(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4912384号
(P4912384)
(45)発行日　平成24年4月11日(2012.4.11)　(24)登録日　平成24年1月27日(2012.1.27)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 320D
G06F 17/30 380E
G06F 17/30 340B

請求項の数 3　(全 13 頁)

(21)出願番号　特願2008-297847(P2008-297847)
(22)出願日　平成20年11月21日(2008.11.21)
(65)公開番号　特開2010-123036(P2010-123036A)
(43)公開日　平成22年6月3日(2010.6.3)
審査請求日　平成22年2月4日(2010.2.4)

(73)特許権者　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(74)代理人　100086232
弁理士　小林　博通
(74)代理人　100104938
弁理士　鵜澤　英久
(74)代理人　100140361
弁理士　山口　幸二
(74)代理人　100096459
弁理士　橋本　剛
(72)発明者　村田　眞哉
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】文書検索装置、文書検索方法、および文書検索プログラム

(57)【特許請求の範囲】
【請求項１】
ユーザ端末から検索指示されたクエリを含む電子文書を検索するときに検索エンジンの検索ログを利用する文書検索装置であって、
前記検索ログに含まれたクエリに応じた検索結果のタイトルおよび概要文から拡張語を生成し、該拡張語を前記クエリの拡張情報として保存するクエリ情報保存手段と、
前記検索ログからクリックされた検索結果の検索時のクエリを判別し、該クエリに関連する拡張語を前記クエリ情報保存手段から求め、検索結果がクリックされたときに投入されたクエリの拡張語群を検索結果の拡張情報として保存する検索結果情報保存手段と、
ユーザ端末から検索指示されたクエリについて、クエリの拡張語を前記クエリ情報保存手段から取得し、ユーザ端末に送る照合処理手段と、
前記検索指示されたクエリを前記拡張語で拡張した拡張クエリの検索で得られた検索結果に対して、前記検索結果情報保存手段に保存された拡張語群を付与して検索結果を拡張し、拡張された検索結果を前記拡張クエリとの類似度により並び替え、この並び替えた結果をリスト化した最終検索結果を前記ユーザ端末に送る検索結果処理手段と、
を備えることを特徴とする文書検索装置。
【請求項２】
ユーザ端末から検索指示されたクエリを含む電子文書を検索するときに検索エンジンの検索ログを利用する文書検索方法であって、
前記検索ログに含まれたクエリに応じた検索結果のタイトルおよび概要文から拡張語を

生成し、該拡張語をクエリ情報保存手段に前記クエリの拡張情報として保存するクエリ情報保存ステップと、

検索結果情報保存手段が、前記検索ログからクリックされた検索結果の検索時のクエリを判別し、該クエリに関連する拡張語を前記クエリ情報保存手段から求め、検索結果がクリックされたときに投入されたクエリの拡張語群を検索結果の拡張情報として保存する検索結果情報保存ステップと、

照合処理手段が、ユーザ端末から検索指示されたクエリについてクエリの拡張語を前記クエリ情報保存手段から取得し、ユーザ端末に送る照合処理ステップと、

検索結果処理手段が、前記検索指示されたクエリを前記拡張語で拡張した拡張クエリの検索で得られた検索結果に対して、前記検索結果情報保存手段に保存された拡張語群を付与して検索結果を拡張し、拡張された検索結果を前記拡張クエリとの類似度により並び替え、この並び替えた結果をリスト化した最終検索結果を前記ユーザ端末に送る検索結果処理ステップと、

を有することを特徴とする文書検索方法。

【請求項３】

請求項１記載の文書検索装置を構成する各手段としてコンピュータを機能させることを特徴とする文書検索プログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、電子文書群中からクエリに該当する電子文書を検索する技術に関する。

【背景技術】

【０００２】

インターネットなどで接続されたクライアント端末から検索語（クエリ）を受信して検索結果を返信する検索エンジンの運営サーバには、時々刻々、検索のログが保存される。このログには、投入されたクエリの情報や、検索結果に対するユーザのクリックの情報などが保存される。

【０００３】

このような検索エンジンのログを利用した文書検索システムが非特許文献１に提案されている。この文書検索システムは、検索エンジンのログのうち、特にクリックに関する情報（クリックログ）を利用して検索結果の精度向上を行うものである。

【０００４】

すなわち、この文書検索システムでは、クリックログを解析することにより、多くの検索結果中からクリックが集中しているサイト（アクセス集中サイト）を的確に判別している。このアクセス集中サイトのタイトルとスニペット（概要文）には有用な情報（キーワード）が含まれていると考えられ、この情報でクエリを拡張し、各クエリの情報要求に沿った高精度な検索を実現するものである。

【非特許文献１】“Ｉｍｐｒｏｖｉｎｇ Ｍｏｂｉｌｅ Ｗｅｂ－ＩＲ Ｕｓｉｎｇ Ａｃｃｅｓｓ Ｃｏｎｃｅｎｔｒａｔｉｏｎ Ｓｉｔｅｓ ｉｎ Ｓｅａｒｃｈ Ｒｅｓｕｌｔｓ.”Ｍａｓａｙａ Ｍｕｒａｔａ，ｅｔ ａｌ．Ｐｒｏｃ．ｏｆ ＷＩＳＥ２００８，ｐｐ ２２１－２３４，２００８.

