(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6431055号
(P6431055)
(45)発行日　平成30年11月28日(2018.11.28)　(24)登録日　平成30年11月9日(2018.11.9)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 310A

請求項の数 17 (全 21 頁)

(21)出願番号	特願2016-521534 (P2016-521534)
(86)(22)出願日	平成26年6月18日 (2014.6.18)
(65)公表番号	特表2016-524766 (P2016-524766A)
(43)公表日	平成28年8月18日 (2016.8.18)
(86)国際出願番号	PCT/US2014/042888
(87)国際公開番号	WO2014/205046
(87)国際公開日	平成26年12月24日 (2014.12.24)
審査請求日	平成29年6月9日 (2017.6.9)
(31)優先権主張番号	61/836,407
(32)優先日	平成25年6月18日 (2013.6.18)
(33)優先権主張国	米国(US)

(73)特許権者　315007628
コピーライト　クリアランス　センター，インク．
アメリカ合衆国　マサチューセッツ州　０１９２３，ダンバーズ，２２２　ローズウッド　ドライブ
(74)代理人　110000659
特許業務法人広江アソシエイツ特許事務所
(72)発明者　マーマニス，バビス
アメリカ合衆国　マサチューセッツ州　０２４５１，ウォルサム，２７　シェリル　レーン

最終頁に続く

(54)【発明の名称】文献のテキストマイニングのシステムおよび方法

(57)【特許請求の範囲】
【請求項１】
ユーザによるアクセスに多様な費用を伴う複数の調査文献の、前記ユーザによるテキストマイニングを容易にするためのシステムであって、
（ａ）前記複数の調査文献を格納するように構成されたコンテンツ収納庫であって、前記コンテンツ収納庫は、前記ユーザからの問合せを受けて、テキストマイニングのための前記複数の調査文献の予備的な集合を選択し、前記コンテンツ収納庫は、ユーザが選択的に問合せを変更して、前記ユーザに向けて最適化された前記複数の調査文献の最終的な集合を与えることを可能にする、前記予備的な集合の中の調査文献に関するコンテンツ展開メトリックを提供する、コンテンツ収納庫と、
（ｂ）導出テキストマイニングデータセットを提供する、調査文献の前記最終的な集合のテキストマイニングのためのテキストマイニングプロセッサと、
を備えるシステム。
【請求項２】
前記複数の調査文献のテキストマイニングを管理するプロジェクトマネージャであって、前記プロジェクトマネージャは前記コンテンツ収納庫およびテキストマイニングプロセッサと電気的にリンクしている、プロジェクトマネージャをさらに備える、請求項１に請求されるシステム。
【請求項３】
前記プロジェクトマネージャが、前記ユーザによる前記システムへの直接のアクセスの

ためのコンピュータインターフェースを提供する、請求項２に請求されたシステム。
【請求項４】
前記コンテンツ収納庫が、集められる前記調査文献の前記コンテンツ展開メトリックに関する１つまたは複数のルールに従って問合せを実行する、請求項３に請求されるシステム。
【請求項５】
前記コンテンツ収納庫が、前記予備的な集合の前記調査文献のコンテンツ展開メトリックに関するレポートを生成する、請求項４に請求されるシステム。
【請求項６】
前記レポートが、リスト、円グラフ、線グラフ、および単一の値で構成されるグループからの少なくとも１つの表示を含む、請求項５に請求されるシステム。
【請求項７】
前記コンテンツ収納庫が、
（ａ）前記複数の調査文献のそれぞれのための書誌的メタデータおよびフルテキストを格納するデータ格納デバイスと、
（ｂ）前記問合せを受け、実行するコンテンツ選択装置であって、前記コンテンツ選択装置が前記データ格納デバイスと電気的に接続されている、コンテンツ選択装置と
を備える、請求項４に請求されるシステム。
【請求項８】
前記データ格納デバイスが、前記コンテンツ収納庫が前記ユーザに対し、前記複数の調査文献のそれぞれのアクセス費用を既定することを可能にするユーザアクセス権のデータベースを含む、請求項７に請求されるシステム。
【請求項９】
前記コンテンツ選択装置が、文献アクセス費用のパラメータを前記問合せでサポートすることが可能である、請求項８に請求されるシステム。
【請求項１０】
前記コンテンツ選択装置が、前記問合せのための前記費用パラメータ内の、前記複数の調査文献のそれぞれのユーザアクセス費用を利用する、請求項９に請求されるシステム。
【請求項１１】
前記コンテンツ選択装置が、文献アクセス費用のパラメータを、前記ユーザによって既定され変更可能である前記問合せでサポートすることが可能である、請求項１０に請求されるシステム。
【請求項１２】
前記コンテンツ選択装置が最大ユーザアクセス費用を前記問合せでサポートする、請求項１１に請求されるシステム。
【請求項１３】
前記コンテンツ収納庫が、前記調査文献の前記予備的な集合の中の調査文献のために、レポートの費用に関するコンテンツ展開メトリックを提供する、請求項５に請求されるシステム。
【請求項１４】
前記コンテンツ選択装置が、将来の容易なテキストマイニング操作のための問合せに関連して引き出された、調査文献の最終的な集合を相互参照および格納する、請求項１に請求されるシステム。
【請求項１５】
前記テキストマイニングプロセッサが、同様のデータ構造の並行クラスタを利用して、前記調査文献の前記最終的な集合のテキストマイニングを実行する、請求項１に請求されるシステム。
【請求項１６】
前記テキストマイニングプロセッサが、標準およびカスタムの両方のテキストマイニングプロセッシングモジュールを構築するアプリケーションプログラミングインターフェー

スを含む、請求項１５に請求されるシステム。
【請求項１７】
前記テキストマイニングプロセッサと接続された導出データ収蔵庫であって、前記前記導出データ収蔵庫は前記導出テキストマイニングデータセットを格納する、導出データ収蔵庫をさらに備える、請求項１に請求されるシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は概して、科学、技術、および医学の分野において発行された調査文献に関し、より詳細には、包括的であるが効率的な方式での調査文献テキストマイニングのためのシステムおよび方法に関する。
【背景技術】
【０００２】
毎年、数千万もの学術文献が世界中で発行される。それらの発行された文献、または記事の大部分が、無料で提供される特定の記事へのアクセス、および各文献の権利を有する主体により示された費用で提供される他の記事へのアクセスにより、調査者によるレビューに電子的に利用可能である。
【０００３】
特定の調査トピックにおいて電子的に利用可能な情報が大量にあるため、調査者にとって、絶えず増え続けるその主題の電子情報を、包括的であるが効率的に検索することがしばしば困難である。特に、調査文献の検索に使用するには、在来の検索エンジンは適していないことがわかっている。この理由は、とりわけ、選択基準の仕様と処理が、関連性について少数の文献を評価するには効果的であるが、すべてが極めて特定の基準にフィットする大量の文献の中から選択する目的には不適当であるためである。このため、特定の主題に関して電子的に利用可能な膨大な量の情報があまりに大量であるため、調査者はしばしば、適切な文献を探し出すことに失敗するリスクがあり、このことは極めて望ましくないものである。
【０００４】
したがって、調査者による大量の発行された記事の検索を補助するために、組織（例えば、発行者および権利のマネジメントサービス）が、当該技術分野で「テキストマイニング」として知られているプロセスを通して、調査文献のテキストから高品質のデータを解析し、抜粋することを可能にするソフトウェアおよびデータベースを構築することがますます慣習的になってきている。テキストマイニングプロセスの解析を通して、数百万もの文献からのテキストを分析し、相互参照することで、コンピュータベースの検索ツールを使用することにより、調査者が適切な発行物をより効率的に特定すること可能にする。
【０００５】
発行された調査文献の効率的なテキストマイニングのプロセスには、多くの課題があり、現在、一定の制限がされている。
【０００６】
第１の課題として、発行された調査文献の効率的なテキストマイニングはまず、関連する文献の全集を大量に集める必要がある。具体的には、より包括的にするために、科学調査のテキストマイニングには、可能なかぎり多くの調査文献にアクセスすることが必要である。同時に、調査文献の集合の権利者は、許可されていない記事の複写および拡散のリスクがあり、このために、権利者が、購読および他の購入されたアクセスの慣習的な形態を通して文献から潜在的に収入を得ることが妨げられることから、テキストマイニングの目的で文献にアクセスすることに同意することをしばしば躊躇する。あらゆる許可されていない文献の複写によるリスクを制限するため、発行者はしばしば、テキストマイニングの目的のために、記事を無作為に（例えば、アルファベット順に並べられた文または単語で）抽出する形式で提供する。しかし、無作為に記事を抽出する形式では、テキストマイニングの特定の機能（例えば、識別された記述パターンに基づいて調査書類と実験記録と

を区別する能力）が制限されることがわかっている。このため、このやり方は理想的ではないことがわかっている。

【０００７】

第２の課題として、発行された調査文献のテキストマイニングには、現在、エンドユーザへの費用を含むことが考慮されていない。上述したように、異なる文献にアクセスするのに別々の費用を伴う。このため、検索経費が制限されている調査者は、無料の発行物の検索に制限することを選択する場合があり、このため、適切な文献を探し出すことに関するリスクがある。同様に、文献にアクセスするのに料金が必要である発行物を含む多くの発行物に検索範囲を広げることを選択した、検索経費が制限されている調査者は、しばしば、過大で法外な調査費用に苦しめられる。

