(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
2010 年 4 月 22 日(22.04.2010)

PCT

(10) 国際公開番号
WO 2010/044123 A1

(51) 国際特許分類:
G06F 17/30 (2006.01)

(21) 国際出願番号: PCT/JP2008/002898

(22) 国際出願日: 2008 年 10 月 14 日(14.10.2008)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人 (米国を除く全ての指定国について): 三菱電機株式会社 (Mitsubishi Electric Corporation) [JP/JP]; 〒1008310 東京都千代田区丸の内二丁目７番３号 Tokyo (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 岡登洋平 (OKATO, Yohei) [JP/JP]; 〒1008310 東京都千代田区丸の内二丁目７番３号 三菱電機株式会社内 Tokyo (JP). 花沢利行 (HANAZAWA, Toshiyuki) [JP/JP]; 〒1008310 東京都千代田区丸の内二丁目７番３号 三菱電機株式会社内 Tokyo (JP).

(74) 代理人: 田澤英昭, 外 (TAZAWA, Hideaki et al.); 〒1000014 東京都千代田区永田町二丁目１２番４号 赤坂山王センタービル５階 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, BR, BW, BY, BZ, CA, CH, 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, KM, 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, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

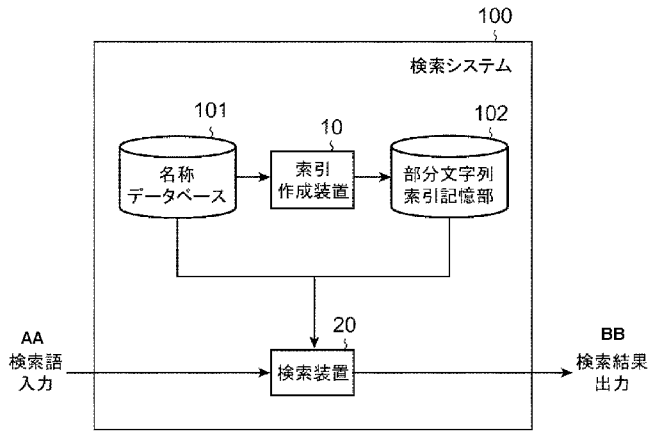
— 国際調査報告（条約第 21 条(3)）

WO 2010/044123 A1

(54) **Title**: SEARCH DEVICE, SEARCH INDEX CREATING DEVICE, AND SEARCH SYSTEM

(54) 発明の名称 : 検索装置、検索用索引作成装置、および検索システム

[図1]



100 SEARCH SYSTEM
101 NAME DATABASE
10 INDEX CREATING DEVICE
102 PARTIAL CHARACTER STRING INDEX STORAGE SECTION
20 SEARCH DEVICE
AA INPUT SEARCH WORD
BB OUTPUT RESULT OF SEARCH

(57) **Abstract**: A search device comprises: a partial character string extracting section (22) for acquiring a partial character string for search from an inputted search query; a partial character string search section (23) for acquiring candidate name text and information on the positions at which the partial character strings in the candidate name text appear on the basis of the partial character string for search; a candidate tallying section (24) for matching so that the positions at which the partial character strings of the candidate name text appear do not overlap on one another in consideration of the information on the positions at which the partial character strings appear for each candidate name text and tallying the matched score; a presentation candidate selecting section (25) for determining presentation candidates on the basis of the matched score; and a candidate presentation section (26) for presenting the presentation candidates.

(57) 要約: 入力された検索用クエリから検索用部分文字列を取得する部分文字列抽出部２２と、検索用部分文字列に基づいて候補名称テキストおよび候補名称テキスト中の部分文字列出現位置情報を取得する部分文字列検索部２３と、候補名称テキストごとに部分文字列出現位置情報を考慮して候補名称テキストの部分文字列の出現位置が重複しないように整合して照合スコアを集計する候補集計部２４と、照合スコアに基づいて提示候補を決める提示候補選択部２５と、提示候補を提示する候補提示部２６を備える。

明　細　書

検索装置、検索用索引作成装置、および検索システム

技術分野

[0001] この発明は、入力された検索語に対する文字列の検索において、特にあいまい性を含む検索語を精度良く検索可能な検索装置、検索用索引作成装置、および検索システムに関するものである。

背景技術

[0002] 従来、予め検索対象となりうる名称のＩＤと名称中の部分文字列との対応関係を記述した部分文字列をキーとした索引を作成し、この索引を参照してあいまい語検索を高速に行う方法が知られている。特許文献１に開示されたあいまい名称検索技術では、検索文字列を長さ「２」の部分文字列に分解し、部分文字列が存在する名称に対してスコアを１点ずつ加算することにより、あいまい語の検索を行っている。さらに、表記および読みを展開して検索文字列を長さ「１」の部分文字列で検索することにより、表記と読みのあいまい性を考慮した検索方法が開示されている。例えば、名称「阿蘇山」に対して、読み「あそさん」の部分文字列である「あ」，「そ」，「さ」，「ん」，「あそ」，「そさ」，「さん」，「阿」，「蘇」，「山」を検索対象に含めることであいまい性を吸収している。

[0003] また、ＯＣＲや音声認識など、あいまい性のある入力を考慮した検索方法に対して高い再現率を得るために、誤認識を考慮して可能な候補を展開することが検討されている。このとき、索引に対して想定される誤認識を展開すると索引が非常に大きくなるため、特許文献２では、音声文書内の音声認識結果の単語が正しくはどの単語の誤りとして出力されるのかを統計的に求めることによって得られる正解単語候補を用いて文書ベクトルの作成を行うことにより、音声認識単語には存在しないユーザの検索質問との類似度を上昇させ、検索の再現率を改善している。

[0004] また、特許文献３では、予め文字を形態的類似性に従ってグループ化した

類似文字群に区別しておき、文字コードを類似文字群を代表する文字に変換して類似文書を検索することにより、誤認識に対する類似判定の精度を向上させて検索の再現率を改善している。

さらに、特許文献４では、あいまい性のある箇所が１つ以上含まれているテキストに対して、あいまいな箇所を可能な候補に展開して、展開されたテキストから特徴情報を抽出し、この特徴情報を用いてあいまいな箇所の候補の組み合わせを選択している。

[0005] 特許文献1：特許第３６６５１１２号

特許文献2：特開２００４－３４８５５２号公報

特許文献3：特開２００７－４８０６１号公報

特許文献4：特開２００７－５８４１５号公報

[0006] 従来のあいまい性を含む名称の検索は以上のように構成されているので、特許文献１では読みを展開した場合の排他性が考慮されない。例えば、「山さん」という入力に対して「阿蘇山」および「あそさん」を見出しに持つ名称の一致度が１００％となる。この検索結果はユーザの違和感が大きく、これらの候補の追加によって、検索結果として提示する候補の妥当性が低下するという課題があった。展開した名称を別に追加すればこの課題は回避可能であるが、その場合登録名称数の増加に比例して索引のサイズが拡大するという課題がある。

[0007] 特に、検索語の入力が音声認識結果である場合、長音化・濁音化・清音化など発音に基づく発声の揺らぎによって、読みを付与する場合にあいまい性が生じる。長音化は、二重母音（／ｏｕ／，／ｅｉ／）が特定の文脈において先行母音の連続（／ｏｏ／，／ｅｅ／）のように発音され易い性質である。例えば、「東京」は、読み「トウキョウ」よりも「トーキョー」に近く発声される。この長音化は、音素配列だけではなく言語的な文脈により生じないケースもある。例えば、「京都魚市場」の読み「キョウトウオイチバ」の場合、「キョウ」は「キョー」に長音化される場合がある一方で、「トウ」は「トー」のように長音化されない。

[0008] 濁音化および清音化も同様に文脈に応じて濁音が濁らない清音になったり、清音が濁る濁音になる。例えば、「研究所」の読み「ケンキュウジョ」は「ケンキュウショ」のように発声されるケースがある。

これらの名称を複数に展開して索引を作成する場合、一般に索引サイズが展開して追加した変形の名称数に比例するため数倍以上のサイズとなってしまう。

[0009] また、特許文献２では、統計的に求められる正解単語候補を用いて文書ベクトルの作成を行うため、当該文書ベクトル作成の処理時間が必要になるという課題があった。特許文献３では、予め文字を形態的類似性に従ってグループ化することにより、例えば「トウ」と「トオ」を区別せずにまとめて扱うため、索引サイズは増加しないものの、前述したように文脈により区別可能な表現が集約されるため検索精度が低下するという課題があった。一方、特許文献４に示されるように、入力されたテキストに対してあいまいな箇所を複数通りの候補に展開する場合、入力テキストの個数に比例した処理時間が必要になるという課題があった。

