(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2009-509259
(P2009-509259A)
(43)公表日　平成21年3月5日(2009.3.5)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	419B	5B075
		G06F	17/30	350C	
		G06F	17/30	170A	

審査請求 未請求　予備審査請求 未請求　（全 19 頁）

(21)出願番号	特願2008-531744 (P2008-531744)
(86)(22)出願日	平成18年9月7日 (2006.9.7)
(85)翻訳文提出日	平成20年5月8日 (2008.5.8)
(86)国際出願番号	PCT/FR2006/050856
(87)国際公開番号	WO2007/034096
(87)国際公開日	平成19年3月29日 (2007.3.29)
(31)優先権主張番号	0552826
(32)優先日	平成17年9月20日 (2005.9.20)
(33)優先権主張国	フランス(FR)

(71)出願人　591034154
フランス　テレコム
フランス国　パリ　７５０１５　プラス　ダルレ　６
(74)代理人　100064908
弁理士　志賀　正武
(74)代理人　100089037
弁理士　渡邊　隆
(74)代理人　100108453
弁理士　村山　靖彦
(74)代理人　100110364
弁理士　実広　信哉
(72)発明者　ジェローム・ガルティエール
フランス・０６０００・ニース・リュ・ヴェルディ・１５

最終頁に続く

(54)【発明の名称】一組の電子文書を並べ替える方法

(57)【要約】
一組の電子文書を並べ替える方法であって、前記一組の電子文書のうち各一対の文書間に存在するハイパーテキストリンク又は共引用を数える段階と、前記数える段階の完了時に得られるリンク数に依存する文書ｕ及びｖ間の相関割合ω（ｕ、ｖ）を、前記一組の電子文書のうち各一対の文書｛ｕ、ｖ｝に関して決定する段階と、Ｒが一組の実数でｄが正の整数である一組のＲ^{d}の表面上に位置する関連点Ｘ（ｕ）を、前記一組の電子文書のうち文書ｕ毎に決定する段階であって、前記一組の電子文書のうち少なくとも一つの文書ｕ１に関して、関連点Ｘ（ｕ１）及びＸ（ｕ２）間の距離Ｒ^{d}は、ｕ２が文書ｕ１及びｕ２間に相関がある文書の場合、文書ｕ１及びｕ２間の相関割合が高くなるにつれ、さらに低くなる、段階と、前記表面上で決定された点の関数として前記一組の電子文書のうち少なくとも一部を並べ替える段階とを具備する。

【特許請求の範囲】
【請求項１】
一組の電子文書を並べ替える方法であって：
－前記一組の電子文書のうち各一対の文書｛ｕ、ｖ｝間に存在するハイパーテキストリンク又は共引用を数える段階と、
－前記数える段階の完了時に得られるリンク数に依存する文書ｕ及びｖ間の相関割合ω（ｕ、ｖ）を、前記一組の電子文書のうち各一対の文書｛ｕ、ｖ｝に関して決定する段階（Ｓ１１０）と、
－Ｒが一組の実数でｄが正の整数である一組のＲ^{d}の表面上に位置する関連点Ｘ（ｕ）を、前記一組の電子文書のうち文書ｕ毎に決定する段階（Ｓ１２０、Ｓ１２５、Ｓ１３０、Ｓ１３５）であって、前記一組の電子文書のうち少なくとも一つの文書ｕ１に関して、関連点Ｘ（ｕ１）及びＸ（ｕ２）間の距離Ｒ^{d}は、ｕ２が文書ｕ１及びｕ２間に相関がある文書の場合、文書ｕ１及びｕ２間の相関割合が高くなるにつれ、さらに低くなる、段階と、
－前記表面上で決定された点の関数として前記一組の電子文書のうち少なくとも一部を並べ替える段階（Ｓ１４０）と
を具備することを特徴とする方法。
【請求項２】
前記文書のうち少なくとも一つは、少なくとも一つの他の文書への少なくとも一つのハイパーテキストリンクを具備し、
ハイパーテキストリンクの数及び／又は共引用リンクの数の関数として決定される２つの文書ｕ及びｖ間の相関割合は、このリンク数が多くなるにつれ、さらに高くなる相関割合を文書ｕ及びｖ間に提供し、相関の不在は、リンクの不在に相当することを特徴とする請求項１に記載の方法。
【請求項３】
２文書ｕ及びｖ間の相関割合は、文書ｕ及びｖの意味的内容の近接の程度に依存し、
相関割合は、前記程度が低くなるにつれ、さらに高くなり、
相関の不在は、所定の閾より低い程度に相当することを特徴とする請求項１又は２に記載の方法。
【請求項４】
－前記表面上に前記一組の電子文書を投射するための初期関数Ｘ$_{0}$を定める段階（Ｓ１２０）と、
－前記表面上に前記一組の電子文書を投射するための関数Ｘを決定する段階（Ｓ１２５、Ｓ１３０、Ｓ１３５）であって、前記投射関数Ｘは、少なくとも一つの反復で初期関数Ｘ$_{0}$に基づいて得られ、各反復は、Ｘ$_{i-1}$（ｕ）だけでなくＸ$_{i-1}$（ｖ）の値と前記一組の電子文書に属する任意の文書ｖに関する文書ｕ及びｖ間の相関割合ω（ｕ、ｖ）とに依存する所定の基準を最適化することを可能にする値Ｘ$_{i}$（ｕ）にＸ$_{i-1}$（ｕ）の値を、前記一組の電子文書のうち少なくとも一つの文書ｕに関して置換することによって、先の反復で得られた関数Ｘ$_{i-1}$に基づいて関数Ｘ$_{i}$を決定することにある、段階と
を具備することを特徴とする請求項１又は２に記載の方法。
【請求項５】
前記所定の基準の最適化は、
【数１】

$$\Delta(u) = \sum_{\{u,v\}\in E} \delta(u,v)\left\| X_{i-1}(u) - X_{i-1}(v)\right\|^2$$

