

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号
特許第4976666号
(P4976666)

(45) 発行日 平成24年7月18日 (2012. 7. 18)

(24) 登録日 平成24年4月20日 (2012. 4. 20)

(51) Int. Cl.

F I

GO 6 F 17/30 (2006. 01) 2 1 0 A

GO 6 F 17/30 4 1 4 B

請求項の数 26 外国語出願 (全 38 頁)

(21) 出願番号	特願2005-216527 (P2005-216527)	(73) 特許権者	505281067
(22) 出願日	平成17年7月26日 (2005. 7. 26)		グーグル インコーポレイテッド
(65) 公開番号	特開2006-48683 (P2006-48683A)		GOOGLE INC.
(43) 公開日	平成18年2月16日 (2006. 2. 16)		アメリカ合衆国 カリフォルニア州 94
審査請求日	平成20年7月23日 (2008. 7. 23)		043 マウンテン ビュー アンフィシ
(31) 優先権主張番号	10/900, 021		アター パークウェイ 1600
(32) 優先日	平成16年7月26日 (2004. 7. 26)	(74) 代理人	100077539
(33) 優先権主張国	米国 (US)		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 951
			20, サンノゼ, ソーンツリー コート
			1127
		審査官	長 由紀子
			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識別方法

(57) 【特許請求の範囲】

【請求項 1】

文書集合内のフレーズを識別するための、コンピュータにより実行される、方法であって、

前記文書集合内の文書から見込フレーズを収集するステップと、

前記文書集合内における前記各見込フレーズの発生頻度に従い、各見込フレーズを良好フレーズまたは不良フレーズとして分類するステップと、

前記文書集合内の或る1対の良好フレーズ (g_j , g_k) について、一方のフレーズ (g_j) に関する他方のフレーズ (g_k) の情報ゲインを、該文書集合内における該両フレーズ (g_j , g_k) の共出現率の期待値及び該両フレーズ (g_j , g_k) の実際の共出現率の関数として、決定するステップと、

前記1対の良好フレーズ (g_j , g_k) における前記一方のフレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが所定の第1の閾値を超えると、前記文書集合内において該一方のフレーズ (g_j) が該他方のフレーズ (g_k) を予測するものとして判別し、該判別したフレーズ (g_j) を、妥当なフレーズとして、選択的に保持するステップと

、

前記選択的に保持された前記妥当なフレーズ (g_j) について、当該フレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが前記第1の閾値よりも高い所定の第2の閾値を超えると、該他方のフレーズ (g_k) を当該フレーズ (g_j) の関連フレーズとして識別するステップと、

10

20

前記妥当なフレーズ及び前記関連フレーズをコンピュータ読み取り可能な記憶媒体に記憶するステップと、
を備える方法。

【請求項 2】

前記見込フレーズを収集する前記ステップは、
マルチワードフレーズ窓により文書のワードをトラバースするステップと、前記窓内の最初のワードで始まる該窓内の全てのワードシーケンスの組み合わせをそれぞれ候補フレーズとして選択するステップと、
を含む請求項 1 の方法。

【請求項 3】

前記フレーズ窓は少なくとも 4 ワードを含む請求項 2 の方法。

【請求項 4】

前記見込フレーズを収集する前記ステップは、
各見込フレーズについて、当該見込フレーズを含む文書数の頻度カウントを保持するステップと、
各見込フレーズについて、当該見込フレーズそれ自体の数の頻度カウントを保持するステップと、
各見込フレーズについて、当該見込フレーズを他と区別するための所定の区別表記を伴う当該見込フレーズの数の頻度カウントを保持するステップと、
を含む請求項 1 の方法。

【請求項 5】

前記区別表記を伴う見込フレーズとは、文法的またはフォーマットマーカにより前記文書内における隣接内容から区別される見込フレーズである請求項 4 の方法。

【請求項 6】

前記各見込フレーズを良好フレーズまたは不良フレーズとして分類する前記ステップは、
、
前記見込フレーズが所定の最小数以上の文書に出現し、かつ、該見込フレーズが前記文書集合において所定の最小数以上出現する場合に、当該見込フレーズを良好フレーズとして分類するステップ、
を含む請求項 1 の方法。