【０００８】

第３の課題としては、発行された調査文献の効率的なテキストマイニングには、検索結果がエンドユーザに、文献の大きな集団のテキストのすべてへのアクセスを提供することが必要である。対照的に、在来の検索エンジンは、人による評価のための制限された文脈上の情報を伴う、個別の記事へのリンクのリストのみを返す。このことは、調査者が各記事の関連性を判断するのには不十分であることがわかっている。

【０００９】

第４の課題としては、発行された調査文献のテキストマイニングは現在、エンドユーザに、検索結果に関する有用な問合せ情報を何も提供していない。さらに、エンドユーザは概して、なぜ最初の検索の際に特定の文献が検索されたのかを判断するための制限されたデータを有する。このように、エンドユーザは、以前の検索からの情報を使用して将来の検索の全体の効率性を上げることが妨げられる。

【発明の概要】

【発明が解決しようとする課題】

【００１０】

本発明の目的は、調査文献のテキストマイニングのための新規の改善されたシステムおよび方法を提供することである。

【００１１】

本発明の別の目的は、包括的で費用効率の高い方式での、調査文献のテキストマイニングのためのシステムおよび方法を提供することである。

【課題を解決するための手段】

【００１２】

したがって、本発明の特徴の１つとして、ユーザによるアクセスに多様な費用を伴う複数の調査文献の、ユーザによるテキストマイニングを簡略化するためのシステムであって、（ａ）複数の調査文献を格納するように構成されたコンテンツ収納庫であって、コンテンツ収納庫は、ユーザからの問合せを受けて、テキストマイニングに向けて複数の調査文献の予備的な集合を選択し、コンテンツ収納庫は、ユーザが選択的に問合せを変更して、ユーザに向けて最適化された複数の調査文献の最終的な集合を与えることを可能にする、予備的な集合の中の調査文献に関するコンテンツ展開メトリックを提供する、コンテンツ収納庫と、（ｂ）得られたテキストマイニングデータセットを提供する、調査文献の最終的な集合のテキストマイニングのためのテキストマイニングプロセッサと、を備えるシステムが提供される。

【００１３】

他の様々な特徴および利点が以下の説明から明らかになる。説明中、添付図面が参照され説明の一部を形成する。その中で、本発明を実施するための実施形態が例として示される。実施形態は、当業者が本発明を実施することを可能にするように十分に詳細に記載され、他の実施形態が利用でき、本発明の範囲を逸脱することなくその構造の変更がされ得ることが理解される。したがって、以下の詳細な説明は限定の意味としてとられず、本発明の範囲は、添付の特許請求の範囲により最適に既定される。

【００１４】

図中、同様の参照符号が同様の部品を示す。
【図面の簡単な説明】
【００１５】
【図１】図１は、文献のテキストマイニングのためのシステムの簡略化されたブロック図であり、このシステムは本発明の教示にしたがって構成されている。
【図２】図２は、図１に示すコンテンツ収納庫に格納された、記事に関する様々な形式のデータ間の実施可能な関係を理解するのに有用な、例示的なデータモデルである。
【図３】図３は、図１に示す文献収納庫内の記事のアクセスドメインの実施を理解するのに有用な、例示的なデータモデルである。
【図４】図４は、図１に示すシステムを使用する文献のテキストマイニングの新規の方法の簡略化されたフローチャートである。
【図５】図５は、図４に示すテキストマイニングの方法のより詳細なフローチャートである。
【図６】図６は、図１に示すコンテンツ選択装置に格納された、展開メトリックに関するデータの実施可能な関係を理解するのに有用な例示的なデータモデルを示す。
【図７ａ】図１に示すシステムの例示的な使用を理解するのに有用な、一連のサンプルスクリーン表示である。
【図７ｂ】図１に示すシステムの例示的な使用を理解するのに有用な、一連のサンプルスクリーン表示である。
【図７ｃ】図１に示すシステムの例示的な使用を理解するのに有用な、一連のサンプルスクリーン表示である。
【図７ｄ】図１に示すシステムの例示的な使用を理解するのに有用な、一連のサンプルスクリーン表示である。
【図７ｅ】図１に示すシステムの例示的な使用を理解するのに有用な、一連のサンプルスクリーン表示である。
【発明を実施するための形態】
【００１６】
テキストマイニングシステム１１
ここで、図１を参照すると、調査文献のテキストマイニングのためのシステムの全体的なブロック図が示されている。このシステムは本発明の教示にしたがって構成され、全体として参照符号１１で識別されている。以下でさらに詳細に説明されるように、システム１１は、とりわけ、（ｉ）それに続くテキストマイニング操作の主題となる、費用パラメータを調査文献の集合の選択のプロセスに組み込み、それによって、（ｉｉ）選択された文献の展開に関する、ユーザが直観的に認識可能なメトリックを提供するように設計される。必要であれば、ユーザは、この後に、テキストマイニングされる調査文献の最適化された集合をもたらすために、メトリックを利用して文献選択プロセスの特定のパラメータを修正することができる。この能力で、システム１１は、包括的であるが費用効率がある、調査文献の集合のテキストマイニングを促進する。このことは、本発明の主要な目的である。
【００１７】
例示の目的のみのために、システム１１は本明細書において、調査文献の大きな収納庫を使用して行われるテキストマイニング作業と関連づけて説明される。しかし、システム１１は調査文献のテキストマイニングに限定されないことが理解される。さらに、システム１１は、あらゆる形式の文献、特に、アクセスするのに費用を伴うあらゆる文献から関連するテキストを識別することを要するあらゆる環境で使用可能であることが理解される。
【００１８】
システム１１は複数のモジュールを含み、この複数のモジュールは全体で、本発明のテキストマイニング操作をエンドユーザ１３に提供する。具体的には、以下に詳細に記載されるように、システム１１は、システム１１の中心の機能的ハブとして働くプロジェクト

マネージャ１５と、テキストマイニングおよび計量されたアクセスのための記事を含む文献収納庫１７と、本発明の主要なテキストマイニング作業を行うテキストマイニングプロセッサ１９と、テキストマイニングプロセッサ１９によって行われるテキストマイニング操作の出力を格納する導出データ収納庫２１とを備える。

【００１９】

プロジェクトマネージャ１５は本明細書において、任意の通信媒体を介して（例えば、インターネットを介して）、エンドユーザ１３のコンピュータデバイスと電子的にリンクしているサーバとして示される。この方式で、プロジェクトマネージャ１５は、エンドユーザ１３に、システム１１にアクセスするための主要なインターフェイスを提供する。以下にさらに詳細に記載されるように、プロジェクトマネージャ１５はエンドユーザ１３に、（ｉ）新たなテキストマイニングプロジェクトを構築すること、（ｉｉ）進行中のプロジェクトの状態および進捗を追跡すること、および、（ｉｉｉ）完了したプロジェクトから返されるデータにアクセスすることを可能にする。

【００２０】

なお、テキストマイニングプロジェクトへのアクセスは、個別、団体ベース、および組織のいずれかのレベルのアクセス権で、プロジェクトマネージャ１５から所与のエンドユーザ１３に与えることができる。この能力で、システム１１が様々な異なる環境で実施され得ることが見込まれる。

【００２１】

文献、またはコンテンツ収納庫１７は、学術記事の大きい集まりの書誌的メタデータとフルテキストとの両方を含むデータ格納デバイス２３－１および２３－２を備え、コンテンツは、速やかに引き出すことを容易にするように、索引がつけられていることが好ましい。

【００２２】

たとえば、ここで図２を参照すると、コンテンツ収納庫１７に格納された、記事に関する様々な形式のデータ間の実施可能な関係を理解するのに有用な、例示的なデータモデルが示されている。このデータモデルは全体として、参照符号２５で識別される。しかし、本発明の趣旨を逸脱することなく、他のデータベース技術の類似のデータモデルが同様に、データベースモデリングの経験を積んだ専門家によって構築され得ることが理解される。

【００２３】

見て取れるように、データモデル２５は、各記事のメタデータを有する記事テーブル２７を含み、このデータは、それに限定されないが、ワークのタイトル、ワークの著作者、および特定のキーワードを含む。記事テーブル２７は、各記事のフルテキスト（すなわち、発行された形式の文献を構成する完全な原本に基づく事項）、ならびに、書誌的事項、引例のリスト、および／または、収納庫１９に置かれているか、もしくは置かれていない別の記事のセットの参照をさらに含むことが好ましい。

【００２４】

著作者テーブル２９は、（記事著作者テーブル３１を介して）記事テーブル２７とリンクしており、学術文献を創作する様々な個人または組織を表す。著作者は文献収納庫１７に、名前によって、また、基本的な識別名の任意のセットをともなって出されることが好ましい。

【００２５】

出所テーブル３３が、記事の一般的なソースに関するデータ（すなわち、どこで記事を見つけられるか）を提供する。ジャーナル（すなわち、記事のセットを発行する学術ワーク）と収蔵庫とは両方とも出所の形式である。したがって、ジャーナルテーブル３５が、その中に見られるタイトル、標準番号、および発行者を含む各ジャーナルの属性とともに、出所テーブル３３とリンクする。同様に、コレクションテーブル３７が出所テーブル３３とリンクし、ジャーナルと集合との両方に潜在的に見られる記事とともに、記事の代替的なソースを提供する。

