(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3781005号
(P3781005)
(45)発行日　平成18年5月31日(2006.5.31)　(24)登録日　平成18年3月17日(2006.3.17)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 220Z
G06F 17/30 370Z

請求項の数 11　(全 12 頁)

(21)出願番号　特願2002-360984(P2002-360984)
(22)出願日　平成14年12月12日(2002.12.12)
(65)公開番号　特開2004-192434(P2004-192434A)
(43)公開日　平成16年7月8日(2004.7.8)
審査請求日　平成14年12月12日(2002.12.12)

(73)特許権者　000002369
セイコーエプソン株式会社
東京都新宿区西新宿２丁目４番１号
(74)代理人　100066980
弁理士　森　哲也
(74)代理人　100075579
弁理士　内藤　嘉昭
(74)代理人　100103850
弁理士　崔　秀▲てつ▼
(74)代理人　100095728
弁理士　上柳　雅誉
(74)代理人　100107076
弁理士　藤綱　英吉
(74)代理人　100107261
弁理士　須澤　修

最終頁に続く

(54)【発明の名称】文書抽出装置及び文書抽出プログラム並びに文書抽出方法

(57)【特許請求の範囲】
【請求項１】
抽出候補となるｎ個の文書を取得し、それら各文書間の全ての類似度を算出する類似度計算手段と、
前記ｎ個の文書の中からｒ個の文書を選ぶ全ての文書の組み合わせを検討し、それら各文書の組み合わせの類似度を合計し、その総和が最も小さいものを抽出する文書抽出手段とを備えたことを特徴とする文書抽出装置。
【請求項２】
上記類似度計算手段は、各文書を所定の文字列毎に分割する文字列分割機能部と、この文字列分割機能部で分割された文字列の出現頻度を基に各文書の文書ベクトルを算出する文字列頻度計算機能部と、この文字列頻度計算機能部で得られた文書ベクトルを基に各文書間の類似度を計算する相互類似度計算機能部とを備えたことを特徴とする請求項１に記載の文書抽出装置。
【請求項３】
上記文字列分割機能部は、形態素解析、ｎ－ｇｒａｍ、ストップワードのいずれかの文字列分割方式を用いて上記各文書を所定の文字列毎に分割することを特徴とする請求項２に記載の文書抽出装置。
【請求項４】
上記文字列頻度計算機能部は、分割された文字列の出現頻度を基に各文書をＴＦＩＤＦで重み付けした文書ベクトルを生成することを特徴とする請求項２または３に記載の文書抽

出装置。
【請求項５】
上記相互類似度計算機能部は、各文書の文書ベクトルを基にベクトル空間法を用いて各文書間の類似度を計算することを特徴とする請求項２～４のいずれかに記載の文書抽出装置。
【請求項６】
コンピュータを、
抽出候補となるｎ個の文書を取得してそれら各文書間の全ての類似度を算出する類似度計算手段と、
前記ｎ個の文書の中からｒ個の文書を選ぶ全ての文書の組み合わせを検討し、それら各文書の組み合わせの類似度を合計し、その総和が最も小さいものを抽出する文書抽出手段として機能させることを特徴とする文書抽出プログラム。
【請求項７】
上記類似度計算手段は、各文書を所定の文字列毎に分割する文字列分割機能と、この文字列分割部で分割された文字列の出現頻度を基に各文書の文書ベクトルを算出する文字列頻度計算機能と、この文字列頻度計算部で得られた文書ベクトルを基に各文書間の類似度を計算する相互類似度計算機能とを備えたことを特徴とする請求項６に記載の文書抽出プログラム。
【請求項８】
上記類似度計算手段は、形態素解析、ｎ－ｇｒａｍ、ストップワードのいずれかの文字列分割方式を用いて上記各文書を文字列毎に分割する文字列分割機能と、分割された文字列の出現頻度を基にＴＦＩＤＦで重み付けした各文書の文書ベクトルを生成する文字列頻度計算機能と、各文書の文書ベクトルを基にベクトル空間法を用いて各文書間の類似度を計算する相互類似度計算機能と、を備えたことを特徴とする請求項６に記載の文書抽出プログラム。
【請求項９】
類似度計算手段が、抽出候補となるｎ個の文書を取得してそれら各文書間の全ての類似度を算出した後、
文書抽出手段が、前記ｎ個の文書の中からｒ個の文書を選ぶ全ての文書の組み合わせを検討してからそれら各文書の組み合わせの類似度を合計し、その総和が最も小さいものを抽出するようにしたことを特徴とする文書抽出方法。
【請求項１０】
上記類似度計算手段による上記抽出候補となる各文書間の類似度は、各文書を所定の文字列毎に分割した後、分割された文字列の出現頻度を計算し、その文字列の出現頻度を基にして各文書の文書ベクトルを算出し、その文書ベクトルを用いて算出するようにしたことを特徴とする請求項９に記載の文書抽出方法。
【請求項１１】
上記類似度計算手段による上記抽出候補となる各文書間の類似度は、形態素分析，ｎ－ｇｒａｍ，ストップワードのいずれかの文字列分割方式を用いて各文書を所定の文字列毎に分割した後、分割された文字列の出現頻度を基にＴＦＩＤＦで重み付けした各文書の文書ベクトルを算出し、その文書ベクトルを基にベクトル空間法によって算出するようにしたことを特徴とする請求項９に記載の文書抽出方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、ニュース等の文書をユーザの好みに応じて自動的に配信する文書配信システム等に係り、特に配信候補となる数多い文書の中から内容的に類似している文書を排除して特徴のある文書のみを抽出するための装置及びその抽出プログラム並びに抽出方法に関するものである。
【０００２】

