(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2019-219764
(P2019-219764A)
(43)公開日　令和1年12月26日(2019.12.26)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 16/00	(2019.01)	G06F 17/30	320C	
G06F 21/62	(2013.01)	G06F 17/30	340Z	
		G06F 21/62	345	

審査請求 未請求　請求項の数 16　ＯＬ　（全 24 頁）

(21)出願番号　特願2018-114944 (P2018-114944)
(22)出願日　平成30年6月15日 (2018.6.15)

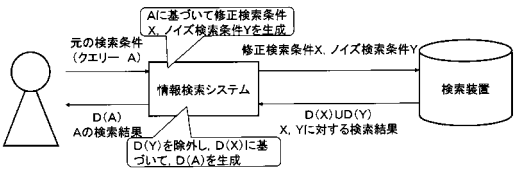
（出願人による申告）平成２４年度，国立研究開発法人科学技術振興機構，戦略的創造研究推進事業・総括実施型研究（ＥＲＡＴＯ）「河原林巨大グラフプロジェクト」に係る委託業務，産業技術力強化法第１９条の適用を受ける特許出願

(71)出願人　504202472
大学共同利用機関法人情報・システム研究機構
東京都立川市緑町１０番３号
(74)代理人　100205084
弁理士　吉浦　洋一
(72)発明者　河原林　健一
東京都千代田区一ツ橋二丁目１番２号　大学共同利用機関法人情報・システム研究機構　国立情報学研究所内
(72)発明者　町出　智也
東京都千代田区一ツ橋二丁目１番２号　大学共同利用機関法人情報・システム研究機構　国立情報学研究所内

最終頁に続く

(54)【発明の名称】情報検索システム

(57)【要約】
【課題】
情報を検索するための情報検索システムを提供することを目的とする。
【解決手段】
ユーザの実際の検索条件である第１の検索単語の意味解析に基づいて，ノイズとなる第２の検索単語を特定するノイズ処理部と，第１の検索単語の意味解析を用いて，第１の検索単語を修正する第３の検索単語を特定する検索条件修正処理部と，第２の検索単語と第３の検索単語とを検索装置に送り，検索結果を受け付ける検索装置処理部と，を有する情報検索システムである。
【選択図】　図１



【特許請求の範囲】
【請求項１】
情報を検索するための情報検索システムであって，
前記情報検索システムは，
ユーザの実際の検索条件である第１の検索単語の意味解析に基づいて，ノイズとなる第２の検索単語を特定するノイズ処理部と，
前記第１の検索単語の意味解析に基づいて，前記第１の検索単語を修正する第３の検索単語を特定する検索条件修正処理部と，
前記第２の検索単語と前記第３の検索単語とを検索装置に送り，検索結果を受け付ける検索装置処理部と，
を有することを特徴とする情報検索システム。
【請求項２】
情報を検索するための情報検索システムであって，
前記情報検索システムは，
ユーザの実際の検索条件である第１の検索単語の意味解析に基づいて，ノイズとなる第２の検索単語を特定するノイズ処理部と，
前記第１の検索単語と前記第２の検索単語とを検索装置に送り，検索結果を受け付ける検索装置処理部と，
を有することを特徴とする情報検索システム。
【請求項３】
前記ノイズ処理部は，
前記第１の検索単語と同じクラスに属する単語を用いて，クラスタリング耐性のある前記第２の検索単語を特定する，
ことを特徴とする請求項１または請求項２に記載の情報検索システム。
【請求項４】
前記ノイズ処理部は，
前記第１の検索単語に基づいて，少なくとも二以上の手法により，クラスタリング耐性のある前記第２の検索単語を特定し，
各手法による前記第２の検索単語の数または割合が変動する，
ことを特徴とする請求項１から請求項３のいずれかに記載の情報検索システム。
【請求項５】
前記ノイズ処理部は，
前記第１の検索単語と同じクラスに属する単語から複数の単語を特定することで単語群を構成し，
前記構成した単語群に対して，高密度クラスタから前記第２の検索単語を特定するクラスタ手法，前記単語群を分割することで前記第２の検索単語を特定する分割手法，前記単語群を構成する単語からランダムに前記第２の検索単語を特定するランダム手法，のいずれか一以上の手法を用いることで，ノイズ単語を特定する，
ことを特徴とする請求項１から請求項４のいずれかに記載の情報検索システム。
【請求項６】
前記ノイズ処理部は，
前記クラスタ手法として，前記構成した単語群を用いて，前記第１の検索単語とは異なるクラスタを構成する複数の単語を特定することで，前記第２の検索単語を特定する，
ことを特徴とする請求項５に記載の情報検索システム。
【請求項７】
前記ノイズ処理部は，
前記クラスタ手法として，前記構成した単語群を用いて，頻出頻度に基づく単語群を構成し，
前記頻出頻度に基づく単語群において，前記第１の検索単語からの距離と類似性に基づいて特定した単語を用いてクラスタを生成することで，前記第２の検索単語を特定する，

ことを特徴とする請求項５または請求項６に記載の情報検索システム。
【請求項８】
前記ノイズ処理部は，
前記分割手法として，前記構成した単語群を用いて，前記第１の検索単語とは非類似であり，かつ類似する単語同士を，前記第２の検索単語として特定する，
ことを特徴とする請求項５から請求項７のいずれかに記載の情報検索システム。
【請求項９】
前記ノイズ処理部は，
前記分割手法として，前記構成した単語群を複数に分割し，分割した単語群における単語と前記第１の検索単語との類似性を用いて，前記第２の検索単語を特定する，
ことを特徴とする請求項５から請求項８のいずれかに記載の情報検索システム。
【請求項１０】
前記検索条件修正処理部は，
ベクトル化した前記第１の検索単語とノイズベクトルとを用いて演算することで，前記第３の検索単語を特定する，
ことを特徴とする請求項１，請求項３から請求項９のいずれかに記載の情報検索システム。
【請求項１１】
前記情報検索システムは，
前記検索装置から受け付けた前記第２の検索単語に対応する検索結果を除外し，前記検索装置から受け付けた前記第１の検索単語または前記第２の検索単語に対応する検索結果に基づいて，前記第１の検索単語に対する検索結果を出力する検索結果処理部，
を有することを特徴とする請求項１から請求項１０のいずれかに記載の情報検索システム。
【請求項１２】
前記情報検索システムは，
前記第２の検索単語と前記第３の検索単語とを出力することで，前記第１の検索単語を推測させる処理部，
を有することを特徴とする請求項１から請求項１１のいずれかに記載の情報検索システム。
【請求項１３】
情報を検索するための情報検索システムであって，
前記情報検索システムは，
ユーザの実際の検索条件であるオリジナル検索条件をベクトル化し，ベクトル化した前記オリジナル検索条件を用いて修正検索条件を特定する検索条件修正処理部と，
前記オリジナル検索条件に基づいて，ノイズとなるノイズ検索条件を特定するノイズ処理部と，
前記修正検索条件と前記ノイズ検索条件とを検索装置に送り，検索結果を受け付ける検索装置処理部，
を有することを特徴とする情報検索システム。
【請求項１４】
コンピュータを，
ユーザの実際の検索条件である第１の検索単語の意味解析に基づいて，ノイズとなる第２の検索単語を特定するノイズ処理部，
前記第１の検索単語の意味解析に基づいて，第３の検索単語を特定する検索条件修正処理部，
前記第２の検索単語と前記第３の検索単語とを検索装置に送り，検索結果を受け付ける検索装置処理部，
として機能させることを特徴とする情報検索プログラム。
【請求項１５】

コンピュータを，
ユーザの実際の検索条件である第１の検索単語の意味解析に基づいて，ノイズとなる第２の検索単語を特定するノイズ処理部，
前記第１の検索単語と前記第２の検索単語とを検索装置に送り，検索結果を受け付ける検索装置処理部，
として機能させることを特徴とする情報検索プログラム。
【請求項１６】
コンピュータを，
ユーザの実際の検索条件であるオリジナル検索条件をベクトル化し，ベクトル化した前記オリジナル検索条件を用いて修正検索条件を特定する検索条件修正処理部，
前記オリジナル検索条件に基づいて，ノイズとなるノイズ検索条件を特定するノイズ処理部
前記修正検索条件と前記ノイズ検索条件とを検索装置に送り，検索結果を受け付ける検索装置処理部，
として機能させることを特徴とする情報検索プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は，情報を検索するための情報検索システムに関する。とくに，検索者が入力をしたキーワードなどの検索条件を，検索エンジンなどの検索装置に知られずに検索を行うことができる情報検索システムに関する。
【背景技術】
【０００２】
インターネットやデータベースから，所望の情報を得るために，検索エンジンなどの検索装置が用いられている。とくにインターネットでの検索エンジンは，無数にあるウェブサイトから検索条件にヒットするウェブサイトを特定するために有益である。
【０００３】
検索装置は有益な面があるものの，検索装置に入力された検索条件を蓄積して解析をすることで，当該検索者の関心や興味の対象，思想などの一定の傾向を把握することが可能となる。そのため検索装置に入力する検索条件を，極力，検索装置に把握されることを回避したい要望がある。しかし，検索装置は，検索条件に基づいて情報の検索を行うので，検索装置に適切な検索条件を入力しないと，所望の情報が記載された検索結果が得られないこととなる。
【０００４】
そこで，検索装置に，ユーザの実際の検索条件を把握されにくくする一方，検索装置からは所望の検索結果を得ることができるためのシステムが検討されており，たとえば特許文献１，特許文献２がある。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開平１１－２５９５１２号公報
【特許文献２】特許第５３０６３５６号
【発明の概要】
【発明が解決しようとする課題】
【０００６】
特許文献１のシステムは，入力された検索言語を，類似語，上位概念語に変換をすることで，データ検索サーバには，直接，入力された検索言語が把握されないようにするシステムである。
【０００７】
特許文献２のシステムは，検索語を文字単位で分解することで抽出した検索語における

正当文字列に，ノイズ語のノイズ文字列を組み合わせてクエリを生成することで，検索エンジンに，検索語が把握されないようにするシステムである。

【０００８】

特許文献１のシステムでは類似語や上位概念語に変換がされているので，クラスタリングをすることで，検索言語を類推することができてしまう課題がある。また，特許文献２のシステムでは，単語単位で文字列を分解していても，特許文献１と同様に，クラスタリングをすることで，元の単語を推定できてしまう課題がある。

【課題を解決するための手段】

【０００９】

そこで本発明者らは，上記課題に鑑み，実際の検索条件の特定が困難である情報検索システムを発明した。特に，クライスタリング耐性の高いノイズを用いる場合には，クラスタリング耐性を高めることができる。

【００１０】

第１の発明は，情報を検索するための情報検索システムであって，前記情報検索システムは，ユーザの実際の検索条件である第１の検索単語の意味解析に基づいて，ノイズとなる第２の検索単語を特定するノイズ処理部と，前記第１の検索単語の意味解析に基づいて，前記第１の検索単語を修正する第３の検索単語を特定する検索条件修正処理部と，前記第２の検索単語と前記第３の検索単語とを検索装置に送り，検索結果を受け付ける検索装置処理部と，を有する情報検索システムである。

【００１１】

第２の発明は，情報を検索するための情報検索システムであって，前記情報検索システムは，ユーザの実際の検索条件である第１の検索単語の意味解析に基づいて，ノイズとなる第２の検索単語を特定するノイズ処理部と，前記第１の検索単語と前記第２の検索単語とを検索装置に送り，検索結果を受け付ける検索装置処理部と，を有する情報検索システムである。

