

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5980520号  
(P5980520)

(45) 発行日 平成28年8月31日(2016.8.31)

(24) 登録日 平成28年8月5日(2016.8.5)

(51) Int.Cl. F I  
**G 0 6 F 17/30 (2006.01)**  
 G 0 6 F 17/30 3 4 0 Z  
 G 0 6 F 17/30 1 7 0 A  
 G 0 6 F 17/30 4 1 4 B

請求項の数 21 (全 16 頁)

(21) 出願番号 特願2012-31022 (P2012-31022)  
 (22) 出願日 平成24年2月15日 (2012.2.15)  
 (65) 公開番号 特開2012-221489 (P2012-221489A)  
 (43) 公開日 平成24年11月12日 (2012.11.12)  
 審査請求日 平成27年2月12日 (2015.2.12)  
 (31) 優先権主張番号 10-2011-0032898  
 (32) 優先日 平成23年4月8日 (2011.4.8)  
 (33) 優先権主張国 韓国 (KR)

(73) 特許権者 390019839  
 三星電子株式会社  
 Samsung Electronics  
 Co., Ltd.  
 大韓民国京畿道水原市靈通区三星路129  
 129, Samsung-ro, Yeon  
 gtong-gu, Suwon-si, G  
 yeonggi-do, Republic  
 of Korea

最終頁に続く

(54) 【発明の名称】 効率的にクエリを処理する方法及び装置

(57) 【特許請求の範囲】

【請求項1】

クエリストリングから、長さの異なる複数個の部分ストリングから構成された有効ストリングセットを生成する段階と、

多数の文書の情報が保存されたデータベースに対する前記有効ストリングセットのサブセットのアクセスコストに基づいて、前記サブセットのうちいずれか一つを候補セットとして決定する段階と、

前記候補セットを利用し、前記データベースに保存された情報から、前記クエリストリングが存在する文書を検索する段階と、を含み、

前記有効ストリングセットを生成する段階は、

前記クエリストリングを長さの異なる複数個の N - gram に分離し、

前記複数個の N - gram において、前記データベースの索引語に含まれる N - gram を選択し、

前記選択された N - gram において、他の N - gram に含まれない N - gram セットを、前記有効ストリングセットとして決定することを含む、クエリ処理方法。

【請求項2】

前記候補セットは、

アクセスコストが基準値以下を有するサブセットであることを特徴とする請求項1に記載のクエリ処理方法。

【請求項3】

10

20

前記基準値は、

前記有効ストリングセットのサブセットに係わるアクセスコストの算出時、既算出のアクセスコストのうち最小値であることを特徴とする請求項 2 に記載のクエリ処理方法。

【請求項 4】

前記アクセスコストは、

前記データベースで、前記サブセットに含まれた部分ストリングそれぞれのポスティングリストをアクセスして読み取るのにかかるコストの和と、前記データベースで、前記サブセットに含まれた有効ストリングのポスティングリストに共通して含まれた文書の識別情報にアクセスして読み取るのにかかるコストとのうち、少なくとも一つであることを特徴とする請求項 1 に記載のクエリ処理方法。

10

【請求項 5】

前記有効ストリングセットの部分ストリングのうち、少なくとも 2 つの部分ストリングの長さは、互いに異なることを特徴とする請求項 1 に記載のクエリ処理方法。

【請求項 6】

前記有効ストリングセットの部分ストリングは、前記有効ストリングセットの他の部分ストリングに含まれないことを特徴とする請求項 1 に記載のクエリ処理方法。

【請求項 7】

前記候補セットは、

前記有効ストリングセットのサブセットのうち、前記アクセスコストが最小であるサブセットとして決定されることを特徴とする請求項 1 に記載のクエリ処理方法。

20

【請求項 8】

前記候補セットは、

前記有効ストリングセットのサブセットのうち、部分ストリングが追加されるときにアクセスコストより、アクセスコストの小さいサブセットでもって決定されることを特徴とする請求項 1 に記載のクエリ処理方法。

【請求項 9】

前記候補セットとして決定する段階は、

前記有効ストリングセットのサブセットをツリー構造に整列し、

深さ優先探索方法で前記ツリー構造でのサブセットを選択し、

前記選択されたサブセットのアクセスコストを算出し、

最小のアクセスコストを有するサブセットを候補セットとして決定することを含むことを特徴とする請求項 1 に記載のクエリ処理方法。

30

【請求項 10】

前記候補セットとして決定する段階は、

前記有効ストリングセットのサブセットのうち、部分ストリングの個数が同一である第 1 サブセットを選択し、

前記第 1 サブセットそれぞれに係わるアクセスコストを算出し、

最小のアクセスコストを有するサブセットを候補セットとして予想し、

前記有効ストリングセットのサブセットのうち、前記予想された候補セットに部分ストリングが追加された第 2 サブセットを選択し、

40

前記第 2 サブセットそれぞれに係わるアクセスコストが、前記予想された候補セットのアクセスコストより大きければ、前記予想された候補セットを候補セットとして決定することを含むことを特徴とする請求項 1 に記載のクエリ処理方法。

【請求項 11】

前記データベースは、

索引ツリー及びポスティングリストを含む逆索引データベースと、

識別情報を有する多数の文書が保存された文書データベースと、を含むことを特徴とする請求項 1 に記載のクエリ処理方法。

【請求項 12】

前記文書を決定する段階は、

50

前記逆索引データベースで、前記候補セットの部分ストリングいずれもとマッチングしていいる文書の識別情報を検索し、

前記文書データベースで、前記文書の識別情報を有する文書を検索することを含むことを特徴とする請求項 1 1 に記載のクエリ処理方法。

【請求項 1 3】

請求項 1 ないし請求項 1 2 のうち、いずれか一項に記載の方法を遂行するためのプログラムが記録されるコンピュータで読み取り可能な記録媒体。

【請求項 1 4】

クエリストリングが入力され、前記クエリストリングが存在する文書が出力されるユーザ・インターフェースと、

多数の文書に係わる情報が保存されたデータベースと、

前記クエリストリングから、長さの異なる複数個の部分ストリングから構成された有効ストリングセットを生成し、前記データベースに対する前記有効ストリングセットのサブセットのアクセスコストに基づいて、前記サブセットのうちいずれか一つを候補セットとして決定し、前記候補セットを利用し、前記データベースに保存された情報から、前記クエリストリングが存在する文書を検索するプロセッサと、

