(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3612125号
(P3612125)
(45)発行日　平成17年1月19日(2005.1.19)　(24)登録日　平成16年10月29日(2004.10.29)

(51)Int.Cl.[7]　F I
G06F 17/30　G06F 17/30 220A
G06F 17/30 170A
G06F 17/30 340B

請求項の数 4　(全 15 頁)

(21)出願番号　特願平7-325907
(22)出願日　平成7年12月14日(1995.12.14)
(65)公開番号　特開平9-167164
(43)公開日　平成9年6月24日(1997.6.24)
審査請求日　平成13年9月14日(2001.9.14)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(74)代理人　100058479
弁理士　鈴江　武彦
(72)発明者　三池　誠司
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(72)発明者　梶浦　正浩
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(72)発明者　酒井　哲也
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

最終頁に続く

(54)【発明の名称】情報フィルタリング方法および情報フィルタリング装置

(57)【特許請求の範囲】
【請求項１】
複数の情報源から配信された、テキストやイメージなどを含む記事を受信する受信手段と、
前記受信手段で受信された記事を記憶する第１の記憶手段と、
ユーザにより予め指定されたテーマに関する記事を検索するための単語とその重み値が指定されている検索条件を記憶する第２の記憶手段と、
前記第１の記憶手段で記憶された記事の中から、ユーザに配信する記事を抽出するとともに、当該記事の抄録を作成する情報フィルタリング処理を行う演算手段と、
前記情報フィルタリング処理を行った結果を配信する配信手段と、
を備えたコンピュータにおける情報フィルタリング方法において、
前記演算手段が、
（１）前記第１の記憶手段で記憶された各記事に含まれる前記検索条件の単語の出現回数と当該単語の前記重み値を用いて、各記事の当該検索条件に対する類似度を算出して、当該類似度の高い順にランク付けされた複数の記事を検索する検索ステップと、
（２）前記検索ステップで検索された各記事を、記事本文の文字数及び段落数が予め定められた第１の閾値よりも少ない第１のグループ、当該第１の閾値以上でかつ当該第１の閾値よりも大きい第２の閾値よりも少ない第２のグループ及び当該第２の閾値以上の第３のグループのうちのいずれか１つに分類するステップと、
（３）前記第１のグループに分類された記事について、その最初の段落を抄録として抽

出する第１の抄録作成ステップと、
（４）前記第２のグループに分類された記事の各段落に対し、当該段落に含まれる前記検索条件の単語の出現回数と当該単語の前記重み値を用いて当該検索条件に対する類似度を算出し、当該記事から最初の段落及び少なくとも当該類似度が最も高い段落を抄録として抽出する第２の抄録作成ステップと、
（５）前記第３のグループに分類された記事の各段落に対し、当該段落に含まれる前記検索条件の単語の出現回数と当該単語の前記重み値を用いて当該検索条件に対する類似度を算出し、当該記事から当該類似度が高い順に所定数の段落を抄録として抽出する第３の抄録作成ステップと、
を実行し、
前記配信手段が、前記検索ステップで検索された各記事のランクと、前記第１乃至第３の抄録作成ステップのうちのいずれかで作成された抄録とを配信する配信ステップと、
を実行する
ことを特徴とする情報フィルタリング方法。
【請求項２】
前記演算手段が、
前記配信ステップで抄録の配信された記事のうち、前記ユーザが関心ありと回答した記事については、その中に含まれる出現頻度の最も高い単語を前記検索条件に追加するステップをさらに有することを特徴とする請求項１記載の情報フィルタリング方法。
【請求項３】
複数の情報源から配信された、テキストやイメージなどを含む記事を受信する受信手段と、
前記受信手段で受信された記事を記憶する第１の記憶手段と、
ユーザにより予め指定されたテーマに関する記事を検索するための単語とその重み値が指定されている検索条件を記憶する第２の記憶手段と、
前記第１の記憶手段で記憶された各記事に含まれる前記検索条件の単語の出現回数と当該単語の前記重み値を用いて、各記事の当該検索条件に対する類似度を算出して、当該類似度の高い順にランク付けされた複数の記事を検索するための検索手段と、
前記検索手段で検索された各記事を、記事本文の文字数及び段落数が予め定められた第１の閾値よりも少ない第１のグループ、当該第１の閾値以上でかつ当該第１の閾値よりも大きい第２の閾値よりも少ない第２のグループ及び当該第２の閾値以上の第３のグループのうちのいずれか１つに分類する手段と、
前記第１のグループに分類された記事について、その最初の段落を抄録として抽出する第１の抄録作成手段と、
前記第２のグループに分類された記事の各段落に対し、当該段落に含まれる前記検索条件の単語の出現回数と当該単語の前記重み値を用いて当該検索条件に対する類似度を算出し、当該記事から最初の段落及び少なくとも当該類似度が最も高い段落を抄録として抽出する第２の抄録作成手段と、
前記第３のグループに分類された記事の各段落に対し、当該段落に含まれる前記検索条件の単語の出現回数と当該単語の前記重み値を用いて当該検索条件に対する類似度を算出し、当該記事から当該類似度が高い順に所定数の段落を抄録として抽出する第３の抄録作成手段と、
前記検索手段で検索された各記事のランクと、前記第１乃至第３の抄録作成手段のうちのいずれかで作成された抄録とを配信する配信手段と、
を具備したことを特徴とする情報フィルタリング装置。
【請求項４】
前記配信手段で抄録の配信された記事のうち、前記ユーザが関心ありと回答した記事については、その中に含まれる出現頻度の最も高い単語を前記検索条件に追加する手段をさらに具備したことを特徴とする請求項３記載の情報フィルタリング装置。
【発明の詳細な説明】