【００１２】

第１の発明，第２の発明を用いることで，実際の検索条件である第１の検索単語に対して意味解析をして特定したノイズとなる第２の検索単語を検索装置に送ることができる。これによって，実際の検索条件の特定を困難とすることができる。

【００１３】

上述の発明において，前記ノイズ処理部は，前記第１の検索単語と同じクラスに属する単語を用いて，クラスタリング耐性のある前記第２の検索単語を特定する，情報検索システムのように構成することができる。

【００１４】

ノイズとする第２の検索単語について，クラスタリング耐性となる単語を用いることで，検索装置側においてクラスタ解析を行ったとしても，実際の検索条件である第１の検索単語を特定することが困難となる。

【００１５】

上述の発明において，前記ノイズ処理部は，前記第１の検索単語に基づいて，少なくとも二以上の手法により，クラスタリング耐性のある前記第２の検索単語を特定し，各手法による前記第２の検索単語の数または割合が変動する，情報検索システムのように構成することができる。

【００１６】

複数の手法を用いて第２の検索単語を特定することで，検索装置側において，実際の検索条件である第１の検索単語を特定することがさらに困難となる。

【００１７】

上述の発明において，前記ノイズ処理部は，前記第１の検索単語と同じクラスに属する単語から複数の単語を特定することで単語群を構成し，前記構成した単語群に対して，高密度クラスタから前記第２の検索単語を特定するクラスタ手法，前記単語群を分割することで前記第２の検索単語を特定する分割手法，前記単語群を構成する単語からランダムに

前記第２の検索単語を特定するランダム手法，のいずれか一以上の手法を用いることで，ノイズ単語を特定する，情報検索システムのように構成することができる。

【００１８】

ノイズとする第２の検索単語を特定するためには，本発明のような方法を一または複数用いるとよい。

【００１９】

上述の発明において，前記ノイズ処理部は，前記クラスタ手法として，前記構成した単語群を用いて，前記第１の検索単語とは異なるクラスタを構成する複数の単語を特定することで，前記第２の検索単語を特定する，情報検索システムのように構成することができる。

【００２０】

上述の発明において，前記ノイズ処理部は，前記クラスタ手法として，前記構成した単語群を用いて，頻出頻度に基づく単語群を構成し，前記頻出頻度に基づく単語群において，前記第１の検索単語からの距離と類似性に基づいて特定した単語を用いてクラスタを生成することで，前記第２の検索単語を特定する，情報検索システムのように構成することができる。

【００２１】

これらの発明の処理を実行することで，ノイズとする第２の検索単語について，第１の検索単語とは異なるクラスタに属する単語により構成することができる。そのため，検索装置側においてクラスタ解析を行ったとしても，実際の検索条件である第１の検索単語を特定することが困難となる。

【００２２】

上述の発明において，前記ノイズ処理部は，前記分割手法として，前記構成した単語群を用いて，前記第１の検索単語とは非類似であり，かつ類似する単語同士を，前記第２の検索単語として特定する，情報検索システムのように構成することができる。

【００２３】

上述の発明において，前記ノイズ処理部は，前記分割手法として，前記構成した単語群を複数に分割し，分割した単語群における単語と前記第１の検索単語との類似性を用いて，前記第２の検索単語を特定する，情報検索システムのように構成することができる。

【００２４】

これらの発明の処理を実行することで，ノイズとする第２の検索単語について，第１の検索単語とは類似していない単語により構成することができる。そのため，検索装置側においてクラスタ解析を行ったとしても，実際の検索条件である第１の検索単語を特定することが困難となる。

【００２５】

上述の発明において，前記検索条件修正処理部は，ベクトル化した前記第１の検索単語とノイズベクトルとを用いて演算することで，前記第３の検索単語を特定する，情報検索システムのように構成することができる。

【００２６】

本発明のように構成することで，第１の検索単語そのものではないが，意味が近い単語を特定し，第３の検索単語を特定することができる。

【００２７】

上述の発明において，前記情報検索システムは，前記検索装置から受け付けた前記第２の検索単語に対応する検索結果を除外し，前記検索装置から受け付けた前記第１の検索単語または前記第２の検索単語に対応する検索結果に基づいて，前記第１の検索単語に対する検索結果を出力する検索結果処理部，情報検索システムのように構成することができる。

【００２８】

第２の検索単語はノイズであるので，その検索結果は不要である。したがって，第２の検索単語による検索結果を除外して，最終的な検索結果を出力すればよい。

【００２９】
上述の発明において，前記情報検索システムは，前記第２の検索単語と前記第３の検索単語とを出力することで，前記第１の検索単語を推測させる処理部，を有する情報検索システムのように構成することができる。
【００３０】
本発明の情報検索システムの効果は，そのままで認識しにくい。そこで，本発明のように構成することで，本発明の効果を認識させることができる。
【００３１】
第１３の発明は，情報を検索するための情報検索システムであって，前記情報検索システムは，ユーザの実際の検索条件であるオリジナル検索条件をベクトル化し，ベクトル化した前記オリジナル検索条件を用いて修正検索条件を特定する検索条件修正処理部と，前記オリジナル検索条件に基づいて，ノイズとなるノイズ検索条件を特定するノイズ処理部と，前記修正検索条件と前記ノイズ検索条件とを検索装置に送り，検索結果を受け付ける検索装置処理部と，を有する情報検索システムである。
【００３２】
本発明を用いることで，検索装置において，実際の検索条件である第１の検索条件の特定を困難とすることができる。この場合，検索条件としてはベクトル表現できる情報であればよく，単語に限らず，画像情報，音情報であっても同様に実現することができる。
【００３３】
第１の発明は，本発明のプログラムをコンピュータに読み込ませて実行することで実現できる。すなわち，コンピュータを，ユーザの実際の検索条件である第１の検索単語の意味解析に基づいて，ノイズとなる第２の検索単語を特定するノイズ処理部，前記第１の検索単語の意味解析に基づいて，第３の検索単語を特定する検索条件修正処理部，前記第２の検索単語と前記第３の検索単語とを検索装置に送り，検索結果を受け付ける検索装置処理部，として機能させる情報検索プログラムのように構成することができる。
【００３４】
第２の発明は，本発明のプログラムをコンピュータに読み込ませて実行することで実現できる。すなわち，コンピュータを，ユーザの実際の検索条件である第１の検索単語の意味解析に基づいて，ノイズとなる第２の検索単語を特定するノイズ処理部，前記第１の検索単語と前記第２の検索単語とを検索装置に送り，検索結果を受け付ける検索装置処理部，として機能させる情報検索プログラムのように構成することができる。
【００３５】
第１３の発明は，本発明のプログラムをコンピュータに読み込ませて実行することで実現できる。すなわち，コンピュータを，ユーザの実際の検索条件であるオリジナル検索条件をベクトル化し，ベクトル化した前記オリジナル検索条件を用いて修正検索条件を特定する検索条件修正処理部，前記オリジナル検索条件に基づいて，ノイズとなるノイズ検索条件を特定するノイズ処理部，前記修正検索条件と前記ノイズ検索条件とを検索装置に送り，検索結果を受け付ける検索装置処理部，として機能させる情報検索プログラムのように構成することができる。
【発明の効果】
【００３６】
本発明の情報検索システムを用いることによって，実際の検索条件の特定を困難とする情報検索システムを発明した。特に，クライスタリング耐性の高いノイズを用いる場合には，クラスタリング耐性を高めることができる。
【図面の簡単な説明】
【００３７】
【図１】本発明の情報検索システムの概念の一例を示す図である。
【図２】本発明の情報検索システムのシステム構成を示すブロック図の一例である。
【図３】本発明の情報検索システムを実現するコンピュータのハードウェア構成の一例を示す図である。

【図４】本発明の情報検索システムの処理プロセスの一例を示すフローチャートである。
【図５】検索条件修正処理の処理プロセスの一例を示すフローチャートである。
【図６】ノイズ処理の全体の処理プロセスの一例を示すフローチャートである。
【図７】クラスタ手法の処理プロセスの一例を示すフローチャートである。
【図８】分割手法の処理プロセスの一例を示すフローチャートである。
【図９】ランダム手法の処理プロセスの一例を示すフローチャートである。
【図１０】修正検索条件を特定する処理の一例を模式的に示す図である。
【図１１】クラスタ手法の処理の一例を模式的に示す図である。
【図１２】分割手法の処理の一例を模式的に示す図である。
【図１３】単語群Ｘのみを用いた場合の匿名性と再構築可能性の精度の関係を示す図である。
【図１４】単語群Ｘと単語群Ｙとを用いた場合の匿名性と再構築可能性の精度の関係を図である。
【図１５】情報検索システムによる検索結果と，実際の検索条件「kyoto」を入力した場合の比較例を示す図である。
【図１６】情報検索システムによる検索結果と，実際の検索条件「kyoto」を入力した場合の比較例を示すほかの図である。
【図１７】情報検索システムによる検索結果と，実際の検索条件「kyoto」を入力した場合の比較例を示すほかの図である。
【図１８】情報検索システムによる検索結果と，実際の検索条件「kyoto」を入力した場合の比較例を示すほかの図である。
【図１９】情報検索システムによる検索結果と，実際の検索条件「nagasaki」を入力した場合の比較例を示す図である。
【図２０】情報検索システムによる検索結果と，実際の検索条件「nagasaki」を入力した場合の比較例を示す図である。
【図２１】実施例２において，単語群Ｘの単語，単語群Ｙの単語をそれぞれ表示した状態の画面を示す図である。
【図２２】実施例２において，単語群Ｙの単語を削除し，単語群Ｘの単語のみを表示した状態の画面を示す図である。
【図２３】実施例２において，正解を表示した状態を示す画面である。
【図２４】単語Ａの検索結果Ｄ（Ａ）と単語群Ｘの検索結果Ｄ（X_i）との関係を示す図である。
【発明を実施するための形態】
【００３８】
本発明の情報検索システム１の全体の概念の一例を図１に示す。また，本発明の情報検索システム１のシステム構成のブロック図の一例を図２に示す。情報検索システム１では，情報の検索を行うユーザが利用するユーザ端末４と，情報の検索を行う検索サーバなどの検索装置３と，情報検索システム１の各処理を実行するための制御端末２とを用いる。情報検索システム１における制御端末２は，コンピュータによって実現される。コンピュータのハードウェア構成の一例を図３に示す。なお，制御端末２とユーザ端末４，制御端末２と検索装置３，制御端末２とユーザ端末４と検索装置３とが一体的に構成されていてもよい。
【００３９】
コンピュータはプログラムの演算処理を実行するＣＰＵなどの演算装置７０と，情報を記憶するＲＡＭやハードディスクなどの記憶装置７１と，ディスプレイなどの表示装置７２と，情報の入力を行う入力装置７３と，演算装置７０の処理結果や記憶装置７１に記憶する情報などの各種情報を通信する通信装置７４とを有している。なお，コンピュータがタッチパネルディスプレイを備えている場合には表示装置７２と入力装置７３とが一体的に構成されていてもよい。タッチパネルディスプレイは，携帯電話やスマートフォン，タブレット型コンピュータなどの可搬型通信端末などで利用されることが多いが，それに限

定するものではない。

【００４０】

タッチパネルディスプレイは，そのディスプレイ上で，直接，所定の入力デバイス（タッチパネル用のペンなど）や指などによって入力を行える点で，表示装置７２と入力装置７３の機能が一体化した装置である。