【００２６】
最後に、発行テーブル３９が記事テーブル２７および出所テーブル３３内のデータ間の関係を構築する。発行テーブル３９は、発行者から直接、しばしば高額での記事の利用可能性を示すデータを含む。たとえば、特定の記事は、そのもともとの発行者から４０ドルで、また、ドキュメントの収納庫から５ドルで、利用可能である場合がある。
【００２７】
したがって、例示的なデータモデル２５の構造を使用することで、他の事を含め、（ｉ）著作者もしくは著作者のセット、（ｉｉ）記事のタイトル、（ｉｉｉ）キーワードもしくは他の同様のメタデータの領域、（ｉｖ）発行物もしくは発行物のセット、（ｖ）ジャーナルもしくはジャーナルのセット、（ｉｖ）集合もしくは集合のセット、および／または（ｖｉｉ）発行データの範囲に関するデータを使用して検索の問合せを容易に処理できることが明らかである。
【００２８】
少なくとも１つのデータ格納デバイス２３が、ユーザのアクセス権のデータベースをさらに含むことが理解される。したがって、権利の付与、またそれによって、問合せ、ジョブ、およびユーザによる記事レベルのアクセス記録に基づき、文献収納庫１７は各ユーザのアクセス権を追跡することができる。
【００２９】
たとえば、図３を参照すると、文献収納庫１７内の記事のアクセスドメインの実施を提供する例示的なデータモデルを示しているこのデータモデルは全体として参照符号４１で識別される。見て取れるように、データモデル４１は、エンドユーザテーブル４３と団体テーブル４５とを（団体ユーザテーブル４７を介して）相互参照する。この理由は、各団体は通常、複数の異なるユーザを含むためである。さらに、団体はしばしば、購読を複数購入するため、団体テーブル４５は購読テーブル４９とリンクしている。記事のソース（たとえば、その中の記事が購入可能な異なる集合）を明示する出所テーブル５１は、そのため、購読アイテムテーブル５３を介して購読テーブル４９とリンクしている。したがって、システム１１は、エンドユーザ１３に文献収納庫１７に含まれる大量の記事を効率的にテキストマイニングすることを可能にするのみならず、各エンドユーザ１３がどの記事を購読しているかを容易に確認する。このことは、非常に望ましいことである。
【００３０】
図１に戻ると、文献収納庫１７は、データ格納デバイス２３とプロジェクトマネージャ１５との両方に接続された、コンテンツ選択装置、または問合せプロセッサ５５をさらに有する。したがって、以下にさらに記載されるように、コンテンツ選択装置５５はデータ格納デバイス２３から調査文献にアクセスし、様々な異なるフルテキストおよびメタデータの問合せを実施することにより、記事の、最適化されたサブセット、またはクラスタを選択する。結果として得られた文献のクラスタは、その後、将来の問合せを容易にするために、コンテンツ選択装置５５に格納され、文献のクラスタは、元の問合せが繰り返された際に、必要に応じてアップデートされる。
【００３１】
本発明の主要な特徴としては、コンテンツ選択装置５５が、データ格納デバイス２３から最初の文献の集団を与えるために、費用パラメータをフルテキストおよびメタデータの問合せに含むことできる。さらに、コンテンツ選択装置２５は、エンドユーザ１３に、最初の問合せから得られた選択された文献の展開に関する直観的メトリックを与える。この方式で、ユーザは、以下にさらに説明するように、続いてテキストマイニングされる調査文献の、包括的であるが効率的な展開を得るために、必要に応じて問合せを改善できる。
【００３２】
上記で簡潔に参照したように、テキストマイニングプロセッサ１９は、本発明の主要なテキストマイニング作業を担っている。言い換えると、テキストマイニングプロセッサ１９は調査者に、収納庫１９から引き出された文献の関連する集合のテキストマイニングジョブを特定することを可能にし、そのジョブを、そのジョブの要求と非同期的に実行し、

そして、完了した後に調査者に知らせる。

【００３３】

本明細書で述べるように、テキストマイニングプロセッサ１９は、標準化されたアーキテクチャにしたがって、テキストマイニングプログラムを並行して実行するように設計された、スタックされた複数の計算装置５７－１～５７－３を備える。具体的には、テキストマイニングソフトウェアが入力データを導出データ収納庫２１内の計算装置５９－１～５９－３から受け取り（すなわち、前回のテキストマイニング操作の出力）、文献のセット内に特定された集合に向けて、文献のメタデータおよびフルテキストのテキストマイニング操作を並行して実施して、その後に導出データ収納庫２１内の名前が付されたデータセットに格納される出力を与える。各作業に向けた処理資源の分配は、テキストマイニングプロセッサ１９によって内部で追跡されることが好ましい。

【００３４】

システム１１を使用するテキストマイニング法１１１

上記で簡潔に参照したように、システム１１は調査文献のテキストマイニングの新規の方法に関わるように設計される。具体的には、ここで図４および５を参照すると、システム１１を使用するテキストマイニングのための文献の選択、購入、および処理の新規の方法の、簡略化されたフローチャート、およびわずかにより詳細なフローチャートがそれぞれ示されている。この方法は本明細書において、全体として参照符号１１１で識別されている。

【００３５】

以下にさらに詳細に説明するように、本発明のテキストマイニング法は、最初に、検索変数、またはパラメータ値のセットを使用して調査文献の集団、またはプールを集めて、潜在的に関連する調査文献の広範囲の集合を与える。言い換えると、最初の集合は、問合せの基準に最も合う単一の文献を見つけることを試みるような、人間の選択のために、関連性により優先順位をつけられた文献を返すことを求めていない。代わりに、結果のセットは審査のために提供されるのではなく、むしろそれに続くテキストマイニング処理のために集められる。

【００３６】

前述の文献選択プロセスは、複数の記事の回りに「フェンス」を投じて、集合のサブセットを形成することに類似している。そのため、このフェンスの構成は、後に、コンテンツの広げられたメトリック（たとえば、なぜ特定の文献が最初に選択されたかについての情報）を使用して、ユーザによって変更され、エンドユーザ１３にとって最も適切で望ましい選択（たとえば、費用、発行者など）に、元の調査文献のプールを再び既定、または狭めることができる。この方式で、それらのすべてが特定の性質に従う調査文献の高品質の選択が、それに続く、効率的かつ費用効率のよい方式のテキストマイニング操作のために集められる。

【００３７】

なお、テキストマイニングジョブは、プロジェクトマネージャ１５にアップロードされるプログラムコードで構成される。

【００３８】

プロセス１１１を開始するために、エンドユーザ１３はまず、テキストマイニングプロジェクトを既定、または構築する。このプロジェクト既定ステップは、全体として参照符号１１３で識別される。具体的には、プロジェクト既定ステップ１１３の一部として、エンドユーザ１３は、（ｉ）テキストマイニング操作に利用される文献セット（すなわち、収蔵庫１９内のコンテンツの選択）、（ｉｉ）プロセスの明細（すなわち、文献のトークン分解、特有の属性の計算、および同様のデータ構造の並行クラスタリング）、ならびに（ｉｉｉ）明細の報告（すなわち、ユーザにテキストマイニングの結果を提供するための特定の手段）を特定する。

【００３９】

なお、文献セットは、（ｉ）文献の識別名、著作者、共同制作者、機関、および発行者

など（もしくは、上述の属性の任意のリストまたは集合）の明細を使用する文献の問合せを通して、または（ｉｉ）あらかじめ既定された文献セット（すなわち、以前の問合せの結果として得られた文献セット）を使用することで、特定することができる。

【００４０】

ステップ１１３が完了すると、コンテンツ選択装置５５は、ステップ１１３で特定された、あらゆるコンテンツ展開条件（たとえば、「Ｃ．Ｅｌｅｇａｎｓ」とうい用語を含むが、発行者Ｘによる記事は除くすべての文献を特定する）を受けて、ジョブに向けて調査文献を選択する。この文献選択ステップは、全体として参照符号１１５で識別されている。

【００４１】

文献選択ステップ１１５の一部として、システム１１は、エンドユーザ１３が、最初の文献の集合に関連する展開メトリックを識別し、分析することを可能にするユーザインターフェイスを生成する。この能力で、エンドユーザ１３は予備的な問合せの特定のパラメータを変更して、テキストマイニングされる、より最適な文献の集合を与えることができる。

【００４２】

対照的に、在来のテキストベースの検索結果は通常、説明がされない。言い換えると、ユーザは概して、なぜ検索結果が特定され、特定の順番で並べられるのかを理解しない。しかし、調査フィールドでは、調査者は検索要求からの任意のコンテンツの選択を利用できない。大量の調査記事が利用可能であるため、調査者は、なぜ特定の記事が選択されたのか、より重要には、どのように検索パラメータの重要性、または詳細を変更して検索結果に影響を与えたのかを知る必要がある。

【００４３】

したがって、上記で簡潔に参照したように、問合せプロセッサ５５は、選択された検索メトリック（すなわち、コンテンツ、発行者、費用などによる検索結果の内訳）に基づき、ユーザに向けてレポートを生成する。この方式で、エンドユーザ１３は検索結果に影響を与えた要因をよりよく特定することができる。したがって、システム１１はエンドユーザ１３が、この後、作動中に（ｏｎ　ｔｈｅ　ｆｌｙ）検索パラメータを調整し、それに続く文献の第二の集合を実施して、前回の集合内に発見されたあらゆる不十分さを順応させることを可能にする。

【００４４】

ステップ１１５で最初に集められた調査文献の集団が拡大したことにより、文献処理ステップが、その中の文献の最適化されたグループ、またはサブセット（すなわち、識別された特定のキーワードに関して最も近似の文献）を既定、または識別し始める。この文献処理ステップは全体として参照符号１１７で識別される。