【０００１】
【発明の属する技術分野】
この発明は、膨大なテキスト記事からユーザの要求・興味にあったものを検索して定期的にユーザに提供する情報フィルタリング方法および情報フィルタリング装置。
【０００２】
【従来の技術】
近年、ワードプロセッサーや電子計算機の普及、インターネットなどの計算機ネットワークを介した電子メールや電子ニュースの普及に伴い、文書の電子化は加速的に進みつつある。
【０００３】
電子出版という言葉が示すように、今後は新聞、雑誌や本の情報も電子的に提供されることが一般的になると考えられる。これにより、個人にとってリアルタイムで入手可能となるテキスト情報の量は膨大になっていくと予測される。
【０００４】
これに伴い、新聞や雑誌などの膨大なテキスト記事からユーザの要求・興味にあったものを選出して定期的にユーザに提供する情報フィルタリングシステムあるいは情報フィルタリングサービスの需要が高まりつつある。
【０００５】
一方、大量の情報をコンパクトにするという観点から、文書の抄録を作成する方法についても、従来から研究が行なわれてきた。従来は、最初の段落のみを抽出する方法や予め登録しておいたキーワードを用いてキーワードを含む文のみを抽出する方法などがとられてきた。
【０００６】
【発明が解決しようとする課題】
しかしながら、最初の段落のみを抽出する方法では、必ずしもそこにユーザが必要とする情報が含まれているとは限らず、適切な方法ではなかった。また、キーワードを含む文を並べただけでは、それらの文のつながりがわからないという問題があった。
【０００７】
また、一旦登録したキーワードは、ユーザが変更しなければ追加や修正が行なわれず、これをユーザが行なうことは面倒な作業であった。
そこで、本発明は、上記問題点に鑑みてなされたものであり、情報フィルタリングにより検索された記事について、ユーザへの負担がなく、各ユーザの所望するテーマに合致する読みやすい抄録を生成できる情報フィルタリング方法および情報フィルタリング装置を提供することを目的とする。
【０００８】
【課題を解決するための手段】
本発明の情報フィルタリング方法は、複数の情報源からテキストやイメージなどの記事の配信を受け、それら配信された記事の中から、ユーザ毎に予め指定されたテーマに関する記事を検索してユーザに配信する情報フィルタリング方法において、ユーザ毎に予め指定されたテーマに基づく検索情報を基に、前記情報源から配信された記事からユーザが所望する記事を検索し、この検索された記事の抄録を前記検索情報を基に生成し、この生成された抄録をユーザに配信することにより、抄録を生成する際に、ユーザの興味や関心に関する情報を格納したプロファイルを用いるので、ユーザそれぞれが必要とする情報を含んだ抄録を生成することが可能になる。
【０００９】
また、本発明の情報フィルタリング方法は、複数の情報源からテキストやイメージなどの記事の配信を受け、それら配信された記事の中から、ユーザ毎に予め指定されたテーマに関する記事を検索してユーザに配信する情報フィルタリング方法において、ユーザ毎に予め指定されたテーマに基づく検索情報を基に、前記情報源から配信された記事からユーザが所望する記事を検索し、この検索された記事の特徴と前記検索情報を基に前記検索され

た記事の抄録を生成し、この生成された抄録をユーザに配信することにより、記事の特徴に応じて、当該文書に適切な抄録生成方法を用いて当該文書の抄録を生成することができる。

【００１０】

また、本発明の情報フィルタリング装置は、複数の情報源からテキストやイメージなどの記事の配信を受け、それら配信された記事の中から、ユーザ毎に予め指定されたテーマに関する記事を検索してユーザに配信する情報フィルタリング装置において、ユーザ毎に予め指定されたテーマに基づく検索情報を基に、前記情報源から配信された記事からユーザが所望する記事を検索する検索手段と、前記検索情報を基に前記検索手段で検索された記事の抄録を生成する抄録生成手段と、この抄録生成手段で生成された抄録をユーザに配信する配信手段とを具備することにより、抄録を生成する際に、ユーザの興味や関心に関する情報を格納したプロファイルを用いるので、ユーザそれぞれが必要とする情報を含んだ抄録を生成することが可能になる。

【００１１】

さらに、本発明の情報フィルタリング装置は、複数の情報源からテキストやイメージなどの記事の配信を受け、それら配信された記事の中から、ユーザ毎に予め指定されたテーマに関する記事を検索してユーザに配信する情報フィルタリング装置において、ユーザ毎に予め指定されたテーマに基づく検索情報を基に、前記情報源から配信された記事からユーザが所望する記事を検索する検索手段と、この検索手段で検索された記事の特徴に応じて、前記検索情報を基に前記検索手段で検索された記事の抄録を生成する複数の抄録生成手段と、前記検索手段で検索された記事の特徴を検出する特徴検出手段と、この特徴検出手段で検出された記事の特徴に基づき前記複数の抄録生成手段のうちの１つを選択する選択手段と、この選択手段で選択された抄録生成手段で生成された抄録をユーザに配信する配信手段とを具備することにより、記事の特徴に応じて、当該文書に適切な抄録生成方法を用いて当該文書の抄録を生成することができる。

【００１２】

【発明の実施の形態】

本発明の一実施形態について図面を参照して説明する。

まず、図１を参照して、本発明の情報フィルタリング装置を用いた情報フィルタリングシステム全体の構成について説明する。

【００１３】

情報フィルタリングシステムは、新聞社、通信社、または出版社などの複数の情報源２からテキスト記事の配信を受け、それを定期的に加入ユーザ端末３それぞれに送信する情報提供システムであり、このシステムの情報提供サービスは情報フィルタリングセンタ１によって実現されている。