【請求項 7】

前記各見込フレーズを良好フレーズまたは不良フレーズとして分類する前記ステップは、
、
前記見込フレーズが、当該見込フレーズを他と区別するための所定の区別表記を伴うフレーズとして、前記文書集合において所定の最小数以上出現する場合に、当該見込フレーズを良好フレーズとして分類するステップ、
を含む請求項 1 の方法。

【請求項 8】

前記一方のフレーズ (g_j) に関する他方のフレーズ (g_k) の前記情報ゲインは、
$$I(j, k) = A(j, k) / E(j, k)$$
であり、
ここで、
 $A(j, k)$ は、前記一方のフレーズ (g_j) および他方のフレーズ (g_k) の実際の共出現率であり、
 $E(j, k)$ は、前記一方のフレーズ (g_j) および他方のフレーズ (g_k) の共出現率の期待値である、
請求項 1 乃至 7 のいずれかの方法。

【請求項 9】

前記第 1 の閾値は 1 である請求項 8 の方法。

【請求項 10】

良好フレーズ g_j および g_k は、 g_j および g_k が互いのワードの所定数内である場合、文

10

20

30

40

50

書内に共出現する請求項 8 の方法。

【請求項 1 1】

前記判別したフレーズ (g_j) を、妥当なフレーズとして、選択的に保持する前記ステップは、

前記 1 対の良好フレーズ (g_j , g_k) における前記一方のフレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが前記所定の第 1 の閾値未満であれば、該一方のフレーズ (g_j) を選択的に保持しないことを特徴とする、

請求項 1 の方法。

【請求項 1 2】

前記妥当なフレーズの中から不完全フレーズを判定するステップと、
判定した不完全フレーズを前記妥当なフレーズから削除するステップと、
を更に備え、ここで、前記不完全フレーズとは、該不完全フレーズによって予測される前記他方のフレーズがフレーズ拡張のみからなるフレーズであり、前記フレーズ拡張とは、該不完全フレーズで始まりかつ該不完全フレーズより長いフレーズである、請求項 1 乃至 1 1 のいずれかの方法。

10

【請求項 1 3】

各不完全フレーズ毎に、該不完全フレーズによって予測される少なくとも 1 つの前記フレーズ拡張を記憶保持するステップと、
検索クエリーを受信するステップと、
受信した検索クエリー内に前記不完全フレーズが含まれているかを判定するステップと

20

、
前記検索クエリー内に含まれている前記不完全フレーズに応じて、該不完全フレーズによって予測される前記少なくとも 1 つのフレーズ拡張を該検索クエリーに追加するステップと、

を含む請求項 1 2 の方法。

【請求項 1 4】

前記所定の第 2 の閾値が約 1 0 0 である請求項 1 乃至 1 3 のいずれかの方法。

【請求項 1 5】

各フレーズ (g_j) 毎に、該フレーズおよび該フレーズに対応する少なくとも 1 つの前記関連フレーズ (g_k) を含むクラスタを識別するステップを更に備える含む請求項 1 乃至 1 4 のいずれかの方法。

30

【請求項 1 6】

前記各フレーズ (g_j) 毎に、該フレーズに対応する複数の前記関連フレーズを含む集合 R を識別するステップと、

該集合 R 内の前記関連フレーズの各対毎に、該関連フレーズ対の情報ゲインを判定するステップと、

前記各フレーズ (g_j) の前記関連フレーズのクラスタとして、該フレーズ (g_j) と、前記集合 R 内の他の各フレーズに関するゼロではない情報ゲインを有する該集合 R 内の前記各関連フレーズとを識別するステップと、

