、図１９に示すスライドバー等のオブジェクトに加えて、ユーザのオブジェクト操作による入力に応じて選択されるキーワードを表示してもよい。例えば、検索サーバ１０のキーワード選択部３３は、クライアント装置８０の表示ユニットに、スライドバーの値に応じて選択される排除条件のキーワードの一覧を更に表示する。
【００９８】
これにより、ユーザは、スライドバー等のオブジェクトの値の変化に応じて選択されるキーワードと、当該キーワードを含むクラスタ毎の文書量とを同時に把握しながら、オブジェクトの値を調整することができる。したがって、クエリの排除条件となるキーワード自体を把握しながらオブジェクトを操作して、キーワード数を選択したいユーザにとって利便性が高い。
【００９９】
なお、上記の例では、検索サーバ１０が、複数のクラスタを生成し、排除クラスタ、選択クラスタのユーザによる指定に基づいて、キーワードリスト２２を生成し、ユーザのスライドバー等のオブジェクトの操作に応じて選択されるキーワードを含む文書の量をクラスタ毎に表示する。ただし、例えば、検索サーバ１０が検索対象となる文書集合２０を格納し、クライアント装置８０が、文書集合２０の検索結果に基づいて、クラスタの生成、キーワードリスト２２の生成、及び、ユーザのオブジェクトの操作に応じて選択されるキーワードを含む文書の量の表示を行ってもよい。
【０１００】
以上の実施の形態をまとめると、次の付記のとおりである。
【０１０１】
（付記１）
処理ユニットが、
入力された検索語に基づいて複数の文書を検索し、前記検索された複数の文書を類似度にしたがって複数の文書集合に分類し、前記複数の文書集合を識別する情報を表示ユニットに表示する第１の工程と、
前記表示された複数の文書集合のうち排除すべき文書集合として指定された、排除文書集合内の文字列を抽出し、前記抽出した文字列毎の、前記検索された複数の文書における前記排除文書集合内での出現分布率を示すスコアを計算し、前記スコアの降順にソートされた前記文字列の文字列リストを生成する第２の工程と、
前記文字列リスト内の文字列数に比例する入力に応じて、前記文字列リストの上位数分の前記文字列を排除条件のクエリの候補として選択し、前記選択した文字列を含む文書の割合を文書集合毎に計算し前記表示ユニットに表示する第３の工程と、
を実行することを特徴とするクエリ生成方法。
【０１０２】
（付記２）
付記１において、
前記第２の工程は、前記抽出した文字列毎に、前記排除文書集合内の当該文字列の頻度に応じて、前記検索された複数の文書における前記排除文書集合内での前記出現分布値を加算して前記スコアを算出するクエリ生成方法。
【０１０３】

（付記３）
付記２において、
前記第２の工程は、前記抽出した文字列毎に、さらに、前記検索された複数の文書における前記排除文書集合以外の文書集合内での前記出現分布値を前記スコアから減算して、前記スコアを算出するクエリ生成方法。
【０１０４】
（付記４）
付記１乃至３のいずれかにおいて、
前記第３の工程は、前記文字列リスト内の文字列数に比例する値を示すオブジェクトを前記表示ユニットに表示し、前記オブジェクトに対する前記ユーザの入力に比例する、前記文字列リストの上位数分の前記文字列を選択するクエリ生成方法。
【０１０５】
（付記５）
付記４において、
前記第３の工程は、前記オブジェクトに対する入力に応じて選択される前記文字列を前記表示ユニットに更に表示するクエリ生成方法。
【０１０６】
（付記６）
付記４または５において、
前記オブジェクトは、スライドバーであるクエリ生成方法。
【０１０７】
（付記７）
付記１乃至６のいずれかにおいて、
前記第２の工程は、前記文書集合の特徴を示す特徴情報を前記表示ユニットに更に表示するクエリ生成方法。
【０１０８】
（付記８）
付記７において、
前記特徴情報は、前記文書集合の文書に含まれる前記検索語の使用文字列、前記文書集合の文書に含まれる頻出文字列、前記文書集合の文書に含まれる文字列であって前記検索された複数の文書における前記書集合内での出現分布率が高い文字列、のうち少なくともいずれかであるクエリ生成方法。
【０１０９】
（付記９）
付記１乃至８のいずれかにおいて、
前記第２の工程は、前記排除文書集合の指定を受け付けるオブジェクトを更に前記表示ユニットに表示するクエリ生成方法。
【０１１０】
（付記１０）
入力された検索語に基づいて複数の文書を検索し、前記検索された複数の文書を類似度にしたがって複数の文書集合に分類し、前記複数の文書集合を識別する情報を表示ユニットに表示し、
前記表示された複数の文書集合のうち排除すべき文書集合として指定された、排除文書集合内の文字列を抽出し、前記抽出した文字列毎の、前記検索された複数の文書における前記排除文書集合内での出現分布率を示すスコアを計算し、前記スコアの降順にソートされた前記文字列の文字列リストを生成し、
前記文字列リスト内の文字列数に比例する入力に応じて、前記文字列リストの上位数分の前記文字列を排除条件のクエリの候補として選択し、前記選択した文字列を含む文書の割合を文書集合毎に計算し前記表示ユニットに表示する、
処理をコンピュータに実行させるクエリ生成プログラム。