【００１４】

情報フィルタリングセンタ１は、通信網を介して複数の情報源２および複数の加入ユーザ端末３に接続された１つの情報フィルタリング装置から構成され、ここには、情報フィルタリングのための制御や処理を行う中央処理装置４、プログラム並びにデータを格納する半導体メモリ、磁気ディスク、光ディスクなどの記憶装置５、回線や電波などの通信網を介して情報源２からテキスト記事を受信する受信部６、回線や電波などの通信網を介してユーザ端末３にテキスト記事を送信したり、ユーザ端末３からの回答等を受信する送受信部７などから構成されている。

【００１５】

各ユーザ端末３は、例えばパーソナルコンピュータやワークステーションなどの情報処理端末であり、情報フィルタリングセンタ１から送信されたテキスト記事を受信したり、情報フィルタリングセンタ１にテキストデータを送信するテキスト情報送受信部８と、受信したテキスト記事を画面表示する表示部９などを備えている。

【００１６】

情報フィルタリングセンタ１は、図２に示されているように、ユーザプロファイル１０と

称する一種の検索条件をユーザ毎に保持しており、そのユーザプロファイル１０に従って該当するユーザに提供すべき記事を検索する。ユーザプロファイル１０は、ユーザによって指定された複数のテーマ（トピック）などを基に構成されており、それらテーマに合致する記事が検索および選出されてユーザに送られる。

【００１７】

次に、図３を参照して情報フィルタリングセンタ１の構成について説明する。図３において、ユーザプロファイル格納部５０には、記事を検索するための検索条件に対応するプロファイルが格納されている。ここで、プロファイルの記憶形式とその具体例について図４を参照して説明する。図４（ａ）に示すように、プロファイルには、利用者がほしいと思う記事に含まれると考えられる単語とその重みが記述されている。具体的には、図４（ｂ）に示すように、各ユーザ毎に複数の単語とその重みが記憶されている。なお、プロファイル作成の方法は本発明の主眼ではなく、例えば、人が利用者からどのような記事がほしいかを聞き、該当する記事に含まれると思われる単語を列挙する方法でもよい。

【００１８】

記事格納部５１には、情報源２から配信され、受信部６で受信された記事が格納される。ここで、記事格納部５１に格納されている記事データの記憶形式とその具体例について図５を参照して説明する。

【００１９】

図５（ａ）および（ｂ）に示すように、記事データは、センタ１が管理する全ての記事についてその記事を識別するための記事ＩＤ（例えば、「００１」、「００２」、「００３」…）、記事の発行元（例えば、「Ａ新聞社」、「Ｂ新聞社」、「Ｃ出版社」…）、記事の見出し（例えば、「マルチメディアパソコン発売」、「半導体売行き好調」…）、およびその記事の本文の格納位置を示す記事格納部５１に対するポインタ（例えば、１２３４５６、１２３４５７、１２３４５８）から構成されて、記事格納部５１に記憶されている。

【００２０】

記事検索部５２は、ユーザプロファイル格納部５０に記憶されているユーザプロファイルを用いて、記事格納部５１に格納されている記事を検索する。ここでの検索方法としては、例えば「ＳＭＡＲＴ情報検索システム」（ジェラルド・サルトン編、神保健二監訳、企画センター）に記載されている公知の方法を用いることができる。すなわち、この方法（ベクトル空間法）よれば、たとえば、プロファイルに記憶されている各単語のベクトル空間を仮定し、このベクトル空間上において各単語の重みベクトルと検索対象の記事における対応する単語の出現回数ベクトルとの内積を求めることによりプロファイルに対する類似度を算出し、この類似度の大きい順にランク付けされ、検索された記事を並べて出力することが可能である。

【００２１】

図６に記事検索部５２から出力される検索結果の具体例を示す。図６には、あるユーザプロファイルに対して４件の記事が検索された場合について示している。すなわち、ユーザプロファイルのＩＤとともに、算出された類似度に基づく記事のランク（例えば、「１」、「２」、「３」、「４」）と記事ＩＤから構成される検索結果が、検索結果格納部５３に格納される。

【００２２】

従来のフィルタリングシステムでは、検索結果格納部５３に格納された結果と、記事格納部５１に格納された記事データから、最終的に利用者へ送るフィルタリング結果が作成されて、検索結果格納部５３の所定の記憶領域に記憶されるようになっている。

【００２３】

図７は、利用者へ送られる従来のフィルタリング結果の具体例を示したもので、記事のランク、記事の見出し、その記事の本文などから構成されている。

抄録生成部１００は、検索結果格納部５３に格納されたフィルタリング結果について、フィルタリング結果に含まれる記事の抄録を生成するものである。抄録生成部１００は、図

３に示すように、抄録生成制御部１０１、記事特徴検出部１０２、および複数（ここでは、例えば３つ）の生成部１０３－１、１０３－２、１０３－３から構成される。

【００２４】

記事特徴検出部１０２は、記事検索部５２で検索された記事から、例えば、記事の文字数、段落数といった記事の特徴を検出するようになっている。

第１の作成部１０３－１、第２の作成部１０３－２、第３の作成部１０３－３は、それぞれ記事特徴検出部１０２で検出された記事の特徴に適した抄録を生成するようになっている。

