

(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
2014 年 4 月 3 日(03.04.2014)



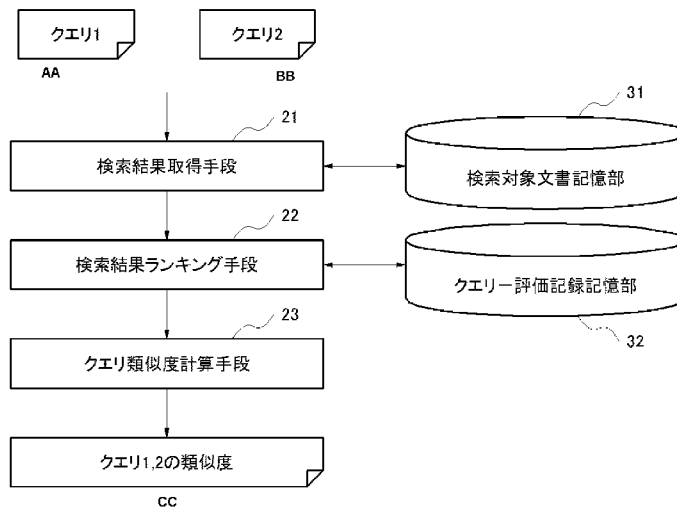
(10) 国際公開番号
WO 2014/050002 A1

- (51) 国際特許分類:
G06F 17/30 (2006.01)
- (21) 国際出願番号: PCT/JP2013/005406
- (22) 国際出願日: 2013 年 9 月 12 日(12.09.2013)
- (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 2012-217118 2012 年 9 月 28 日(28.09.2012) JP
- (71) 出願人: 日本電気株式会社(NEC CORPORATION)
[JP/JP]; 〒1088001 東京都港区芝五丁目 7 番 1 号
Tokyo (JP).
- (72) 発明者: 村岡 優輔(MURAOKA, Yusuke); 〒1088001 東京都港区芝五丁目 7 番 1 号日本電気株式会社内 Tokyo (JP). 楠村 幸貴(KUSUMURA, Yukitaka); 〒1088001 東京都港区芝五丁目 7 番 1 号日本電気株式会社内 Tokyo (JP). 水口 弘紀(MIZUGUCHI, Hironori); 〒1088001 東京都港区芝五丁目 7 番 1 号日本電気株式会社内 Tokyo (JP). 久寿居 大(KUSUI, Dai); 〒1088001 東京都港区芝五丁目 7 番 1 号日本電気株式会社内 Tokyo (JP).
- (74) 代理人: 下坂 直樹(SHIMOSAKA, Naoki); 〒1088001 東京都港区芝五丁目 7 番 1 号日本電気株式会社内 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, BN, 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, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK,

[続葉有]

(54) Title: QUERY DEGREE-OF-SIMILARITY EVALUATION SYSTEM, EVALUATION METHOD, AND PROGRAM

(54) 発明の名称: クエリ類似度評価システム、評価方法、及びプログラム



(57) Abstract: [Problem] As a result of similarities of queries being assessed on the basis of similarities of documents unrelated to an intent of a search, it is not possible to assess queries with similar intents of search. [Solution] The invention is provided with: search result ranking means which, on the basis of evaluation results of each of a plurality of documents searched by a first query, determines a first degree of importance for each of the plurality of documents, and on the basis of evaluation results of each of a plurality of documents searched by a second query, determines a second degree of importance for each of the plurality of documents; and query degree-of-similarity calculation means which calculates the degree of similarity of two search results, which have been imparted with degrees of importance, so as to increase in accordance with the similarities of documents of high importance. As a result, by calculating similarities of documents in cases in which the intents of searches are the same, the problem is resolved.

(57) 要約:

[続葉有]

- 21 Search result acquisition means
22 Search result ranking means
23 Query degree-of-similarity calculation means
31 Searched document storage unit
32 Query evaluation record storage unit
AA Query 1
BB Query 2
CC Degrees of similarity between queries 1 and 2

WO 2014/050002 A1



SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA,
UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, 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, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告 (条約第 21 条(3))

〔課題〕 検索意図と関係ない文書の類似を基にクエリの類似を判定してしまうために、検索意図が似ているクエリを判定できない。〔解決手段〕 第 1 のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第 1 の重要度を決定し、第 2 のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第 2 の重要度を決定する検索結果ランキング手段と、重要度の付けられた 2 つの検索結果の類似度を、重要度が高い文書が類似するほど大きく計算するクエリ類似度計算手段と、を備えることにより、検索意図が同じであった場合の文書の類似を計算することにより課題を解決する。

明 細 書

発明の名称：

クエリ類似度評価システム、評価方法、及びプログラム

技術分野

[0001] 本発明は、クエリ類似度評価システム、評価方法、プログラム及び記憶媒体に関する。

背景技術

[0002] 検索システムにおいては、ユーザが目的の文書を迅速に見つけ出せることが重要である。ここで、例えば、「my sqlでメモリサイズの設定方法を知りたい」、「my sqlでの検索速度を上げる方法を知りたい」といった、検索者が探している記述内容を検索意図と呼ぶこととする。

ユーザがクエリを入力した際、検索システムがユーザの検索意図に似ているクエリをユーザに推薦することや、検索意図が似ているクエリで目的の文書を上位とするような検索された結果の文書（以下、「検索結果文書」と記載する）に対するランキングは、検索意図を含む文書を探す場合に有効である。また、検索システムは、入力されたクエリの結果だけでなく、検索意図が似ているクエリの結果も表示することで、検索漏れを防ぐことができる。

また、ユーザが検索意図を含む文書を検索する際に、過去の検索時の文書へのアクセスログ、または評価ログを用いると、検索システムは検索結果文書に対するランキングを改善できるが、上記ログがすべてのクエリに対しては十分に存在しない場合がある。上記ログが十分でないクエリに対して、当該クエリのログだけでなく、検索意図が似ているクエリのログを用いることで、より多くのクエリに対して検索結果文書のランキングの改善が可能となる。

こうした応用のために、検索意図の似ているクエリを判定することが必要となる。複数のクエリに対し、検索意図が似ているかを判定するための手法として、それぞれのクエリの検索結果文書を利用する手法が知られている。検

