(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5202524号
(P5202524)
(45)発行日　平成25年6月5日(2013.6.5)　(24)登録日　平成25年2月22日(2013.2.22)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 220A
G06F 17/21 (2006.01)　G06F 17/30 170A
G06F 17/30 380E
G06F 17/21 550A

請求項の数 20　(全 19 頁)

(21)出願番号　特願2009-522777 (P2009-522777)
(86)(22)出願日　平成19年7月20日 (2007.7.20)
(65)公表番号　特表2009-545808 (P2009-545808A)
(43)公表日　平成21年12月24日 (2009.12.24)
(86)国際出願番号　PCT/US2007/016435
(87)国際公開番号　WO2008/016491
(87)国際公開日　平成20年2月7日 (2008.2.7)
審査請求日　平成22年7月20日 (2010.7.20)
(31)優先権主張番号　11/496,650
(32)優先日　平成18年7月31日 (2006.7.31)
(33)優先権主張国　米国(US)

(73)特許権者　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100077481
弁理士　谷　義一
(74)代理人　100088915
弁理士　阿部　和夫
(72)発明者　サリハ　アザム
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション　インターナショナル　パテンツ内

最終頁に続く

(54)【発明の名称】多段アプローチを使用した事実の抽出の最適化

(57)【特許請求の範囲】
【請求項１】
電子リソース内で意見と事実との区別をプロセッサにより実施するコンピュータ実装方法であって、
検索用語を受け取ること、
前記検索用語に一致する関連する電子リソースを発見すること、
前記検索用語に一致する単語を含む前記関連する電子リソースのリスト及び前記リスト内の前記電子リソースの断片を表示すること、
前記検索用語と、事実の表現を示唆するように判定された動詞のリストを含むように構成された事実－単語表に一致する１以上の動詞とからなる文書の事実の記述を発見するために、関連する電子リソースをスキャンすること、
前記関連する電子リソースの部分であって、前記検索用語と前記事実－単語表の単語と一致しない単語を含む部分を事実の抽出処理から、削除すること、
前記関連する電子リソースの部分を削除することの後に、前記事実の記述の言語成分を識別するために、前記発見した事実の記述を調査すること、
前記識別した言語成分に基づいた事実として事実の記述を提示するか否かを決定すること、
前記検索用語と、前記検索用語に関連する事実であると判定された事実の記述とを含む文書の少なくとも一部を表現すること
から成ることを特徴とする方法。

【請求項２】
前記識別した言語成分に基づいた事実として事実の記述を提示するか否かを決定することは、
ある事実の記述を考慮からはずすために、前記事実の記述の前記言語成分に関する除外規則を適用すること、
前記事実の記述にスコアをつけること、
考慮するために残っている事実の記述の各々の前記スコアを閾値と比較すること、
前記閾値を超えるスコアを有する事実の記述の各々に対して、事実として前記事実の記述を含む文章の少なくとも一部を提示すること
から成ることを特徴とする請求項１に記載の方法。
【請求項３】
更に会話の部分と共に前記事実の記載の単語にタグを付けることから成ることを特徴とする請求項２に記載の方法。
【請求項４】
会話の部分と共に前記事実の記載の単語にタグを付けることは、単語が動詞か名詞のいずれかであるとき、名詞タグを適用することから成ることを特徴とする請求項３に記載の方法。
【請求項５】
前記除外規則を適用することは、主語の役目を有する統語上の句のための規則の第１のセットを適用することと、目的語の役目を有する統語上の句のための規則の第２のセットを適用することから成ることを特徴とする請求項４に記載の方法。
【請求項６】
規則の前記第１のセットを適用することは、主語又は目的語の意見又は偏った修飾語句を有する名詞句を除外することから成ることを特徴とする請求項５に記載の方法。
【請求項７】
規則の前記第２のセットを適用することは、
固有名詞でない限定記述を含む主語名詞句を除外すること、
代名詞を含む名詞句を除外すること、文書の冒頭に現れない主語名詞句を除外することから成ることを特徴とする請求項５に記載の方法。
【請求項８】
前記名詞句の役目に関係なく、更に規則の第３のセットを適用することから成ることを特徴とする請求項５に記載の方法。
【請求項９】
規則の前記第３のセットを適用することは、前記文章の句読点が疑問符である事実の記述を除外することと、ストップワードを含む句を有する文章を除外することから成ることを特徴とする請求項８に記載の方法。
【請求項１０】
前記事実の記述にスコアを付けることは、前記除外規則の適用後、又は適用中のどちらかに考慮するために残っているこれらの事実の記述だけにスコアを付けることから成ることを特徴とする請求項２に記載の方法。
【請求項１１】
コンピュータストレージ媒体であって、
検索用語を受け取ること、
前記検索用語に一致する関連する電子リソースを発見すること、
前記検索用語に一致する単語を含む前記関連する電子リソースのリスト及び前記リスト内の前記電子リソースの断片を表示すること、
前記検索用語と、事実の表現を示唆するように判定された動詞のリストを含むように構成された事実－単語表の単語に一致する１以上の動詞とからなる文書の事実の記述を発見するために、複数の関連する電子文書を構文解析すること、
前記関連する電子リソースの部分であって、前記検索用語と前記事実－単語表の単語と