【発明の開示】

【発明が解決しようとする課題】

【０００５】

非特許文献１の文書検索システムでは、クリックログを解析し、検索結果中においてアクセスが集中しているサイトを的確に判別することで、そこから有用な情報（キーワード）を抽出する。そしてこのキーワード群を基にクエリを拡張し、検索を行う。これは、通常１語か２語であるクエリの少ない単語数を、その情報要求を的確に表現する他のキーワード群で補い、ユーザの検索を補助するものである。

【０００６】

一方、各検索結果が満たすことのできる情報要求の表現については、それらの内容、すなわち文書の作成者が記した本文のみが利用され、ユーザがその検索結果を実際に見て、それが要求通りか否かをどのように判断したのかに関する情報は考慮されていない。このような文書を利用するユーザ側からの情報を基に各検索結果を拡張することで、検索エンジンに対してクエリを発行したユーザが持つ情報要求（クエリの情報要求）と各検索結果が満たすことができる情報要求レベルでのマッチングが可能となる。

【０００７】

そこで本発明は、このような問題に鑑み、クエリの情報要求と、検索システムに登録されている各文書が満たすことができる情報要求とを考慮した検索の実現を解決課題としている。

【課題を解決するための手段】

【０００８】

本発明は、前記課題を解決するため、検索エンジンのログを利用してクエリおよび検索結果を拡張することによりそれぞれの満たす情報要求の表現を行い、これらの間の関係性を基に検索結果のランキングを行うことで、高精度の検索を実現している。

【０００９】

具体的には、請求項１記載の発明は、ユーザ端末から検索指示されたクエリを含む電子文書を検索するときに検索エンジンの検索ログを利用する文書検索装置であって、前記検索ログに含まれたクエリに応じた検索結果のタイトルおよび概要文から拡張語を生成し、該拡張語を前記クエリの拡張情報として保存するクエリ情報保存手段と、前記検索ログからクリックされた検索結果の検索時のクエリを判別し、該クエリに関連する拡張語を前記クエリ情報保存手段から求め、検索結果がクリックされたときに投入されたクエリの拡張語群を検索結果の拡張情報として保存する検索結果情報保存手段と、ユーザ端末から検索指示されたクエリについて、クエリの拡張語を前記クエリ情報保存手段から取得し、ユーザ端末に送る照合処理手段と、前記検索指示されたクエリを前記拡張語で拡張した拡張クエリの検索で得られた検索結果に対して、前記検索結果情報保存手段に保存された拡張語群を付与して検索結果を拡張し、拡張された検索結果を前記拡張クエリとの類似度により並び替え、この並び替えた結果をリスト化した最終検索結果を前記ユーザ端末に送る検索結果処理手段と、を備えることを特徴としている。

【００１２】

請求項２記載の発明は、ユーザ端末から検索指示されたクエリを含む電子文書を検索するときに検索エンジンの検索ログを利用する文書検索方法であって、前記検索ログに含まれたクエリに応じた検索結果のタイトルおよび概要文から拡張語を生成し、該拡張語をクエリ情報保存手段に前記クエリの拡張情報として保存するクエリ情報保存ステップと、検索結果情報保存手段が、前記検索ログからクリックされた検索結果の検索時のクエリを判別し、該クエリに関連する拡張語を前記クエリ情報保存手段から求め、検索結果がクリックされたときに投入されたクエリの拡張語群を検索結果の拡張情報として保存する検索結果情報保存ステップと、照合処理手段が、ユーザ端末から検索指示されたクエリについてクエリの拡張語を前記クエリ情報保存手段から取得し、ユーザ端末に送る照合処理ステップと、検索結果処理手段が、前記検索指示されたクエリを前記拡張語で拡張した拡張クエリの検索で得られた検索結果に対して、前記検索結果情報保存手段に保存された拡張語群を付与して検索結果を拡張し、拡張された検索結果を前記拡張クエリとの類似度により並び替え、この並び替えた結果をリスト化した最終検索結果を前記ユーザ端末に送る検索結果処理ステップと、を有することを特徴としている。

【００１５】

請求項３記載の発明は、文書検索プログラムであって、請求項１記載の文書検索装置を構成する各手段としてコンピュータを機能させることを特徴としている。

【発明の効果】

【００１６】

請求項１～３記載の発明によれば、検索エンジンのログを利用してクエリおよび検索結

果を拡張し、これらの関係性に基づいて検索結果をランキングすることで、検索の精度が向上する。

【発明を実施するための最良の形態】

【００１７】

本発明は、検索エンジンのログ（クリックログ）を利用することで、クエリ拡張によるクエリの情報要求の表現、検索結果の拡張による検索結果の満たす情報要求の表現を行い、これらの間の関係性を基に検索結果のランキングを行っている。

【００１８】

すなわち、多くのユーザが有用だと判断し、アクセスが集中しているサイトのタイトルとスニペットを拡張語の取得源とみなすことにより、クエリに対する高い適合性を持った拡張語を抽出する。この拡張語を用いたクエリ拡張により、クエリの情報要求を表現する。

【００１９】

また、クリックログを解析することで、検索結果がクリック（閲覧）されたときに投入されたクエリを判別し、この判別したクエリの拡張語で検索結果を拡張することにより、この検索結果が満たすことができる情報要求を表現する。

【００２０】

そして、それぞれ拡張されたクエリと検索結果同士をキーワードベースで比較することで、クエリと検索結果が潜在的に持つ情報要求に沿った高精度な検索を可能としている。以下、図面に基づき本発明の実施形態に係る文書検索装置１を説明する。

【００２１】

図１は、本発明の実施形態に係る文書検索装置１の構成例を示している。この文書検索装置１は、インターネット経由で複数のユーザ端末１３とネットワーク接続されている。このユーザ端末１３をもってユーザはクエリを送信し文書検索を行う。

