(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4243376号
(P4243376)
(45)発行日　平成21年3月25日(2009.3.25)　(24)登録日　平成21年1月9日(2009.1.9)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 210D
G06F 17/30 419A
G06F 17/30 220B

請求項の数 7　(全 18 頁)

(21)出願番号　特願平11-17644
(22)出願日　平成11年1月26日(1999.1.26)
(65)公開番号　特開平11-316768
(43)公開日　平成11年11月16日(1999.11.16)
審査請求日　平成18年1月23日(2006.1.23)
(31)優先権主張番号　09/013,668
(32)優先日　平成10年1月26日(1998.1.26)
(33)優先権主張国　米国(US)

(73)特許権者　596170170
ゼロックス　コーポレイション
ＸＥＲＯＸ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国、コネチカット州　０６８５６、ノーウォーク、ピーオーボックス　４５０５、グローバー・アヴェニュー　４５
(74)代理人　100075258
弁理士　吉田　研二
(74)代理人　100096976
弁理士　石田　純
(72)発明者　クレイグ　デー　シルバースタイン
アメリカ合衆国　カリフォルニア州　スタンフォード　サンタ　フェ　アベニュー　８１７

最終頁に続く

(54)【発明の名称】任意のコーパスサブセットをほぼ一定時間でクラスタ化するための方法および装置

(57)【特許請求の範囲】
【請求項１】
関連する複数のドキュメントがメタドキュメントにまとめられ、このメタドキュメントが階層的に関連づけられて電子的に記憶されるコーパスを処理して、事前に識別された関心に関連するドキュメントをクラスタ化する方法であって、
プロセッサが、
ユーザの入力に応じて、少なくとも一つの初期メタドキュメントであって下位のメタドキュメントを含む初期メタドキュメントを選択する初期選択ステップと、
選択された初期メタドキュメントを階層的関連に基づいて下層側に拡張し、次の階層の複数メタドキュメントからなるフォーカスセットを選択する拡張ステップと、
前記フォーカスセット内のメタドキュメントの良好度を検査して最悪メタドキュメントを選択する最悪選択ステップと、
前記選択された最悪メタドキュメントをその下層側の子孫のメタドキュメントに置き換える置き換えステップと、
関心のあるドキュメントを含まない子孫のメタドキュメントを除去する除去ステップと、
前記フォーカスセットのメタドキュメントの数が所定数になるまで、前記最悪選択ステップと、置き換えステップと、除去ステップを繰り返すステップと、
を実行し、
得られたフォーカスセット内のメタドキュメントをクラスタとして、所定数にクラスタ

化された電子的に記憶されたドキュメントのコーパスを得ることを特徴とする方法。
【請求項２】
請求項１に記載の方法において、さらに、
前記メタドキュメントの要約を確定するステップと、
前記要約をユーザに提示するステップと、
を含むことを特徴とする方法。
【請求項３】
請求項２に記載の方法において、前記要約は、
各メタドキュメントにおいて最も頻繁に現れる固定数の話題のワードと、
各メタドキュメント内の少なくとも一つの典型的なドキュメントの名称と、
を含むことを特徴とする方法。
【請求項４】
請求項１に記載の方法において、前記拡張ステップは、さらに、選択されたメタドキュメント内の関心のあるドキュメントの数がカットオフ値を超えるかを決定するステップを含むことを特徴とする方法。
【請求項５】
請求項４に記載の方法において、前記選択されたメタドキュメント内の関心のあるドキュメントの数がカットオフ値未満である場合、前記ドキュメントは別のドキュメントセットに加えられることを特徴とする方法。
【請求項６】
請求項５に記載の方法において、前記拡張ステップは、さらに、前記別のドキュメントセットを前記次のメタドキュメントに加えるステップを含むことを特徴とする方法。
【請求項７】
関連する複数のドキュメントがメタドキュメントにまとめられ、このメタドキュメントが階層的に関連づけられて電子的に記憶されるコーパスを処理して、事前に識別された関心に関連するドキュメントをクラスタ化する装置であって、
ユーザの入力に応じて、少なくとも一つの初期メタドキュメントであって下位のメタドキュメントを含む初期メタドキュメントを選択する初期選択手段と、
選択された初期メタドキュメントを階層的関連に基づいて下層側に拡張し、次の階層の複数メタドキュメントからなるフォーカスセットを選択する拡張手段と、
前記フォーカスセット内のメタドキュメントの良好度を検査して最悪メタドキュメントを選択する最悪選択手段、
前記選択された最悪メタドキュメントをその下層側の子孫のメタドキュメントに置き換える置き換え手段と、
関心のあるドキュメントを含まない子孫のメタドキュメントを除去する除去手段と、
前記フォーカスセットのメタドキュメントの数が所定数になるまで、前記最悪選択手段と、置き換え手段と、除去手段の処理を繰り返させる繰り返し手段と、
を含み、
得られたフォーカスセット内のメタドキュメントをクラスタとして、所定数にクラスタ化された電子的に記憶されたドキュメントのコーパスを得ることを特徴とする装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、電子ドキュメントをほぼ一定時間でクラスタ化するための方法および装置に関する。特に、本発明は、電子ドキュメントの大きなコーパス（記録されたテキストの集合）をほとんど一定時間で、それに比べて非常に小さなクラスタのセットに分割することを目的とする。
【０００２】
【従来の技術】
ドキュメントブラウジングは、大きなテキストコレクションにアクセスするために使用さ

れる有力なツールである。ブラウジングは、クエリー（質問）がないため検索と識別され、余りにも一般的であるかまたは余りにも漠然としているためのいずれかによって、幾つかの検索言語によって有効に表現できない情報ニーズに対して、都合よく作動する。たとえば、ユーザが、関心のある話題を適切な言語で記述することに不慣れである場合、またはワードの特定の選択にコミットする（明言する）ことを希望しない場合がある。実際、ユーザは、特定のものは何も要求しないが、コレクションの一般的な（全般的な）情報コンテンツ（内容）を探求したい場合がある。

【０００３】

この状況において、情報アクセスシステムは有用である。情報アクセスシステムは、コレクションのコンテンツを提示すること、および、ユーザがコンテンツの幾つかの話題に固有のサブセットに関心を集中することの両者をナビゲーションできるコレクションのアウトラインを含む。このようなブラウジングシステムは、Ｐｅｄｅｒｓｅｎらによる米国特許第５，４４２，７７８号（分散／集合：Scatter/Gather）および米国特許第５，４８３，６５０号に開示されており、これらの各特許は引用により本願に援用する。

【０００４】

分散／集合においては、注意は、常に、ドキュメントのフォーカスセット、特にユーザの関心を引く可能性のあるサブセットに向けられる。最初に、フォーカスセットは、ドキュメントコレクション全体であるかもしれない。フォーカスセット内のドキュメントは、ドキュメントの少数の話題固有の（トピックに密着したtopic-coherent）サブセット、またはドキュメントのクラスタにクラスタ化される。用語「クラスタ化（clustering）」および「分散（scattering）」は、同義語として使用される。したがって、フォーカスセット内のドキュメントは、クラスタに分散される、ということができる。

【０００５】

分散／集合においては、クラスタ要約（サマリー）が作成（develop）され、ユーザに提示される。クラスタ要約は、通常、フォーカスセットのドキュメントの輪郭（アウトライン）を示すコンテンツの表（テーブル）である。クラスタ要約は、各クラスタ内のドキュメントから自動的に決定される示唆に富むテキストを含む。各クラスタ要約は、２種類の情報を含む。すなわち、クラスタのドキュメントに最も頻繁に出現する話題の（topical）ワード、およびクラスタ内の数個の典型的なドキュメントの名称である。要約は、クラスタプロファイルに基づく。このクラスタプロファイルはクラスタのドキュメントに出現するワードを反映する。