【００４５】

文献処理ステップ１１７は好ましくは、大きなデータセットのバッチ処理に使用される、パイプライン方式のマップの様々な縮小パラダイムを利用する。好ましくは、テキストマイニングプロセッサ１９は、カスタムマップの構築のためのアプリケーションプログラミングインターフェイス（ＡＰＩ）を提供し、モジュールを削減する。

【００４６】

具体的には、個別の文献の操作を行って各文献を他の形態に変換する「マップ」処理が特定され得る。たとえば、プロセスにより、遺伝子配列リサーチを記載する文書が、各文書によって言及される特定の遺伝子のリストに変換され得る。

【００４７】

さらに、「削減」プロセスは、変換された文書のリストを集合の形態に結合する。たとえば、プロセスにより、調査文書の集合によって言及される遺伝子のリストを取り、したがって、調査を行う機関により集計された遺伝子のリストを返すことができる。削減変換の第二の段階が、第一の段階の出力を操作して、機関による遺伝子のセットを取り、機関による集計を繰り返す。これは「結合」変換と呼ばれる。このやり方でプロセスを分割す

ることで、ジョブを並行して実施することを補助するのに役立つ。

【００４８】

本発明の新規の特徴として、文献処理ステップ１１７は、標準処理モジュール１１９とカスタム処理モジュール１２１との両方を補助し、以下にさらに説明されるように、それらからの出力はさらに処理されて独自の属性を見出す。

【００４９】

標準処理モジュール１１９は、すべてのエンドユーザ１３が使用するために、テキストマイニングプロセッサ１９によって提供される。標準処理モジュール１１９の例は、調査作業の専門性を高めるために、（ｉ）記事の、セクション、段落、文、および単語の階層へのトークン分解（すなわち、分解または分裂）、（ｉｉ）スピーチのタグ付け（すなわち、単語を名詞や動詞などとして識別すること）の一部、（ｉｉｉ）引例の抜粋（すなわち、記事の書誌的事項を記事のメタデータまたは記事の参照のリストに変換すること）、ならびに（ｉｖ）因子の抜粋（すなわち、ＨＯＸＡ１、ＢＲＣＡ１などのＨＵＧＯ　ｇｅｎｅ　ｎｏｍｅｎｃｌａｔｕｒｅ　ｓｙｓｔｅｍによる、記事の単語形式のタグ付け）を含む。

【００５０】

カスタム処理モジュール１２１は、繰り返し使用するために、特定のエンドユーザ１３により構築され、モジュールのアプリケーションプログラミングインターフェイス（ＡＰＩ）に係るプログラムとして実施される。本発明の特徴として、カスタム処理モジュール１２１は、その構築を担うエンドユーザにより個人的使用のために保持されるか、匿名または記名形式ですべてのエンドユーザ１３による広範囲の使用のために発行される。多くの顧客に頻繁に利用されるカスタム処理モジュール１２１により、その創作者に特別な特典または金銭的利益を与えられる場合があることが理解される。

【００５１】

文献の最初の集合がテキストマイニング処理モジュール１１９および１２１により、分解され、タグが付され、および／または変換されると、次いで、独特の、ユーザにより特定された属性が識別されてデータセット１２３を形成する。次いで、データセット１２３は、以下にさらに詳細に説明されるように、データ削減、または関連データを並行してクラスタにする集合処理ステップ１２５の際に更に削減される。

【００５２】

データ削減ステップ１２５は、標準データセット処理モジュール１２７およびカスタムデータセット処理モジュール１２９へのアクセスにより、それぞれモジュール１１９および１２１を増大して、標準データセットおよびカスタムデータセットを与える。

【００５３】

標準データセットは、対になっているデータ（すなわち、名前および数値）の集合であり、したがって、任意のモジュールから名前でアクセスすることができる。標準データセットの例には、それに限定されないが、ＩＳＯ国名コード、ＨＵＧＯ　ｇｅｎｅ　ｎｏｍｅｎｃｌａｔｕｒｅ、および元素周期表が含まれる。

【００５４】

カスタムデータセットは標準データセットに似ているが、システム１１の個別のエンドユーザ１３により与えられる。カスタムモジュールのように、カスタムデータセットは個人的使用のために保持されるか、または、匿名か記名のいずれかでシステムのすべてのエンドユーザ１３による使用のために発行される。再度、多くの顧客に頻繁に利用されるカスタムデータセットは、その創作者に特別な特典または金銭的利益を与える場合があることが理解される。

【００５５】

データセット処理モジュール１２７および１２９は、パイプライン、またはクラスタに結合される。モジュール１２７および１２９の出力は、別のデータセット処理モジュールに直接流れることができる。または、いくつかのデータセット処理モジュールの出力が集合およびフィルタリング操作を使用して結合されることができる。

【００５６】
ステップ１２５における関連するデータの並列クラスタリングが完了すると、報告ステップ１３１の一部として、テキストマイニング操作の結果がユーザ１３に報告される。報告ステップ１３１において、標準報告モジュール１３３およびカスタム報告モジュール１３５が、テキストマイニング操作から最も適切と思われる文献の書誌データを生成する。この書誌データは収納庫２１内に導出データセットとして格納される。この導出データセットはこの後、プロジェクトマネージャ１５を介する調査の過程の中で、エンドユーザ１３によって引き出され、検査することが可能である。
【００５７】
コンテンツ選択装置５５の費用計算モジュール
上記で簡潔に参照したように、コンテンツ選択装置５５は、エンドユーザ１３に、テキストマイニングのために文献の最適な集合を引き出すことを確実にする、相互作用するコンテンツ選択プロセスに従事することを可能にする。本発明の特徴として、コンテンツ選択装置５５は新規の費用計算モジュールを使用したフルテキストおよびメタデータの問合せから引き出された文献の最初の集団を改善または最適化することができる。言い換えると、記事にアクセスする費用を考慮しつつ（すなわち、ユーザがどの記事を購読しているか、検索経費の最大値がいくらか、など）、コンテンツ選択装置５５は、エンドユーザ１３が記事のプールを（たとえば、特定のキーワードに基づいて、記事の言語によって、および／または特定の著作者によって）選択することを可能にするようにプログラムされている。
【００５８】
認識できるように、費用ベースの文献の集合の選択は、調査者にかなりの経済的負担を課し得る。より詳細には、文献収納庫１７は、ユーザ１３が購読していないが、必要なアクセス費用を支払うことで利用可能である非常に多くの記事のテキストを含むか、アクセスできることが好ましい。しかし、在来のテキストマイニングプロセスが通常、エンドユーザに、著作者が読むことを望むか、望み得るよりも多くの文献のアクセスを提供することを考慮すると、的確さが不十分である文献選択の問合せは、実施するには費用が法外になる場合がある。
【００５９】
したがって、コンテンツ選択装置５５には、とりわけ、さらなる検索の制約が存在する一方で、各テキストマイニングジョブのためのコンテンツ費用の最大値を設定し、それを履行するのに使用できる費用計算モジュールが提供される。
【００６０】
テキストマイニングジョブのためのコンテンツ費用の最大値を設定するために、コンテンツ選択装置２５により以下の式が利用され得る。

$$\sum^{n} F(d_i), \qquad (1)$$

式中、ｎは集合の中の文献の数、Ｆ（ｄ）は、発行テーブル３９の例示的スキームで定められたように（すなわち、既存の記事の購読／購入を除いて）、各文献ｄを得るための費用を定める関数である。
【００６１】
しかし、式（１）は、ユーザがすでにアクセスする権利が与えられている文献を考慮していない。文献の異なる出所（すなわち、ソース）が異なる平均金額を付けるが、同時に、すべての出所がすべての文献を差し出すわけではないことを考慮することも、有用である。たとえば、文献は（ｉ）ユーザが現在購読をしている出所から費用無しで、（ｉｉ）ＪＳＴＯＲ（商標）電子図書館などの公共の文献収蔵庫から低い定額で、および、（ｉｉｉ）個人の発行者から比較的高額で、使用可能となり得る。したがって、コンテンツ選択装置５５によって利用される、より有用な費用計算式には、以下に示すように、すべての使用可能な出所から取る場合に、各記事のすべての異なる費用の合計を考慮する。

$$\sum_{j}^{p}\sum_{i}^{n} F_j(d_i) \qquad (2)$$

式中、ｎは集合の中の文献の数、Ｆ（ｄ）は、発行テーブル３９の例示的スキームで定められたように、各出所ｊから各文献ｄを得るための費用を定める関数である。

【００６２】

式（２）を利用して、以下に示すように、問合せセットの制約を足し合わせることで、テキストマイニングジョブのためのコンテンツ費用、または経費（予算）の最大値Ｂが定められる。

$$\sum_{j}^{p}\sum_{i}^{n} F_j(d_i) < B \qquad (3)$$

【００６３】

任意選択的に、テキストマイニング調査は、例外を減らすし、別の方法で結果の統計上の信頼性を上げるために、選択された調査文献のプールを最大化することを試みる。経費の制約を満たす１つの方法は、同時に、文献の集団を最大化しつつ、費用を上げることで集合の中の記事をソートすることである。記事はその後、順に、記事の集められたセットが既定の経費に達するまで選択される。