を含み、

前記プロセッサは、

前記クエリストリングを長さの異なる複数個の N - gram に分離し、

前記複数個の N - gram において、前記データベースの索引語に含まれる N - gram を選択し、

前記選択された N - gram において、他の N - gram に含まれない N - gram セットを、前記有効ストリングセットとして決定する、

ことにより前記有効ストリングセットを生成する、クエリ処理装置。

【請求項 1 5】

前記アクセスコストは、

前記データベースで、前記サブセットに含まれた部分ストリングそれぞれのポスティングリストにアクセスして読み取るのにかかるコストの和と、前記データベースで、前記サブセットに含まれた有効ストリングのポスティングリストに共通して含まれた文書の識別情報にアクセスして読み取るのにかかるコストと、のうち少なくとも一つであることを特徴とする請求項 1 4 に記載のクエリ処理装置。

【請求項 1 6】

前記有効ストリングセットの部分ストリングのうち、少なくとも 2 つの部分ストリングの長さは、互いに異なることを特徴とする請求項 1 4 に記載のクエリ処理装置。

【請求項 1 7】

前記有効ストリングセットの部分ストリングは、前記有効ストリングセットの他の部分ストリングに含まれないことを特徴とする請求項 1 4 に記載のクエリ処理装置。

【請求項 1 8】

前記候補セットは、

前記有効ストリングセットのサブセットのうち、前記アクセスコストが最小であるサブセットをもって決定されることを特徴とする請求項 1 4 に記載のクエリ処理装置。

【請求項 1 9】

前記候補セットは、

前記有効ストリングセットのサブセットのうち、部分ストリングが追加されるときにアクセスコストより、アクセスコストの小さいサブセットとして決定されることを特徴とする請求項 1 4 に記載のクエリ処理装置。

【請求項 2 0】

前記データベースは、

索引ツリー及びポスティングリストを含む逆索引データベースと、

識別情報を有する多数の文書が保存された文書データベースと、を含むことを特徴とす

10

20

30

40

50

る請求項 1 4 に記載のクエリ処理装置。

【請求項 2 1】

前記プロセッサは、

前記逆索引データベースで、前記候補セットの部分ストリングいずれともマッチングしている前記文書の識別情報を検索し、

前記文書データベースで、前記文書の識別情報を有する文書を検索することを含むことを特徴とする請求項 2 0 に記載のクエリ処理装置。

【発明の詳細な説明】

【技術分野】

10

【0 0 0 1】

本発明は、効率的にクエリストリング (query string) を処理する方法及びその装置に関する。

【背景技術】

【0 0 0 2】

テキスト文書 (document) の検索 (searching) は、自然言語 (natural language) で表現された情報の検索、遺伝子列分析 (genetic sequence processing) などの多様な分野で広く使われている。蛋白質と DNA との列 (sequence) で特別のアルファベット列 (例えば、DNA の A, C, G, T) は、テキスト文書と見なされる。かようなテキスト文書の検索のための代表的な確率モデルとして、N - gramモデルを挙げることができる。

20

【発明の概要】

【発明が解決しようとする課題】

【0 0 0 3】

本発明は、長さの異なる複数個の部分ストリングを使用し、検索性能を向上させるクエリ処理方法及びその装置を提供するものである。

【0 0 0 4】

本発明はまた、複数個の部分ストリングセットのサブセットを効率的に決定し、クエリ処理性能を向上させるクエリ処理方法及びその装置を提供するものである。

【課題を解決するための手段】

【0 0 0 5】

30

一類型によるクエリ処理方法は、クエリストリングから長さが同じあるか、あるいは異なる複数個の部分ストリングから構成された有効ストリングセットを生成する段階と、多数の文書の情報が保存されたデータベースに対する前記有効ストリングセットのサブセットのアクセスコストに基づいて、前記サブセットのうちいずれか一つを候補セットとして決定する段階と、前記候補セットを利用し、前記データベースに保存された情報から、前記クエリストリングが存在する文書を検索する段階と、を含む。

【0 0 0 6】

また、一類型によるクエリ処理装置は、クエリストリングが入力され、前記クエリストリングが存在する文書が出力されるユーザ・インターフェース；多数の文書に係わる情報が保存されたデータベース；及び前記クエリストリングから、長さの異なる複数個の部分ストリングから構成された有効ストリングセットを生成し、前記データベースに対する前記有効ストリングセットのサブセットのアクセスコストに基づいて、前記サブセットのうちいずれか一つを候補セットとして決定し、前記候補セットを利用し、前記データベースに保存された情報から、前記クエリストリングが存在する文書を検索するプロセッサ；を含む。

40

【発明の効果】

【0 0 0 7】

本発明のクエリ処理方法及びその装置は、長さが固定されていない複数個の部分ストリングを使用するために、検索性能を向上させることができる。

【0 0 0 8】

50

併せて、既存の逆索引データベースの構造を変更させずに、候補セットに含まれる部分  
ストリングまたは候補セットの選定方法を改善したので、既存の逆索引データベースをそ  
のまま活用することができる。