[0010] この発明は、上記のような課題を解決するためになされたもので、索引サイズの拡大および検索時の演算量を抑制すると共に、あいまい性を考慮した検索において検索精度を向上させることを目的とする。

発明の開示

[0011] この発明に係る検索装置は、検索用クエリを取得する入力部と、前記検索用クエリから検索用部分文字列を取得する部分文字列抽出部と、前記検索用部分文字列に基づいて候補名称テキストおよび前記候補名称テキスト中の部分文字列出現位置情報を取得する部分文字列検索部と、前記候補名称テキストごとに前記部分文字列出現位置情報を考慮して前記候補名称テキストの部分文字列の出現位置が重複しないように整合して照合スコアを集計する候補集計部と、前記照合スコアに基づいて提示候補を決める提示候補選択部と、前記提示候補を提示する候補提示部とを備えるものである。

[0012] この発明によれば、候補名称テキストごとに部分文字列出現位置情報を考

慮して候補名称テキストの部分文字列の出現位置が重複しないように整合して照合スコアを集計する候補集計部と、照合スコアに基づいて提示候補を決める提示候補選択部と、提示候補を提示する候補提示部とを備えるように構成したので、検索語のあいまい性を考慮した検索において検索精度を向上させることができる。また、部分文字列索引のサイズの拡大および検索時の演算量を抑制することができる。

図面の簡単な説明

[0013] [図1]実施の形態１に係る検索システムの構成を示すブロック図である。

[図2]実施の形態１に係る名称データベースの一例を示す図である。

[図3]実施の形態１に係る索引作成装置の構成を示すブロック図である。

[図4]実施の形態１に係る言語解析用辞書が有する単語情報の一例を示す図である。

[図5]実施の形態１に係る言語解析用辞書が有する言語規則の一例を示す図である。

[図6]実施の形態１に係る名称展開部が生成する有向グラフの一例を示す図である。

[図7]実施の形態１に係る部分文字列抽出部が抽出する部分文字列情報の一例を示す図である。

[図8]実施の形態１に係る部分文字列索引の一例を示す図である。

[図9]実施の形態１に係る検索装置の構成を示すブロック図である。

[図10]実施の形態１に係る検索装置の動作を示すフローチャートである。

[図11]実施の形態１に係る名称展開部による同義語の展開例を示す図である。

[図12]実施の形態２に係る検索装置の構成を示すブロック図である。

[図13]実施の形態２に係る名称展開部が生成する有向グラフの一例を示す図である。

[図14]実施の形態２に係る部分文字列抽出部が抽出する部分文字列情報の一例を示す図である。

[図15]実施の形態２に係る検索装置の動作を示すフローチャートである。

発明を実施するための最良の形態

[0014] 以下、この発明をより詳細に説明するために、この発明を実施するための最良の形態について、添付の図面に従って説明する。

実施の形態１.

図１は、この発明の実施の形態１に係る検索システムの構成を示すブロック図である。

検索システム１００は、索引作成装置（検索用索引作成装置）１０、検索装置２０、名称データベース１０１、部分文字列索引記憶部１０２で構成されている。

索引作成装置１０は、名称データベース１０１に記憶されている検索対象となりうる名称テキストに基づき事前に部分文字列索引を作成する。検索装置２０は、入力される検索語に応じて部分文字列索引記憶部１０２に記憶されている部分文字索引を用いて検索結果候補を演算し、出力する。

[0015] 名称データベース１０１は、検索対象となりうる名称テキストに関する情報が登録されている。登録情報は、それぞれの名称テキストに関する認識可能な名称ＩＤおよび名称の文字列を表す見出しで構成されている。さらに、この見出しに対応する漢字、アルファベット、数字あるいは記号などを含む表記が含まれていてもよい。図２は、名称データベース１０１の登録情報の一例を示す図である。部分文字列索引記憶部１０２は、索引作成装置１０が作成した部分文字列索引を記憶する。

[0016] 図３は、この発明の実施の形態１に係る索引作成装置の構成を示すブロック図である。

索引作成装置１０は、言語解析用辞書１１、名称展開部１２、部分文字列抽出部１３および部分文字列ソート部１４で構成されている。言語解析用辞書１１は、言語解析を行い見出しの変形を抽出する際に用い、単語情報と単語を結合するための言語規則を有している。図４は言語解析用辞書に登録されている単語情報の一例示し、図５は言語規則の一例を示している。

[0017] 図４に示すように単語情報として、名称データベース１０１から取得可能な見出し、この見出しに対応する表記、品詞などの言語情報、および表記ゆれを示す変形パタンが登録されている。単語とは、読みと表記の少なくとも一方が１文字以上含まれていればよく、言語学的な意味に制約されるものではない。また、変形パタンを構成する読みの長さは、元の見出しの読みの長さと同一とする。また、図５に示すように言語規則として、解析のために必要な情報である品詞や単語を結合するための知識（先行品詞および後続品詞などで示す接続可能性やペナルティ）が登録されている。

[0018] 名称展開部１２は、名称データベース１０１から名称テキストを１つ読み込み、言語解析用辞書１１を参照して、読みの先頭位置と整列した場合の位置情報（出現位置情報）を表す各ノードと、この各ノードの接続関係を示すアークで構成される有向グラフで示される待ち受け表現（表現グラフ）を生成する。図６は、名称展開部が生成する有向グラフの一例を示している。図６の例では、図２で示した名称データベース１０１の名称ＩＤ：０００２の見出し「キョウトウドン」に対して、変形パタン「キョオ」を適用し、長音を複数通りに展開した有向グラフを示している。有向グラフのノード構成単位は、１文字に相当する音節とする。また、ここでは長音は母音で表し、拗音「ァィゥェォャュョ」や促音「ッ」は単独で発音されないため、前の文字とまとめて１単位とする。

[0019] 部分文字列抽出部１３は、名称展開部１２から入力される待ち受け表現の有向グラフから部分文字列を抽出すると共に、その部分文字列に対応する位置情報を付与した部分文字列情報を生成する。図７は、部分文字列抽出部が生成する部分文字列情報の一例を示している。図７の例では、部分文字列を２音節に固定して１音節ずつずらして取得した見出しと、その見出しに対応する名称ＩＤおよび位置情報を対応付けている。部分文字列の取得音節単位は、検索装置に好適な条件で設定可能である。

[0020] 部分文字列ソート部１４は、部分文字列抽出部１３から入力される部分文字列情報に基づき、名称ＩＤおよび位置情報のリストをソートする。さらに

、部分文字列の見出しと、その見出しに対応する名称ＩＤおよび位置情報からなるリストを作成し、部分文字列索引として部分文字列索引記憶部１０２に出力する。図８は、部分文字列ソート部が作成する部分文字列索引の一例を示している。図８の例では、アイウエオ順にソートされた部分文字列の見出しと、この見出しに対応する名称ＩＤ・位置情報リストの組み合わせで構成される部分文字列索引を示している。

上述のようにして事前に作成した部分文字列索引を参照して検索を行うことにより、名称データベースそのものを走査する場合と比べてはるかに短時間で検索結果に合致する候補名称を取得することができる。

[0021] 次に、索引作成装置１０により作成された部分文字列索引を参照して検索語（検索クエリ）の検索を行う検索装置２０について説明する。図９は、この発明の実施の形態１に係る検索装置の構成を示すブロック図である。検索装置２０は、入力部２１、部分文字列抽出部２２、部分文字列検索部２３、候補集計部２４、提示候補選択部２５および候補提示部２６で構成されている。

[0022] 入力部２１は、ユーザからの検索クエリの入力を受け付ける。部分文字列抽出部２２は、入力された検索クエリから検索用部分文字列を抽出する。部分文字列検索部２３は、部分文字列索引記憶部１０２の部分文字列索引を参照し、部分文字列抽出部２２において抽出された検索用部分文字列に対応する候補名称テキストの部分文字列に関する名称ＩＤ・位置情報リストを取得する。