に等しく量Δ（ｕ）の値を文書ｕに関して最大にすることにあり、
ここで、δ（ｕ、ｖ）＝１－ω（ｕ、ｖ）、０≦ω（ｕ、ｖ）≦１、文書ｕ及びｖ間の相関の不在においてω（ｕ、ｖ）＝０であり、

値X_i（ｕ）は、X_i（ｕ）＝－Ｙ（Ｕ）／‖Ｙ（Ｕ）‖であり、
ここで、
【数２】

$$Y(U) = \sum_{v \in V-\{u\}} \delta(u,v)\,\delta(u,v)\,X_{i-1}(v) \quad \text{if } Y(U) \neq 0,$$

値X_i（ｕ）は、Ｙ（Ｕ）＝０の場合にX_{i-1}（ｕ）に等しいことを特徴とする請求項３又は４に記載の方法。
【請求項６】
前記表面と前記表面上に位置する点Ｘ（ｕ）との図形表示を生成する段階をさらに具備することを特徴とする請求項１から５のうち何れか１項に記載の方法。
【請求項７】
端末上に前記図形表示を表示する過程と、
前記図形表示のうち少なくとも一部の図形表示に適切な図形選択ツールを端末のユーザに提供する過程と、
前記ユーザによって選択された前記少なくとも一部を定めるデータを得る過程と、
前記データの関数として前記一組の文書の並べ替えを実行する過程と
からなる段階をさらに具備することを特徴とする請求項６に記載の方法。
【請求項８】
プログラムであって、前記プログラムがコンピュータシステムによって実行される時、請求項１から７のうち何れか１項に記載の方法を実行するためのコンピュータシステムによって読取可能な媒体上に記録されたプログラムコード命令を具備することを特徴とするプログラム。
【請求項９】
請求項１から７のうち何れか１項に記載の方法からなる段階を実行するためのデータ処理手段を具備するデータ処理装置。
【請求項１０】
前記プログラムがコンピュータシステムによって実行される時、請求項１から７のうち何れか１項に記載の方法を実行するためのプログラムコード命令を含むプログラムを具備する、コンピュータシステムによって読取可能な記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、電気通信の分野に関し、具体的に電子文書を検索するための検索エンジンの分野に関する。
【０００２】
より具体的に、本発明は、一組の電子文書を並べ替える方法に関する。そのような一組の電子文書は、インターネットのネットワーク種類上で検索エンジンを用いてユーザが実行する検索から生じ、この場合の電子文書は、局所的にローカル記憶媒体を介して、又は遠隔からネットワークを介してアクセス可能なウェブページ（“World Wide Web”の略称）である。
【背景技術】
【０００３】
検索エンジンは、検索で生ずるページを評価又は並べ替えるためのいくつかの技術を用いる。一組のウェブページを探索するための周知技術のうち、一部の技術は、意味論に依存し、検索対象の言葉又は複数の言葉がページにおいて多く発生する場合に、ページは、さらに関連しているとして評価される。これら技術は、“スパミング”の名で知られる慣習の影響を受け易く、それは、インターネットユーザがその検索クエリで一般に採用した言葉を所定のページにおいて多数にすることを目的とし、これは、ページを関連するものとして頻繁に出現させる効果を有する。

【０００４】
他の技術は、ウェブの位相構造に基づく。これら技術は、考えられるページ間の既存リンクとそのページのプロパティとを同時に考慮し、そのプロパティは、ウェブのネットワークドメイン又はサブドメインにおけるページの会員等である。一般に、これら技術は、処理すべきページの図形形式の表示に基づく。それは、図形に与えられている位相構造を満たすページの分類に適する。これら技術は、所定のページを多数回参照することを目的とする“スパミング”の方法の変形による影響を受け易く、これは、ウェブの図形の位相特性（topological characteristics）を局所的に偽る効果を有する。
【０００５】
ウェブの位相構造を利用する一部の技術は、ページ及び他のページ間の関係に依存するランクを各種ページに割り当てることによってウェブページの分類をもたらす。
【０００６】
用語“ページランク”として知られるそのような手順の一例は、Google™検索エンジンの実行で使用され、1998 Stanford University, Computer Science Department, Technical Report, L. Page, S. Brin, R. Motwani, T. Winogradによる“The PageRank Citation Ranking: Bringing Order on the Web”という文献に記載されている。
【０００７】
ページランク手順は、ウェブ上の視認性の関数としてページを提供する。この手順において、ハイパーテキストリンクに従ってウェブ上をページ毎にランダムに閲覧することがシミュレートされる。この閲覧は、ウェブにアクセスするユーザが引起す閲覧に相当し、その時後者は、もう一方のページにアクセスするために閲覧ページに位置するハイパーテキストリンクのうち一つをランダムに活性化する。この手順は、このようなページ毎のランダムな閲覧の間にユーザが所定のページ上にいる可能性を判断するために、このシミュレートされた閲覧の確率論的解析を行う。ページのランクは、このページが他のページによって引用される回数が多いほど、さらに高くなる。
【０００８】
そのような手順は、ユーザが行った検索に必ずしも関連しない評価ランクを提供し、（最高ランクの）最も高く評価されたページは必ずしも、ユーザの期待に最も相応するページではない。
【０００９】
また、この手順は、興味のあるページにユーザをより早く導ける意味的共同体又は対象となる共同体を一組の文書において識別することができない。
【００１０】
最後に、ユーザが特に興味ある文書が示されていることをユーザが一組の文書において識別する場合、単にそのランクの関数として整理された文書のリストを用いて、興味のある文書に近いか又は後者にリンクされているかの何れかである他の文書が一組の文書に提供されていることを直ちに判断できない。
【発明の開示】
【発明が解決しようとする課題】
【００１１】
故に、本発明の目的は特に、電子文書、例えばウェブページを並べ替えるための技術を提供することによって、従来技術の上記欠点を解消することにあり、特にスパミングの問題を検出することができ、その実行において早い一方で一組のかなり多い文書に適用可能であり、単にランクを割り当てることによってではなく、互いに近い文書の共同体又は文書のサブセットを構築することによって文書を並べ替えることができ、この近接性の概念は、文書の意味的内容、これら文書間のハイパーテキストリンク、又はいくつかの他の方法全体の関数として定められる。
【課題を解決するための手段】
【００１２】
この目的のため、本発明の課題は、第１の実施形態によると、一組の電子文書を並べ替