【図面の簡単な説明】

【0009】

【図1】本発明の一実施形態によるクエリ処理装置のブロック図である。

【図2】本発明の一実施形態によるクエリストリングによる文書を検索する方法について  
説明するフローチャートである。

【図3】本発明の一実施形態によるストリングセット生成部の細部ブロック図である。

【図4】本発明の一実施形態によるツリー構造で具現された索引語の構造を図示した図面  
である。

10

【図5】本発明の一実施形態による有効ストリングを生成する過程について説明するフロ  
ーチャートである。

【図6】本発明の一実施形態による候補セット決定部の細部ブロック図である。

【図7】本発明の第1実施形態による候補セットを決定する方法について説明するフロ  
ーチャートである。

【図8】本発明の一実施形態による第1実施形態を介して候補セットを決定する方法につ  
いて説明するための図面である。

【図9】第2実施形態による候補セットを決定する方法について説明するフローチャート  
である。

20

【図10】本発明の一実施形態による第2実施形態を介して候補セットを決定する方法に  
ついて説明するための図面である。

【発明を実施するための形態】

【0010】

以下、添付された図面を参照しつつ、本発明の実施形態について詳細に説明する。

【0011】

図1は、本発明の一実施形態によるクエリ処理装置 (apparatus for processing qu  
ery) のブロック図である。図1を参照すれば、クエリ処理装置100は、ユーザ・イン  
ターフェース105、プロセッサ135及びストレージ165から構成される。クエリ処  
理装置は、ウェブページなどで表示される検索エンジンを具備し、クエリストリング (qu  
ery string) に係わる検索結果を出力するPC (personal computer)、携帯端末機な  
どであってもよく、ユーザ端末と、インターネットのネットワークとに連結されている別  
途のサーバであってもよい。従って、図1に図示されたクエリ処理装置100は、他のP  
C、他の携帯端末機、他のサーバなどと通信するための通信モジュールなど、他の一般  
的な構成要素をさらに含んでもよい。

30

【0012】

ユーザ・インターフェース105は、ユーザからクエリを入力され、このクエリによる  
文書の検索結果を出力する装置である。ここで、ユーザ・インターフェースは、ユーザ  
からクエリを入力されるためのキーボード、マウスのような入力装置と、ユーザに文書の  
検索結果を表示するための出力装置と、を含み、ウェブページのようなGUI (graphic u  
ser interface) で表現されることもできる。ここで、クエリは、あるストリングを含ん  
でいる文書を検索せよというユーザの要請を意味する。かようなクエリに含まれている  
ストリングを、以下では簡単に、「クエリストリング」と呼ぶ。

40

【0013】

ストレージ165には、任意のストリングを含んでいる多数の文書の情報が保存されて  
いる。例えば、ストレージ165には、索引語が保存されている索引語DB (index ter  
m database) 140が含まれている。ストレージ165にはまた、文書検索に利用され  
る部分ストリングである索引語と、索引語が含まれた文書の識別情報とがマッチングされ  
た逆索引DB (inverted index database) 150が保存されている。この逆索引DB

150は、文書に含まれた文字をN-gramに分離し、N-gramに分離された各部分スト

50

リングを、当該文書に係わる情報とマッチングさせることによって生成される。また、文書の識別情報に対応する文書が保存された文書DB (document database) 160をさらに含むことができる。一方、文書DB 160は、クエリ処理装置に含まれるが、クエリ処理装置と通信可能な外部装置またはサーバに含まれてもよい。

【0014】

プロセッサ135は、クエリストリングが入力されれば、クエリストリングをN-gramに分離し、逆索引DB 150で、クエリストリングに含まれた部分ストリングとマッチングされた文書の識別情報を読み取る。また、プロセッサ135は、検索された文書の識別情報を有する文書を、文書DB 160で読み取る。プロセッサ135は、その機能によって、ストリングセット生成部110、候補セット決定部120及び文書検索部130を含む。各構成要素の機能は、下記文書の検索方法でさらに具体的に説明する。

10

【0015】

図2は、本発明の一実施形態による、プロセッサがクエリストリングによる文書を検索する方法について概略的に説明するフローチャートである。図2を参照すれば、210段階で、入力部からクエリストリングを受信すれば、220段階で、ストリングセット生成部110は、クエリストリングから、文書検索のために利用することができる部分ストリング、すなわち、有効ストリングから構成された有効ストリングセットを生成する。有効ストリングは、長さの異なる複数個の部分ストリングであって、有効ストリングは、他の有効ストリングに含まれないが、これについては後述する。

【0016】

20

そして、230段階で、候補セット決定部120は、有効ストリングセットのサブセットのうちいずれか1つのサブセットを文書を検索するためのセット(以下、「候補セット」とする)として決定し、240段階で、検索部130は、前記した候補セットを利用し、前記クエリストリングが存在する文書を検索する。プロセッサの各構成要素に係わる機能及びクエリ処理方法は、以下でさらに具体的に説明する。

【0017】

図3は、本発明の一実施形態によるストリングセット生成部110の細部ブロック図である。図3に図示されているように、ストリングセット生成部110は、クエリストリングを、長さの異なる複数個のN-gramに分離するN-gram分離部310、及びN-gramにおいて、索引語DB 140に保存された索引語に含まれるN-gramを、有効ストリングとして選択するストリング選択部330を含む。

30

【0018】

索引語DB 140は、索引語に係わる情報が保存されたデータベースであり、索引語になるストリングの長さ範囲も保存されていてもよい。また、索引語DB 140は、ツリー構造に具現されてもよい。例えば、各ノードに単位文字が配置され、長さ範囲が設定される。従って、親ノードから子ノードに順次に連結された長さ範囲内の文字が索引語となる。索引語DB 140に保存された索引語は、逆索引DB 150に保存された索引語と一致しうる。説明の便宜を図るために、索引語DB 140について別途に説明したが、ストリング選択部330は、N-gramにおいて、逆索引DB 150の索引語に含まれたN-gramを、有効ストリングとして選択することができることは言うまでもない。

40

【0019】

図4は、本発明の一実施形態による、ツリー構造で具現された索引語の構造を図示した図面である。図4を参照すれば、ツリー構造の各ノードには、単位文字が配置され、索引語の最小長は3であり、最大長は5であることが分かる。そして、「s」、「u」及び「b」に該当するノードは、順次に連結されているために、「sub」は索引語となる。しかし、「s」、「t」、「r」、「i」及び「c」に該当するノードは、順次に連結されていないために、「stric」は索引語にならない。