【従来の技術】
ユーザ毎にカスタマイズが可能な情報配信システムは、ユーザがフィルタリング条件を設定し、リアルタイムで送られてくるニュース等の各種情報（以下、文字情報を主体とした文書という。）の中からコンピュータが自動的にその設定されたフィルタリング条件に合う文書のみを抽出してユーザに配信する形態が一般的である。
【０００３】
このような形態の文書配信システムの場合、フィルタリング条件によっては、配信される文書が偏りすぎたり、また、同じような内容の文書が繰り返し送られてくるといった問題点がある。特に、後者の問題点に関しては、文書の内容が重複することにより、配信される情報に無駄が多くなったり、文書掲載スペースが限られている場合には他の重要な文書がカットされてしまう等の不都合を招き、文書配信システム自体の利便性や信頼性等を大きく損なう結果となる。
【０００４】
そのため、このような文書の重複配信を防止すべく必要な文書のみを効率的に抽出するためのフィルタリング、あるいは分類技術が極めて重要となっており、これらに関する従来技術として、例えば以下の特許文献１及び２に示すような技術が提案されている。
先ず、特許文献１（特許第３２０３２０３号公報）には、すべての文書にキーワードを付与し、そのキーワードから文書をベクトル化し、ある文書Ａが他の文書Ｂに包含されているときに最大値をとるような類似度評価尺度を導入して、代表文書、従属文書、独立文書等を認識して適宜関係のある文書をまとめる等の技術が開示されている。
【０００５】
一方、特許文献２（特開２０００－１４８７７０号公報）には、分類対象となる文書の特徴量を計算し、それら各特徴量の類似度を求めた後、数学的、統計的なクラスタ分析によって文書を分類する等の技術が開示されている。
【０００６】
【特許文献１】
特開平９－９６４１８号公報
【特許文献２】
特開２０００－１４８７７０号公報
【０００７】
【発明が解決しようとする課題】
ところで、前者の従来技術にあっては、全ての文書にキーワード等の特徴を付与する必要性があるが、全ての文書に対してキーワードを付与する作業には多くのコストがかかる。一方、後者の従来技術で用いているクラスタ分析は階層的な分類やグループ分けをするために適した分析方法であるが、文書数が増えると計算量が極端に多くなってしまい、スループットが大きく低下するといった問題点がある。
【０００８】
そこで、本発明は、このような従来の技術の有する未解決の課題に着目してなされたものであり、その目的は、低コストでかつ抽出に要する計算量も少ない新規な文書抽出装置及び文書抽出プログラム並びに文書抽出方法を提供するものである。
【０００９】
上記課題を解決するために発明１の文書抽出装置は、
抽出候補となるｎ個の文書を取得し、それら各文書間の全ての類似度を算出する類似度計算手段と、前記ｎ個の文書の中からｒ個の文書を選ぶ全ての文書の組み合わせを検討し、それら各文書の組み合わせの類似度を合計し、その総和が最も小さいものを抽出する文書抽出手段とを備えたものである。
【００１０】
このような構成を採用することにより、抽出候補となる複数の文書のなかからいくつかの文書を抽出する際に、類似度が大きい文書同士は一緒に選択されることがなくなるため、内容が類似するような文書を重複して抽出する可能性を大幅に低下させることができる。