える方法であって：
－前記一組の電子文書のうち各一対の文書｛ｕ、ｖ｝間に存在するハイパーテキストリンク又は共引用を数える段階と、
－前記数える段階の完了時に得られるリンク数に依存する文書ｕ及びｖ間の相関割合ω（ｕ、ｖ）を、前記一組の電子文書のうち各一対の文書｛ｕ、ｖ｝に関して決定する段階と、
－Ｒが一組の実数でｄが正の整数である一組のＲ^{d}の表面上に位置する関連点Ｘ（ｕ）を、前記一組の電子文書のうち文書ｕ毎に決定する段階であって、前記一組の電子文書のうち少なくとも一つの文書ｕ１に関して、関連点Ｘ（ｕ１）及びＸ（ｕ２）間の距離Ｒ^{d}は、ｕ２が文書ｕ１及びｕ２間に相関がある文書の場合、文書ｕ１及びｕ２間の相関割合が高くなるにつれ、さらに低くなる、段階と、
－前記表面上で決定された点の関数として前記一組の電子文書のうち少なくとも一部を並べ替える段階とを具備する。

【００１３】
表面を用いて点の位置を決定する事実は、得られた点に関して単に定めることを可能にし、故に関連する文書を定めることを可能にし、互いに関連するこれらの点の位置を定めることを可能にするという点で独創的である。特に、この表示モードにおいて、もう一方に関して好ましい点が無い。結果として、２点間の関連位置、故にこれら２点間の距離を用いて、これら２点に関連付けられた２つの文書間の相関割合を表示することができる。故に、得られた表示は、対象となる文書間の相関又はリンクを反映する。

【００１４】
表面、例えば３次元空間の表面上で一組の文書の表示を採用する事実は、任意種類の並べ替え動作をさらに予想することができ、その動作は、選択、評価、フィルタリング、分類、及び非常に単純な動作であり、なぜなら各文書は、Ｎ次元の空間におけるｎ組の座標によって（例えば、３次元空間における３つの座標によって）のみ表示されるからである。

【００１５】
多数の方法の用途があり、それは一群の文書の構築、文書の評価、又は文書の選択である。これら動作は、文書の投射の空間位置の関数として、又は距離の程度に基づいて空間Ｒ^{d}で実行され、即ち決定された相関割合又は近接度を考慮して実行される。

【００１６】
例えば、本発明による方法を用いて、検索エンジンが実行した検索により生じるウェブページに対する任意種類の並べ替え動作、評価、分類を実行することができ、最初のページ、即ち他のページから最も遠いページが最も高く評価される。

【００１７】
代替又は組合せとして、ページは、グループで並べ替えられ、各グループは、一組のページに相当し、一組のページの関数Ｘによる投射は、空間Ｒ^{d}の表面の所定の空間領域に位置する。この変形において、空間領域に対するこの表面の区分が定められ、文書は、区分の空間領域のうち一つにおけるその投射の会員に従って評価されるのが好ましい。

【００１８】
また、本発明による方法を用いて、“スパム”の存在を検出することができ、即ち互いを示すページを検出することができ、なぜなら表面Ｓ上の全てのこれらページの投射は、互いに実質的に近接するからである。

【００１９】
また、本発明による方法を用いて、検索エンジンで実行した検索により生じるウェブページの視覚表示を生成することができる。

【００２０】
第１の方法の変形によると、前記文書のうち少なくとも一つは、少なくとも一つの他の文書への少なくとも一つのハイパーテキストリンクを示し、ハイパーテキストリンクの数及び／又は共引用リンクの数の関数として決定される２つの文書ｕ及びｖ間の相関割合は

、この数が多くなるにつれ、さらに高くなる相関割合を文書ｕ及びｖ間に提供し、相関の不在は、リンクの不在に相当する。

【００２１】

この第１の変形により、文書間のハイパーテキストリンク又は共引用リンクを考慮に入れた並べ替え動作の実現が有利に可能となる。

【００２２】

第２の方法の変形によると、２文書ｕ及びｖ間の相関割合は、文書ｕ及びｖの意味的内容の近接程度の関数として決定され、相関割合は、この程度が低くなるにつれ、さらに高くなり、相関の不在は、所定の閾より低い程度に相当する。