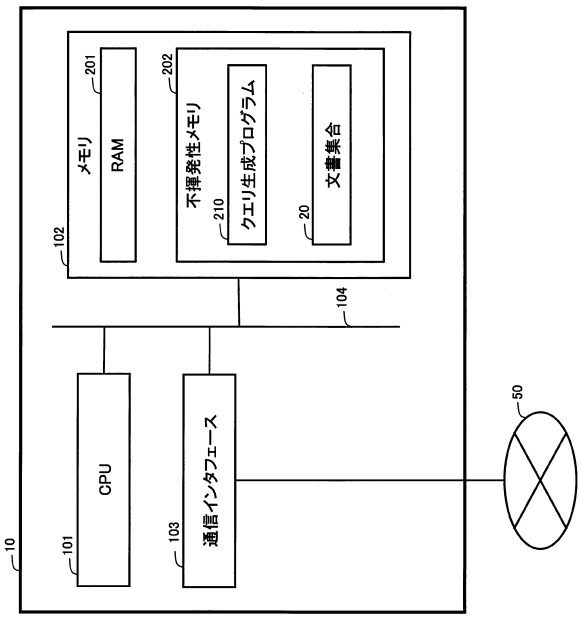
【０１１１】
（付記１１）
付記１０において、
前記抽出した文字列毎に、前記排除文書集合内の当該文字列の頻度に応じて、前記検索された複数の文書における前記排除文書集合内での前記出現分布値を加算して前記スコアを算出するクエリ生成プログラム。
【０１１２】
（付記１２）
付記１１において、
前記抽出した文字列毎に、さらに、前記検索された複数の文書における前記排除文書集合以外の文書集合内での前記出現分布値を前記スコアから減算して、前記スコアを算出するクエリ生成プログラム。
【０１１３】
（付記１３）
付記１０乃至１２のいずれかにおいて、
前記文字列リスト内の文字列数に比例する入力に応じて、前記文字列リスト内の文字列数に比例する値を示すオブジェクトを前記表示ユニットに表示し、前記オブジェクトに対する前記ユーザの入力に比例する、前記文字列リストの上位数分の前記文字列を選択するクエリ生成プログラム。
【０１１４】
（付記１４）
付記１３において、
前記オブジェクトに対する入力に応じて選択される前記文字列を前記表示ユニットに更に表示するクエリ生成プログラム。
【０１１５】
（付記１５）
付記１３または１４において、
前記オブジェクトは、スライドバーであるクエリ生成プログラム。
【０１１６】
（付記１６）
処理ユニットと、
複数の文書を記憶する記憶装置と、
表示装置と、を有し、
入力された検索語に基づいて前記複数の文書を検索し、前記検索された複数の文書を類似度にしたがって複数の文書集合に分類し、前記複数の文書集合を識別する情報を表示ユニットに表示し、前記表示された複数の文書集合のうち排除すべき文書集合として指定された、排除文書集合内の文字列を抽出し、前記抽出した文字列毎の、前記検索された複数の文書における前記排除文書集合内での出現分布率を示すスコアを計算し、前記スコアの降順にソートされた前記文字列の文字列リストを生成し、前記文字列リスト内の文字列数に比例する入力に応じて、前記文字列リストの上位数分の前記文字列を排除条件のクエリの候補として選択し、前記選択した文字列を含む文書の割合を文書集合毎に計算し前記表示ユニットに表示するクエリ生成装置。
【０１１７】
（付記１７）
付記１６において、
前記抽出した文字列毎に、前記排除文書集合内の当該文字列の頻度に応じて、前記検索された複数の文書における前記排除文書集合内での前記出現分布値を加算して前記スコアを算出するクエリ生成方法。
【０１１８】
（付記１８）