また、文書の抽出に際しては各文書にキーワードを付与する等の作業を要しないため、その作業に要するコストが不要となる。また、各文書間の類似度の総和によって文書の組み合わせを抽出するため、文書量が多くなった場合でも計算量が極端に増えてしまうようなこともない。

【００１１】

また、発明２の文書抽出装置は、上記類似度計算手段を、各文書を所定の文字列毎に分割する文字列分割機能部と、この文字列分割機能部で分割された文字列の出現頻度を基に各文書の文書ベクトルを算出する文字列頻度計算機能部と、この文字列頻度計算機能部で得られた文書ベクトルを基に各文書間の類似度を計算する相互類似度計算機能部とで構成したものである。

【００１２】

このような構成を採用することにより、各文書間の類似度を的確に算出することができ、発明１の効果を確実に実現することができる。

また、発明３の文書抽出装置は、上記文字列分割機能部が、形態素解析、ｎ－ｇｒａｍ、ストップワードのいずれかの文字列分割方式を用いて上記各文書を文字列毎に分割するようにしたものである。

【００１３】

すなわち、これら形態素解析、ｎ－ｇｒａｍ、ストップワードといった文字列分割方式は従来から多用されている信頼性に優れた方式であり、これらを本発明の文字列分割機能部として用いることにより、各文書を的確に文字列に分割できることは勿論、これらいずれかの方式を用いることにより様々な形態の文書にも的確に対応することができる。

【００１４】

また、発明４の文書抽出装置は、上記文字列頻度計算機能部が、分割された文字列の出現頻度を基にＴＦＩＤＦで重み付けした各文書の文書ベクトルを生成するようにしたものである。

すなわち、各文書の文書ベクトルを生成するに際し、分割された文字列の出現回数をそのまま用いても良いが、後述するＴＦＩＤＦという文字列の重要度を反映した公知の重み付け方法を用いると各文書の特徴を良く表現した文書ベクトルを生成できる。

【００１５】

また、発明５の文書抽出装置は、上記相互類似度計算部が、各文書の文書ベクトルを基にベクトル空間法を用いて各文書間の類似度を計算するようにしたものである。

すなわち、各文書間の類似度を計算する方式としてベクトル空間法を用いれば、２つのベクトルの類似度は２つのベクトルのなす角の余弦（０～１）として定量的に表現することが可能となり、後の文書抽出をより的確に行うことが可能となる。

【００１６】

発明６の文書抽出プログラムは、コンピュータを、

抽出候補となるｎ個の文書を取得してそれら各文書間の全ての類似度を算出する類似度計算手段と、前記ｎ個の文書の中からｒ個の文書を選ぶ全ての文書の組み合わせを検討し、それら各文書の組み合わせの類似度を合計し、その総和が最も小さいものを抽出する文書抽出手段として機能させることを特徴とするものである。

【００１７】

これにより、上述した各手段を実施するに際し、専用のハードウエアを用意することなくソフトウェアによって安価な汎用のパソコンをそのまま使用することができ、実施等に要するコストの大幅な削減や実施までの準備期間等を大幅に短縮することが可能となる。

また、発明７の文書抽出プログラムは、上記類似度計算手段を、各文書を所定の文字列毎に分割する文字列分割機能と、この文字列分割部で分割された文字列の出現頻度を基に各文書の文書ベクトルを算出する文字列頻度計算機能と、この文字列頻度計算部で得られた文書ベクトルを基に各文書間の類似度を計算する相互類似度計算機能とで実現したものである。