一致しない単語を含む部分を事実の抽出処理から、削除すること、
前記関連する電子文書の部分を削除することの後に、前記事実の記述の言語成分を識別するために、前記発見した事実の記述を調査すること、
前記言語成分に関する候補となる事実の記述に除外規則を適用することにより、前記識別した言語成分に基づいた前記検索用語に関連する事実として事実の記述を提示するか否かを決定すること、
一致する事実－単語表に基づき、かつ、主語と目的語の個々の重みに基づき、候補となる事実の記述をスコアリングすること、
前記除外規則および事実の記述のスコアリングに従って、前記候補となる事実の記述を考慮からはずすこと、
前記検索用語と、前記検索用語に関連する事実であると判定された事実の記述とを含む文書の少なくとも一部を表現すること
から成ることを特徴とする行為を、プロセッサに実行させる実行可能プログラム命令を含むコンピュータストレージ媒体。
【請求項１２】
前記行為は、さらに電子文書の集合を検索して、前記検索用語を含むこれらの文書を発見することにより前記複数の文書を得ることから成り、
前記集合は、前記複数の電子文書を解析する前に前記検索用語を含むこれらの文書を発見するために検索されること
を特徴とする請求項１１に記載のコンピュータストレージ媒体。
【請求項１３】
前記行為は、さらに前記電子文書を入手して前記検索用語を受け取る前に事実の記述を提示すること、又前記電子文書と事実の記述を検索して、これらの電子文書と前記検索用語に関連する対応する事実の記述を見つけることから成ることを特徴とする請求項１１に記載のコンピュータストレージ媒体。
【請求項１４】
前記行為は、さらに考慮するために残っている事実の記述の各々の前記スコアを閾値に対して比較すること、
前記検索用語を含み、前記閾値を超えるスコアを有する電子文書から取られた事実の記述の各々に対して、前記検索用語に関連する事実として前記事実の記述を含む前記文章の少なくとも一部を提示することから成ることを特徴とする請求項１１に記載のコンピュータストレージ媒体。
【請求項１５】
前記事実の記述にスコアを付けることは、前記除外規則を適用した後に考慮するために残っているこれらの事実の記述にだけスコアをつけること
から成ることを特徴とする請求項１４に記載のコンピュータストレージ媒体。
【請求項１６】
本文情報から成る複数の電子リソースを含むストレージと、
プロセッサとから成るコンピュータシステムであって、
前記プロセッサは、検索用語を受け取り、前記検索用語と一致する関連する電子リソースを発見し、前記検索用語に一致する単語を含む前記関連する電子リソースのリスト及び前記リスト内の前記電子リソースの断片を表示し、電子文書のセットから前記検索用語に関する事実を提示するための要求を受け取り、前記検索用語と、事実の表現を示唆するように判定された動詞のリストを含むように構成された事実－単語表の単語に一致する１以上の動詞とからなる文書の事実の記述を発見するために、前記関連する電子文書を構文解析し、前記関連する電子リソースの部分であって、前記検索用語と前記事実－単語表の単語と一致しない単語を含む部分を事実の抽出処理から、削除すること、前記関連する電子文書の部分を削除した後に、前記事実の記述の言語成分を識別するために、前記発見した事実の記述を調査し、前記識別した言語成分に基づいた事実として事実の記述を提示するか否かを決定し、前記事実として提示されると判定された事実の記述と、前記検索用語に