付記１７において、
前記抽出した文字列毎に、さらに、前記検索された複数の文書における前記排除文書集合以外の文書集合内での前記出現分布値を前記スコアから減算して、前記スコアを算出するクエリ生成方法。
【０１１９】
（付記１９）
付記１６乃至１８のいずれかにおいて、
前記文字列リスト内の文字列数に比例する入力に応じて、前記文字列リスト内の文字列数に比例する値を示すオブジェクトを前記表示ユニットに表示し、前記オブジェクトに対する前記ユーザの入力に比例する、前記文字列リストの上位数分の前記文字列を選択するクエリ生成方法。
【０１２０】
（付記２０）
付記１９において、
前記オブジェクトに対する入力に応じて選択される前記文字列を前記表示ユニットに更に表示するクエリ生成方法。
【符号の説明】
【０１２１】
１０：検索サーバ、８０：クライアント装置、２１０：クエリ生成プログラム、Ｃ１～Ｃ４：クラスタ、１１：文書検索部、１２：クラスタリング部、１３：キーワード生成部、３１：クラスタ指定受付部、３２：キーワードリスト生成部、３３：キーワード選択部

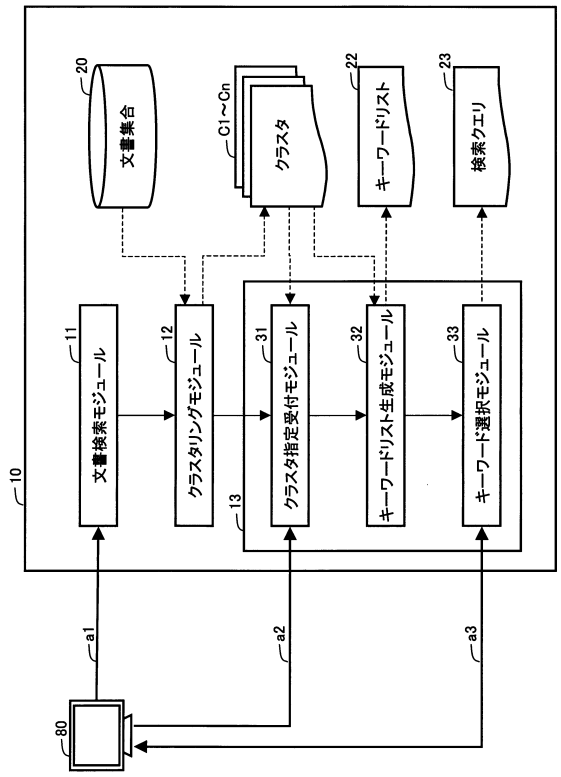