【００１８】

これにより、発明２と同様に各文書間の類似度をソフトウェアによって的確に算出することができる。
また、発明８の文書抽出プログラムは、上記類似度計算手段を、形態素解析、ｎ－ｇｒａｍ、ストップワードのいずれかの文字列分割方式を用いて上記各文書を文字列毎に分割する文字列分割機能と、分割された文字列の出現頻度を基にＴＦＩＤＦを用いて各文書の文書ベクトルを生成する文字列頻度計算機能と、各文書の文書ベクトルを基にベクトル空間法を用いて各文書間の類似度を計算する相互類似度計算機能とで実現したものである。
【００１９】
これにより、発明３～５と同様な作用・効果をソフトウェア上で確実に達成することができる。
発明９の文書抽出方法は、
文書抽出手段が、抽出候補となるｎ個の文書を取得してそれら各文書間の全ての類似度を算出した後、文書抽出手段が、前記ｎ個の文書の中からｒ個の文書を選ぶ全ての文書の組み合わせを検討してからそれら各文書の組み合わせの類似度を合計し、その総和が最も小さいものを抽出するようにしたことを特徴とするものである。
【００２０】
これによって、発明１の文書抽出装置と同様に、内容が類似（重複）するような文書を抽出する可能性を大幅に低下させることができると共に、文書抽出処理に要するコストが安価となり、文書量が多くなった場合でも計算量が大幅に増えてしまうようなこともない。
また、発明１０の文書抽出方法は、
上記類似度計算手段による上記抽出候補となる各文書間の類似度は、各文書を所定の文字列毎に分割した後、分割された文字列の出現頻度を計算し、その文字列の出現頻度を基にして各文書の文書ベクトルを算出し、その文書ベクトルを用いて算出するようにしたことを特徴とするものである。
【００２１】
これにより、発明２と同様に、各文書間の類似度を的確に算出することができる。
また、発明１０の文書抽出方法は、
上記類似度計算手段による上記抽出候補となる各文書間の類似度は、形態素分析，ｎ－ｇｒａｍ，ストップワードのいずれかの文字列分割方式を用いて各文書を所定の文字列毎に分割した後、分割された文字列の出現頻度を基にＴＦＩＤＦで重み付けした各文書の文書ベクトルを算出し、その文書ベクトルを基にベクトル空間法によって算出するようにしたことを特徴とするものである。
【００２２】
これにより、発明３～５と同様な作用・効果を確実に達成することができる。
【００２３】
【発明の実施の形態】
以下、本発明の実施の形態を添付図面を参照しながら詳述する。
先ず、図１は本発明に係る文書抽出装置１０の実施の一形態を示したものである。図示するように、この文書抽出装置１０はインターネット等の情報通信網内にある情報供給源Ｓから供給されるいくつかの情報をそれぞれ文書として一時的に記憶しておく情報記憶手段１２と、この情報記憶手段１２に記憶された複数の文書をまとめて取得してそれら各文書間の類似度を算出する類似度計算手段１４と、この類似度計算手段１４で得られた各文書間の類似度を基にしてその文書群の中からいくつかの文書のみを抽出する文書抽出手段１６とから主に構成されている。
【００２４】
また、図示するようにこの類似度計算手段１４はさらに、文字列分割機能部１８と、文字列頻度計算機能部２０と、相互類似度計算機能部２２とから構成されており、後に詳述するが、情報記憶手段１２から取得された各文書に対し、文字列分割機能部１８によって各文書を文字列毎に分割した後、分割された各文字列の出現頻度を文字列頻度計算機能部２

０によって算出して各文書のベクトルを算出し、その後、この文字列頻度計算機能部２０で得られた各文書の文書ベクトル相互の類似度を相互類似度計算機能部２２によって算出してそのデータを求めるようになっている。

【００２５】

具体的にはこの文書抽出装置１０は図２に示すような構成をしたコンピュータ１００により実現されることになる。

図示するように、このコンピュータ１００は、制御プログラムに基づいて演算および装置全体を制御するＣＰＵ３０と、所定領域にあらかじめＣＰＵ３０の制御プログラム等を格納しているＲＯＭ３２と、ＲＯＭ３２等から読み出したデータやＣＰＵ３０の演算過程で必要な演算結果を格納するためのＲＡＭ３４と、外部装置に対してデータの入出力を媒介するＩ／Ｆ３８とで構成されており、これらは、データを転送するための信号線であるバス３９で相互にかつデータ授受可能に接続されている。

【００２６】

Ｉ／Ｆ３８には、外部装置として、データ入力が可能なキーボードやマウス等からなる入力装置４０と、画像信号に基づいて画面を表示する表示装置４２と、前述したように情報供給源Ｓから供給される情報を所定の文書データとして一時的に記憶するための情報記憶手段１２とが接続されている。この情報記憶手段１２は、例えばハードディスク等の外部記憶装置であり、インターネット等の情報供給源Ｓから所定の情報が定期的にあるいは随時供給されるようになっている。