【００２２】

前記文書検索装置１は、主に２つの処理部、すなわちユーザから投入されたクエリに対する検索結果を取得する検索エンジン１００と、前記検索エンジン１００の取得した検索結果を適切なランキングに並べ替え（ｒｅ－ｒａｎｋｉｎｇ）、前記ユーザ端末１３へ返信する支援処理部１２５とで構成されている。

【００２３】

前記検索エンジン１００は、「Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂ（WWW）」もしくは「Ｍｏｂｉｌｅ　Ｗｅｂ（MW）」１０１から各サイトのデータを随時ダウンロードし、そのインデックスをインデックスＤＢ１０２に格納する。そして、前記ユーザ端末１３からの検索指示に従って前記インデックスＤＢ１０２を検索し、検索結果を取得する。

【００２４】

前記支援処理部１２５は、前記検索エンジン１００の検索結果を適切なランキングに並べ替えて前記ユーザ端末１３へ返信する。この支援処理部１２５は、図１に示すように、ログＤＢ１１０，解析処理部１１１，クエリ情報要求生成部１１４，検索結果情報要求生成部１１５，検索結果情報要求ＤＢ１１６，クエリ情報要求ＤＢ１１７，照合処理部１１９，ランキング処理部１２３として機能している。

【００２５】

ここで、前記ログＤＢ１１０には、ユーザの検索ログに含まれたクエリと、該クエリの検索結果から実際にユーザがクリックして閲覧した電子文書のＵＲＬとを対応付けたクリックログが格納されている。

【００２６】

前記解析処理部１１１は、前記クリックログを解析して、使用頻度が上位のクエリ（以下、頻度上位クエリとする）１１２を求める。そして、該頻度上位クエリ１１２を前記検索エンジン１００に送信し、各クエリに対する検索結果集合１１３を取得する。この取得した検索結果集合１１３を前記各情報要求生成部１１４．１１５に送信する。

【００２７】

前記クエリ情報要求生成部１１４は、前記クリックログを用いて前記検索結果集合１１３の解析を行い、クエリの情報要求を生成して前記クエリ情報要求ＤＢ１１７へ格納する。

【００２８】

ここでは、クエリの情報要求は、前記各頻度上位クエリ１１２に対する拡張語の集合として求められる。すなわち、前記クエリ情報要求ＤＢ１１７には、前記各頻度上位クエリ１１２と、該各クエリに対する拡張語の集合とが対応して格納される。

【００２９】

前記検索結果情報要求生成部１１５は、前記クリックログを用いて前記検索結果集合１１３の解析を行い、検索結果の情報要求を生成して前記検索結果情報要求ＤＢ１１６へ格納する。

【００３０】

ここでは、前記検索結果集合１１３からユーザの閲覧した文書のＵＲＬを含む検索結果を探索し、該検索結果の検索時のクエリを判別する。そして、この判別したクエリに対する拡張語の集合を該検索結果の情報要求として求める。すなわち、前記検索結果情報要求ＤＢ１１６には、検索結果の電子文書のＵＲＬと、該電子文書がクリックされたときに投入されたクエリの拡張語の集合とが対応して格納される。

【００３１】

ここまでの前記両ＤＢ１１６．１１７の生成処理は、図２に示すように、前記ユーザ端末１３と未接続のオフライン状態で実施される。この生成処理後に前記ユーザ端末１３と接続され、オンライン状態の処理が実施される。

【００３２】

前記ユーザ端末１３は、ネットワークに接続可能なブラウザなどのユーザインタフェース１３０を備えていればよい。例えば、パーソナルコンピュータ（ＰＣ）や携帯電話などが該当する。前記ユーザインタフェース１３０には、ユーザがクエリを入力するクエリ入力画面１３１、および検索結果を表示する検索結果表示画面１３２が表示される。

【００３３】

ここでは、ユーザは前記クエリ入力画面１３１にてクエリ１１８を投入する。投入されたクエリ１１８は前記照合処理部１１９へ送信される。前記照合処理部１１９は、前記クエリ１１８を受信すると、対応する情報要求１２０を前記クエリ情報要求ＤＢ１１７から取得する。この取得したクエリの情報要求１２０を前記ユーザ端末１３へ返信し、クエリ入力画面１３１へ表示させる。そして、該クエリの情報要求１２０と前記クエリ１１８とを組み合わせた拡張クエリ１２１が、前記クエリ入力画面１３１から前記検索エンジン１００に送信される。

【００３４】

前記検索エンジン１００は、受信した拡張クエリ１２１をもって前記インデックスＤＢ１０２を検索し、検索結果１２２を取得する。そして、取得した検索結果１２２を前記ランキング処理部１２３へ送信する。

【００３５】

前記ランキング処理部１２３は、前記検索結果１２２に応じた情報要求を前記検索結果情報要求ＤＢ１１６から取得し、該情報要求を用いて前記検索結果１２２を拡張する。そして、拡張した検索結果１２２を前記拡張クエリ１２１との関係性に基づいて並べ替え、ランク付けされた最終検索結果１２４を生成する。そして、生成した最終検索結果１２４を前記ユーザ端末１３へ返信して検索結果表示画面１３２に表示させ、ユーザに提示する。

【００３６】

ユーザは、前記検索結果表示画面１３２に表示された最終検索結果１２４をクリックして任意の電子文書を閲覧する。クエリ１１８の投入から最終検索結果１２４のクリックまでの操作情報は、検索ログ記録部１３３で随時取得され、該取得情報は前記ログＤＢ１１０へ蓄積される。ここまでの処理フローを図３に示す。