【００２５】

抄録生成制御部１０１は、抄録生成部１００全体の制御を司るものである。

次に、図８に示すフローチャートを参照して、抄録生成制御部１０１の処理動作について説明する。

【００２６】

抄録生成制御部１０１は、例えば、記事のランク（ｎ）が高い順に検索結果格納部５３に格納されているフィルタリング結果を取り出し（ステップＳ１～ステップＳ３）、当該のフィルタリング結果の記事ＩＤについて、順に以下の処理を行なう。まず、記事ＩＤを記事特徴検出部１０２へ渡し、記事特徴検出部１０２から選択すべき生成部の番号（この場合、「１」（第１の生成部１０３－１）、「２」（第２の生成部１０３－２）、「３」（第３の生成部１０３－３）のいずれか）を受け取る（ステップＳ４～ステップＳ５）。

【００２７】

次に、当該の番号の生成部へ、記事ＩＤとユーザプロファイルのＩＤを渡す（ステップＳ６）。そして、当該の生成部で生成された抄録を受け取り（ステップＳ７）、当該の記事について、図６に示したような検索結果格納部５３に格納されたフィルタリング結果に含まれる記事のランク、記事ＩＤにより記事格納部５１から取り出した記事見出し、および、当該記事の抄録が検索結果格納部５３の所定の記憶領域に記憶され、最終的に利用者へ配信するフィルタリング結果が生成される。

【００２８】

次に、図９に示すフローチャートを参照して記事特徴検出部１０２の処理動作について説明する。

記事特徴検出部１０２は、抄録生成制御部１０１から渡された記事ＩＤを用いて、記事格納部５１から、記事本文へのポインタをたどって、記事本文を取り出す（ステップＳ１０）。次に、取り出した記事本文の文字数と段落数をカウントする（ステップＳ１１）。段落は、行の先頭が空白文字であることを検出して、段落の始まりであるとすることができる。さらに、予め格納されている抄録生成条件テーブルを参照しながら処理を進める。

【００２９】

図１０に、記事特徴検出部１０２に記憶されている抄録生成条件テーブルの形式（図１０（ａ））とその具体例（図１０（ｂ））を示す。抄録生成条件テーブルは、複数の生成部を識別するＩＤ番号（例えば、「１」、「２」、「３」）と、その生成部にて適用される記事の文字数の最大値（例えば、「４００」、「８００」、「１００００」）および段落数の最大値（例えば、「５」、「１０」、「１００」）が格納されている。

【００３０】

さて、記事特徴検出部１０２は、図１０に示したような抄録生成条件テーブルの最初の行に示されているＩＤ番号「１」の第１の生成部１０３－１の文字数および段落数を取り出し、当該の記事本文の文字数および段落数と比較する（ステップＳ１２）。当該の記事本文の文字数と段落数の両方が、抄録作成部１のそれぞれの値より小さい場合には「１」を返して終了する（ステップＳ１３）。

【００３１】

もし、この条件に合わない場合には、抄録生成条件テーブルの次の行、すなわち、ＩＤ番号「２」の第２の生成部１０３－２の文字数および段落数を取り出して同様の処理を行ない、条件に合う場合には「２」を返して終了する（ステップＳ１４～ステップＳ１５）。

さらに、この条件に合わない場合には第３の生成部１０３－３のＩＤ番号「３」を返して終了する（ステップＳ１６）。

【００３２】

抄録生成制御部１０１は、記事特徴検出部１０２から返された数字に対応して、第１～第３の生成部のうちの１つを選択し、選択した生成部へ記事ＩＤとユーザプロファイルを渡す。

【００３３】

次に、第１～第３の生成部（１０３－１～１０３－３）の処理動作について説明する。

まず、第１の生成部１０３－１の処理動作について説明する。第１の生成部１０３－１は、抄録生成制御部１０１から渡された記事ＩＤを用いて、記事格納部５１から記事の最初の段落を取り出す。文字数や段落数が少ない記事では、第１段落に主な情報が書かれていることが多いので、記事格納部５１から取り出された記事の第１段落を取り出す処理を行なっている。

【００３４】

次に、第２の生成部１０３－２の処理動作について図１１に示すフローチャートを参照して説明する。第２の生成部１０３－２は、まず、抄録生成制御部１０１から渡された記事ＩＤを用いて、記事格納部５１から記事の本文を取り出し格納し（ステップＳ２０）、その本文について各段落の類似度を算出する（ステップＳ２１）。

【００３５】

ここで、段落類似度の算出処理動作について、図１２に示すフローチャートを参照して説明する。

まず、記事格納部５１から取り出された記事本文の第１段落から順に、ひとつずつ段落を取り出し（ステップＳ３０～ステップＳ３３）、ユーザプロファイルのＩＤを用いてユーザプロファイル格納部５０から所望のユーザプロファイルを取り出して、このユーザプロファイルを用いて段落の類似度を算出し（ステップＳ３４）、その結果を蓄えていく（ステップＳ３５）。

【００３６】

段落の類似度の算出は、前述の記事検索部５２が記事全体を対象として行なう類似度の算出処理（ベクトル空間法）を、段落を対象にして行なうことによって実現できる。

【００３７】

全ての段落について類似度を算出してその結果が蓄えられたら（ステップＳ３６、ステップＳ３１）、最後に、蓄えた結果を段落の類似度の順に並べ換える（ステップＳ３２）。なお、この段落類似度の算出結果は、第２の生成部１０３－２に割り当てられた検索結果格納部５３の所定の記憶領域に、例えば、図１３に示すように記憶される。図１３（ａ）に示すように、段落類似度の算出結果は、段落の番号、段落の類似度、段落の内容から構成されていて、段落の類似度が高いものから順に並んでいる（図１３（ｂ））。