【0020】

図5は、本発明の一実施形態による有効ストリングを生成する過程について説明するフローチャートである。

50

## 【0021】

図5に図示されているように、クエリストリングが入力されれば、図5の510段階で、N-gram分離部310はクエリストリングを長さの異なる複数個のN-gramに分離する。N-gram長は、既定の長さ範囲に含まれる。例えば、クエリストリングがm個の文字から構成されており、長さ範囲がiからkであるならば(このとき、 $2 \leq i < k \leq m$ である)、N-gram分離部310は、クエリストリングを、長さがiであるN-gram、長さがi+1であるN-gram、...、長さがk-1であるN-gram、長さがkであるN-gramに分離する。

## 【0022】

例えば、「substring」というクエリストリングが入力されれば、索引語DB 140に保存された長さ範囲が3ないし5であるから、N-gram分離部310は、クエリストリングを、長さが3以上5以下である第1部分ストリングに分離する。すなわち、第1部分ストリングは、「sub」、「ubs」、「bst」、「str」、「tri」、「rin」、「ing」、「subs」、「ubst」、「bstr」、「stri」、「trin」、「ring」、「subst」、「ubstr」、「bstri」、「strin」、「tring」である。

## 【0023】

そして、530段階で、ストリング選択部330は、N-gramから、索引語DB 140に保存された索引語に含まれるN-gramを選択する。すなわち、N-gramにおいて、「sub」、「ubs」、「bst」、「str」、「stri」、「strin」、「tri」、「trin」、「tring」、「rin」、「ring」、「ing」が索引語に含まれるので、ストリング選択部330は、前記のN-gramを選択することができる。

## 【0024】

さらに、550段階で、ストリング選択部330は、索引語に含まれたN-gramにおいて、他のN-gramに含まれないN-gramを有効ストリングとして決定できる。それにより、ストリング選択部330は、「sub」、「ubs」、「bst」、「strin」、「tring」を有効ストリングとして決定することができる。

## 【0025】

前記の通り、長さの異なる複数個の部分ストリングを利用して文書を検索すれば、長さが同じ部分ストリングを利用して文書を検索するより、クエリ処理速度を向上させることができる。それだけではなく、長さの異なる部分ストリングが互いに重複されなければ、重複した部分ストリングを利用した検索より、クエリ処理時間を短縮させることができる。

## 【0026】

図6は、本発明の一実施形態による候補セット決定部の細部ブロック図である。図6を参照すれば、候補セット決定部120は、有効ストリングに係わるセット(以下、「有効ストリングセット」とする)のサブセットを、サブセット選択部610及び逆索引DB 150に係わるサブセットのアクセスコストに基づいて、有効ストリングセットのサブセットのうちいずれか1つのサブセットを候補セットとして予想する候補セット予想部630を含む。

## 【0027】

本実施形態で逆索引DB 150は、索引ツリーとポスティングリストとから構成される。索引ツリーは、部分ストリング形態である索引語が、リーフノードに存在するB+ツリー構造であり、ポスティングリストは、特定索引語を含む文書の識別情報と、索引語が文書に示された位置情報とのリストである。そして、ポスティングリストを構成する各要素の文書識別情報と位置情報とを通称し、ポスティングという。

## 【0028】

まず、候補セット予想部630のアクセスコストを算出する方法について説明する。候補セット予想部630は、下記のような式(1)を利用し、アクセスコストを算出するこ

10

20

30

40

50

とができる。

【 0 0 2 9 】

【 数 1 】

$$Cost(Q) = Q_a + Q_b = \sum_{(g_i, p_i) \in Q} (h + l_i - 1) + f_B(|\Pi_{rid}(\bigcap_{(g_i, p_i) \in Q} L_i)|) \quad (1)$$

10

ここで、Qは、有効ストリングセットの特定サブセット、 $g_i$ は、Qを構成するi番目の部分ストリング、 $p_i$ は、索引ツリーでの $g_i$ の位置情報、 $h - 1$ は、索引ツリーの高さ、 $l_i$ は、 $g_i$ に係わるポスティングリストを含む索引構造でのリーフノードの個数、 $L_i$ は、 $g_i$ のポスティングリストである。

【 0 0 3 0 】

また、式(1)に適用される関数 $f_B$ は、下記式(2)の通りである。

【 0 0 3 1 】

【 数 2 】

20

$$f_B(n) = \begin{cases} |R| \cdot [1 - (1 - 1/|R|)^n], & \text{if } n < k_0, \\ B + (n - k_0)(1 - B/|R|), & \text{if } n \geq k_0 \end{cases} \quad (2)$$

ここで、

【 0 0 3 2 】

【 数 3 】

30

$|R|$

は、文書DBでの文書の個数、 $B$ は、逆索引DBを読み取るときに使われるバッファの大きさである。そして、

【 0 0 3 3 】

【 数 4 】

40

$$k_0 = \ln(1 - B/|R|) / \ln(1 - 1/|R|)$$

である。

【 0 0 3 4 】

前記のようなアクセスコストは、サブセットに含まれた有効ストリングそれぞれについて、逆索引DBで、有効ストリングのポスティングリストにアクセスして読み取るのにか

50



かるコスト（またはデータ量）の和（以下、Q a 値とする）と、文書DBで、サブセットに含まれた有効ストリングのポスティングリストに共通して含まれた文書の識別情報に対応する文書にアクセスして読み取るのにかかるコスト（以下、Q b 値とする）とに区分される。

【0035】

アクセスコストのうちQ a 値は、有効ストリングと関係なしに、固定コストであるために、常に増加する。一方、アクセスコストのうちQ b 値は、サブセットの種類によって、増減する。

