(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4944405号
(P4944405)
(45)発行日　平成24年5月30日(2012.5.30)　(24)登録日　平成24年3月9日(2012.3.9)

(51)Int.Cl.　F I
*G06F 17/30 (2006.01)*　G06F 17/30 414B
G06F 17/30 220Z
G06F 17/30 170A

請求項の数 21　外国語出願　(全 38 頁)

(21)出願番号　特願2005-216529(P2005-216529)
(22)出願日　平成17年7月26日(2005.7.26)
(65)公開番号　特開2006-48685(P2006-48685A)
(43)公開日　平成18年2月16日(2006.2.16)
審査請求日　平成20年7月23日(2008.7.23)
(31)優先権主張番号　10/900,055
(32)優先日　平成16年7月26日(2004.7.26)
(33)優先権主張国　米国(US)

(73)特許権者　505281067
グーグル　インコーポレイテッド
GOOGLE　INC.
アメリカ合衆国　カリフォルニア州　94043　マウンテン　ビュー　アンフィシアター　パークウェイ　1600
(74)代理人　100077539
弁理士　飯塚　義仁
(72)発明者　アンナ　エル．パターソン
アメリカ合衆国　カリフォルニア　95120，サンノゼ，ソーンツリー　コート　1127

審査官　吉田　誠

最終頁に続く

(54)【発明の名称】情報検索システムにおけるフレーズに基づくインデックス化方法

(57)【特許請求の範囲】
【請求項１】
文書コレクション内の文書をインデックス化する方法であって、前記各文書は対応する識別子を有しており、
コンピュータシステム内のプロセッサの処理によって、前記文書コレクション内の複数文書から収集されたフレーズのリストを参照し、前記文書コレクション内の或る文書について、該文書内に存在するフレーズを、前記フレーズのリストから識別するステップと、
コンピュータシステム内のプロセッサの処理によって、前記文書内の前記識別された各フレーズ毎に、同じ前記文書内に存在する関連フレーズを識別するステップと、ここで、該識別するステップは、前記文書内の前記識別されたフレーズのうち第１及び第２フレーズ（$g_j$，$g_k$）について、第１フレーズ（$g_j$）に関する第２フレーズ（$g_k$）の情報ゲイン（Ｉ）を、両フレーズ（$g_j$，$g_k$）の共出現率の期待値Ｅ(ｊ，ｋ)及び該両フレーズ（$g_j$，$g_k$）の実際の共出現率Ａ(ｊ，ｋ)の関数として、決定し、前記第１フレーズ（$g_j$）に関する前記第２フレーズ（$g_k$）の前記情報ゲイン（Ｉ）が所定の閾値を超えるとき、前記第２フレーズ（$g_k$）を前記第１フレーズ（$g_j$）の前記関連フレーズ($g_k$)として識別することを含み、
コンピュータシステム内のプロセッサの処理によって、前記文書内の前記識別された各フレーズ($g_j$)毎に、該識別された各フレーズ($g_j$)のポスティングリスト内に、該文書の前記識別子と該文書内に存在する前記識別された前記関連フレーズ($g_k$)を示す情報とを格納することにより、該文書をインデックス化するステップと、

を備える方法。
【請求項２】
コンピュータシステム内のプロセッサの処理によって、前記文書コレクション内の複数文書からフレーズを収集し、収集したフレーズにより前記リストを構築するステップを更に備える請求項１の方法。
【請求項３】
該文書内に存在するフレーズを、前記フレーズのリストから識別する前記ステップは、
コンピュータシステム内のプロセッサの処理によって、複数ワードを抽出するフレーズ窓によって前記文書内の連続する複数ワードを抽出し、該フレーズ窓内の複数ワードから１以上の候補フレーズを特定し、該フレーズ窓を順次移動させることで更に候補フレーズを特定すること、を含む請求項１又は２の方法。
【請求項４】
前記所定の閾値が略１００である請求項１乃至３のいずれかの方法。
【請求項５】
前記第１フレーズ（$g_j$）に関する前記第２フレーズ（$g_k$）の前記情報ゲイン（Ｉ）は、Ｉ（ｊ、ｋ）＝Ａ（ｊ、ｋ）／Ｅ（ｊ、ｋ）である、請求項１乃至４のいずれかの方法。
【請求項６】
前記格納するステップは、前記文書内に存在する前記各関連フレーズ毎に当該関連フレーズを１次関連フレーズとして、該１次関連フレーズに関連する該文書内に存在する少なくとも１つの２次関連フレーズを示す情報を、前記ポスティングリスト内に更に格納する、請求項１乃至５のいずれかの方法。
【請求項７】
前記２次関連フレーズを示す情報は、それに対応する前記１次関連フレーズの前記情報ゲインが減少する順に格納される、請求項６の方法。
【請求項８】
コンピュータシステム内のプロセッサの処理によって、前記１次関連フレーズとそれに対応する前記２次関連フレーズが前記文書内に存在するならば１次主題として判定し、前記１次関連フレーズに対応する前記２次関連フレーズが前記文書内に存在しないならば２次主題として判定するステップを更に備える、
請求項６又は７の方法。
【請求項９】
前記文書内の前記識別された各フレーズ毎に、該識別されたフレーズが、或る目標文書を指し示すために前記文書内でハイパーリンクのアンカーテキストとなっているかを判定するステップと、
前記文書内の前記識別された各フレーズ毎に、該識別されたフレーズが前記アンカーテキストとなっているとの判定に応じて、該識別されたフレーズの関連フレーズに基づき該識別されたフレーズのリンクスコアを決定するステップと、
前記文書内の前記識別された各フレーズ毎に、前記識別されたフレーズの前記ポスティングリスト内に前記リンクスコアを格納し前記文書に関連付けるステップと
を更に具備する、請求項１乃至８のいずれかの方法。
【請求項１０】
コンピュータシステム内のプロセッサの処理によって、前記文書内の前記関連フレーズの数をカウントするステップを更に具備する、請求項１乃至９のいずれかの方法。
【請求項１１】
文書コレクション内の文書をインデックス化する方法であって、前記各文書は対応する識別子を有しており、
コンピュータシステムにおいてアクセス可能に、フレーズのリストを提供するステップであって、前記リスト内の各フレーズは、前記文書コレクション内において所定最低回数は出現し、かつ、少なくとも１つの他のフレーズを予測するものであり、前記他のフレー

ズ（$g_k$）を予測するフレーズ（$g_j$）とは、当該フレーズ（$g_j$）に関する前記他のフレーズ（$g_k$）の情報ゲイン（Ｉ）が所定の閾値を超えるものであり、前記情報ゲイン（Ｉ）とは、両フレーズ（$g_j$，$g_k$）の共出現率の期待値Ｅ(ｊ，ｋ)及び該両フレーズ（$g_j$，$g_k$）の実際の共出現率Ａ(ｊ，ｋ)の関数であり、

コンピュータシステム内のプロセッサの処理によって、前記文書コレクション内の複数の文書にアクセスするステップと、

コンピュータシステム内のプロセッサの処理によって、アクセスした文書それぞれについて、該文書内に存在するフレーズを、前記フレーズのリストから識別するステップと、

コンピュータシステム内のプロセッサの処理によって、前記文書内の前記識別された各フレーズ毎に、該識別された各フレーズのポスティングリスト内に、該文書の前記識別子を格納することにより、該文書をインデックス化するステップと、

を備える方法。

【請求項１２】

前記所定の閾値が略１．５である請求項１１の方法。

【請求項１３】

文書コレクション内の文書をインデックス化するための、コンピュータを装備したシステムであって、前記各文書は対応する識別子を有しており、

フレーズ用のインデックスであって、記憶媒体内に格納されており、かつ、フレーズリスト及び複数のフレーズポスティングリストを含み、各フレーズポスティングリストは、前記フレーズリスト内のフレーズに関連付けられている、前記インデックスと、

インデックス化システムであって、

前記フレーズリストを参照し、前記文書コレクション内の或る文書について、該文書内に存在するフレーズを、前記フレーズリストから識別し、

前記文書内の前記識別された各フレーズ毎に、同じ前記文書内に存在する関連フレーズを識別し、ここで、該関連フレーズを識別することは、前記文書内の前記識別されたフレーズのうち第１及び第２フレーズ（$g_j$，$g_k$）について、第１フレーズ（$g_j$）に関する第２フレーズ（$g_k$）の情報ゲイン（Ｉ）を、両フレーズ（$g_j$，$g_k$）の共出現率の期待値Ｅ(ｊ，ｋ)及び該両フレーズ（$g_j$，$g_k$）の実際の共出現率Ａ(ｊ，ｋ)の関数として、決定し、前記第１フレーズ（$g_j$）に関する前記第２フレーズ（$g_k$）の前記情報ゲイン（Ｉ）が所定の閾値を超えるとき、前記第２フレーズ（$g_k$）を前記第１フレーズ（$g_j$）の前記関連フレーズ($g_k$)として識別することを含み、

前記文書内の前記識別された各フレーズ($g_j$)毎に、前記フレーズ用のインデックス内の、該識別されたフレーズに関連付けられたポスティングリスト内に、該文書の前記識別子と該文書内に存在する前記識別された前記関連フレーズ($g_k$)を示す情報とを格納することにより、該文書をインデックス化する、

ように構成された前記インデックス化システムと

を備えるシステム。

【請求項１４】

文書コレクション内の文書をインデックス化するための、コンピュータシステム内のプロセッサによって実行されるプログラムであって、前記各文書は対応する識別子を有しており、該プロセッサに、

前記文書コレクション内の複数文書から収集されたフレーズのリストを参照し、前記文書コレクション内の或る文書について、該文書内に存在するフレーズを、前記フレーズのリストから識別する手順と、

前記文書内の前記識別された各フレーズ毎に、同じ前記文書内に存在する関連フレーズを識別する手順と、ここで、該識別するステップは、前記文書内の前記識別されたフレーズのうち第１及び第２フレーズ（$g_j$，$g_k$）について、第１フレーズ（$g_j$）に関する第２フレーズ（$g_k$）の情報ゲイン（Ｉ）を、両フレーズ（$g_j$，$g_k$）の共出現率の期待値Ｅ(ｊ，ｋ)及び該両フレーズ（$g_j$，$g_k$）の実際の共出現率Ａ(ｊ，ｋ)の関数として、決定し、前記第１フレーズ（$g_j$）に関する前記第２フレーズ（$g_k$）の前記情報ゲイン（

Ｉ）が所定の閾値を超えるとき、前記第２フレーズ（$g_k$）を前記第１フレーズ（$g_j$）の前記関連フレーズ($g_k$)として識別することを含み、
前記文書内の前記識別された各フレーズ($g_j$)毎に、該識別された各フレーズ($g_j$)のポスティングリスト内に、該文書の前記識別子と該文書内に存在する前記識別された前記関連フレーズ($g_k$)を示す情報とを格納することにより、該文書をインデックス化する手順と、
を実行させるためのプログラム。
【請求項１５】
文書コレクション内の文書をインデックス化するための、コンピュータを装備したシステムであって、前記各文書は対応する識別子を有しており、
フレーズ用のインデックスであって、記憶媒体内に格納されており、かつ、フレーズリスト及び複数のフレーズポスティングリストを含む、前記インデックスと、ここで、
前記フレーズリスト内の各フレーズは、前記文書コレクション内において所定最低回数は出現し、かつ、少なくとも１つの他のフレーズを予測するものであり、前記他のフレーズ（$g_k$）を予測するフレーズ（$g_j$）とは、当該フレーズ（$g_j$）に関する前記他のフレーズ（$g_k$）の情報ゲイン（Ｉ）が所定の閾値を超えるものであり、前記情報ゲイン（Ｉ）とは、両フレーズ（$g_j$，$g_k$）の共出現率の期待値Ｅ(ｊ，ｋ)及び該両フレーズ（$g_j$，$g_k$）の実際の共出現率Ａ(ｊ，ｋ)の関数であり、また、
前記各フレーズポスティングリストは、前記フレーズリスト内のフレーズに関連付けられており、
インデックス化システムであって、
前記文書コレクション内の複数の文書にアクセスし、
アクセスした文書それぞれについて、該文書内に存在するフレーズを、前記フレーズのリストから識別し、
前記文書内の前記識別された各フレーズ毎に、該識別された各フレーズのポスティングリスト内に、該文書の前記識別子を格納することにより、該文書をインデックス化する、
ように構成された前記インデックス化システムと
を備えるシステム。
【請求項１６】
文書コレクション内の文書をインデックス化するための、コンピュータシステム内のプロセッサによって実行されるプログラムであって、前記各文書は対応する識別子を有しており、該プロセッサに、
前記文書コレクション内の複数文書からフレーズを収集して、収集したフレーズでなるフレーズのリストを構築する手順と、ここで、前記リスト内の各フレーズは、前記文書コレクション内において所定最低回数は出現し、かつ、少なくとも１つの他のフレーズを予測するものであり、前記他のフレーズ（$g_k$）を予測するフレーズ（$g_j$）とは、当該フレーズ（$g_j$）に関する前記他のフレーズ（$g_k$）の情報ゲイン（Ｉ）が所定の閾値を超えるものであり、前記情報ゲイン（Ｉ）とは、両フレーズ（$g_j$，$g_k$）の共出現率の期待値Ｅ(ｊ，ｋ)及び該両フレーズ（$g_j$，$g_k$）の実際の共出現率Ａ(ｊ，ｋ)の関数であり、
前記文書コレクション内の複数の文書にアクセスする手順と、
アクセスした文書それぞれについて、該文書内に存在するフレーズを、前記フレーズのリストから識別する手順と、
前記文書内の前記識別された各フレーズ毎に、該識別された各フレーズのポスティングリスト内に、該文書の前記識別子を格納することにより、該文書をインデックス化する手順と、
を実行させるためのプログラム。
【請求項１７】
文書をインデックス化するためにコンピュータによって実行される方法であって、前記文書は対応する識別子を有しており、

フレーズのリストを格納するステップと、ここで、前記リスト内には各フレーズ（$g_j$）に対応する関連フレーズを含んでおり、各フレーズ（$g_j$）毎に、該フレーズ（$g_j$）に関する前記関連フレーズ（$g_k$）の情報ゲイン（Ｉ）は所定の閾値を超えており、該情報ゲイン（Ｉ）は、両フレーズ（$g_j$，$g_k$）の共出現率の期待値Ｅ(ｊ，ｋ)及び該両フレーズ（$g_j$，$g_k$）の実際の共出現率Ａ(ｊ，ｋ)の関数であり、

インデックス記憶装置内に複数のフレーズポスティングリストを格納するステップと、ここで、前記各フレーズポスティングリストは、前記フレーズのリスト内の各フレーズに対応付けられており、

コンピュータシステム内のプロセッサの処理によって、当該フレーズの少なくとも１つのインスタンスが前記文書内に存在している、前記フレーズのリスト内の１つのフレーズを、識別するステップと、

コンピュータシステム内のプロセッサの処理によって、前記識別されたフレーズに対応する前記関連フレーズのインスタンスが前記同じ文書内に存在している当該関連フレーズを示す情報を、前記識別されたフレーズの前記フレーズポスティングリスト内に格納するステップと、

を備える方法。

【請求項１８】

文書コレクション内の文書をインデックス化するためにコンピュータによって実行される方法であって、前記各文書は対応する識別子を有しており、

コンピュータシステム内のプロセッサの処理によって、前記文書コレクション内の複数文書から収集されたフレーズのリストを参照し、前記文書コレクション内の或る文書について、該文書内に存在するフレーズを、前記フレーズのリストから識別するステップと、

コンピュータシステム内のプロセッサの処理によって、前記文書内の前記識別された各フレーズ毎に、同じ前記文書内に存在する関連フレーズを識別するステップと、

コンピュータシステム内のプロセッサの処理によって、前記文書内の前記識別された各フレーズ毎に、該識別された各フレーズのポスティングリスト内に、該文書の前記識別子と該文書内に存在する前記関連フレーズを示す情報とを格納することにより、該文書をインデックス化するステップと、