[0023] 候補集計部２４は、名称ＩＤ毎の累積スコア（照合スコア）および参照した位置情報を格納する集計用メモリ２４ａを有している。部分文字列検索部２３から入力される名称ＩＤ・位置情報リストから候補名称テキストの部分文字列の名称ＩＤと位置情報を読み出し、該位置情報と検索用部分文字列の位置情報とに基づき部分文字列の出現位置が重複しないように整合して集計用メモリ２４ａの累積スコアを更新する。提示候補選択部２５は、部分文字列の累積スコアと位置情報に基づき最終スコアを算出し、この最終スコアを

ソートして検索結果として提示する上位候補を決定する。さらに、この上位候補の名称ＩＤに該当する名称テキストを名称データベース１０１から読み出し、検索結果名称テキストとして出力する。候補提示部２６は、提示候補選択部２５から入力される検索結果名称テキストをユーザに提示する。

[0024] 次に、この発明の実施の形態１に係る検索装置の動作について説明する。図１０は、実施の形態１に係る検索装置の検索処理動作を示すフローチャートである。

候補集計部２４は、集計用メモリ２４ａを初期化する（ステップＳＴ１）。入力部２１は、ユーザにより入力された検索クエリを読み込み、部分文字列抽出部２２に出力する（ステップＳＴ２）。部分文字列抽出部２２は、ステップＳＴ２において入力された検索クエリから検索用部分文字列ｓ［ｉ］を抽出し、部分文字列検索部２３に出力する（ステップＳＴ３）。なお、ここではM個の検索用部分文字列ｓ［１］，ｓ［２］，・・・，ｓ［M］を抽出するものとする。また、部分文字列の初期値は「１」として、部分文字列抽出開始時にｉ＝１として初期化する。

[0025] 部分文字列検索部２３は、部分文字列索引記憶部１０２の部分文字列索引を参照して、ステップＳＴ３において入力された検索用部分文字列ｓ［ｉ］に対応する候補名称テキストの部分文字列に関する名称ＩＤ・位置情報リスト（ｉｄ［ｊ］，ｏｆｓ［ｊ］）を取得し、候補集計部２４に出力する（ステップＳＴ４）。なお、長さＮの名称ＩＤ・位置情報リストは（ｉｄ［１］，ｏｆｓ［１］），（ｉｄ［２］，ｏｆｓ［２］），・・・，（ｉｄ［Ｎ］，ｏｆｓ［Ｎ］）と表し、ｉｄ［ｊ］はｊ番目の候補名称テキストの名称ＩＤ、ｏｆｓ［ｊ］はｊ番目の候補名称テキスト中の部分文字列の出現位置を表している。また、リスト長さの初期値は「１」として、部分文字列検索開始時にｊ＝１として初期化する。

[0026] 候補集計部２４は、集計用メモリ２４ａを参照し、ステップＳＴ４において入力された候補名称テキストの部分文字列の名称ＩＤおよび位置情報が累積スコアに加算済みであるか否か判定する（ステップＳＴ５）。ステップＳ

Ｔ５において、累積スコアに未加算であると判定された場合には、ｉｄ［ｊ］の累積スコアを「１」加算し、重複加算防止のために集計メモリのｉｄ［ｊ］についてｏｆｓ［ｊ］が加算済であるフラグをセットする（ステップＳＴ６）。一方、ステップＳＴ５において、累積スコアに加算済であると判定された場合にはステップＳＴ７の処理に進む。

[0027] 候補集計部２４は、名称ＩＤ・位置情報リストの「ｊ」に１を加算し（ステップＳＴ７）、ｊがＮ以下であるか否か判定する（ステップＳＴ８）。ステップＳＴ８において、ｊがＮ以下であると判定された場合にはステップＳＴ５に戻り、次の名称ＩＤ・位置情報リスト項目（ｊ＋１した項目）に対して上述の処理を繰り返す。一方、ステップＳＴ８において、ｊがＮ以下でないと判定され、全ての名称ＩＤ・位置情報リスト項目の処理が終了した場合には、部分文字列の「ｉ」にも１加算し（ステップＳＴ９）、ｉがＭ以下であるか否か判定する（ステップＳＴ１０）。ステップＳＴ１０において、ｉがＭ以下であると判定された場合にはステップＳＴ４に戻り、次の部分文字列（ｉ＋１した項目）に対して上述の処理を繰り返す。

[0028] 一方、ステップＳＴ１０において、ｉがＭ以下出ないと判定され、全ての部分文字列の処理が終了した場合には、提示候補選択部２５が名称ＩＤ毎に累積スコアをソートし、ユーザに提示する上位候補を抽出すると共に、名称データベース１０１を参照して抽出した上位候補の名称ＩＤに該当する名称テキストを読み出して候補提示部２６に出力する（ステップＳＴ１１）。このとき、名称の長さ、入力の長さ、部分照合のパタンなどを考慮してスコアを正規化してもよい。候補提示部２６は、ステップＳＴ１１において入力された検索結果である名称テキストをユーザに提示する（ステップＳＴ１２）。

[0029] 図１０のフローチャートに基づく検索処理を実行することにより、図７に示した部分文字列情報の例では「キョウトウドン」および「キョオトウドン」という２通りの表現を受理するものの、部分文字列索引記憶部１０２のサイズは５項目から７項目への２項目の増加に留まり、検索処理の高速化を可

能にしている。

また、検索処理時に位置情報の重複判定に基づいて累積スコアを集計するので、索引作成処理時に複数通りの部分文字列へ展開した場合にもそれらを重複して集計することがなく、検索精度を向上させることができる。具体的には、図７で展開された見出しに対して入力「キョウキョオ」は、「キョウ」または「キョオ」の一方を加算した時点でｏｆｓ［１］へフラグがセットされるので、２度の重複カウントを避けることができる。

[0030] 次に、対応付けのあいまい性に対する処理について説明を行う。図１０のフローチャートにおける候補集計部２４のステップＳＴ５の処理において、部分文字列索引記憶部１０２を構成する名称テキストの部分文字列と検索用部分文字列との対応付けにあいまい性が生じる場合がある。

具体的には、検索用部分文字列が名称テキストの部分文字列中の複数の位置と対応付け可能な場合（条件Ａ）、複数の検索用部分文字列が名称テキスト中の部分文字列中の一つの位置のみと対応付け可能な場合（条件Ｂ）に対応付けにあいまい性が生じる。

[0031] まず、条件Ａの検索用部分文字列と名称テキストの部分文字列の対応付けに関して説明する。検索用部分文字列の出現回数が、名称テキストの部分文字列中の出現回数と同一あるいは多い場合、名称テキストの部分文字列の全ての位置情報と対応付けを行う。

一方、検索用部分文字列の出現回数が、名称テキストの部分文字列の出現回数よりも少ない場合、対応付けのあいまい性が生じる。例えば、名称テキスト「ホオホオ」という文字列には長さが「２」の部分文字列「ホオ」が２回出現する。検索クエリが「ホオ」の場合、集計処理の過程ではどちらの部分文字列と対応付けすべきかあいまいとなる。

[0032] 次に、条件Ｂの複数の検索用部分文字列が名称テキストの部分文字列中の一つの位置のみと対応付けされる場合、具体例としては長音化前後の表現が両方とも検索クエリに出現する場合、名称「ホウ」について長音化した「ホオ」が同一位置で索引に登録されており、検索クエリが「ホウホオ」の場合

、対応付けがあいまいとなる。

[0033] 上述した条件Ａおよび条件Ｂのあいまい性が生じる場合、候補集計部２４は規則により優先順位を決めて対応付けする（方法１）、可能性のある組み合わせについて照合候補を展開する（方法２）、照合履歴に基づいて対応付けを決める（方法３）の周知の方法を用いて対応付け可能である。また、これらの方法を組み合わせることも可能である。

[0034] まず、方法１では、あいまい性が生じた場合の適応順序を規則として予め決定しておく。例えば、条件Ａにおいて同一名称内で複数回部分文字列が出現した場合、先頭から順に対応付けを行うように決定しておく。また、条件Ｂにおいて集計を行う部分文字列の順番を予め決定しておく。見出し部分文字列の展開内容が長音化の場合、長音から非長音への一方向の変換であるため、長音でない部分文字列を先に対応付けることにより対応数の集計誤りを防ぐことができる。