【０００６】

そこで、ユーザは、最も関心を持たせるように見える複数のクラスタを識別し、選択する。選択されたクラスタは、一緒に集合され、新しい小さいフォーカスセットを形成する。すなわち、新しいフォーカスセットは、選択されたクラスタ内のドキュメントの合併（ユニオン：union）である。ユーザがドキュメントに個別にアクセスすること、またはクエリーに基づく探索方法を使用することを求めるまで、この処理が所望の回数反復される。

【０００７】

分散／集合は、必ずしも独立型情報アクセスツールである必要はない。むしろ、分散／集合は、論理探索または類似性探索などの探索方法と連携して使用することができる。類似する例は、リファレンスブック（reference book）であり、リファレンスブックは二つのアクセス方法を提供する。一つは、ブラウジングのための前にある目次（a table of contents）であり、他は、さらに直接的な探索のための最後にある索引（index）である。分類／集合は、必ずしも特定のドキュメントを探索するために使用されるとは限らない。むしろ、クラスタ要約に存在する用語（ボキャブラリー）を公開することによって、分散／集合は、相補的な探索方法を補助する。たとえば、クラスタプロファイルを、類似性探索においてコレクション全体に対するクエリーとして使用することができる。逆に言えば、分散／集合を使用し、過剰の多数のドキュメントを検索するワードを基礎とするクエリーの結果を編成することができる。

【０００８】

図９は、１９９０年８月のＮｅｗ　Ｙｏｒｋ　Ｔｉｍｅｓ　Ｎｅｗｓ　Ｓｅｒｖｉｃｅに掲載された約５，０００記事のテキストコレクションに適用される分散／集合方法を示す図である。図９においては、分散／集合方法を一層簡単に提示するために、実際のクラスタ要約の代わりに単独のワード文字（ラベル）が示されている。

【０００９】

図９に示す例においては、ユーザの情報ニーズは、１９９０年に発生したことを一般的に決定することである。特定の話題記述は全く存在しないため、この情報ニーズを効果的に表現するワードに基づくクエリーを構成することは困難である。ユーザは、一般的な話題、たとえば、「国際的事件」を考えるが、この話題記述は、国際事件に関する記事は、通常、これらのワードを使用しないため、有効ではない。

【００１０】

分散／集合によって、ある用語を提供することが強制されるのではなく、ユーザは、クラスタ要約のセット、すなわちコレクションの輪郭を提供される。ユーザニーズは、関心のある話題に関連する可能性があると考えられるクラスタを選択する。図９に示す分散／集合処理においては、その月の主要な新聞記事は、最初の分散からすぐに明らかになり、イラクのクエート侵入およびドイツ再統合問題である。これによって、ユーザは、国際問題に焦点を絞るようになり、「イラク」、「ドイツ」、および「石油」クラスタを選択する。これらの３クラスタは一緒に集合され、より小さなフォーカスセットを形成する。

【００１１】

次に、この比較的小さなフォーカスセットは、クラスタ化、すなわち分散され、減少されたコレクションを包含する８個の新しいクラスタを形成する。減少されたコレクションは、記事のサブセットのみを含むので、これらの新クラスタによって、元の８個のクラスタより細かいレベルの詳細が明らかになる。イラク侵入に関する記事および石油記事の一部は、米国軍展開、石油市場に対するイラク侵入の影響、およびクエートにおける人質についてのクラスタに分離される。

【００１２】

ユーザが、これらの主要な新聞記事を適切に理解するが、世界の他の部分で何が発生したかを見つけたいと思う場合、ユーザは、たとえば、「パキスタン」クラスタを選択、－このクラスタも他の外国政治新聞記事を含む－およびアフリカに関する記事を含むクラスタを選択することができる。これらのクラスタを分散することによって、多数の特定の国際状況ならびに多方面にわたる国際記事の小さなコレクションが明らかになる。このようにして、ユーザは、パキスタンにおける政変およびトリニダードで発生した人質について知る。これらの記事は、別の状態では、その月の一層重要な記事の中に埋没してしまうものである。

【００１３】

図１０は、分散／集合の操作を示す図である。図１０に示す例においては、テキストコレクション（またはフォーカスセット）２０は、グロリエ（Ｇｒｏｌｉｅｒ）の百科事典のオンライン版である。フォーカスセット内の２，７００，０００記事は、それぞれ、独立のドキュメントとして処理される。図１０に示す例においては、ユーザは、宇宙開発における女性の役割を調査することに関心がある。この情報ニーズを正式の（formal）クーリエによって表現しようとするのでなく、ユーザは、代わりに、クラスタの記述から、関心のある話題に関連すると考えられる多数のトップレベルのクラスタ２２Ａ～２２Ｉを提供される。次に、ユーザは、軍事経過（ヒストリー）クラスタ２２Ａ、科学および産業検出子２２Ｃおよびアメリカ社会クラスタ２２Ｈを選択し、グロリエの事典から得られる記事の指示されたサブセットの減少されたコーパス（またはフォーカスセット）２４を形成する。

【００１４】

次に、減少されたコーパスは、浮動によって（on the fly）もう一度クラスタ化され、減少されたコーパス２４を対象にする新しいクラスタのセット２６Ａ～２６Ｊを生成する。減少されたコーパスはグロリエの事典の記事のサブセットを含むため、これらの新クラス

タは、トップレベルクラスタ２２Ａ～２２Ｉより細かいレベルの詳細である。ユーザは、再度、関心のあるクラスタを選択する。この場合、選択されたクラスタは、軍用機クラスタ２６Ｅ、工業技術クラスタ２６Ｇ、および物理クラスタ２６Ｈである。再度、さらに減少されたコーパス２８が形成され、再クラスタ化される。最終セットのクラスタ３０Ａ～３０Ｆは、軍用機クラスタ３０Ａ、アポロ計画クラスタ３０Ｂ、航空宇宙産業クラスタ３０Ｃ、天候クラスタ３０Ｄ、天文学クラスタ３０Ｅ、および民間航空機クラスタ３０Ｆを含む。この段階において、クラスタは、十分に小さく、記事名称の網羅的なリストを通じて直接に精読することができる。関心のある少なくとも一つの記事が見出されると仮定すると、ユーザは、同じくラスタ内に類似の特性の記事をさらに見出すこと、またはことによると探し当てた記事またはクラスタ記述の用語集（ボキャブラリー）に基づいて方向を持った（directed）探索方法を使用して追加記事を見出すことができる。

【００１５】

【発明が解決しようとする課題】

ドキュメントクラスタ化に関する以前の成果は、線形時間(linear-time)法、たとえば、分散／集合および米国特許第５，４８３，６５０号に記載の線形時間法を含み、この方法によれば、クラスタ化のために要する時間は僅か数分に減少される。これは、広範囲のワードに基づくクエリーを使用し、中程度の大きさのコレクションを探索するのに十分な速さである。たとえば、毎秒およそ３０００ドキュメントの速度を、分散／集合を使用し、サンマイクロシステムズ（ＳｕｎＭｉｃｒｏｓｙｓｙｔｅｍｓ）のＳＰＡＲＣＳＴＡＴＩＯＮ２上において、実現できる。しかし、線形時間クラスタ化でさえも、非常に大きなドキュメントコレクションの対話型ブラウジングを支援するためには遅すぎる。このことは、約７５０，０００のドキュメントを含むテキスト検索評価のために、分散／集合をＴＩＰＳＴＥＲコレクション、ＤＡＲＰＡ標準に適用する場合を考慮すれば、特に、明らかである。毎秒３０００ドキュメントの速度において、これは、分散するために４時間以上を必要とし、対話型にとっては長すぎると考えられる。したがって、ドキュメントをクラスタ化するために、一層迅速なさらに効率的な方法を見出すことが必要とされる。

【００１６】

本発明は、顧客対応可能な時間／精度トレードオフを持つコーパスサブセットをほぼ一定時間でクラスタ化するための方法および装置を提供するものである。

【００１７】

本発明は、基礎的なブラウジング方法、たとえば、分散／集合に使用することが可能であり、大きなドキュメントコレクションを関連のあるドキュメントのクラスタに効率的に分割するほぼ一定時間でクラスタ化するための方法も提供するものである。