関連する事実の記述とを含む文書の少なくとも一部を表現する
ことを特徴とするコンピュータシステム。
【請求項１７】
表示装置を更に備え、前記表示装置上に前記文章の少なくとも前記部分を表示することにより、前記プロセッサが前記文章の少なくとも前記部分を提示することを特徴とする請求項１６に記載のコンピュータシステム。
【請求項１８】
ネットワークインタフェースを更に備え、前記ネットワークインタフェースを介してこれらの部分を他のコンピュータに出力することにより、前記プロセッサが前記文章の少なくとも前記部分を提示することを特徴とする請求項１６に記載のコンピュータシステム。
【請求項１９】
ネットワークインタフェースを更に備え、前記ストレージは前記ネットワークインタフェースを介して前記プロセッサによりアクセス可能なことを特徴とする請求項１６に記載のコンピュータシステム。
【請求項２０】
前記事実の記述の前記言語成分に関連して除外規則を適用して前記事実の記述の一部を考慮から外すこと、
前記事実の記述にスコアを付けること、
閾値に対して考慮するために残存する事実の記述の各々のスコアを比較すること、
前記検索用語を含み、前記閾値を超えるスコアを有する事実の記述の各々に対して、前記検索用語に関連する事実として前記事実の記述を含む前記文章の少なくとも前記部分を提示すること、
により、事実として事実の記述を提示するか否かを前記プロセッサが決定することを特徴とする請求項１６に記載のコンピュータシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、多段アプローチを使用した事実の抽出の最適化に関する。
【背景技術】
【０００２】
電子文書は、事実と意見の混合体を含む。読者は、事実に興味があるだけのこともあれば、事実を識別したいだけの時もある。例えば、情報を求めてオンライン検索しているユーザは、できるだけ早く、効率的に特定の主題についての事実が欲しいと思っている。けれども、使用する検索用語に関連するウェッブページ、又は他の電子文書のリストを提示するためには、ユーザは個別にウェッブページ、又は他の電子文書の各々を調べて、事実と意見、又は主題となる情報を区別することが必要になる。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
事実の抽出をしようとする試みがなされてきたが、サーバコンピュータがいくら速くても、正確な事実の抽出には時間がかかり、効率の悪いプロセスになってしまう。そのような事実を抽出する試みは、一般に言語分析を電子文書の内容全体に適用して電子文書に含まれる可能性がある事実を抽出する。事実の抽出を何百何千という電子文書に適用した場合、結果が得られるまでにかかる時間は受け入れられるものではないかもしれない。
【課題を解決するための手段】
【０００４】
多段アプローチを使用することによる事実抽出を最適化する実施形態を提供する。電子文書を読み取り、事実単語表を使用して事実を含んでいそうな事実の記述を見つけて、電子文書の文章内の用語と合わせて事実の記述のセットを入手する。文書全体の解析ではなく、その事実の記述セットの近くで、例えば統語論の成分、及び／又は意味論のような言

語成分を決定することを含む解析をさらに行う。従って、関係する電子文書毎に文書全体を、複雑に辞書的、又統語的に解析するのを避けることにより時間が節約できる。

【０００５】

この要約は、以下の発明を実施するための最良の形態で詳述する概念の選択を、簡単な形式で紹介するためのものである。この要約は、発明の主題の主要な特徴、又は本質的な特徴を識別するためのものではなく、又は発明の主題の範囲を決定するための支援として使用するためのものでもない。

【図面の簡単な説明】

【０００６】

【図１】実施形態を実装するためのコンピュータシステムの一例を示す図である。

【図２】検索前に抽出された事実の提示も含む検索の操作フローの一例を示す図である。

【図３】検索中に抽出された事実の提示も含む検索の操作フローの一例を示す図である。

【図４】事実の抽出の多段ステップの操作フローの一例を示す図である。

【図５】事実の抽出の多段ステップの更に詳細な操作フローの一例を示す図である。

【図６】検索により発見された電子文書から得られた事実の提示を含む検索結果を提供する画像表示の一例を示す図である。

【発明を実施するための形態】

【０００７】

実施形態は、関係する文書全体の複雑な解析をしないようにするために多段ステージを使用する事実抽出を提供する。文書の事実の記述は、初期段階において事実－単語表に関連して認識される。これらの事実の記述は、会話の部分、名詞か動詞かどちらかでタグ付けされる。そしてこれらの事実の記述に関して、その後の段階でさらに詳細な解析が行うことが可能で、その際関連する文書全体にわたってそのような詳細な解析を回避している。事実の記述の各々のための言語成分が決定され、除外とスコアを使用して、事実になりそうにない事実の記述を除去する。除外してスコアをつけてから、残りの事実の記述が事実として提示される。