[0035] 方法２では、条件Ａのあいまい性が生じた場合、該当する名称ＩＤの累積スコアおよび参照した位置情報が格納された集計用メモリ２４ａをコピーし、複数の対応付けそれぞれの累積スコアを算出する。最終的に各名称ＩＤについて最大の累積スコアとなる対応付けを採用する。

[0036] 方法３では、名称ＩＤ毎に直前に加算された位置情報を集計用メモリ２４ａに保持して条件Ａのあいまい性を解消する。名称ＩＤ毎の位置情報は初期値を０とする。部分文字列索引記憶部１０２の部分文字列索引において当該名称ＩＤについて複数の位置情報候補が含まれる場合「集計用メモリ２４ａに保持された位置情報＋１」から最も近い位置を対応付け結果とする。これにより、連続性のある位置情報を優先する対応付けを行うことができる。

[0037] 以上のように、この実施の形態１によれば、検索処理時に位置情報の重複判定を行い累積スコアを集計する候補集計部２４を設けるように構成したので、複数通りの部分文字列に展開された索引を用いた場合にも、当該複数通りの部分文字列を重複して集計することがなく、検索精度を向上させることができる。

[0038] また、この実施の形態１によれば、検索語と索引の部分文字列の対応付けにあいまい性が生じる場合に、候補集計部２４が規則により優先順位を決めて対応付けする（方法１）、可能性のある組み合わせについて照合候補を展開する（方法２）、照合履歴に基づいて対応付けを決める（方法３）などの方法を用いて照合関係を求めるように構成したので、より検索精度を向上させることができる。

[0039] また、この実施の形態１によれば、元の名称データベース１０１内に出現したオリジナル表現である検索語の名称読みに変形が想定される場合に、当該変形可能名称読みに同一の位置情報を付して複数の経路へ展開した有向グラフを作成する名称展開部１２を設けるように構成したので、部分文字列索引のサイズの増加を抑制し、検索処理の高速化を実現することができる。

[0040] また、この実施の形態１によれば、変形可能な名称読みが想定される文字列を展開した部分文字列索引を参照して、検索語の検索を行うように構成したので、名称データベースそのものを走査する場合と比較して短時間で検索結果に合致する名称テキストを取得することができる。

[0041] なお、上記実施の形態１では、部分文字列を２音節として説明を行ったが、形態素を単位として処理してもよい。この場合、発音の揺れだけではなく、同義語表現の重複も吸収可能である。図１１は、この形態素を単位とした場合の同義語の展開例を示す図である。「トウキョウ／カントリー／クラブ」、「トウキョウ／ゴルフ／クラブ」の２通りについて重複を考慮して索引作成および検索処理が可能になる。

[0042] 実施の形態２.

図１２は、この発明の実施の形態２に係る検索装置の構成を示すブロック図である。実施の形態２に係る検索装置は、実施の形態１の検索装置に入力方法識別部を追加して設けている。以下、実施の形態１と同一の構成には図９で使用した符号と同一の符号を付し、説明を省略または簡略化する。

[0043] 入力方法識別部３１は、入力部２１への検索クエリの入力が音声であり部分文字列検索部２３に音声認識結果が入力されるか、あるいは入力がキーボ

ードなどであり部分文字列検索部２３に検索クエリの読みがテキストのまま直接入力されるかを識別し、識別結果を部分文字列検索部２３に出力する。

検索クエリが音声入力であるか、テキスト入力であるかを識別することにより、検索クエリに対する読みの長音化の展開処理が必要であるか、不要であるか判断することが可能となる。テキスト入力である場合には、検索クエリの読みがテキストとして直接入力されているため、読みの長音化の展開処理が不要となる。これに合わせて、部分文字列索引記憶部１０２の部分文字列索引において、認識入力のために追加した見出しを区別しておき、検索クエリの入力方式に応じて検索表現を切替可能に構成している。

[0044] 図１３は、この発明の実施の形態２に係る名称展開部が生成する有向グラフの一例を示す図である。実施の形態２に係る名称展開部１２は、名称「キョウトウドン」について、音節を単位とした名称の読みとその長音化を展開した有向グラフを生成する。長音化した部分については「オ＊」と記載して展開して生成した結果であることを明記しておく。

[0045] 図１４は、図１３の有向グラフに基づき部分文字列抽出部が生成する部分文字列情報の一例を示している。図１３の展開結果を参照して名称「キョウトウドン」を文字列長さ「２」の部分文字列に分解した見出し、名称ＩＤ（図１４においては０００２とする）その見出しが出現する位置情報を示す。ただし、展開結果であることを示す記号「＊」はそのまま付与しておく。これにより、部分文字列索引記憶部１０２の索引において、同一の読みであっても名称の読みによって生成された見出しと、名称の読みの長音化により生成された見出しとを区別することができる。

[0046] 次に、この発明の実施の形態２に係る検索装置の動作について説明する。図１５は、実施の形態２に係る検索装置の検索処理動作を示すフローチャートであり、以下このフローチャートに従って説明する。実施の形態１に係る検索装置と同一の処理を行うステップには図１０で使用した符号と同一の符号を付し、説明を省略する。

[0047] ステップＳＴ１において集計用メモリが初期化されると、入力方法識別部

３１が検索クエリの入力が音声入力であるか、テキスト入力であるかを識別して識別結果を部分文字列検索部２３に出力すると共に、入力部２１がユーザにより入力された検索クエリを読み込み、部分文字列抽出部２２に出力する（ステップＳＴ２１）。

[0048] 部分文字列抽出部２２は、ステップＳＴ２において入力された検索クエリから検索用部分文字列ｓ［ｉ］を抽出し、部分文字列検索部２３に出力する（ステップＳＴ３）。なお、ここではM個の検索用部分文字列ｓ［１］，ｓ［２］，・・・，ｓ［M］を抽出するものとする。また、部分文字列の初期値は「１」として、部分文字列抽出開始時にｉ＝１として初期化する。

[0049] 部分文字列検索部２３は、ステップＳＴ３において部分文字列抽出部２２から入力された検索用部分文字列ｓ［ｉ］、およびステップＳＴ２１において入力方法識別３１から入力された識別結果である検索クリエの入力方法に対応する候補名称テキストの部分文字列に関する名称ＩＤ・位置情報リスト（ｉｄ［ｊ］，ｏｆｓ［ｊ］）を取得し、候補集計部２４に出力する（ステップＳＴ２２）。なお、ここで索引リストの長さは「Ｎ」とする。また、部分文字列検索開始時にｊ＝１として初期化する。

ステップＳＴ２２において、検索クエリが音声入力である場合、読みの展開結果である見出し（図１４における「キョオ」→「キョオ＊」）を追加して参照する。検索クエリがテキスト入力である場合、展開結果を反映せずに部分文字列通りの見出しのみを参照する。

[0050] 候補集計部２４は、集計用メモリ２４ａを参照し、ステップＳＴ２２において入力された候補名称テキストの部分文字列の名称ＩＤおよび位置情報が累積スコアに加算済であるか否か判定する（ステップＳＴ５）。その後、実施の形態１において説明したステップＳＴ６からステップＳＴ１２までと同様の処理を行い、検索結果を出力する。

[0051] 以上のように、この実施の形態２によれば、検索語の入力方法を識別する入力方法識別部３１を設け、索引作成装置１０が索引作成時に識別子を付して入力方法を区別可能な索引を作成し、部分文字列検索部２３が入力方法識

別部３１において識別された入力方法に応じて参照する部分文字列の見出しを展開するように構成したので、展開による見出しの増加を除くと部分文字列索引の記載内容は見出しを展開して作成した場合と同等であり、入力方法に応じて２つの部分文字列索引を作成する場合と比べて部分文字列索引の総ファイルサイズを縮小することができる。

[0052] また、この実施の形態２によれば、元の名称データベース１０１内に出現したオリジナル表現である検索語の名称読みと、部分文字列索引作成時に付与した追加表現である展開結果を区別するように構成したので、検索時にまず検索語の名称読みの部分文字列索引と照合し、次に展開結果の部分文字列索引と照合することにより、オリジナル表現である検索語の名称読みとの対応付けを優先して照合することができる。