【００３８】

図１１の説明にもどり、第２の生成部１０３－２は、図１３に示したような段落類似度算出結果を基に、最も類似度が高かった段落の番号を取り出す（ステップＳ２２）。もし、当該の段落の番号が「１」である場合には、第１段落のみを抄録生成制御部１０１へ返す（ステップＳ２３、ステップＳ２５）。そうでないない場合には、格納しておいた第１段落と、最も類似度が高かった段落の両方を、抄録生成制御部１０１へ返す（ステップＳ２３、ステップＳ２４）。

【００３９】

このように、第２の生成部１０３－２では、主な情報が書かれている第１段落と、利用者の関心が高い（類似度の最も高い）段落の両方を取り出す処理を行なっている。これは、例えば、国や地方自治体の予算に関する記事では、予算全体に関する情報が最初の段落に書かれ、予算の個々の細目については、第２段落以降の段落に分けて書かれるので、ある利用者がマルチメディアに関心がある場合、プロファイルを用いることによって、マルチメディアに関する予算の割当などについて書かれた段落を取り出すことができ、第１段落

とともに利用者へ送られることになる。従って、利用者は、第１段落により、記事の全体の情報を得るとともに、自分の関心がある情報についても知ることができる。

【００４０】

なお、第１段落の他には、最も類似度の高い段落だけでなく、最も類似度の高い段落から複数の段落を選ぶようにすることは容易に実現できる。

次に、第３の生成部１０３－３の処理動作について図１４に示すフローチャートを参照して説明する。

【００４１】

第３の生成部１０３－３は、抄録生成制御部１０１から受け取った記事ＩＤを用いて、記事格納部５１から記事本文を取り出し（ステップＳ４０）、図１２に示したユーザプロファイルを用いた段落類似度算出処理を行って、全ての段落の類似度を算出し、第３の生成部１０３－２に割り当てられた検索結果格納部５３の所定の記憶領域に、例えば、図１３に示すように記憶される（ステップＳ４１）。

【００４２】

その結果をもとに、次に、例えば、段落の総数を３で割った値（整数）の数だけ、類似度の最も大きい段落から順に取り出す。段落の総数を「３」で割った値を用いることにより、約１／３の長さの抄録を生成することができる（ステップＳ４２）。例えば、記事の全体が１４の段落からなる場合には、類似度の最も高い段落から４つの段落が取り出されることになる。このように第３の生成部１０３－３は、雑誌の記事のように、記事全体の長さが長い場合に、利用者が関心のある情報が含まれると考えられる段落を取り出して抄録を作成する処理を行なっている。

【００４３】

なお、段落の総数を割る値は「３」に限るものではなく、他の値でもよい。また、利用者が予め指定した値を設定するようにすることも容易に実現できる。

以上の第１～第３の生成部１０３－１～１０３－３のいづれかの生成部で生成された抄録を抄録生成制御部１０１が受け取り（図８のステップＳ７）、当該の記事についての記事ランク、記事見出し、および、当該記事の抄録等が、検索結果格納部５３の所定の記憶領域に記憶され、最終的に各ユーザへ配信されるフィルタリング結果が生成される。

【００４４】

図１５に抄録生成部１００で生成されたフィルタリング結果の具体例を示す。図１５に示すように、フィルタリング結果は、記事のランク、記事の見出し、記事の全文から一部の段落を取り出して作成した抄録から構成されている。

【００４５】

このような構成で検索結果格納部５３に記憶されたフィルタリング結果は、テキスト情報送受信部７および所定の通信網を介して、例えば、電子メールやＦＡＸで利用者へ送られる。利用者は、送られてきた記事を読み、自分にとって関心がある記事であるかどうかの観点で「○」や「×」を付けて情報フィルタリングセンタ１へ送り返すことができる。すなわち、ユーザ端末から所定の通信網およびテキスト情報送受信部７を介して送り返されてきた回答は、ユーザ回答格納部６０へ格納される。

【００４６】

図１６にユーザ回答格納部６０に格納されたユーザの回答データの具体例を示す。利用者は、関心のあった記事に「○」、関心のなかった記事に「×」、どちらでもない記事はそのままとしている。ユーザの回答は、さらにユーザが必要とする記事を選択できるように、プロファイルを修正するために利用される。

【００４７】

プロファイル修正部６１は、図１６に示したようなユーザの回答と、ユーザ回答中の記事ＩＤを用いて記事格納部５１から取り出した記事本文とから、ユーザプロファイル格納部５０に格納されているユーザプロファイルを修正する。プロファイルを修正する方法は、例えば、「ＳＭＡＲＴ情報検索システム」（ジェラルド・サルトン編、神保健二監訳、企画センター）に記載されている方法が適用できよう。すなわち、この方法は、ユーザが関

心のあった記事の中から単語の出現頻度を計数して、その計数値が最も大きい単語をユーザプロファイルに新たに追加するというものである。

【００４８】

以上説明したように、上記実施形態によれば、ユーザプロファイル格納部５０に格納されている各ユーザ毎のプロファイルは、ユーザの回答に基づいて常に修正・調整され、このプロファイルを用いて記事検索部５２で記事格納庫５１に格納された記事との類似度を算出してユーザが所望する情報を含んだ記事の検索を行い、さらに、この検索された記事について、抄録生成部１００で同じくプロファイルを用いて類似度を算出することにより抄録を生成することにより、ユーザが必要とする情報を含んだ抄録の作成が容易に行える。