【００１８】

【課題を解決するための手段】

本発明による再クラスタ化方法および装置においては、入力は、全体のドキュメントの複数のメタドキュメントへのクラスタ化であり、複数のメタドキュメントから「最悪」メタドキュメントが選択される。「最悪」メタドキュメントは、その子のメタドキュメントによって置換され、関心のあるドキュメントを含まないこれらの子は除去される（pruned）。次に、残りのメタドキュメントは一緒に集合され再クラスタ化される。ユーザが所望の程度の特定性を得るまで、この処理が反復される。

【００１９】

このクラスタ化方法は従来の方法より速く、この方法においては、クラスタは本来の資質（in their own right）でドキュメントとして処理され、既存の階層（hierarchy）を使用しクラスタの新しいセットを生成する。すなわち、本発明による再クラスタ化方法および装置においては、クラスタは、大きな個別ドキュメントであるかのように、クラスタ化する必要があるメタドキュメントして処理され、クラスタ化される。したがって、ファンアウトｋを有するクラスタ階層の場合、本発明による再クラスタ化方法および装置は、最小のクラスタから開始し、各クラスタをそのｋ個の子によって置換する。親クラスタは検査され、最悪クラスタが除去される。すなわち、「最悪」親クラスタは、そのｋ個の子に

よって置換される。
【００２０】
本発明のこれらおよび他の特徴および利益は、以下の好適な実施形態に関する詳細記述に記載され、明らかとなる。
【００２１】
以下、本発明を添付図面を参照して詳細に述べる。図面において、同じ符号は、同じ構成要素を示す。
【００２２】
【発明の実施の形態】
図１は、本発明による再クラスタ化システム１０の一実施形態を示すブロック図である。システム１０は、プロセッサ１１、ＲＯＭ１２、ＲＡＭ１３、不揮発性メモリ１４、コーパス入力１５、ユーザ入力装置１６、ディスプレイ装置１７、および出力装置１８を備える。
【００２３】
ブラウジング手順を実行する前に、ドキュメントコーパスがコーパス入力１５から入力される。次に、ドキュメントコーパスは、プロセッサ１１によって分割される。分割手順の結果は、ディスプレイ装置１７に表示される。操作者は、ユーザ入力装置１６、たとえば、マウス、キーボード、タッチスクリーン、スタイラス、またはこれらの要素の組合せなどを使用し、コマンドおよびデータを入力することができる。ユーザは、ドキュメントのハードコピーのみでなくクラスタダイジェスト要約（サマリー）のプリント出力も出力装置１８、たとえばプリンタに出力することができる。
【００２４】
従来は、プロセッサ１１によって、ドキュメントの初期順序付け（initial ordering）が準備される。初期順序付けは、たとえば、分散／集合に記載の分別法を使用して準備される。プロセッサ１１によって、コーパスの最初の順序付けの要約も決定され、この要約はディスプレイ装置１７に表示、または出力装置１８によってユーザに出力することができる。この要約は、たとえば、分散／集合に記載されているクラスタダイジェスト法を使用し、決定することができる。
【００２５】
ユーザからユーザ入力装置１６を経由して適切な命令を受領後、プロセッサ１１は、コーパスのさらに進んだ順序付けを実行することができる。このさらに進んだ順序付けは、たとえば、分散／集合に記載されているバックショット(buckshot：大きめの散弾)法を使用し、形成される。次に、このステップの所望の数の反復が実行され、コーパスがさらに狭くされる。結局、個別のドキュメントが検査され、または幾つかの有向探索ツールが限定コーパスに適用される場合がある。
【００２６】
図２は、本発明による再クラスタ化の一実施形態の輪郭（アウトライン）を示す図である。処理は、ステップＳ１００において開始され、ステップＳ２００に続く。ステップＳ２００において、ユーザは、全ドキュメントコレクションの一部またはコーパスの一部を表現するドキュメントセットを選択する。後のステップにおける反復のために、フォーカスセットはメタドキュメントを含み、メタドキュメントは、それぞれ、コレクションの一部のみを表現する（代表する）。メタドキュメントセット中のメタドキュメントの数は、ほぼ所定の最大数に等しく、最大数は、たとえば、５００または１０００とすることができる。次に、ステップＳ３００において、最初のメタドキュメントセットは、プロセッサ１１によって選択され、クラスタ化される。好適には、メタドキュメントクラスタの所定数は、１０である。一般に、必要とされることは、新メタドキュメントの所定数は、その後のメタドキュメントの所定最大数より小さいことが必要であるということのみである。メタドキュメントを選択し、クラスタ化する処理は、図３および図４に関連して、以下に述べる。次に、制御はステップＳ４００に続く。
【００２７】

ステップＳ４００において、新メタドキュメントは、要約されて利用できる形式になる。次に、ステップＳ５００において、たとえば、ディスプレイ装置１７または出力装置１８を使用し、ユーザに提示される。次に、処理は、ステップＳ６００に続き、ステップＳ６００において、処理は停止する。

【００２８】

メタドキュメントセットはクラスタ階層Ｈを有し、クラスタ階層Ｈは、ｋ個の子のファンアウト（fan-out）およびルートノードｒを有する。階層は、クラスタのツリー構造であり、クラスタはノードと呼ばれ、ノード１のｋ個の子の合併はノード１自体と同じドキュメントを有するように、ノードはメタドキュメントを表現する。ドキュメントのセットＳは、クラスタ化ルーチンに入力される。この処理の結果、ｋ個のクラスタのセットとなり、このクラスタはＳ中のドキュメントを正確に含む。

【００２９】

図３は、図２のメタドキュメント選択およびクラスタ化ステップＳ３００の第１実施形態のさらに詳細な輪郭を示す図である。ステップＳ３００から始まり、制御はステップＳ３２０に進む。ステップＳ３２０において、収集する必要があるノードの最大数Ｍが設定される。次に、ステップＳ３３０において、初期フォーカスセットＴが、階層Ｈのルートノードｒとして設定される。次に、ルートノードは、そのｋ個の子によって直ちに置換される。次に、制御はステップＳ３４０に進む。

【００３０】

ステップＳ３４０～Ｓ３６０において、ある方法において「良好」である関心のあるノードがクラスタ階層中に見出される。ノードの良好度を決定する方法について、以下に詳細に述べる。

【００３１】

ステップＳ３４０において、フォーカスセットＴのｋ個のノードは検査され、「最悪」ノードがピックされる。「最悪」ノードは、以下の述べる「良好度」検査によって決定される。次に、ステップＳ３５０において、「最悪」ノードは除去され、そのｋ個の子のノードによって置換され、子は関心のあるドキュメントを含む。関心のあるドキュメントを含まない子は含まれず、効果的に除去される(pruned)。

【００３２】

次に、ステップＳ３６０において、制御ルーチンは、フォーカスセットＴが収集する必要があるノードの最大数Ｍに等しいノード数またはそれより大きいノード数を有するかを決定する。フォーカスセットＴのノード数が収集する必要がある最大ノード数Ｍ未満である場合、制御はステップＳ３４０に跳び戻る。そうではなく、フォーカスセットのノード数が少なくともＭに等しい場合、制御は、ステップＳ３７０に続く。

【００３３】

ステップＳ３７０において、フォーカスセットＴはクラスタ化され、クラスタＰのセットが得られる。次に、ステップＳ３８０において、クラスタＰのこのセットの各ノードは、クラスタ内の、Ｓにおいては存在しなかったドキュメントを削除するために、関心のあるドキュメントI_S（ｎ）によって置換される。次に、制御はステップＳ３９０に続き、ここで制御はステップＳ４００に戻る。

【００３４】

前述したクラスタ化ステップにおいて、見出されたＭ個のノードは、線形時間クラスタ化方法を使用し、クラスタ化される。選択されるノード数が限定される限り、これによって、一定時間（constant-time）のクラスタ化が与えられる。