【００２７】

そして、ＣＰＵ３０は、マイクロプロセッシングユニットＭＰＵ等からなり、ＲＯＭ３２の所定領域に格納されている文書抽出プログラムを起動させ、その文書抽出プログラムに従って上記類似度計算手段１４に相当する処理と上記文書抽出手段１６に相当する処理をそれぞれ時分割で実行するようになっている。

以上において、本実施の形態の動作を説明する。

【００２８】

図１に示すように、先ず情報記憶手段１２には情報供給源Ｓからユーザの好みに対応した文書が一定の時間毎もしくは不定期に供給されて一時的に蓄積され、その文書数が所定数に達したとき、あるいは保存時間が一定時間経過したならば、一旦、その蓄積された文書の全てが類似度計算手段１４に送られ、ここで各文書間の類似度が算出される。

【００２９】

すなわち、類似度計算手段１４に送られてきた各文書は、先ず文字列分割機能部１８によって文字列に分割される。この文字列の分割方式（手法）は特に限定されるものではないが、図３に示すように各文書D_1～D_mを文字列に分割するに際して形態素解析を用いた場合は、形態素解析辞書を参照しながら文法的な区切りで文字列（単語）に分割することができる。ここで、この形態素解析には様々な手法があり、辞書の善し悪しによっても結果は異なるが、例えば同図の「無線／の／セキュリティ／が／話題／に／なっ／ている／。／…」等のように、名詞、動詞、形容詞助詞、助動詞等の単語に分けることができる。また、この形態素解析は分割の精度が良いが、以前では精度を維持するために辞書の作成やメンテナンスにコストは掛かるといった欠点があったが、最近では長年十分に作り込まれてきた辞書が資産として使えるため、コストの問題も次第に解消されてきており、現在最もよく使われる文字列分割方法である。ただし、この形態素解析は日本語に限って使用できるものであり、英語や中国語などの他の言語には使えないといった不利な面もある。

【００３０】

また、このように各文書D_1～D_mを文字列に分割するに際して形態素解析ではなく、一定間隔毎に文字列を切っていくｎ－ｇｒａｍと言う文字列分割方式を用いることも可能であり、このｎ－ｇｒａｍ方式を用いた場合、上記文書は図４に示すように分割される。すなわち、このｎ－ｇｒａｍの「ｎ」とは何バイト毎（または何文字毎）かを表す数字で、図４の場合では２文字毎なので２－ｇｒａｍと書くことができる。ただし、日本語などの２バイト文字の場合、２文字＝４バイトなので４－ｇｒａｍと書く場合もあるかもしれな

いが、ここではその数字の正確さは問題にすることではない。ｎ－ｇｒａｍは、意味のある単語を塊として切り出すことは困難であるが、分割したものをそのまま統計的に処理するだけであれば必ずしも意味のある単語が塊になっている必要がない場合もある。また、このｎ－ｇｒａｍは上記形態素解析に比べてアルゴリズムが単純でどの言語に対しても使えるというメリットがある。

【００３１】

また、他の文字列分割方式として図５に示すようなストップワードという方式を使用しても良い。このストップワード方式とは文書の中で切れ目となる文字や規則を登録し、それに従って分割していく方法である。例えば、図５に示す例では、▲１▼助詞だと思われる「の」「は」「が」「に」「を」「や」、▲２▼句読点「、」「。」、▲３▼漢字、カタカナ、アルファベット等の字種の変わり目、等といった３つのルールのいずれかが成立するところで分割したものである。尚、このストップワードはある程度意味のある単語を抜き出すことが可能であるが、「情報通信技術」等といった長い熟語や「インターネットテクノロジー」等といった長いカタカナの複合語などは分割できないという問題もある。また、英語であれば、▲１▼スペース、▲２▼カンマ、ピリオド、コロン、セミコロン、その他の記号、▲３▼アルファベット、数字、記号などの字種の変わり目等といったルールをもとに、単語の活用形を落とすステミングという手法を使うことである程度の文字列分割を行うことができる。

【００３２】

このように上記文字列分割機能部１８によって全ての文書D_1～D_mについての文字列分割が行われたならば、次に文字列頻度計算機能部２０によって文字列頻度の計算を行い、図７に示すような文字列－文書行列を作成する。この文字列－文書行列は各文書D_1～D_mとユニークな文字列T_1～T_nとの対応関係を示したものであり、各文字列T_1～T_nが各文書D_1～D_m中に何回出現するかを数え、それを示したものである。例えば、文字列分割方式として形態素解析を用いた分割結果の場合では、図６に示すように、文書D_1中には「無線」（網掛け文字）という文字列（T_1）は３回出現しており、その$W_{1\ 1}$に相当する行列の要素は、その出現回数をそのまま用いた場合では「３」となる。