【００４１】

情報検索システム１の制御端末２は一台のコンピュータによって実現されていてもよいが，その機能が複数のコンピュータによって実現されていてもよい。この場合のコンピュータとして，たとえばクラウドサーバであってもよい。

【００４２】

さらに，本発明の情報検索システム１における各処理部は，その機能が論理的に区別されているのみであって，物理上あるいは事実上は同一の領域を為していても良い。

【００４３】

検索装置３は，インターネットの情報を検索するための検索エンジンサーバや，各種の情報を記憶するデータベースサーバなど，情報を検索するための装置である。なお，検索装置３としては，検索エンジンサーバやデータベースサーバに限定するものではなく，情報を検索するための装置であればよい。

【００４４】

情報検索システム１における制御端末２は，検索条件受付処理部２１と検索条件修正処理部２２とノイズ処理部２３と検索装置処理部２４と検索結果処理部２５とを有する。

【００４５】

検索条件受付処理部２１は，ユーザ端末４から，ユーザが実際に検索をしたい検索条件を含むクエリの入力を受け付ける。検索条件としては，ベクトル表現できる情報であればいかなる情報であってもよい。本明細書では，検索条件として，キーワードなどの単語の場合を説明するが，画像情報，音情報などでも同様の処理を実行することで実現できる。たとえば単語が画像情報，音情報になっている場合には，画像情報をＯＣＲ認識してテキスト化した後に処理を実行し，音情報を音声認識技術に基づいてテキスト化した後に処理を実行してもよい。また，画像情報における各画素の色情報に基づいてベクトル化してもよいし，音情報における周波数情報に基づいてベクトル化し，以降の処理を実行してもよい。

【００４６】

検索条件修正処理部２２は，検索条件受付処理部２１で受け付けた実際の検索条件に基づいて，検索装置３に入力するための修正した検索条件（修正検索条件）を特定する処理である。修正検索条件を特定する処理にはさまざまな方法を用いることができる。たとえば実際の検索条件が単語である場合，その単語そのものではないが，その単語に近い単語を修正検索条件として特定する。この場合，検索条件修正処理部２２は，実際の検索条件の単語の意味解析に基づいて，修正検索条件となる単語を特定する。なお，意味解析とは，単語エンベディング（Ｗｏｒｄ　ｅｍｂｅｄｄｉｎｇ）であって，自然言語解析における技術である。すなわち，ある単語とほかの単語の意味や概念などが類似しているか否かなど，単語の意味関係を自動的に解析するための技術である。

【００４７】

検索条件受付処理部２１で受け付けた実際の検索条件に基づいてコサイン類似度を用いて，修正検索条件を特定することができる。すなわち，検索条件修正処理部２２は，検索条件受付処理部２１で入力を受け付けた実際の検索条件をベクトル化し，それにノイズベクトルを演算，たとえば加算する。そして，演算したベクトルの点からコサイン類似度に基づき近傍検索（コサイン類似度が一定の範囲内にあるか）をすることで，修正検索条件を特定する。修正検索条件を特定する処理の一例を模式的に示すのが図１０である。

【００４８】

たとえば，以下のような処理を実行すればよい。検索条件受付処理部２１で受け付けたクエリにおける実際の検索条件における単語をＡとした場合，検索条件修正処理部２２は

，単語Ａに対応するベクトルｖ（ｖ＝ｖ$_{1}$，ｖ$_{2}$，・・・，ｖ$_{300}$）を取得する。単語Ａに対応するベクトルｖの取得方法としては，ＧｌｏＶｅ（Ｇｌｏｂａｌ Ｖｅｃｔｏｒｓ ｆｏｒ Ｗｏｒｄ Ｒｅｐｒｅｓｅｎｔａｔｉｏｎ），ｗｏｒｄ２ｖｅｃ，ｆａｓｔｔｅｘｔなどを用いる方法があるが，それに限定するものではない。また，本明細書では，単語のベクトル化にＧｌｏＶｅを用いるので，３００次元のベクトルで説明するが，それに限定するものではない。

【００４９】

そして検索条件修正処理部２２では，単語Ａに対応するベクトルｖに対して，同次元のノイズベクトルｎ（ｎ＝ｎ$_{1}$，ｎ$_{2}$，・・・，ｎ$_{300}$）を取得する。なお，ノイズベクトルにおける各実数ｎ$_{i}$は，たとえばガウス分布のノイズを用いることができるが，それに限定するものではない。

【００５０】

以上のようにして検索条件修正処理部２２で単語Ａに対応するベクトルｖとノイズベクトルｎとを取得すると，それぞれを演算，たとえば加算することでベクトルｖ’を算出する。そして，ベクトルｖ’の点から距離，たとえばコサイン類似度（コサイン距離）が近い（ベクトルｖ’の点からコサイン類似度が一定範囲内にある），任意のｍ個のベクトルｘ（ｘ$_{1}$，ｘ$_{2}$，・・・，ｘ$_{m}$）を，上述のＧｌｏＶｅのデータセット（単語Ａをベクトル化した際のモデルのデータセット）から特定する。そして，特定したベクトルｘ（ｘ$_{1}$，ｘ$_{2}$，・・・，ｘ$_{m}$）に対応する単語Ｘ$_{1}$，Ｘ$_{2}$，・・・，Ｘ$_{m}$の単語群Ｘを特定することで，検索条件である単語Ａ（第１の検索単語）に対応する修正検索条件である単語（第３の検索単語）の単語群Ｘを特定することができる。単語群Ｘを構成する単語の数ｍは，任意の数でよく，複数，たとえば１０個から２０個程度とすることができるが，それに限定するものではない。

【００５１】

なお，検索条件修正処理部２２は，上述の処理のほか，たとえば，単語とそれに類似，関連する単語，上位概念の単語をあらかじめ対応づけて記憶しておき，その対応関係に基づいて，検索条件受付処理部２１で受け付けた検索条件における単語Ａに対応する単語の単語群Ｘを特定してもよい。

【００５２】

ノイズ処理部２３は，検索条件受付処理部２１で入力を受け付けた検索条件に対するノイズとなる検索条件を特定する処理である。ノイズ処理としては，無関係の検索条件を付加するほか，クラスタリングが困難となるノイズを付加することが好ましい。たとえば実際の検索条件が単語である場合，実際の検索条件の単語の意味解析に基づいて，ノイズとする単語を特定することができる。

【００５３】

ノイズ処理部２３におけるノイズ処理としては，高密度クラスタからノイズを選択するクラスタ手法，分割手法，ランダム手法などがあり，これらの手法のいずれか一以上によって得られた検索条件を特定するとよい。また，上記の３手法に限定するものではなく，上記の３手法以外，あるいは上記の３手法と組み合わせて，ほかの手法を用いることも可能である。

【００５４】

上述と同様に，検索条件受付処理部２１で受け付けたクエリにおける実際の検索条件における単語をＡとした場合，ノイズ処理部２３は，単語Ａと同じクラスに属する単語から，ノイズの候補となる所定数，たとえば１０００個の単語ｗ$_{1}$，・・・，ｗ$_{1000}$をランダムに特定し，単語群Ｗを構成する。なお，単語Ａと同じクラスに属する単語ｗは，たとえば，ウィキペディアのＯｎｔｏｌｏｇｙクラスのデータセットを用いることなどで特定することができるが，それに限定するものではない。

【００５５】

そして，ノイズ処理部２３は，特定した単語群Ｗに対して，クラスタ手法，分割手法，ランダム手法のいずれかまたは複数の手法による処理を実行し，ノイズとなる単語Ｙ$_{1}$，

Y_2，・・・，Y_zによる単語群Ｙを取得する。単語群Ｙを構成する単語の数ｚは任意の数とすることができ，好ましくは複数，たとえば１０個～２０個程度とすることができるが，それに限定するものではない。ノイズとなる単語群Ｙは，クラスタリングに対する耐性が高い単語により構成されることが好ましい。

【００５６】

ノイズ処理部２３は，複数の手法を用いる場合，ノイズとなる単語群Ｙの各単語について，各手法により得られる単語の単語数を任意の割合または数として設定することができる。たとえば単語群Ｙの単語数を１０個とする場合，クラスタ手法による単語を５個（５０％），分割手法による単語を３個（３０％），ランダム手法による単語を２個(２０％）のように設定することができる。各手法による単語の割合や数は，毎回，変更してもよいし，固定でもよい。

【００５７】

ノイズ処理部２３におけるクラスタ手法は，検索条件となる単語Ａと同じクラスに属する単語群Ｗの単語ｗに基づいて，単語Ａとは異なるクラスタを構成可能な複数の単語を特定してノイズとする単語群Ｙを構成することで，クラスタリングに対する耐性を高める。ノイズ処理部２３におけるクラスタ手法は，以下のように実行する。クラスタ手法によるノイズとなる検索条件を特定する処理の一例を模式的に示すのが図１１である。

【００５８】

まず，単語群Ｗの単語ｗのなかから，単語Ａと頻出頻度が近い単語を特定することで，クラスタの中心の候補となる単語の単語群Ｓ（s_1，s_2，・・・，s_i）を構成する。そして，特定した単語群Ｓの単語ｓのなかから，使用する単語群Ｓ’（s'_1，s'_2，・・・，s'_j）（ただしｊ＜ｉ）を特定する。この特定の際には，単語Ａのベクトルから適度に離れており，意味が類似している単語ｓ’を優先して特定をすることが好ましい。すなわち，単語群Ｓ’における単語ｓ’の特定は，単語Ａのベクトルに対して，同次元のノイズベクトル（単語Ａから適度に離れる値として設定するベクトル）を加算等の演算をして算出し，その加算したベクトルの点から，コサイン類似度（コサイン距離）が一定の閾値以上である単語群Ｓにおける単語ｓを，単語ｓ’として特定する。

【００５９】

特定した単語群Ｓ’において，それぞれの単語s'_1，s'_2，・・・，s'_jに近い単語を所定数特定し，単語のクラスタＣを生成する。たとえば各クラスタの単語数は３～１０個とするが，それに限定するものではない。そして，密集度の高いクラスタＣから順番に，クラスタにおける単語を特定し，あらかじめ設定した数になったら，それらをノイズの単語群Ｙ（Y_1，Y_2，・・・，Y_z）として特定をする。

【００６０】

このような処理を実行することで，単語Ａとは相違する密集度の高いクラスタＣにおける単語を，ノイズの単語として特定できるので，仮に検索装置３側でクラスタリングをしたとしても，実際の単語Ａの特定が困難となり，クラスタリングに対する耐性が高くなる。

【００６１】

また，ノイズ処理部２３における分割手法は，検索条件となる単語Ａと同じクラスに属する多数の単語を分割，たとえば２分割（ただし単語数は同数ではない）し，その分割によって構成される単語群における単語を用いて，検索条件となる単語Ａとは似ていない単語同士の単語群を生成することを，所定条件を充足するまで繰り返し，条件充足後の単語群から，単語Ａと似ている複数の単語を特定してノイズとする単語群Ｙを構成することで，クラスタリングに対する耐性を高める。ノイズ処理部２３における分割手法は，以下のように実行する。分割手法によるノイズとなる検索条件を特定する処理の一例を模式的に示すのが図１２である。

【００６２】

まず実際の検索条件である単語Ａの点を取る超平面をランダムに特定し，その超平面において，単語群Ｗ（w_1，・・・，w_{10000}）を，単語群Ｗ１，単語群Ｗ２の２つに分