【００２３】

この第２の変形により、文書の意味的内容を考慮に入れた並べ替え動作の実現が有利に可能となる。

【００２４】

第３の変形によると、相関割合は、複数のユーザが定めた好ましいページの関数として決定される。この場合、各ユーザは、一組の文書（自分の嗜好ページ）に関連付けられており、２文書ｕ及びｖ間の相関割合は、文書ｕ及びｖが属する当該一組の文書の数であるとして決定される。

【００２５】

この第３の変形は、ページ間の相関割合を決定するのにユーザプロフィールを考慮に入れることを有利に可能にする。

【００２６】

また、第３の変形を共に結合して、ハイパーテキストリンク、意味的内容、及び／又は好ましくはユーザを同時に考慮に入れる相関割合を決定することができる。２文書間の任意のその他種類のリンクも、相関割合を定めるのに使用可能である。

【００２７】

特定の実施形態によると、方法はさらに、

－前記表面上に前記一組の電子文書を投射するための初期関数X_0を定める段階と、

－前記表面上に前記一組の電子文書を投射するための関数Ｘを決定する段階であって、前記投射関数Ｘは、少なくとも一つの反復で初期関数X_0に基づいて得られ、各反復は、X_{i-1}（ｕ）だけでなくX_{i-1}（ｖ）の値と前記一組の電子文書に属する任意の文書ｖに関する文書ｕ及びｖ間の相関割合ω（ｕ、ｖ）とに依存する所定の基準を最適化することを可能にする値X_i（ｕ）にX_{i-1}（ｕ）の値を、前記一組の電子文書のうち少なくとも一つの文書ｕに関して置換することによって、先の反復で得られた関数X_{i-1}に基づいて関数X_iを決定することにある、段階とを具備する。

【００２８】

本発明による方法は、関数Ｘの反復決定に適するので、その実行を容易にし、方法の収束を正確に制御することができる。

【００２９】

関数X_0は、ランダムな方法で定められるのが好ましい。ランダム関数から開始する事実は、所望の関数Ｘに収束する速度を統計的に改善し、得るべき関数の演繹的知識を必要とすることなく改善する。

【００３０】

この実施形態において、前記所定の基準の最適化は、

【００３１】

【数１】

$$\Delta(u) = \sum_{\{u,v\}\in E} \delta(u,v)\left\| X_{i-1}(u) - X_{i-1}(v)\right\|^2$$

に等しく量Δ（ｕ）の値を文書ｕに関して最大にすることにあり、

ここで、δ（ｕ、ｖ）＝１－ω（ｕ、ｖ）、０≦ω（ｕ、ｖ）≦１、文書ｕ及びｖ間の

相関の不在においてω（ｕ、ｖ）＝０であり、
値Ｘ$_{i}$（ｕ）は、Ｘ$_{i}$（ｕ）＝－Ｙ（Ｕ）／‖Ｙ（Ｕ）‖であり、
ここで、
【００３２】
【数２】

$$Y(U) = \sum_{v \in V-\{u\}} \delta(u,v) X_{i-1}(v) \quad \text{if } Y(U) \neq 0,$$

値Ｘ$_{i}$（ｕ）は、Ｙ（Ｕ）＝０の場合にＸ$_{i-1}$（ｕ）に等しい。
【００３３】
また、本発明の要旨は、コンピュータシステムによって読取可能な情報媒体上のコンピュータプログラムであり、前記プログラムは、このプログラムが読込まれてコンピュータシステムで実行される時、上記で簡潔に定めたような本発明による方法を実行するための命令を具備する。
【００３４】
また、本発明の要旨は、データ処理装置であり、本発明による方法からなる段階を実行するためのデータ処理手段を具備する。当該装置は、例えば文書検索エンジンを実行するコンピュータサーバである。
【００３５】
また、本発明の要旨は、コンピュータシステムによって読取可能な記録媒体であり、前記プログラムがコンピュータシステムによって実行される時、本発明による方法を実行するプログラムコード命令を含むプログラムを具備する。
【発明を実施するための最良の形態】
【００３６】
本発明の他の目的、特徴、及び利点は、非限定的な例によってのみ与えられ、添付図面を参照して提供される以下の詳細な説明を通じて明らかとなり、図１は、本発明による一つの実施形態のフローチャートである。
【００３７】
本発明による方法は、一組の電子文書、特に一組のウェブページに適用され、それらの一部が、一つ以上の他のページへの一つ以上のハイパーテキストリンクを含む。
【００３８】
図示された好ましい実施形態において、一組の文書Ｖのうち２つの文書ｕとｖとの間の相関割合は、文書ｕ及びｖ間に存在するハイパーテキストリンクと共引用リンクとの数の関数として決定される。
【００３９】
２つの文書間のハイパーテキストリンク数の決定に関して、ハイパーテキストリンクの方向（sense）が考慮されず“対称化された”ハイパーテキストリンクが考慮され、即ち文書ｕが文書ｖへのリンクを含み、文書ｖが文書ｕへのリンクを含む場合に、同一の処理が適用される。
【００４０】
２つの文書ｕ及びｖは、少なくとも一つの他の文書ｗが存在する場合に共引用リンクを有するので：
－ｗからｕを示す少なくとも一つのハイパーテキストリンクが存在し、
－ｗからｖを示す少なくとも一つのハイパーテキストリンクが存在する。
【００４１】
ここで、本発明による方法のステップは、図１を参照して詳細に説明される。
【００４２】
ステップＳ１００は、一組のＶのうち任意の一対の文書（ｕ、ｖ）に関して重みω$_{1}$（ｕ、ｖ）を決定することであり、その重みは、文書ｕ及びｖ間のハイパーテキストリンクの数に依存する。関数ω$_{1}$（ｕ、ｖ）は、文書ｕ及びｖ間のハイパーテキストリンクの数