索結果文書を利用して、同様の検索意図を表すクエリを判定するシステムの一例が、非特許文献 1 に記載されている。

図 1 1 に示すように、非特許文献 1 に記載のクエリ類似度判定システムは、類似度を評価したいクエリ（クエリ 1、クエリ 2）それぞれの検索結果を取得する検索結果取得手段と、その検索結果の類似度を計算する検索結果類似度計算手段と、を有する。このような構成を有する従来のクエリ類似度判定システムは、次のように動作する。

まず、検索結果取得手段は、入力された 2 つのクエリそれぞれの検索結果文書を検索対象文書記憶部から取得する。次に、検索結果取得手段が取得した 2 つの検索結果文書の集合を入力とし、検索結果類似度計算手段は、検索結果文書の一致または文書に含まれる単語の一致に基づいて、一致する個数が多いほど大きく類似度を計算し、出力する。

先行技術文献

非特許文献

- [0003] 非特許文献 1：“Finding similar queries to satisfy searches based on query traces”，Zaiane, O. and Strilets, A., Advances in Object-Oriented Information Systems, (2002)

発明の概要

発明が解決しようとする課題

- [0004] しかし、上述の非特許文献 1 に記載されたクエリ類似度判定システムは、クエリから取得される検索結果の文書の類似度を算出するので、次のような問題点がある。その問題点は、閲覧されていない文書と、検索意図に沿っていない文書との一致によってクエリを類似していると判定してしまうことである。その結果、検索意図の似ていないクエリが不当に似ていると判定されてしまうという問題があった。言い換えれば、非特許文献 1 に記載されたクエ

り類似度判定システムは、クエリの類似度を判定する精度が甘く、改善の余地がある。

そこで、本発明の目的の一例は、入力された複数のクエリの検索意図が似ているかを高い精度で判定するクエリ類似度評価システム、評価方法、及びプログラムを提供することにある。

課題を解決するための手段

[0005] 上記目的を達成するため、本発明の一形態にかかるクエリ類似度評価システムは、第1のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第1の重要度を決定し、第2のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第2の重要度を決定する検索結果ランキング手段と、前記文書集合の各文書の第1及び第2の重要度に基づき、前記複数のクエリの類似度を計算するクエリ類似度計算手段と、を備える。

また、上記目的を達成するため、本発明の一形態にかかるクエリ類似度評価方法は、第1のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第1の重要度を決定し、第2のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第2の重要度を決定する検索結果ランキングステップと、前記文書集合の各文書の第1及び第2の重要度に基づき、前記複数のクエリの類似度を計算するクエリ類似度計算ステップと、を備える。

更に、上記目的を達成するため、本発明の一形態にかかるプログラムは、コンピュータによって、第1のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第1の重要度を決定し、第2のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第2の重要度を決定し、前記文書集合の各文書の第1及び第2の重要度に基づき、前記複数のクエリの類似度を計算するクエリ類似度計算ステップとして機能させる。

発明の効果

[0006] 以上のように、本発明におけるクエリ評価システム、クエリ評価方法、及びプログラムによれば、検索意図が似ているクエリを高い精度で特定することができる。

図面の簡単な説明

[0007] [図1]図1は、本発明を実施するための最良の形態の構成を示すブロック図である。

[図2]図2は、本発明を実施するための最良の動作を示す流れ図である。

[図3]図3は、本発明を実施するための最良の形態の構成を実現するコンピュータの一例を示すブロック図である。

[図4]図4は、検索対象文書記憶部31のデータの具体例を示す図である。

[図5]図5は、クエリー評価記録記憶部32のデータの具体例を示す図である。

[図6]図6は、検索結果取得部21による出力の具体例を示す図である。

[図7]図7は、検索結果取得部21による出力の具体例を示す図である。

[図8]図8は、検索結果ランキング部22による出力の具体例を示す図である。

[図9]図9は、検索結果ランキング部22による出力の具体例を示す図である。

[図10]図10は、クエリー評価記録記憶部32が記憶するデータの例を示す図である。

[図11]図11は、従来技術のブロック図である。

発明を実施するための形態

[0008] 発明を実施するための最良の形態について図面を参照して詳細に説明する。本願で使用される用語「評価」は、検索エンジンの使用者が取った行動のうち、文書を求めていたか、求めていなかったかの手掛かりとなる行動を表す。評価とは、例えば、(1) 検索中に文書が役に立ったかを使用者にアンケートした結果に基づく検索システムに登録された文書への評価、または(2) 検索時の文書の閲覧である。アンケートや評価で「役に立つ」と回答され

るという行動、および文書が使用者に閲覧されるという行動は、その文書を求めていたことを示す手掛かりであり、それぞれ評価が高いとする。逆に「役に立たなかった」と回答されるという行動、および画面に文書リンクを表示したにもかかわらず文書が使用者に閲覧されないという行動は、その文書を求めていなかったことを示す手掛かりであり、それぞれ評価が低いとする。

図 1 を用いて、本発明を実施するための最良の形態におけるクエリ類似度評価システムの構成について説明する。図 1 は、本発明を実施するための最良の形態の構成を示すブロック図である。

図 1 を参照すると、本発明を実施するための最良の形態におけるクエリ類似度評価システムは、検索結果取得部 2 1、検索結果ランキング部 2 2、クエリ類似度計算部 2 3、検索対象文書記憶部 3 1、クエリー評価記録記憶部 3 2 から構成されている。

検索対象文書記憶部 3 1 は、検索システムで検索対象となる文書を記憶している。検索対象文書記憶部 3 1 は、例えば、文書テキストそのもの、文書に対して付けられたメタデータ（文書 ID、文書の更新日時、筆者、特定のタグが付いたテキスト、文書を参照する文書の ID、文書に付けられたスコア等）、文書テキスト中の単語に対して付けられた転置インデックス等を記憶する。

クエリー評価記録記憶部 3 2 は、クエリとそのクエリに対する評価の記録（以下、「評価記録」と記載する）を互いに関連付けた情報を記憶する。クエリー評価記録記憶部 3 2 は、例えば、図 10 に示すように、過去に検索エンジンに使用者から入力されたクエリ（以下、「クエリ」と記載する）と、当該クエリによって検索された文書、および当該文書への評価とを対応付けした情報を記録する。ここで、クエリー評価記録記憶部 3 2 が記憶するデータは、例えば、検索システムで、クエリと閲覧された文書を記述したログを出力させることで、作成されることにより、あらかじめ記憶されておいてよい。