【図１】



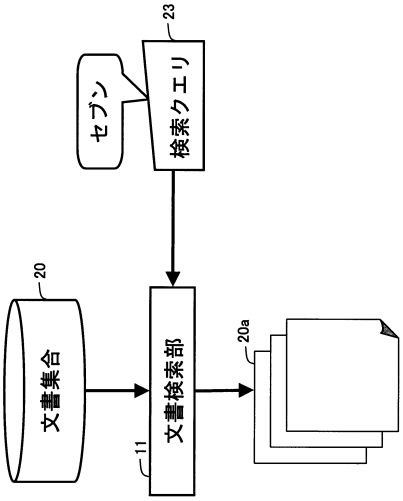
【図２】



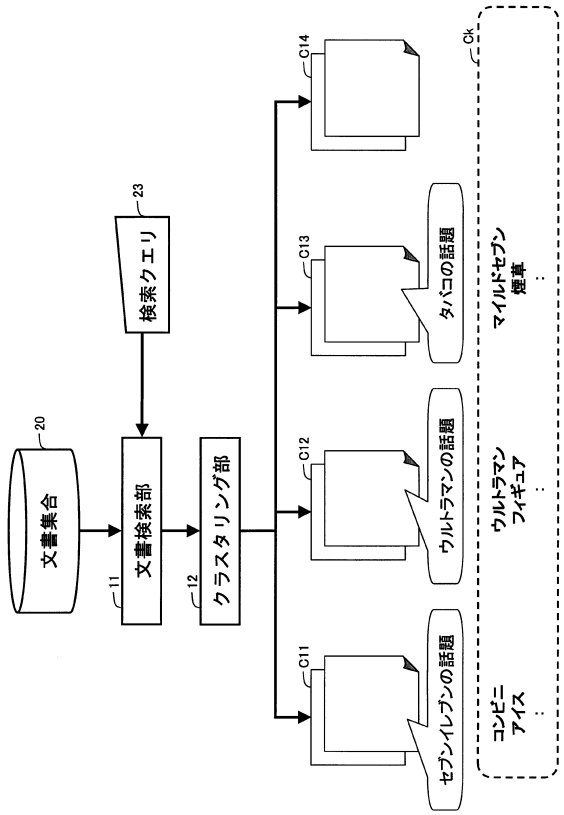
【図３】


10
文書検索モジュール
11
クラスタリングモジュール
12
13
クラスタ指定受付モジュール
31
キーワードリスト生成モジュール
32
キーワード選択モジュール
33
20
文書集合
C1~Cn
クラスタ
22
キーワードリスト
23
検索クエリ
a1
a2
a3
80

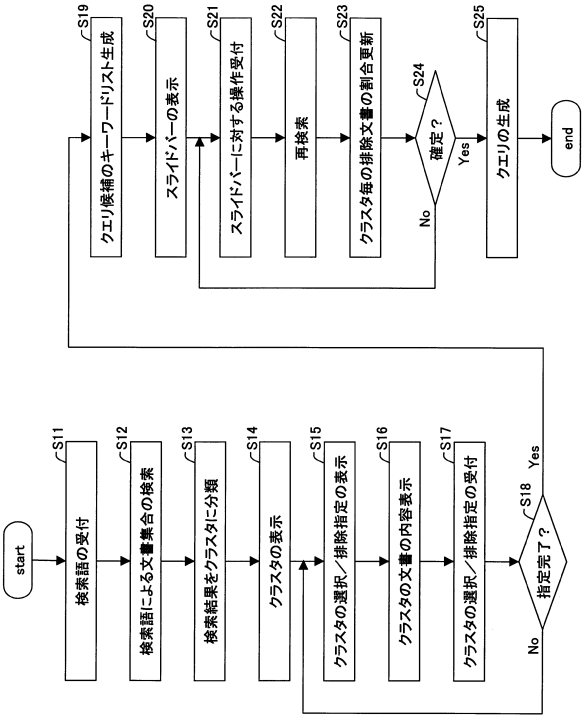
【図４】


セブン
検索クエリ
23
20
文書集合
文書検索部
11
20a

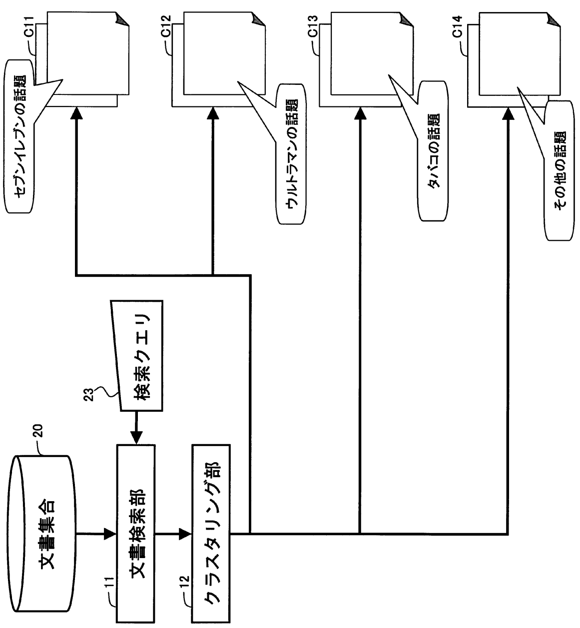
【図５】


20
文書集合
検索クエリ
23
文書検索部
11
クラスタリング部
12
C11
C12
C13
C14
Ck
セブンイレブンの話題
コンビニ
アイス
ウルトラマンの話題
ウルトラマン
フィギュア
タバコの話題
マイルドセブン
煙草