コンピュータシステム内のプロセッサの処理によって、前記文書内の前記識別された各フレーズ毎に、該識別されたフレーズが、或る目標文書を指し示すために前記文書内でハイパーリンクのアンカーテキストとなっているかを判定するステップと、

前記文書内の前記識別された各フレーズ毎に、該識別されたフレーズが前記アンカーテキストとなっているとの判定に応じて、該識別されたフレーズの関連フレーズに基づき該識別されたフレーズのリンクスコアを決定するステップと、

前記文書内の前記識別された各フレーズ毎に、前記識別されたフレーズの前記ポスティングリスト内に前記リンクスコアを格納し前記文書に関連付けるステップと

を具備する方法。

【請求項１９】

前記識別されたフレーズのリンクスコアを決定する前記ステップは、前記目標文書内に存在する前記識別されたフレーズの関連フレーズに基づき、該識別されたフレーズのインリンクスコアを決定する、請求項１８の方法。

【請求項２０】

前記識別されたフレーズのリンクスコアを決定する前記ステップは、前記識別されたフレーズと同じ前記文書内に存在する該識別されたフレーズの関連フレーズに基づき、該識別されたフレーズのアウトリンクスコアを決定する、請求項１８の方法。

【請求項２１】

文書コレクション内の文書をインデックス化するための、コンピュータを装備したシステムであって、前記各文書は対応する識別子を有しており、

フレーズ用のインデックスであって、記憶媒体内に格納されており、かつ、フレーズリスト及び複数のフレーズポスティングリストを含み、前記各フレーズポスティングリスト

は、前記フレーズリスト内のフレーズに対応付けられている、前記インデックスと、
インデックス化システムであって、
前記文書コレクション内の或る文書について、該文書内に存在するフレーズを、前記フレーズリストから識別し、
前記文書内の前記識別された各フレーズ毎に、同じ前記文書内に存在する関連フレーズを識別し、
前記文書内の前記識別された各フレーズ毎に、前記インデックス内の該識別された各フレーズに対応するポスティングリスト内に、該文書の前記識別子と該文書内に存在する前記関連フレーズを示す情報とを格納することにより、該文書をインデックス化し、
前記文書内の前記識別された各フレーズ毎に、該識別されたフレーズが、或る目標文書を指し示すために前記文書内でハイパーリンクのアンカーテキストとなっているかを判定し、
前記文書内の前記識別された各フレーズ毎に、該識別されたフレーズが前記アンカーテキストとなっているとの判定に応じて、該識別されたフレーズの関連フレーズに基づき該識別されたフレーズのリンクスコアを決定し、
前記文書内の前記識別された各フレーズ毎に、前記識別されたフレーズの前記ポスティングリスト内に前記リンクスコアを格納し前記文書に関連付ける、
ように構成された前記インデックス化システムと、
を備えるシステム。

【発明の詳細な説明】

【技術分野】

【０００１】

（関連出願の相互参照）

この出願は、下記の係属中の出願に関連している。

出願日２００４年７月２６日、出願番号第１０／ｘｘｘ、ｘｘｘ号、名称「情報検索システムにおけるフレーズ同定化」、

出願日２００４年７月２６日、出願番号第１０／ｘｘｘ、ｘｘｘ号、名称「情報検索システムにおけるフレーズに基づく検索」、

出願日２００４年７月２６日、出願番号第１０／ｘｘｘ、ｘｘｘ号、名称「情報検索システムにおけるフレーズに基づく検索の個人化」、

出願日２００４年７月２６日、出願番号第１０／ｘｘｘ、ｘｘｘ号、名称「フレーズを用いる検索結果の自動ソート生成」、

出願日２００４年７月２６日、出願番号第１０／ｘｘｘ、ｘｘｘ号、名称「フレーズに基づく文書説明の生成」および、

出願日２００４年７月２６日、出願番号第１０／ｘｘｘ、ｘｘｘ号、名称「情報検索システムにおけるフレーズに基づく複製文書の検出」、

これらは全て共に所有されており、本明細書中に引用して組み込む。

【０００２】

本発明は、インターネット等の大規模コーパスにおける文書をインデックス化、検索、およびソートするための情報検索システムに関する。

【背景技術】

【０００３】

検索エンジンと一般に呼ばれる情報検索システムは、現在では、インターネット等の大規模で、多様な、肥大化するコーパスにおいて情報を探し出すために不可欠なツールである。一般に、検索エンジンは、各文書内にある個々の言葉に文書（または「ページ」）を関連させるインデックスを作成する。文書は、幾つかのクエリー用語を含むクエリー（問い）に応じて検索され、普通は、その文書内にある幾つかのクエリー用語を有していることに基づいて検索される。次いで、検索文書は、クエリー用語、ホストドメイン、リンク解析等の発生頻度等の他の統計的指標に従ってランク付けされる。次いで、検索された文書は、それ以上グループ化または階層化せずに、ランク付けした順序でユーザーに提示する

のが典型的である。場合によっては、文書のテキストの選択部分を提示してユーザーに文書の内容を一瞥させる。

【０００４】

クエリー用語の直接的な「ブール」一致には限界があるのは周知であり、特に、クエリー用語をもたないが関連ワードをもつ文書を識別しない。例えば、普通のブールシステムでは、「Australian Shepherds（オーストラリアンシェパード）」に関する検索は、正確なクエリー用語を持たないボーダーコリー等の他の牧羊犬に関する文書を返さない。というよりも、このようなシステムは、Australia（オーストラリア）に関する文書（犬に関するものを何も含まない）、および一般的な「shepherds（シェパード）」に関する文書を検索し、高いランクにしがちである。

【０００５】

ここでの問題は、従来システムが、概念ではなく、個々の用語に基づいて文書をインデックス化することにある。概念は、「Australian Shepherds」、「President of the United States（米国大統領）」、または「Sundance Film Festival（サンダンス映画フェスティバル）」等のフレーズで表わされることが多い。良くても、旧来のシステムには、所定の非常に限られた「既知」フレーズに関する文書をインデックス化するものもあるが、人間のオペレータが選択するのが普通である。フレーズのインデックス化は普通避けられるが、例えば３、４または５ワード以上からなる全ての可能性のあるフレーズを識別するための認識される計算要件およびメモリ要件が理由である。例えば、任意の５ワードがフレーズを構成し、大規模なコーパスが少なくとも２００，０００の一意語を有すると仮定すると、約３．２×$10^{26}$の可能性のあるフレーズが存在することになり、あらゆる既存システムのメモリに格納できる限界、またはプログラムで操作できる限界を超えるのは明白である。別の問題は、新しい個々のワードが創作されるよりも、フレーズが、使用という点において継続的に用語集を出入りする頻度が高いことにある。新しいフレーズは、技術、芸術、世界的イベント、および法律等の供給源から常に生成されている。他のフレーズは、時間が経つにつれて使用されなくなっていく。

【０００６】

既存の情報検索システムでは、個々のワードの共出現パターンを用いることにより、概念検索を行うよう試みるものがある。これらのシステムでは、「President」等の１ワードの検索により、「White」および「House」等の、「President」とともに出現する頻度が高い他のワードを有する文書も検索する。この手法は、個々のワードレベルで概念的に関連する文書を有する検索結果を生成できるかもしれないが、共出現フレーズ間に内在する主題性のある関係を取得しないのが普通である。

【０００７】

従って、大規模コーパス内のフレーズを包括的に識別し、フレーズにより文書をインデックス化し、フレーズにより文書を検索およびランク付けし、そして、文書に関する追加のクラスタ化および記載情報を提供できる情報検索システムおよび方法論に対するニーズがある。

【発明の開示】

【０００８】

情報検索システムおよび方法論は、フレーズを用いて、文書コレクションにおいて文書をインデックス化、検索、ランク付け、および説明する。システムは、文書コレクションにおいて頻度が十分高く、および／または際だった使用法を持つフレーズを識別して、それらが「妥当」または「良好」なフレーズであることを示すようにしている。この方法で、多数のワードフレーズ、例えば、４語、５語以上からなるフレーズを識別できる。これにより、所与の幾つかのワードで可能なシーケンス全てから得られる可能なフレーズ毎に識別、インデックス化しなければならないという問題を回避できる。

【０００９】

本システムは、更に、フレーズが文書内の他のフレーズの存在を予測する能力に基づいて、相互に関連するフレーズを識別するようになっている。より詳細には、２フレーズの

実際の共出現率を、その２フレーズの期待される共出現率と関連させる予測指標を用いる。実際の共出現率と期待される共出現率との比としての情報ゲインは、このような予測指標の１つである。予測指標が所定の閾値を超える場合に２フレーズを関連付ける。その場合、第２フレーズは、第１フレーズに対して大きな情報ゲインをもつ。意味的に、関連するフレーズは、「President of the United States」と「White House」等の所与の主題または概念を検討または説明するために普通に用いるものである。所与のフレーズに対して、それぞれの予測指標に基づく関連性または重要性に従って、関連フレーズをランク付けできる。

【００１０】

情報検索システムは、文書コレクションにおいて文書を妥当または良好なフレーズによりインデックス化する。フレーズ毎に、ポスティングリストがそのフレーズを含む文書を識別する。更に、所与のフレーズに対して、第２リスト、ベクトル、または他の構成を用いて、所与のフレーズのどの関連フレーズが、所与のフレーズを含む各文書にも出現するかを示すデータを格納する。本方法では、システムは、検索クエリーに応じて、どの文書がどのフレーズを含むかだけではなく、どの文書がクエリーフレーズに関連するフレーズも含むか、従って、クエリーフレーズで表現される主題または概念に関してより具体的な可能性を直ちに識別できる。

【００１１】

フレーズおよび関連フレーズを用いることにより、更に、意味規制的に意味があるフレーズのグループ化を表す関連フレーズのクラスタ作成および使用が提供される。クラスタは、クラスタ内の全てのフレーズ間で非常に高い予測指標を有する関連フレーズにより特定される。クラスタを用いて、検索結果およびこれらの順位に含まれる文書の選択、および検索結果からの文書削除をともに含む検索結果の整理が可能となる。

【００１２】

本情報検索システムは、クエリーに応じて文書を検索する場合のフレーズ使用にも適合している。クエリーを処理して、クエリーフレーズに対する関係ポスティングリスト、および関連フレーズ情報を検索するように、クエリーに存在する何れかのフレーズを識別する。更に、幾つかの例では、ユーザーは、検索クエリーに「President of the」等の不完全なフレーズを入力してもよい。このような不完全なフレーズは、「President of the United States」等のフレーズ拡張により識別、置換してもよい。これは、最も可能性のあるユーザーの検索を実際に確実に実行するのに役立つ。

【００１３】

システムは関連フレーズ情報を用いて、どの文書が検索結果を含むかを識別、選択することもできる。関連フレーズ情報は、所与のフレーズおよび所与の文書に対して、所与のフレーズのどの関連フレーズが所与の文書内にあるかを示す。従って、２つのクエリーフレーズを含むクエリーに対して、第１のクエリーフレーズに対するポスティングリストを処理して、第１のクエリーフレーズを含む文書を識別し、次いで、関連フレーズ情報を処理して、これらの内のどの文書が第２のクエリーフレーズも含んでいるかを識別する。次いで、これらの後者の文書を検索結果に含める。これにより、システムが次に第２クエリーフレーズのポスティングリストを別々に処理する必要がなくなる。それによって、検索時間をより高速にできる。言うまでもなく、この手法は、クエリー内の任意のフレーズ数に拡張でき、計算および時間が著しく節約される。

【００１４】

システムは、更に、フレーズおよび関連フレーズ情報を用いて、検索結果集合内の文書をランク付けするようにしてもよい。所与のフレーズの関連フレーズ情報は、所与のフレーズに対する各関連フレーズの相対的重要度を表すビットベクトル等の形式で格納するのが好ましい。例えば、関連フレーズビットベクトルは、所与のフレーズの各関連フレーズにビットを有し、そのビットは、関連フレーズに対する予測指標（例えば、情報ゲイン）によりランク付けされる。関連フレーズビットベクトルの最上位ビットは、最大予測指標を有する関連フレーズと関係付けられ、最下位ビットは、最小予測指標を有する関連フレー

ズと関係付けられる。本手法では、所与の文書および所与のフレーズに対して、関連フレーズ情報を用いて文書をスコア付けすることができる。ビットベクトル自体の値（数値として）を、文書スコアとして用いてもよい。本手法では、クエリーフレーズの高順位の関連フレーズを含む文書の方が、低順位の関連フレーズを有する文書よりも、クエリーに主題的に関連する可能性がある。ビットベクトル値も、もっと複雑なスコア付け関数の成分として用いてもよく、更に重み付けしてもよい。次いで、文書は、文書スコアによりランク付けできる。

【００１５】

フレーズ情報を情報検索システムに用いて、ユーザーの検索を個人化することもできる。ユーザーは、例えば、ユーザーがアクセスした文書（例えば、画面で見た文書、印刷した文書、格納した文書等）から導き出されたフレーズ集としてモデル化される。より詳細には、ユーザーがアクセスした文書ついてみると、この文書内にある関連フレーズは、ユーザーモデルまたはプロファイルに含まれる。後続の検索をしている間、ユーザーモデルのフレーズを用いて、検索クエリーのフレーズをフィルタ処理し、検索文書の文書スコアに重み付けする。

【００１６】

また、フレーズ情報を情報検索システムで用いて、例えば、検索結果集合に含まれる文書等の文書説明を作成してもよい。検索クエリーについてみると、システムは、クエリーにあるフレーズを、それらの関連フレーズおよびフレーズ拡張とともに識別する。所与の文書に対して、文書の各文は、クエリーフレーズ、関連フレーズ、およびフレーズ拡張がその文に幾つあるかのカウントを有する。文書の文はこれらカウントによりランク付けでき（個々に、または組み合わせて）、幾つかの上位ランクの文（例えば、５つの文）を選択して文書説明を形成する。次に、文書が検索結果に含まれる場合、文書説明をユーザーに提示できるので、ユーザーはクエリーに関する文書を良く理解することができる。

【００１７】

文書説明生成のこのプロセスを更に改良することにより、システムは、ユーザーの関心を反映する個人化した説明を提供できる。上記のように、ユーザーモデルは、ユーザーが関心をもつ関連フレーズを識別する情報を格納する。このユーザーモデルは、クエリーフレーズに関連するフレーズのリストと共通部分を求めて、両グループに共通なフレーズを識別する。次いで、関連フレーズ情報により共通部分をランク付ける。次に、得られた関連フレーズ集合を用いて、各文書にあるこれらの関連フレーズのインスタンス数により文書の文ランク付けする。共通関連フレーズの最大数を有する幾つかの文は、個人化文書説明として選択される。

【００１８】

情報検索システムは、文書コレクションをインデックス化（クローリング）しながら、または検索クエリーを処理する場合のいずれかに、フレーズ情報を用いて、複製文書を識別および削除することもできる。所与の文書に対して、文書の各文は、関連フレーズが文中に幾つあるかのカウントを有する。文書の文をこのカウントによりランク付けし、幾つかの上位ランクの文（例えば、５つの文）が選択されて、文書説明を形成する。次いで、この説明を、例えば、文の文字列またはハッシュとして、文書に関係付けて格納する。インデックス化中に、新規にクローリングした文書は同じ方法で処理されて、文書説明を生成する。新規の文書説明を以前の文書説明と一致（例えば、ハッシュ化）させることができ、一致が見付かった場合、新規文書は複製である。同様に、検索クエリー結果の作成中に、検索結果集合の文書を処理して、複製を削除する。

【００１９】

本発明は更に、システムおよびソフトウエアアーキテクチャ、コンピュータプログラム製品およびコンピュータ実装方法、ならびにコンピュータ生成ユーザーインターフェースおよび提示の実施の形態を有する。

【００２０】

前述のことは、フレーズに基づく情報検索システムおよび方法論の特長を内のほんの一部

に過ぎない。情報検索の従来技術に精通する者には言うまでもないが、多くのフレーズ情報のフレキシビリティにより、インデックス化、文書注釈化、検索、ランク付け、ならびに文書解析および処理の他領域における多様な使用および用途が可能になる。