を更に備える請求項 1 乃至 1 4 のいずれかの方法。

40

【請求項 1 7】

前記各クラスタに、該クラスタに含まれる前記関連フレーズの関数としての一意のクラスタ番号を割り当てるステップを更に備える請求項 1 6 の方法。

【請求項 1 8】

前記クラスタの前記関連フレーズの最大情報ゲインを有する前記関連フレーズを含む名称を前記クラスタに割り当てるステップを更に備える請求項 1 6 の方法。

【請求項 1 9】

或るフレーズ (g_j) について、各ビット位置が妥当なフレーズに対応するとともに、該各ビット位置に対応するビットが、それに対応する前記妥当なフレーズが当該或るフレーズ (g_j) の関連フレーズであるかどうかを示している、ビットベクトルを格納するス

50

トップを更に備える請求項 1 乃至 18 のいずれかの方法。

【請求項 20】

文書集合内のフレーズを識別するためのコンピュータプログラムであって、コンピュータに、

前記文書集合内の文書から見込フレーズを収集する手順と、

前記文書集合内における前記各見込フレーズの発生頻度に従い、各見込フレーズを良好フレーズまたは不良フレーズとして分類する手順と、

前記文書集合内の或る 1 対の良好フレーズ (g_j , g_k) について、一方のフレーズ (g_j) に関する他方のフレーズ (g_k) の情報ゲインを、該文書集合内における該両フレーズ (g_j , g_k) の共出現率の期待値及び該両フレーズ (g_j , g_k) の実際の共出現率の関数として、決定する手順と、

10

前記 1 対の良好フレーズ (g_j , g_k) における前記一方のフレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが所定の第 1 の閾値を超えると、前記文書集合内において該一方のフレーズ (g_j) が該他方のフレーズ (g_k) を予測するものとして判別し、該判別したフレーズ (g_j) を、妥当なフレーズとして、選択的に保持する手順と、

前記選択的に保持された前記妥当なフレーズ (g_j) について、当該フレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが前記第 1 の閾値よりも高い所定の第 2 の閾値を超えると、該他方のフレーズ (g_k) を当該フレーズ (g_j) の関連フレーズとして識別する手順と、

前記妥当なフレーズ及び前記関連フレーズをコンピュータ読み取り可能な記憶媒体に記憶する手順と、

20

を実行させるためのコンピュータプログラム。

【請求項 21】

前記一方のフレーズ (g_j) に関する他方のフレーズ (g_k) の前記情報ゲインは、

$$I(j, k) = A(j, k) / E(j, k) \text{ であり、}$$

ここで、

$A(j, k)$ は、前記一方のフレーズ (g_j) および他方のフレーズ (g_k) の実際の共出現率であり、

$E(j, k)$ は、前記一方のフレーズ (g_j) および他方のフレーズ (g_k) の共出現率の期待値である、

30

請求項 20 のコンピュータプログラム。

【請求項 22】

文書集合内のフレーズを識別するためのコンピュータシステムであって、

前記文書集合内の文書から見込フレーズを収集する手段と、

前記文書集合内における前記各見込フレーズの発生頻度に従い、各見込フレーズを良好フレーズまたは不良フレーズとして分類する手段と、

前記文書集合内の或る 1 対の良好フレーズ (g_j , g_k) について、一方のフレーズ (g_j) に関する他方のフレーズ (g_k) の情報ゲインを、該文書集合内における該両フレーズ (g_j , g_k) の共出現率の期待値及び該両フレーズ (g_j , g_k) の実際の共出現率の関数として、決定する手段と、

40

前記 1 対の良好フレーズ (g_j , g_k) における前記一方のフレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが所定の第 1 の閾値を超えると、前記文書集合内において該一方のフレーズ (g_j) が該他方のフレーズ (g_k) を予測するものとして判別し、該判別したフレーズ (g_j) を、妥当なフレーズとして、選択的に保持する手段と、