【００３３】

ここで、$W_{m\ n}$に相当する行列の各要素は、文字列の出現回数をそのまま用いても良いが、ＴＦＩＤＦ（Term Frequency & Inverse Document Frequency）という文字列の重要度を反映した重み付け方法を用いると、各文書の特徴を良く表現した文書ベクトルが生成できることが知られており、後の相互類似度計算で活用することができる。

【００３４】

すなわち、このＴＦＩＤＦは、以下の数式１に示すように、ある文書Ｄ内での文字列Ｔの出現頻度（ＴＦ：Term Frequency）と、文書集合全体で文字列Ｔが出現する文書数の頻度を逆数（ＩＤＦ：Inverse Document Frequency）の積で求め、数値が大きいほどその文字列Ｔが重要であることを表すものである。ＴＦは頻出する文字列は重要であるという指標であり、ある文書中に文字列が出現する頻度が増加すると大きくなる性質を持っている。ＩＤＦは多くの文書中に出現する文字列は重要でない、つまり特定の文書に出現する文字列が重要であるという指標であり、ある文字列が使われている文書数が減少すると大きくなると言う性質を持っている。従って、このＴＦＩＤＦの値は、頻繁に出現するが多くの文書に出現する文字列（接続詞や助詞等）や、特定の文書にのみ出現するがその文書に高頻度で出現する文字列に対しては大きくなる性質を持っており、このＴＦＩＤＦによって文書中の文字列は数値化され、その数値を要素として文書をベクトル化することができる。

【００３５】

【数１】

$$W(t,d) = TF(t,d) \times IDF(t)$$

但し、$TF(t,d)$＝文書dに文字列tが出現する頻度

$$IDF(t) = \log\left(\frac{D}{DF(t)}\right)$$

$DF(t)$＝文書全体で文字列tが出現する文書数の頻度
D＝全文書数

【００３６】
このような文書取得から文字列計算までの流れを表したのが図８に示すフローチャートである。図示するように、ステップＳ１００では情報記憶手段１２にストックされた文書を一つずつ取得し、ステップＳ１０２では取得したその文書を文字列毎に分割し、ステップ１０４ではその文書と文字列との対応関係を示す文字列－文書行列に文字列の頻度情報を記憶してステップＳ１０６に移行する。ステップＳ１０６では情報記憶手段１２にストックされている文書が残っているか否かを判定し、残っている場合（ＹＥＳ）には、その文書を取得して同様な処理を行い、全ての文書がなくなるまで繰り返す。
【００３７】
一方、ステップＳ１０６において情報記憶手段１２にストックされている文書が残っていないと判定した場合（ＮＯ）にはステップＳ１０８に移行し、完成した文字列－文書行列の頻度情報を基にＴＦＩＤＦによって重み付けし直した文字列－文書行列を作成する。これによって全ての文書はそれらに出現するユニークな文字列の数と同じ次元（数千～数十万）のベクトルとして表現できることになる。
【００３８】
そして、このようにして全ての文書がベクトル化されたならば、相互類似度計算機能部２２によって各文書間の類似度が求められる。具体的には、この相互類似度計算機能部２２は公知のベクトル空間法を採用するものであり、上記ＴＦＩＤＦによって求められた各文書ベクトルは、このベクトル空間法によって相互の類似度が定義されることになる。すなわち、対比する２つの文書のベクトル類似度は、図９に示すように２つのベクトルのなす角θの余弦値（０～１）として定義できることから、各文書同士の類似度は図１０に示すような対称行列で表現できる。
【００３９】
その後、その対称行列をもとに類似情報のグルーピングやカットを行うことで類似文書を排除したフィルタリングが実現可能となる。例えば、図１０のような対称行列では文書D_1と文書D_2の類似度は図１１に示すように０．９，文書D_1と文書D_3の類似度は０．３というように各文書間の類似度が定量的に示される。
【００４０】
次に、このように類似度計算手段１４によって各文書間の類似度が定量的に求められたならば、文書抽出手段１６が、その文書群のなかから各文書D_1～D_n間の類似度の総和が最も小さくなる文書の組み合わせを抽出することになる。
具体的に説明すると、この文書抽出手段１６は、取得された全ての文書ｎ個の中からｒ個（文書の配信量やレイアウトの都合等によって決定される。）の文書を抽出することになるが、このｒ個の文書を抽出するに際しては、以下の数式２に示すようにｎ個の文書の中からｒ個を選ぶ全ての文書の組み合わせを検討し、それら各組み合わせの類似度を合計し、その総和が最も小さいものを抽出することになる。
【００４１】
【数２】