【００４９】

また、抄録生成部１００で検索された記事の抄録を生成する際、まず、記事特徴検出部１０２で、その記事の特徴、具体的には、例えば、文字数、段落数を求めて、第１～第３の生成部１０３－１～１０３－３のうちの１つを選択して、その記事の特徴に適した抄録生成方法を選択することにより、ユーザが必要とする情報を含んだ抄録の作成が容易に行える。

【００５０】

さらに、このようにして生成された抄録を各ユーザに配信することにより、情報フィルタリングにより検索されて配信された記事内容から各利用者が情報を得るための時間を短縮し、労力を軽減することができる。

【００５１】

【発明の効果】

以上説明したように、本発明によれば、情報フィルタリングにより検索された記事について、ユーザが必要とする情報を含んだ抄録の作成が容易に行える情報フィルタリング方法および情報フィルタリング装置を提供できる。

【図面の簡単な説明】

【図１】本発明の一実施形態に係る情報フィルタリング装置のシステム全体の構成を示すブロック図。

【図２】図１のシステムの運用形態を概念的に示す図。

【図３】情報フィルタリング装置の構成を示したブロック図。

【図４】ユーザプロファイルの記憶形式とその具体例について説明するための図。

【図５】記事格納部に格納される記事データの記憶形式とその具体例について説明するための図。

【図６】記事検索部から出力され、検索結果格納部に格納される検索結果の記憶形式とその具体例について説明するための図。

【図７】従来のフィルタリング装置の出力形態の具体例を示した図。

【図８】抄録生成制御部の処理動作を説明するためのフローチャート。

【図９】抄録生成部の記事特徴検出部の処理動作を説明するためのフローチャート。

【図１０】記事特徴検出部が参照する抄録生成条件テーブルの記憶形式とその具体例について説明するための図。

【図１１】抄録生成部の第２の生成部の処理動作を説明するためのフローチャート。

【図１２】段落類似度算出処理について説明するためのフローチャート。

【図１３】段落類似度の算出結果の記憶形式とその具体例について説明するための図。

【図１４】抄録生成部の第３の生成部の処理動作について説明するためのフローチャート。

【図１５】抄録生成部で生成された抄録に基づき生成されて検索結果格納部に格納され各ユーザに配信される情報フィルタリング結果の具体例を示した図。

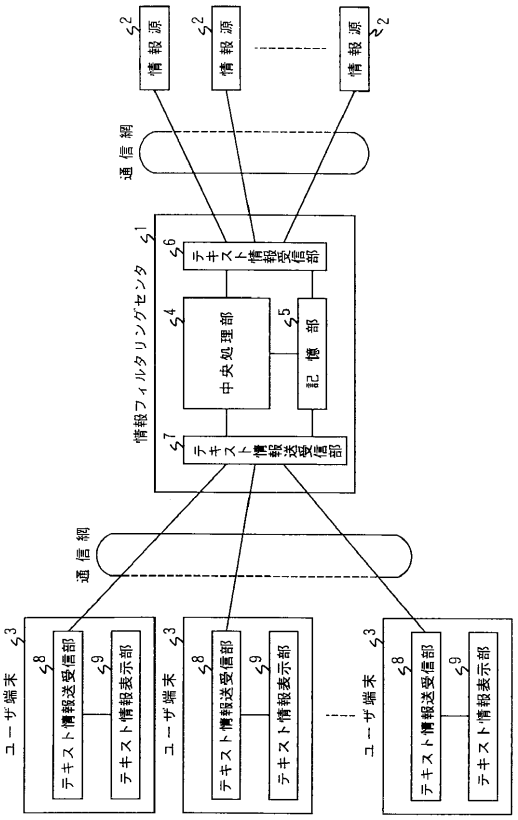
【図１６】ユーザ回答格納部に格納されたユーザの回答データの具体例を示した図。

【符号の説明】

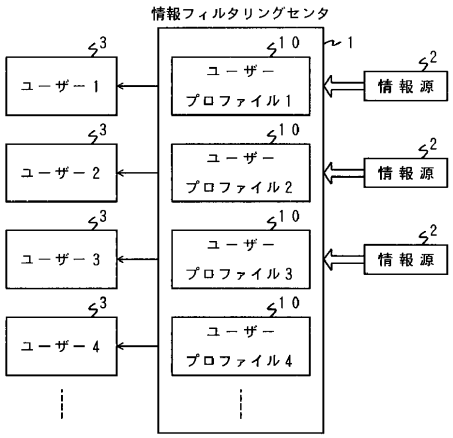
１…情報フィルタリングセンタ、２…情報源、３…ユーザ端末、４…中央処理部、５…記憶部、６…テキスト情報受信部、７…テキスト情報送受信部、５０…ユーザプロファイル

格納部、５１…記事格納部、５２…記事検索部、５３…検索結果格納部、６０…ユーザ回答格納部、６１…プロファイル修正部、１００…抄録生成部、１０１…抄録生成制御部、１０２…記事特徴検出部、１０３－１…第１の生成部、１０３－２…第２の生成部、１０３－３…第３の生成部。