前記選択的に保持された前記妥当なフレーズ (g_j) について、当該フレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが前記第 1 の閾値よりも高い所定の第 2 の閾値を超えると、該他方のフレーズ (g_k) を当該フレーズ (g_j) の関連フレーズとして識別する手段と、

前記妥当なフレーズ及び前記関連フレーズをコンピュータ読み取り可能な記憶媒体に記憶する手段と、

50

を備えるコンピュータシステム。

【請求項 2 3】

文書集合内の文書をインデックス化するための、コンピュータにより実行される、方法であって、

前記文書集合内の文書から見込フレーズを収集するステップと、

前記文書集合内における前記各見込フレーズの発生頻度に従い、各見込フレーズを良好フレーズまたは不良フレーズとして分類するステップと、

前記文書集合内の或る 1 対の良好フレーズ (g_j , g_k) について、一方のフレーズ (g_j) に関する他方のフレーズ (g_k) の情報ゲインを、該文書集合内における該両フレーズ (g_j , g_k) の共出現率の期待値及び該両フレーズ (g_j , g_k) の実際の共出現率の関数として、決定するステップと、

10

前記 1 対の良好フレーズ (g_j , g_k) における前記一方のフレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが所定の第 1 の閾値を超えると、前記文書集合内において該一方のフレーズ (g_j) が該他方のフレーズ (g_k) を予測するものとして判別し、該判別したフレーズ (g_j) を、妥当なフレーズとして、記憶媒体に記憶するステップと、

前記選択的に保持された前記妥当なフレーズ (g_j) について、当該フレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが前記第 1 の閾値よりも高い所定の第 2 の閾値を超えると、該他方のフレーズ (g_k) を当該フレーズ (g_j) の関連フレーズとして識別するステップと、

20

前記妥当なフレーズ及び前記関連フレーズに関係付けて前記文書集合内の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 4】

前記一方のフレーズ (g_j) に関する他方のフレーズ (g_k) の前記情報ゲインは、

$$I(j, k) = A(j, k) / E(j, k) \text{ であり、}$$

ここで、

$A(j, k)$ は、前記一方のフレーズ (g_j) および他方のフレーズ (g_k) の実際の共出現率であり、

$E(j, k)$ は、前記一方のフレーズ (g_j) および他方のフレーズ (g_k) の共出現率の期待値である、

30

請求項 2 3 の方法。

【請求項 2 5】

前記見込フレーズを収集する前記ステップは、

各見込フレーズについて、当該見込フレーズを含む文書数の頻度カウントを保持するステップと、

各見込フレーズについて、当該見込フレーズの数 の頻度カウントを保持するステップと、

各見込フレーズについて、当該見込フレーズを他と区別するための文法的またはフォーマットマーカによる区別表記を伴う当該見込フレーズの数 の頻度カウントを保持するステップと、

40

を含む請求項 2 3 又は 2 4 の方法。

【請求項 2 6】

文書集合内の文書をインデックス化するためのコンピュータプログラムであって、コンピュータに、

前記文書集合内の文書から見込フレーズを収集する手順と、

前記文書集合内における前記各見込フレーズの発生頻度に従い、各見込フレーズを良好フレーズまたは不良フレーズとして分類する手順と、

前記文書集合内の或る 1 対の良好フレーズ (g_j , g_k) について、一方のフレーズ (g_j) に関する他方のフレーズ (g_k) の情報ゲインを、該文書集合内における該両フレーズ (g_j , g_k) の共出現率の期待値及び該両フレーズ (g_j , g_k) の実際の共出現率の関数として、決定するステップと、

50

g_j, g_k) の共出現率の期待値及び該両フレーズ (g_j, g_k) の実際の共出現率の関数として、決定する手順と、

前記1対の良好フレーズ (g_j, g_k) における前記一方のフレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが所定の第1の閾値を超えると、前記文書集合内において該一方のフレーズ (g_j) が該他方のフレーズ (g_k) を予測するものとして判別し、該判別したフレーズ (g_j) を、妥当なフレーズとして、記憶媒体に記憶する手順と、