【００３７】
前記文書検索装置１の各機能ブロック１００．１０２．１１０．１１１．１１４～１１７．１１９．１２３の機能は、コンピュータのハードウェアとソフトウェアの協働で実現されている。また、前記文書検索装置１は、コンピュータの通常の構成要素、例えば図示省略の処理データなどを一時記憶する書き換え可能なメモリ（ＲＡＭ）と、前記ユーザ端末１３とのネットワーク接続に使用する通信デバイスと、ハードディスクドライブ装置などの保存部などを備え、前記各ＤＢ１０２．１１０．１１６．１１７は前記ハードディスクドライブ装置上に構築されている。以下、前記文書検索装置１の動作例を説明する。
【００３８】
＜動作例＞
前記文書検索装置１が実行する一連の処理は、主にオフラインで行われる情報要求生成フェーズと、オンラインで行われる検索実行フェーズから構成されている。以下、両フェーズの処理について、図４～８に基づき説明する。
【００３９】
（１）情報要求生成フェーズ
情報要求生成フェーズでは、クリックログを解析して、クエリの情報要求および該クエリの検索結果が満たす情報要求を生成する。この情報要求生成フェーズは通常、前記ユーザ端末１３と接続されないオフライン状態で、ユーザからの検索要求を受け付ける前に行われる。
【００４０】
図４．５は、情報要求生成フェーズの処理フローを示している。まず、前記解析処理部１１１は、前記ログＤＢ１１０に格納されたクリックログに含まれるクエリを使用頻度順に並べ、頻度上位クエリ１１２を得る。そして、該頻度上位クエリ１１２を前記検索エンジン１００に送信し、それぞれのクエリの上位Ｋ件の検索結果集合１１３を得る。ここでは、前記頻度上位クエリ１１２の任意のクエリｑに対する検索結果集合１１３を例に説明する。
【００４１】
前記検索結果集合１１３は、前記検索エンジン１００から前記クエリ情報要求生成部１１４および前記検索結果情報要求生成部１１５にそれぞれ送信され、クエリの情報要求、および検索結果が満たすことができる情報要求の算出が開始される。
【００４２】
＜クエリの情報要求＞
前記クエリ情報要求生成部１１４は、前記クエリｑに対する検索結果集合１１３の各検索結果sr_i（ｓｅａｒｃｈ　ｒｅｓｕｌｔ）（ｉ＝１，…，ｋ）において、そのタイトルとスニペットを形態素解析して内容語（キーワード）ｔを抽出し、このキーワード集合をベクトルＶ（sr_i）で表現する。
【００４３】
このベクトルＶ（sr_i）には、キーワードｔのｔｆ（ｔ）・ｉｄｆ（ｔ）に基づく重みが含まれている。ｔｆ（ｔ）は、その検索結果sr_iのタイトルとスニペットにおけるキーワードｔの出現頻度（Ｔｅｒｍ　Ｆｒｅｑｕｅｎｃｙ）、ｉｄｆはあるドキュメント集合におけるキーワードｔの出現頻度（Ｉｎｖｅｒｓｅ　Ｄｏｃｕｍｅｎｔ　Ｆｒｅｑｕｅｎｃｙ）である。
【００４４】
このベクトルＶ（sr_i）を、検索結果sr_iに対するアクセス集中度合ＡＣＤ（sr_i）（Ａｃｃｅｓｓ　Ｃｏｎｃｅｎｔｒａｔｉｏｎ　Ｄｅｇｒｅｅ）で加重平均する。その結果得られるベクトルを、クエリｑの情報要求ベクトルV_{IN}（ｑ）（Ｉｎｆｏｒｍａｔｉｏｎ　Ｎｅｅｄ　Ｖｅｃｔｏｒ）とする。具体的には、クエリｑの情報要求ベクトルV_{IN}（ｑ）は以下の式（１）で与えられる。この式（１）は、プログラムなどに定義されていればよい。
【００４５】

【数１】

$$V_{IN}(q) = \frac{\sum_i (ACD(sr_i) \times V(sr_i))}{\sum_i ACD(sr_i)} \quad \cdots\cdots \quad (1)$$

【００４６】

このように算出されたクエリｑの情報要求ベクトルＶ$_{IN}$（ｑ）は、クエリｑの拡張語の集合を表現するベクトルであり、前記クエリ情報要求ＤＢ１１７へ格納される。なお、アクセス集中度合ＡＣＤ（ｓｒ$_i$）は、例えば非特許文献１の手法により求めることができる。また、ここでは処理を簡単にするため、クリックログを解析することで得られる文書の絶対的クリック回数Ｃ（ｓｒ$_i$）をアクセス集中度合ＡＣＤ（ｓｒ$_i$）に置き換えてもよい。

【００４７】

＜検索結果の情報要求＞

前記検索結果情報要求生成部１１５は、前記クリックログを解析することで、前記検索結果集合１１３のある検索結果ｓｒがクリックされたときに投入されたクエリの集合ｑ$_j$（ｊ＝１，…，ｍ）を求める。また、該集合の各クエリから検索結果ｓｒに対するアクセス集中度合ＡＣＤ（ｓｒ，ｑ$_j$）を求める。