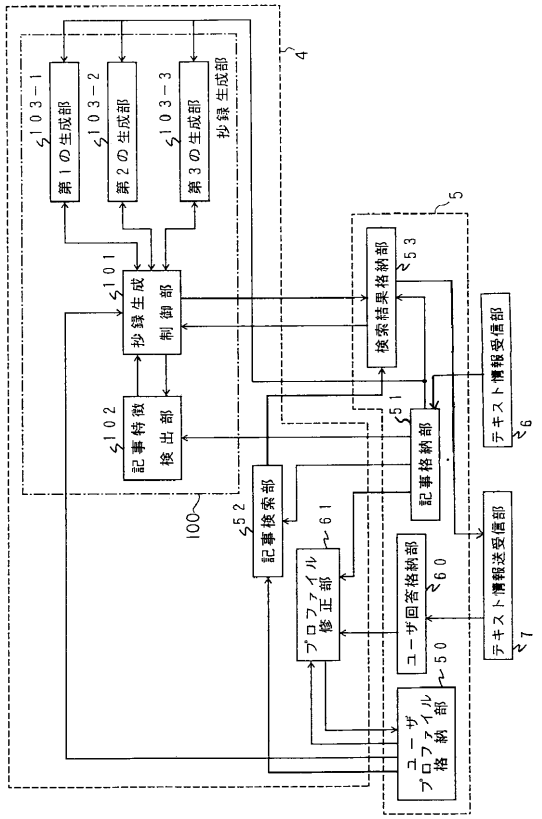
【図１】



【図２】



【図３】



【図４】

(a) ＜形式＞

単語：（重み）

(b) ＜例＞

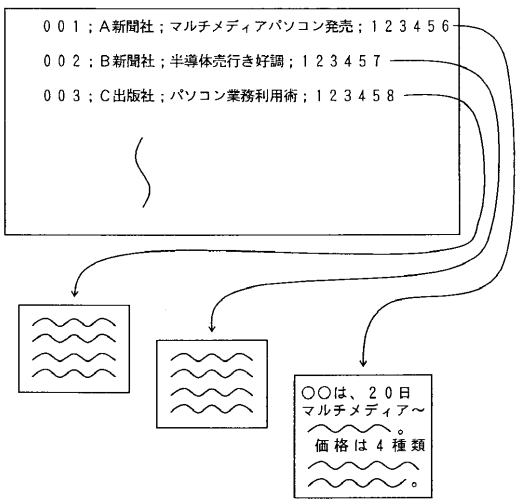
マルチメディア（1）
CD−ROM（2）
パソコン（3）
動　画（1）
音　声（1）
～