【００２１】

【００２２】

【００２３】

【００２４】

【００２５】

【００２６】

【００２７】

【００２８】

【００２９】

図は、本発明の好適な実施の形態を説明するためだけに示す。当該技術に精通する者には以下の検討から直ちに理解できるのは言うまでもないが、本明細書に示す構造および方法の代替の実施の形態は、本明細書内で説明する本発明の原理から逸脱することなく利用されてもよい。

【発明を実施するための最良の形態】

【００３０】

Ｉ．システム概要

図１について説明する。本発明の一実施の形態による検索システム１００の実施の形態のソフトウエアアーキテクチャを示す。本実施の形態では、システムには、インデックス化システム１１０、検索システム１２０、提示システム１３０、およびフロントエンドサーバ１４０が含まれる。

【００３１】

インデックス化システム１１０は、各種のウェブサイト１９０、および他の文書コレクションにアクセスすることにより、文書中のフレーズを識別し、それらのフレーズにより文書のインデックス化を担当する。フロントエンドサーバ１４０は、クライアント１７０のユーザーからクエリーを受け取り、検索システム１２０にこれらのクエリーを提供する。検索システム１２０は、検索クエリーに関する文書（検索結果）の検索を担当するが、これには、検索クエリーの任意のフレーズを識別すること、次いで、フレーズの存在を用いて検索結果中の文書をランク付けしてランクに影響を与えることが含まれる。検索システム１２０は、検索結果を提示システム１３０に提供する。提示システム１３０は、複製に近い文書を削除し、文書の主題説明を生成し、そして、クライアント１７０に結果を提供するフロントエンドサーバ１４０に修正した検索結果を返すことを含む検索結果の修正を担当する。システム１００は更に、文書に関係するインデックス情報を格納するインデックス１５０、ならびにフレーズおよび関連統計情報を格納するフレーズデータ格納部１６０を含む。

【００３２】

本出願の文脈では、「文書」は、検索エンジンがインデックス化し、検索することができる任意の種類の媒体であると理解され、ウェブ文書、画像、マルチメディアファイル、テキスト文書、ＰＤＦまたは他の画像フォーマットのファイル等が含まれる。文書は、内容および種類に応じて、一枚以上のページ、パーティション、セグメント、または他の構成を有していてもよい。同等に、文書は、インターネット上の文書を通常呼ぶのに用いるように、「ページ」と呼んでもよい。一般的な用語「文書」の使用により、本発明の範囲に関する制限を意味することはない。検索システム１００は、インターネットおよびワールドワイドウェブ等の文書の大規模コーパス上で動作するが、図書館または個人企業の文書コレクションに対する等の限定されるコレクションでも用いることもできる。どちらの文脈でも、文書は、多くの異なるコンピュータシステムおよびサイトにわたって分散しているのが普通であるのは言うまでもない。また、一般性を失うことなく、フォーマットまたは場所とは無関係に（例えば、どのウェブサイトまたはデータベースか）、文書を一般的

に、コーパスまたは文書コレクションと総称して呼ぶ。各文書は、文書を一意に識別する関係識別子を有し、その識別子はＵＲＬであることが好ましいが、他の種類の識別子（例えば、文書番号）を同じように用いてもよい。本開示では、文書を識別するのにＵＲＬの使用を想定している。

II．インデックス化システム

【００３３】

一実施の形態では、インデックス化システム１１０は、３つの主要機能動作：１）フレーズおよび関連フレーズの識別、２）フレーズに関連する文書のインデックス化、および３）フレーズに基づくソートの生成および保持、を提供する。当該技術に精通する者には言うまでもないが、インデックス化システム１１０は、従来のインデックス化機能も同様にサポートする等、他の機能も実行するが、本明細書ではこれら他の動作についてはこれ以上説明しない。インデックス化システム１１０は、フレーズデータのインデックス１５０、およびデータレポジトリ１６０に関して動作する。これらデータレポジトリについて以下に更に説明する。

１．フレーズ識別

【００３４】

インデックス化システム１１０のフレーズ識別動作は、文書のインデックス化および検索に有用な文書コレクション内の「良好」および「不良」フレーズを識別する。一局面では、良好フレーズは、文書コレクション内のある割合を超える文書に発生する傾向があるフレーズであり、および／または、マークアップタグ、もしくは他の形態学的マーカ、フォーマットマーカ、または文法的マーカにより区切られるような、かかる文書に著しく出現するとして示されるフレーズである。良好フレーズの別の局面は、それらが他の良好フレーズの予測であり、単に用語集に出現するワードのシーケンスというだけではない、ということである。例えば、フレーズ「President of the United States（米国大統領）」は、「George Bush（ジョージ・ブッシュ）」、および「Bill Clinton（ビル・クリントン）」等の他のフレーズを予測させるフレーズである。しかし、「fell down the stairs（階段から落ちた）」または「top of the morning（最高の朝）」、「out of the blue（前触れもなく）」等の、他のフレーズは予測できない。なぜなら、このような慣用句および話しことばは、他の異なる無関係な多くのフレーズとともに出現する傾向をもつからである。従って、フレーズ識別は、どのフレーズが良好で、どのフレーズが不良（すなわち、予測力が欠如している）かを判定する。

【００３５】

図２について説明する。フレーズ識別処理は以下の機能的な段階を有する：

【００３６】

２００：見込みがあり、良好なフレーズを、フレーズの頻度および共出現の統計値とともに収集する。

【００３７】

２０２：頻度統計値に基づいて、見込フレーズを良好または不良フレーズにソートする。

【００３８】

２０４：共出現統計値から導かれた予測指標に基づいて良好フレーズリストを絞り込む。

【００３９】

これら段階のそれぞれについて更に詳細に説明する。

【００４０】

第１段階２００は、インデックス化システム１１０が文書コレクション内の文書の集合をクロールするプロセスであり、時間が経つと、文書コレクションの繰り返しパーティションを作成する。一パーティションが、一通過毎に処理される。一通過毎にクロールされる文書の数は、変化でき、パーティションあたり約１，０００，０００が好ましい。全ての文書が処理されるまで、または他の終了基準が満たされるまで、前回クロールしなかった文書だけを、各パーティションで処理するのが好ましい。実施において、新規文書が文書コレクションに連続的に加えられるようにクローリングを続ける。インデックス化システ

ム１１０は、クロールする文書毎に以下のステップをとる。

【００４１】

フレーズ窓長さｎをもつ文書のワードをトラバースする。ここで、ｎは所望の最大フレーズ長である。窓の長さは、少なくとも２語、好ましくは、４または５語（ワード）とするのが普通である。フレーズが、フレーズ窓内の全てのワードを含み、その他に、「a」、「the」等のストップワードとして特徴付けられるものを含むことが好ましい。フレーズ窓は、行端、段落リターン、マークアップタグ、または内容またはフォーマットの変更を示す他の印で終了してもよい。

【００４２】

図３はトラバース中の文書３００の一部を示し、ワード「stock」で開始し、右に５ワード拡張するフレーズ窓３０２を示す。窓３０２の最初の窓は、候補のフレーズｉであり、各シーケンスｉ＋１、ｉ＋２、ｉ＋３、ｉ＋４、およびｉ＋５も同様に候補フレーズである。このように、本実施例では、候補フレーズは：「stock」、「stock dogs」、「stock dogs for」、「stock dogs for the」、「stock dogs for the Basque」、および「stock dogs for the Basque shepherds（バスクシェパードに対する家畜犬）」である。

【００４３】

それぞれのフレーズ窓３０２では、それぞれの候補フレーズをチェックして、次に、良好フレーズリスト２０８または見込フレーズリスト２０６に既にあるかどうかを判定する。候補フレーズが、良好フレーズリスト２０８または見込フレーズリスト２０６のどちらにもない場合、候補は既に「不良」であると判定されており、スキップされる。

【００４４】

候補フレーズが、エントリ$g_j$として、良好フレーズリスト２０８にある場合、フレーズ$g_j$に対するインデックス１５０のエントリを更新して、その文書（例えば、ＵＲＬまたは他の文書識別子）を含め、この候補フレーズ$g_j$が現在の文書に出現することを示す。フレーズ$g_j$（または語）に対するインデックス１５０のエントリは、フレーズ$g_j$のポスティングリストと呼ばれる。ポスティングリストには、フレーズが発生する文書ｄのリスト（文書識別子、例えば、文書番号または代替としてのＵＲＬによる）が含まれる。

【００４５】

更に、共出現マトリックス２１２を、以下に更に示すように更新する。最初の通過では、良好リストおよび不良リストは空なので、従って、大部分のフレーズは見込フレーズリスト２０６に追加される傾向にある。

【００４６】

候補フレーズが良好フレーズリスト２０８にない場合、それを見込フレーズリスト２０６に追加する。但し、見込フレーズリストに既にある場合を除く。見込フレーズリスト２０６の各エントリｐは、３つの関係付けられたカウントを有する。

【００４７】

Ｐ（ｐ）：見込フレーズが出現する文書の数；

【００４８】

Ｓ（ｐ）：見込フレーズの全てのインスタンスの数；

【００４９】

Ｍ（ｐ）：見込フレーズの関心のあるインスタンスの数。見込フレーズが、文法的またはフォーマットマーカにより、例えば、太字、または下線、またはハイパーリンクのアンカーテキスト、または引用符により、文書内の隣接内容と区別される場合、見込フレーズのインスタンスは「関心」がある。これらの（および、他の）区別されている出現は、各種のＨＴＭＬマークアップ言語タグ、および文法的マーカにより示される。良好フレーズリスト２０８に配置される場合、これらの統計値は、フレーズに対して保持される。

【００５０】

各種のリストに加えて、良好フレーズに対する共出現マトリックス２１２（Ｇ）が保持される。マトリックスＧは、ｍ×ｍの次元を有し、ここで、ｍは良好フレーズの数である。マトリックスにおける各エントリＧ（ｊ、ｋ）は、一対の良好フレーズ（$g_j$、$g_k$）を表

す。共出現マトリックス２１２は、現在のワードｉの中心にあり、＋／－ｈワードを拡張する２次窓３０４に関する良好フレーズの各対（$g_j$、$g_k$）に対する３つの別々のカウントを論理的に（必ずしも物理的ではないが）保持する。図３に示すような、一実施の形態では、２次窓３０４は３０ワードである。従って、共出現マトリックス２１２は以下を保持する：

【００５１】

Ｒ（ｊ、ｋ）：生の共出現カウント。フレーズ$g_j$がフレーズ$g_k$とともに２次窓３０４に出現する回数；

【００５２】

Ｄ（ｊ、ｋ）：離接的関心カウント。フレーズ$g_j$またはフレーズ$g_k$のいずれかが２次窓に区別されたテキストとして出現する回数；および、

【００５３】

Ｃ（ｊ、ｋ）：接続的関心カウント；$g_j$およびフレーズ$g_k$がともに２次窓に区別されたテキストとして出現する回数。接続的関心カウントの使用は、フレーズ（例えば、著作権の注記）がサイドバー、フッタ、またはヘッダに頻繁に出現するので、実際には、他のテキストを予測しない場合の状況を回避するのに特に有益である。

【００５４】

図３の実施例について説明する。「stock dogs」は、フレーズ「Australian Shepherd」および「Australian Shepard Club of America」と同様に、良好フレーズリスト２０８にあると仮定する。これら後者のフレーズはともに、現在のフレーズ「stock dogs」周りの２次窓３０４内に出現する。しかし、フレーズ「Australian Shepherd Club of America」は、ウェブサイトへのハイパーリンク（下線で示す）のためのアンカーテキストとして出現する。従って、対となる{「stock dogs」、「Australian Shepherd」}に対する生の共出現カウントは増大し、{「stock dogs」、「Australian Shepherd Club of America」}に対する生の出現カウント、および離接的関心カウントは、後者が区別されたテキストとして出現するので、ともに増大する。

【００５５】

シーケンス窓３０２および２次窓３０４両方による各文書トラバース処理は、パーティションにある文書毎に繰り返される。

【００５６】

パーティションにある文書をトラバースすると、インデックス化動作の次の段階は、見込フレーズリスト２０６により良好フレーズリスト２０８を更新する（２０２）ことである。フレーズの出現頻度、およびそのフレーズが出現する文書の数が、そのフレーズがセマンティックな意味のあるフレーズとして十分に使用されていることを示す場合、見込フレーズリスト２０６の見込フレーズｐを良好フレーズリスト２０８に移動する。

【００５７】

一実施の形態では、これは以下のように検査される。見込フレーズｐは、以下の場合、見込フレーズリスト２０６から削除し、良好フレーズリスト２０８に配置する：

【００５８】

ａ）Ｐ（ｐ）＞１０、かつＳ（ｐ）＞２０（フレーズｐを含む文書の数が１０を越え、かつフレーズｐの発生回数が２０を超える）、または；

【００５９】

ｂ）Ｍ（ｐ）＞５、（フレーズｐの関心インスタンス数が５を超える）。

【００６０】

これらの閾値は、パーティションの文書数により縮小拡大される；例えば、２，０００，０００の文書をパーティション内でクロールする場合、閾値は約２倍となる。もちろん、当該技術に精通する者には言うまでもないが、閾値の特定の値、またはそれらを検査するロジックは、要望に応じて変更できる。

【００６１】

フレーズｐが良好フレーズリスト２０８に対して適格でない場合、不良フレーズとしての

適性をチェックする。以下の場合、フレーズｐは不良フレーズである：

【００６２】

ａ）フレーズを含む文書数、Ｐ（ｐ）＜２；および、

【００６３】

ｂ）フレーズの関心インスタンス数、Ｍ（ｐ）＝０。

【００６４】

これらの条件は、フレーズの頻度が低く、かつ重要ではない内容として用いられていることを示す。繰り返しになるが、これらの閾値は、パーティションの文書数により縮小拡大される。

【００６５】

良好フレーズリスト２０８は、上記のようにワード数が多いフレーズの他に、フレーズとして個別のワードを当然含むということに留意されたい。この理由は、フレーズ窓３０２の最初の各ワードが、常に候補フレーズであり、適切なインスタンスカウントが累積されるからである。従って、インデックス化システム１１０は、個別のワード（すなわち、単一ワードをもつフレーズ）およびワード数が多いフレーズ両方を自動的にインデックス化できる。良好フレーズリスト２０８も、ｍフレーズの全ての可能性のある組合せによる理論的最大値より、かなり短くなる。代表的な本実施の形態では、良好フレーズリスト２０８は、約６．５×$10^5$フレーズを含む。システムが見込フレーズおよび良好フレーズのトラックを保持してさえいればよいとき、不良フレーズリストは格納不要である。

【００６６】

文書コレクションを最終的に通過させることにより、見込フレーズリストは、大規模コーパスのフレーズ使用の期待される分布により、比較的短くなる。従って、例えば１０回の通過（例えば、１０，０００，０００文書）の場合、フレーズは、まさに最初の回だけに出現し、その時点で良好フレーズとなる可能性はほとんどない。そのフレーズは使用されるようになったばかりの新規フレーズかもしれず、従って、後続のクロールをしている間に次第に共通となる。この場合、それぞれのカウントは増大し、最終的に良好フレーズとなる閾値を満たすことになる。

【００６７】

インデックス化動作の第３段階は、共出現マトリックス２１２から導き出された予測指標を用いる良好フレーズリスト２０８の絞り込み２０４である。絞り込みをしなければ、用語集に正当に出現する間、それら自体が他のフレーズの存在を十分に予測しないか、またはそれら自体がさらに長いフレーズの続きである多くのフレーズを、良好フレーズリスト２０８が含む可能性もあり得る。これらの弱い良好フレーズを削除することにより、良好フレーズが非常にロバストになる可能性がある。良好フレーズを識別するために、別のフレーズが存在する文書に出現する或るフレーズの可能性が高いことを表す予測指標を用いる。一実施の形態では、これは以下のように成される：