割をする。ただし，単語群Ｗ１の単語数は，単語群Ｗ２の単語数より多いとする。そして，単語群Ｗ１において，単語Ａと類似していない単語を，単語群Ｗ１から所定割合または所定数，たとえば１０％だけ消去し，消去した単語を新たに単語群Ｗとして構成する。なお，単語Ａと類似している単語か否かは，単語Ａと，単語群Ｗ１における比較対象となる単語とのコサイン類似度（コサイン距離）が一定の閾値以上であるかで特定可能である。

【００６３】

以上の処理を所定条件，たとえば｜Ｗ｜＜２ｙ（ｙは任意の値）となるまで繰り返す。

【００６４】

上記の所定条件を充足した場合，その単語群Ｗの単語のなかから，単語Ａに類似している単語をｚ個特定し，それらを単語群Ｙ（Y_1，Y_2，・・・，Y_z）として特定をする。

【００６５】

以上のような分割手法を用いることで，実際の検索条件における単語Ａ（第１の検索単語）とは非類似であって，かつ，また単語群Ｙを構成する単語自体は意味が近い単語（第２の検索単語）をノイズの単語として特定できるので，単語Ａとは異なるクラスタとなりやすい単語をノイズの単語とすることができる。そのため，仮に検索装置３側でクラスタリングをしたとしても，実際の単語Ａの特定が困難となり，クラスタリングに対する耐性が高くなる。

【００６６】

さらに，ノイズ処理部２３におけるランダム手法は，単語群Ｗ（w_1，・・・，w_{1000}）の中から，ランダムにｚ個の単語を特定し，それらを単語群Ｙ（Y_1，Y_2，・・・，Y_z）とする。

【００６７】

ランダムにノイズとなる単語を特定することで，クラスタリングに対する耐性を高めることができる。

【００６８】

以上のような処理をノイズ処理部２３が実行することで，ノイズとする検索条件Ｙを特定できる。とくに，一つの手法のみならず，複数の手法を組み合わせることで，クラスタリングに対する耐性は，一層，高くすることができる。

【００６９】

検索装置処理部２４は，検索条件修正処理部２２で特定した単語群Ｘにおける単語と，ノイズ処理部２３で特定した単語群Ｙにおける単語とを，それぞれ検索装置３に送ることで検索処理を実行させる。この際には，単語群Ｘにおける単語，単語群Ｙにおける単語をランダムな順番で検索装置３に送るとよい。そして，各単語に対する検索結果を受け付ける。なお，少なくとも，検索装置３に送った単語群Ｘにおける単語と，その検索結果とを対応づけて記憶しておく。

【００７０】

検索結果処理部２５は，検索装置３から受け付けた検索結果に基づいて，ユーザ端末４に送る検索結果を出力する。検索結果処理部２５は，検索装置３から受け付けた単語群Ｙの単語Y_1，Y_2，・・・，Y_zに対する検索結果Ｄ（Y_i）（$1 \leqq i \leqq z$）をユーザ端末４に送る検索結果から除外し，単語群Ｘにおける単語X_1，X_2，・・・，X_mに対する検索装置３での各検索結果Ｄ（X_i）（$1 \leqq i \leqq m$）に基づいて検索結果を生成する。たとえば各検索結果Ｄ（X_1），Ｄ（X_2），・・・，Ｄ（X_m）をソートすることで，検索結果を生成する。また検索結果を生成する際に，検索結果におけるページランクを用いてもよいし，ページランクの重み付けなどを用いてソートをしてもよい。さらに検索結果処理部２５は，検索装置３から受け付けた検索結果の群Ｄ（X_i）に対して，単語Ａに基づいて検索を行うことで，検索結果を生成してもよい。なお，検索結果の生成は，公知の方法を用いることができる。

【００７１】

検索結果Ｄ（X_i）は，単語群Ｘの単語X_1，X_2，・・・，X_mに対する検索装置３

での検索結果である。そして単語群Ｘにおける単語$Ｘ_{1}$，$Ｘ_{2}$，・・・，$Ｘ_{m}$は，単語Ａに対応する単語ベクトルに近いものを特定している。すなわち，単語Ａと単語群Ｘの単語とは共起性が高い（同一の文に同時に現れやすい）。そのため，単語Ａの検索結果Ｄ（Ａ）は，共起性の高い単語群Ｘに対する検索結果Ｄ（$Ｘ_{i}$）に基づいて生成することができる。この関係を模式的に示すのが図２４である。

【００７２】

検索結果処理部２５は，以上のように生成した検索結果をユーザ端末４に送る。

【００７３】

以上のような処理を実行することで，ユーザが入力をした検索条件は検索装置３側に知られることなく，精度のよい検索結果を得ることができる。

【実施例１】

【００７４】

つぎに本発明の情報検索システム１を用いて情報の検索を行う場合の処理プロセスの一例を，図４乃至図９のフローチャートを用いて説明する。なお，本発明の処理は一例であって，その処理，とくに検索条件修正処理部２２，ノイズ処理部２３の処理などの順序を適宜，変更することは可能である。

【００７５】

ユーザが，自らが入力する検索条件としての単語を知られずに検索装置３で検索を行うことを所望する場合，ユーザ端末４において実際の検索条件としての単語Ａを入力すると，単語Ａを含むクエリがユーザ端末４から制御端末２に送られる。そして，制御端末２の検索条件受付処理部２１で，単語Ａを含むクエリを受け付け（Ｓ１００），検索条件修正処理部２２において，検索条件としての単語Ａを修正する，検索条件修正処理を実行する（Ｓ１１０）。

【００７６】

すなわち，検索条件修正処理部２２は，ＧｌｏＶｅなどの公知のモデルを用いることで，単語Ａの３００次元の単語ベクトルｖ（$ｖ＝ｖ_{1}，ｖ_{2}，・・・，ｖ_{300}$）を取得する（Ｓ２００）。また，検索条件修正処理部２２は，３００次元のノイズベクトルｎ（$ｎ＝ｎ_{1}，ｎ_{2}，・・・，ｎ_{300}$）を取得する（Ｓ２１０）。

【００７７】

このように取得した単語ベクトルｖとノイズベクトルｎとをそれぞれ加算することでベクトルｖ’（$ｖ’＝ｖ’_{1}，ｖ’_{2}，・・・，ｖ’_{300}$）を算出し（Ｓ２２０），ベクトルｖ’の点からコサイン類似度（コサイン距離）が一定の範囲内にある，任意のｍ個のベクトルｘ（$ｘ_{1}，ｘ_{2}，・・・，ｘ_{m}$）を，ＧｌｏＶｅのデータセットを参照することで特定をする（Ｓ２３０）。そして検索条件修正処理部２２は，特定した各ベクトルｘ（$ｘ_{1}，ｘ_{2}，・・・，ｘ_{m}$）に対応する単語$Ｘ_{1}$，$Ｘ_{2}$，・・・，$Ｘ_{m}$を特定し，それらを修正検索条件の単語群Ｘとする（Ｓ２４０）。たとえば，ｍは１０個とすることができるが，数を増減してもよい。

【００７８】

以上のように修正検索条件の単語群Ｘの単語$Ｘ_{1}$，$Ｘ_{2}$，・・・，$Ｘ_{m}$を特定する。

【００７９】

また，ノイズ処理部２３は，単語Ａに基づいて，ノイズとする単語の単語群Ｙを特定するノイズ処理を実行する（Ｓ１２０）。

【００８０】

ノイズ処理部２３は，まず，ウィキペディアのＯｎｔｏｌｏｇｙクラスのデータセットを参照し，単語Ａと同じクラスに属する単語から，十分に大きな数，たとえば１０００個程度以上の単語ｗを特定する（Ｓ３００）。これらの単語ｗによって構成される単語群を，単語群Ｗとする。たとえば単語群Ｗは，単語$ｗ_{1}$，・・・，$ｗ_{1000}$により構成される。

【００８１】

そしてノイズ処理部２３は，ノイズとする単語群Ｙの単語数を１０個とし，その比率を

，たとえばクラスタ手法による単語数が５個，分割手法が３個，ランダム手法が２個と決定をすると，各手法によって，単語群Ｙにおけるノイズとする単語$Ｙ_{1}$，$Ｙ_{2}$，・・・，$Ｙ_{10}$を特定する処理を実行する（Ｓ３１０，Ｓ３２０，Ｓ３３０）。

【００８２】

まずクラスタ手法によりノイズとする単語$Ｙ_{1}$，$Ｙ_{2}$，・・・，$Ｙ_{5}$を特定するには（Ｓ３１０），ノイズ処理部２３は，単語Ａの頻出頻度と，単語$ｗ_{1}$，・・・，$ｗ_{1000}$のそれぞれの頻出頻度とを比較することで，単語Ａの頻出頻度から所定範囲内の頻出頻度にある単語ｗを特定し，その特定した単語により単語群Ｓ（$ｓ_{1}$，$ｓ_{2}$，・・・，$ｓ_{i}$）を構成する（Ｓ４００）。この単語群Ｓにおける単語$ｓ_{1}$，$ｓ_{2}$，・・・，$ｓ_{i}$は，クラスタの中心の候補となる単語である。

【００８３】

そして，単語Ａの単語ベクトルに，任意に設定する同次元のノイズベクトルを加算する。そして，その加算したベクトルの点から，コサイン類似度（コサイン距離）が一定の閾値以上である単語群Ｓの単語を特定し，特定した単語により単語群Ｓ’（$ｓ’_{1}$，$ｓ’_{2}$，・・・，$ｓ’_{j}$）（ただしｊ＜ｉ）を構成する（Ｓ４１０）。

【００８４】

以上のように特定した単語群Ｓ’を構成する各単語$ｓ’_{1}$，$ｓ’_{2}$，・・・，$ｓ’_{j}$のうち，これらの各単語の単語ベクトルの点からコサイン類似度（コサイン距離）が一定の範囲内にある任意の数（たとえば３～１０個程度）のベクトルを，ＧｌｏＶｅのデータセットを参照することで特定をする。そして特定した各ベクトルに対応する単語を特定することで，一つのクラスタＣを構成する。そして，単語のクラスタＣを一または複数構成する（Ｓ４２０）。このようにすることで，単語群Ｓ’を構成する各単語$ｓ’_{1}$，$ｓ’_{2}$，・・・，$ｓ’_{j}$に近い単語に基づいて単語のクラスタＣを構成することができる。

【００８５】

そして各クラスタＣにおける単語の密集度が高いクラスタから順番に，そのクラスタにおける単語を特定し，あらかじめ設定した数，ここでは５個になったら，それらをノイズの単語群Ｙ（$Ｙ_{1}$，$Ｙ_{2}$，・・・，$Ｙ_{5}$）として特定をする（Ｓ４３０）。

【００８６】

以上のような処理をノイズ処理部２３が実行することで，クラスタ手法によるノイズとする単語$Ｙ_{1}$，$Ｙ_{2}$，・・・，$Ｙ_{5}$を特定できる。

【００８７】

つぎに，ノイズ処理部２３が分割手法により，ノイズとする単語$Ｙ_{6}$，$Ｙ_{7}$，$Ｙ_{8}$を特定するには（Ｓ３２０），ノイズ処理部２３は，まず，単語Ａのベクトルの点を通る超平面をランダムに特定することで（Ｓ５００），Ｓ３００で特定した単語群Ｗにおける単語ｗについて，単語群Ｗ１，単語群Ｗ２に分割をする（Ｓ５１０）。このとき，単語数が多い領域を単語群Ｗ１，少ない領域を単語群Ｗ２とする。