産業上の利用可能性

[0053] 以上のように、この発明はあいまい性を有する検索語入力に対して精度の高い検索結果を表示する検索装置、検索の際に参照する索引のファイルサイズを縮小可能とする検索用索引作成装置およびこれらを有する検索システムに幅広く適用することができる。

請求の範囲

[1] 検索用クエリを取得する入力部と、

前記検索用クエリから検索用部分文字列を取得する部分文字列抽出部と、

前記検索用部分文字列に基づいて候補名称テキストおよび前記候補名称テキスト中の部分文字列出現位置情報を取得する部分文字列検索部と、

前記候補名称テキストごとに前記部分文字列出現位置情報を考慮して前記候補名称テキストの部分文字列の出現位置が重複しないように整合して照合スコアを集計する候補集計部と、

前記照合スコアに基づいて提示候補を決める提示候補選択部と、

前記提示候補を提示する候補提示部とを備えることを特徴とする検索装置。

[2] 検索用クエリの入力方法を識別する入力方法識別部を設け、

部分文字列検索部は、識別された入力方法および検索用部分文字列に基づいて候補名称テキストおよび前記候補名称テキスト中の部分文字列出現位置情報を取得することを特徴とする請求項１記載の検索装置。

[3] 候補集計部は、検索用クエリと候補名称テキストの部分文字列の位置対応付けにあいまい性が存在する場合に、事前に決めた照合順序で照合する、候補ごと別の照合候補を生成する、照合履歴に基づいて照合関係を求めることの少なくとも一つの方法を用いることを特徴とする請求項１記載の検索装置。

[4] 候補集計部は、検索用クエリと候補名称テキストの部分文字列の位置対応付けにあいまい性が存在する場合に、事前に決めた照合順序で照合する、候補ごと別の照合候補を生成する、照合履歴に基づいて照合関係を求めることの少なくとも一つの方法を用いることを特徴とする請求項２記載の検索装置。

[5] 名称テキストを解析し、入力の名称変形が想定される場合に同一位置情報を付した複数の経路へ展開した入力表現グラフを生成する名称展開部と、

展開した名称テキストから部分文字列と出現位置情報を取得する部分文字

列抽出部と、

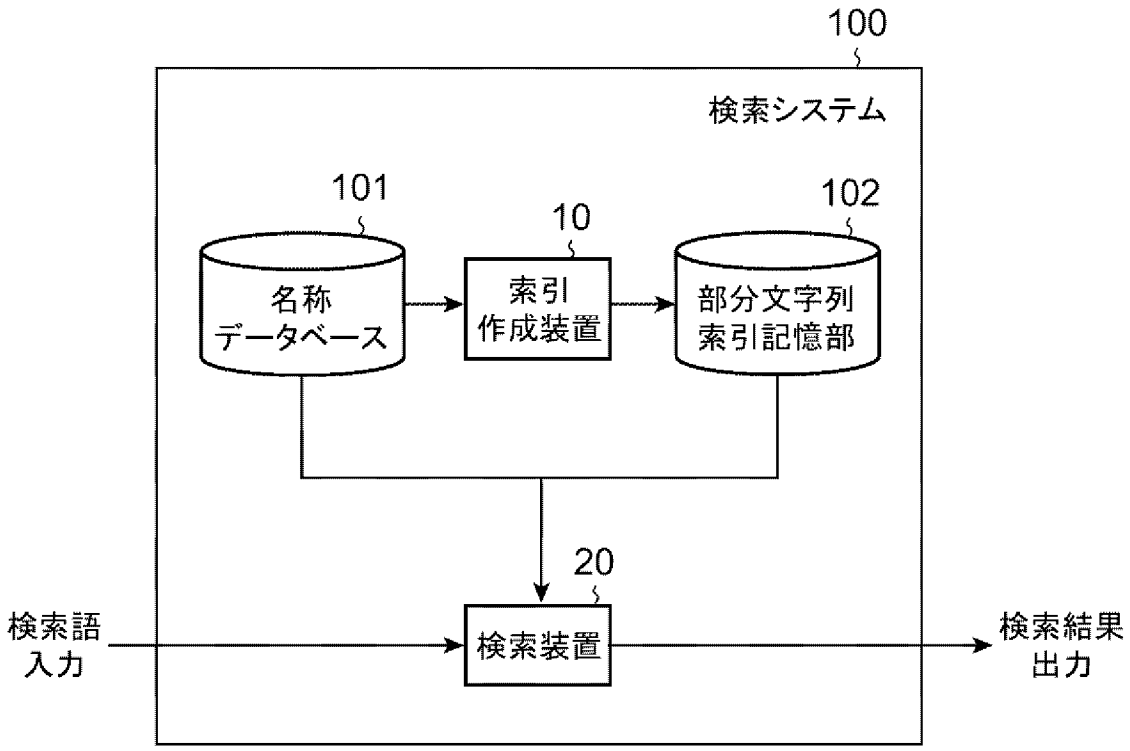
前記部分文字列、前記名称テキストおよび前記出現位置情報をソートして名称テキスト検索のための部分文字列索引を生成する部分文字列ソート部からなる検索用索引作成装置。

[6] 名称展開部は、入力の名称変形がある場合は名称変形であることを付した符号を付与することを特徴とする請求項５記載の検索用索引作成装置。

[7] 請求項１記載の検索装置および請求項５記載の検索用索引作成装置を備え、

名称テキストを蓄積する名称データベースおよび前記検索用索引作成装置において作成された部分文字列索引を蓄積する部分文字列索引記憶部を備えたことを特徴とする検索システム。

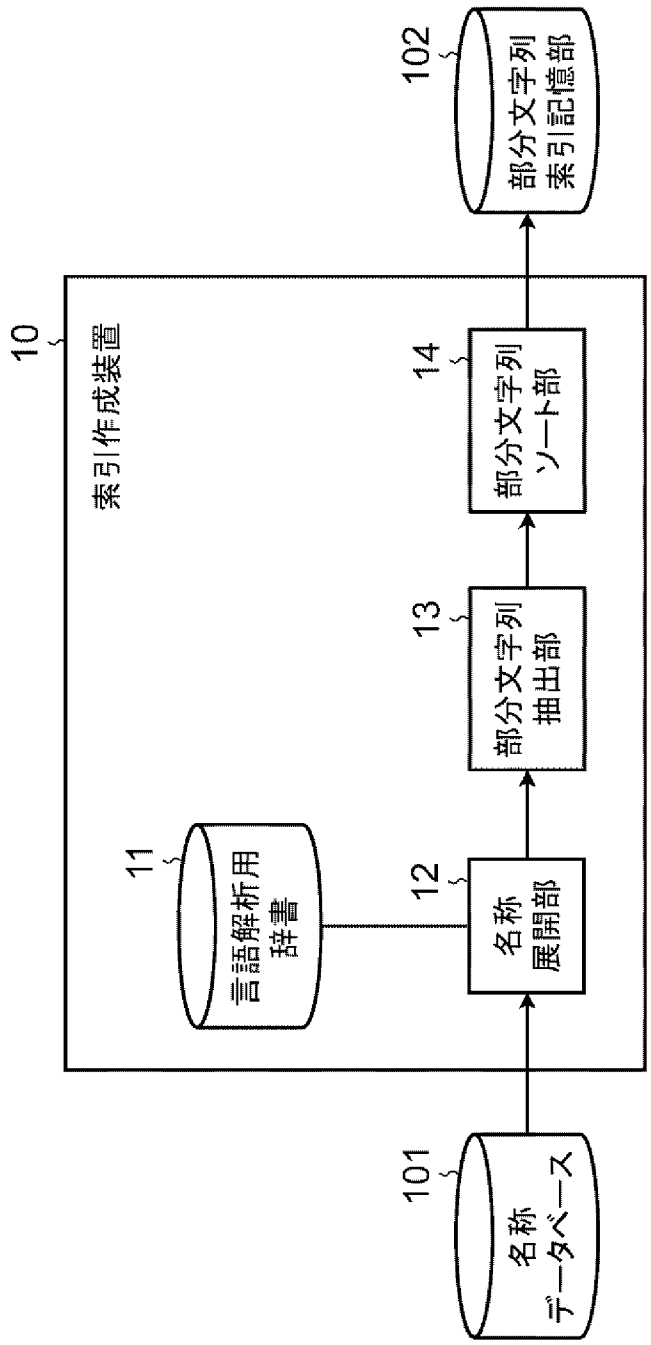
[図1]