の増加関数であることが好ましい。

【００４３】

ω_1（ｕ、ｖ）の値は、所定の最小値（通常０）と所定の最大値（通常１）との間にある。この場合、最小値は、文書ｕ及びｖ間にハイパーテキストリンクが存在しないことに相当し、最大値は、例えば文書ｕ及びｖ間にハイパーテキストリンクが所定の最小数存在することに対応する。

【００４４】

第１の例によると、ω_1（ｕ、ｖ）の値は、文書ｕ及びｖ間にハイパーテキストリンクが存在しない場合には０に等しく選択され、少なくとも一つのハイパーテキストリンクが存在する場合には１に等しく選択される。

【００４５】

第２の例によると、ω_1（ｕ、ｖ）の値は、文書ｕ及びｖ間にハイパーテキストリンクが存在しない場合には０に等しく選択され、単一のハイパーテキストリンクが存在する場合には０．５に等しく選択され、文書ｕ及びｖ間に２つ以上のハイパーテキストリンクが存在する場合には１に等しく選択される。

【００４６】

第３の例によると、ω_1（ｕ、ｖ）の値は、文書ｕ及びｖ間のハイパーテキストリンクの数N_hの連続増加関数として定義され、例えば

【００４７】

【数３】

$$\omega_1(u,v)=\sqrt{\frac{N_h}{N_{h\max}}}$$

ここで、N_{hmax}は、ハイパーテキストリンクの数N_hに対する最大閾値である。

【００４８】

ステップＳ１０５は、一組のＶのうち任意の一対の文書（ｕ、ｖ）に関して重みω_2（ｕ、ｖ）を決定することであり、その重みは、文書ｕ及びｖ間の共引用リンクの数に依存する。関数ω_2（ｕ、ｖ）は、文書ｕ及びｖ間の共引用リンクの数の増加関数であるのが好ましい。

【００４９】

ω_1（ｕ、ｖ）に提供される関数定義の例は、ω_2（ｕ、ｖ）に置換可能である。例えば、ω_2（ｕ、ｖ）の値は、文書ｕ及びｖ間に共引用リンクが存在しない場合には０に等しく選択され、少なくとも一つの共引用リンクが存在する場合には１に等しく選択される。

【００５０】

ステップＳ１１０は、以下の関係によって一対の{ｕ、ｖ}に関連付けられる相関ω（ｕ、ｖ）の割合を任意の一対の文書（ｕ、ｖ）に関して決定することであり、

【００５１】

【数４】

$$\omega(u,v) = k_1\omega_1(u,v) + k_2\omega_2(u,v),$$

ｋ１及びｋ２は、実係数なので

【００５２】

【数５】

$$0 \leq k_1 \leq 1,$$
$$0 \leq k_2 \leq 1,$$
$$k_1 + k_2 = 1.$$

【００５３】
故に、相関ω（ｕ、ｖ）の割合は、０と１との間にある実数を取り、値０は、リンクの不在に相当する。

【００５４】
係数ｋ１に提供される値は、ハイパーテキストリンクの存在を重要視する時、さらに高く選択される。一方、係数ｋ２に提供される値は、共引用リンクの存在を重要視する時、さらに高く選択される。

【００５５】
文書間の相関の割合を決定するためのこの手順は、ハイパーテキストリンク及び共引用リンクという、文書間における２種類のリンクを、本発明による文書分類方法で考慮することを可能にする。

【００５６】
この手順は、他の種類のリンクに一般化される。例えば、ｕからｖへ通過することを可能にする一つ以上のハイパーテキストリンクが存在する場合に間接ハイパーテキストリンクによって２つの文書ｕ及びｖを連結することを定義でき、ハイパーテキストリンクの数は、この場合２以上である。

【００５７】
もう一つの例によると、文書間の意味的種類のリンクを考慮することができる。この場合、２文書間の相関割合の決定は、２文書の意味的内容の分析及び比較に基づき実行される。このために、意味的内容を比較するための周知手順が適用可能である。そして、相関の割合は、２文書間の意味的近接の程度を示す。意味的相関の割合は、例えば各文書に含まれる言葉の意味的分析及び比較に基づき決定することができる。一つの変形として、２文書間の距離が狭まるほどこれら文書間の相関割合が高くなるように、２文書間の距離を定義すること、及び定義された距離の減少関数として相関の割合を定義することができる。

【００５８】
手順は、最終的に任意数のリンクに、それら種類に関係なく一般化可能である。そして、２文書間の相関の割合は、相関の基本割合の加重和、例えば２文書間の共引用リンクの数に依存する相関割合の和、及び２文書間の意味的内容に依存する相関割合の和として決定される。手順は、文書間のハイパーテキストリンクによって、及び文書の意味的内容によって与えられる情報を同時に考慮することができる。

【００５９】
図１に戻ると、以下のステップＳ１２０からＳ１３５は、一組の文書Ｖと一組R^d（Ｒのｄデカルト乗であって、Ｒは、一組の実数を意味し、ｄは、正の整数である）の表面Ｓとの間を投射するための関数Ｘを決定することである。ｄは、２又は３に等しく選択されるのが好ましい。

【００６０】
決定された関数Ｘによって、少なくとも一つの文書ｕに関して、ｖが文書ｕ及びｖ間に相関がある文書である場合における２点Ｘ（ｕ）及びＸ（ｖ）間の距離R^dは、さらに低く、相関の割合が比較的高い。