【0036】

一方、候補セット決定部120は、多様な方法で候補セットを決定することができる。

10

【0037】

例えば、候補セット決定部120は、有効ストリングセットのサブセットのうち、アクセスコストが最小であるサブセットを候補セットとして決定することができる。

【0038】

図7は、本発明の第1実施形態による候補セットを決定する方法について説明するフローチャートである。図7を参照すれば、710段階で、サブセット選択部610は、有効ストリングセットのサブセットをツリー構造に配列する。ツリー構造に配列するにおいて、サブセット選択部610は、子ノードに該当するサブセットをして親ノードのサブセットに含まれるようにする。

【0039】

20

720段階で、サブセット選択部610は、ツリー構造に配列されたサブセットで、深さ優先探索（depth first search）方法で選択するサブセットがあるか否かを判断する。優先探索法というのは、ルートノードから出発し、ルートノードから可能な限り遠くにある下位ノードまで探索し、子ノードを有さないノードがあれば、バックトラッキング（backtracking）して他のノードを探索する。

【0040】

深さ優先探索で選択するサブセットがあれば、730段階で、サブセット選択部610は、前記のサブセットを選択する。サブセット選択部610が有効ストリングが一つであるサブセットをまず選択し、前記の有効ストリングを含む他のサブセットを選択する方法で選択するサブセットがあるか否かを判断する。そして、候補セット予想部630は、選択されたサブセットのアクセスコストのうちQ a 値を算出する。アクセスコストの算出方法は、前述の通りであり、具体的な説明は省略する。

30

【0041】

Q a 値が基準値以上であるならば、候補セット予想部630は、選択されたサブセットだけではなく、選択されたサブセットに対応するノードの下位ノードに含まれたサブセットのアクセスコストも算出せず、720ないし750段階を遂行する。ここで、基準値というのは、すでに選択されて算出されたサブセットのアクセスコストのうち最小値を意味する。それにより、既選択のサブセットがない場合、現在選択されたサブセットのアクセスコストが基準値になり、現在選択されたサブセットが予想候補セットになる。すなわち、深さ優先探索方法によって、初めにサブセットが選択されれば、候補セット予想部630は、740ないし770段階を遂行せずに、初めに選択されたサブセットのアクセスコストを基準値として、初めに選択されたサブセットを予想候補セットとする。

40

【0042】

一方、アクセスコストのうちQ a 値は、常に増加するために、予想候補セットのアクセスコストより、現在選択されたサブセットのQ a 値が大きければ、現在選択されたサブセットのアクセスコストは、予想候補セットのアクセスコストより常に大きい。現在選択されたサブセットを含むあらゆるサブセットのアクセスコストも、予想候補セットのアクセスコストより大きい。結局、現在選択されたサブセットだけではなく、前記のサブセットを含むサブセットについて、アクセスコストは算出しなくても差し支えなく、前記のサブセットを含むサブセットについての探索を終了し、サブセット選択部610は、バック

50

ラッキングし、他のストリングを含むサブセットがあるか否かを判断する。

【0043】

Q a 値が予想候補セットのアクセスコスト未満であるならば、760段階で、候補セット予想部630は、選択されたサブセットのアクセスコストを算出し、アクセスコストと、予想候補セットのアクセスコストとを比較する。

【0044】

選択されたサブセットのアクセスコストが、予想候補セットのアクセスコスト以下であるならば、780段階で、候補セット予想部630は、予想候補セットを、選択されたサブセットにアップデートする。

【0045】

前記の通り、候補セット決定部120は、深さ優先探索方法でサブセットを選択し、アクセスコストに基づいて、候補セットを予想する。深さ優先探索方法で選択するノードがなければ、790段階で、候補セット決定部120は、予想候補セット、すなわち、最小のアクセスコストを有するサブセットを候補セットとして決定する。

【0046】

図8は、本発明の一実施形態による、第1実施形態を介して候補セットを決定する方法について説明するための図面である。「三星総合技術院」というクエリストリングが入力されれば、ストリングセット生成部110は、索引語DB 140を基に、「三星」、「星綜」、「総合技術」及び「術院」という有効ストリングを生成する。ストリングセット生成部110は、前記の有効ストリングセットを候補セット決定部120に印加する。

【0047】

サブセット選択部610は、図8に図示されているようなツリー構造に、有効ストリングセットのサブセットを配列する。ツリー構造の各ノードには、有効ストリングセットのサブセットが配置される。

【0048】

候補セット予想部630は、上位ノードから下位ノードへの順に、各ノードに係わるアクセスコストを算出する。コスト算出方法は、前述の通りであり、具体的な説明は省略する。

【0049】

図8で、 $readCost(Q')$ は、サブセット( $Q'$ )に係わるアクセスコストのうちQ a 値を意味し、 $Cost(Q')$ は、サブセット( $Q'$ )に係わるアクセスコストを意味する。候補セット予想部630は、{三星}、{三星、星綜}、{三星、星綜、総合技術}、{三星、星綜、総合技術、術院}、{三星、星綜、術院}、{三星、総合技術}、...のような順にアクセスコストを算出する。

【0050】

一方、候補セット決定部120は、{三星、星綜}のアクセスコストのうち、Q a 値が25であり、予想候補セットのアクセスコスト、すなわち、最小アクセスコストが24であるならば、{三星、星綜}に、任意の有効ストリングが追加したサブセットも、最小アクセスコストより大きいアクセスコストを有する。従って、候補セット決定部120は、{三星、星綜}に該当するノード及び下位ノードについて、それ以上アクセスコストを算出しない。

【0051】

一方、候補セット決定部120は、有効ストリングセットのサブセットのうち、有効ストリングが一つ追加されるときアクセスコストより小さいアクセスコストを有するサブセットを候補セットを決定することもできる。

【0052】

図9は、第2実施形態による、候補セットを決定する方法について説明するフローチャートである。図9の910段階で、サブセット選択部610は、サブセットのうち、有効ストリングの個数が1であるサブセットを選択し、選択されたサブセットを候補セット予想部630に印加する。