【００６４】

しかし、上述のような、費用が高くなる選択プロセスを利用することは、特に、文献毎の費用が明らかに異なる多くのプールで構成される多数の文献の場合、多くの調査ジョブの要求に対して極めて不十分である。最も明白には、経費が制約された選択では、無料のコンテンツ、ユーザによって購読されるコンテンツ、および公共の文献収蔵庫のより古いコンテンツが強く重視され、したがって、信頼性および関連性が低い大量の文献を含む検索結果が与えられる。

【００６５】

したがって、本発明は、コンテンツの支出の制約を尊重するが、同時に、特定の無料または低価格の出所または他のメタデータのフィールド値への不公平な割当を回避する文献の集団を特定し、選択するための機構を含む。

【００６６】

本明細書において既定されるように、「コンテンツ展開」は、文献の集団が、出所によってなど、特定の条件を満たすものの中に広く分配された範囲を示す。たとえば、無料と支払いがされたものとの両方を含み、および様々な異なる発行者からの集合を伴う、多くの異なるソースからの公正な代表を含む調査文献の集団は、比較的広い、または広域のコンテンツ展開を考慮する。

【００６７】

コンテンツ選択装置５５により最初の文献の集合が完了したが、対応するテキストマイニングジョブの実際のスケジューリングおよび実施の前に、コンテンツ選択装置５５は、様々な予め既定されたメトリック、またはルールを使用してコンテンツ展開を算定する。したがって、コンテンツ選択装置５５は、１つまたは複数のユーザインターフェイス（ＵＩ）のレビュースクリーンを通して、算定されたコンテンツ展開を表示する。この方式で、エンドユーザ１３は様々なメトリック（たとえば、費用、ソースなど）にわたるコンテンツ展開を分析することができ、必要であれば、テキストマイニング操作のスケジューリングの前に、調整された文献の集合のセットを与えるように、検索パラメータを変更する。

【００６８】

コンテンツ展開メトリックは、構成可能な警告の閾値、およびユーザへのメッセージ表示をサポートして、それに続くテキストマイニング操作の際に最適化された文献の集合が利用されることを確実にすることができる。さらに、ユーザは属性、および合計または平均などの集計機能を選択することで、集合の中の文献の様々な異なる属性の中のコンテン

ツ展開を調査することができる。したがって、コンテンツ選択装置５５は、セットの要素にわたる集計を計算する。

【００６９】

ここで図６を参照すると、コンテンツ展開メトリックのそれぞれ、および展開メトリックのルールのそれぞれの結果を実施および表示する手段に関連する、既定、またはルールの柔軟な性質をサポートする例示的なデータモデルが示されている。このデータモデルは、全体として参照符号２１１で識別される。見て取れるように、展開のメトリックテーブル２１３のそれぞれは、複数の変更可能なルール２１５によって既定され、それによりユーザが（閾値テーブル２１７を介して）閾値を使用して展開メトリックを創作して、特定のコンテンツ選択方法に対処することを可能にする。したがって、変更可能なルール２１５のそれぞれは、ユーザに、実施された展開メトリックルールのそれぞれを（たとえば、リスト、円グラフ、線グラフ、および／または単一の数値で）表示する好ましい手段を確立することを可能にする。

【００７０】

コンテンツ選択装置５５による展開メトリックルールの利用は、複数ステップのプロセスを要する。プロセスの第１のステップでは、適切と考えられれば、変更に利用可能であるメトリックに実施される各ルールの規定により、エンドユーザ１３が、コンテンツ選択プロセスの際に利用される、関連する展開メトリックを選択する。展開メトリックテーブル２１３は好ましくは、エンドユーザ１３に利用可能なすべての展開メトリックを列挙する。

【００７１】

特定の展開メトリックの選択がされると、必要であれば、その展開メトリックのための対応する展開メトリックルールが試験および変更のために利用可能になる。展開メトリックルールを既定するための例示的な擬似コードを以下に示す。

```
return true
If count(article) > 1000
    return true
If metric-columns includes-any
(article.author, article.author.institution)
    return true
```

【００７２】

各展開メトリックテーブル２１３の関連性表示欄は、テキストマイニングジョブの既定に対して実行されて、所与の展開メトリックの関連性について「ｔｒｕｅ」または「ｆａｌｓｅ」の値を返すプログラムコードを含む。言い換えると、上記で与えられたルールの第１段階に基づいて、「ｔｒｕｅ」の値が、そのルールに関連性がある、またはそのルールが適用されるべきであることを示す。

【００７３】

ルールの第２段階では、そのルールのパラメータが既定される。提示の例では、１０００より多くの記事がコンテンツ展開の中に存在するかが定められる。このルールは、ジョブの既定に対して実行される集計機能に基づき、関連すると見なされる。

【００７４】

ルールの第３段階では、測定の属性が定められる。その後、前述のプロセスがすべての実行される展開メトリックのルール（すなわち、「ｔｒｕｅ」として識別される関連性の表示を有する各ルールに対して繰り返される。

【００７５】

プロセスの第２のステップでは、すべての関連する展開メトリック（すなわち、コンテンツ選択プロセスに適用されるメトリック）がコンテンツ選択装置５５により引き出されることで、それに従って実行される。なお、所与の展開メトリックは、１つまたは複数の展開メトリックルールを含むことができる。

【００７６】
このルールの表示欄は、ジョブの既定およびその関連する文献の集合に対して実行され得るプログラムコードを含む。例示的な擬似コードを以下に示す。

```
Ｓｅｌｅｃｔ ａｒｔｉｃｌｅ．ｐｕｂｌｉｃａｔｉｏｎ．ｏｒｉｇｉｎ，
    ｃｏｕｎｔ（ｄｉｓｔｉｎｃｔ ａｒｔｉｃｌｅ．ｐｕｂｌｉｃａｔｉｏｎ．ｏｒｉｇｉｎ）
        ／ｃｏｕｎｔ（ａｒｔｉｃｌｅ）
ｆｒｏｍ ｊｏｂ．ａｒｔｉｃｌｅｓ
```

【００７７】
上記の例示的コードでは、文献のソースのリストが、集団全体におけるその割合によりソートされ、それに応じて表示される。これにより、調査者が、特定のジョブのための文献の集合の中で特定の記事のソースが大きな比率を占めているかを判断することができる。

【００７８】
さらなる例示的な擬似コードを以下に示す。

```
Ｓｅｌｅｃｔ ｓｕｍ（ａｒｔｉｃｌｅ．ｐｕｂｌｉｃａｔｉｏｎ．ｐｒｉｃｅ）
ｆｒｏｍ ｊｏｂ．ａｒｔｉｃｌｅｓ
```

【００７９】
上記の例示的コードでは、特定のジョブに含まれる記事のためのコンテンツ取得費用全体がユーザに向けて表示される。

【００８０】
このプロセスの最終ステップでは、実施された展開メトリックのそれぞれのリンクが表示されることで、ユーザが展開メトリックルールに示す表示方法にしたがって結果を精査することができる。一例として、円グラフの表示方法により、ルールが、割合として解釈され得る｛記事名、記事の値｝の対のリストを返すことが示される。別の例として、単一の値の表示方法により、ルールが、メッセージの属性（たとえば、Ｃ言語のストリングである「Ｔｈｅ ｔｏｔａｌ ｃｏｓｔ ｏｆ ｔｈｅ ｊｏｂ ｉｓ ％ｄ」、％ｄのパラメータは表示のために、ルールの表示によって返される値と入れ替えられる）と合わせられ得る単一の値を返すことが示される。

【００８１】
上述のジョブの集合のコンテンツ選択プロセスは、強制されたプログラミングまたは最適化技術を使用して達成され得ることが理解される。したがって、知識を有する専門家が、シンプレックス、ミニマックス、および非線形反復法を含む様々な数学上の最適化の方法を利用して、最適に文献収蔵庫１９からコンテンツを選択できる。

【００８２】
テキストマイニングシステム１１および方法１１１の例示的使用
ここで図７（ａ）－（ｅ）を参照すると、本発明の原理を理解するのに有用である一連のサンプルスクリーン表示が示されている。

【００８３】
上記で参照したように、方法１１１の第１ステップ１１３は、エンドユーザ１３がテキストマイニングジョブを既定することを要する。ステップ１１５で集められる記事の選択を補助するために、システム１１は、コンテンツを選択するためのユーザインターフェイスを生成する。ユーザインターフェイスの例示的なスクリーン表示が図７（ａ）に示されており、全体として参照符号３１１で識別される。

【００８４】
見て取れるように、コンテンツ選択ユーザインターフェイス３１１は、新しい、または以前に規定されたテキストマイニングプロジェクトへのアクセスを提供する複数のタブ３１３－１および３１３－２を含む。各プロジェクトスクリーンは、ジョブを識別するためのプロジェクト名ウインドウ３１５、ジョブの範囲を簡潔にまとめるための解説ウインドウ３１７、コンテンツ選択プロセスで使用されるキーワードを入力のためのキーワードウ

インドウ３１９、コンテンツ選択プロセスから選択された著作者を含むか引き出すための著作者ウインドウ３２１、コンテンツ選択プロセスから選択された発行者を含むか引き出すための発行者ウインドウ３２３、および既定の期間内に発行された記事にコンテンツ選択プロセスを制限するための期日ウインドウ３２５を含む。ともに、スクリーン３１１に提供される様々な検索パラメータ、または要素がコンテンツ選択装置５５に渡されてテキストマイニングジョブのために記事の集合を取り込む。