前記選択的に保持された前記妥当なフレーズ (g_j) について、当該フレーズ (g_j) に関する前記他方のフレーズ (g_k) の前記情報ゲインが前記第1の閾値よりも高い所定の第2の閾値を超えると、該他方のフレーズ (g_k) を当該フレーズ (g_j) の関連フレーズとして識別する手順と、

10

前記妥当なフレーズ及び前記関連フレーズに係り付けて前記文書集合内の文書をインデックス化する手順と、

を実行させるためのコンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

(関連出願の相互参照)

出願日2004年7月26日、出願番号第10/x x x、x x x号、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」、

出願日2004年7月26日、出願番号第10/x x x、x x x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」、

20

出願日2004年7月26日、出願番号第10/x x x、x x x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/x x x、x x x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/x x x、x x x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/x x x、x x x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。

30

【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、幾つかのクエリー語を含むクエリーに応じて検索され、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。

40

【0004】

クエリー語の直接的な「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

50

、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、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 3×10^{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。

10

【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。

20

【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。

30

【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。

40

【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フレーズに関連付ける。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

50

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。

10

【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と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。

20

【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。

30

【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。

40

50

【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。

10

【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。

20

【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。

30

【0019】

本発明は更に、システムおよびソフトウェアアーキテクチャ、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。

40

【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】

50

【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】

10

【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ェ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。

20

30

【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。

40

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

50

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ルを続ける。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。

10

【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。

20

【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；

30

【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ーにより、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ーにより示される。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】

40

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、 $m \times m$ 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

50

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、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 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

10

20

30

40

50

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。

10

【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ない、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。

20

【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 $m \cdot m$ 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。

30

【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：

40

【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 $E(j, k)$ より

50

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】

$I(j, k) > \text{情報ゲイン閾値}$ 。

【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 United 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

10

20

30

40

50

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。

10

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。

20

【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：

30

【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。

40

【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】

50

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 R_j であり、ここで、 $R = \{g_k, g_l, \dots, g_m\}$ 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ー）」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	クラスター 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

10

【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ー 14 にある。

20

【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：

30

【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。

40

【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

50

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、 g_1 は第1良好フレーズ、 g_2 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、 g_1 「President」、および g_4 「President of A TT」)、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

10

20

30

40

50

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。

10

【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。

20

【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 g_k 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。

30

【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；

40

【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) の内のどれかが文書 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 110 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

50

り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对のスコア、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 のアウトリンクを經由して) 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

10

20

30

40

50

していると言することができる。図 8 b は、この場合を示し、フレーズ A は、URL 0（アンカーテキストとして）、および URL 1 の本文の両方に出現する。この場合、URL 1 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 URL 0 から URL 1 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 URL 1 本文に出現しない場合（図 8 a 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 URL 1 にあるかのように）フレーズ A の URL 1 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 URL 1 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、URL 0 から URL 1 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 URL 0 および URL 1 に出現すると仮定する：

【表 2】

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	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）。URL 0 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、URL 1 の行は、リンクのインリンクスコアである。ここで、URL 0 は URL 1 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 Shepherd」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 URL 0 に出現する。従って、URL 0 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、URL 1 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、URL 0 および URL 1 の両方に出現するので、URL 0 に対するアウトリンクスコア、および URL 1 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 URL 1 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 URL 1 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 URL 1 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、URL 1 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（文書の種類）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。

10

【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。

20

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。

30

【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。

40

【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。

50

【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
10 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）。
20

【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。
30

【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。
40

【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）」
50

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

10

【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

20

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。

30

【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

40

【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】

50

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 Q_r がある。上記のように、関連クエリーフレーズ Q_r 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 Q_{r1} （すなわち、Qによる最大情報ゲインをもつ関連フレーズ Q_r ）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 Q_{r2} に対する $N-1$ ポイント、次いで、 Q_{r3} 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 Q_{rN} が1ポイントを割り当てられる。