10

20

30

40

50

## 【 0 0 5 3 】

9 2 0 段階で、候補セット予想部 6 3 0 は、選択されたサブセットそれぞれについて、アクセスコストを算出し、9 3 0 段階で、最小のアクセスコストを有するサブセットを候補セットとして予想し、前記の最小のアクセスコストを基準値とする。

## 【 0 0 5 4 】

一方、9 4 0 段階でサブセット選択部 6 1 0 は、サブセットのうち、予想された候補セット、すなわち、予想候補セットに有効ストリングが一つ追加されるサブセットが存在するならば、9 5 0 段階で、サブセット選択部 6 1 0 は、予想候補セットに、有効ストリングが一つ追加されたサブセットを選択し、選択されたサブセットを候補セット予想部 6 3 0 に印加する。

10

## 【 0 0 5 5 】

9 6 0 段階で、候補セット予想部 6 3 0 は、選択されたサブセットそれぞれについてアクセスコストを算出し、算出されたアクセスコストの最小値と基準値とを比較する(9 7 0 段階)。

## 【 0 0 5 6 】

アクセスコストの最小値が基準値以下であるならば、9 8 0 段階で、候補セット予想部 6 3 0 は、予想候補セットを最小のアクセスコストを有するサブセットにアップデートする。そして、S 9 4 0 段階ないし S 9 7 0 段階を反復的に遂行する。

## 【 0 0 5 7 】

一方、アクセスコストの最小値が基準値を超えるか、あるいは予想候補セットに有効ストリングが一つ追加されたサブセットが存在しなければ、9 9 0 段階で、候補セット予想部 6 3 0 は、予想候補セットを候補セットとして最終的に決定する。

20

## 【 0 0 5 8 】

本実施形態で、有効ストリング個数が 1 個であるサブセットから選択し、最小クエリ処理コストを有するサブセットを候補セットとして決定するとしたが、これに限定されるものではない。有効ストリング個数が多数である場合、最小予想候補セットは、有効ストリング個数が 2 またはそれ以上のサブセットから選択し、最小クエリ処理コストを有するサブセットを候補セットとして決定することもできる。

## 【 0 0 5 9 】

図 1 0 は、本発明の一実施形態による第 2 実施形態を介して候補セットを決定する方法について説明するための図面である。

30

## 【 0 0 6 0 】

クエリストリング「三星総合技術院」に係わる有効ストリングセットが、{三星、星綜、総合技術、術院}であると与えられたとすれば、まず、サブセット選択部 6 1 0 は、有効ストリングを一つ有するサブセット、すなわち{三星}、{星綜}、{総合技術}、{術院}を選択し、候補セット予想部 6 3 0 に印加する。候補セット予想部 6 3 0 は、各サブセットに係わるアクセスコストを算出する。サブセットに係わるアクセスコストが、それぞれ 1 0、2 0、2 5、3 0 であるとすれば、候補セット予想部 6 3 0 は、{三星}を候補セットとして予想する。それにより、サブセット選択部 6 1 0 は、予想候補セットである{三星}に、有効ストリングを一つ追加したサブセット、すなわち、{三星、星綜}、{三星、総合技術}、{三星、術院}を選択し、候補セット予想部 6 3 0 に印加する。候補セット予想部 6 3 0 は、選択されたサブセットそれぞれについてアクセスコストを算出する。サブセットに係わるコストがそれぞれ 1 2、1 4、8 であるならば、{三星、術院}であるサブセットのアクセスコストが予想候補セットのアクセスコストより小さいので、候補セット予想部 6 3 0 は、予想候補セットを{三星、術院}にアップデートする。

40

## 【 0 0 6 1 】

さらにサブセット選択部 6 1 0 は、{三星、術院}のサブセットに他のストリングが一つ追加されたサブセットである{三星、術院、星綜}、{三星、術院、総合技術}を選択し、候補セット予想部 6 3 0 に印加する。候補セット予想部 6 3 0 は、選択されたサブセットそれぞれについてアクセスコストを算出する。{三星、術院、星綜}、{三星、術院

50

、総合技術}のアクセスコストがそれぞれ14、15であるとするならば、予想候補セットのアクセスコストより大きいので、候補セット予想部630は、{三星、術院}を候補セットとして最終的に決定し、候補セット決定を終了する。

【0062】

候補セット決定部120が、第2実施形態による候補セット決定方法を、第1実施形態による候補セット決定方法の方よりも、候補セット決定方法として決定すれば、候補セットを決定するのにかかる時間を短縮することができる。しかし、第2実施形態による候補セット決定方法は、第1実施形態による候補セット決定方法より、正確度が多少低下する。従って、候補セット決定部120は、第1実施形態による候補セット決定方法及び第2実施形態による候補セット決定方法を選択的に使用することができる。

10

【0063】

例えば、有効ストリングの個数が基準個数以下であるならば、候補セット決定部120は、第1実施形態の候補セット決定方法で候補セットを決定することができ、有効ストリング個数が基準個数を超えれば、候補セット決定部120は、第2実施形態の候補セット決定方法で候補セットを決定することができる。

【0064】

最後に、逆索引DB 150で、候補セットの有効ストリングいずれともマッチングされている文書の識別情報を決定した後、文書検索部130は、文書DB 160で、前記の文書の識別情報とマッチングしている文書を検索することによって、クエリが存在する文書を検索する。

20

【0065】

プロセッサは、各機能によって、別途の構成要素に分離されたが、それは、説明の便宜を図るためのものであり、1つのチップまたはそれ以上のチップで具現できることは言うまでもない。

【0066】

本発明の実施形態による方法は、多様なコンピュータ手段を介して遂行されるプログラム命令形態で具現され、コンピュータで読み取り可能な媒体に記録される。前記コンピュータで読み取り可能な媒体は、プログラム命令、データファイル、データ構造などを単独で、または組み合わせて含んでもよい。前記媒体に記録されるプログラム命令は、本発明のために特別に設計されて構成されたものであってもよく、コンピュータソフトウェア当業者に公知されて使用可能なものであってもよい。