【００３５】

クラスタ階層のノードの数は大きい場合があるため、すべてのノードを検査して「良好」ノードを見出すことはできない。その代わり、クラスタ階層は、トップからファンアウトする。階層Ｈのルートノードから始まり、ルートノードは、直ちにその子によって置換される。得られるセットのｋ個のノードは検査され、「最悪」ノードがピックされる。「最悪」ノードは除去され、そのｋ個の子によって置換される。この処理は、今、検討中の２

ｋ－１のノードについて反復される。実際は、すべてのｋ個の子ノードは、必ずしも含まれない。むしろ、子ノードのサブセットのみが、検討される。Ｍ個のノードが収集されると、処理は停止される。

【００３６】

この時点において、共通集合（積集合）テーブルI_Sが生成される。任意のノードｎに対して、そのノードの共通集合I_S（ｎ）は、Ｓ∩ｎにおけるドキュメントのセットである。すなわち、I_S（ｎ）は、ドキュメントセットＳとノードｎに含まれるドキュメント間の共通集合である。したがって、共通集合テーブルI_Sによって、ドキュメントセットＳおよびノードｎの両者に含まれる関心のあるドキュメントのみが、提供される。I_Sは、｜Ｓ｜ｌｏｇ（ｎ）時間内に作成される。共通集合テーブルI_Sを使用し、結果として得られる各ノードが、共通集合I_S（ｎ）によって置換され、ドキュメントセットＳに存在しない、クラスタ中のドキュメントが削除される。得られるノードは、クラスタ化され、ｋ個のクラスタとなり、各ノードはなお単独の実体（エンティティ：entity）として処理される。

【００３７】

任意のノードｎに対するＳおよびｎの共通集合を求めるために、ドキュメントを処理し、ドキュメントを含む階層Ｈのすべてのノードを戻す関数が使用される。この関数は、ドキュメントセットＳに従属せず、階層Ｈが決定されると同時に決定されることができる。階層Ｈは、一定のｋのファンアウトを有するので、階層Ｈは深度ｌｏｇｎを有し、したがって、各ドキュメントは、ｌｏｇｎノードにある。

【００３８】

I_Sを求めるために、テーブルが構成され、ノードによって索引される。テーブルの各項目は、原始状態においては、空である。ドキュメントセットＳの各ドキュメントに対して、事前に計算された前述した関数を使用し、どのノードがドキュメントを包含するかを見出す。次に、ドキュメントセットをこのような各ノードに対するテーブル項目に追加する。理論上は、一定時間内に任意のサイズの空テーブルを構成することが可能であるが、実際には、明白な線形時間（リニアー・タイム）アルゴリズムは極めて迅速である。テーブル更新は、ドキュメント当たり時間ｌｏｇｎ、または全体で時間｜Ｓ｜ｌｏｇｎを要する。得られるテーブルは、I_S、すなわち必要とされる共通集合計算ツールである。

【００３９】

図４は、図２のメタドキュメント選択およびクラスタ化ステップＳ３００の第２実施形態のさらに詳細な輪郭を示す図であり、如何にして、追加されるカットオフ値を有する任意のデータセットに対するノードが決定されるかを示す。ステップＳ３００において開始され、制御はステップＳ１３０５に続く。次に、ステップＳ１３０５において、カットオフ値が、ｃに設定され、その結果、ｃ未満のドキュメントを含むノードは単独のドキュメントノードによって置換されることができる。再度、収集する必要があるノードの最大数Ｍも、設定される。次に、ステップＳ１３１５において、初期フォーカスセットＴが、階層Ｈのルートノードとして設定される。次に、制御はステップＳ１３２０に続く。

【００４０】

ステップＳ１３２０において、小さいドキュメントセットＥは、ゼロに設定される。次に、ステップ１３２５において、フォーカスセットのｋノードが検査され、「最悪」ノードがピックされる。次に、ステップＳ１３３０において、「最悪」ノードは、照合され、そのノードが、カットオフ値ｃ未満の数のドキュメントを含むかまたはそれに等しい数のドキュメントを含むかが、決定される。ノードのドキュメントの数がカットオフ値ｃ未満である場合、制御は、ステップＳ１３３５に続く。そうではなく、選択されるノードがカットオフ値ｃ未満の数のドキュメントを含まない場合、制御は、ステップＳ１３４０に跳ぶ。

【００４１】

ステップ１３３５において、ノード内のその数のドキュメントが、小さなドキュメントセットＥに加えられる。次に、制御は、ステップＳ１３４５に跳ぶ。ステップＳ１３４０に

おいて、関心のあるドキュメントを含むノードの子がフォーカスセットＴに加えられる。関心のあるドキュメントを含まない子は包含されず、効果的に、「除去される：pruned」。次に、制御は、ステップＳ１３４５に続く。

【００４２】

ステップＳ１３４５において、フォーカスセットは、照合され、フォーカスセットが収集する必要がある最大数Ｍに達しているかが決定される。収集する必要があるノードの最大数Ｍに達している場合、制御は、ステップＡ１３５０に続く。その他の場合は、制御は、ステップＳ１３２５に跳び戻り、次の最悪ノードを見出す。

【００４３】

次に、ステップＳ１３５０において、小さいドキュメントセットＥがフォーカスセットＴに加えられる。次に、ステップＳ１３５５において、フォーカスセットＴは、クラスタ化され、クラスタＰのセットが得られる。次に、ステップＳ１３６０において、各ノードＰは、関心のあるドキュメントI_S（ｎ）によって、置換される。次に、制御は、ステップＳ１３６５に続く。ステップＳ１３６５において、制御は、図２のステップＳ４００に戻る。

【００４４】

このように、追加されるカットオフ値を有する任意のデータセットに対して、ノードがドキュメントセットＳから得られる数個のドキュメントのみを含む場合、これらのドキュメントは、ノードを拡張する時間を消費する代わりに別のセットＥに追加される。

【００４５】

ノードをその子によって置換する場合、「空」の子、すなわちドキュメントセットＳにいかなるドキュメントも含まない子は、明白に回避することができる。「単集合（シングルトン：Singleton）」子、すなわち、ドキュメントＳから得られる一つのドキュメントのみしか包含しない子も、特別に取り扱うことができる。ノード内に一つのドキュメントしか存在しない場合は、ノード全体が包含されない。ドキュメントが簡単に取り出され、それ自体がノードとして処理される。これは、適切な終端子孫（リーフディセンデント：leaf descendent）によって子ノードを置換することと等価である。一般に、カットオフ値ｃ未満のドキュメントを包含するノードは、ｃ個の単独ドキュメントノードによって置換することができる。一定の数のノードのみが検査されるため、この方法によって生成される新しいノードの数も一定である。

【００４６】

如何にして多数のノードが拡張されるかにｃの値が影響を及ぼすようにすることは望ましくないので、単独のドキュメントノードは、通常のノードと別に数えられる。すなわち、単独のドキュメントノードをフォーカスセットＴ内に保持するのではなく、単独のドキュメントノードは、別のセットＥに移動される。この処理は、フォーカスセットＴが所定のサイズに達するまで続く。｜Ｅ｜は、定数によって限定されるので、この値は実行時間の解析に影響を及ぼさない。

【００４７】

たとえば、図３のステップＳ３４０および図４のステップＳ１３２５において、「最悪」ノードを決定するために使用される幾つかの「良好度」検査がある。使用することができる一つの「良好度」検査は、適合度検査または割合（ＲＡＴＩＯ）検査である。ノードが包含する大部分のドキュメントもドキュメントセットＳから得られる関心のあるドキュメントである場合、ノードは、「良好」である。