次に、本発明を実施するための最良の形態におけるクエリ類似度評価システムの動作について説明する。

検索結果取得部21は、検索対象文書記憶部31を参照し、2つのクエリ（第1のクエリ、第2のクエリ）に対する検索結果をそれぞれ特定する。例えば、検索されたクエリが文書中に含まれる文書を特定する。検索結果取得部21は、特定された2つの検索結果文書の集合（以下、「検索結果文書集合」または「検索結果文書集合1、検索結果集合2」と記載する）を、検索結果ランキング部22に出力する。検索結果ランキング部22は、検索結果取得部21が出力した2つのクエリとそれぞれに対応する2つの検索結果文書集合の組に対し、クエリー評価記録記憶部32を参照して、クエリに対する評価記録が含まれるか否かを調べる。もし、いずれの評価記録もクエリー評価記録記憶部32に含まれない場合、検索結果ランキング部22は、検索結果文書とクエリのみから計算されるランキングスコア（例えば、クエリ単語が含まれる回数、Page Rank等の文書スコア）に基づいて2つの検索結果文書集合の各文書に対し重要度を算出し、クエリ類似度計算部23に算出した重要度を出力する。

いずれかの評価記録が、クエリー評価記録記憶部32に含まれる場合、検索結果ランキング部22はクエリー評価記録記憶部32を参照する。検索結果ランキング部22は、参照した結果を基に、2つの検索結果文書集合の各文書に対する重要度を算出する。例えば、検索結果ランキング部22は、クエリに対応する文書の評価が高くなるほど重要度がより高く、また文書の評価が低くなるほど重要度がより低くなるよう算出する。検索結果ランキング部22は、その算出した結果をクエリ類似度計算部23に出力する。

上記の重要度を算出する方法（以下、「重要度算出方法」と記載する）は、例えば、高評価された文書で出現頻度が高く、低評価された文書で出現頻度が低い語（特徴語）を特定し、並べ替えたい文書に対し、上で特定された単語の頻度が大きいほど高い重要度を算出する、という方法であってもよい。また、重要度算出方法は、例えば、クエリと文書の組に対して、文書中のク

エリキーワードの出現頻度、文書に付与されたメタデータ（文書の更新日時、文書の長さ等）の値を特徴ベクトルとして、入力文書の特徴ベクトルと、高評価された文書の特徴ベクトルとのユークリッド距離を計算し、距離が小さいほど高い重要度を算出する、という方法であってもよい。

もし、両方の評価記録がクエリー評価記録記憶部 32 に含まれるならば、検索結果ランキング部 22 はそれぞれのクエリーに対して、クエリー評価記録記憶部 32 を参照する。検索結果ランキング部 22 は、参照した結果を基に、クエリーに対応する評価された文書を上位に、評価されていない文書を下位にするように 2 つの検索結果文書集合を並べ替える。検索結果ランキング部 22 は、それぞれの並べ替えによる、2 組の 2 つの検索結果文書集合の組をクエリー類似度計算部 23 に出力する。

クエリー類似度計算部 23 は、検索結果ランキング部 22 から出力された、1 組または 2 組の並べ替えられた検索結果文書集合の組に対し、それぞれの文書で高い重要度を算出された文書同士の類似を重視するように、検索結果文書集合間の類似度を計算する。

[数 1]

$$\sum_{d_1 \in S_1} \sum_{d_2 \in S_2} w_1(d_1)w_2(d_2)\text{sim}(d_1, d_2)$$

[0009] 数式 1 は、検索結果集合 1 を S_1 、検索結果集合 2 を S_2 、文書 d_1 の検索結果集合 1 での重要度を $w_1(d_1)$ 、文書 d_2 の検索結果集合 2 での重要度を $w_2(d_2)$ 、文書 d_1 と文書 d_2 の類似度を $\text{sim}(d_1, d_2)$ で表したものである。

数式 1 は、検索結果集合 1、検索結果集合 2 に含まれる文書の組み合わせそれぞれについて、検索結果集合 1 での重要度と、検索結果集合 2 での重要度との積が大きいほど大きい重みをつけて、類似度を足し合わせたものである。2 組入力された場合には、数式 1 は、それぞれの組で計算された値の平均を用いる。

特に、 $\text{sim}(d_1, d_2)$ を文書の一致で判断する場合、類似度は以下の式で計算される。

[数 2]

$$\sum_{d \in S_1 \cap S_2} w_1(d)w_2(d)$$

[0010] 数式 2 では、クエリ類似度計算部 23 は、文書の ID の一致により文書類似度を判断したが、文書内容の類似で判断してもよい。例えば、クエリ類似度計算部 23 は、文書本文の単語ベクトルのコサイン類似度や、メタデータの差分のノルムを用いてもよい。

[クエリ類似度評価システムの動作]

次に、本発明を実施するための最良の形態におけるクエリ類似度評価システムの動作について、図 1 を適宜参酌しつつ、図 2 を用いて説明する。なお、本発明の実施形態では、クエリ類似度評価システムを動作させることによってクエリ類似度評価方法が実施されるため、本発明の実施形態におけるクエリ類似度評価方法の説明は、以下のクエリ類似度評価システムの動作説明に代える。

次に、図 2 を参照して本発明を実施するための最良の形態におけるクエリ類似度評価システムの全体の動作について詳細に説明する。図 2 は、本発明の実施形態に係るクエリ類似度評価システムの処理を表すフローチャートである。

まず、検索結果取得部 21 は、2 つのクエリに対する検索結果文書集合を、検索対象文書記憶部 31 から参照して特定し、2 つのクエリとそれぞれのクエリに対する検索結果文書集合を検索結果ランキング部 22 に出力する（ステップ A1）。

次に、ステップ A1 での 2 つのクエリとそれぞれの検索結果について、検索結果ランキング部 22 は、クエリー評価記録記憶部 32 に、評価記録が存在するかどうかを判定する。クエリー評価記録記憶部 32 に、評価記録が存在

するならば、処理はステップA 4に進む。クエリー評価記録記憶部3 2に、評価記録が存在しないならば、処理はステップA 3に進む（ステップA 2）。

次に、検索結果ランキング部2 2は、ステップA 1での2つのクエリとそれぞれのクエリに対する検索結果文書の集合に対し、重要度を算出する（ステップA 3）。例えば、ステップA 1での2つのクエリとそれぞれのクエリに対する検索結果ランキング部2 2は、検索結果文書の集合に対して、検索結果の並べ替えを行う等である。

次に、検索結果ランキング部2 2は、にステップA 1での2つのクエリとそれぞれのクエリに対する検索結果文書の集合に対し、クエリー評価記録記憶部3 2に存在する評価記録を特定する（ステップA 4）。