【００６８】

上記のように、共出現マトリックス２１２は、良好フレーズと関係付けられる格納データのｍ・ｍマトリックスである。マトリックスの各行ｊは、良好フレーズ$g_j$を表し、各列ｋは良好フレーズ$g_k$を表す。各良好フレーズ$g_j$について期待値Ｅ（$g_j$）が計算される。期待値Ｅは、$g_j$を含むと期待されるコレクション内の文書の割合である。例えば、これは、$g_j$を含む文書の数と、クロールされているコレクション内の文書の総数Ｔとの比、Ｐ（ｊ）／Ｔとして計算される。

【００６９】

上記のように、$g_j$を含む文書の数は、$g_j$が１文書に出現する毎に更新される。Ｅ（$g_j$）の値は、$g_j$のカウントが１つ増加する毎に、またはこの第３段階の間に更新できる。

【００７０】

次に、他の良好フレーズ$g_k$（例えば、マトリックスの列）それぞれに対して、$g_j$が$g_k$を予測するかどうかを判定する。$g_j$の予測指標は以下のように決定される：

【００７１】

i）期待値Ｅ（$g_k$）を計算する。$g_j$および$g_k$の期待される共出現率Ｅ（ｊ、ｋ）は、両者が関連性のないフレーズの場合、Ｅ（$g_j$）＊Ｅ（$g_k$）である；

【００７２】

ii）$g_j$および$g_k$の実際の共出現率Ａ（ｊ、ｋ）を計算する。これは文書総数Ｔで除した生の共出現カウントＲ（ｊ、ｋ）である；

【００７３】

iii）実際の共出現率Ａ（ｊ、ｋ）が或る閾値量だけ前記期待される共出現率Ｅ（ｊ、ｋ）を超える場合、$g_j$は$g_k$を予測するといえる。

【００７４】

一実施の形態では、予測指標は情報ゲインである。従って、$g_j$が存在しているところで$g_k$の情報ゲインＩが閾値を超える場合、フレーズ$g_j$は別のフレーズ$g_k$を予測する。一実施の形態では、これは以下のように計算される：

【００７５】

Ｉ（ｊ、ｋ）＝Ａ（ｊ、ｋ）／Ｅ（ｊ、ｋ）

【００７６】

そして、次の場合、良好フレーズ$g_j$は、良好フレーズ$g_k$を予測する：

【００７７】

Ｉ（ｊ、ｋ）＞情報ゲイン閾値。

【００７８】

一実施の形態では、情報ゲイン閾値は１．５であるが、１．１と１．７との間が好ましい。閾値を１．０以上に上げるのは、２つの別々の関連がないフレーズがランダムに予測される以上に共出現する可能性を低下させるのに役立つ。

【００７９】

注記したように、情報ゲインの計算は、所与の行ｊに関するマトリックスＧの各列ｋに対して繰り返す。行が完全になると、どの良好フレーズ$g_k$の情報ゲインも情報ゲイン閾値を越えない場合、フレーズ$g_j$が他の良好フレーズをどれも予測しないことを意味する。その場合、$g_j$は、良好フレーズリスト２０８から削除され、基本的に不良フレーズになる。注意すべきは、フレーズ$g_j$に対する列ｊは削除しない、ということである。このフレーズ自体は他の良好フレーズにより予測できるからである。

【００８０】

このステップは、共出現マトリックス２１２の全ての行を評価した場合、終了する。

【００８１】

この段階の最終ステップは、良好フレーズリスト２０８を絞り込んで、不完全なフレーズを削除することである。不完全なフレーズは、フレーズ拡張を予測するだけのフレーズであり、フレーズの一番左側（すなわち、フレーズの先頭）で開始する。フレーズｐの「フレーズ拡張」は、フレーズｐで始まるスーパーシーケンスである。例えば、フレーズ「President of」は、「President of the United States」、「President of Mexico」、「President of AT&T」等を予測する。これら後者のフレーズ全ては、「President of」で始まり、そのスーパーシーケンスであるためフレーズ「President of」のフレーズ拡張である。

【００８２】

従って、良好フレーズリスト２０８に残る各フレーズ$g_j$は、先に説明した情報ゲイン閾値に基づいて、幾つかの数の他のフレーズを予測する。ここで、各フレーズ$g_j$に対して、インデックス化システム１１０は、予測されるフレーズ$g_k$それぞれとの文字列一致を実行する。文字列一致は、各予測フレーズ$g_k$がフレーズ$g_j$のフレーズ拡張かどうかを検査する。予測フレーズ$g_k$全てが、フレーズ$g_j$のフレーズ拡張である場合、フレーズ$g_j$は不完全であり、良好フレーズリスト２０８から削除され、不完全フレーズリスト２１６に加えられる。従って、$g_j$の拡張ではない少なくとも１つのフレーズ$g_k$がある場合、$g_j$は完全であり、良好フレーズリスト２０８に保持される。例えば、「President of the United」は、予測する他のフレーズが、フレーズの拡張である「President of the Unite

d States」だけなので、不完全フレーズである。

【００８３】

不完全フレーズリスト２１６自体は、実際の検索中には非常に有用である。検索クエリーを受け取った場合、それを不完全フレーズリスト２１６と比較できる。クエリー（またはその一部）がリスト内のエントリと一致する場合、検索システム１２０は、不完全フレーズの最も可能性が高いフレーズ拡張（不完全フレーズを与えられた最大の情報ゲインを有するフレーズ拡張）を参照し、このフレーズ拡張をユーザーに示すか、またはそのフレーズ拡張を自動的に検索できる。例えば、検索クエリーが「President of the United」である場合、検索システム１２０は、ユーザーに検索クエリーとして「President of the United States」を自動的に示すことができる。

【００８４】

インデックス化処理の最後の段階が完了すると、良好フレーズリスト２０８は、コーパス中で発見された多数の良好フレーズを含む。これらの良好フレーズはそれぞれ、そのフレーズ拡張ではない少なくとも１つの他のフレーズを予測する。すなわち、各良好フレーズを十分な頻度で、かつ独立して用いて、コーパス中で表わされる意味のある概念またはアイディアを示す。所定の、または手動で選択したフレーズを用いる既存システムとは違って、良好フレーズリストは、コーパスで実際に用いられているフレーズを反映する。更に、クローリングおよびインデックス化の上記処理は、新規文書を文書コレクションに追加する際に周期的に繰り返されるので、それらが用語集に入力されるように、インデックス化システム１１０は自動的に新規フレーズを検出する。

２．関連フレーズ、および関連フレーズのクラスタの識別

【００８５】

図４について説明する。関連フレーズ識別処理には、以下の機能的動作が含まれる。

【００８６】

４００：高い情報ゲイン値を有する関連フレーズを識別する。

【００８７】

４０２：関連フレーズのクラスタを識別する。

【００８８】

４０４：クラスタービットベクトルおよびクラスタ番号を格納する。

【００８９】

これら動作をそれぞれ詳細に説明する。

【００９０】

最初に、思い出すべきは、共出現マトリックス２１２が良好フレーズ$g_j$を含み、そのそれぞれは、情報ゲイン閾値より大きな情報ゲインをもつ少なくとも１つの他の良好フレーズ$g_k$を予測するということである。次いで、関連フレーズを識別する（４００）ために、良好フレーズの各対（$g_j$、$g_k$）毎に、その情報ゲインを関連フレーズ閾値、例えば１００、と比較する。すなわち、$g_j$および$g_k$は、次の場合、関連フレーズである：

【００９１】

$I(g_j, g_k) > 100$

【００９２】

この高い閾値を用いて、統計的期待率を十分超える良好フレーズの共出現を識別する。統計的に、これはフレーズ$g_j$および$g_k$が、期待される共出現率より１００倍多く共出現する、ということを意味する。例えば、文書や「Monica Lewinsky（モニカ・ルインスキー）」というフレーズを考えてみると、フレーズ「Bill Clinton（ビル・クリントン）」は同一文書に１００倍以上出現する可能性があり、ひいては、フレーズ「Bill Clinton」は、任意のランダムに選択される文書にも出現する可能性がある。言いかえると、出現率が１００：１なので、予測精度は９９．９９９％である。

【００９３】

従って、関連フレーズ閾値より少ない任意のエントリ（$g_j$、$g_k$）はゼロになり、フレーズ$g_j$、$g_k$が関連していないことを示す。共出現マトリックス２１２の残りのエントリは

いずれも、全ての関連フレーズを示すことになる。

【００９４】

次いで、共出現マトリックス２１２の各行$g_j$の列$g_k$は、情報ゲイン値Ｉ（$g_j$、$g_k$）でソートされるので、最大情報ゲインをもつ関連フレーズ$g_k$は最初にリストに挙げられる。従って、このソートは、情報ゲインの点からみて他のどのフレーズが最も関連する可能性があるかを、所与のフレーズ$g_j$に対して識別する。

【００９５】

次のステップは、関連フレーズのクラスタを互いに形成しているのはどの関連フレーズかを決定する（４０２）ことである。クラスタは、各フレーズが、少なくとも１つの他のフレーズに対して、高い情報ゲインをもつ関連フレーズの集合である。一実施の形態では、クラスタは以下のように識別される。

【００９６】

マトリックスの各行$g_j$には、フレーズ$g_j$に関連する１つの以上の他のフレーズがある。この集合は、関連フレーズ集合Ｒjであり、ここで、Ｒ＝{$g_k$、$g_l$、．．$g_m$}である。

【００９７】

$R_j$の各関連フレーズｍについて、インデックス化システム１１０は、Ｒの他の関連フレーズそれぞれが$g_j$にも関連しているかどうかを判定する。従って、Ｉ（$g_j$、$g_k$）がゼロではない場合も、$g_j$、$g_k$、および$g_l$はクラスタの一部である。このクラスタ検査は、Ｒの各対（$g_l$、$g_m$）について繰り返す。

【００９８】

例えば、良好フレーズ「Bill Clinton」がフレーズ「President」、「Monica Lewinsky」と関連していると仮定する。なぜなら、「Bill Clinton」に対するこれらの各フレーズの情報ゲインは関連フレーズ閾値を超えるからである。更に、フレーズ「Monica Lewinsky」がフレーズ「purse designer（ハンドバッグデザイナ）」に関連していると仮定する。従って、これらのフレーズは、集合Ｒを形成する。クラスタを判定するために、インデックス化システム１１０は、これらの各フレーズの他のフレーズに対する情報ゲインを、対応する情報ゲインを判定することにより評価する。こうして、インデックス化システム１１０は、Ｒの全ての対について、情報ゲインＩ（“President”、“Monica Lewinsky”）、Ｉ（“President”、“purse designer”）等を判定する。本実施例では、「Bill Clinton」、「President」および「Monica Lewinsky」は、１つのクラスタを形成し、「Bill Clinton」および「President」は第２のクラスタを形成し、「Monica Lewinsky」および「Purse designer」は第３のクラスタを形成し、そして、「Monica Lewinsky」、「Bill Clinton」および「purse designer」は第４のクラスタを形成する。この理由は、「Bill Clinton」は、十分な情報ゲインをもつ「Purse designer」を予測しないが、「Monica Lewinsky」は、これらのフレーズの両方を予測するからである。

【００９９】

クラスタ情報を記録する（４０４）ために、各クラスタは、一意のクラスタ番号（クラスタＩＤ）を割り当てられる。次いで、この情報は、各良好フレーズ$g_j$とともに記録される。

【０１００】

一実施の形態では、クラスタ番号は、フレーズ間の直交関係をも示すクラスタービットベクトルにより決定される。クラスタービットベクトルは、良好フレーズリスト２０８内の良好フレーズ数ｎの長さのビットのシーケンスである。所与の良好フレーズ$g_j$について、ビット位置はソートした$g_j$の関連フレーズＲに対応する。ビットは、Ｒの関連フレーズ$g_k$がフレーズ$g_j$と同一のクラスタにある場合に設定される。より一般的には、これは、$g_j$と$g_k$との間のいずれかの方向の情報ゲインがある場合に、クラスタービットベクトルの対応するビットが設定される、ということを意味している。

【０１０１】

次いで、クラスタ番号は、結果的に得られるビット列の値となる。この実装は、多くの方向、または一方向の情報ゲインをもつ関連フレーズが同一クラスタに出現するという特

性をもつ。

【０１０２】

クラスタービットベクトルの例は、上記フレーズを用いて、以下のようになる：

【表１】

| | Bill Clinton | President | Monica Lewinsky | purse designer | クラスタＩＤ |
|---|---|---|---|---|---|
| Bill Clinton | 1 | 1 | 1 | 0 | 14 |
| President | 1 | 1 | 0 | 0 | 12 |
| Monica Lewinsky | 1 | 0 | 1 | 1 | 11 |
| purse designer | 0 | 0 | 1 | 1 | 3 |

【０１０３】

要約すると、この処理の後、各良好フレーズ$g_j$について関連フレーズＲの集合が識別され、この集合は、情報ゲインＩ（$g_j$、$g_k$）の最大値から最小値の順にソートされる。更に、各良好フレーズ$g_j$について、クラスタービットベクトルがあり、その値は、フレーズ$g_j$がメンバーである一次クラスタを識別するクラスタ番号であり、そして、Ｒの関連フレーズのどれが$g_j$をもつ共通クラスタにあるかを示す直交値（各ビット位置について１または０）がある。従って、上記実施例では、「Bill Clinton」、「President」、および「Monica Lewinsky」は、フレーズ「Bill Clinton」の行のビット値に基づくクラスタ１４にある。

【０１０４】

この情報を格納するために、２つの基本的な表現が利用できる。第１は、上記のように、情報を共出現マトリックス２１２に格納してもよく、その場合は：

【０１０５】

エントリＧ［行ｊ、列ｋ］＝（Ｉ（ｊ、ｋ）、クラスタ番号、クラスタービットベクトル）

【０１０６】

代替として、マトリックス表現を回避することができ、全ての情報は、良好フレーズリスト２０８に格納され、そこでは、各行が良好フレーズ$g_j$を表す：

【０１０７】

フレーズ行j＝リスト［フレーズ$g_k$、（Ｉ（ｊ、ｋ）、クラスタ番号、クラスタービットベクトル）］。

【０１０８】

本手法は、クラスタに対する有用な整理を提供する。最初に、本手法は、主題および概念を厳密にそして頻繁には任意に定義する階層というより、主題が複雑な関係グラフを、関連フレーズが示すように形成するということがわかる。ここで、幾つかのフレーズは、多くの他のフレーズに関連付けられるとともに、幾つかのフレーズは、より制限された範囲を有し、その関係は相互関係であるか（各フレーズが他のフレーズを予測する）、または一方向のこともある（１フレーズが他のフレーズを予測するが、その逆はない）。その結果は、クラスタが各良好フレーズに「ローカル」を特徴付けることができる、ということであり、クラスタの幾つかは、１つ以上の共通関連フレーズにより重複する。

【０１０９】

次いで、所与の良好フレーズ$g_j$について、情報ゲインによる関連フレーズのランク付けが、フレーズのクラスタを名付けるためのソートを提供し、クラスタ名は、最大情報ゲインを有するクラスタにある関連フレーズの名前である。

【０１１０】

上記処理は、文書コレクションに出現する顕著なフレーズを識別する非常にロバストな方法を提供し、有益に、これらの関連フレーズが、実施において自然な「クラスタ」で共に使用される方法を提供する。結果的に、関連フレーズのこのデータ駆動クラスタは、多くのシステムに共通するような、関連語および概念の手動で指導する「編集」の選択における固有の偏向を回避することができる。

３．フレーズおよび関連フレーズによる文書インデックス化

【０１１１】

関連フレーズおよびクラスタに関する情報を含む良好フレーズリスト２０８を考えると、インデックス化システム１１０の次の機能的動作は、良好フレーズおよびクラスタに関する文書コレクション内の文書をインデックス化し、インデックス１５０に更新情報を格納することである。図５は、文書をインデックス化するための以下の機能段階があるこの処理を示す：