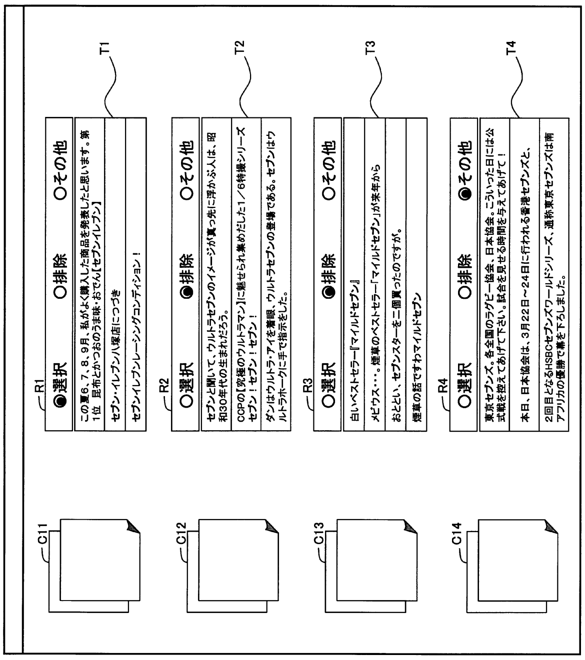
【図６】


start
S11
検索語の受付
S12
検索語による文書集合の検索
S13
検索結果をクラスタに分類
S14
クラスタの表示
S15
クラスタの選択／排除指定の表示
S16
クラスタの文書の内容表示
S17
クラスタの選択／排除指定の受付
S18
指定完了？
No
Yes
S19
クエリ候補のキーワードリスト生成
S20
スライドバーの表示
S21
スライドバーに対する操作受付
S22
再検索
S23
クラスタ毎の排除文書の割合更新
S24
確定？
No
Yes
S25
クエリの生成
end

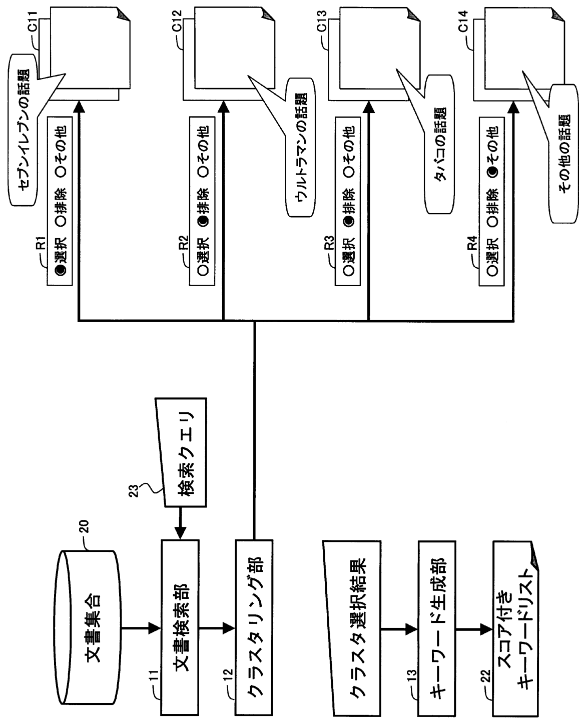
【図７】


文書集合
20
文書検索部
11
クラスタリング部
12
検索クエリ
23
C11
C12
C13
C14
セブンイレブンの話題
ウルトラマンの話題
タバコの話題
その他の話題