次に、検索結果ランキング部2 2は、ステップA 4で特定された、評価記録、クエリ、クエリに対する検索結果文書の集合に対し、クエリに対する検索結果文書の集合2つの各文書に対し、評価記録で評価された文書ほど高くなるように重要度を算出する。2つの各文書の評価記録が特定された場合には、検索結果ランキング部2 2は、2種類の重要度を算出する。検索結果ランキング部2 2は、それぞれの評価記録に基づき重要度を算出された、2つの検索結果文書集合の組、1組または2組を、クエリ類似度計算部2 3に出力する（ステップA 5）。

次に、クエリ類似度計算部2 3は、ステップA 3ないし、ステップA 5での、1組または2組の2つの検索結果文書集合に対し、高い重要度の文書同士の類似を重視するよう、類似度を計算する。クエリ類似度計算部2 3は、2組の2つの検索結果文書集合が出力された場合には、部それぞれの組の類似度の平均を出力する（ステップA 6）。

[プログラム]

本発明を実施するための最良の形態におけるクエリ類似度評価システムのプログラムは、コンピュータに、図2に示すステップA 1～A 6を実行させるプログラムであればよい。このプログラムをコンピュータに導入し、実行す

ることによって、本発明を実施するための最良の形態におけるクエリ類似度評価システムと、クエリ類似度評価方法と、を実現することができる。

[コンピュータ]

図3を用いて、本発明を実施するための最良の形態におけるクエリ類似度評価システムを実現するコンピュータについて説明する。図3は、本発明を実施するための最良の形態の構成を実現するコンピュータの一例を示すブロック図である。

図3は、本発明を実施するための最良の形態におけるクエリ類似度評価システムのハードウェア構成図である。図3に示すように、クエリ類似度評価システムは、例えばCPU (Central Processing Unit) 1、RAM (Random Access Memory) 2、記憶装置3、通信インターフェース4、入力装置5、出力装置6等を含む。

検索結果取得部21、検索結果ランキング部22等は、例えば、CPU1が、プログラムをRAM2に読み出し、実行することによって実現される。検索結果取得部21、検索結果ランキング部22等が情報の送受信を行う動作は、例えばOS (Operating System) が提供する機能を使ってアプリケーションプログラムが通信インターフェース4を制御することによって実現される。記憶装置3は、例えば、ハードディスクや、フラッシュメモリである。入力装置5は、例えばキーボードやマウス等である。出力装置6は、例えばディスプレイ等である。

具体的な例を用いて本発明の実施形態の動作を説明する。

図4に示すように、検索対象文書記憶部31は、検索対象文書データを記憶している。ここで、図4に示す検索対象文書データは、例えば、6つの各文書に対してのデータ集合を示す。例えば、検索対象文書データは、文書のID、文書のタイトル、文書の更新日時が現在から何日前なのか、文書の被リンク数、文書の長さ（文字数）等の、データ集合である。

図5に示すように、クエリー評価記録記憶部32は、クエリと当該クエリに対する評価記録（クエリー評価記録）を記憶している。

ここで、図5に示すクエリー評価記録は、例えば、クエリ「mysql メモリ 設定」を入力して検索している際に行われた評価1回につき、クエリ、評価された文書のID、評価内容（Goodなら探していた文書であることを表し、Badなら探していた文書と異なっていることを表す）等の、データ集合である。

以下、「mysql メモリ 設定」と「my.cnf cache size」の2つのクエリが入力された場合（case 1）と、「mysql メモリ 設定」と、「mysql インデックス作成」の2つのクエリが入力された場合（case 2）との、クエリ類似度を計算する際の具体的な処理を記述する。

case 1においては、どちらのクエリもmysqlのメモリに関する設定方法の検索を意図しており、検索意図が似ている。case 2においては、「mysql メモリ 設定」はメモリの設定方法の検索を意図しており、「mysql インデックス作成」はフィールドのインデックスの作成方法を意図しているため、検索意図が異なる。ただし、case 2のクエリは、どちらも処理速度を上げるための方法であるため、同一の文書に記述があることもある。

まず、検索結果取得部21は、検索対象文書記憶部31を参照して、それぞれのクエリにより検索される文書を特定する。例えば、図6に示すように、例えば、case 1の場合では、検索結果取得部21は、クエリが本文中に含まれる文書を特定し、クエリ「mysql メモリ 設定」に対しては文書ID 0、1、2、3、5の文書を、クエリ「my.cnf cache size」に対しては文書ID 0、2、3の文書を検索結果として特定する。

図7に示すように、例えば、case 2の場合では、検索結果取得部21は、クエリ「mysql メモリ 設定」に対しては文書ID 0、1、2、3、5の文書を、クエリ「mysql インデックス作成」に対しては文書ID 0、1、4、5の文書を検索結果として特定する。検索結果取得部2

1 は、それぞれのクエリと検索結果文書 ID の集合を検索結果ランキング部 22 に出力する。

次に、検索結果ランキング部 22 は、クエリー評価記録記憶部 32 を参照し、case 1、case 2 とともに、検索結果取得部 21 によって出力された 2 つのクエリのうち、「mysql メモリ 設定」の評価記録のみが存在することを特定する。

ここでは、具体的な例として、クエリが完全に一致する評価記録を用いたが、以下のクエリ類似度を計算する際の具体的な処理では、クエリをキーワードに分解し（例えば、「mysql メモリ 設定」を「mysql」、「メモリ」、「設定」に分解）、キーワードが含まれる評価記録を用いるようにしても良い。

次に、検索結果ランキング部 22 は、評価記録が存在したクエリ「mysql メモリ 重い」の評価記録（評価記録 ID 0、1）に基づき、評価記録で高評価の（Good と評価された）文書 ID 3 の文書の重要度を高く、評価記録で低評価の（Bad と評価された）文書 ID 5 の文書に重要度を低く出力された 2 つの検索結果のランキングを行う。