【００４８】

次に、クエリｑ$_j$をもって前記クエリ情報要求ＤＢ１１７を検索し、該クエリｑ$_j$に対応する情報要求ベクトルＶ$_{IN}$（ｑ$_j$）を取得する。そして、この情報要求ベクトルＶ$_{IN}$（ｑ$_j$）を前記アクセス集中度合ＡＣＤ（ｓｒ，ｑ$_j$）で加重平均して得られるベクトルを、検索結果ｓｒが満たすことができる（できた）情報要求ベクトルＶ$_{IN}$（ｓｒ）とみなす。この情報要求ベクトルＶ$_{IN}$（ｓｒ）は、以下の式（２）で与えられる。この式（２）もプログラムなどに定義されていればよい。

【００４９】

【数２】

$$V_{IN}(sr) = \frac{\sum_j (ACD(sr, q_j) \times V_{IN}(q_j))}{\sum_j ACD(sr, q_j)} \quad \cdots\cdots \quad (2)$$

【００５０】

このように算出された検索結果の情報要求ベクトルＶ$_{IN}$（ｓｒ）は、検索結果ｓｒがクリックされたときに投入されたクエリの拡張語集合を表現するベクトルであり、前記検索結果情報要求ＤＢ１１６へ格納される。なお、処理を簡単にするため、クリックログを解析することで得られる文書の絶対的クリック回数Ｃ（ｓｒ，ｑ$_j$）をアクセス集中度合ＡＣＤ（ｓｒ，ｑ$_j$）に置き換えてもよい。また、クリックログを解析することで得られるクエリｑ$_j$の全投入回数ＴＩＮ（ｑ$_j$）（Ｔｏｔａｌ　Ｉｎｐｕｔ　Ｎｕｍｂｅｒ）で絶対的クリック回数Ｃ（ｓｒ，ｑ$_j$）を正規化した値で置き換えてもよい。

【００５１】

ここで、前記クエリｑ$_j$をもって前記クエリ情報要求ＤＢ１１７を検索した際、該クエリｑ$_j$に対応する情報要求ベクトルＶ$_{IN}$（ｑ$_j$）が存在しない場合もある。そのような場合、前記検索結果情報要求生成部１１５は、このクエリｑ$_j$を前記解析処理部１１１へ送信し、前述したクエリの情報要求の生成処理と同様の手順で、該クエリｑ$_j$に対応する情報要求ベクトルＶ$_{IN}$（ｑ$_j$）の算出を行えばよい。この算出した情報要求ベクトルＶ$_{IN}$（ｑ$_j$

）を前記クエリ情報要求ＤＢ１１７へ格納し、これを用いて前記式（２）により検索結果ｓｒの情報要求ベクトルV_{IN}（ｓｒ）を算出すればよい。以上の情報要求生成フェーズにおけるデータ例を図６に示す。

【００５２】

（２）検索実行フェーズ

検索実行フェーズでは、情報要求生成フェーズで生成したクエリおよび検索結果の情報要求を用いて、ユーザの投入したクエリを拡張するとともに、該拡張クエリに対する検索結果も拡張する。そして、拡張されたクエリと検索結果との間の関係性に基づき検索結果を並べ替え、最終検索結果を生成している。この検索実行フェーズは、前記ユーザ端末１３と接続されたオンライン状態で行われる。

【００５３】

図７は、検索実行フェーズの処理フローを示している。まず、前記照合処理部１１９は、ユーザが前記クエリ入力画面１３１をもって投入したクエリ１１８を受信する。

【００５４】

前記照合処理部１１９は、前記クエリ情報要求ＤＢ１１７から前記クエリ１１８に対応する情報要求ベクトルV_{IN}（ｑ）を取得し、これをクエリの情報要求１２０として前記ユーザ端末１３へ返信しクエリ入力画面１３１へ表示させる。このクエリの情報要求１２０によって、前記クエリ１１８が拡張される。

【００５５】

すなわち、前記クエリ１１８と前記クエリの情報要求１２０との組み合わせが、拡張クエリ１２１として前記クエリ入力画面１３１から前記検索エンジン１００へ送信される。前記検索エンジン１００は、受信した拡張クエリ１２１を用いて検索を行い、取得した検索結果１２２を前記ランキング処理部１２３へ送信する。

【００５６】

ここで、通常の検索結果は、ランク付けされた検索結果のタイトルの一部、本文の一部（スニペット）、およびＵＲＬが返されるが、ここではランク付けされた検索結果のタイトルの全文、全本文、およびＵＲＬが返される。この時点でのランク付けは、従来の全文検索アルゴリズムに沿って行われる。

【００５７】

前記ランキング処理部１２３は、前記検索結果１２２の各検索結果ｓｒに対応する情報要求ベクトルV_{IN}（ｓｒ）を前記検索結果情報要求ＤＢ１１６から取得する。そして、取得した情報要求ベクトルV_{IN}（ｓｒ）で表現される拡張語群を各検索結果ｓｒのタイトルと本文に付与する。これにより、検索結果１２２の拡張が行われる。

【００５８】

そして、このように拡張された検索結果１２２を、拡張クエリ１２１とのキーワードベースでの類似度を考慮に入れて並べ替える（ｒｅ－ｒａｎｋｉｎｇ）。そして、この並べ替えた結果をリスト化した最終検索結果１２４を前記ユーザ端末１３へ返信し、検索結果表示画面１３２へ表示させてユーザへ提示する。ユーザは、提示された前記最終検索結果１２４から任意の電子文書をクリックして閲覧する。