10

【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 Q_r 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 Q_r 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。

20

【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 Q_r 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。

30

【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 Q_p 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。

40

【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

50

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】

$R_i \cdot Q$ ・関連フレーズビットベクトル $\times D \cdot Q$ ・関連フレーズビットベクトル。

10

【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。

20

【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。

30

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）。入力の実行クエリーに対して、クエリーは、クエリーフレーズ Q 、関連フレーズ Q_r 、およびクエリーフレーズ Q_p のフレーズ拡張 Q_e 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）

40

50

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に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。

10

20

30

40

50

【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。

10

【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。

20

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。

30

【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、入力と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

40

【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

50

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 Q_r 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 Q_r に対して、Q_r 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 Q_r 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 Q_r 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 Q_r 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 Q_r を伴いつつ各上位クラスタが識別される。

10

【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。

20

【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」;

30

【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。

40

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。

50

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Q_r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Q_r、およびフレーズ拡張Q_p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。

10

【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Q_r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Q_p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。

20

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Q_r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U_r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Q_r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Q_uと呼ぶ。

30

【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Q_u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Q_r、およびフレーズ拡張Q_p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U_rの順序となる。

40

【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U_r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

50

【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。

10

【0237】

提示システム130は、検索結果集合701にある各文書进行处理する。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。

20

【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。

30

【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、ハードウェアおよびソフトウェアの組合せを介して、または全体をハードウェ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。

40

【0241】

上記説明の幾つかの部分、情報操作のアルゴリズム、およびシンボリック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

50

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。

10

【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ェア、ファームウェア、またはハードウェアで実施でき、ソフトウェ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。

20

30

【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広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。

40

【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】

50

【0248】

【図1】図1は、本発明の一実施の形態のソフトウェ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。

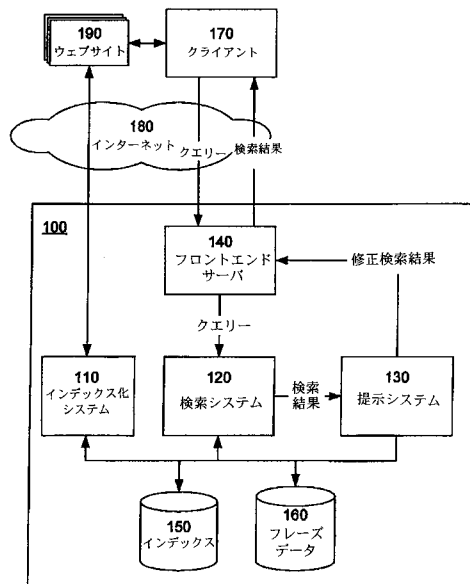
【符号の説明】

【0249】

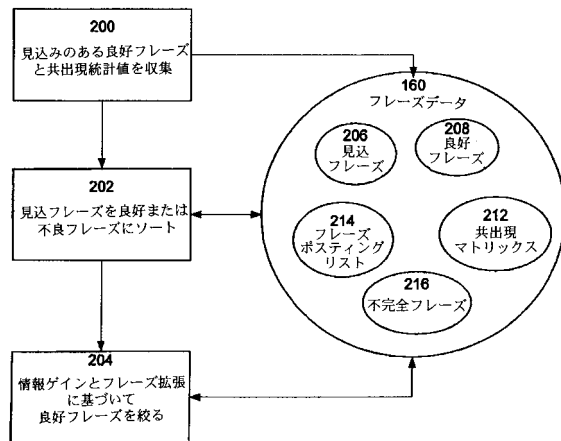
- 100 検索システム
- 170 クライアント
- 180 インターネット
- 190 ウェブサイト