例えば、検索結果ランキング部 22 は、高評価の文書 ID 3 の文書で頻度が高く、低評価の文書 ID 5 の文書で頻度が低い語「buffer」、「pool」、「設定ファイル」を特徴語として特定し、「buffer」、「pool」、「設定ファイル」の本文での出現頻度の和を重要度として算出する。そして、図 8 に示すように、例えば、case 1 では、検索結果ランキング部 22 は、クエリ「mysql メモリ 設定」の検索結果文書集合と、クエリ「my.cnf cache size」の検索結果文書集合に対する、順位、文書 ID、スコア等のランキング結果を得る。図 9 に示すように、例えば、case 2 では、検索結果ランキング部 22 は、クエリ「mysql メモリ 設定」の検索結果文書集合と、クエリ「mysql インデックス作成」の検索結果文書集合に対する、順位、文書 ID、スコア等のランキング結果を得る。

ここで、検索結果ランキング部 22 の評価方法としては、逆に低評価された文書のみで頻度が高い語を特定し、その語の頻度が小さいほど大きい重要度を算出してもよい。また、検索結果ランキング部 22 の評価方法としては、メタデータを用い、高評価された文書のスコアを +1、低評価された文書のスコアを -1 として、メタデータ（例だと、更新日時、被リンク数、長さ）からスコアを出力する関数を学習し、関数の出力する値を重要度としてもよい。

ここでは、検索結果 S の中での文書 d の重要度は、検索結果 S 内での順位 $order(d)$ を利用して以下のように計算される。また、検索結果 S_1 の中での文書 d_1 の重要度は順位 $order_1(d)$ を、検索結果 S_2 の中での文書 d_2 の重要度は順位 $order_2(d)$ を利用して計算される。

[数 3]

$$w(d) = \frac{e^{-order(d)}}{\sum_{d \in S} e^{-order(d)}}$$

[0011] また、文書の重要度に基づいたクエリ類似度は、以下のように計算される。

[数 4]

$$\frac{\sum_{d \in S_1 \cap S_2} w_1(d) w_2(d)}{\min \left(\sum_{d \in S_1} w_1(d)^2, \sum_{d \in S_2} w_2(d)^2 \right)}$$

[0012] [数 5]

$$\frac{\sum_{d \in S_1 \cap S_2} e^{-(order_1(d) + order_2(d))}}{\sum_{i=1}^{\min(|S_1|, |S_2|)} e^{-2i}}$$

[0013] 数式 5 は、数式 4 に数式 3 を代入すると得られる式である。

次に、クエリ類似度計算部 23 は、検索結果ランキング部 22 から入力された図 8 または図 9 の重要度のついた検索結果文書 2 つを入力として、以下の

ように類似度を計算する。

[数 6]

$$\frac{e^{-(1+1)} + e^{-(2+2)} + e^{-(3+3)}}{\sum_{i=1}^3 e^{-2i}} = \frac{0.1561}{0.1561} = 1.0$$

[0014] c a s e 1 の場合は、クエリ類似度計算部 2 3 は、数式 6 のように計算結果 1. 0 を出力する。

[数 7]

$$\frac{e^{-(2+1)} + e^{-(4+2)} + e^{-(5+4)}}{\sum_{i=1}^4 e^{-2i}} = \frac{0.0524}{0.1565} = 0.335$$

[0015] c a s e 2 の場合は、クエリ類似度計算部 2 3 は、数式 7 のように計算結果 0. 3 3 5 を出力する。

従来手法の場合では、検索結果の共通の文書の割合では、c a s e 1 でそれぞれの検索結果の 3 / 5、3 / 3 であり、平均すると 0. 8、c a s e 2 ではそれぞれの検索結果の 3 / 5、3 / 4 であり平均すると 0. 6 7 5 となり検索意図が異なるクエリに対しても、類似度を大きく計算してしまっていた。

一方、本発明の実施形態では、検索意図が同じ c a s e 1 では 1. 0、検索意図が異なる c a s e 2 では 0. 3 3 5 と、検索意図が異なるクエリに対してより小さい類似度を計算することができる。

以上、実施形態を用いて本願発明を説明したが、本願発明は、上記実施形態に限定されるものではない。本願発明の構成や詳細には、本願発明のスコープ内で当業者が理解しうる様々な変更をすることができる。

上記の実施形態の一部又は全部は、以下の付記のようにも記載されうるが、以下には限られない。この出願は、2 0 1 2 年 9 月 2 8 日に出願された日本出願特願 2 0 1 2 - 2 1 7 1 1 8 を基礎とする優先権を主張し、その開示の全てをここに取り込む。

産業上の利用可能性

[0016] 本発明は、クエリ推薦システム、文書ランキングシステムといった用途に適用できる。

符号の説明

- [0017] 1 CPU
2 RAM
3 記憶装置
4 通信インターフェース
5 入力装置
6 出力装置
21 検索結果取得部
22 検索結果ランキング部
23 クエリ類似度計算部
31 検索対象文書記憶部
32 クエリー評価記録記憶部

請求の範囲

- [請求項1] 第1のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第1の重要度を決定し、第2のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第2の重要度を決定する検索結果ランキング手段と、
前記文書集合の各文書の第1及び第2の重要度に基づき、前記複数のクエリの類似度を計算するクエリ類似度計算手段と、
を備えることを特徴とするクエリ類似度評価システム。
- [請求項2] 前記検索結果ランキング手段は、
少なくとも前記第1のクエリと前記第2のクエリを含む複数のクエリの類似度を評価する際に、前記各クエリによって得られる結果の文書集合のそれぞれに対して、前記クエリの過去の文書集合の評価結果と今回の文書集合を比較することによって、当該文書集合に含まれる各文書の重要度を算出することを特徴とする請求項1に記載のクエリ類似度評価システム。
- [請求項3] 前記クエリ類似度計算手段は、前記検索結果ランキング手段が、評価が高い文書と評価が低い文書それぞれの特徴語を特定し、評価が高い文書の特徴語の出現頻度が高い文書に対しては重要度を高く、評価が低い文書の特徴語の出現頻度が高い文書に対しては重要度を低く算出することを特徴とする請求項1または2に記載のクエリ類似度評価システム。
- [請求項4] 前記検索結果ランキング手段は、評価が高い文書と評価が低い文書それぞれに付与されたメタデータを参照し、評価が高い文書とメタデータの値が近い文書ほど重要度を高く、評価が低い文書のメタデータと近い文書ほど重要度を低く算出することを特徴とする請求項1乃至3に記載のクエリ類似度評価システム。
- [請求項5] 前記クエリ類似度計算手段は、検索結果集合1を S_1 、検索結果集合

2を S_2 、文書 d の検索結果集合1での重要度（検索結果集合1内の文書での総和が1となるように正規化されていることとする）を $w_1(d)$ 、文書 d の検索結果集合2での重要度を $w_2(d)$ 、文書 d_1 と文書 d_2 の類似度を $\text{sim}(d_1, d_2)$ として、アルゴリズム