【００４８】

たとえば、ｎは、ｄドキュメントを有する場合、ｎの良好度は、下式によって表される。

【００４９】

【数１】

$g = |I_S(n)| / d$

関数ｆ（Ｓ，Ｔ）によって、フォーカスセットＴ内の最低の良好度を有するノードは返される。この関数は、僅かしか一致しないノード、すなわち一致しない子を有する可能性の

あるノードに有利であるので、この良好度検査は、結果として、広い範囲の除去（pruned）となり、結果が改善される。他方、かなり良好な割合を有する大きなノードは、絶対値の項に多数の非一致ドキュメントを含む場合でも、フォーカスセットＴ内にそのままで留まる。

【００５０】

一つの大きなノードが、ドキュメントセットＳ内に多数のドキュメントを包含する場合、割合検査は、このノードに有利である。このことはクラスタ化の場合に問題となる場合があり、その理由は、クラスタ化方法は、ノード内のドキュメントすべてを単独の実体として処理し、不均衡なクラスタサイズとなる可能性があるためである。このような大きなノードの拡張は、良好度値に重みを付けることによって促進される。たとえば、ノードｎは、ｄドキュメントを有する場合、ノードｎの加重良好度ｇ′は、下式によって表される。

【００５１】

【数２】

$$g' = \sqrt{|I_S(n)|} / d$$

この場合、ドキュメントセットＳ内に多数のドキュメントを有することは、良好な割合の保証にはならない。実際に、比較的少数のドキュメントｄを有することが、一層有利である。このことによって、出力ノードは、すべて、ドキュメントセットＳから得られるほぼ等しい数のドキュメントを有することを保証することが容易になる。

【００５２】

良好度を決定する他の手法は、情報理論による測定を使用する。ノードの子がノード自体より多くのドキュメントセットＳに関する情報をコード化する場合、そのノードは、その子によって置換される良い候補である。このことは、親における一致は、子の間に不均一に分散され、その結果、劣悪な子は除去され、良好な子が維持されることを暗に示す。

【００５３】

たとえば、ノードｎが、サイズｄを有する場合、ノードn_iは、ノードｎの子であり、サイズd_iを有する。ノードｎ内の情報Ｉ（ｎ）は、下式で表される。

【００５４】

【数３】

$$I(n) = -(|I_s(n)|/d) \cdot \log_2(|I_s(n)|/d)$$

ノードｎに対する情報ゲインＧ（ｎ）は、下式で表される。

【００５５】

【数４】

$$G(n) = I(n) - \Sigma\{(|d_i|/|d|) \cdot I(n_i)\}$$

ここで、Σは、ｉについてのサメンションである。

【００５６】

ノードｎに対する適切な良好度測定は、Ｇ（ｎ）によって与えられる。関数ｆ（Ｓ，Ｔ）によって、フォーカスセットＴ内の最高の情報ゲインを有するノードは返される。このことは、その子によって置換されることにより最も利益が得られるノードがピックされるという利点を有する。不都合なことに、これらの一致が子の間に均一に分散される場合、このことは、僅かな一致しか有しない大きなノードを無視することになる。

【００５７】

本発明においては、非所定数の個別ドキュメントの代わりに、所定数のメタドキュメントが、クラスタ化または分散のための手順において使用される。メタドキュメントは、ツリー、たとえば、図５から図８までのツリーなどのメタドキュメントから得られる降順の複数の個別ドキュメントを表現する。

【００５８】

図５から図８までの以下の討議の場合、本発明に従って、たとえば前述した割合検査などの幾つかの「良好度」検査の一つを使用し、「最悪」メタドキュメントを選択することができる。しかし、討議を容易にするために、図５から図８までにおいて、「最悪」メタドキュメントは、最低数の関心のあるドキュメントを有するメタドキュメントを選択するこ

とによって簡単に選択されるものとする。

【００５９】

図５において、ツリー８１のノード８２～８６は、個別ドキュメント、たとえば、ドキュメント８８などのコレクションを表現するメタドキュメントである。たとえば、図５において、ノード８９は３個の子、ドキュメント８８ａ、８８ｂ、８８ｃを有する内部ノードである。内部ノード８９も、内部ノード８４の子であり、内部ノード８４自体はルートノード８２の子である。ルートノード８２は、ドキュメントコレクション全体を表現するメタドキュメントである。メタドキュメント８３～８６はメタドキュメント８２から直接に得られる子である。さらに、メタドキュメント８９のレベル８７は、メタドキュメント８３～８６から直接に得られる子である。最後に、個別ドキュメント８８、すなわちツリーの葉は、メタドキュメント８７から直接に得られる子である。ツリー８１は、説明上、非常に簡単にしてある。実際には、大きなコーパスは非常に多数の個別ドキュメントおよび便利に示す必要があるメタドキュメントのレベルを有する。

【００６０】

一例として、１０，０００のドキュメントをクラスタ化し、１０の話題に関連するグループ、すなわちクラスタとする場合を考える。この例の場合、同じ１０，０００ドキュメントの、たとえば５００クラスタへの原型のクラスタ化は、既に利用可能である。互いに極端に類似しているドキュメントは、通常、同じクラスタに現れるので、５００のクラスタの内の所定のクラスタのすべてのドキュメントは、所望の１０のクラスタの内の同じクラスタに同様に出現するものとする。言い換えれば、細粒度クラスタ化において一緒にクラスタ化されるほど十分に類似しているドキュメントは、粗粒度クラスタ化において、一緒にクラスタ化されることになる。これは、米国特許第５，４８３，６５０号に開示されているクラスタリファインメント（refinement）仮説である。

【００６１】

本発明は、既存のクラスタをメタドキュメントとして処理し、このメタドキュメントは全体としてコーパス全体の圧縮表現を形成する。すべての個別ドキュメントを直接にクラスタ化する代わりに、本発明は、すべての個別ドキュメントを表現するメタドキュメントをクラスタ化する。前述した例において、１０，０００の個別ドキュメントをクラスタ化する代わりに、本発明によれば、５００のメタドキュメントをクラスタ化することができる。クラスタ洗練仮説によれば、メタドキュメントクラスタ化および個別ドキュメントクラスタ化は、同様な結果を生成する。

【００６２】

たとえば、ステップＳ３４０からＳ３６０までの第１反復の場合、図５のフォーカスセット１００は、ドキュメントコレクション全体を表現するルートノードすなわちメタドキュメント８２のみを含む。当然、第１反復中は、このメタドキュメント８２は、フォーカスセットＴの唯一のメタドキュメントであるので、ステップＳ３４０において選択される。ステップＳ３５０において、メタドキュメント８２は、その直接の子孫、すなわち子であるメタドキュメント８３～８６に拡張される。次に、これらの子メタドキュメント８３～８６を使用し、図６に示すように、フォーカスセット１００において、メタドキュメント８２を置換する。このようにして、フォーカスセット１００は、子孫のメタドキュメント８３～８６を含む。

【００６３】

次に、ステップＳ３４０が、図６のフォーカスセット１００に関して反復される。フォーカスセット１００内のメタドキュメントの数がステップＳ３６０における所定の最大数未満である限り、クラスタ化処理はステップＳ３４０～Ｓ３６０を経由して循環を継続する。メタドキュメント８３～８６の内、メタドキュメント８４は、最低数の個別ドキュメント８８を表現する。すなわち、メタドキュメント８４は、６個の個別ドキュメントを表現し、一方、メタドキュメント８３、８５、および８６は、それぞれ、７、８、および９個の個別ドキュメントを表現する。したがって、メタドキュメント８４は、図７に示すように、選択され、その子孫、すなわち孫、メタドキュメント８９～９２に拡張される。しか

し、孫メタドキュメント９０および９１は、関心のあるドキュメントを含まないので除去される。したがって、フォーカスセット１００は、今度は、メタドキュメント８３、８５～８６、８９、および９２を含む。