【００６１】
特定の実施形態によると、反復処理を使用して関数Ｘを決定する。この反復処理の各反復は、所定の基準を最適化することを可能にするX_i（ｕ）の値にX_{i-1}（ｕ）の値を一組Ｖのうち少なくとも一つの文書ｕに関して置換することによって、先のステップで得

られた関数X_{i-1}に基づき関数X_iを決定することである。この基準は一方では、一組Ｖのうち考えられる文書ｕのために得られたX_{i-1}（ｕ）の値と任意の文書ｖのために得られたX_{i-1}（ｖ）の値とに依存し、他方では、一組Ｖのうち文書ｕ及び任意の文書ｖ間の相関割合ω（ｕ、ｖ）に依存する。基準は、上記説明したプロパティを示す関数Ｘに一連の関数X_iを収束させるように選択される。

【００６２】

前記所定の基準に対する最適化は、以下のような量Δ（ｕ）の値を所定の文書ｕに関して最大化することであるのが好ましい。

【００６３】

【数６】

$$\Delta(u) = \sum_{\{u,v\}\in E} \delta(u,v)\left\|X(u)-X(v)\right\|^2$$

ここで、δ（ｕ、ｖ）＝１－ω（ｕ、ｖ）、０≦ω（ｕ、ｖ）≦１、及び文書ｕ及びｖ間の相関がない時にω（ｕ、ｖ）＝０である。

【００６４】

ステップＳ１２０において、初期投射関数X_0が決定される。初期関数X_0は、表面Ｓ上でランダム値をとるのが好ましい。その後反復処理は、現在の関数$X_i = X_0$に適用される。

【００６５】

投射関数Ｘを決定する反復処理は、ステップＳ１２５から開始する。反復は、ステップＳ１２５、Ｓ１３０、及びＳ１３５の実行に相当する。反復は、インデクスｉでインデクスを付けられる。ステップＳ１２０の終了に、インデクスｉは、その初期値をとり、０に等しい。

【００６６】

ステップＳ１２５で、このインデクスは、ｉ＝ｉ＋１増分される。

【００６７】

ステップＳ１３０で、以下の動作は、少なくとも一つの文書ｕに関して実行される。

－

【００６８】

【数７】

$$Y(U) = \sum_{v\in V-\{u\}} \delta(u,v)X_{i-1}(v)$$

の値が決定され

－もしＹ（Ｕ）≠０の場合、その時X_i（ｕ）は、X_i（ｕ）＝－Ｙ（Ｕ）／‖Ｙ（Ｕ）‖によってＹ（Ｕ）から計算され

－もしＹ（Ｕ）＝０の場合、その時X_i（ｕ）は、X_{i-1}（ｕ）に等しい値をとる。

【００６９】

ステップＳ１３５で、反復処理が終了したかが判断される。処理は、一組Ｖのうち文書ｕ毎に関数Ｘが少なくとも１回変更されるのに十分な回数反復されるのが好ましい。

【００７０】

一連の関数X_iは、ランダム開始関数（random starting function）と同等に早く収束し、一組の文書上で限定された回数反復することができる。

【００７１】

また、反復を停止するための決定は：

－既に実行された反復回数

－各反復後に実行された、関数の収束の測定

に基づく。

【００７２】
この収束の測定は、各反復後に以下の方法で合計Δ_iを計算することによって
【００７３】
【数８】

$$\Delta_i = \sum_{u \in V} \left\| X_i(u) - X_{i-1}(u) \right\|$$

及び、一組Ｖのうち文書ｕの数に依存しうる閾値を固定することによって実行可能であり、それ以下であると反復処理が停止する。
【００７４】
反復処理を停止するための決定がステップＳ１３５で行われる場合、その後ステップＳ１４０が実行され、そうでなければステップＳ１２５から開始する以下の反復が実行される。
【００７５】
ステップＳ１４０で、並べ替え動作は、最後の反復で得られた関数Ｘによって取られた値の関数として一組の文書Ｖのうち少なくとも一部の上で実行される。
【００７６】
決定された投射関数Ｘによって、表面Ｓ上の点Ｘ（ｕ）の位置は、文書ｕ及び他の文書間のリンクに依存する。特に、２点間の距離は、これら２点に相当する文書間の相関割合を示す。
【００７７】
そのような関数に初期ランダム関数を収束させるために他の数学的基準を使用することが考えられる。
【００７８】
一組Ｖが検索エンジンにより実行された検索からもたらされる一組のウェブページである場合、この並べ替え動作は以下のことを目的としうる：
－他の投射から最も遠い投射を有するページを検出することによって、最もオリジナルなページを選択すること
－投射が一群のページの投射に実質的に近接するページを検出することによって、“スパム”を含むページ（互いを指すページ）をフィルタリングすること
－投射が決定された基準を満たすページを選択すること。
【００７９】
第１の変形によると、並べ替え動作は、以下の動作を含む：
－距離ｄ（ｕ、ｖ）＝‖Ｘ（ｕ）－Ｘ（ｖ）‖の値を一組Ｖのうち任意の一対の｛ｕ、ｖ｝に関して計算
－例えば所定の閾より上又は下であることによって、値ｄ（ｕ、ｖ）が所定の基準を満たす一組Ｖのうち少なくとも一つのサブセットV_1の決定
【００８０】
この第１の変形は、表面上における一群の点の検出を可能にするので、対応する一群の文書を決定することができる。
【００８１】
第２の変形によると、並べ替え動作は、決定された一組に任意の点Ｘ（ｕ）が属する、例えばR^dの空間における所定の区画にそれが属するサブセットV_1を決定する動作を含む。
【００８２】
例えばこの区画は、表面の、立方体の内部量、あるいはR^dの表面Ｓ上に定義された表面にすることができる。いくつかの所定の区画にこの動作を反復することによって、一組の文書の分離又は区分を構成することができる。
【００８３】
故に、本発明による方法は、決定された関数Ｘによって得られた値に基づき、一組の文