【０１１２】

５００：文書内に見付かった良好フレーズのポスティングリストに文書を書き込む。

【０１１３】

５０２：インスタンスカウント、ならびに関連フレーズおよび２次関連フレーズに対する関連フレーズビットベクトルを更新する。

【０１１４】

５０４：関連フレーズ情報をもつ文書に注釈を付ける。

【０１１５】

５０６：ポスティングリストサイズによりインデックスエントリを並べ替える。

【０１１６】

これらの段階を更に詳細に説明する。

【０１１７】

文書集合を上記のようにトラバースするか、またはクロールする。これは同一または異なる文書集合であってもよい。所与の文書ｄに対して、上記の方法で、位置ｉから、長さｎのシーケンス窓３０２によりワード毎に文書をトラバースする（５００）。

【０１１８】

所与のフレーズ窓３０２では、窓内の全ての良好フレーズを識別し、位置ｉで開始する。各良好フレーズは$g_i$として記す。従って、ｇ１は第１良好フレーズ、ｇ２は第２良好フレーズ等となる。

【０１１９】

各良好フレーズ$g_i$に対して（例えば、ｇ１「President」、およびｇ４「President of ATT」）、インデックス１５０の良好フレーズ$g_i$について、ポスティングリストに文書識別子（例えば、ＵＲＬ）を書き込む。この更新は、良好フレーズ$g_i$がこの特定文書に出現することを識別する。

【０１２０】

一実施の形態では、フレーズ$g_j$に対するポスティングリストは、以下の論理的形式をとる：

【０１２１】

フレーズ$g_j$：リスト（文書ｄ、［リスト：関連フレーズカウント］［関連フレーズ情報］）

【０１２２】

各フレーズ$g_j$に対して、フレーズが出現する文書ｄのリストがある。各文書に対して、文書ｄにも出現するフレーズ$g_j$の関連フレーズＲの発生数のカウントのリストがある。

【０１２３】

一実施の形態では、関連フレーズ情報は、関連フレーズビットベクトルである。このビットベクトルは、「バイビット」ベクトルとして特徴付けることができ、各関連フレーズ$g_k$に対して、２つのビット位置$g_k-1$、$g_k-2$がある。第１ビット位置は、関連フレーズ$g_k$が文書ｄにあるか（すなわち、文書ｄ内の$g_k$に対するカウントが０を超えるか）ど

うかを示すフラグを格納する。第２ビット位置は、$g_k$の関連フレーズ$g_l$も文書ｄにあるかどうかを示すフラグを格納する。フレーズ$g_j$の関連フレーズ$g_k$の関連フレーズ$g_l$は、本明細書では「$g_j$の２次関連フレーズ」と称する。カウントおよびビット位置は、Ｒのフレーズの正準順位（情報ゲイン減少順にソートされる）と対応する。このソート順は、関連フレーズビットベクトルの最上位ビットと関係付けられる$g_j$により最大予測される関連フレーズ$g_k$、および最下位ビットと関係付けられる$g_j$により最小予測される関連フレーズ$g_l$を作成する効果を有する。

【０１２４】

有用な注意点は、所与のフレーズｇについて、関連フレーズビットベクトルの長さ、およびベクトルの個々のビットに対する関連フレーズの関係が、ｇを含む全ての文書に関して同一である、ということである。この実装は、システムが、ｇを含む任意の（または全ての）文書に対する関連フレーズビットベクトルを直ちに比較して、どの文書が所与の関連フレーズを有するかを調べることができるという特性を有する。これにより、検索処理が容易になり検索クエリーに応じて文書を識別するのに有益なものとなる。従って、所与の文書は、多くの異なるフレーズのポスティングリストに出現し、そのような各ポスティングリストでは、その文書に対する関連フレーズベクトルは、ポスティングリストを所有するフレーズに特有である。この局面は個々のフレーズおよび文書に関する関連フレーズビットベクトルの局所性を保つ。

【０１２５】

従って、次の段階５０２には、文書の現在のインデックス位置の２次窓３０４（先に説明したように、＋／－Ｋ語、例えば３０語、の２次窓）、例えば、ｉ－Ｋからｉ＋Ｋまでをトラバースすることが含まれる。２次窓３０４に出現する$g_i$の各関連フレーズ$g_k$に対して、インデックス化システム１１０は、関連フレーズカウントにある文書ｄに関する$g_k$のカウントを増加させる。$g_i$が文書の後ろの方に出現し、関連フレーズが後ろの２次窓内に再度見付かる場合は、再度カウントを増加させる。

【０１２６】

注記したように、関連フレーズビットマップにある対応第１ビット$g_k-1$は、カウントに基づいて、$g_k$に対するカウントが＞０である場合１にビット設定し、またはカウントが０の場合０に設定する。

【０１２７】

次に、第２ビット$g_k-2$はインデックス１５０において関連フレーズ$g_k$を参照することにより設定し、文書ｄに対するエントリをｇkのポスティングリストにおいて識別し、次いで、$g_k$の２次関連フレーズカウント（またはビット）を任意のその関連フレーズに対してチェックする。これらの２次関連フレーズカウント／ビットの何れかが設定された場合、これにより示されるのは、$g_j$の２次関連フレーズが文書ｄにもあるということである。

【０１２８】

文書ｄをこの方法で完全に処理した場合、インデックス化システム１１０は以下を識別することになる：

【０１２９】

ｉ）文書ｄ内の各良好フレーズ$g_i$；

【０１３０】

ii）各良好フレーズ$g_j$に対して、その関連フレーズ$g_k$の内のどれかが文書ｄにある；

【０１３１】

iii）文書ｄにある各良好フレーズ$g_k$に対して、その関連フレーズ$g_l$（$g_j$の２次関連フレーズ）の内の何れかが文書ｄにあるか。

ａ）文書の主題を判定

【０１３２】

フレーズによる文書のインデックス化、およびクラスター情報の使用により更に、関連フレーズ情報を基に、文書が関係する主題を判定できるというインデックス化システム１１

０の別の利点を提供する。
【０１３３】
所与の良好フレーズ$g_j$、および所与の文書ｄを仮定すると、ポスティングリストエントリは以下の通りである：
【０１３４】
$g_j$：文書ｄ：関連フレーズカウント：＝{3,4,3,0,0,2,1,1,0}
【０１３５】
関連フレーズビットベクトル：＝{11 11 10 00 00 10 10 10 01}
【０１３６】
ここで、関連フレーズビットベクトルはバイビット対で示される。
【０１３７】
関連フレーズビットベクトルから、文書ｄの一次および２次の主題を判定できる。一次主題はビット対（１、１）により示され、２次主題はビット対（１、０）により示される。（１、１）の関連フレーズビット対が示すのは、ビット対に対する関連フレーズ$g_k$が文書ｄにあることと、２次関連フレーズ$g_l$も同様であるという両方である。これにより解釈できる意味は、文書ｄの著者が、文書の草案において幾つかの関連フレーズ$g_j$、$g_k$、および$g_l$をともに用いた、ということである。（１、０）のビット対により示されることは、$g_j$および$g_k$はともに存在するが、$g_k$からの２次関連フレーズはそれ以上なく、従って、これはあまり重要な主題ではないということである。
ｂ）改良型ランク付けのための文書注釈
【０１３８】
インデックス化システム１１０の更なる局面では、後続の検索中に改良型ランク付けを提供する情報によりインデックス化処理中の各文書ｄに注釈５０４を付けることができる。注釈処理５０６は以下の通りである。
【０１３９】
文書コレクションの所与の文書ｄは、他の文書への幾つかのアウトリンクを有することができる。各アウトリンク（ハイパーリンク）には、アンカーテキスト、および目標文書の文書識別子が含まれる。説明のために、処理中の現在の文書ｄを、ＵＲＬ０と呼び、文書ｄのアウトリンクの目標文書をＵＲＬ１と呼ぶことにする。幾つかの他のＵＲＬｉを指し示すＵＲＬ０の全てのリンクについて、検索結果におけるランク付き文書で後で用いるために、インデックス化システム１１０は、ＵＲＬ０に関するリンクのアンカーフレーズに対するアウトリンクスコア、およびＵＲＬｉに関するアンカーフレーズに対するインリンクスコアを作成する。すなわち、文書コレクションの各リンクは一対のスコア、すなわちアウトリンクスコア、およびインリンクスコアを有する。これらのスコアは以下のように計算する：
【０１４０】
所与の文書ＵＲＬ０について、インデックス化システム１１０は、アンカーテキストＡが良好フレーズリスト２０８のフレーズである別の文書ＵＲＬ１への各アウトリンクを識別する。図８ａに、文書ＵＲＬ０のアンカーテキスト「Ａ」をハイパーリンク８００で用いる場合の関係を略図で示す。
【０１４１】
フレーズＡのポスティングリストでは、ＵＲＬ０がフレーズＡのアウトリンクとして書き込まれ、ＵＲＬ１がフレーズＡのインリンクとして書き込まれる。ＵＲＬ０に対して、関連フレーズビットベクトルが上記のように完成されて、ＵＲＬ０にある関連フレーズ、および２次関連フレーズが識別される。この関連フレーズビットベクトルは、アンカーフレーズＡを含むＵＲＬ０からＵＲＬ１へのリンクに対するアウトリンクスコアとして用いられる。
【０１４２】
次に、インリンクスコアを以下のように決定する。アンカーフレーズを含むＵＲＬ１への各リンクに対して、インデックス化システム１１０は、ＵＲＬ１を走査し、フレーズＡが

ＵＲＬ１の本文に出現するかどうかを判定する。フレーズＡが（ＵＲＬ０のアウトリンクを経由して）ＵＲＬ１を指すだけでなく、ＵＲＬ１自体の内容に出現する場合、これにより示されるのは、ＵＲＬ１がフレーズＡによって表される概念にＵＲＬ１が包括的に関連していると言うことができる。図８ｂは、この場合を示し、フレーズＡは、ＵＲＬ０（アンカーテキストとして）、およびＵＲＬ１の本文の両方に出現する。この場合、ＵＲＬ１のフレーズＡに対する関連フレーズビットベクトルは、フレーズＡを含むＵＲＬ０からＵＲＬ１へのリンクに対するインリンクスコアとして用いられる。

【０１４３】

アンカーフレーズＡがＵＲＬ１本文に出現しない場合（図８ａのように）、異なるステップをとってインリンクスコアを決定する。この場合、インデックス化システム１１０は、（フレーズＡがあたかもＵＲＬ１にあるかのように）フレーズＡのＵＲＬ１に対する関連フレーズビットベクトルを作成し、フレーズＡのどの関連フレーズがＵＲＬ１に出現するかを示す。次に、この関連フレーズビットベクトルは、ＵＲＬ０からＵＲＬ１へのリンクに対するインリンクスコアとして用いられる。

【０１４４】

例えば、以下のフレーズが最初にＵＲＬ０およびＵＲＬ１に出現すると仮定する：

【表２】

| | アンカーフレーズ | 関連フレーズビットベクトル | | | | |
|---|---|---|---|---|---|---|
| 文書 | Australian Shepherd | Aussie | blue merle | red merle | tricolor | agility training |
| URL0 | 1 | 1 | 0 | 0 | 0 | 0 |
| URL1 | 1 | 0 | 1 | 1 | 1 | 0 |

【０１４５】

（上の表および以下の表には、２次関連フレーズビットは示されていない）。ＵＲＬ０の行は、アンカーテキストＡからのリンクのアウトリンクスコアであり、ＵＲＬ１の行は、リンクのインリンクスコアである。ここで、ＵＲＬ０はＵＲＬ１を目標とするアンカーフレーズ「Australian Shepard（オーストラリアンシェパード）」を含む。「Australian Shepard」の５つの関連フレーズのうち、「Aussie（オーストラリアの）」だけがＵＲＬ０に出現する。従って、ＵＲＬ０だけがAustralian Shepherdsに関して関連が薄いことが直感的に分かる。それと比較して、ＵＲＬ１には、文書本文に出現するフレーズ「Australian Shepherd」があるだけでなく、「blue merle（青灰色の地に青斑点）」、「red merle（青灰色の地に赤斑点）」、および「tricolor（三色）」のように多くの関連フレーズが出現する。従って、アンカーフレーズ「Australian Shepard」は、ＵＲＬ０およびＵＲＬ１の両方に出現するので、ＵＲＬ０に対するアウトリンクスコア、およびＵＲＬ１に対するインリンクスコアは、それぞれ上に示す行となる。

【０１４６】

上記第２の場合は、アンカーフレーズＡがＵＲＬ１に出現しない場合である。その場合は、インデックス化システム１１０はＵＲＬ１を走査し、関連フレーズ「Aussie」、「blue merle」、「red merle」、「tricolor」、および「agility training（犬の障害物競争訓練）」のうちのどれがＵＲＬ１に出現するかを判定し、それにより、関連フレーズビットベクトルを作成する。例えば：

【表３】

| 文書 | アンカーフレーズ | 関連フレーズビットベクトル | | | | |
|---|---|---|---|---|---|---|
| | Australian Shepherd | Aussie | blue merle | red merle | tricolor | agility training |
| URL0 | 1 | 1 | 0 | 0 | 0 | 0 |
| URL1 | 0 | 0 | 1 | 1 | 1 | 0 |

【０１４７】
ここに示されることは、ＵＲＬ１がアンカーフレーズ「Australian Shepard」を含まないが、関連フレーズ「blue merle」、「red merle」、および「tricolor」を含むということである。
【０１４８】
この手法は、検索結果を歪曲するためのウェブページ（文書の一種）のある種の操作を全体に防ぐという利点がある。所与の文書を指すリンク数に依って文書に順位を付けるランク付けアルゴリズムを用いる検索エンジンを、その後所望のページを指す所与のアンカーテキストを持つ大量のページを人工的に作成する、ことにより「爆撃」することができる。その結果、アンカーテキストを用いる検索クエリーが入力された場合、実際にそのページがアンカーテキストに関することをほとんど、または全く含まなくても、所望のページを返すのが普通である。文書ＵＲＬ０のフレーズＡ関連フレーズビットベクトルに、目標文書ＵＲＬ１からの関連ビットベクトルをインポートすることにより、検索システムが、重要性のインジケータとして、ＵＲＬ１を指すＵＲＬ０のフレーズＡ、またはアンカーテキストフレーズに対するＵＲＬ１の関係だけに頼ることが排除される。
【０１４９】
インデックス１５０の各フレーズは、コーパスに発生する頻度に基づくフレーズ番号も与えられる。フレーズが共通になればなるほど、インデックスの順位を受け取るフレーズ番号が小さくなる。次いで、インデックス化システム１１０は、各ポスティングリスト内のフレーズ番号をリストアップした文書の数に従って降順で、インデックス１５０中のポスティングリストを全てソートする（５０６）。それにより、最も発生頻度が高いフレーズを最初にリストアップする。次いで、フレーズ番号を用いて特定フレーズを参照できる。

III．検索システム

【０１５０】
検索システム１２０は、クエリーに関する文書に対するクエリーおよび検索を受け取るよう動作し、一組の検索結果における（文書にリンクする）これらの文書のリストを提供する。図６は、検索システム１２０の主な機能的動作を示す。
【０１５１】
６００：クエリー内のフレーズを識別する。
【０１５２】
６０２：クエリーフレーズに関する文書を検索する。
【０１５３】
６０４：検索結果の文書をフレーズによりランク付けする。
【０１５４】
これら段階のそれぞれの詳細を以下に示す：

１．クエリーおよびクエリー拡張のフレーズ識別

【０１５５】
検索システム１２０の第１段階６００は、インデックスを効率的に検索するためにクエリー内に存在するフレーズを識別することである。以下の用語は本セクションで用いる：
【０１５６】
ｑ：検索システム１２０によって入力され、受け取られるクエリー。
【０１５７】