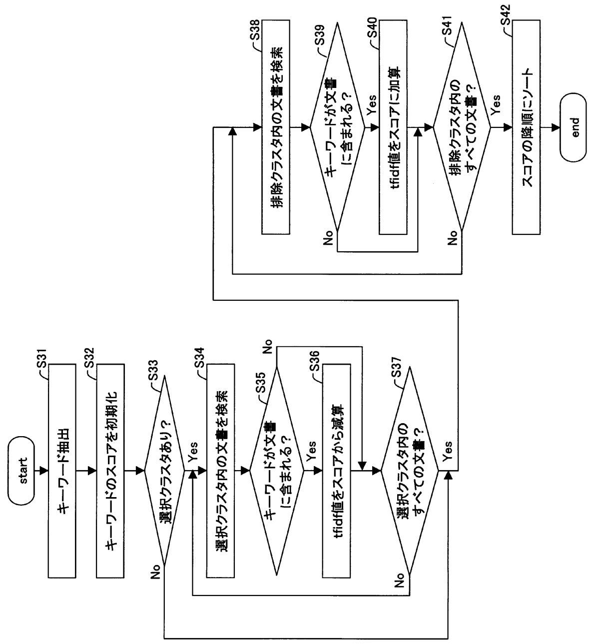
【図８】


C11
C12
C13
C14
R1
R2
R3
R4
T1
T2
T3
T4
選択
排除
その他

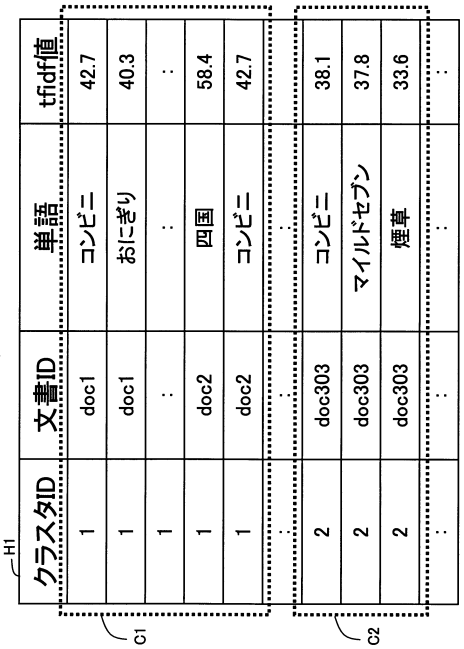
【図９】


文書集合
20
文書検索部
11
クラスタリング部
12
検索クエリ
23
クラスタ選択結果
キーワード生成部
13
スコア付き
キーワードリスト
22
C11
C12
C13
C14
R1
R2
R3
R4
選択
排除
その他
セブンイレブンの話題
ウルトラマンの話題
タバコの話題
その他の話題

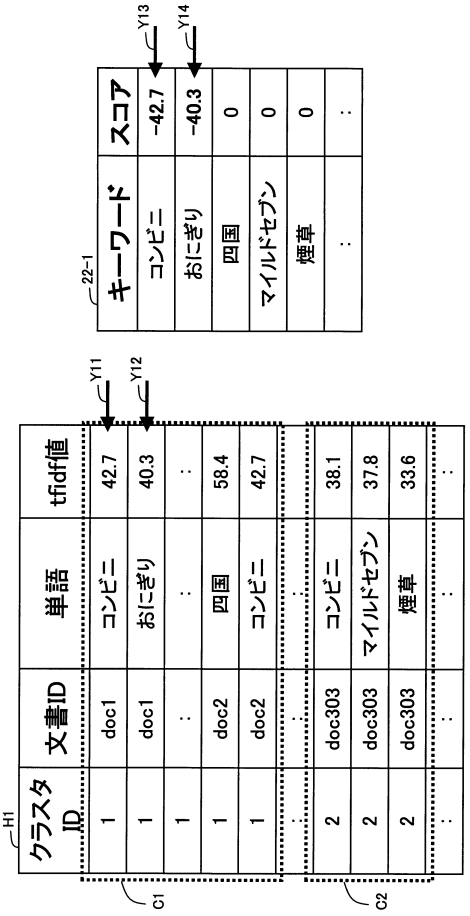
【図１０】


start
キーワード抽出
S31
キーワードのスコアを初期化
S32
選択クラスタあり？
S33
選択クラスタ内の文書を検索
S34
キーワードが文書に含まれる？
S35
tfidf値をスコアから減算
S36
選択クラスタ内のすべての文書？
S37
排除クラスタ内の文書を検索
S38
キーワードが文書に含まれる？
S39
tfidf値をスコアに加算
S40
排除クラスタ内のすべての文書？
S41
スコアの降順にソート
S42
end
Yes
No