【００５９】

なお、検索実行フェーズにおけるユーザのクエリの投入から最終検索結果のクリックまでの行動情報は、前記検索ログ記録部１３３で常に監視・取得される。この行動情報は、前記文書検索装置１へ送信され、新たなクリックログとして前記ログＤＢ１１０に蓄積される。これにより、ユーザの行動情報が以降の情報要求生成フェーズに随時反映され、時々刻々と変化するユーザの情報要求を適切に把握することが可能となる。なお、この検索ログ記録部１３３は、前記文書検索装置１内に実装されていてもよい。ここまでの検索実行フェーズにおけるデータ例を図８に示す。

【００６０】

＜発明の効果＞

以上のように、文書検索システムに本発明の前記文書検索装置１を配置し、検索結果の

精度評価を行った実験の結果を表１に示す。

【００６１】

【表１】

	5	10	15	20	30	50
BM25 (baseline1)	42.4 **	37.5 **	37.6 **	37.6 **	36.5 **	36.1 *
RCN (baseline2)	34.9 **	34.1 **	33.7 **	31.2 **	30.7 **	30.2 **
QEC (baseline3)	46.3 *	43.7 **	43.5	42.4 *	42.5	39.4
本発明	54.1	47.3	44.8	45.2	42.9	40.2

【００６２】

精度評価の指標は「Ｐｒｅｃｉｓｉｏｎ@Ｘ」と呼ばれるものを使用している。これは、クエリに対して正解であるサイトが検索結果の上位Ｘ件に多く入るほど高い数値を出す指標である。

【００６３】

なお、本発明との比較対象の手法は、（１）ＢＭ２５，（２）クリック回数の多いサイトで検索結果を並べ替える方法（Ｒｅ－ｒａｎｋｉｎｇ ｂｙ Ｃｌｉｃｋ Ｎｕｍｂｅｒ，ＲＣＮ），（３）クリック回数に基づくクエリ拡張法（Ｑｕｅｒｙ Ｅｘｐａｎｓｉｏｎ ｍｅｔｈｏｄ ｕｓｉｎｇ Ｃｌｉｃｋ ｎｕｍｂｅｒ，ＱＥＣ）の３種類としている。

【００６４】

（１）のＢＭ２５は、クエリーサイト間のキーワードマッチングベースのランキングとして幅広く用いられている手法である。（２）のＲＣＮは、単純に検索結果をそのクリック回数順で並べ替える手法である。（３）のＱＥＣは、クリック回数の多い検索結果のタイトルとスニペットからキーワードを抽出し、抽出したキーワードでクエリ拡張を行い、その拡張されたクエリとサイト間の類似度を基にランキングを行う手法である。

【００６５】

表１中の「*」，「**」は、それぞれウィルコクソンの符号付順位和検定において、本発明の手法と各比較手法との統計的有意差が５％，１％であった結果である。太文字になっている手法と数値が「Ｐｒｅｃｉｓｉｏｎ@Ｘ」に対する最大値である。

【００６６】

表１に示すように、本発明の手法は全ての検索結果ランク@Ｘにおいて最大の精度を達成しており、特に検索結果の上位ランク１≦Ｘ≦２０の領域における精度を著しく向上させている。

【００６７】

なお、本発明は、コンピュータを前記文書検索装置１の各機能ブロック１００．１０２．１１０．１１１．１１４～１１７．１１９．１２３として機能させる文書検索プログラムとしても提供することができる。このプログラムは、各機能ブロック１００．１０２．１１０．１１１．１１４～１１７．１１９．１２３の全ての機能を実現させるものでもよく、あるいは一部の機能を実現させるものであってもよい。