30

【0067】

コンピュータで読み取り可能な可能記録媒体の例としては、ハードディスク、フロッピー(登録商標)ディスク及び磁気テープのような磁気媒体(magnetic media)、CD-ROM、DVD(digital versatile disc)のような光記録媒体、フロプティカルディスク(floptical disk)のような磁気-光媒体(magneto-optical media)、及びROM(read-only memory)、RAM(random-access memory)、フラッシュメモリのようなプログラム命令を保存して遂行するように特別に構成されたハードウェア装置が含まれる。前記媒体は、プログラム命令、データ構造などを指定する信号を伝送する搬送波を含む光、金属線または導波管などの伝送媒体であってもよい。プログラム命令の例としては、コンパイラによって作られるような機械語コードだけではなく、インタープリタなどを使用し、コンピュータによって実行される高級言語コードを含む。前記のハードウェア装置は、本発明の動作を遂行するために、一つ以上のソフトウェア・モジュールとして作動するように構成され、その逆も同様である。

40

【0068】

以上、本発明について、たとえ限定された実施形態と図面とによって説明したが、本発明は、前記の実施形態に限定されるものではなく、本発明が属する分野で当業者であるならば、かような記載から多様な修正及び変形が可能であろう。

【0069】

従って、本発明の範囲は、説明された実施形態に限定されるものではなく、特許請求の

50

範囲だけではなくして、該特許請求の範囲と均等なものなどによって決まるものである。

【産業上の利用可能性】

【0070】

本発明の効率的にクエリを処理する方法及び装置は、例えば、検索関連の技術分野に効果的に適用可能である。

【符号の説明】

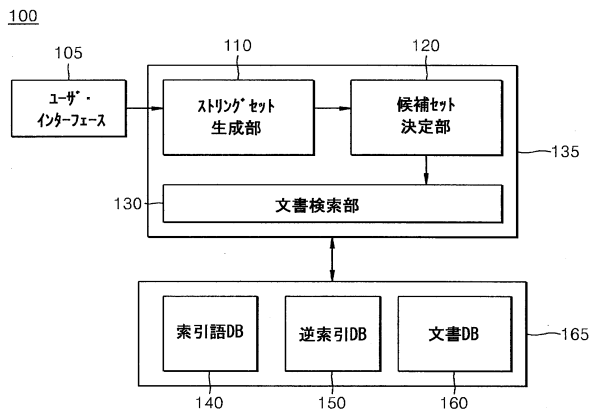
【0071】

- 100 クエリ処理装置
- 105 ユーザ・インターフェース
- 110 スtringセット生成部
- 120 候補セット決定部
- 130 文書検索部
- 135
- 140 索引語DB
- 150 逆索引DB
- 160 文書DB
- 165
- 310 N-gram分離部
- 330 スtring選択部
- 610 サブセット選択部
- 630 候補セット予想部