【０００８】

図１は、実施形態の作動環境を提供するコンピュータシステム１００の一例を示す。ここに示すコンピュータシステム１００は、マスストレージ１１２、メモリ１０４、ディスプレイアダプタ１０８、そしてキーボード、キーパッド、マウス等の１つ又は複数の入力装置１１０を含む様々なコンポーネントと同様プロセッサを備えた、標準の汎用のプログラム可能なコンピュータシステム１００である。プロセッサ１０２は、データ信号バス１０６を介してコンポーネントの各々と通信する。

【０００９】

コンピュータシステム１００は、データネットワークを介してコンピュータシステム１００と他のコンピュータシステムが通信できるようにする例えば有線、又は無線接続のようなネットワークインタフェース１２４を備える。あるいは又、コンピュータシステム１００は、１つ又は複数の実施形態を実装する有線のアプリケーションに特化した装置であってもよい。

【００１０】

図１の例では、プロセッサ１０２は、オペレーティングシステム１１４の形式でマスストレージ１１２内に記憶されている命令を実装している。この例でのオペレーティングシステム１１４は、コンピュータシステム１００のコンポーネントを利用するためにその上に実装されている様々なアプリケーションの基盤を提供する。特定の状況に関連する電子文書を見つけるために、コンピュータシステム１００は、検索エンジン１１８、又は類似のアプリケーションを実装している。例えば、検索エンジン１１８は、コンピュータシステム１００のユーザから入力装置１１０を介して直接入力された検索用語を受信可能であり、又ネットワークインタフェース１２２を介して受信したリモートコンピュータのユーザにより提出された検索用語を受信することもできる。

【００１１】

検索、及び／又は事実の抽出は、ウェッブページ、標準のワープロ文書、スプレッドシート等のような原文情報を含む１つ、又は複数の電子文書のセットに関連して発生する場合がある。これらの電子文書は、電子文書セット１１６としてローカルに保存される。電子文書セット１２６を含むネットワークベースのストレージ１２４のようなローカルでない場所でも、これらの電子文書は保存されることもある。ネットワークベースのストレージ１２４は、ローカルネットワークストレージ、インターネットのオンラインストレージの場所等を表す。ネットワークベースのストレージ１２４は、ネットワークインタフェース１２２を介してアクセスが可能である。

【００１２】

その上、これらの実施形態は、電子文書１１６，１２６から事実を抽出するために、プロセッサ１０２による実装のための論理を提供する。事実抽出ツール１２０は、オペレーティングシステム１１４のコンポーネント、検索エンジン１１８又は他のアプリケーションのコンポーネントとして、又は自身の独立した結果を生成することができるスタンドアローンのアプリケーションとしてのどちらかでローカルのストレージ装置１１２上に存在することができる。事実の抽出ツール１２０の実施形態により行われる論理操作は、図２から図５に関連して以下で議論する。

【００１３】

図１のコンピュータシステム１００は、様々なコンピュータが読取り可能な媒体を備える。そのようなコンピュータが読取り可能な媒体は、コンピュータシステムを操作するための、又本明細書で議論される実施形態を実装するための命令を含む。コンピュータが読取り可能な媒体は、コンピュータ１００がアクセス可能であり、揮発性と不揮発性媒体、可搬と非可般媒体の両方を含む入手可能な媒体であれば何でも良い。制限するものでなく一例として、コンピュータが読取り可能な媒体は、コンピュータストレージ媒体と通信媒体から成る。

【００１４】

コンピュータストレージ媒体は、コンピュータが読取り可能な命令、データ構造、プログラムモジュール、又は他のデータのような情報を保存するための如何なる方法や技術にでも実装される揮発性及び不揮発性、可搬型及び非可搬型媒体を含む。コンピュータストレージ媒体は、これに限らないが、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ，フラッシュメモリ又は他のメモリ技術、ＣＤ－ＲＯＭ，ＤＶＤ又は他の光ディスクストレージ、磁気カセット、磁気テープ、磁気ディスクストレージ又は他の磁気ストレージ装置、又は必要な情報を記憶するのに使用可能な、又はコンピュータシステム１００がアクセス可能な他のどんな媒体でも含む。