【００８８】

そして，ノイズ処理部２３は，単語Ａと，単語群Ｗ１における各単語とのコサイン類似度（コサイン距離）を比較し，コサイン類似度に基づいてソートをする。そして，コサイン類似度が低い順に下から，たとえば１０％程度の単語群Ｗ１における単語を，単語Ａに類似していない単語として，単語群Ｗ１から消去する。そしてこの消去した各単語を，新たな単語群Ｗとして構成する（Ｓ５２０）。

【００８９】

Ｓ５２０で特定した単語群Ｗの単語に基づいて，Ｓ５００乃至Ｓ５２０の処理を，所定条件，たとえば｜Ｗ｜＜２ｙ（ｙは任意の値）となるまで繰り返す（Ｓ５３０）。

【００９０】

そして所定条件を充足した場合，最終的な単語群Ｗにおける単語と，単語Ａとのコサイン類似度（コサイン距離）を比較し，コサイン類似度に基づいてソートをする。そして，コサイン類似度が高い順に上から，分割手法によるノイズの単語数分（ここでは３個）の単語を特定することで，分割手法によるノイズとする単語$Ｙ_{6}$，$Ｙ_{7}$，$Ｙ_{8}$を特定する（

Ｓ５４０）。

【００９１】

さらに，ノイズ処理部２３がランダム手法により，ノイズとする単語Y_9，Y_{10}を特定するには（Ｓ３３０），Ｓ３００で特定した単語群W（w_1，・・・，w_{1000}）の中から，ランダムに２個の単語を特定し，それらをランダム手法によるノイズとする単語Y_9，Y_{10}とする（Ｓ６００）。

【００９２】

ノイズ処理部２３が以上のような処理を実行することで，ノイズとする単語群Ｙを構成する単語Y_1，Y_2，・・・，Y_{10}を特定することができる。

【００９３】

検索装置処理部２４は，検索条件修正処理部２２で特定した修正検索条件の単語群Ｘの単語X_1，X_2，・・・，X_{10}，ノイズ処理部２３で特定したノイズとする単語群Ｙの単語Y_1，Y_2，・・・，Y_{10}を，たとえばランダムや所定の規則に基づいて検索装置３に送ることで，各単語に基づく検索処理を検索装置３に実行させる（Ｓ１３０）。なおこの際に，検索装置処理部２４は，単語群Ｘ，単語群Ｙの各単語のほかに，単語Ａを検索装置３に送ってもよい。

【００９４】

そして，検索装置処理部２４は，検索装置３に送った各単語に基づく検索結果を受け付け（Ｓ１４０），検索結果処理部２５が，ユーザ端末４に送る検索結果の生成処理を行う（Ｓ１５０）。すなわち，検索結果処理部２５は，検索装置処理部２４で受け付けた検索結果のうち，単語群Ｙにおける単語Y_1，Y_2，・・・，Y_{10}に対する検索結果Ｄ（Y_1），Ｄ（Y_2），・・・，Ｄ（Y_{10}）を除外し，単語群Ｘにおける単語X_1，X_2，・・・，X_{10}に対する検索装置３での各検索結果Ｄ（X_1），Ｄ（X_2），・・・，Ｄ（X_{10}）に基づいてソートするなど公知の手法を用いることで，単語Ａに対する検索結果Ｄ（Ａ）を生成する。そして検索結果処理部２５は，Ｓ１５０で生成した検索結果を，単語Ａに対する検索結果Ｄ（Ａ）として，ユーザ端末４に送る（Ｓ１６０）。

【００９５】

ユーザ端末４でこの検索結果を受け付けることで，ユーザは，自らが入力した単語Ａに対する検索結果Ｄ（Ａ）を取得することができる。

【００９６】

検索条件の匿名化（検索装置３に検索条件を知られないようにすること）と，検索結果の精度とはトレードオフの関係にある。本発明の情報検索システム１において，匿名性αは，実際の検索条件である単語Ａとの間の平均コサイン類似度を用い，以下の数１で示される。

（数１）

$$\alpha = 1 - \frac{1}{|Q(A)|}\sum_{i=1}^{n} sim(v(A),v(X_i))$$

ここで，ｖ（Ａ）は単語Ａのベクトルであり，ｖ（X_i）は修正検索条件である単語Ｘのベクトルであり，Ｑ（Ａ）は，単語Ａに基づく修正検索条件の単語群Ｘ（X_1，X_2，・・・，X_n）である。

【００９７】

また，修正検索条件Ｘに基づく検索結果による，検索結果の再構築可能性の精度ρは，以下の数２で示される。

（数２）

$$\rho = \frac{\left|D(A) \bigcap D'(A)\right|}{|D(A)|}$$

ここでＤ（Ａ）は，単語Ａに基づく検索結果であり，Ｄ’（Ａ）は，以下の数３で示される修正検索条件Ｘによる検索結果を用いて再構成された検索結果である。

（数３）

$$D'(A) = \bigcup_{i=1}^{n} D(X_i)$$

【００９８】

そして，匿名性αと，検索結果の再構築可能性の精度ρは，以下の数４の関係性が成立する。

（数４）

$$log\ \rho = \frac{ck}{2d}(c + 2(1 - \alpha)\ ||v(A)||_2) - log\ Z$$

【００９９】

出願人による本発明の情報検索システム１における，単語群Ｘのみを用いた場合の匿名性と再構築可能性の精度の関係を図１３に，単語群Ｘと単語群Ｙとを用いた場合の匿名性と再構築可能性の精度の関係を図１４に示す。図１３（ａ）および図１４（ａ）は修正検索条件の単語Ｘを特定するにあたりノイズベクトルｎを用いない場合であり，図１３（ｂ）および図１４（ｂ）はノイズベクトルが小さい場合であり，図１３（ｃ）および図１４（ｃ）はノイズベクトルが大きい場合である。

【０１００】

図１３と図１４は，本発明が実験的に確認できていることを示している。つまり，匿名性と再構築性がトレードオフの関係にあり，匿名性を上げれば再構築性は小さくなり，匿名性を下げれば再構築性は大きくなる。 匿名性はノイズの大小で制御できるため，ノイズの選び方により，匿名性か再構築性のどちらを重要視したいかを選択できる。図１３と図１４の比較から，単語群Ｙを使用した方（図１４）が使用しない方（図１３）より全般的に匿名性が向上することがわかる。

【０１０１】

また，図１５乃至図２０に，本発明の情報検索システム１による検索結果と，実際の検索条件を入力した場合の比較例を示す。なお，図１５乃至図２０で用いた検索装置３はwikipediaである。図１５乃至図１８は実際の検索条件として「kyoto」を用いており，図１５および図１６では強いノイズ（ノイズベクトルが大きい）を，図１７および図１８では弱いノイズ（ノイズベクトルが小さい）の場合を示している。また図１５乃至図１８ではノイズ処理部２３が，分割手法とランダム手法を用いて単語群Ｙを特定している。

【０１０２】

図１５では単語群Ｘの単語として「tokyo,copenhagen,hokkaido,nagoya,osaka,japan,kansai,seoul,fukuoka,chiba」が，単語群Ｙの単語として「arkansas,pueblo,saitama,conway,john,rosario,owen souund,armenia,patti,lyons,laporte,knowle west,columbus,north berwick,surat,patterson,millbrook,san diego,gill,walnut」が特定されている。そして，実際の検索条件「kyoto」で検索した場合と比較して，１０個中９個の検索結果が一致している（左側のコラムが実際の検索条件に基づく検索結果，右側のコラムが本発明の情報検索システム１に基づく検索結果であり，左側のコラムに表示される「Ｅ」が一致している検索結果である）。また，匿名性αは０．７７８，検索結果の再構築可能性ρは０．４２１（ただし上位１００の検索結果では０．７１）である。

【０１０３】

また図１６では単語群Ｘの単語として「vasteras,nagoya,seoul,cmom,waseda,osaka,ginza,joad,tokyo,yokohama」が，単語群Ｙの単語として「ina,valley,islampur,rudbar,qasemabad,habibabad,alexander,wollongong,first,mehrdasht,humboldt,price,lara,perth,hayden,dauphin,hat,kuhsar,jahanabad,nosratabad」が特定されている。そして，実際の検索条件「kyoto」で検索した場合と比較して，１０個中７個の検索結果が一致している。また，匿名性αは０．７９６，検索結果の再構築可能性ρは０．３５（ただし上位１００の検索結果では０．５９）である。

【０１０４】

また図１７では単語群Ｘの単語として「nagoya,osaka,japan,copenhagen,seoul,tokyo,oslo,unfccc,treaty,nara」が，単語群Ｙの単語として「fernando,street,lugo,gray,walton,madhubani,stoney,mineral,english,nassau,sulphur,spring,durango,clay,rain,buena vista,gatineau,mari,lacey,foster」が特定されている。そして実際の検索条件「kyoto」で検索した場合と比較して，すべての検索結果が一致している。また，匿名性αは０．７６９，検索結果の再構築可能性ρは０．４４２（ただし上位１００の検索結果では０．７６）である。

【０１０５】

また図１８では単語群Ｘの単語として「japan,hiroshima,osaka,nagasaki,oslo,seoul,nagoya,tokyo,copenhagen,treaty」が，単語群Ｙの単語として「columbus,saint-louis,henderson,sidney,murray,roy,wolf,fox,hunter,clarksville,fountain,madison,marsa,monroe,seneca,southside,belvedere,huntingdon,avondale,afonso」が特定されている。そして，実際の検索条件「kyoto」で検索した場合と比較して，１０個中９個の検索結果が一致している。また，匿名性αは０．７８７，検索結果の再構築可能性ρは０．４０９（ただし上位１００の検索結果では０．６６）である。

【０１０６】

さらに，図１９および図２０は実際の検索条件として「nagasaki」を用いており，いずれも強いノイズ（ノイズベクトルが大きい）の場合を示している。また図１９および図２０ではノイズ処理部２３が，クラスタ手法とランダム手法を用いて単語群Ｙを特定している。

【０１０７】

図１９では単語群Ｘの単語として，「iacono,niigata,bombing,bombed,hiroshima,bombs,a-bombing,osaka,sanfecce,hijrah」が，単語群Ｙの単語として「anjar,clark,jennings,lakeland,alexander,marshall,apple,valley,james,belleair,jefferson,band,thompson,ripley,morrison,taft,minneapolis,brooklyn heights,franklin,anderson」が特定されている。そして，実際の検索条件「nagasaki」で検索した場合と比較して，１０個中６個の検索結果が一致している。また，匿名性αは０．８３３，検索結果の再構築可能性ρは０．５３９（ただし上位１００の検索結果では０．４７）である。

【０１０８】

図２０では単語群Ｘの単語として，「devastated,kiel,suburbs,niigata,prefecture,incinerated,bombings,inundated,bombed,hiroshima」が，単語群Ｙの単語として「grants,lakeview,howard,on, woodland,horton,lakeside,rudbar,nosratabad,baker,melrose park,davis, valley, victor, logan, manor, haymana, va, wollongong, bloomfield」が特定されている。そして，実際の検索条件「nagasaki」で検索した場合と比較して，１０個中７個の検索結果が一致している。また，匿名性αは０．８３６，検索結果の再構築可能性ρは０．５３９（ただし上位１００の検索結果では０．５１）である。

【０１０９】

以上のように，図１５乃至図２０の比較結果に基づけば，実際の検索条件を匿名化した上で，検索結果の再構築可能性も一定の精度を有している。とくにウェブサイトを検索する検索エンジンの場合には，検索結果としてせいぜい上位１０位程度しか参照しないことも多い。そうすると，検索結果の再構築可能性も６割以上であるなど，十分に実用に耐えることができる。