【００６４】

所定の最大数のメタドキュメントが、ステップＳ３６０において、まだ実現されない場合、ステップＳ３４０が、図７に示すフォーカスセット１００に関して反復される。最低数の個別ドキュメントを表現する子メタドキュメント８３が、ステップＳ３５０において選択され、図８に示すように、その子孫、すなわち孫、メタドキュメント８７、および９３～９５に拡張される。しかし、メタドキュメント９５は、関心のあるドキュメントを包含しないので、メタドキュメント９５は除去される。したがって、フォーカスセット１００は、ここで、子孫メタドキュメント８５～８７、８９、および９２～９４を包含する。

【００６５】

図２、図３または図４、および図５に輪郭を示す処理は、フォーカスセット内のメタドキュメントの数が所定の最大数未満である限り継続される。所定の最大数が充分に高い場合、フォーカスセットは、実質上、個別ドキュメントを含む。その場合、ステップＳ３６０によって、メタドキュメントおよび個別ドキュメントの全数が所定の最大数未満であるかが決定される。しかし、この状況は、通常発生せず、特に、処理の僅かしかない第１反復中には発生しない。

【００６６】

図１に示すように、再クラスタ化システム１０は、好適には、プログラム式汎用コンピュータ上において実現される。しかし、再クラスタ化システム１０は、専用コンピュータ、プログラム式マイクロプロセッサまたはマイクロコントローラおよび周辺一体型回路構成要素、ＡＳＩＣまたは他の一体型回路、ディジタル信号プロセッサ、有線（ハードワイヤード：hardwired）電子または論理回路たとえば個別要素(ディスクリートエレメント：discrete element)回路、ＰＬＤ、ＰＬＡ、ＦＰＧＡ、ＰＡＬなどのプログラマブル論理装置、などによっても実現することができる。一般に、図２から図５に示す流れ図を実行することができる有限状態機械(finite state machine)を実現できるいかなる装置を使用しても、再クラスタ化システム１０を実現することができる。

【００６７】

以上、特定の実施形態について述べたが、多数の代替方法、変形、および異形は当業者には明らかであることは、明白である。したがって、前述した本発明の好適な実施形態は、説明を目的とするものであり、これに限定されるものではない。特許請求の範囲によって規定される本発明の思想および範囲を離脱することなく、種々の変化を実施し得る。

［付記］

［付記１］　電子的に記憶されるドキュメントのコーパスを処理し、一つ以上の事前に識別された関心のあるドキュメントをクラスタ化する方法であって、

複数のドキュメントを代表する少なくとも一つの初期メタドキュメントを含むフォーカスセットを拡張し、複数の次のメタドキュメントとするステップであって、それぞれの次のメタドキュメントは前記初期メタドキュメントのサブセットであるステップと、

前記フォーカスセット内のメタドキュメントを選択するステップと、

前記選択されたメタドキュメントを子孫のメタドキュメントに拡張するステップと、

少なくとも一つの関心のあるドキュメントを含まない子孫のメタドキュメントを除去するステップと、

前記次のメタドキュメントの数が少なくとも所定の最大数に等しくなるまで、前記選択および拡張ステップを反復するステップと、

を含む拡張ステップと、

前記次のメタドキュメントをクラスタ化し、所定数のクラスタとするステップと、

を含むことを特徴とする方法。

［付記２］　付記１に記載の方法において、クラスタの前記所定数は、前記所定最大数未満であることを特徴とする方法。

［付記３］　付記１に記載の方法において、前記少なくとも一つの初期メタドキュメントは、ドキュメントのコーパス全体を代表する単独のメタドキュメントであることを特徴とする方法。

［付記４］　付記１に記載の方法において、前記所定最大数は、前記拡張および選択ステップが、共に、時間制約内に完了するように決定されることを特徴とする方法。

［付記５］　付記１に記載の方法において、さらに、
前記新メタドキュメントの要約を確定するステップと、
前記要約をユーザに提示するステップと、
を含むことを特徴とする方法。

［付記６］　付記５に記載の方法において、前記要約は、
各新メタドキュメントにおいて最も頻繁に現れる固定数の話題のワードと、
各新メタドキュメント内の少なくとも一つの典型的なドキュメントの名称と、
を含むことを特徴とする方法。

［付記７］　付記１に記載の方法において、前記クラスタ化ステップは、各メタドキュメントによって表現されるドキュメントの数に関係なく、多くても、所定量の時間を要することを特徴とする方法。

［付記８］　付記１に記載の方法において、前記拡張ステップは、さらに、選択されたメタドキュメント内の関心のあるドキュメントの数がカットオフ値を超えるかを決定するステップを含むことを特徴とする方法。

［付記９］　付記８に記載の方法において、前記選択されたメタドキュメント内の関心のあるドキュメントの数がカットオフ値未満である場合、前記ドキュメントは別のドキュメントセットに加えられることを特徴とする方法。

［付記１０］　付記９に記載の方法において、前記拡張ステップは、さらに、前記クラスタ化ステップにおいてクラスタ化するために、前記別のドキュメントセットを前記次のメタドキュメントに加えるステップを含むことを特徴とする方法。

［付記１１］　電子的に記憶されるドキュメントのコーパスを処理し、少なくとも一つの事前に識別された関心のあるドキュメントをクラスタ化する装置であって、
複数のドキュメントを代表する少なくとも一つの初期メタドキュメントを含むフォーカスセットを拡張し、複数の次のメタドキュメントとする拡張手段であって、それぞれの次のメタドキュメントは前記少なくとも一つの初期メタドキュメントのサブセットである拡張手段と、
前記フォーカスセット内のメタドキュメントを選択するための選択手段であって、選択されたメタドキュメントは拡張手段によってその子孫のメタドキュメントに拡張される選択手段と、
少なくとも一つの関心のあるドキュメントを含まない子孫のメタドキュメントを除去するための除去手段と、
を含む拡張手段と、
前記次のメタドキュメントをクラスタ化し、所定数のクラスタとするためのクラスタ化手段と、
を備え、
前記拡張手段は、前記次のメタドキュメントの数が所定の最大数に少なくとも等しくなるまで、前記フォーカスセットを拡張することを特徴とする装置。

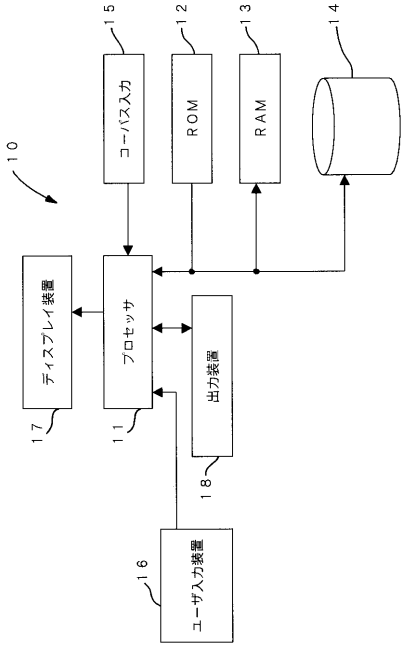
［付記１２］　付記１１に記載の装置において、新メタドキュメントの前記所定数は、前記所定最大数未満であることを特徴とする装置。

［付記１３］　付記１１に記載の装置において、前記少なくとも一つの初期メタドキュメントは、ドキュメントのコーパス全体を代表する単独のメタドキュメントであることを特徴とする装置。

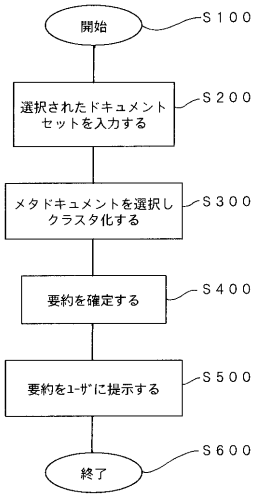
［付記１４］　付記１１に記載の装置において、前記所定最大数は、前記クラスタ化手段が前記次のメタドキュメントのクラスタ化を時間制約内に完了するように決定されることを特徴とする装置。