[図2]

名称ID	見出し(検索キー)	表記
0001	トウキョウタワー	東京・タワー
0002	キョウトウドン	京都うどん
0002	ギンザインターチェンジ	銀座インターチェンジ
0003	コクリツケンキュウジョ	国立研究所

[図3]


10
索引作成装置
11
言語解析用
辞書
12
名称
展開部
13
部分文字列
抽出部
14
部分文字列
ソート部
101
名称
データベース
102
部分文字列
索引記憶部

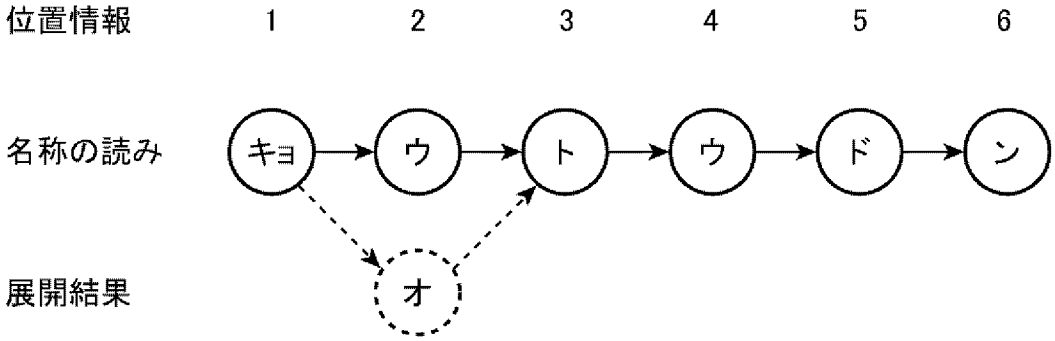
[図4]

見出し	表記	言語情報(品詞)	変形パタン
トウ	東	文字(漢字)	トオ
キョウ	京	文字(漢字)	キョオ
ト	都	文字(漢字)	
うどん	うどん	形態素(名詞)	

[図5]

先行品詞	後続品詞	ペナルティ
文字(漢字)	文字(漢字)	0
文字(漢字)	記号	1
記号	形態素(名詞)	0
形態素(名詞)	形態素(名詞)	0

[図6]



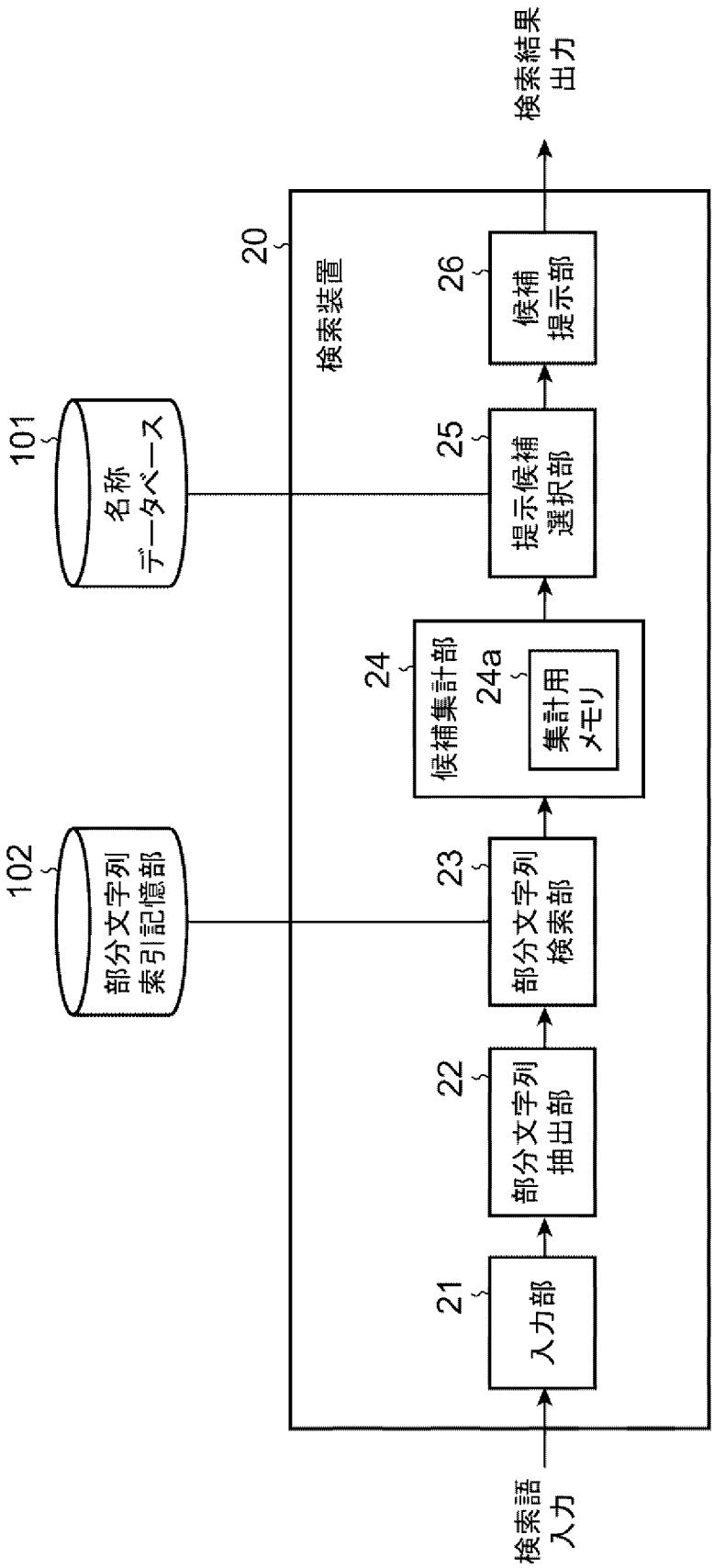
[図7]

見出し	名称ID	位置情報
キョウ	0002	1
キョオ	0002	1
ウト	0002	2
オト	0002	2
トウ	0002	3
ウド	0002	4
ドン	0002	5

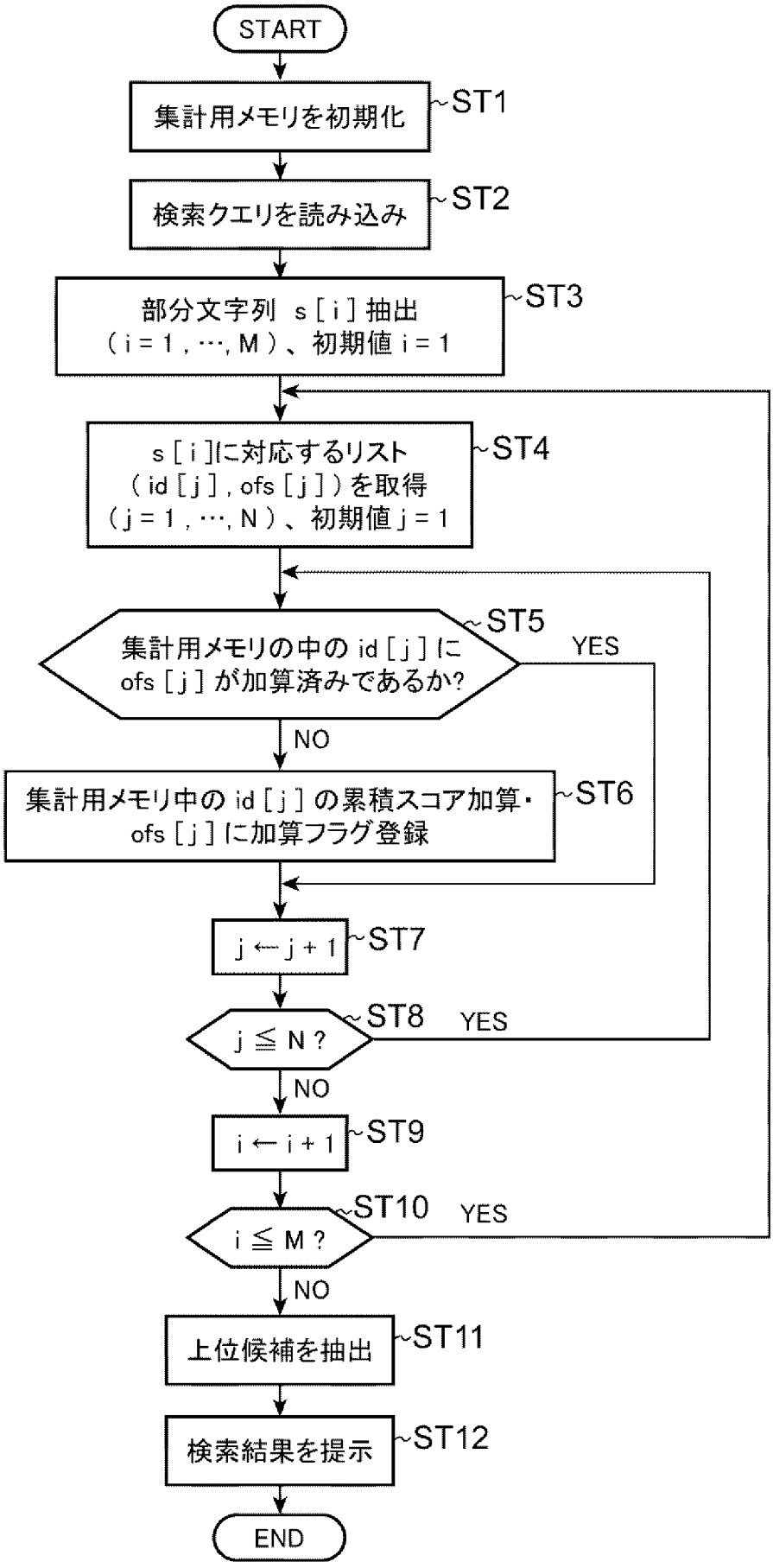
[図8]