書上で任意の種類の並べ替え動作を実行することが可能となる。

【００８４】

また、関数Ｘを決定する処理が早く収束することを証明できる。

【００８５】

また、この処理の反復に対する計算時間は、相関割合がこのハイパーテキストリンクの数の関数として決定される時、ハイパーテキストリンクの数に比例する。故に本発明の方法は、多数のページ上で使用可能である。

【００８６】

最後に一組の電子文書が（文書追加、文書削除、又は文書間のリンクの変更によって）変更される場合、それは、変更されない一組に関して得られた関数Ｘから開始して、その後２～３の選択された文書に関して（好ましくは、少なくとも変更を受けたか又は追加された文書に関して）ステップＳ１３０の動作を実行するので、変更された一組の電子文書を考慮に入れる訂正された関数Ｘを決定するのに十分である。故に、本発明は、一部が規則的に更新される多数の文書からなる一組の処理に特に適している。

【００８７】

本発明による方法の変形において、関数Ｘの図形表示、即ち前記表面の表示及び前記表面に適した点Ｘ（ｕ）の表示が生成される。そのような図形表示を生成する事実は、一組の関連文書をユーザが選択するのを容易にすることができる。この表示は、例えば２次元の地図表示（cartographic representation）の形式で行われ、各文書は、この文書に関して決定された関数Ｘの値に相当する図形によって区画される。

【００８８】

故に、本発明は、この図形表示がユーザのコンピュータ端末上に表示される実施形態に役立ち、その端末は、ディスプレイ画面及び図形選択ツール（例えば、画面上の図形区画を定めることができるポインタと組み合せて用いるマウス）を含み、このツールは、図形表示の少なくとも一部を選択するのに適する。

【００８９】

その後ユーザは、ユーザが選択した一組以上の文書に相当する図形表示のうち一つ以上の部分を選択することができる。端末は、選択部分を定めるデータを、図形選択ツールを介して得る。これらデータの関数として、端末は、一組の文書Ｖを並べ替える。それは例えば、ユーザが選択した一部に投射がある文書に相当する、低減された文書のリストを生成する。代わりに、投射が選択部分にある文書は、これに反して削除される。ユーザによって保持される文書のリストに基づき、追加の並べ替え動作が実行可能であり、これら動作は、文書のプロパティ又はその相関割合に基づき自動的に実行されるか、又は初期に選択された部分内で選択された新たな部分に基づき手動で実行される。

【００９０】

検索エンジンによって実行される検索の結果を閲覧するこのモードは、ユーザにとって特に人間工学的である。それは、一組の相互に近接した点の形式で、文書の共同体を明らかにする。

【００９１】

故に、本発明で定められるような投射ベースの表示は、図形選択ツールを用いて視覚的及び手動で、又は生成された表示におけるこれら文書の位置にリンクされた所定の基準によって自動的に、並べ替え又は分類を実行することができる。

【００９２】

好ましい実施形態によると、本発明による電子文書の並べ替え方法のステップは、コンピュータプログラムの命令によって決定される。

【００９３】

ここで、“コンピュータプログラム”は、適切なコンピュータシステムによって実行される時、目的が本発明の実施である一組を形成する一つ以上のコンピュータプログラム（ソフトウェア）の意味に理解される。

【００９４】

その後本発明による方法は、例えばインターネットのネットワーク種類に適切でインターネットブラウザソフトウェアを備える場合にリンクされるユーザ端末に組込まれたコンピュータ手段に前記プログラムが読込まれる時に実行される。

【００９５】

結果として、本発明の目的は、そのようなコンピュータプログラムであり、特に情報媒体に記憶されるソフトウェアの形式である。そのような情報媒体は、本発明によりプログラムを並べ替えることができる任意のエンティティ又は装置からなることができる。

【００９６】

例えば、対象となる媒体は、ＲＯＭのようなハードウェア記憶手段、例えばＣＤ　ＲＯＭ若しくは超小型電子回路ＲＯＭ、又は磁気記録手段、例えばハードディスクを含むことができる。変形として、情報媒体は、プログラムが組込まれた集積回路にすることができ、その回路は、対象となる方法を実行するのに適合されるか又はその実行に用いられるように適合される。

【００９７】

また、情報媒体は、無線によって又は他の手段によって電気又は光ケーブルを介してトランクされうる電気又は光信号のような、伝達可能な非ハードウェア媒体でもよい。本発明によるプログラムは、特にインターネットのネットワーク種類からダウンロード可能である。

【００９８】

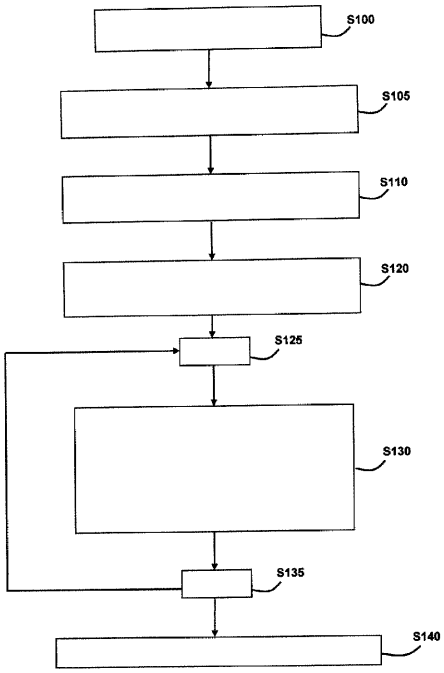
設計的観点では、本発明によるコンピュータプログラムは、任意のプログラム言語を使用でき、ソースコード、オブジェクトコード、又はソースコード及びオブジェクトコード間の中間コードの形式（例えば、部分的にコンパイルされた形式）でもよく、又は本発明による方法を実行する任意の他の好ましい形式でもよい。