【実施例２】

【０１１０】

本発明の情報検索システム１を用いた，別の実施態様として，たとえば検索条件修正処理部２２による単語群Ｘの単語と，ノイズ処理部２３による単語群Ｙの単語とを表示させ，実際の検索条件を推測させる処理を設けることも可能である。この場合，実際の検索条件は，制御端末２が任意に特定をすることで，それに基づいて検索条件修正処理部２２で単語群Ｘの単語を，ノイズ処理部２３で単語群Ｙの単語を，それぞれ特定する。そして単語群Ｘ，単語群Ｙの各単語をユーザ端末４にランダムの順番で表示させることで，それら

の単語から，実際に入力された単語を推測させる，一種のゲーム感覚で，情報検索システム１における効果を体感することも可能である。

【０１１１】

図２１は単語群Ｘの単語，単語群Ｙの単語をそれぞれ表示した状態を示す画面である。そして，図２２は，図２１の状態で分からなかった場合（分からないことを示す操作を受け付けた，または回答として入力を受けた単語が誤っていた場合）に，単語群Ｙの単語を削除し，単語群Ｘの単語のみを表示した状態を示す画面である。図２３は正解を表示した状態を示す画面である。

【０１１２】

このように，本実施例の処理を実行することで，ゲーム感覚で本発明の情報検索システム１の効果を体感することもできる。

【産業上の利用可能性】

【０１１３】

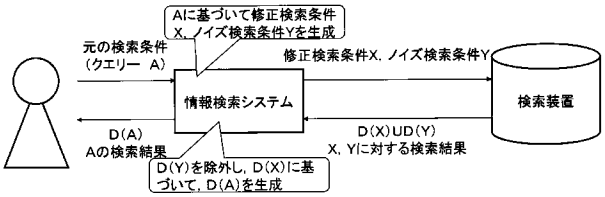
本発明の情報検索システム１を用いることによって，実際の検索条件の特定を困難とする情報検索システム１を発明した。特に，クライスタリング耐性の高いノイズを用いる場合には，クラスタリング耐性を高めることができる。

【符号の説明】

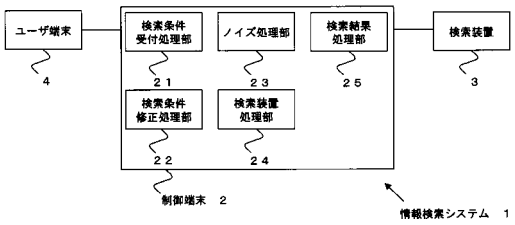
【０１１４】

１：情報検索システム
２：制御端末
３：検索装置
４：ユーザ端末
２１：検索条件受付処理部
２２：検索条件修正処理部
２３：ノイズ処理部
２４：検索装置処理部
２５：検索結果処理部
７０：演算装置
７１：記憶装置
７２：表示装置
７３：入力装置
７４：通信装置