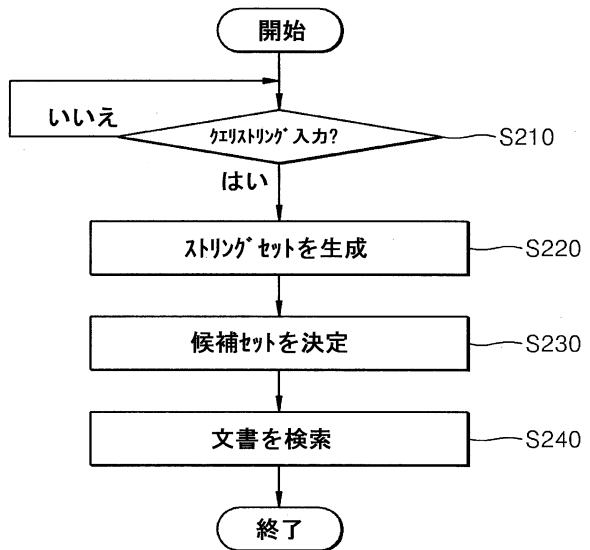
10

20

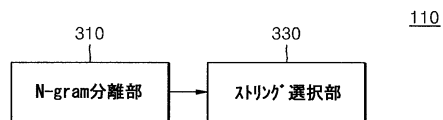
【図1】



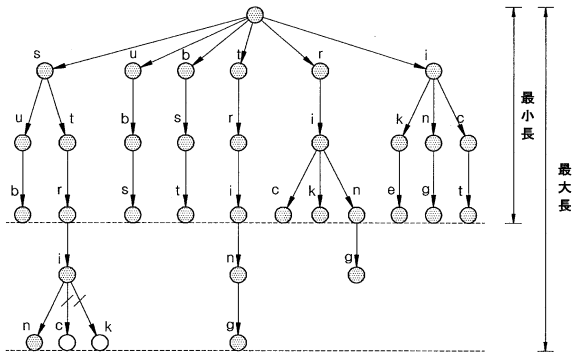
【図2】



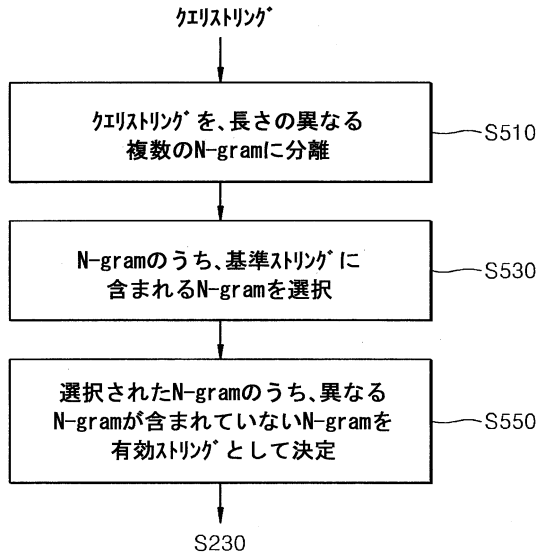
【図3】



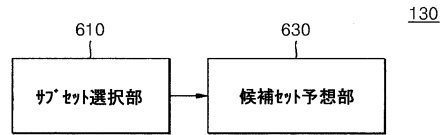
【図4】



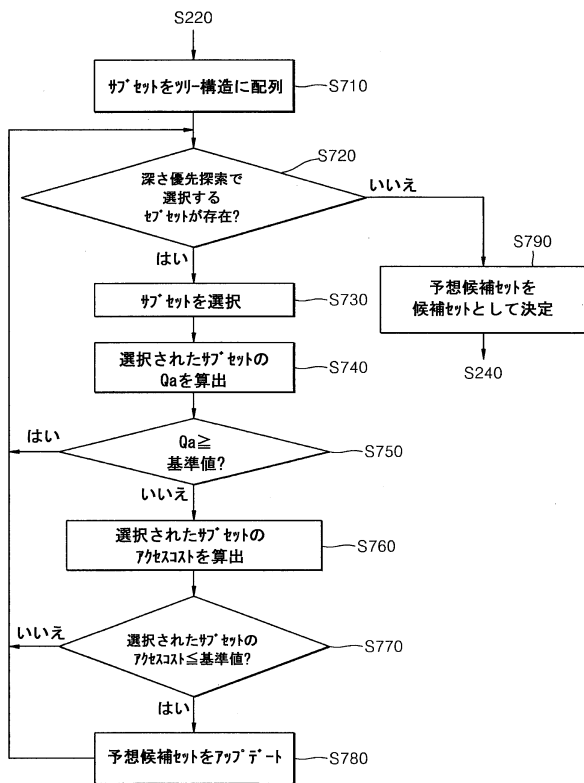
【図5】



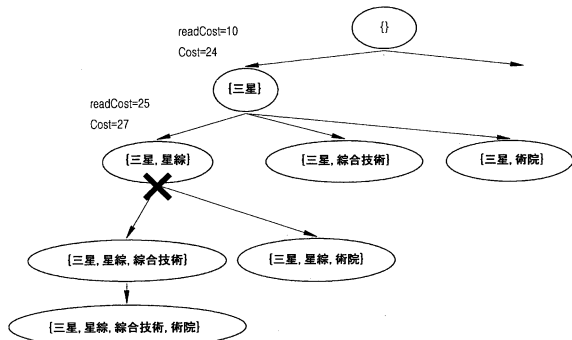
【図6】



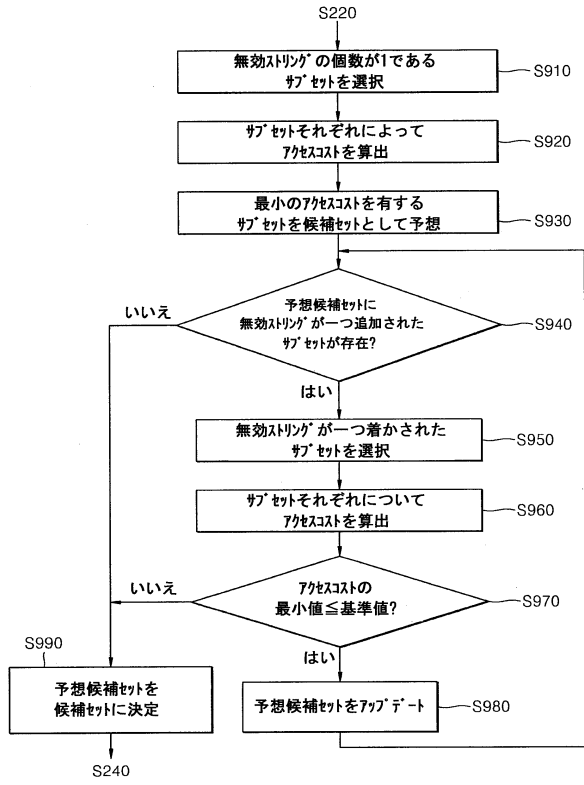
【図7】



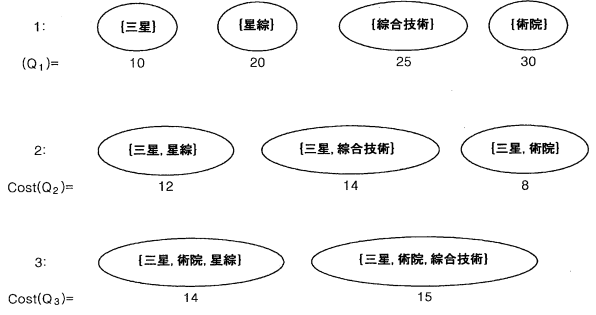
【図8】



【図9】



【図10】



## フロントページの続き

(73)特許権者 508298075

ソウル大学校産学協力団

SEOUL NATIONAL UNIVERSITY R&DB FOUNDATION

大韓民国ソウル特別市冠岳区新林洞山56-1

San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-015  
, Republic of Korea

(74)代理人 100070150

弁理士 伊東 忠彦

(74)代理人 100091214

弁理士 大貫 進介

(74)代理人 100107766

弁理士 伊東 忠重

(72)発明者 金 永 勳

大韓民国ソウル市冠岳区新林洞山56-1番地 ソウル大学校産学協力団内

(72)発明者 朴 ひょん 旻

大韓民国ソウル市冠岳区新林洞山56-1番地 ソウル大学校産学協力団内

(72)発明者 沈 揆 錫

大韓民国ソウル市冠岳区新林洞山56-1番地 ソウル大学校産学協力団内

(72)発明者 禹 景 久

大韓民国京畿道龍仁市器興区農書洞山14-1番地 三星綜合技術院内

審査官 樋口 龍弥

(56)参考文献 米国特許出願公開第2010/0241622(US, A1)

特開平08-194718(JP, A)

(58)調査した分野(Int.Cl., DB名)

G06F 17/30