見出し	名称ID・位置情報リスト
アア	(10,1) ...
アイ	(10,2) ...
...	...
イン	(02,5) (03,1) (05,7) ...
ウド	(02,4) (11,1) (12,1) ...
ウト	(02,2) ...
...	...

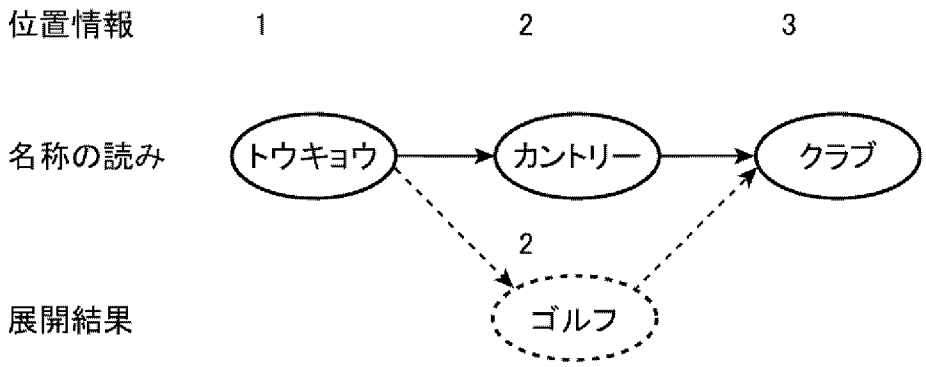
[図9]


検索語
入力
検索結果
出力
20
検索装置
21
入力部
22
部分文字列
抽出部
23
部分文字列
検索部
24
候補集計部
24a
集計用
メモリ
25
提示候補
選択部
26
候補
提示部
101
名称
データベース
102
部分文字列
索引記憶部

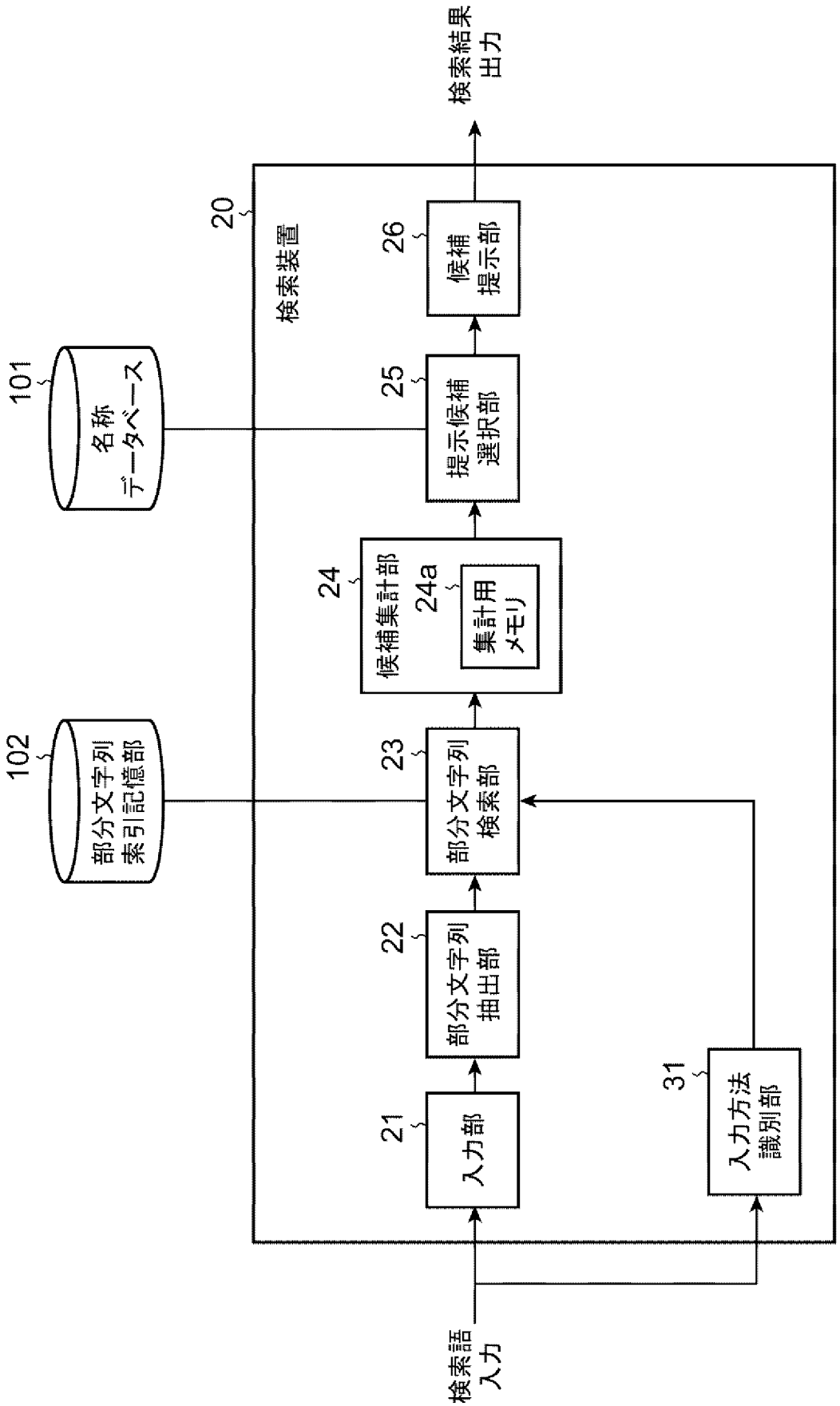
[図10]


START
集計用メモリを初期化
ST1
検索クエリを読み込み
ST2
部分文字列 s [i] 抽出
(i = 1 , ⋯, M)、初期値 i = 1
ST3
s [i]に対応するリスト
(id [j] , ofs [j]) を取得
(j = 1 , ⋯, N)、初期値 j = 1
ST4
集計用メモリの中の id [j] に
ofs [j] が加算済みであるか?
ST5
YES
NO
集計用メモリ中の id [j] の累積スコア加算・
ofs [j] に加算フラグ登録
ST6
j ← j + 1
ST7
j ≦ N ?
ST8
YES
NO
i ← i + 1
ST9
i ≦ M ?
ST10
YES
NO
上位候補を抽出
ST11
検索結果を提示
ST12
END