［付記１５］　付記１１に記載の装置において、さらに、
新メタドキュメントの要約を確定し、前記要約をユーザに提示する要約手段を備えることを特徴とする装置。
［付記１６］　付記１５に記載の装置において、前記要約は、
各新メタドキュメントにおいて最も頻繁に現れる固定数の話題のワードと、
各新メタドキュメント内の少なくとも一つの典型的なドキュメントの名称と、
を含むことを特徴とする装置。
［付記１７］　付記１１に記載の装置において、前記クラスタ化手段は、各メタドキュメントによって表現されるドキュメントの数に関係なく、多くても、所定量の時間を要することを特徴とする方法。
［付記１８］　付記１１に記載の装置において、前記拡張手段は、前記メタドキュメント内の関心のあるドキュメントの数がカットオフ値を超えるかを決定することを特徴とする装置。
［付記１９］　付記１８に記載の装置において、前記メタドキュメント内の関心のあるドキュメントの数がカットオフ値未満である場合、前記ドキュメントは別のドキュメントセットに加えられることを特徴とする装置。
［付記２０］　付記１９に記載の装置において、前記拡張手段は、クラスタ化手段によってクラスタ化するために、前記別のドキュメントセットを前記次のメタドキュメントに加えることを特徴とする装置。
【図面の簡単な説明】
【図１】　本発明による装置の一実施形態を示すブロック図である。
【図２】　本発明による再クラスタ化方法の一実施形態の輪郭を示す流れ図である。
【図３】　図２のメタドキュメント拡張ステップの第１実施形態の輪郭をより詳細に示す流れ図である。
【図４】　図２のメタドキュメント拡張ステップの第２実施形態の輪郭をより詳細に示す流れ図である。
【図５】　本発明の好適な実施形態によるフォーカスセットのツリーおよび変化するコンテンツを示す図である。
【図６】　本発明の好適な実施形態によるフォーカスセットのツリーおよび変化するコンテンツを示す図である。
【図７】　本発明の好適な実施形態によるフォーカスセットのツリーおよび変化するコンテンツを示す図である。
【図８】　本発明の好適な実施形態によるフォーカスセットのツリーおよび変化するコンテンツを示す図である。
【図９】　分散／集合手順を広く示す図である。
【図１０】　従来の分散／集合ドキュメントブラウジング法を、ドキュメントの特定のコーパスに適用する場合を示す図である。
【符号の説明】
１０　再クラスタ化システム、１１　プロセッサ、１２　ＲＯＭ、１３　ＲＡＭ、１４　不揮発性メモリ、１５　コーパス入力、１６　ユーザ入力装置、１７ディスプレイ装置、１８　出力装置、８１　ツリー、８２　ルートノード（メタドキュメント）、８３～８７，８９～９５　ノード（メタドキュメント）、８８　ドキュメント、１００　フォーカスセット。

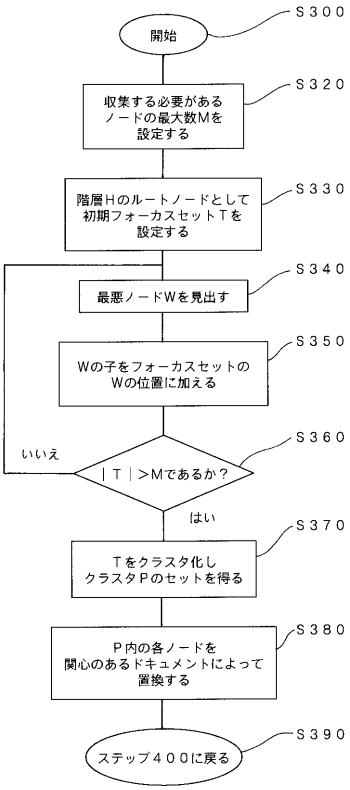
【図１】


10
15
コーパス入力
12
ROM
13
RAM
14
ディスプレイ装置
17
プロセッサ
11
出力装置
18
16
ユーザ入力装置

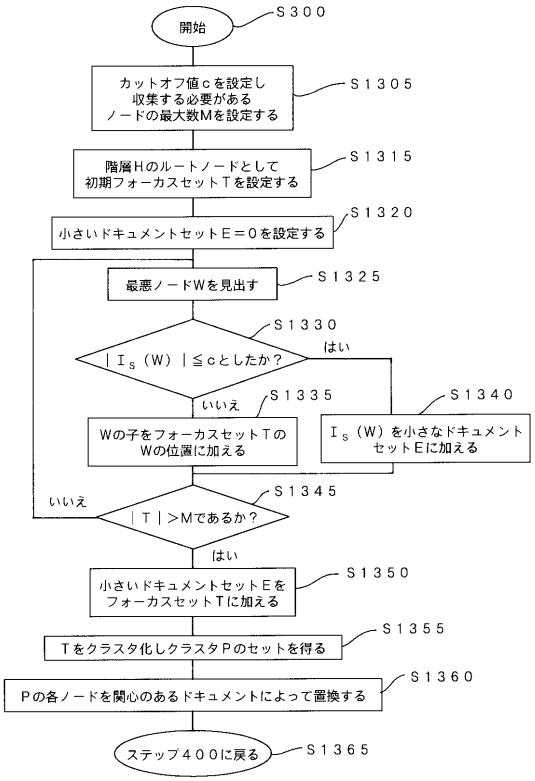
【図２】


開始
S100
選択されたドキュメントセットを入力する
S200
メタドキュメントを選択しクラスタ化する
S300
要約を確定する
S400
要約をﾕｰｻﾞに提示する
S500
終了
S600

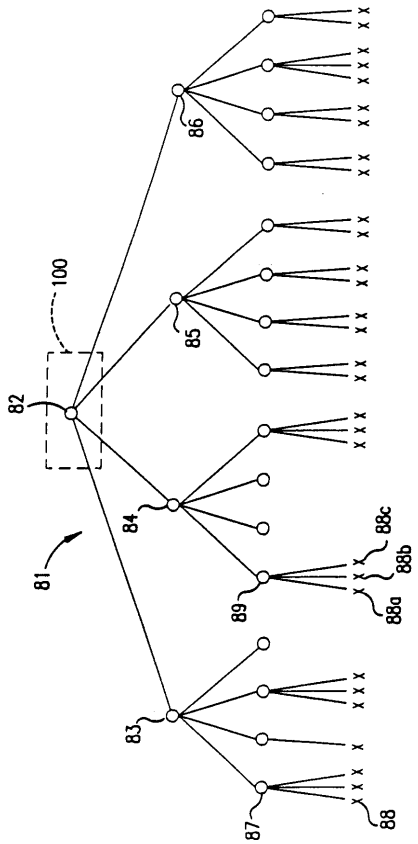
【図３】


開始
S300
収集する必要があるノードの最大数Mを設定する
S320
階層Hのルートノードとして初期フォーカスセットTを設定する
S330
最悪ノードWを見出す
S340
Wの子をフォーカスセットのWの位置に加える
S350
｜T｜＞Mであるか？
S360
いいえ
はい
Tをクラスタ化しクラスタPのセットを得る
S370
P内の各ノードを関心のあるドキュメントによって置換する
S380
ステップ４００に戻る
S390