【００６８】

このプログラムは、Ｗｅｂサイトなどからのダウンロードによってコンピュータに提供される。また、前記プログラムは、ＣＤ－ＲＯＭ，ＤＶＤ－ＲＯＭ，ＣＤ－Ｒ，ＣＤ－Ｒ

Ｗ，ＤＶＤ－Ｒ，ＤＶＤ－ＲＷ，ＭＯ，ＨＤＤ，Ｂｌｕ－ｒａｙ　Ｄｉｓｋ（登録商標）などの記録媒体に格納してコンピュータに提供してもよい。

【図面の簡単な説明】

【００６９】

【図１】本発明の実施形態に係る文書検索装置の構成図。

【図２】同　情報要求生成フェーズの概略図。

【図３】同　検索実行フェーズの概略図。

【図４】同　情報要求生成フェーズ前半の処理フロー図。

【図５】同　情報要求生成フェーズ後半の処理フロー図。

【図６】同　情報要求生成フェーズのデータ例。

【図７】同　検索実行フェーズの処理フロー図。

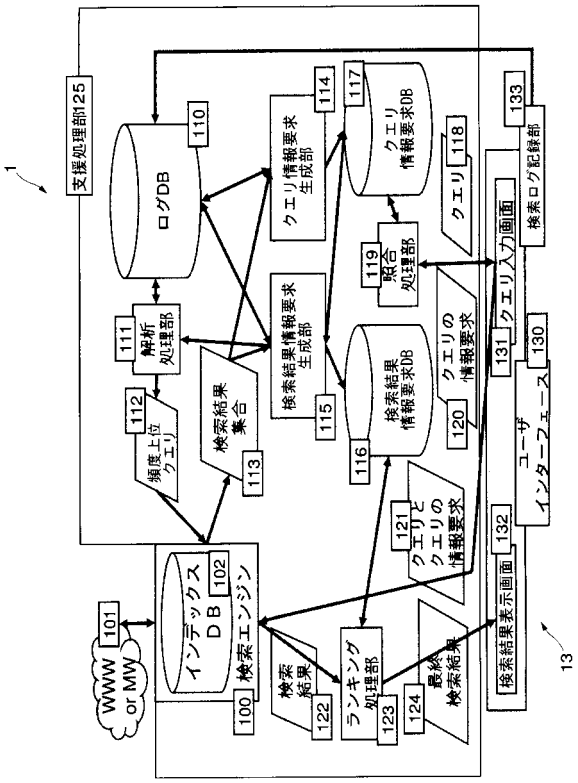
【図８】同　検索実行フェーズのデータ例。

【符号の説明】

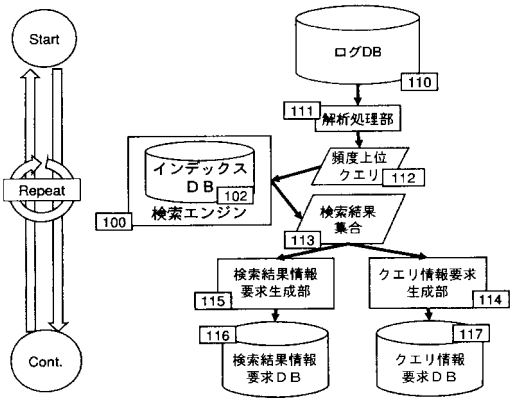
【００７０】

１…文書検索装置
１３…ユーザ端末
１００…検索エンジン
１０１…Ｗｏｒｌｄ　Ｗｉｄｅ　ＷｅｂもしくはＭｏｂｉｌｅ　Ｗｅｂ
１０２…インデックスＤＢ
１１０…ログＤＢ
１１１…解析処理部
１１２…頻度上位クエリ
１１３…検索結果集合
１１４…クエリ情報要求生成部
１１５…検索結果情報要求生成部
１１６…検索結果情報要求ＤＢ
１１７…クエリ情報要求ＤＢ
１１８…クエリ
１１９…照合処理部
１２０…クエリの情報要求
１２１…クエリとクエリの情報要求（拡張クエリ）
１２２…検索結果
１２３…ランキング処理部（検索結果処理手段）
１２４…最終検索結果
１２５…支援処理部
１３０…ユーザインタフェース
１３１…クエリ入力画面
１３２…検索結果表示画面
１３３…検索ログ記録部

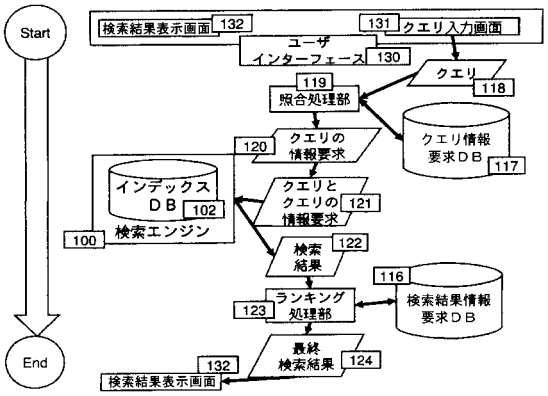
【図１】


1
支援処理部125
ログDB
110
クエリ情報要求生成部
114
クエリ情報要求DB
117
クエリ
118
照合処理部
119
検索ログ記録部
133
クエリ入力画面
131
ユーザインターフェース
130
検索結果情報要求生成部
115
検索結果情報要求DB
116
クエリの情報要求
120
解析処理部
111
検索結果集合
113
頻度上位クエリ
112
クエリとクエリの情報要求
121
検索結果表示画面
132
インデックスDB
102
検索エンジン
100
WWW or MW
101
検索結果
122
ランキング処理部
123
最終検索結果
124
13

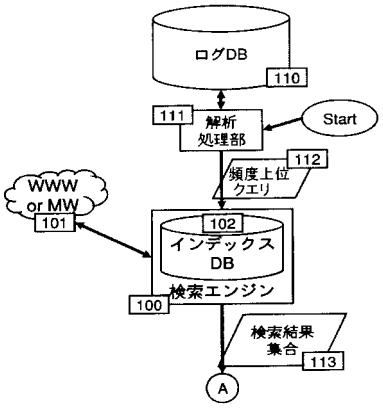
【図２】


Start
Repeat
Cont.
ログDB
110
解析処理部
111
頻度上位クエリ
112
インデックスDB
102
検索エンジン
100
検索結果集合
113
検索結果情報要求生成部
115
クエリ情報要求生成部
114
検索結果情報要求DB
116
クエリ情報要求DB
117

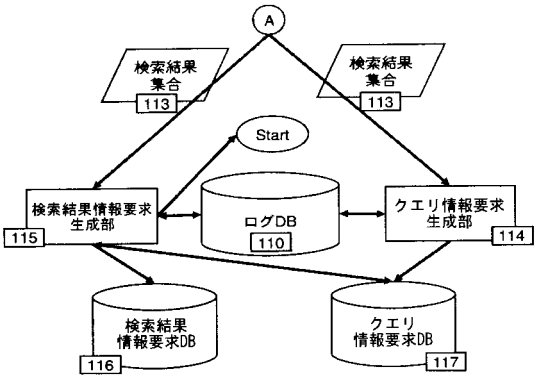
【図３】


Start
End
検索結果表示画面
132
131
クエリ入力画面
ユーザインターフェース
130
クエリ
118
照合処理部
119
クエリの情報要求
120
クエリ情報要求DB
117
インデックスDB
102
検索エンジン
100
クエリとクエリの情報要求
121
検索結果
122
116
検索結果情報要求DB
ランキング処理部
123
最終検索結果
124
検索結果表示画面
132