【図面の簡単な説明】

【００９９】

【図１】図１は、本発明による一つの実施形態のフローチャートである。

【図１】



【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/FR2006/050856

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GB 2 338 324 A (* BRUNEL UNIVERSITY) 15 December 1999 (1999-12-15) pages 7-21; claims; figures pages 26-28	1-10
X	US 5 987 470 A (MEYERS ET AL) 16 November 1999 (1999-11-16) the whole document	1-10
X	US 5 808 615 A (HILL ET AL) 15 September 1998 (1998-09-15) columns 1-3 column 5, line 40 - column 6, line 67; figures	1-10
A	US 2002/156760 A1 (LAWRENCE STEPHEN R ET AL) 24 October 2002 (2002-10-24) paragraphs [0026] - [0030], [0078]	1-10

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 January 2007	23/01/2007

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Herry, Tzvetanka

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/FR2006/050856

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
GB 2338324 A	15-12-1999	NONE	
US 5987470 A	16-11-1999	NONE	
US 5808615 A	15-09-1998	NONE	
US 2002156760 A1	24-10-2002	US 6289342 B1	11-09-2001

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale n°
PCT/FR2006/050856

A. CLASSEMENT DE L'OBJET DE LA DEMANDE
INV. G06F17/30

Selon la classification internationale des brevets (CIB) ou à la fois selon la classification nationale et la CIB

B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE

Documentation minimale consultée (système de classification suivi des symboles de classement)
G06F

Documentation consultée autre que la documentation minimale dans la mesure où ces documents relèvent des domaines sur lesquels a porté la recherche

Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si cela estréalisable, termes de recherche utilisés)
EPO-Internal

C. DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
X	GB 2 338 324 A (* BRUNEL UNIVERSITY) 15 décembre 1999 (1999-12-15) pages 7-21; revendications; figures pages 26-28	1-10
X	US 5 987 470 A (MEYERS ET AL) 16 novembre 1999 (1999-11-16) le document en entier	1-10
X	US 5 808 615 A (HILL ET AL) 15 septembre 1998 (1998-09-15) colonnes 1-3 colonne 5, ligne 40 - colonne 6, ligne 67; figures	1-10
A	US 2002/156760 A1 (LAWRENCE STEPHEN R ET AL) 24 octobre 2002 (2002-10-24) alinéas [0026] - [0030], [0078]	1-10

☐ Voir la suite du cadre C pour la fin de la liste des documents

☒ Les documents de familles de brevets sont indiqués en annexe

* Catégories spéciales de documents cités:

"A" document définissant l'état général de la technique, non considéré comme particulièrement pertinent

"E" document antérieur, mais publié à la date de dépôt international ou après cette date

"L" document pouvant jeter un doute sur une revendication de priorité ou cité pour déterminer la date de publication d'une autre citation ou pour une raison spéciale (telle qu'indiquée)

"O" document se référant à une divulgation orale, à un usage, à une exposition ou tous autres moyens

"P" document publié avant la date de dépôt international, mais postérieurement à la date de priorité revendiquée

"T" document ultérieur publié après la date de dépôt international ou la date de priorité et n'appartenenant pas à l'état de la technique pertinent, mais cité pour comprendre le principe ou la théorie constituant la base de l'invention

"X" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme nouvelle ou comme impliquant une activité inventive par rapport au document considéré isolément

"Y" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme impliquant une activité inventive lorsque le document est associé à un ou plusieurs autres documents de même nature, cette combinaison étant évidente pour une personne du métier

"&" document qui fait partie de la même famille de brevets

Date à laquelle la recherche internationale a été effectivement achevée	Date d'expédition du présent rapport de recherche internationale
10 janvier 2007	23/01/2007

Nom et adresse postale de l'administration chargée de la recherche internationale	Fonctionnaire autorisé
Office Européen des Brevets, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Herry, Tzvetanka

Formulaire PCT/ISA/210 (deuxième feuille) (avril 2005)

RAPPORT DE RECHERCHE INTERNATIONALE

Renseignements relatifs aux membres de familles de brevets

Demande internationale n°
PCT/FR2006/050856

Document brevet cité au rapport de recherche		Date de publication	Membre(s) de la famille de brevet(s)		Date de publication
GB 2338324	A	15-12-1999	AUCUN		
US 5987470	A	16-11-1999	AUCUN		
US 5808615	A	15-09-1998	AUCUN		
US 2002156760	A1	24-10-2002	US	6289342 B1	11-09-2001

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LS,MW,MZ,NA,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IS,IT,LT,LU,LV,MC,NL,PL,PT,RO,SE,SI,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BW,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,HN,HR,HU,ID,IL,IN,IS,JP,KE,KG,KM,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LT,LU,LV,LY,MA,MD,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PG,PH,PL,PT,RO,RS,RU,SC,SD,SE,SG,SK,SL,SM,SV,SY,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,ZA,ZM,ZW

Ｆターム(参考) 5B075 KK02 ND36 PR06 UU40