Ｑｐ：クエリー内に存在するフレーズ。
【０１５８】
Ｑｒ：Ｑｐの関連フレーズ。
【０１５９】
Ｑｅ：Ｑｐのフレーズ拡張。
【０１６０】
Ｑ：ＱｐおよびＱｒの連合。
【０１６１】
クエリーｑは、クライアント１９０から受け取り、ある最大数未満の文字またはワードを有する。
【０１６２】
検索システム１２０はサイズＮ（例えば、５）のフレーズ窓を用いて、クエリーｑの用語をトラバースする。フレーズ窓はクエリーの第１語により開始し、右にＮ語拡げる。次いで、この窓は、右にＭ－Ｎ回シフトする。ここでＭは、クエリー中の用語の数である。
【０１６３】
各窓位置で、窓にＮ語（または、それ未満）があることになる。これらの用語は見込クエリーフレーズを構成する。見込フレーズは良好フレーズリスト２０８で参照して、良好フレーズかどうかを判定する。見込フレーズが良好フレーズリスト２０８にある場合、フレーズ番号がフレーズに返され、見込フレーズをここで候補フレーズとする。
【０１６４】
各窓の全ての見込フレーズを検査して、それらが良好候補フレーズであるかどうかを判定してから、検索システム１２０は、クエリー内の対応するフレーズに対するフレーズ番号集合を有することになる。次いで、これらのフレーズ番号をソートする（降順で）。
【０１６５】
第１候補フレーズとして最大フレーズ番号で開始し、検索システム１２０は、別の候補フレーズが、ソートリストの固定数距離内にあるかどうか、すなわち、フレーズ番号間の差が、閾値の量、例えば、２０，０００内にあるかどうかを判定する。もしあれば、クエリーの最左端にあるフレーズを妥当なクエリーフレーズＱｐとして選択する。このクエリーフレーズおよびそのサブフレーズ全てを候補のリストから削除し、リストを再ソートし、その処理を繰り返す。この処理結果は妥当なクエリーフレーズＱｐ集合である。
【０１６６】
例えば、検索クエリーが「Hillary Rodham Clinton Bill on the Senate Floor（上院議会でのヒラリー・ロドハム・クリントン・ビル）」であると仮定する。検索システム１２０は以下の候補フレーズ、「Hillary Rodham Clinton Bill on」、「Hillary Rodham Clinton Bill」、および「Hillary Rodham Clinton」を識別する。最初の２つは破棄し、最後の１つを妥当なクエリーフレーズとして保持する。次に検索システム１２０は、「Bill on the Senate Floor」、およびサブフレーズ「Bill on the Senate」、「Bill on the」、「Bill on」、「Bill」を識別し、妥当なクエリーフレーズＱｐとして「Bill」を選択する。最後に検索システム１２０は、「on the Senate Floor」を分析し、「Senate Floor」を妥当なクエリーフレーズとして識別する。
【０１６７】
次に、検索システム１２０は、大文字使用について妥当フレーズＱｐを調整する。クエリーを分析する場合、検索システム１２０は、妥当フレーズ毎に可能性のある大文字使用を識別する。これは、「united states」を「United States」と大文字化する等の、既知の大文字使用表を用いて、または文法に基づく大文字化アルゴリズムを用いて成される。これは適切な大文字化クエリーフレーズ集合を生成する。
【０１６８】
次いで、検索システム１２０は、大文字化フレーズへ第２の通過を行い、フレーズおよびそのサブフレーズが共に集合にある場合、これらのフレーズが左端にあって、大文字の場合だけを選択する。例えば、「president of the united states」の検索は「President

of the United States」と大文字化する。

【０１６９】

次の段階では、検索システム１２０は、クエリーフレーズＱに関する文書を識別する（６０２）。次いで、検索システム１２０は、クエリーフレーズＱのポスティングリストを検索し、これらのリストの共通部分を求めて、どの文書がクエリーフレーズに対する全て（または幾つかの）のポスティングリストに出現するかを判定する。クエリー内のフレーズＱがフレーズ拡張Ｑｅの集合をもつ場合（以下に更に説明するように）、検索システム１２０は、最初にフレーズ拡張のポスティングリスト連合を形成してから、ポスティングリストにより共通部分を求める。検索システム１２０は、上記のように不完全フレーズリスト２１６の各クエリーフレーズＱを参照することにより、フレーズ拡張を識別する。

【０１７０】

共通部分を求めた結果は、クエリーに関する文書集合である。フレーズおよび関連フレーズによる文書インデックス化、クエリー内のフレーズＱの識別、およびフレーズ拡張を含むクエリー拡張は、クエリー語を含む文書しか選択されない従来のブール型検索システムで結果的に得られる場合よりも、クエリーに関係が深い文書集合の選択が得られる。

【０１７１】

一実施の形態では、検索システム１２０は、クエリーフレーズＱのポスティングリスト全ての共通部分を求めずに、最適化メカニズムを用いてクエリーに応じた文書を識別できる。インデックス１５０の構造の結果として、各フレーズ$g_j$に対して、関連フレーズ$g_k$は、$g_k$に対する関連フレーズビットベクトル内で既知であり、識別される。従って、この情報を用いて、２つ以上のクエリーフレーズが相互に関連するフレーズであるか、または共通の関連フレーズを有する場合、共通部分を求める処理をショートカットできる。これらの場合、関連フレーズビットベクトルは、直接アクセスでき、次いで、それを用いて対応する文書を検索できる。この処理は、以下に更に完全に説明する。

【０１７２】

任意の２つのクエリーフレーズＱ１およびＱ２を考えると、その関係には３つの可能性がある。

【０１７３】

１）Ｑ２はＱ１の関連フレーズである。

【０１７４】

２）Ｑ２はＱ１の関連フレーズではなく、それぞれの関連フレーズＱｒ１、およびＱｒ２は共通部分をもたない（すなわち、関連フレーズが共通しない）。

【０１７５】

３）Ｑ２はＱ１の関連フレーズではないが、それぞれの関連フレーズＱｒ１およびＱｒ２は共通部分をもつ。

【０１７６】

クエリーフレーズの各対に対して、検索システム１２０は、クエリーフレーズＱｐの関連フレーズビットベクトルを参照することによりどの場合が適切かを判定する。

【０１７７】

Ｑ１を含む文書を含み、これらの文書それぞれについて、関連フレーズビットベクトルを含むクエリーフレーズＱ１に対するポスティングリストを検索することにより、検索システム１２０は進行する。Ｑ１に対する関連フレーズビットベクトルは、フレーズＱ２（および、もしあれば、残りの各クエリーフレーズ）がＱ１の関連フレーズであり、文書にあるかどうかを示す。

【０１７８】

第１の場合がＱ２に適用されるのは、検索システム１２０が、Ｑ１のポスティングリストにある各文書ｄについての関連フレーズビットベクトルを走査して、Ｑ２に対するビット集合をもつかどうかを判定する場合である。このビットがＱ１のポスティングリストにある文書ｄに対して設定されていない場合、Ｑ２がその文書に出現しないことを意味する。その結果、この文書は直ちにそれ以上の考察から排除される。次いで、残りの文書につい

てスコアを付ける。これが更に意味することは、Ｑ２のポスティングリストを処理して、どの文書にあるのかを検索システム１２０で調べるのが不要となり、従って、計算時間を節約できるということである。

【０１７９】

第２の場合がＱ２に適用されると、２つのフレーズが互いに関連しない。例えば、クエリー「cheap bolt action rifle（安物のボルトアクション型ライフル）」は、２つのフレーズ「cheap」および「bolt action rifle」を有する。これらのフレーズのどちらも互いに関連せず、更にこれらそれぞれの関連フレーズは重ならない。すなわち、「cheap（安物）」は、関連フレーズ「low cost（低価格）」、「inexpensive（安価な）」、「discount（値引き）」、「bargain basement（地下特売場）」、および「lousy（お粗末な）」を有し、一方、「bolt action rifle」は、リストが共通部分を有しない関連フレーズ「gun（銃）」、「22 caliber（２２口径）」、「magazine fed（供給されたマガジン）」、および「Armalite AR-30M（アーマライトＡＲ－３０Ｍ）」を有する。この場合、検索システム１２０は、Ｑ１およびＱ２のポスティングリストの規則的な共通部分を求めて、スコア付けするための文書を得る。

【０１８０】

第３の場合が適用されるのは、関連していないが、少なくとも１つの関連フレーズを共通にもつ２つのフレーズＱ１およびＱ２の場合である。例えば、フレーズ「bolt action rifle」および「22」は共に関連フレーズとして「gun」を有する。この場合、検索システム１２０は、両フレーズＱ１およびＱ２のポスティングリストを検索し、リストの共通部分を求めて両フレーズを含む文書リストを生成する。

【０１８１】

次いで、検索システム１２０は、得られる文書それぞれを迅速にスコア付けすることができる。最初に、検索システム１２０は、各文書についてのスコア調整値を決定する。スコア調整値は、文書に対する関連フレーズビットベクトルのクエリーフレーズＱ１およびＱ２に対応する位置のビットから形成されるマスクである。例えば、仮定として、Ｑ１およびＱ２が、文書ｄに対する関連フレーズビットベクトルの第３および第６のバイビット位置に対応し、第３位置のビット値は（１、１）、第６対のビット値は（１、０）であり、スコア調整値はビットマスク「00 00 11 00 00 10」であるとする。次いで、スコア調整値を用いて、文書に対する関連フレーズビットベクトルをマスクし、修正フレーズビットベクトルを、文書に対する本文スコアを計算するのに用いられるランク付け機能に渡す（次で説明する）。

２．ランク付け

ａ）含まれるフレーズに基づく文書ランク付け

【０１８２】

検索システム１２０は、クエリーフレーズに対する各文書の関連フレーズビットベクトル、およびクラスタービットベクトルのフレーズ情報を用いて、検索結果の文書をランク付けするランク付け段階６０４を提供する。この手法は、文書に含まれるフレーズ、または略称「本文ヒット」により、ランク付けする。

【０１８３】

上記のように、任意の所与のフレーズ$g_j$に対して、$g_j$のポスティングリストにある各文書ｄは、どの関連フレーズ$g_k$、およびどの２次関連フレーズ$g_l$が文書ｄにあるかを識別する関係付けられた関連フレーズビットベクトルを有する。関連フレーズおよび２次関連フレーズが所与の文書に多くあればあるほど、所与のフレーズに対する文書の関連フレーズビットベクトルに設定されるビットは多くなる。設定されるビットが多くなればなるほど、関連フレーズビットベクトルの数値は大きくなる。

【０１８４】

従って、一実施の形態では、検索システム１２０は、関連フレーズビットベクトルの値により検索結果内の文書をソートする。クエリーフレーズＱに最も関連するフレーズを含む文書は、最大値の関連フレーズビットベクトルを有し、これら文書は、検索結果内の最高

ランク文書となる。
【０１８５】
この手法は、これら文書がクエリーフレーズに関して意味的に最も主題に富むので望ましい。注記すべきは、この手法は、文書が高頻度の入力クエリー語ｑを含まない場合でも、関連フレーズ情報を用いて関連文書を識別し、次いで、これら文書をランク付けするので、高度に関連する文書を提供する。低頻度の入力クエリー語をもつ文書が、依然としてクエリー語およびフレーズに関連する多数のフレーズを有し、従って、高頻度のクエリー語およびフレーズだけを有し、関連フレーズをもたない文書より関連が高くなることもある。
【０１８６】
第２の実施の形態では、検索システム１２０は、それが含むクエリーフレーズＱの関連フレーズにより結果集合にある各文書をスコア付けする。これは以下のようになされる：
【０１８７】
各クエリーフレーズＱについて考えると、フレーズ識別処理中に識別されるような、関連フレーズに関連する幾つかの数Ｎのフレーズ Ｑｒがある。上記のように、関連クエリーフレーズＱｒは、クエリーフレーズＱからの情報ゲインによりランク付けられる。次いで、これらの関連フレーズは、最初の関連フレーズＱｒ１（すなわち、Ｑによる最大情報ゲインをもつ関連フレーズＱｒ）に対するＮポイントにより開始されるポイント、次いで、次の関連フレーズＱｒ２に対するＮ－１ポイント、次いで、Ｑｒ３に対するＮ－２ポイント等を割り当てられ、それにより、最後の関連フレーズＱｒＮが１ポイントを割り当てられる。
【０１８８】
次いで、検索結果内の各文書は、クエリーフレーズＱのどの関連フレーズＱｒがあるかを決定することにより、そしてそのような関連フレーズＱｒそれぞれに割り当てられるポイントを文書に与えることによりスコア付けされる。次いで、文書は最大値から最小値のスコアによりソートされる。
【０１８９】
更なる改良として、検索システム１２０は、結果集合から特定文書を抜粋できる。場合によっては、文書は多くの異なる主題に関連し、特に長い文書ほど関連する。多くの場合、ユーザーは、多くの様々な主題に関する文書全体に、クエリーにおいて表現される単一の主題に関する強い示唆がある文書を好む。
【０１９０】
これら後者の種類の文書を抜粋するために、検索システム１２０は、クエリーフレーズのクラスタービットベクトル内のクラスタ情報を用い、文書のクラスタが閾値数を超える文書は削除する。例えば、検索システム１２０は、２クラスタを超えるクラスタを含む文書を削除できる。このクラスタ閾値は、検索パラメータとしてユーザーが事前に決定、または設定できる。
ｂ）アンカーフレーズに基づく文書のランク付け
【０１９１】
一実施の形態におけるクエリーフレーズＱの本文ヒットに基づく検索結果内の文書をランク付けするのに加えて、検索システム１２０は、他の文書へのアンカーのクエリーフレーズＱおよび関連フレーズＱｒの出現に基づいて文書のランク付けも行う。一実施の形態では、検索システム１２０は、２種類のスコア、つまり本文ヒットスコア、およびアンカーヒットスコアの関数（例えば、線形結合）であるスコアをそれぞれの文書に対して計算する。
【０１９２】
例えば、所与の文書に対する文書スコアは、以下のように計算できる：
【０１９３】
スコア＝.30＊（本文ヒットスコア）＋.70＊（アンカーヒットスコア）。
【０１９４】

.30および.70の重み付けは、要望に応じて調整できる。文書に対する本文ヒットスコアは、クエリーフレーズＱｐを考えると、上記方法では、文書に対する最大値の関連フレーズビットベクトルの数値である。代替として、この値は、インデックス１５０の各クエリーフレーズＱを参照し、クエリーフレーズＱのポスティングリストから文書にアクセスし、次いで、関連フレーズビットベクトルにアクセスすることにより、検索システム１２０が直接取得できる。

【０１９５】

文書ｄのアンカーヒットスコアは、クエリーフレーズＱの関連フレーズビットベクトルの関数であり、ここでＱは、文書ｄを参照する文書のアンカー語である。インデックス化システム１１０が文書コレクション内の文書をインデックス化する場合、それは、フレーズがアウトリンクのアンカーテキストである文書のリストを各フレーズに対して保持し、同様に、各文書に対して他の文書からのインリンクのリスト（および関係付けられたアンカーテキスト）を保持する。文書に対するインリンクは、他の文書（参照している文書）から、所与の文書への参照（例えば、ハイパーリンク）である。

【０１９６】

次いで、所与の文書ｄに対するアンカーヒットスコアを決定するために、検索システム１２０は、アンカーフレーズＱによりインデックスにリストアップされた参照文書Ｒ（ｉ＝１から参照文書の数まで）の集合全体にわたって繰り返し、以下の積を合計する：

【０１９７】

$R_i$.Ｑ.関連フレーズビットベクトル＊Ｄ.Ｑ．関連フレーズビットベクトル。

【０１９８】

ここで積算値は、アンカーフレーズＱが文書Ｄに対してどれ位主題性があるかのスコアである。このスコアは、本明細書では「インバウンドスコア成分」と呼ぶ。この積は、参照文書Ｒのアンカーフレーズの関連ビットベクトルにより現在の文書Ｄの関連ビットベクトルに効率的に重み付けをする。参照文書Ｒ自体がクエリーフレーズＱに関連している（つまり、より高い値の関連フレーズビットベクトルを有する）場合、これにより、現在の文書Ｄのスコアは高くなる。次いで、上記のように、本文ヒットスコア、およびアンカーヒットスコアを結合して文書点数を作成する。