$$ {}_nC_r = \frac{n!}{r!(n-r)!} $$

【００４２】
例えば、抽出候補となる文書が４つ（D_1，D_2，D_3，D_4）あり、その中から３つの文書を抽出する組み合わせは以下の数式３に示すように４通りとなる。
【００４３】
【数３】

$$ {}_4C_3 = \frac{4!}{3!(4-3)!} = \frac{4\times3\times2\times1}{(3\times2\times1)\times(1)} = 4 $$

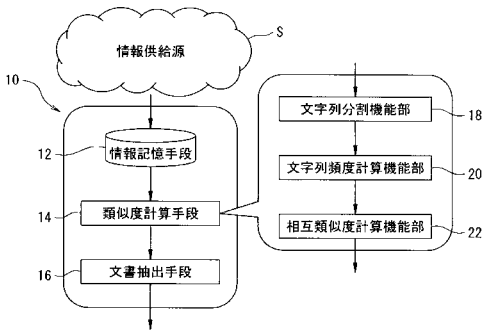
【００４４】
そして、それぞれの組み合わせからできる２つ組の類似度を図１１に示す類似度対称行列からピックアップしていき、以下の表１に示すように各組み合わせI、II、III、IV毎にそれらの類似度を単純に足してその総和を計算する。
【００４５】
【表１】

I	D1, D2, D3⇒(D1, D2)=0.9, (D1, D3)=0.3, (D2, D3)=0.2⇒0.9+0.3+0.2=1.4
II	D1, D2, D4⇒(D1, D2)=0.9, (D1, D4)=0.5, (D2, D4)=0.8⇒0.9+0.5+0.8=2.2
III	D1, D3, D4⇒(D1, D3)=0.3, (D1, D4)=0.5, (D3, D4)=0.3⇒0.3+0.5+0.3=1.1
IV	D2, D3, D4⇒(D2, D3)=0.2, (D2, D4)=0.8, (D3, D4)=0.3⇒0.2+0.8+0.3=1.3

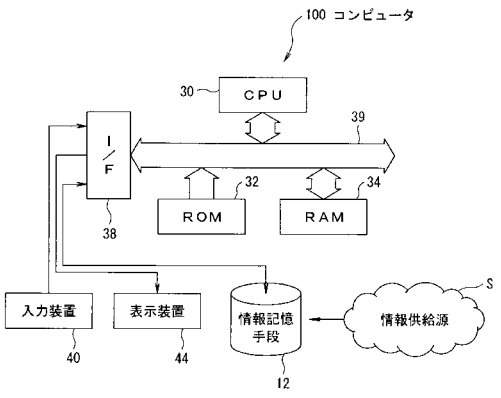
【００４６】
表１に示す例では、４つの組み合わせI、II、III、IVのうちII（D_1，D_2，D_4）の組み合わせの類似度の総和が２．２と最も高く、III（D_1，D_3，D_4）の組み合わせの類似度の総和が１．１と最も低くなったことから、４つの文書（D_1，D_2，D_3，D_4）のなかから文書D_1、D_3、D_4の文書を組み合わせて抽出することになる。
【００４７】
このように本発明は各文書間の類似度の総和が最も小さくなる文書の組み合わせを抽出するようにしたことから、各文書の類似度が小さくなり、内容が似通った文書を同時に抽出する可能性を大幅に減少することができる。
この結果、これを前述した文書配信システム等に適用した場合、重複した内容の文書がユーザに配信される等といった不都合を未然に回避することが可能となると共に、抽出に際して文書毎にキーワードを付与するなどといった煩わしい作業が不要となり、文書抽出処理に要するコストの削減に大きく貢献することができる。また、各文書間の類似度の算出に際しては計算量が大きくなる可能性のあるクラスタ分析を用いる必要がなくなり、非力なコンピュータ等によってもその機能を十分発揮することが可能となる。
【００４８】
尚、本実施の形態では類似度の総和として、定量的に求められた各文書間の類似度を単に足した数値を用いたが、その他に二乗和、対数和等を用いても良い。ただし、本実施の形態のように文書ベクトルの類似度は余弦を使って０～１の間の値として正規化されているので、敢えて二乗和や対数和等の非線形な関数を使う必要はなく、そのまま足し算することで十分である。
【図面の簡単な説明】
【図１】文書抽出装置の構成を示すブロック図である。
【図２】コンピュータの構成を示すブロック図である。