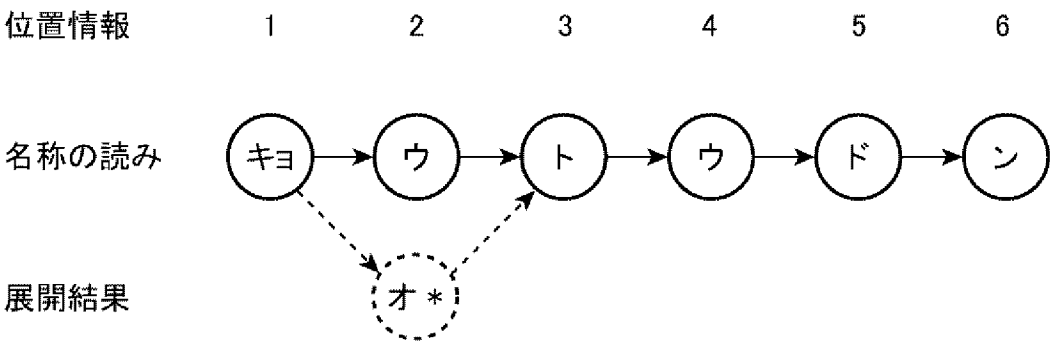
[図11]



[図12]


検索結果
出力
20
検索装置
26
候補
提示部
101
名称
データベース
25
提示候補
選択部
24
候補集計部
24a
集計用
メモリ
102
部分文字列
索引記憶部
23
部分文字列
検索部
22
部分文字列
抽出部
21
入力部
31
入力方法
識別部
検索語
入力

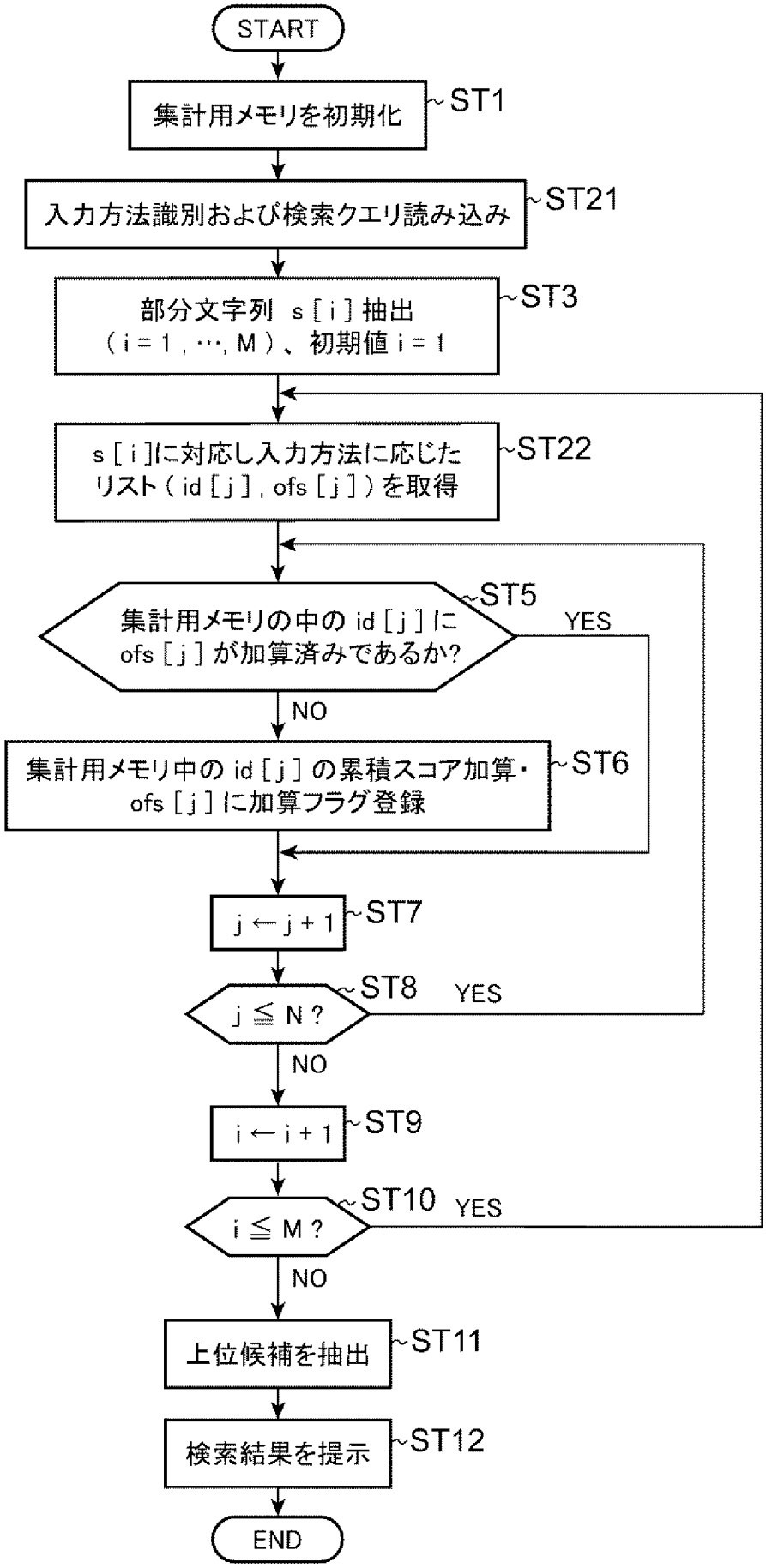
[図13]



[図14]

見出し	名称ID	位置情報
キョウ	0002	1
キョオ＊	0002	1
ウト	0002	2
オ＊ト	0002	2
トウ	0002	3
ウド	0002	4
ドン	0002	5

[図15]


START
集計用メモリを初期化
ST1
入力方法識別および検索クエリ読み込み
ST21
部分文字列 s[i]抽出
(i=1,…,M)、初期値 i=1
ST3
s[i]に対応し入力方法に応じた
リスト(id[j], ofs[j])を取得
ST22
集計用メモリの中の id[j]に
ofs[j]が加算済みであるか?
ST5
YES
NO
集計用メモリ中の id[j]の累積スコア加算・
ofs[j]に加算フラグ登録
ST6
j←j+1
ST7
j≦N?
ST8
YES
NO
i←i+1
ST9
i≦M?
ST10
YES
NO
上位候補を抽出
ST11
検索結果を提示
ST12
END

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2008/002898

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 1996-2008
Kokai Jitsuyo Shinan Koho 1971-2008 Toroku Jitsuyo Shinan Koho 1994-2008

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.
Y	JP 2001-337989 A (Ricoh Co., Ltd.), 07 December, 2001 (07.12.01), Par. Nos. [0015] to [0021] (Family: none)	1-7
Y	JP 2004-206608 A (Nippon Telegraph And Telephone Corp.), 22 July, 2004 (22.07.04), Par. Nos. [0049] to [0051] (Family: none)	1-7
Y	JP 2002-312365 A (Fujitsu Ltd.), 25 October, 2002 (25.10.02), Par. Nos. [0033] to [0046] & US 7142716 B2 & US 2002/0154817 A1	5-7

☒ 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	Date of mailing of the international search report
06 November, 2008 (06.11.08)	18 November, 2008 (18.11.08)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2008/002898

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 2006-65675 A (Canon Inc.), 09 March, 2006 (09.03.06), Par. Nos. [0033] to [0035] & US 2006/0047647 A1	5-7

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２００８／００２８９８

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

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（ＩＰＣ））

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	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２００８年
日本国実用新案登録公報	１９９６－２００８年
日本国登録実用新案公報	１９９４－２００８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	関連する請求の範囲の番号
Y	JP 2001-337989 A（株式会社リコー）2001.12.07,【００１５】－【００２１】（ファミリーなし）	1-7
Y	JP 2004-206608 A（日本電信電話株式会社）2004.07.22,【００４９】－【００５１】（ファミリーなし）	1-7
Y	JP 2002-312365 A（富士通株式会社）2002.10.25,【００３３】－【００４６】 & US 7142716 B2 & US 2002/0154817 A1	5-7

☑　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 ０６．１１．２００８	国際調査報告の発送日 １８．１１．２００８
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 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）　５Ｍ　４０５２ 鈴木　和樹 電話番号　０３－３５８１－１１０１　内線　３５９９

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	関連する請求の範囲の番号
A	JP 2006-65675 A（キヤノン株式会社）2006.03.09, 【００３３】－【００３５】 & US 2006/0047647 A1	5-7