【図５】

(a) ＜形式＞

記事ＩＤ；発行元；記事見出し；記事本文へのポインタ

(b) ＜例＞



【図６】

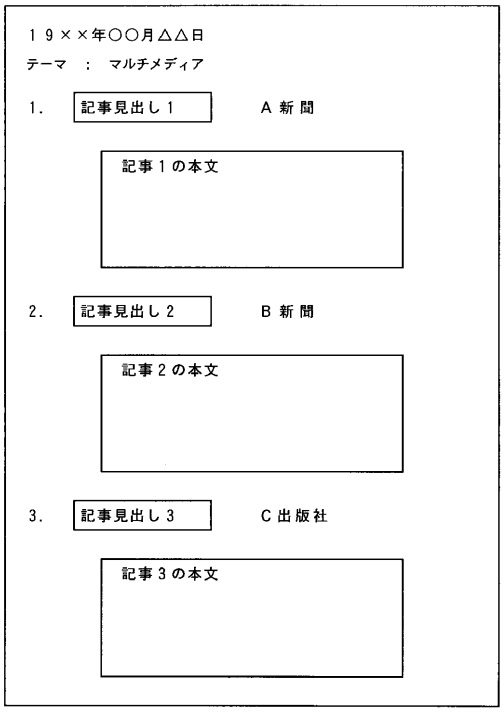
(a) ＜形式＞

記事のランク；記事ＩＤ

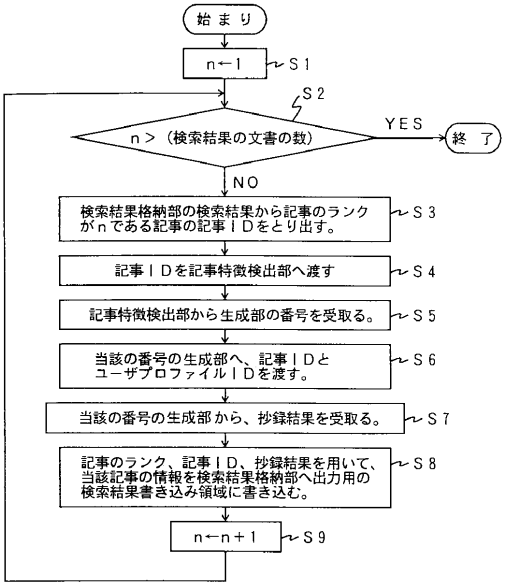
(b) ＜例＞

１　；　００１
２　；　００２
３　；　００３
４　；　００４

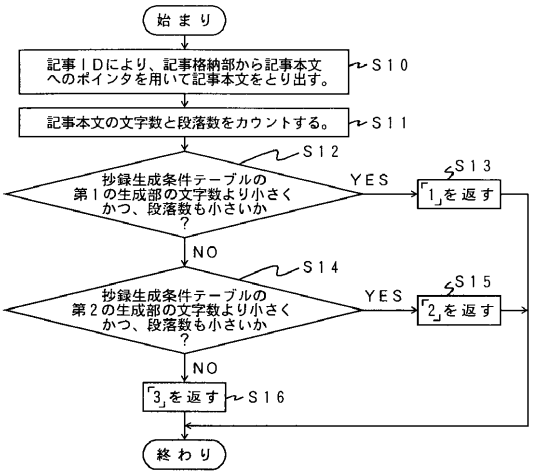
【図７】



【図８】



【図９】



【図１０】

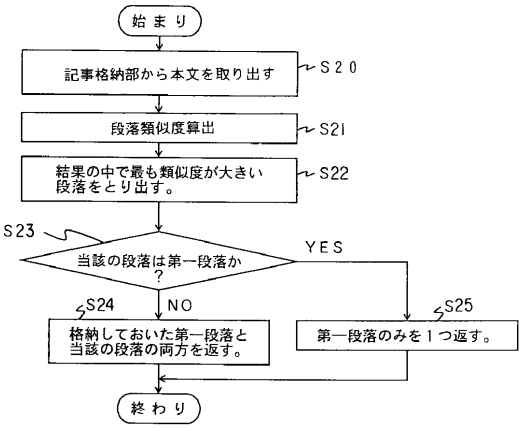
(a) ＜形式＞

抄録作成部のＩＤ；文字数；段落数

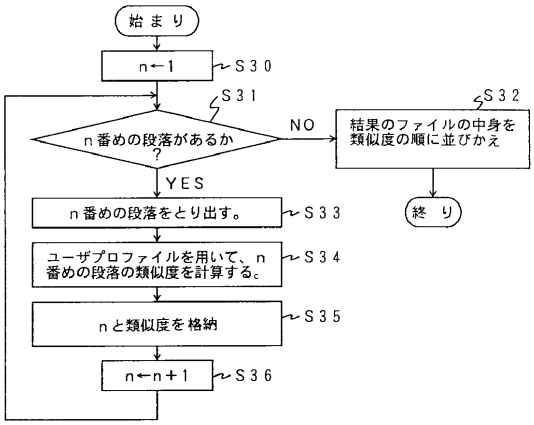
(b) ＜例＞

1	；	400	；	5
2	；	800	；	10
3	；	10000	；	100

【図１１】



【図１２】



【図１３】

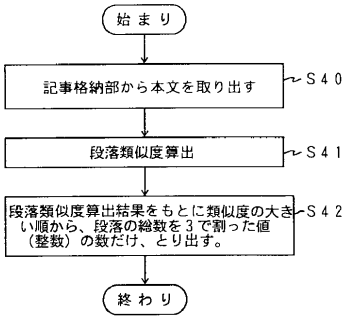
＜形式＞

段落の番号；段落の類似度；段落の内容

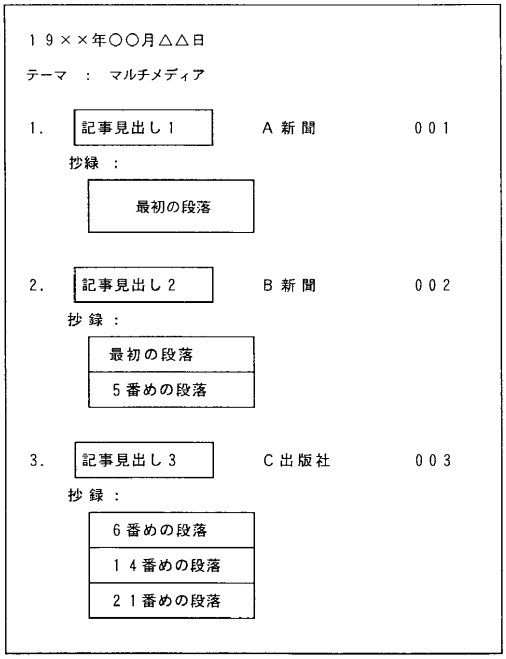
＜例＞

```
4 ; 0.45 ; マルチメディアへの投資は、～。
5 ; 0.32 ; ～。
3 ; 0.29 ; ～。
1 ; 0.13 ; ○○市の予算が決定し、～。
2 ; 0.12 ; ～。
6 ; 0.05 ; ～。
7 ; 0.04 ; ～。
8 ; 0.00 ; ～。
```

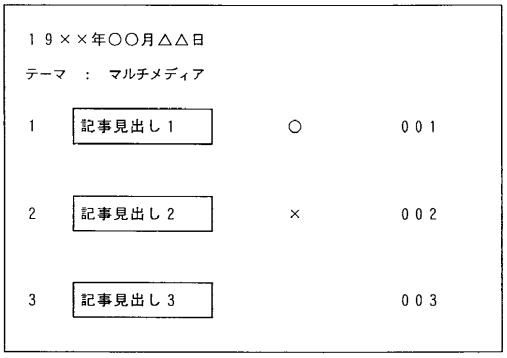
【図１４】



【図１５】


１９××年○○月△△日
テーマ ： マルチメディア
1. 記事見出し１ A新聞 001
抄録 ：
最初の段落
2. 記事見出し２ B新聞 002
抄録 ：
最初の段落
５番めの段落
3. 記事見出し３ C出版社 003
抄録 ：
６番めの段落
１４番めの段落
２１番めの段落

【図１６】


１９××年○○月△△日
テーマ ： マルチメディア
1 記事見出し１ ○ 001
2 記事見出し２ × 002
3 記事見出し３ 003

フロントページの続き

(72)発明者　小野　顕司
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(72)発明者　住田　一男
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

審査官　野崎　大進

(56)参考文献　特開平０４－２８１５５９（ＪＰ，Ａ）
特開平０６－０１２４５４（ＪＰ，Ａ）
特開平０４－２８１５６６（ＪＰ，Ａ）
特開平０６－２１５０４９（ＪＰ，Ａ）
住田一男，自然言語処理に基づいたネットワーク環境下での文書処理，自然言語処理の応用に関するシンポジウム－ネットワーク時代の自然言語処理－論文集，日本，社団法人情報処理学会，１９９５年１０月１９日，第９５巻　第6号，第６５頁～第７２頁

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F 17/30
JSTPlus(JOIS)