[数1]

$$\sum_{d_1 \in S_1} \sum_{d_2 \in S_2} w_1(d_1) w_2(d_2) \text{sim}(d_1, d_2)$$

を用いてクエリ類似度を計算することを特徴とする請求項1乃至5に記載のクエリ類似度評価システム。

[請求項6] 第1のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの重要度を決定し、第2のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの重要度を決定する検索結果ランキングステップと、前記文書集合の各文書の第1及び第2の重要度に基づき、前記複数のクエリの類似度を計算するクエリ類似度計算ステップと、を備えることを特徴とするクエリ類似度評価方法。

[請求項7] 前記検索結果ランキングステップは、少なくとも前記第1のクエリと前記第2のクエリを含む複数のクエリの類似度を評価する際に前記各クエリによって得られる結果の文書集合のそれぞれに対して、前記クエリの過去の文書集合の評価結果と今回の文書集合を比較することによって、当該文書集合に含まれる各文書の重要度を算出することを特徴とする請求項6に記載のクエリ類似度評価方法。

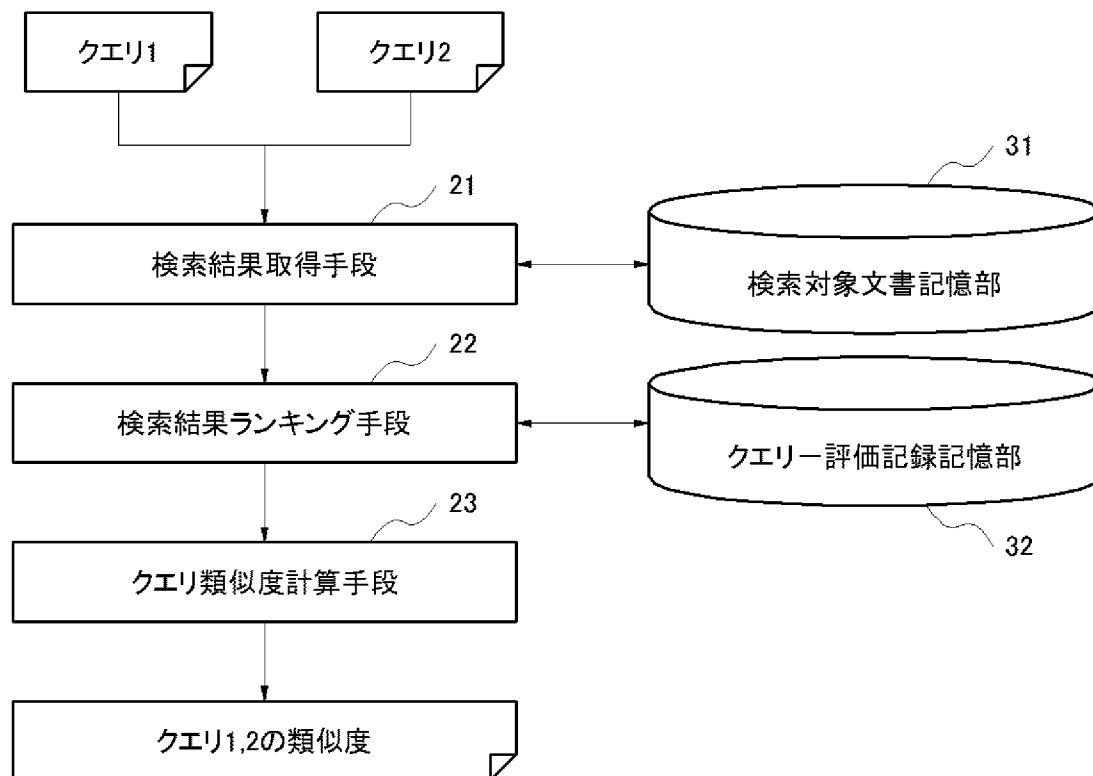
[請求項8] 前記クエリ類似度計算ステップは、前記検索結果ランキングステップが、評価が高い文書と評価が低い文書それぞれの特徴語を特定し、評価が高い文書の特徴語の出現頻度が高い文書に対しては重要度を高く、評価が低い文書の特徴語の出現頻度が高い文書に対しては重要度を低く算出することを特徴とする請求項6または7に記載のクエリ類似