【０１９９】

次に、参照文書Ｒそれぞれに対して、各アンカーフレーズに対する関連フレーズビットベクトルを取得する。これは、アンカーフレーズＱが文書Ｒにどれ位主題性があるかの尺度である。この値は、本明細書ではアウトバウンドスコア成分と呼ぶ。

【０２００】

次いで、インデックス１５０から、アンカーフレーズＱに対する全ての（参照している文書、参照される文書の）対が抽出される。次いで、これらの対は、関係付けられる（アウトバウンドスコア成分、インバウンドスコア成分）値によりソートされる。実施の形態によっては、これらの成分のどちらかが一次ソートキーとなり、他方が２次ソートキーとなる。次いで、ソート結果をユーザーに提示する。アウトバウンドスコア成分で文書をソートすることにより、アンカーヒットとしてクエリーに対して多くの関連フレーズを有する文書を最上位にランク付けするので、これらの文書を「エキスパート」文書と表す。インバウンド文書スコアでソートすることにより、アンカー語により高頻度で参照される文書を最上位にランク付ける。

３．フレーズに基づく検索個人化

【０２０１】

検索システム１２０の別の局面は、ユーザー特定の関心のモデルによる検索結果のランク付けを個人化する（６０６）能力、つまりカスタム化する能力である。この方法では、ユーザーの関心に大きく関連している可能性がある文書の方が、検索結果で高位にランク付けられる。検索結果の個人化は以下の通りである。

【０２０２】

事前に、ともにフレーズで表すことができるクエリーおよび文書の観点から、ユーザーの

関心を定義すると便利である（例えば、ユーザーモデル）。入力の検索クエリーに対して、クエリーは、クエリーフレーズＱ、関連フレーズＱｒ、およびクエリーフレーズＱｐのフレーズ拡張Ｑｅにより表わされる。従って、用語およびフレーズのこの集合はクエリーの意味を表す。次に、文書の意味は、ページと関係付けられるフレーズにより表わされる。上記のように、クエリーおよび文書を考えると、文書に対する関連フレーズは、その文書にインデックス化される全フレーズに対する本文点数（関連ビットベクトル）により決定される。最終的に、文書集合をもつクエリー集合の連合として、これらの各要素を表すフレーズによりユーザーを表すことができる。ユーザーを表す集合に含まれる特定文書は、ユーザーの動作および宛先を監視するクライアント側ツールを用いて、以前の検索結果、またはコーパスの一般的な閲覧（例えば、インターネットの文書にアクセスすること）において、どの文書をユーザーが選択したかにより決定できる。

【０２０３】

個人化ランク付けのためのユーザーモデルを構築し、使用する処理は以下の通りである。

【０２０４】

最初に、所与のユーザーについて、最後のＫ個のクエリー、およびアクセスされるＰ個の文書のリストが保持される。ここでＫおよびＰは、それぞれ２５０程度が好ましい。リストは、ログイン、またはブラウザクッキーによりユーザーが認証されるユーザーアカウントのデータベースに保持してもよい。所与のユーザーについて、ユーザーがクエリーを提供する初回には、リストは空である。

【０２０５】

次に、クエリーｑをユーザーから受け取る。ｑの関連フレーズＱｒをフレーズ拡張とともに、上記の方法で検索する。これはクエリーモデルを形成する。

【０２０６】

最初の通過（例えば、ユーザーに対する格納クエリー情報がない場合）では、検索システム１２０は、検索結果のユーザークエリーに関連した文書を単に返すだけで、それ以上のカスタム化のランク付けをせずに動作する。

【０２０７】

クライアント側ブラウザツールは、ユーザーが、検索結果の内のどの文書に、例えば、検索結果の文書リンクのクリック等によりアクセスするかを監視する。これらのアクセス文書は、どのフレーズがユーザーモデルの一部となるかを選択する基準となる。このようなアクセス文書それぞれに対して、検索システム１２０は、文書に関連するフレーズのリストである文書モデルをその文書の代わりに検索する。アクセス文書に関連する各フレーズは、ユーザーモデルに追加される。

【０２０８】

次に、アクセス文書に関連するフレーズを考えると、これらのフレーズと関係付けられるクラスタは、各フレーズに対するクラスタービットベクトルにより決定される。各クラスタに対して、クラスタのメンバーである各フレーズは、クラスタ番号、または上記のようなクラスタービットベクトル表現を含む関連フレーズ表で、フレーズを参照することにより決定する。次いで、クラスタ番号をユーザーモデルに加える。更に、このような各クラスタではカウンタ値が保持され、そのクラスタのフレーズがユーザーモデルに追加されるたびに値が１つ増加する。これらのカウント値を以下に説明するように、重み付けに用いる。従って、ユーザーモデルは、ユーザーが文書にアクセスすることにより関心を示した文書にあるクラスタに含まれるフレーズにより作成される。

【０２０９】

同様な一般的手法をもっと正確に絞り込んで、ユーザーが文書に単にアクセスする以上に関心が高いレベルを示すフレーズ情報を取得することができる（ユーザーは実際にその文書が関連しているかどうかを判定するだけでよい）。例えば、ユーザーモデルに収集したフレーズは、ユーザーが印刷した文書、保存した文書、お気に入りに格納した文書もしくはリンクした文書、別のユーザーにメールした文書、またはある時間以上（例えば、１０分間）ブラウザ窓に開いたままの文書に限ってもよい。これらの、およびその他の動作は

、文書の関心レベルが高いことを示す。

【０２１０】

別のクエリーをユーザーから受け取った場合、関連クエリーフレーズＱｒを検索する。これらの関連クエリーフレーズＱｒと、ユーザーモデルにリストアップされるフレーズとの共通部分を求めて、どのフレーズがクエリーおよびユーザーモデルの両方にあるかを判定する。マスクビットベクトルをクエリーＱｒの関連フレーズに対して初期化する。このビットベクトルは上記のようにビット－ビットベクトルである。ユーザーモデルにも現れるクエリーの各関連フレーズＱｒに対して、この関連フレーズに対する両ビットをマスクビットベクトルに設定する。従って、マスクビットベクトルは、クエリーおよびユーザーモデル両方に現れる関連フレーズを表す。

【０２１１】

次いで、マスクビットベクトルを用いて、関連フレーズビットベクトルとマスクビットベクトルとの論理積をとることにより、検索結果の現在の集合にある各文書に対する関連フレーズビットベクトルをマスクする。これは本文スコアおよびアンカーヒットスコアをマスクビットベクトルにより調整する効果をもつ。次いで、先のように本文スコアおよびアンカースコアについて文書をスコア付けし、ユーザーに提示する。この手法では、文書が高位にランク付けられるためには、ユーザーモデルに含まれるクエリーフレーズをもつことが基本的に必要である。

【０２１２】

前述の厳しい制約を与えない代替の実施の形態として、マスクビットベクトルは、各ビットを用いて、ユーザーモデルの関連フレーズに対するクラスタカウントに重み付けできるように、アレイに入れることができる。つまり、それぞれのクラスタカウントに０または１を乗じて、効率的にカウントをゼロにし、または保持する。次に、これらのカウント自体を重み付けに用い、スコア付けしている各文書に対して関連フレーズを乗じるのにも用いる。この手法は関連フレーズとしてクエリーフレーズをもたない文書でも適切にスコア付けできる利点を有する。

【０２１３】

最終的に、ユーザーモデルは、検索のアクティブ時間を時間間隔とする現在のセッションに限定してもよく、そのセッションのあとでユーザーモデルを放出する。代替として、所与のユーザーに対するユーザーモデルは、全時間にわたって持続して、そのあと重み付けを下げるか、またはやめてもよい。

Ⅳ．結果提示

【０２１４】

提示システム１３０は、検索システム１２０から、スコア付けし、ソートした検索結果を受け取り、整理、注釈、およびクラスタ操作を更に実行してからユーザーに結果を提示する。これらの操作は、検索結果内容のユーザー理解を助け、複製を削除し、そして検索結果のより代表的なサンプリングを提供する。図７は、提示システム１３０の主な機能的操作を示す。

【０２１５】

７００：主題クラスタにより文書をクラスタ化する。

【０２１６】

７０２：文書説明を生成する。

【０２１７】

７０４：複製文書を削除する。

【０２１８】

これらの操作は、それぞれ入力として検索結果７０１をとり、修正した検索結果７０３を出力する。図７で示唆するように、これら操作の順序は独立し、所与の実施の形態で要望に応じて変更でき、従って、入力は図に示すような並列式の代わりにパイプライン式としてもよい。

１．提示のための動的ソート生成

【０２１９】
所与のクエリーについて、クエリーを満たす数百の文書を、たとえそれが数千であっても、返すのが普通である。多くの場合、特定の文書は、互いに異なる内容をもつが関連性が十分あり、関連文書の意味のあるグループ、基本的にクラスタ、を形成している。しかしながら、ほとんどのユーザーは、検索結果の最初の３０または４０文書を超えてまで検討しない。従って、最初の１００文書が、例えば、３クラスタからもたらされるが、次の１００文書が追加の４クラスタを表す場合、それ以上の調整をせずに、ユーザーは、クエリーに関連する多様な主題を実際に表すので、ユーザーのクエリーに完全に関連しているかもしれないこれら後者の文書を検討しないのが普通である。従って、ここで各クラスタからの文書サンプルをユーザーに提供して、ユーザーが検索結果からの様々な文書をより広く選択できるようにすることが望まれている。提示システム１３０はこれを以下のように行う。
【０２２０】
システム１０の他の局面として、提示システム１３０は、検索結果の各文書ｄに対する関連フレーズビットベクトルを利用する。より詳細には、各クエリーフレーズＱに対して、およびＱのポスティングリストの各文書ｄに対して、関連フレーズビットベクトルは、どの関連フレーズＱｒがその文書にあるかを示す。検索結果の文書集合全体にわたって、次いで、各関連フレーズＱｒに対して、Ｑｒに対応するビット位置のビット値を加算することにより、幾つの文書が関連フレーズＱｒを含むかについてカウントを判定する。検索結果全体にわたって合計し、ソートすると、最高頻度で発生する関連フレーズＱｒが示され、それぞれは、文書のクラスタである。最高頻度で発生する関連フレーズは、その名称としてその関連フレーズＱｒをとる第１クラスタであり、上位３～５クラスタについて以下同様である。このように、クラスタの名称またはヘッダとしてフレーズＱｒを伴いつつ各上位クラスタが識別される。
【０２２１】
ここで、各クラスタからの文書は、様々な方法でユーザーに提示できる。１つの用途では、各クラスタからの固定数の文書、例えば、各クラスタの上位スコア１０文書を提示してもよい。別の用途では、各クラスタからの比例数の文書を提示してもよい。従って、検索結果に１００文書があり、クラスタ１に５０文書、クラスタ２に３０文書、クラスタ３に１０文書、クラスタ４に７文書、およびクラスタ５に３文書あり、２０文書だけが提示を要望されている場合、文書は以下のように選択される：クラスタ１から１０文書、クラスタ２から７文書、クラスタ３から２文書、およびクラスタ４から１文書である。次いで、文書をユーザーに示すことができ、ヘッダとしての適切なクラスタ名称のもとでしかるべくグループ化される。
【０２２２】
例えば、検索システム１２０が１００文書を検索する「blue merle agility training」の検索クエリーを仮定する。検索システム１２０は、すでにクエリーフレーズとして「blue merle」および「agility training」を識別している。これらのクエリーフレーズの関連フレーズは次の通りである:
【０２２３】
「blue merle」::「Australian Shepherd」,「red merle」,「tricolor」,「aussie」;
【０２２４】
「agility training」::「weave poles(編んだ柱)」,「teeter（シーソー）」,「tunnel（トンネル）」,「obstacle(障害)」,「border collie(ボーダーコリー）」。
【０２２５】
次いで、提示システム１３０は、各クエリーフレーズの上記関連フレーズそれぞれについて、このようなフレーズを含む文書数のカウントを決定する。例えば、仮定として、フレーズ「weave poles」は１００文書のうちの７５文書に出現し、「teeter」は６０文書に出現し、「red merle」は５０文書に出現するとする。第１クラスタは「weave poles」と

名称が付けられ、そのクラスタからの選択文書数が提示され；第２クラスタは「teeter」と名称が付けられ、そのクラスタからの選択文書数が同様に提示され、以下同様である。固定した提示では、各クラスタからの１０文書が選択されてもよい。比例提示は、全文書数に対する各クラスタからの比例文書数を用いる。

２．主題に基づく文書説明

【０２２６】

提示システム１３０の第２の機能は、各文書の検索結果提示に挿入できる文書説明の作成（７０２）である。これらの説明は、各文書にある関連フレーズに基づき、従って、検索に文脈的に関連している方法で文書が関係していることへのユーザーの理解を助ける。文書説明は、一般化、またはユーザーに個人化されるかのどちらでもよい。

ａ）一般的な主題文書説明

【０２２７】

先に説明したように、クエリーを考えると、検索システム１２０は関連クエリーフレーズＱｒを決定し、クエリーフレーズのフレーズ拡張を同様に決定し、次いで、クエリーに対する関連文書を識別した。提示システム１３０は、検索結果の各文書にアクセスし、以下の操作を実行する。

【０２２８】

最初に、提示システム１３０は、文書の文をクエリーフレーズＱ、関連クエリーフレーズＱｒ、およびフレーズ拡張Ｑｐのインスタンス数によりランク付けし、文書の各文について、これら３つの局面のカウントを保持する。

【０２２９】

次いで、文をこれらカウントにより、クエリーフレーズＱのカウントである第１ソートキー、関連クエリーフレーズＱｒのカウントである第２ソートキー、そして最後にフレーズ拡張Ｑｐのカウントである最終ソートキーでソートする。

【０２３０】

最終的に、ソートに従う上位Ｎ（例えば、５）の文を文書の説明として用いる。この文集合は、フォーマット化でき、修正検索結果７０３の文書の提示に含めることができる。この処理を検索結果にある幾つかの文書数について繰り返し、ユーザーがその結果の次のページを要求する時にオンデマンドで成してもよい。

ｂ）個人化した主題に基づく文書説明

【０２３１】

検索結果の個人化が提供される実施の形態では、文書説明も同様に個人化でき、ユーザーモデルで説明したようにユーザーの関心を反映することができる。提示システム１３０はこれを以下のように行う。

【０２３２】

最初に、提示システム１３０は、上記のように、クエリー関連フレーズＱｒと、（ユーザーがアクセスした文書に発生するフレーズをリストアップした）ユーザーモデルとの共通部分を求めることにより、ユーザーに関連する関連フレーズを決定する。

【０２３３】

次いで、提示システム１３０は、ビットベクトル自体の値によりユーザー関連フレーズＵｒのこの集合を安定ソートし、そのソートリストをクエリー関連フレーズＱｒのリストの先頭に追加し、どの複製フレーズも削除する。安定ソートは、等しくランク付けたフレーズの既存の順位を保持する。これは、クエリーまたはユーザーに関連した関連フレーズ集合を生み出し、集合Ｑｕと呼ぶ。

【０２３４】

ここで、提示システム１３０は、検索結果にある各文書の文をランク付けるための基準として、上記の一般文書説明処理と類似の方法で、このフレーズ順位リストを用いる。従って、所与の文書に対して、提示システム１３０は、ユーザー関連フレーズ、およびクエリー関連フレーズＱｕそれぞれのインスタンス数により、文書の文をランク付けし、ランク付けした文をクエリーカウントによりソートし、最後に、このようなフレーズそれぞれに

対するフレーズ拡張数に基づいてソートする。以前は、ソートキーは、クエリーフレーズＱ、関連クエリーフレーズＱｒ、およびフレーズ拡張Ｑｐの順序であったが、ここでは、ソートキーは最高ランクから最低ランクのユーザー関連フレーズＵｒの順序となる。

【０２３５】

この処理を検索結果の文書について再度繰り返す（要求時、または事前に）。こうして、このような文書それぞれについて、得られた文書説明は、文書からの上位Ｎ位の文を備える。ここで、これらの文は、ユーザー関連フレーズＵｒの最大数をもつ１つであり、従って、ユーザーに最も関連する概念および主題（少なくともユーザーモデルで取得した情報による）を表す文書のキー文を表す。