【００８５】

なお、コンテンツ選択ユーザインターフェイス３１１にはさらに、ユーザが特定のテキストマイニング処理属性を選択および変更することを可能にする属性セットドロップダウンウインドウ３２７が与えられる。たとえば、ウインドウ３２７の「ｖａｌｕｅ」の語をクリックすると、エンドユーザ１３は、テキストマイニング操作のための検索費用キャップが発効され得る別のスクリーンに移される。

【００８６】

具体的には、ここで図７（ｂ）を参照すると、コンテンツ展開の範囲を設定するためのユーザインターフェイスのサンプルスクリーン表示が示されている。この例示的スクリーン表示は全体として参照符号３３１で識別される。見て取れるように、様々な費用に関連するルールが文献選択ステップ１１５に包含され得る。ユーザインターフェイス３３１を通して、エンドユーザ１３はリストからルールを選択することで費用の範囲を設定することができ、したがって、ルールのために返された値に対して実行される式を特定する。

【００８７】

たとえば、第一ルール３３３では、式は結果となる最大値が５０になることを示す。言い換えると、記事の集団全体の５０％より多くを構成するソースは無い。コンテンツ選択ステップ１１５の実行中、コンテンツ選択装置５５は集合のための記事の選択を制限して特定された範囲を受け入れる（すなわち、単一の記事のコンテンツのホットスポットを防止する）。この制限は、したがって、集合に現れる文献数の合計に影響する。

【００８８】

第二のルール３３５では、式が、このルールによって計算される記事の費用全体が１０００ドルを超えないだろうことを示している。コンテンツ選択ステップ１１５の実行中、コンテンツ選択装置５５は集合のための記事の選択を制限して、記事の費用全体がこの値を超えないようにしている。この制限により、したがって、集合における記事のソースの関連表示と記事の合計数との両方に影響を与える。

【００８９】

なお、ジョブのためのコンテンツ展開の範囲のすべてがそれに追従して実行されなければならない。たとえば、上記に与えられた例を使用して、コンテンツの選択は、（ｉ）どの１つのソースも記事の５０％を超えて構成しないように、様々なソースからの記事で構成しなければならず、また、（ｉｉ）調査者のアクセス費を伴う記事（すなわち、ユーザが購読しておらず、または、公共に無料で利用可能ではない記事）を取得するために必要な支出が１０００ドル以下でなければならない。

【００９０】

なお、また、上記に示されたルールは、可能性のあるコンテンツ展開の範囲のルールの一例でしかない。他のタイプのコンテンツ展開の範囲のルールが、本発明の要旨から逸脱することなく、同様に既定および利用され得る。

【００９１】

なお、さらに、本明細書ではコンテンツの費用がドルで示されているが、技術を有する専門家が、本発明の要旨を逸脱することなく、国際通貨および関連する通過の兌換で費用のサポートを付けることもできることが理解される。

【００９２】

様々な問合せのルールが既定されると、コンテンツ選択装置５５は、それに続くテキストマイニング操作に使用される予備的な文献の集合を選択する。エンドユーザ１３がテキストマイニングの前に予備的な文献の集合の質を評価することを可能にするために、コン

テンツ選択装置５５は、コンテンツ展開の詳細なメトリックを提供するＵＩレビュースクリーンを生成する。サンプルのＵＩレビュースクリーン表示は図７（ｃ）に示され、全体として参照符号３４１で識別される。

【００９３】

例示的なスクリーン表示３４１では、示されたソースのコンテンツ展開が、コンテンツ展開を視覚化するのに有用であるテーブル、またはリスト３４３、および円グラフ３４５として提供されている。見て取れるように、集められたれたコンテンツの４２％が単一のソース（無料のソースであるＰｕｂＭｅｄ）から得られる。さらに、集められたコンテンツの約７０％が、両方とも無料のソースである、上位２つのソース（ＰｕｂＭｅｄとＰＬｏＳ）から得られる。

【００９４】

このことから、ユーザ１３はコンテンツ展開が非常に狭い（すなわち、十分なソースが適切に表示されていない）という結論にすぐに達することができる。この結果は、ユーザ１５に（ｉ）ソース数が少ないこと、および、（ｉｉ）単一のソースが多くを占めていることを知らせる警告３４７によりサポートされる。

【００９５】

ユーザにより、コンテンツ展開が非常に狭く、この理由は、他のことの中で、経費が非常に制限されているためだと判断される場合がある。その結果、ユーザはより良いコンテンツ展開を得るためにコンテンツの費用を増やすことを決める場合がある。

【００９６】

ユーザにより、コンテンツ展開が非常に狭く、この理由は、他のことの中で、問合せが非常に幅広く、このため最初の文献のプールをかなり多く得たためであると判断する場合もある。その結果、ユーザが検索パラメータの範囲を狭めることを決める場合がある。

【００９７】

本明細書にはソースのコンテンツ展開が示されるが、コンテンツ展開の代替的な属性（たとえば、発行日、タイトル、出所の国、記事の言語、および費用の内訳など）が同様に、レビューのためにユーザ１３に提供されることが理解される。この相互作用する直観的プロセスを通して、エンドユーザ１３は、最終的に最適なコンテンツ展開が得られるまで（たとえば、最適化されたコンテンツ展開が予め既定された経費の範囲に入るまで）、文献の集団を変更することができる。

【００９８】

最適化されたコンテンツ展開が得られると、具体化されたスケジュールにしたがって、テキストマイニングプロセッサ１９によりテキストマイニング操作の処理ステップが実施される。完了すると、結果として得られた書誌的データが収蔵庫２１内に導出データセットとして格納される。この獲得されたデータセットは、次いで、プロジェクトマネージャ１５を介する調査の中でエンドユーザ１３により、引出し、および試験を行うことが可能である。

【００９９】

具体的には、ここで図７（ｄ）を参照すると、全体として参照符号３５１で識別された、テキストマイニングの結果のリストのサンプルスクリーン表示が示されている。見て取れるように、スクリーン表示３５１には、テキストマイニングプロジェクトの一部として識別される一連の調査文献３５３－１～３５３－５のそれぞれの情報（たとえば、書誌的データ、ユーザのアクセス費用、概要など）が含まれる。さらに、このリストに提供された各文献には、ユーザ１３に無料または既定の費用で使用可能であれば、記事のフルテキストにアクセスするリンクが含まれる。この方式で、ユーザ１３は効率的に、特定のトピックの適切な調査記事をユーザ既定の費用でアクセスおよびレビューすることが可能であり、このことは、本発明の主要な目的である。

【０１００】

定期的に、エンドユーザ１３は様々なテキストおよびデータのマイニングプロジェクトの状態を、プロジェクトマネージャ１５によって提供される適切なユーザインターフェイ

スを通してレビューおよびモニターすることができる。具体的には、ここで図７（ｅ）を参照すると、エンドユーザ１３によって始められた現在および過去のテキストマイニングプロジェクトのレビューのための、ユーザインターフェイスのサンプルスクリーン表示が示されている。この例示的スクリーン表示は全体として参照符号３６１で識別される。スクリーン表示３６１では、システム１１の、ログインしたエンドユーザ１３に利用可能である、開始されたテキストマイニングジョブのテーブル３６３が示されている。

【０１０１】

見て取れるように、エンドユーザ１３に関する様々なプロジェクトが、コンテンツ選択インターフェイス３１１を介してユーザによって事前に提供されたプロジェクト名３６５および詳細情報３６７を使用して列挙される。さらに、テーブル３６３は、各プロジェクトのための作品期日ウインドウ３６９、およびジョブの状態（すなわち、完了した、開かれた、失敗した、処理中など）をユーザに知らせるステータスウインドウ３７１を含む。さらに、ワンクリック実行ボタン３７３をクリックすることで、各ジョブに関する特定の機能を取ることができる。

【０１０２】

上記に示された実施形態は、単に例示であることが意図され、当業者であれば、本発明の要旨を逸脱することなく、多数の変形および変更を加えることができる。そのような変形および変更は、添付の特許請求の範囲に既定される本発明の範囲内にあることが意図される。

【符号の説明】

【０１０３】

１１　システム
１３　エンドユーザ
１５　プロジェクトマネージャ
１７　文献収納庫
１９　テキストマイニングプロセッサ
２１　導出データ収納庫
５５　コンテンツ選択装置