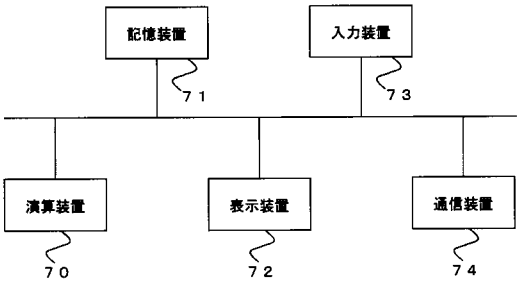
【図１】



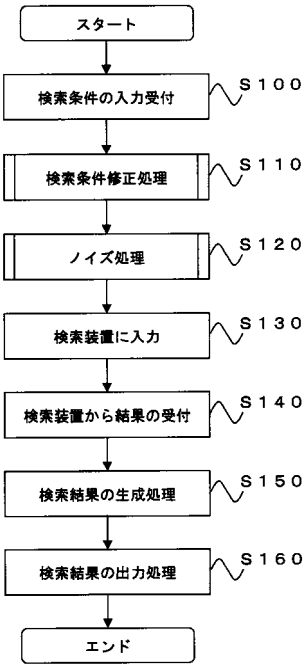
【図２】



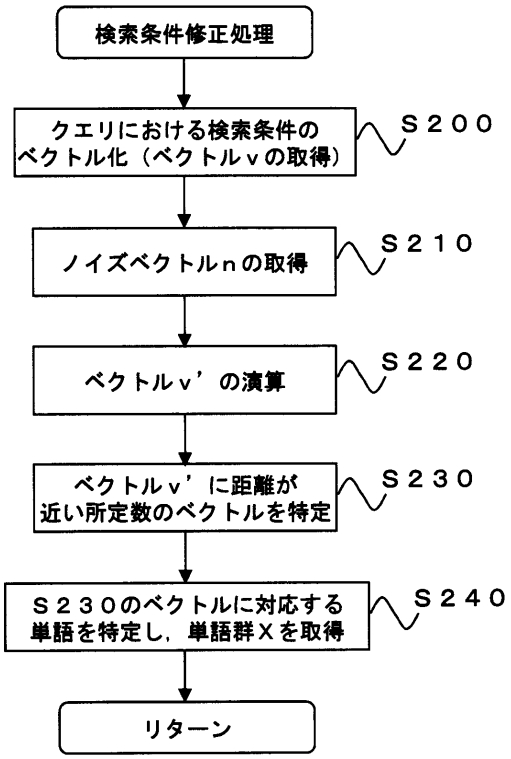
【図３】



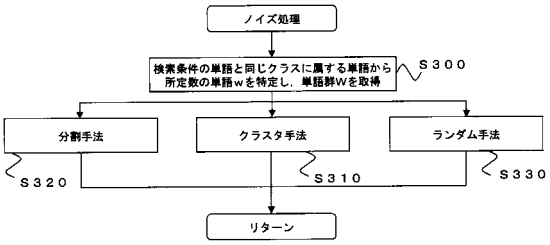
【図４】



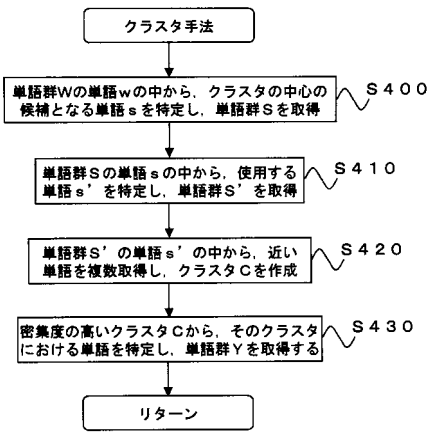
【図５】



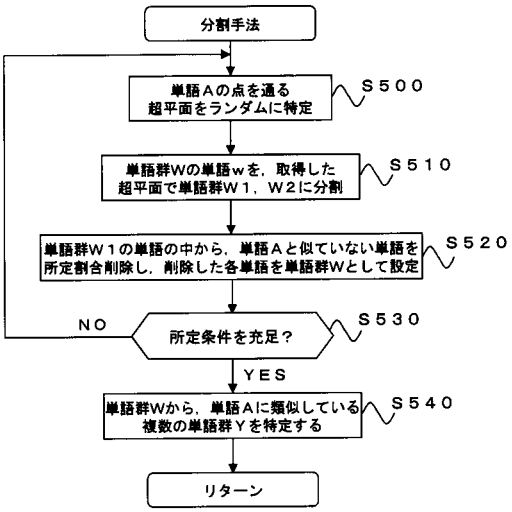
【図６】



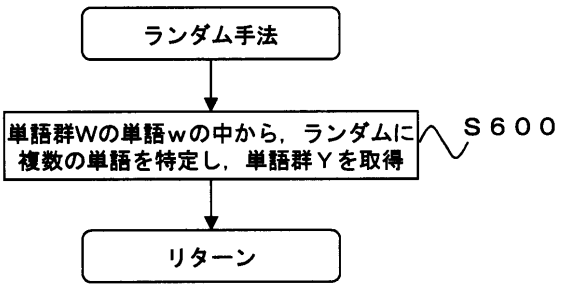
【図７】



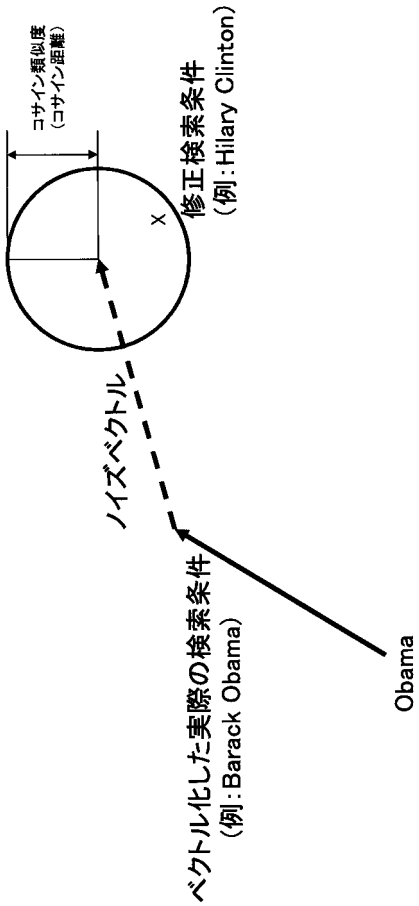
【図８】



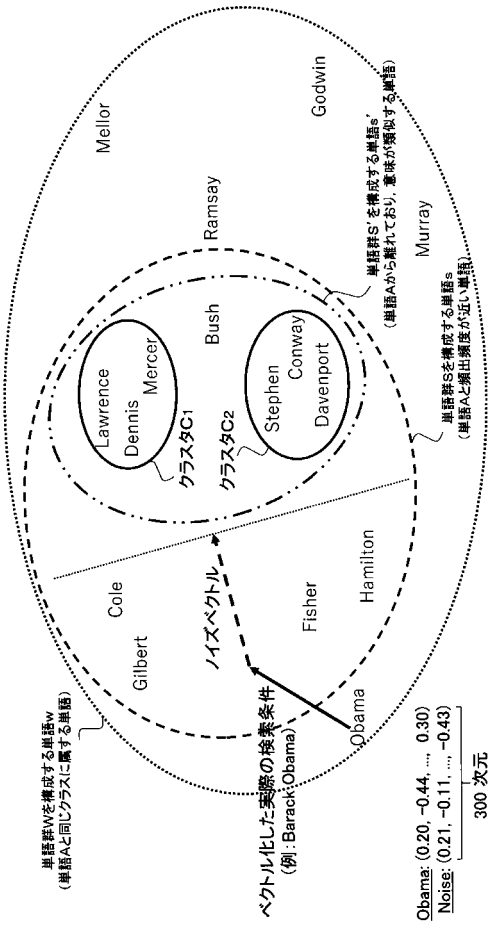
【図９】



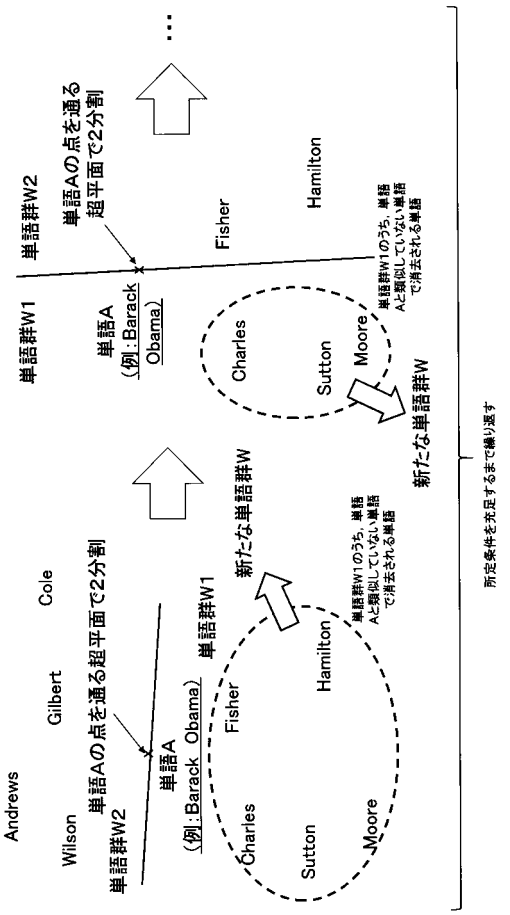
【図１０】



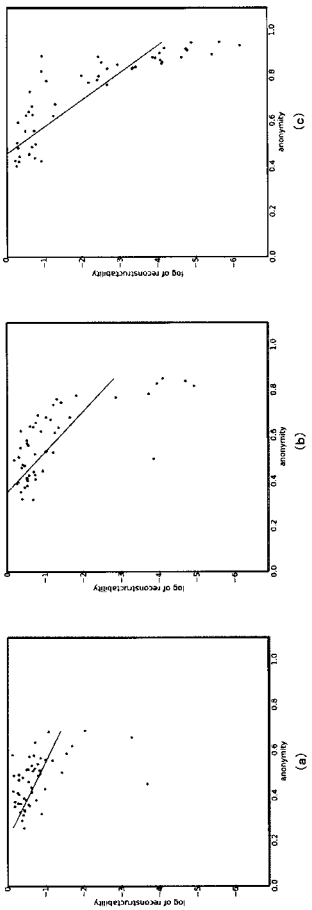
【図１１】



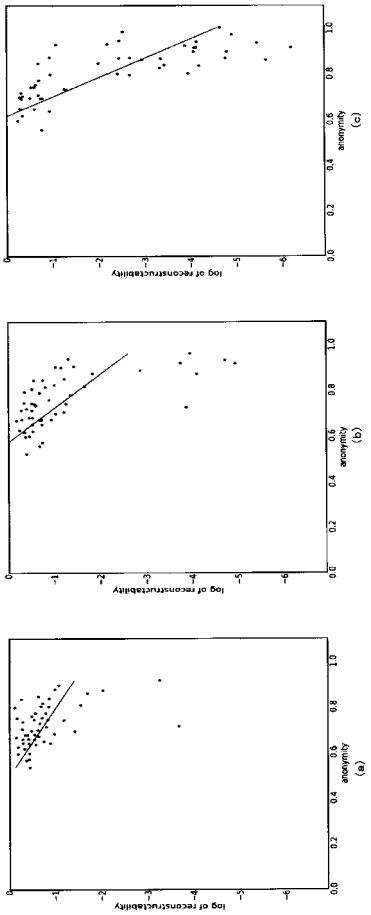
【図１２】



【図１３】



【図１４】



【図１５】

query: kyoto (up to three words available)
noise level of semantically related terms: high-level
number of unrelated distractor terms: 20 distractor method: 2-division+random

Search Results in wikipage titles

	Results of input query		Results of our system
1	Kyoto	E	Kyoto
2	Kyoto Protocol		Kyoto Prefecture
3	Kyoto Prefecture	E	Kyoto Protocol and government action
4	Kyoto Protocol and government action	E	Kyoto Station
5	Kyoto Station	E	Japanese garden
6	Japanese garden	E	Nara Line
7	Nara Line	E	Views on the Kyoto Protocol
8	Views on the Kyoto Protocol	E	Karasuma Line
9	Karasuma Line	E	List of National Treasures of Japan (crafts: others)
10	List of National Treasures of Japan (crafts: others)	E	Kansai dialect

('E' means "exist in the right results of our system")

Note: Ranking is sorted by the frequency of the input query in the text of each wikipage

Anonymity : 0.778 (0.493), Reconstructability : 0.421 (0.71)

Note 1: Value (0.493) in the left parentheses is the anonimity in case of semantically related terms: the actual value is 0.778.
Note 2: Value (0.71) in the right parentheses is the reconstructability in case of top 100 search results: the actual value is 0.421.
Note 3: Go to hyperlink pages for the definitions of the anonymity and reconstructability. (Click each word).

Semantically related terms (10 items), Unrelated distractor terms (20 items)

tokyo, arkansas, pueblo, copenhagen, saitama, conway, hokkaido, john, rosario, owen sound, armenia, patti, nagoya, osaka, lyons, laporte, knowle west, columbus, north berwick, surat, japan, patterson, millbrook, san diego, kansai, gill, seoul, walnut, fukuoka, chiba

【図１６】

query: kyoto (up to three words available)
noise level of semantically related terms: high-level
number of unrelated distractor terms: 20 distractor method: 2-division+random

Search Results in wikipage titles

	Results of input query		Results of our system
1	Kyoto	E	Kyoto
2	Kyoto Protocol		Kyoto Prefecture
3	Kyoto Prefecture	E	Kyoto Station
4	Kyoto Protocol and government action		Japanese garden
5	Kyoto Station	E	Nara Line
6	Japanese garden	E	Karasuma Line
7	Nara Line	E	List of National Treasures of Japan (crafts: others)
8	Views on the Kyoto Protocol		Kansai dialect
9	Karasuma Line	E	Kyoto University
10	List of National Treasures of Japan (crafts: others)	E	Japan National Route 1

('E' means "exist in the right results of our system")

Note: Ranking is sorted by the frequency of the input query in the text of each wikipage

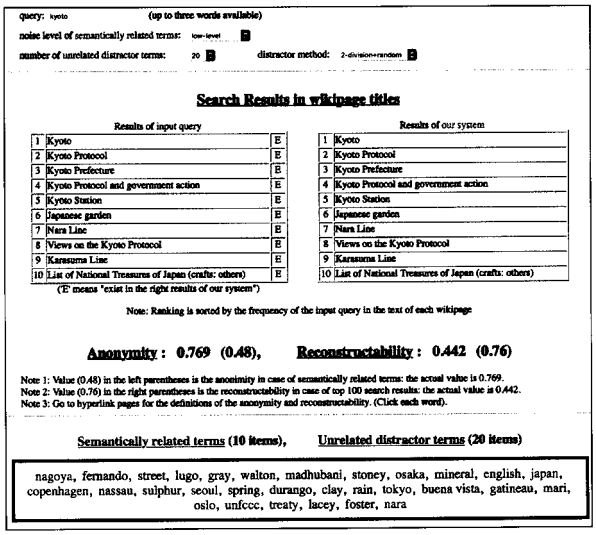
Anonymity : 0.796 (0.624), Reconstructability : 0.35 (0.59)

Note 1: Value (0.624) in the left parentheses is the anonimity in case of semantically related terms: the actual value is 0.796.
Note 2: Value (0.59) in the right parentheses is the reconstructability in case of top 100 search results: the actual value is 0.35.
Note 3: Go to hyperlink pages for the definitions of the anonymity and reconstructability. (Click each word).

Semantically related terms (10 items), Unrelated distractor terms (20 items)

ina, valley, västerås, nagoya, islampur, seoul, rudbar, qasemabad, cmom, waseda, osaka, habibabad, alexander, wollongong, first, mehrdasht, humboldt, price, lara, perth, ginza, joad, tokyo, hayden, dauphin, hat, kuhsar, jahanabad, yokohama, nosratabad

【図１７】



query: kyoto (up to three words available)

noise level of semantically related terms: low-level

number of unrelated distractor terms: 20 distractor method: 2-division+random

Search Results in wikipage titles

	Results of input query		Results of our system
1	Kyoto	E	Kyoto
2	Kyoto Protocol	E	Kyoto Protocol
3	Kyoto Prefecture	E	Kyoto Prefecture
4	Kyoto Protocol and government action	E	Kyoto Protocol and government action
5	Kyoto Station	E	Kyoto Station
6	Japanese garden	E	Japanese garden
7	Nara Line	E	Nara Line
8	Views on the Kyoto Protocol	E	Views on the Kyoto Protocol
9	Karasuma Line	E	Karasuma Line
10	List of National Treasures of Japan (crafts: others)	E	List of National Treasures of Japan (crafts: others)

('E' means "exist in the right results of our system")

Note: Ranking is sorted by the frequency of the input query in the text of each wikipage

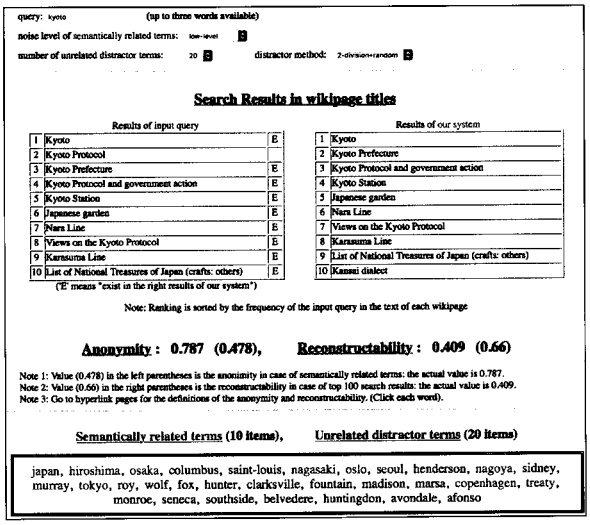
Anonymity : 0.769 (0.48), Reconstructability : 0.442 (0.76)

Note 1: Value (0.48) in the left parentheses is the anonimity in case of semantically related terms: the actual value is 0.769.
Note 2: Value (0.76) in the right parentheses is the reconstructability in case of top 100 search results: the actual value is 0.442.
Note 3: Go to hyperlink pages for the definitions of the anonymity and reconstructability. (Click each word).

Semantically related terms (10 items), Unrelated distractor terms (20 items)

nagoya, fernando, street, lugo, gray, walton, madhubani, stoney, osaka, mineral, english, japan, copenhagen, nassau, sulphur, seoul, spring, durango, clay, rain, tokyo, buena vista, gatineau, mari, oslo, unfccc, treaty, lacey, foster, nara

【図１８】



query: kyoto (up to three words available)

noise level of semantically related terms: low-level

number of unrelated distractor terms: 20 distractor method: 2-division+random

Search Results in wikipage titles

	Results of input query		Results of our system
1	Kyoto	E	Kyoto
2	Kyoto Protocol		Kyoto Prefecture
3	Kyoto Prefecture	E	Kyoto Protocol and government action
4	Kyoto Protocol and government action	E	Kyoto Station
5	Kyoto Station	E	Japanese garden
6	Japanese garden	E	Nara Line
7	Nara Line	E	Views on the Kyoto Protocol
8	Views on the Kyoto Protocol	E	Karasuma Line
9	Karasuma Line	E	List of National Treasures of Japan (crafts: others)
10	List of National Treasures of Japan (crafts: others)	E	Kansai dialect

('E' means "exist in the right results of our system")

Note: Ranking is sorted by the frequency of the input query in the text of each wikipage

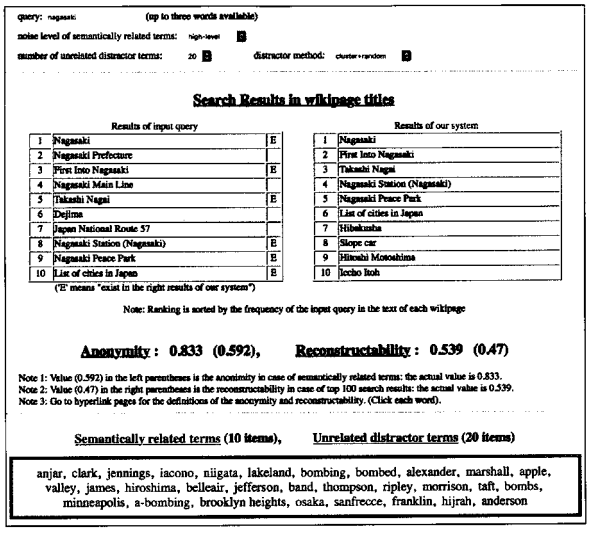
Anonymity : 0.787 (0.478), Reconstructability : 0.409 (0.66)

Note 1: Value (0.478) in the left parentheses is the anonimity in case of semantically related terms: the actual value is 0.787.
Note 2: Value (0.66) in the right parentheses is the reconstructability in case of top 100 search results: the actual value is 0.409.
Note 3: Go to hyperlink pages for the definitions of the anonymity and reconstructability. (Click each word).