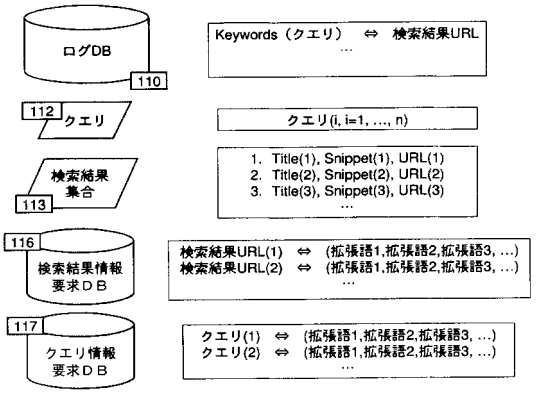
【図４】


ログDB
110
111
解析処理部
Start
112
頻度上位クエリ
WWW or MW
101
102
インデックスDB
検索エンジン
100
検索結果集合
113
A

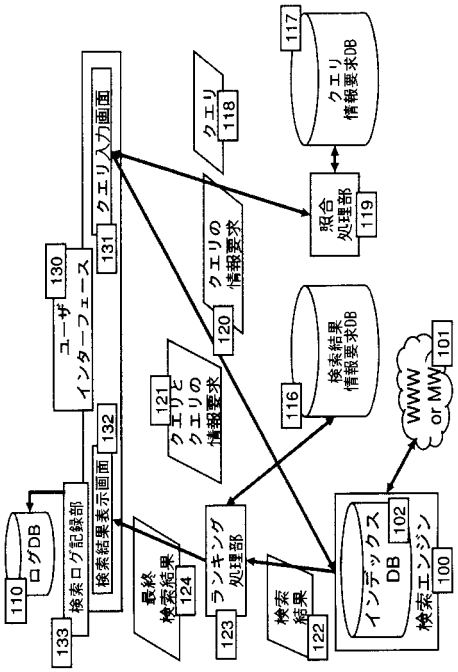
【図５】


A
検索結果
集合
113
検索結果
集合
113
Start
検索結果情報要求
生成部
115
ログDB
110
クエリ情報要求
生成部
114
検索結果
情報要求DB
116
クエリ
情報要求DB
117

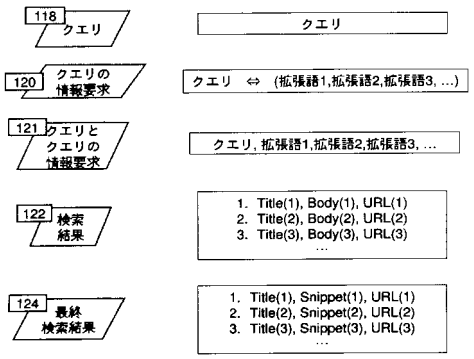
【図６】


ログDB
110
Keywords（クエリ） ⇔ 検索結果URL
…
112
クエリ
クエリ(i, i=1, …, n)
検索結果
集合
113
1. Title(1), Snippet(1), URL(1)
2. Title(2), Snippet(2), URL(2)
3. Title(3), Snippet(3), URL(3)
…
116
検索結果情報
要求ＤＢ
検索結果URL(1) ⇔ (拡張語1,拡張語2,拡張語3, …)
検索結果URL(2) ⇔ (拡張語1,拡張語2,拡張語3, …)
…
117
クエリ情報
要求ＤＢ
クエリ(1) ⇔ (拡張語1,拡張語2,拡張語3, …)
クエリ(2) ⇔ (拡張語1,拡張語2,拡張語3, …)
…

【図７】


117
クエリ
情報要求DB
クエリ
118
照合
処理部
119
クエリ入力画面
131
ユーザ
インターフェース
130
クエリの
情報要求
120
検索結果
情報要求DB
116
121
クエリと
クエリの
情報要求
WWW
or MW
101
132
検索結果表示画面
ログDB
110
検索ログ記録部
133
最終
検索結果
124
ランキング
処理部
123
検索
結果
122
インデックス
DB
102
検索エンジン
100

【図８】


118
クエリ
クエリ
120
クエリの
情報要求
クエリ ⇔ (拡張語1,拡張語2,拡張語3, …)
121
クエリと
クエリの
情報要求
クエリ, 拡張語1,拡張語2,拡張語3, …
122
検索
結果
1. Title(1), Body(1), URL(1)
2. Title(2), Body(2), URL(2)
3. Title(3), Body(3), URL(3)
…
124
最終
検索結果
1. Title(1), Snippet(1), URL(1)
2. Title(2), Snippet(2), URL(2)
3. Title(3), Snippet(3), URL(3)
…

フロントページの続き

(72)発明者　戸田　浩之
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　松浦　由美子
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　片岡　良治
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

審査官　岩間　直純

(56)参考文献　村田　眞哉，ほか，検索結果中のアクセス集中サイトを利用したクエリ拡張法の提案，データベースとＷｅｂ情報システムに関するシンポジウム　情報処理学会シンポジウムシリーズ，日本，社団法人情報処理学会，２００７年１１月２７日，Ｖｏｌ．２００７，Ｎｏ．３，ｐｐ．１－７
村田　眞哉，ほか，検索結果中のアクセス集中サイトを利用したクエリ拡張法の提案，日本データベース学会Ｌｅｔｔｅｒｓ，日本，日本データベース学会，２００８年　3月２１日，Ｖｏｌ．６，Ｎｏ．４，ｐｐ．４５－４８

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０