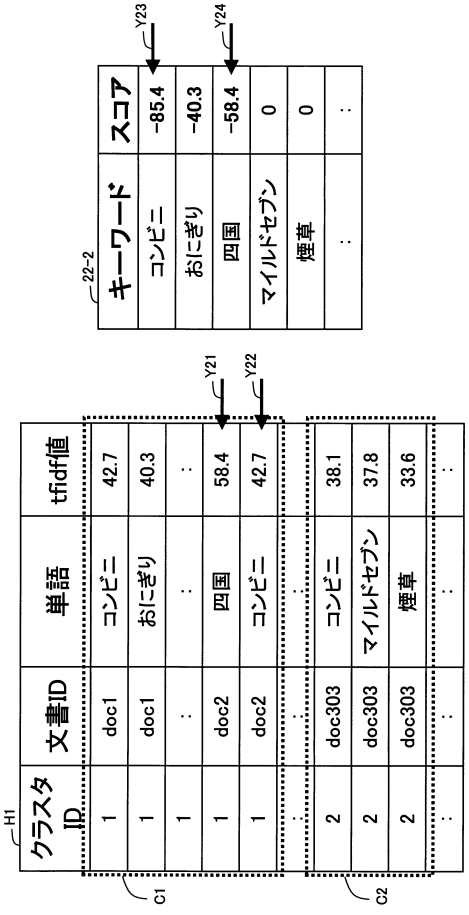
【図１１】



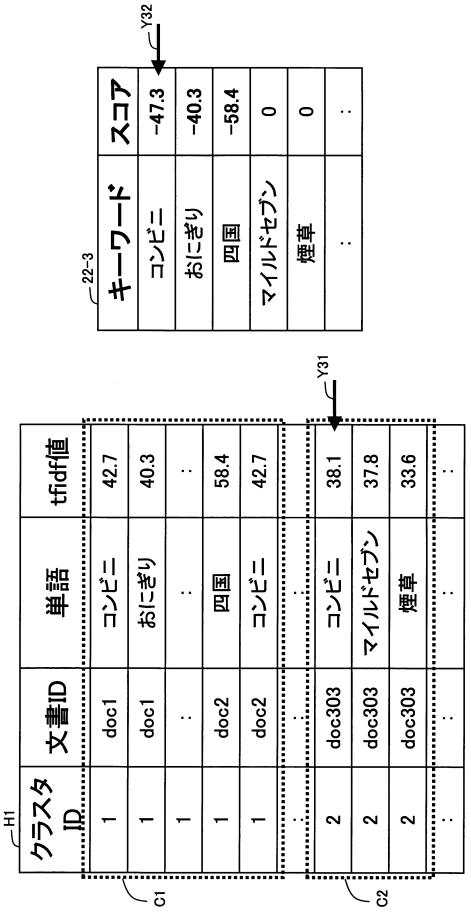
【図１２】



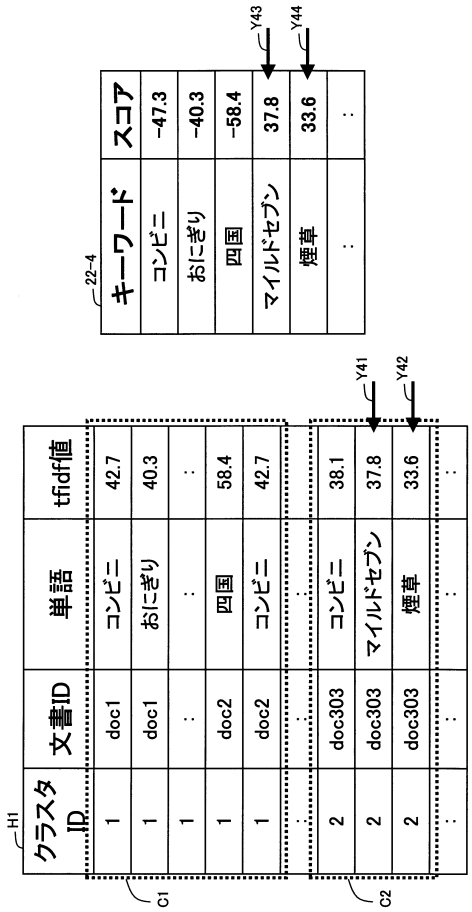
【図１３】



【図１４】



【図１５】



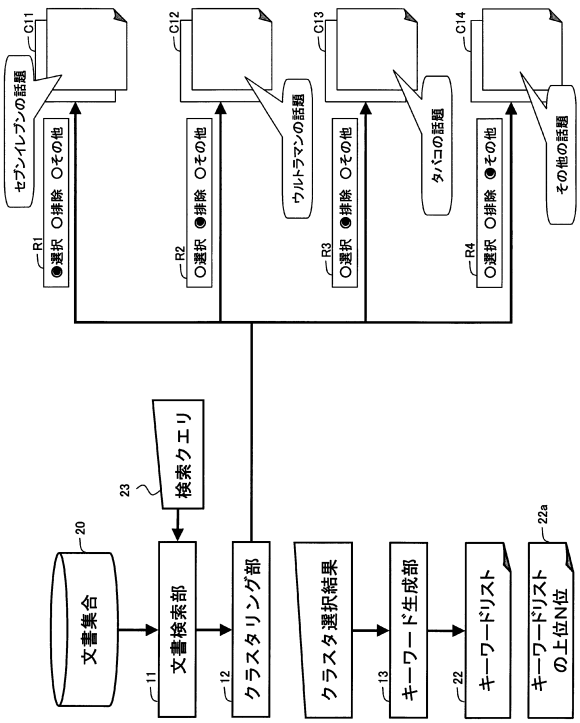
クラスタID	文書ID	単語	tfidf値
1	doc1	コンビニ	42.7
1	doc1	おにぎり	40.3
1	：	：	：
1	doc2	四国	58.4
1	doc2	コンビニ	42.7
：	：	：	：
2	doc303	コンビニ	38.1
2	doc303	マイルドセブン	37.8
2	doc303	煙草	33.6
：	：	：	：