【図３】形態素解析による文字列分割の一例を示す図である。
【図４】ｎ－ｇｒａｍによる文字列分割の一例を示す図である。
【図５】ストップワードによる文字列分割の一例を示す図である。
【図６】形態素解析による文字列分割結果を示す図である。
【図７】文字列－文書行列を示す図である。
【図８】文字列－文書行列を得るための流れを示すフローチャート図である。
【図９】文書ベクトル及びその相関関係を示す図である。
【図１０】文書－文書間の対称行列を示す図である。
【図１１】文書－文書間の対称行列を示す図である。
【符号の説明】
10…文書抽出装置、12…情報記憶手段、14…類似度計算手段、16…文書抽出手段、18…文字列分割機能部、20…文字列頻度計算機能部、22…相互類似度計算機能部、100…コンピュータ、D_1～D_m…文書、S…情報供給源

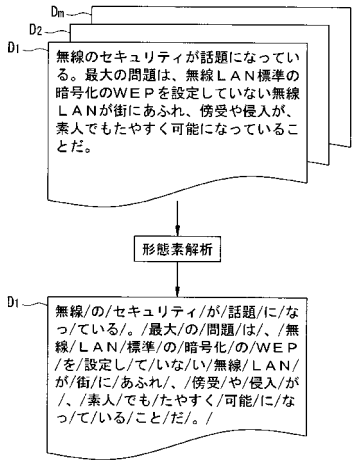
【図１】



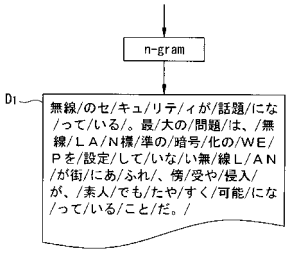
【図２】



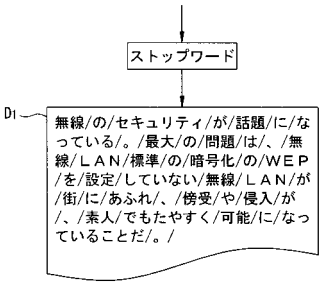
【図３】



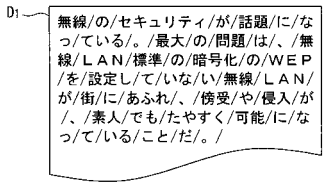
【図４】



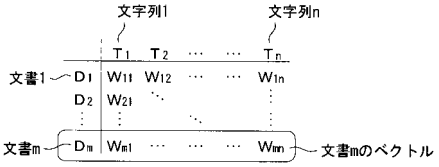
【図５】



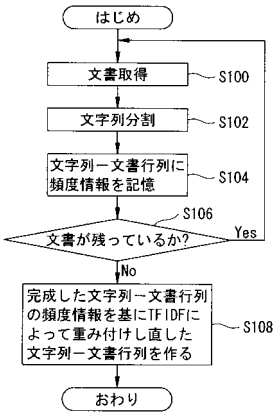
【図６】



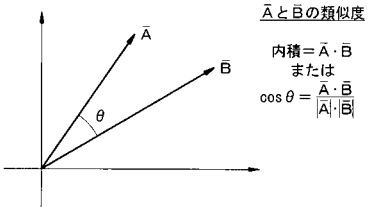
【図７】



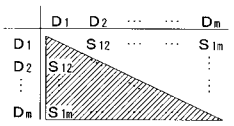
【図８】



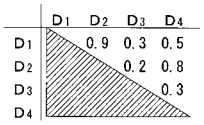
【図９】



【図１０】



【図１１】



フロントページの続き

(72)発明者 萱原 直樹
長野県諏訪市大和３丁目３番５号 セイコーエプソン株式会社内
(72)発明者 大橋 洋貴
長野県諏訪市大和３丁目３番５号 セイコーエプソン株式会社内

審査官 深津 始

(56)参考文献 特開平０９－２３１２３８（ＪＰ，Ａ）
特開２００１－２７３３０２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/30
JICSTファイル(JOIS)