【図１】

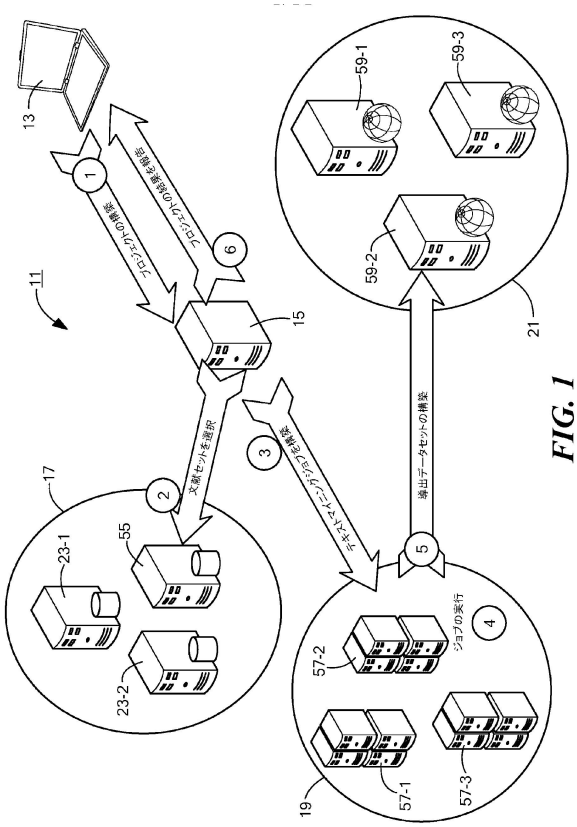


FIG. 1

【図２】

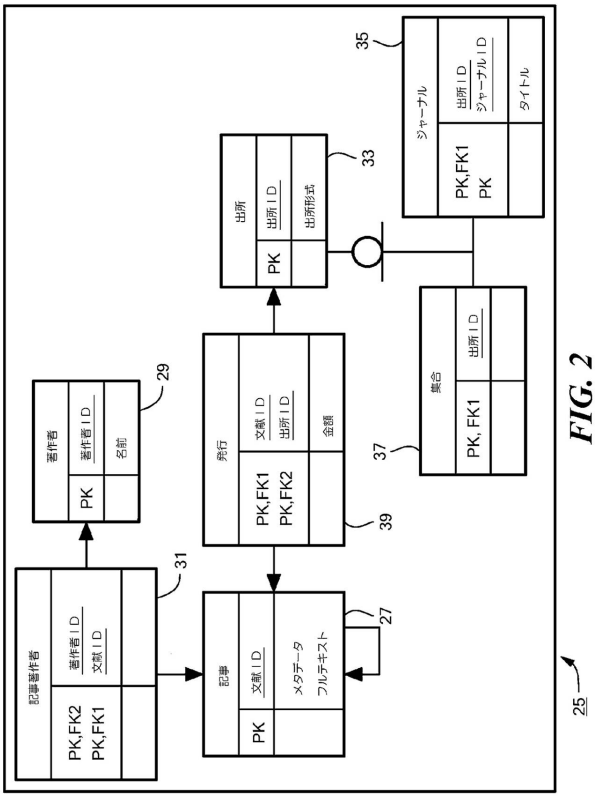


FIG. 2

【図３】

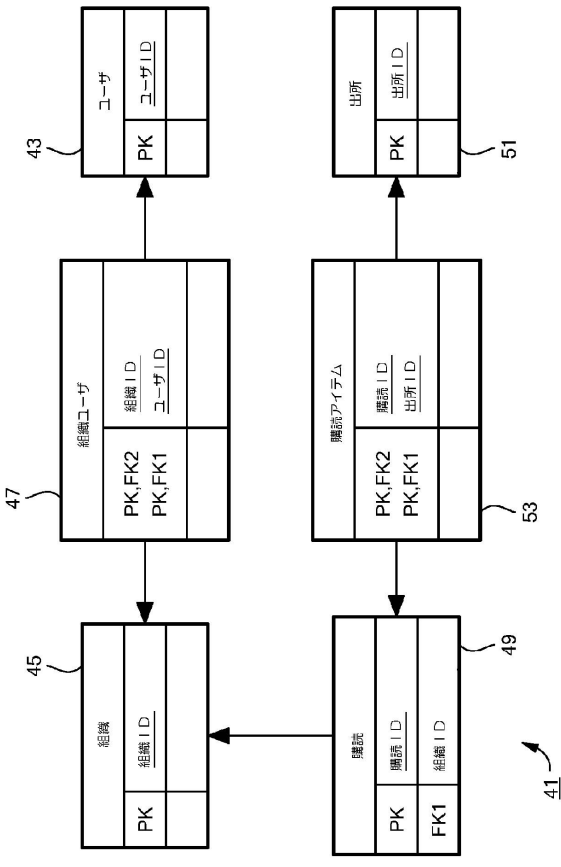


FIG. 3

【図４】

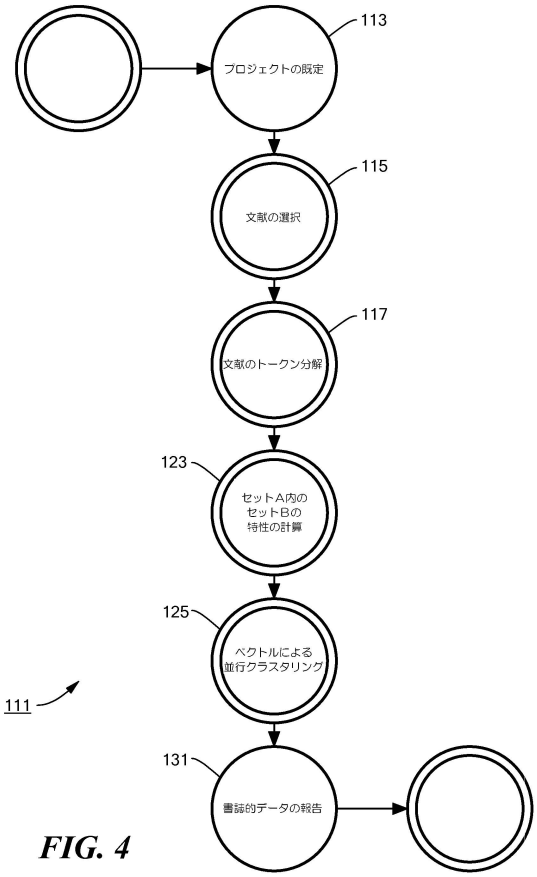


FIG. 4

【図５】

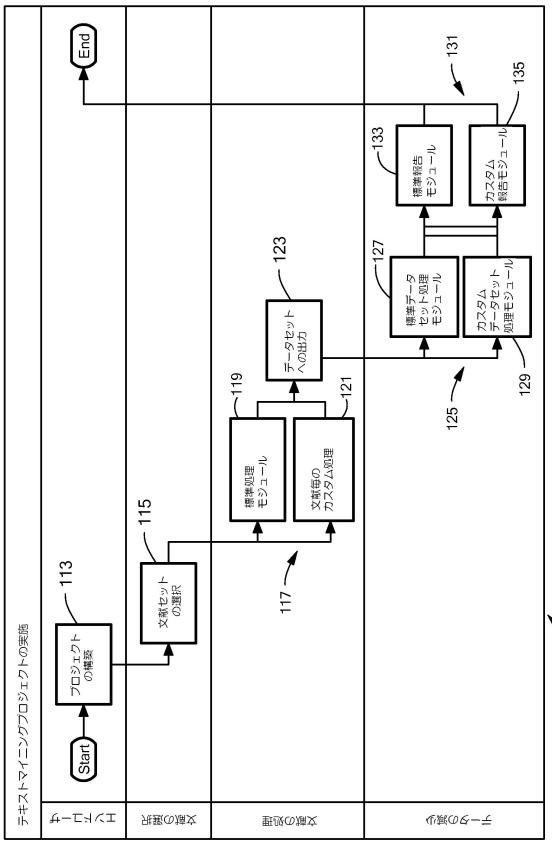


FIG. 5

【図６】

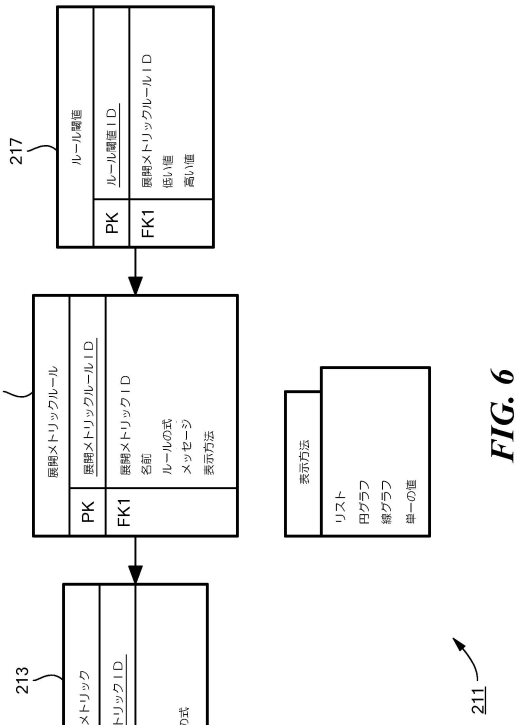


FIG. 6

【図７ａ】

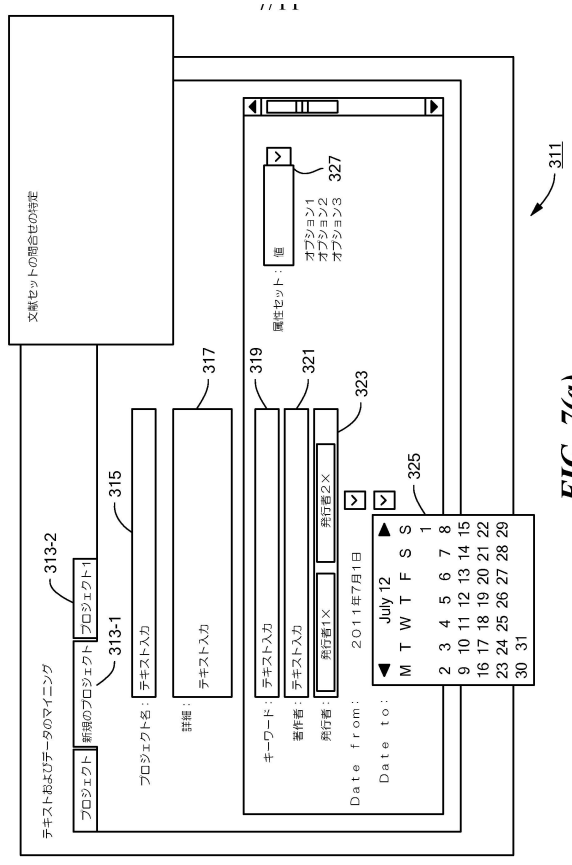


FIG. 7(a)