10

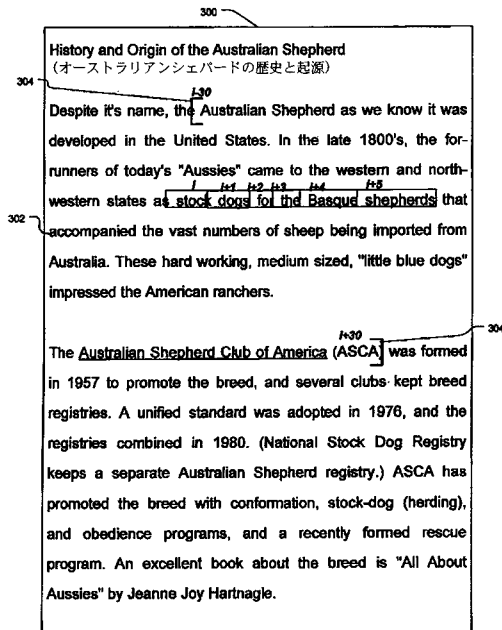
【図1】



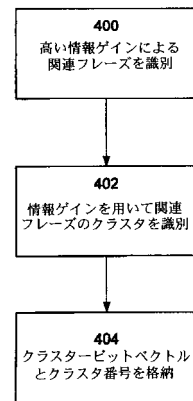
【図2】



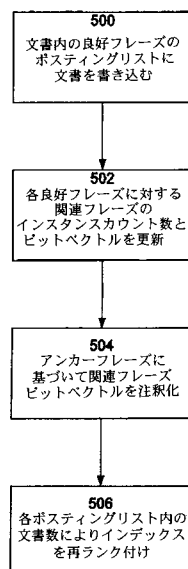
【図 3】



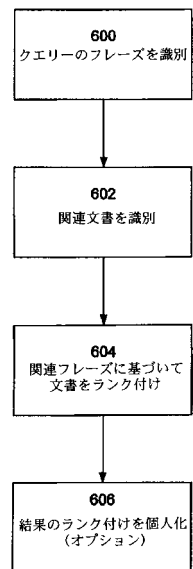
【図 4】



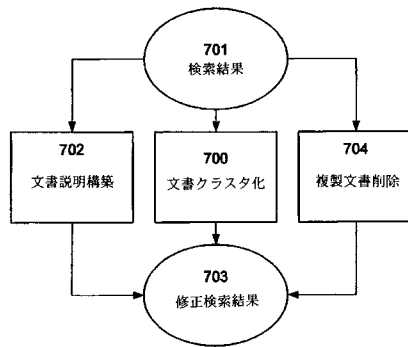
【図 5】



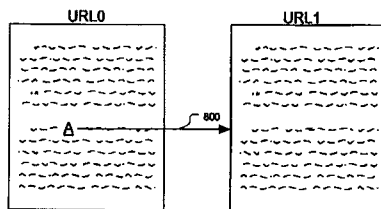
【図 6】



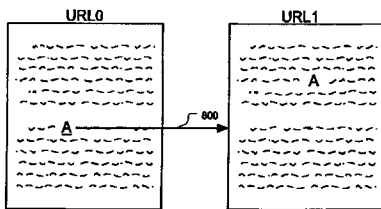
【図 7】



【図 8 a】



【図 8 b】



フロントページの続き

(56)参考文献 特開2004-005600 (JP, A)

特開平08-305730 (JP, A)

特開2000-331032 (JP, A)

AHONEN-MYKA H et al, 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, [平成22年10月28日検索]、インターネット<<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.4013>> Download as a PDF

神林 隆 他, インターネット情報探索に適したキーワード抽出, 情報処理学会研究報告 Vol. 97 No. 29 IPSJ SIG Notes, 日本, 社団法人情報処理, 1997年 3月22日, 第97巻第29号, 79-84

安部 潤一郎 他, テキストデータからの高速データマイニング-探索的文書ブラウジングとウェブデータへの応用-, 人工知能学会誌 第15巻 第4号, 日本, 社団法人人工知能学会, 2000年 7月 1日, 第15巻第4号, 618-628

(58)調査した分野(Int.Cl., DB名)

G06F 17/30