キーワード	スコア
コンビニ	−47.3
おにぎり	−40.3
四国	−58.4
マイルドセブン	37.8
煙草	33.6
：	：

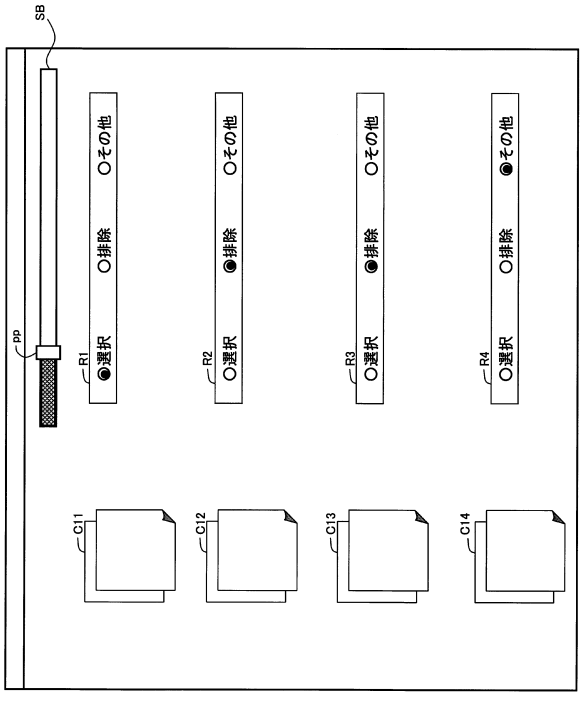
【図１６】

キーワード	スコア
マイルドセブン	37.8
煙草	33.6
おにぎり	−40.3
コンビニ	−47.3
四国	−58.4
：	：

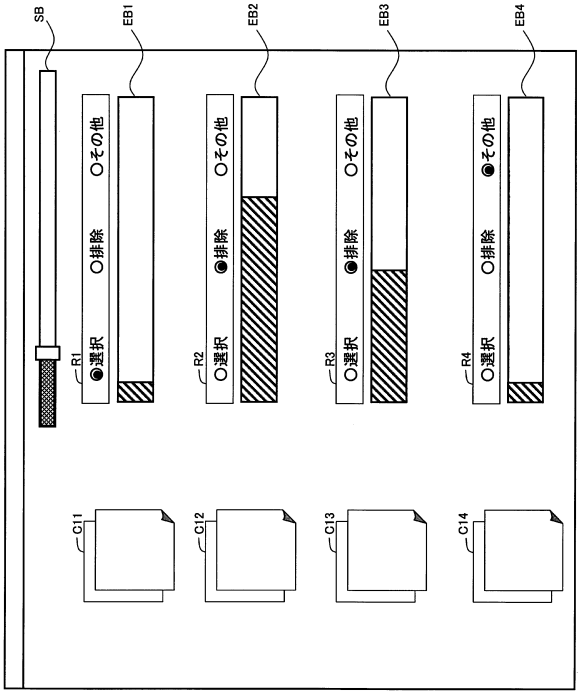
【図１７】



【図１８】



【図１９】


SB
EB1
EB2
EB3
EB4
R1
●選択
○排除
○その他
R2
○選択
●排除
○その他
R3
○選択
●排除
○その他
R4
○選択
○排除
●その他
C11
C12
C13
C14

フロントページの続き

(56)参考文献 特開２００４－３４１７５３（ＪＰ，Ａ）
特開２００７－１７２６１６（ＪＰ，Ａ）
特開２００７－３１０７３４（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０