【図７ｂ】

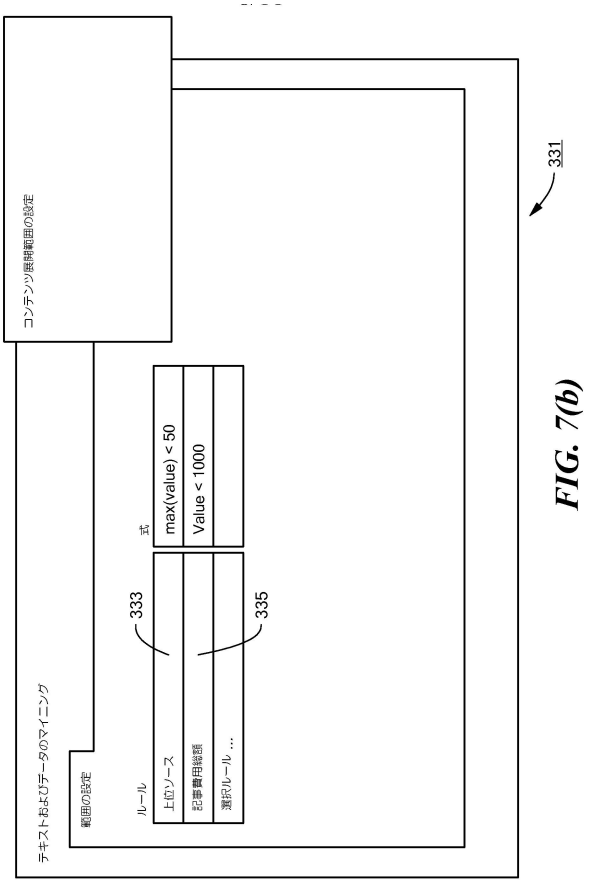


FIG. 7(b)

【図７ｃ】

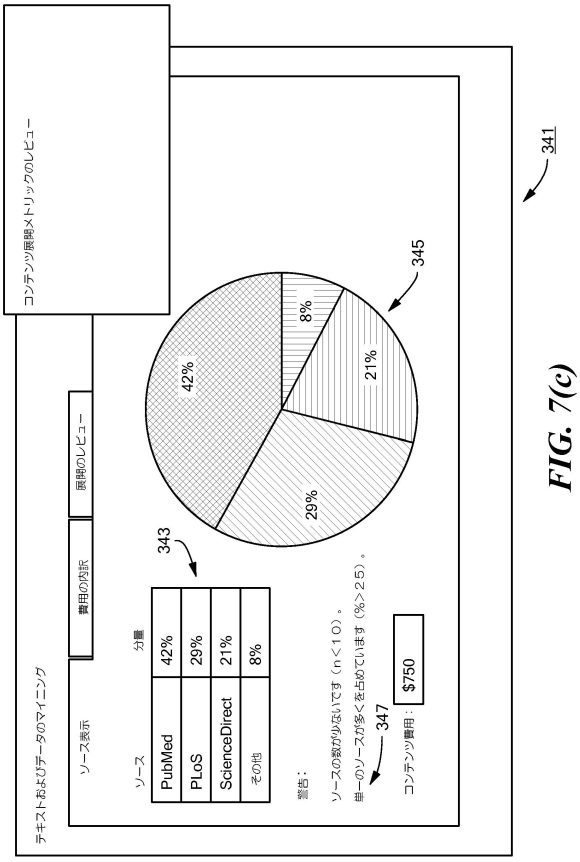


FIG. 7(c)

【図７ｄ】

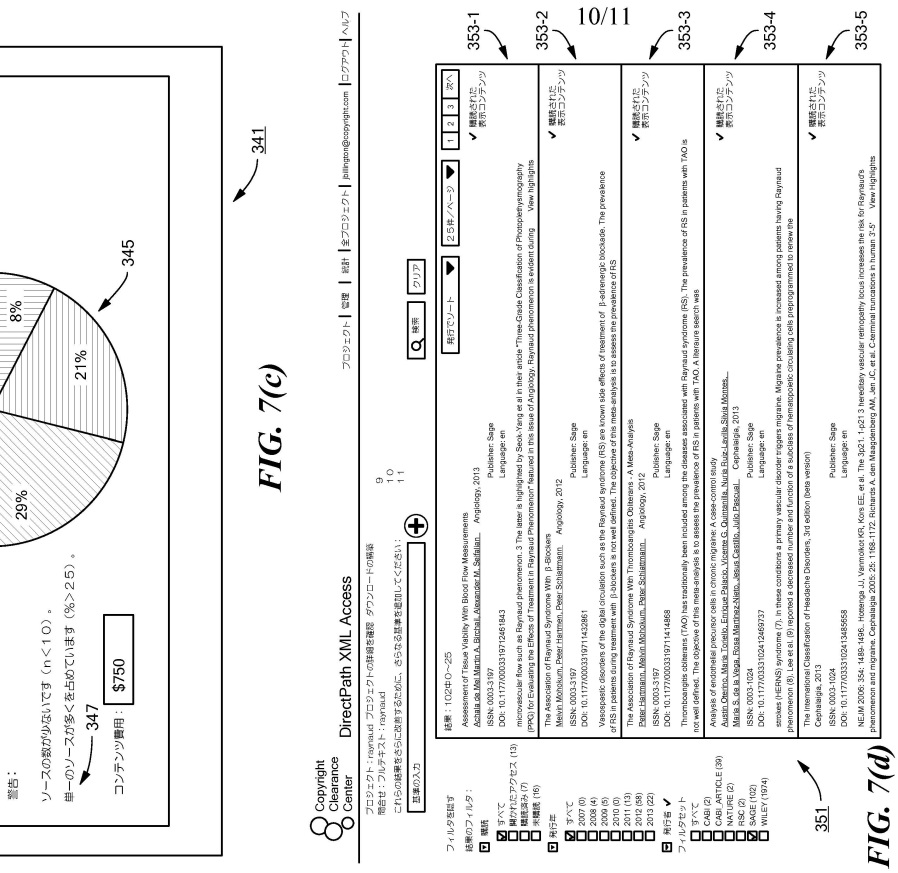
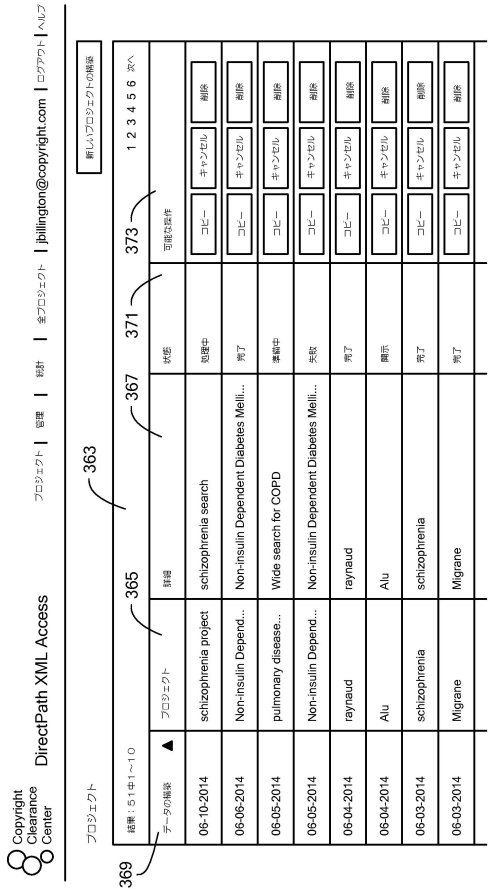


FIG. 7(d)

【図７ｅ】



データの編集	プロジェクト	詳細	状態	可能な操作
06-10-2014	schizophrenia project	schizophrenia search	処理中	コピー キャンセル 削除
06-06-2014	Non-insulin Depend...	Non-insulin Dependent Diabetes Melli...	完了	コピー キャンセル 削除
06-05-2014	pulmonary disease...	Wide search for COPD	準備中	コピー キャンセル 削除
06-05-2014	Non-insulin Depend...	Non-insulin Dependent Diabetes Melli...	失敗	コピー キャンセル 削除
06-04-2014	raynaud	raynaud	完了	コピー キャンセル 削除
06-04-2014	Alu	Alu	開示	コピー キャンセル 削除
06-03-2014	schizophrenia	schizophrenia	完了	コピー キャンセル 削除
06-03-2014	Migrane	Migrane	完了	コピー キャンセル 削除

FIG. 7(e)

フロントページの続き

(72)発明者　クレベ，スコット
アメリカ合衆国　マサチューセッツ州　０２４２０，レキシントン，８　ハンコック　アベニュー
(72)発明者　ビリントン，ジョン
アメリカ合衆国　マサチューセッツ州　０１８３３，ジョージタウン，４２　プロスペクト　ストリート

審査官　松尾　真人

(56)参考文献　特開２００３－２１６６４５（ＪＰ，Ａ）
特開２００９－１２３１３９（ＪＰ，Ａ）
米国特許出願公開第２０１２／０２２１５５３（ＵＳ，Ａ１）
特開２００８－１３５０５７（ＪＰ，Ａ）
特表２００７－５１７３４３（ＪＰ，Ａ）
米国特許出願公開第２０１２／００８９６４２（ＵＳ，Ａ１）
特表２００８－５４２９２６（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０