３．複製文書の検出および削除

【０２３６】

インターネットなどの大規模コーパスでは、同一文書の多くのインスタンスがあるか、または多くの別の場所に文書の一部があるのが極く普通である。例えば、Associated Press等の報道局が生成した所与のニュース記事は、１ダース以上の個人新聞のウェブサイトに複製されている。検索クエリーへの応答に際して、これらの複製文書全てを含めるのは、冗長な情報でユーザーを悩ませるだけであり、クエリーへの応答は有益でない。従って、提示システム１３０は、互いに複製または複製に近いと思われる文書を識別し、検索結果にあるこれらの内の１つだけを含む更なる能力７０４を提供する。その結果、ユーザーは、ずっと多様化し、ロバストな結果集合を受取り、互いに複製された文書を検討する時間を浪費する必要がない。提示システム１３０は、以下の機能を提供する。

【０２３７】

提示システム１３０は、検索結果集合７０１にある各文書を処理する。文書ｄそれぞれについて、提示システム１３０は、文書と関係付けられる関連フレーズＲのリストを最初に決定する。これらの関連フレーズそれぞれについて、提示システム１３０は、これらのフレーズそれぞれの発生頻度により文書の文をランク付けし、次いで、上位Ｎ（例えば、５から１０）にランク付けられた文を選択する。次に、この文集合は文書に関係付けて格納する。これを行う一方法は、選択した文を連結することであり、次にハッシュテーブルを利用して文書識別子を格納する。

【０２３８】

次に、提示システム１３０は、文書ｄそれぞれの選択文を、検索結果７０１の他の文書の選択文と比較し、選択文が一致した場合（許容範囲内で）、その文書は複製であると推定し、それらの内の１つを検索結果から削除する。例えば、提示システム１３０は、連結した文をハッシュでき、ハッシュテーブルが既にハッシュ値に対するエントリを有する場合、これは、現在の文書、および今ハッシュされた文書が複製であるということを示す。次いで、提示システム１３０は、文書の内の１つの文書ＩＤで表を更新できる。好ましくは、提示システム１３０は、より高いページランク、または文書の重要性を表すクエリーとは無関係な他の尺度をもつ文書を保存する。更に、提示システム１３０は、インデックス１５０を修正して、複製文書を削除できるので、どんなクエリーに対しても、その文書は将来の検索結果に出現しない。

【０２３９】

同じ複製削除処理を、インデックス化システム１１０により直接適用してもよい。文書をクロールする場合、上記の文書説明処理を実行して、選択文、次いで、これら文書のハッシュを取得する。ハッシュテーブルが満たされる場合、先の場合と同様に、新規にクロールした文書を以前の文書の複製であると見なす。同様に、インデックス化システム１１０は、より高いページランク、またはクエリーとは無関係な他の尺度をもつ文書を保存できる。

【０２４０】

本発明を、１つの可能性のある実施の形態に関して特に詳細に説明した。当該技術に精通する者には言うまでもないが、本発明は他の実施の形態でも実施できる。第１に、構成の特定名称、用語の大文字化、属性、データ構造、または任意の他のプログラミング、もし

くは構造の態様は、必須でも重要でもなく、本発明またはその特徴を実装するメカニズムは、別の名称、フォーマット、またはプロトコルを有してもよい。更に本システムは、上記のように、ハードウエアおよびソフトウエアの組合せを介して、または全体をハードウエア素子を介して実装してもよい。同様に、本明細書で説明した様々なシステム構成間の特定の機能分割は、単なる例示であり、義務ではなく、単一システム構成体により実行される機能は、代替として多数の構成体で実行してもよく、多数の構成体により実行される機能は代替として単一の構成体で実行してもよい。

【０２４１】

上記説明の幾つかの部分は、情報操作のアルゴリズム、およびシンボル的表現により本発明の特徴を示す。これらのアルゴリズム記述および表現は、データ処理技術に精通する者が用いて、彼らの業務の実体を他の当該技術に精通する者へ最も効率的に伝える手段である。これらの操作は、機能的または論理的に説明したが、コンピュータプログラミングにより実装されるものであることは言うまでもない。更に、一般性を失うことなく、これらの操作の編成をモジュールと称し、または機能名称で呼ぶことが、時として便利であることも証明した。

【０２４２】

上記説明から明らかなように、特に説明しない限り、言うまでもなく、本説明全体を通じて、「処理」、「演算」、または「計算」、または「決定」、または「ディスプレイ」等の用語を用いる説明は、コンピュータシステムのメモリやレジスタ、もしくは他のそのような情報の格納装置、伝送装置、またはディスプレイ装置内の物理（電子）量として表わされるデータを操作し、変換するコンピュータシステムもしくは類似の電子計算装置の動作および処理を指す。

【０２４３】

本発明の特定の態様には、本明細書で説明したアルゴリズム形式の処理ステップおよび命令が含まれる。注意すべきは、本発明の処理ステップおよび命令がソフトウエア、ファームウエア、またはハードウエアで実施でき、ソフトウエアで実施する場合は、ダウンロードして取込み、リアルタイムネットワーク動作システムが用いる別のプラットフォームにより操作することができるということである。

【０２４４】

本発明は本明細書の操作を実行するための装置にも関連している。本装置は要求される目的のために特別に構成するか、またはコンピュータがアクセスできるコンピュータ可読媒体に格納されたコンピュータプログラムにより選択的に起動されるか再構成される汎用コンピュータを備えてもよい。このようなコンピュータプログラムは、フレキシブル磁気ディスク、光ディスク、ＣＤ－ＲＯＭ、光磁気ディスク、リードオンリーメモリ（ＲＯＭ）、ランダムアクセスメモリ（ＲＡＭ）、ＥＰＲＯＭ、ＥＥＰＲＯＭ、磁気または光カード、アプリケーション特定集積回路（ＡＳＩＣ）、または電子的命令を格納するために適し、任意の種類のコンピュータシステムバスにそれぞれ接続される媒体を含む任意の種類のディスク等の、無論これだけには留まらないが、コンピュータ可読格納媒体に格納してもよい。更に、本明細書で言うコンピュータは、単一プロセッサを含むか、または計算能力を強化するマルチプロセッサ設計を利用するアーキテクチャでもよい。

【０２４５】

本明細書で提示したアルゴリズムおよび動作は、任意の特定コンピュータまたは他の装置に本質的には関連しない。各種の汎用システムを本明細書の教示に従ったプログラムとともに用いることもできるし、またはより特殊化した装置を構築して本方法の要求ステップを実行すると便利になることもある。これら各種システムに要求される構成が、等価の改変を伴うことは、当該技術に精通する者には言うまでもない。更に、本発明の説明は、何らかの特定プログラム言語に基づいていない。各種のプログラム言語を用いて本明細書で説明した本発明の教示を実装し、任意の特定言語に基づいて本発明の可能性および最上のモードの開示を提供できるのは言うまでもない。

【０２４６】

本発明は、無数のトポロジーを越えて拡がる広範囲なコンピュータネットワークシステムに良く適合している。この分野では、大規模ネットワークの構成および管理は、インターネット等のネットワークを越えて、異種のコンピュータおよび格納装置に接続して通信する格納装置およびコンピュータを備える。

【０２４７】

最後に留意すべきは、本明細書で用いる言語は基本的に可読性および説明目的のために選択し、発明性のある主題を描写、または制限するために選択したものではないということである。従って、本発明の開示は、説明を意図したものであって、以下の請求範囲で述べる本発明の範囲の制限を意図したものではない。

【図面の簡単な説明】

【０２４８】

【図１】図１は、本発明の一実施の形態のソフトウエアアーキテクチャのブロック図である。

【図２】図２は文書内のフレーズを識別する方法を示す。

【図３】図３はフレーズ窓および２次窓をもつ文書を示す。

【図４】図４は、関連フレーズを識別する方法を示す。

【図５】図５は、関連フレーズに対して文書をインデックス化する方法を示す。

【図６】図６は、フレーズに基づいて文書を検索する方法を示す。

【図７】図７は、検索結果を提示するための提示システムの動作を示す。

【図８ａ】図８ａは、参照している文書、および参照される文書の間の関係を示す。

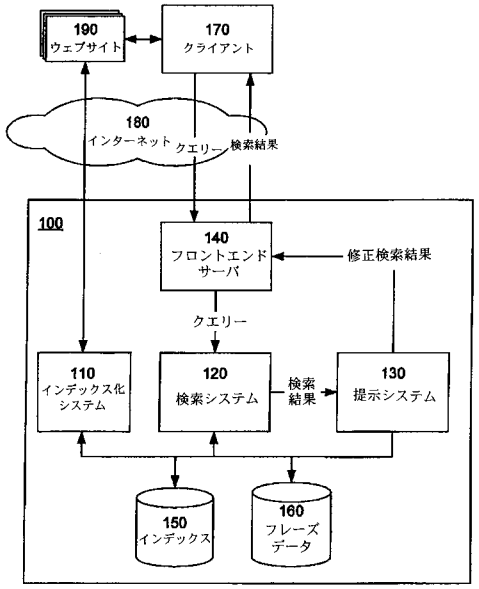
【図８ｂ】図８ｂは、参照している文書、および参照される文書の間の関係を示す。

【符号の説明】

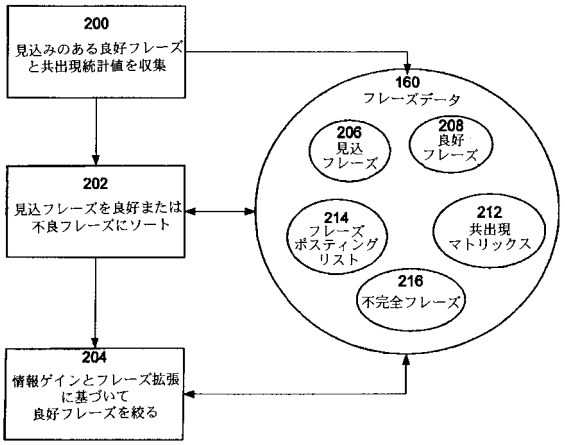
【０２４９】

１００　検索システム
１７０　クライアント
１８０　インターネット
１９０　ウェブサイト

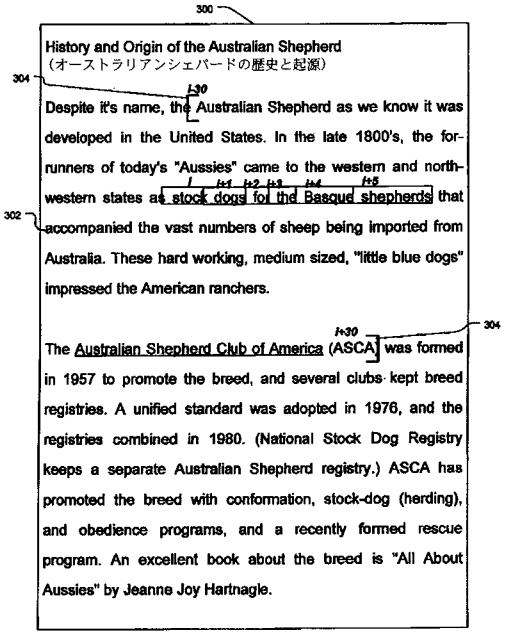
【図１】


190
ウェブサイト
170
クライアント
180
インターネット
クエリー
検索結果
100
140
フロントエンド
サーバ
修正検索結果
クエリー
110
インデックス化
システム
120
検索システム
検索
結果
130
提示システム
150
インデックス
160
フレーズ
データ

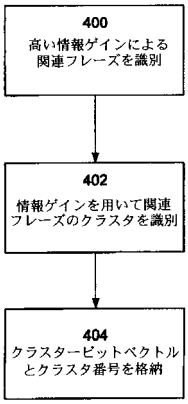
【図２】


200
見込みのある良好フレーズ
と共出現統計値を収集
160
フレーズデータ
206
見込
フレーズ
208
良好
フレーズ
202
見込フレーズを良好または
不良フレーズにソート
214
フレーズ
ポスティング
リスト
212
共出現
マトリックス
216
不完全フレーズ
204
情報ゲインとフレーズ拡張
に基づいて
良好フレーズを絞る

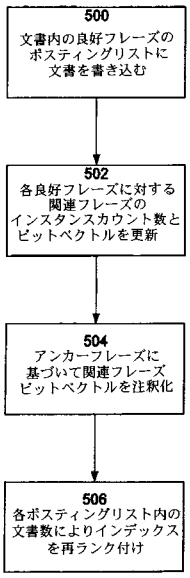
【図３】


300
History and Origin of the Australian Shepherd
(オーストラリアンシェパードの歴史と起源)
304
i-30
Despite it's name, the Australian Shepherd as we know it was developed in the United States. In the late 1800's, the forrunners of today's "Aussies" came to the western and northwestern states as stock dogs for the Basque shepherds that
i
i+1
i+2
i+3
i+4
i+5
302
accompanied the vast numbers of sheep being imported from Australia. These hard working, medium sized, "little blue dogs" impressed the American ranchers.
i+30
304
The Australian Shepherd Club of America (ASCA) was formed in 1957 to promote the breed, and several clubs kept breed registries. A unified standard was adopted in 1976, and the registries combined in 1980. (National Stock Dog Registry keeps a separate Australian Shepherd registry.) ASCA has promoted the breed with conformation, stock-dog (herding), and obedience programs, and a recently formed rescue program. An excellent book about the breed is "All About Aussies" by Jeanne Joy Hartnagle.

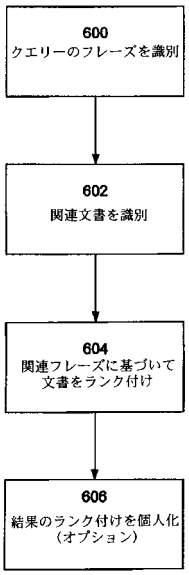
【図４】


400
高い情報ゲインによる
関連フレーズを識別
402
情報ゲインを用いて関連
フレーズのクラスタを識別
404
クラスタービットベクトル
とクラスタ番号を格納

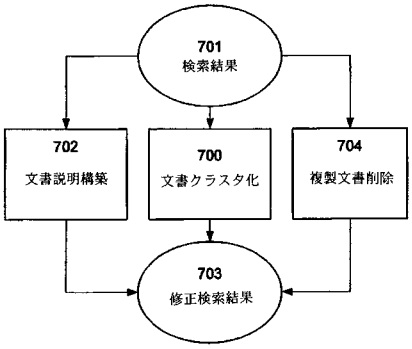
【図５】



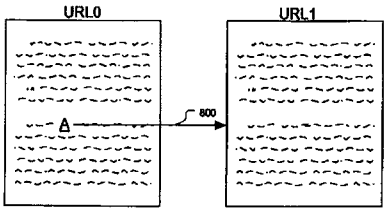

【図６】



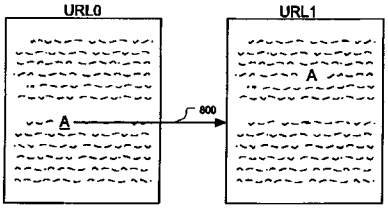

【図７】




【図８ａ】




【図８ｂ】



フロントページの続き

(56)参考文献 特開２００２－１３２７８９（ＪＰ，Ａ）
特開２００４－０４６４３８（ＪＰ，Ａ）
片岡 充照，テキスト情報を対象としたキーワード抽出と関連情報収集システム，日本ファジィ学会誌，日本，日本ファジィ学会，１９９７年１０月１５日，第９巻 第５号，７１０－７１６ページ
Helena AHONEN-MYKA, Finding co-occurring text phrases by combining sequence and frequent set discovery, PROCEEDINGS OF 16TH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE IJCAI-99 WORKSHOP ON TEXT MINING: FOUNDATIONS,TECHNIQUES AND APPLICATIONS, １９９９年 ８月１９日，１－９ページ

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０