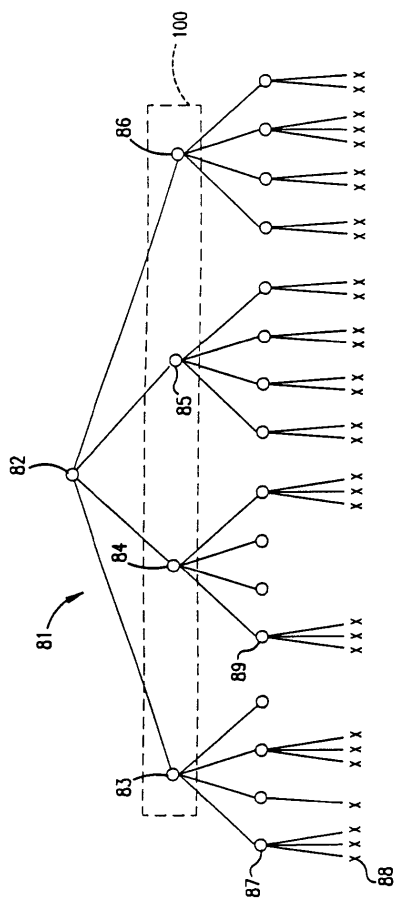
【図４】


開始
S300
カットオフ値ｃを設定し収集する必要があるノードの最大数Mを設定する
S1305
階層Hのルートノードとして初期フォーカスセットTを設定する
S1315
小さいドキュメントセットE＝０を設定する
S1320
最悪ノードWを見出す
S1325
｜Is（W）｜≦cとしたか？
S1330
はい
いいえ
Wの子をフォーカスセットTのWの位置に加える
S1335
Is（W）を小さなドキュメントセットEに加える
S1340
｜T｜＞Mであるか？
S1345
いいえ
はい
小さいドキュメントセットEをフォーカスセットTに加える
S1350
Tをクラスタ化しクラスタPのセットを得る
S1355
Pの各ノードを関心のあるドキュメントによって置換する
S1360
ステップ４００に戻る
S1365

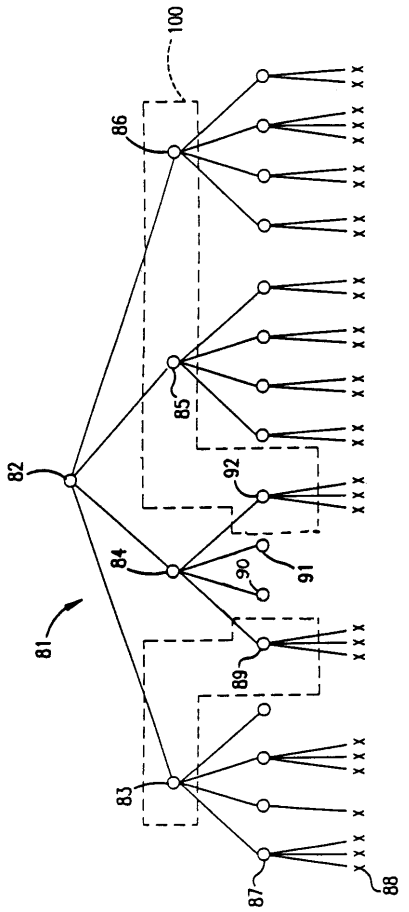
【図５】


100
82
81
84
83
85
86
87
88
88a
88b
88c
89

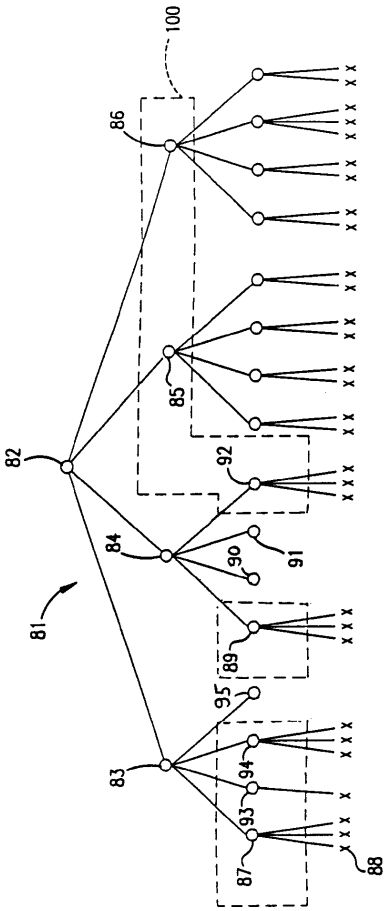
【図６】


100
82
81
84
83
85
86
87
88
89

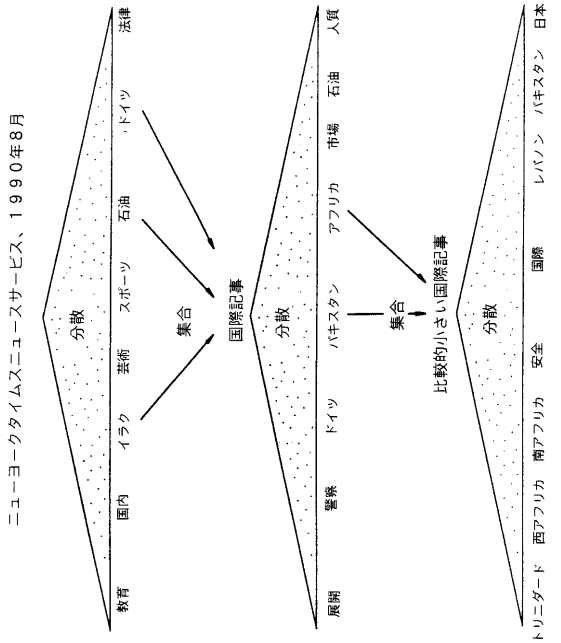
【図７】


100
82
81
84
83
85
86
87
88
89
90
91
92

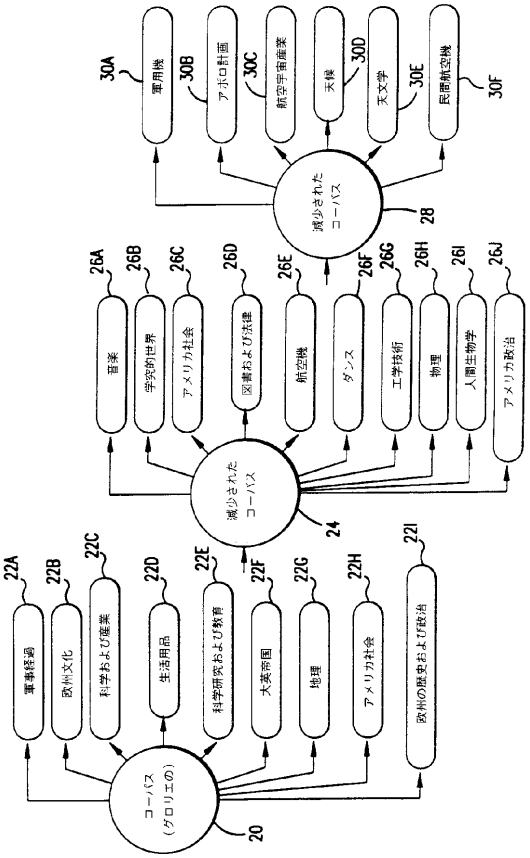
【図８】


100
82
81
84
83
85
86
87
88
89
90
91
92
93
94
95

【図９】


ニューヨークタイムスニュースサービス、１９９０年８月
分散
教育
国内
イラク
芸術
スポーツ
石油
ドイツ
法律
集合
国際記事
分散
展開
警察
ドイツ
パキスタン
アフリカ
市場
石油
人質
集合
比較的小さい国際記事
分散
トリニダード
西アフリカ
南アフリカ
安全
国際
レバノン
パキスタン
日本

【図１０】


コーパス
（グロリエの）
20
減少されたコーパス
24
減少されたコーパス
28
22A
軍事経過
22B
欧州文化
22C
科学および産業
22D
生活用品
22E
科学研究および教育
22F
大英帝国
22G
地理
22H
アメリカ社会
22I
欧州の歴史および政治
26A
音楽
26B
学究的世界
26C
アメリカ社会
26D
図書および法律
26E
航空機
26F
ダンス
26G
工学技術
26H
物理
26I
人間生物学
26J
アメリカ政治
30A
軍用機
30B
アポロ計画
30C
航空宇宙産業
30D
天体
30E
天文学
30F
民間航空機

フロントページの続き

審査官　辻本　泰隆

(56)参考文献　特開平０８－１５３１２１（ＪＰ，Ａ）
特開平０５－２２５２５６（ＪＰ，Ａ）
特開平８－２３５１９８（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/30