Semantically related terms (10 items), Unrelated distractor terms (20 items)

japan, hiroshima, osaka, columbus, saint-louis, nagasaki, oslo, seoul, henderson, nagoya, sidney, murray, tokyo, roy, wolf, fox, hunter, clarksville, fountain, madison, marsa, copenhagen, treaty, monroe, seneca, southside, belvedere, huntingdon, avondale, afonso

【図１９】



query: nagasaki (up to three words available)

noise level of semantically related terms: high-level

number of unrelated distractor terms: 20 distractor method: cluster+random

Search Results in wikipage titles

	Results of input query		Results of our system
1	Nagasaki	E	Nagasaki
2	Nagasaki Prefecture		First Into Nagasaki
3	First Into Nagasaki	E	Takashi Nagai
4	Nagasaki Main Line		Nagasaki Station (Nagasaki)
5	Takashi Nagai	E	Nagasaki Peace Park
6	Dejima		List of cities in Japan
7	Japan National Route 57		Hibakusha
8	Nagasaki Station (Nagasaki)	E	Slope car
9	Nagasaki Peace Park	E	Hitoshi Motoshima
10	List of cities in Japan	E	Iccho Itoh

('E' means "exist in the right results of our system")

Note: Ranking is sorted by the frequency of the input query in the text of each wikipage

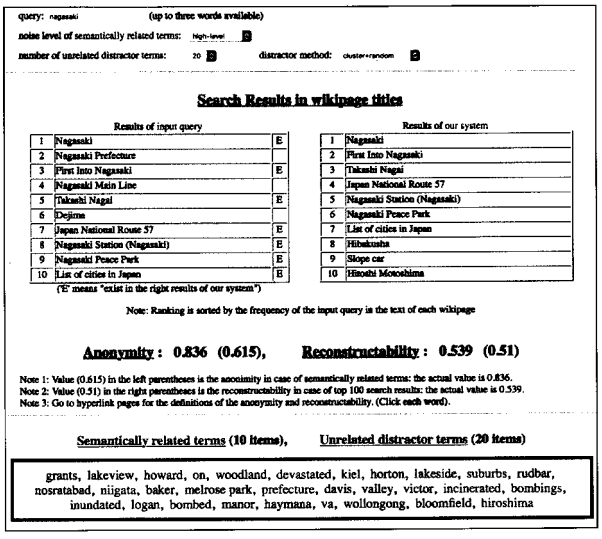
Anonymity : 0.833 (0.592), Reconstructability : 0.539 (0.47)

Note 1: Value (0.592) in the left parentheses is the anonimity in case of semantically related terms: the actual value is 0.833.
Note 2: Value (0.47) in the right parentheses is the reconstructability in case of top 100 search results: the actual value is 0.539.
Note 3: Go to hyperlink pages for the definitions of the anonymity and reconstructability. (Click each word).

Semantically related terms (10 items), Unrelated distractor terms (20 items)

anjar, clark, jennings, iacono, niigata, lakeland, bombing, bombed, alexander, marshall, apple, valley, james, hiroshima, belleair, jefferson, band, thompson, ripley, morrison, taft, bombs, minneapolis, a-bombing, brooklyn heights, osaka, sanfrecce, franklin, hijrah, anderson

【図２０】



query: nagasaki (up to three words available)

noise level of semantically related terms: high-level

number of unrelated distractor terms: 20 distractor method: cluster+random

Search Results in wikipage titles

	Results of input query		Results of our system
1	Nagasaki	E	Nagasaki
2	Nagasaki Prefecture		First Into Nagasaki
3	First Into Nagasaki	E	Takashi Nagai
4	Nagasaki Main Line		Japan National Route 57
5	Takashi Nagai	E	Nagasaki Station (Nagasaki)
6	Dejima		Nagasaki Peace Park
7	Japan National Route 57	E	List of cities in Japan
8	Nagasaki Station (Nagasaki)	E	Hibakusha
9	Nagasaki Peace Park	E	Slope car
10	List of cities in Japan	E	Hitoshi Motoshima

('E' means "exist in the right results of our system")

Note: Ranking is sorted by the frequency of the input query in the text of each wikipage

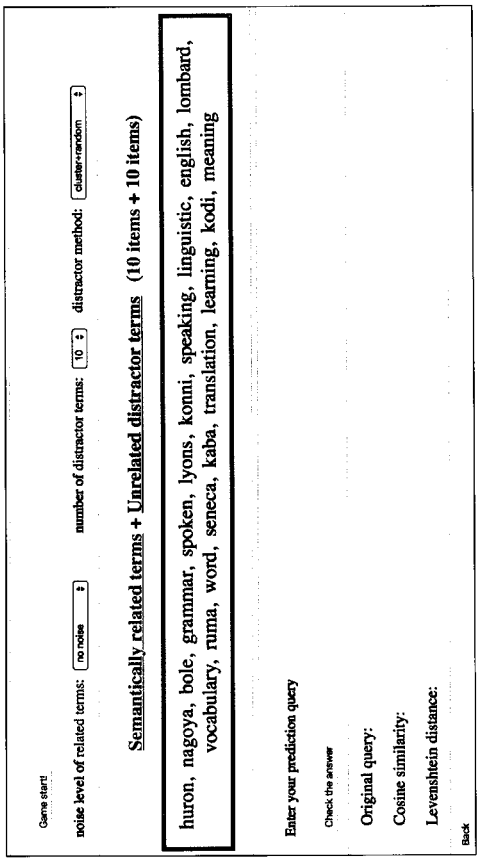
Anonymity : 0.836 (0.615), Reconstructability : 0.539 (0.51)

Note 1: Value (0.615) in the left parentheses is the anonimity in case of semantically related terms: the actual value is 0.836.
Note 2: Value (0.51) in the right parentheses is the reconstructability in case of top 100 search results: the actual value is 0.539.
Note 3: Go to hyperlink pages for the definitions of the anonymity and reconstructability. (Click each word).

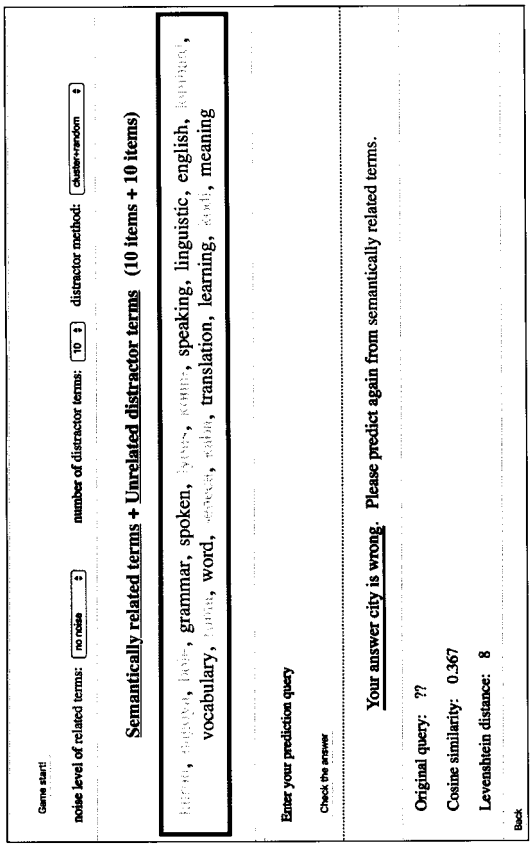
Semantically related terms (10 items), Unrelated distractor terms (20 items)

grants, lakeview, howard, on, woodland, devastated, kiel, horton, lakeside, suburbs, rudbar, nosratabad, niigata, baker, melrose park, prefecture, davis, valley, victor, incinerated, bombings, inundated, logan, bombed, manor, haymana, va, wollongong, bloomfield, hiroshima

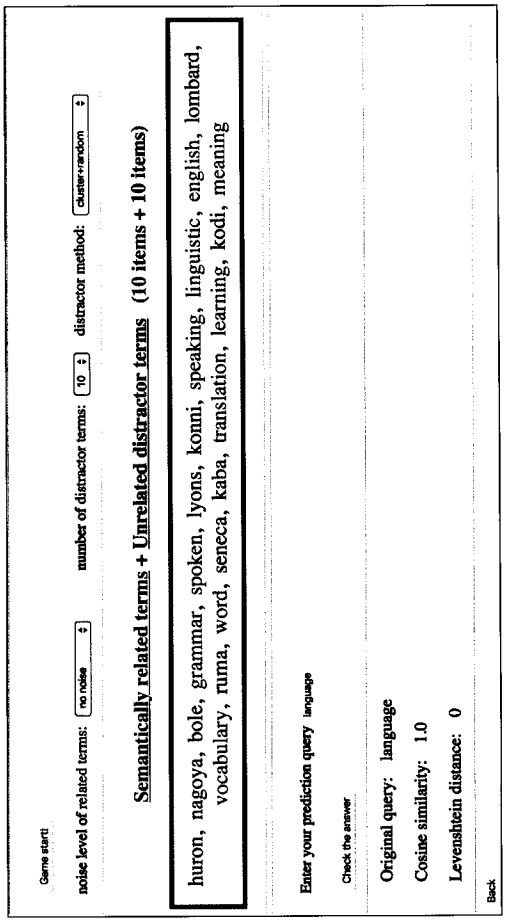
【図２１】



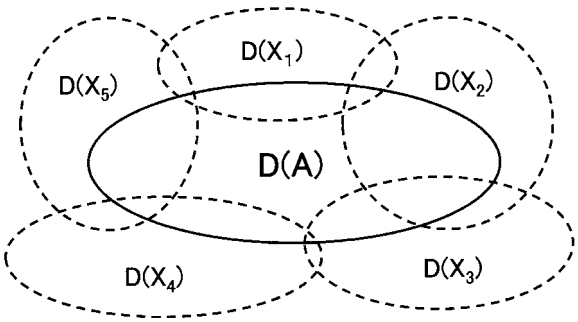
【図２２】



【図２３】



【図２４】



フロントページの続き

(72)発明者　ボレガラ　ダヌシカ
東京都千代田区一ツ橋二丁目１番２号　大学共同利用機関法人情報・システム研究機構　国立情報学研究所内