度評価方法。

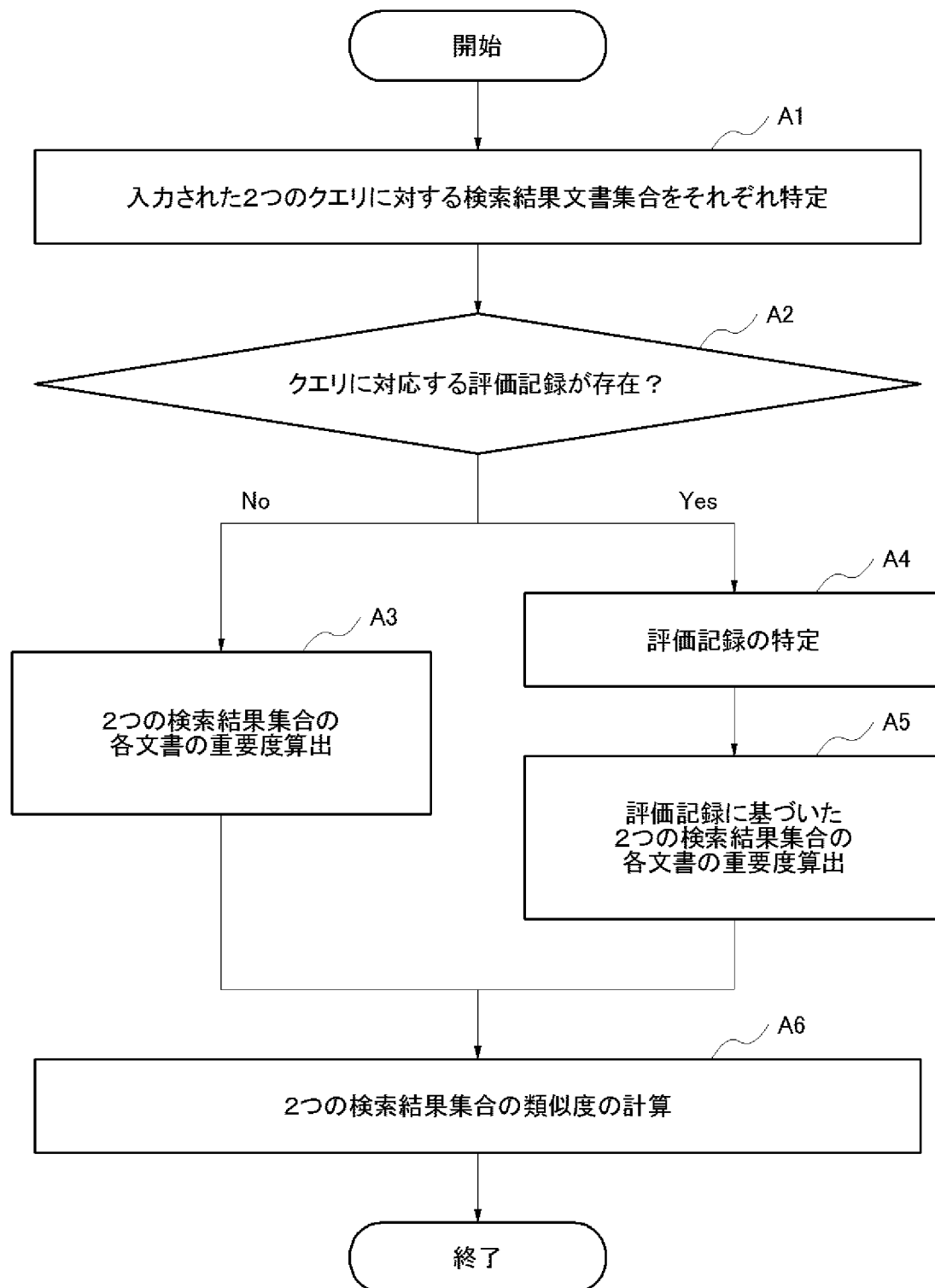
[請求項9] 前記検索結果ランキングステップは、評価が高い文書と評価が低い文書それぞれに付与されたメタデータを参照し、評価が高い文書とメタデータの値が近い文書ほど重要度を高く、評価が低い文書のメタデータと近い文書ほど重要度を低く算出することを特徴とする請求項6乃至8に記載のクエリ類似度評価方法。

[請求項10] コンピュータによって、第1のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第1の重要度を決定し、第2のクエリが検索された複数の文書のそれぞれの評価結果に基づいて、前記複数の文書のそれぞれの第2の重要度を決定し、前記文書集合の各文書の第1及び第2の重要度に基づき、前記複数のクエリの類似度を計算するクエリ類似度計算ステップとして機能させるためのプログラム。

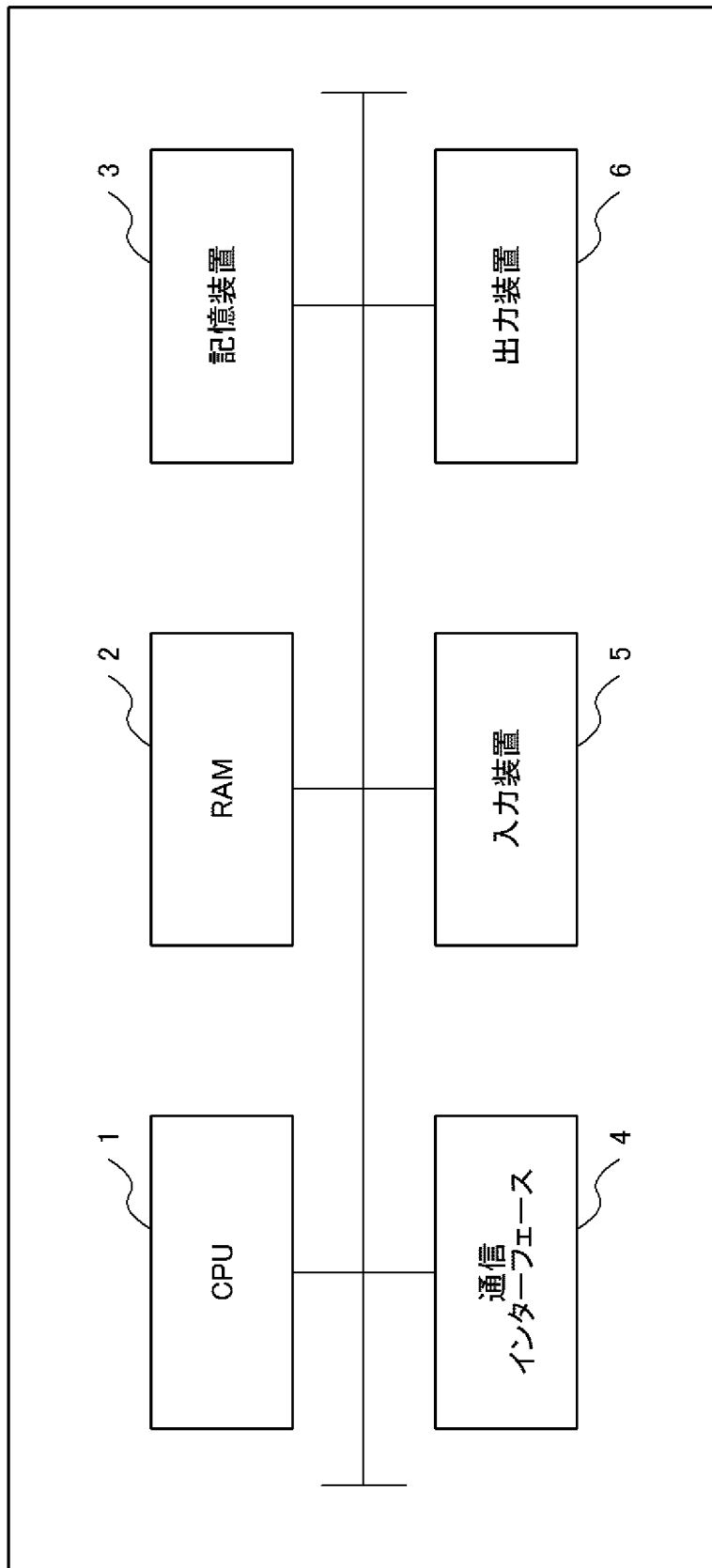
[図1]



[図2]



[図3]



[図4]

文書ID	文書タイトル	更新日時	被リンク数	文書長さ
0	mysqlの処理速度向上施策	1日	10	1k
1	mysqlで検索時間を短縮したい	5日	1	9k
2	mysql環境設定ログ	3日	2	10k
3	mysqlのキャッシュサイズの設定	40日	30	3k
4	mysqlインデックス作成Tips	30日	10	5k
5	はじめてのmysql	10日	5	4k

[図5]

記録ID	クエリ	文書ID	評価
0	mysql メモリ 設定	3	Good
1	mysql メモリ 設定	5	Bad

[図6]

検索クエリ	検索結果文書ID
「mysql メモリ 設定」	0,1,2,3,5
「my.cnf cache size」	0,2,3

[図7]

検索クエリ	検索結果文書ID
「mysql メモリ 設定」	0,1,2,3,5
「mysql インデックス 作成」	0,1,4,5

[図8]

クエリ「mysql メモリ 設定」の検索結果文書集合

順位	文書ID	スコア
1	3	10
2	0	8
3	2	7
4	1	4
5	5	0

クエリ「my.cnf cache size」の検索結果文書集合

順位	文書ID	スコア
1	3	10
2	0	8
3	2	7

[図9]

クエリ「mysql メモリ 設定」の検索結果文書集合

順位	文書ID	スコア
1	3	10
2	0	8
3	2	7
4	1	4
5	5	0

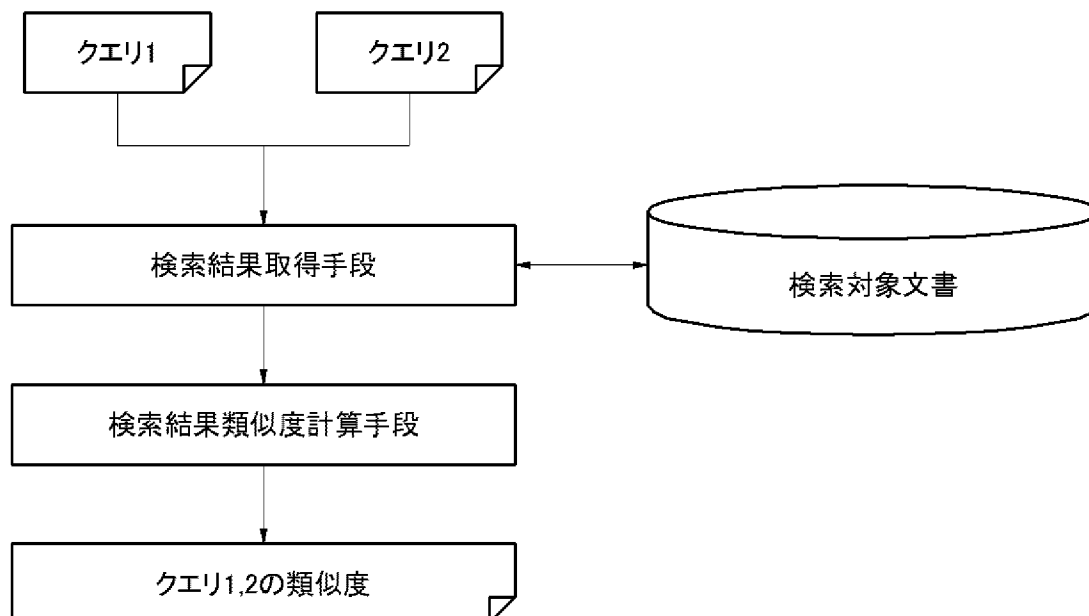
クエリ「mysql インデックス 作成」の検索結果文書集合

順位	文書ID	スコア
1	0	8
2	1	4
3	4	1
4	5	0

[図10]

検索クエリ	クエリによって検索された文書	評価
A	文書A	高
	文書B	低
	文書C	低
	文書D	高
	文書E	低
B	文書A	低
	文書B	高
	文書C	高
	文書D	高
	文書E	高

[図11]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2013/005406

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	1996-2013
Kokai Jitsuyo Shinan Koho	1971-2013	Toroku Jitsuyo Shinan Koho	1994-2013

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 A	JP 2011-209999 A (Yahoo Japan Corp.), 20 October 2011 (20.10.2011), paragraphs [0034] to [0056], [0091] to [0096]; fig. 1 to 6, 11 (Family: none)	1, 6, 10 2, 7 3-5, 8, 9
Y	JP 2001-202390 A (Xerox Corp.), 27 July 2001 (27.07.2001), paragraph [0019]; fig. 2 & US 6732088 B1	2, 7
A	JP 2010-067011 A (Fujitsu Ltd.), 25 March 2010 (25.03.2010), entire text; all drawings (Family: none)	1-10

☒ 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
18 October, 2013 (18.10.13)

Date of mailing of the international search report
29 October, 2013 (29.10.13)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2013/005406

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2010-072909 A (Nippon Telegraph and Telephone Corp.), 02 April 2010 (02.04.2010), entire text; all drawings (Family: none)	1-10
A	JP 2010-122932 A (Nippon Telegraph and Telephone Corp.), 03 June 2010 (03.06.2010), entire text; all drawings (Family: none)	1-10

A. 発明の属する分野の分類（国際特許分類（I P C））

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（I P C））

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	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 1 3 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 3 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 3 年

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

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
X	JP 2011-209999 A（ヤフー株式会社）	1, 6, 10
Y	2011.10.20, 段落【0034】 - 【0056】 , 【0091】 - 【0096】 ,	2, 7
A	【図 1】 - 【図 6】 , 【図 11】（ファミリーなし）	3-5, 8, 9
Y	JP 2001-202390 A（ゼロックス コーポレーション） 2001.07.27, 段落【0019】 , 【図 2】 & US 6732088 B1	2, 7

☒ 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

1 8 . 1 0 . 2 0 1 3

国際調査報告の発送日

2 9 . 1 0 . 2 0 1 3

国際調査機関の名称及びあて先

日本国特許庁（I S A / J P）

郵便番号 1 0 0 - 8 9 1 5

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）

齊藤 貴孝

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 9 9

5 M

4 7 7 4

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
A	JP 2010-067011 A (富士通株式会社) 2010.03.25, 全文, 全図 (ファミリーなし)	1-10
A	JP 2010-072909 A (日本電信電話株式会社) 2010.04.02, 全文, 全図 (ファミリーなし)	1-10
A	JP 2010-122932 A (日本電信電話株式会社) 2010.06.03, 全文, 全図 (ファミリーなし)	1-10