【００１５】

通信媒体は、普通、コンピュータが読取り可能な命令、データ構造、プログラムモジュール、又は搬送波、又は他の輸送機構のような変調されたデータ信号内の他のデータを具体化しており、どのような情報配信媒体も含む。「変調されたデータ信号」という用語は、信号中の情報を符号化する様に設定、又は変更された一つ、又は複数のその特徴を備えた信号を意味する。制限するものでなく一例として、通信媒体は、有線ネットワーク、又は直結接続のような有線媒体、及び音響、高周波、赤外、又他の無線媒体のような無線媒体を含む。上記をどれでも組み合わせたものも、コンピュータが読取り可能な媒体の範囲内に含まれる。

【００１６】

図２は、事実の抽出ツール１２０に関連して検索エンジン１１８により行われる論理操作の一例を示す。この例では、検索する電子文書内に存在する事実のライブラリを生成するために、検索する前に事実の抽出ツール１２０が利用される。このようにして、事実を抽出するのに処理時間はかからないが、その代わりにこれらの事実は既に抽出されていて、入力された検索用語に基いて事実のライブラリから引き出されている。

【００１７】

論理操作は、電子文書の集合が得られるか、又はそうでなければアクセスが達成される

集合操作２０２で開始される。例えば、最終的に検索される電子文書は、ローカルストレージに記憶されるか、オンラインアクセスを介して入手できる。事実の抽出ツール１２０は、次いで、これらの電子文書の各々に作用して、電子文書内にある事実の全てを抽出しようとする。事実抽出ツール１２０は、対応する電子文書に関連して保存され、将来検索する際のアクセスに利用可能な事実のライブラリを生成することができる。例えば、表１は関連するそのようなライブラリを示す。

【００１８】

【表１】

電子文書	事実
www.sample1.com	事実Ａ 事実Ｂ 事実Ｃ
www.sample2.com	事実ＡＡ 事実ＢＢ 事実ＣＣ
www.sample3.com	事実ＡＡＡ

表１

【００１９】

図２の操作フローについて続けると、関連する電子文書を見つけるために、特にこれらの電子文書から関連する事実を見つけるために検索しようとするユーザは、用語操作２０６で検索用語を検索エンジン１１８に入力する。この例では、次いで、検索エンジン１１８が検索用語を探して電子文書を検索して、文書操作２０８で一致する文書を見つけ出す。検索エンジンは又、検索用語と一致する電子文書から検索用語と一致する以前に抽出した事実を見つけ出し、関連する文書、又はその文書へのリンクを、関連する事実と共に表示操作２１０で表示する。例えば、ある検索用語はwww.sample1.comで見つけることができるかもしれないし、www.sample1.comへのリンクが事実Ａと事実Ｂと一緒に表示されるように、その検索用語は、又事実Ａと事実Ｂと一致することが見つけられるかもしれない。従って、ユーザは入力した検索用語に関する事実を迅速に提供される。そのような画像表示の一例を、以下図６に関連して議論する。

【００２０】

勿論、検索するのは電子文書そのものというよりもむしろ、過去に抽出した事実の検索だけにするという選択もできる。更に、事実を含んでいる電子文書が検索用語と一致するか否かにかかわらず、環境によっては、過去に抽出した事実が検索用語と一致する場合もある。

【００２１】

図３は、事実の抽出ツール１２０に関連して、検索エンジン１１８により行われる論理操作の別の例を示す。この例では、検索によって事実が見つけられるので、電子文書中に存在する事実を見つけるため、検索中に事実の抽出ツール１２０を利用している。このようにして、事実の抽出の事前検索は必要ないし、事実のライブラリを記憶する必要もない。そのようなシナリオでは、事実の抽出ツールは文書の断片、又は概要をスキャンして迅速な結果を提供するか、又は文書全体も又スキャンされて、可能性がある事実全てを抽出する。

【００２２】

ユーザが検索用語を検索エンジン１１８に入力する用語操作３０２にて論理操作が開始される。この例では、次いで、検索エンジン１１８は検索用語を探して電子文書を検索して、文書操作３０４にて一致する文書を見つける。次いで、検索用語に関連するこれらの文書から事実を抽出するために、検索によって見つけられた電子文書を解析するために、抽出操作３０６にて抽出ツール１２０が利用される。抽出操作３０６の結果は、表１に示すように電子文書と事実の間の関係の一時的なセットを生成することがあるが、将来発生するこれら検索用語の検索を予想して、比較的長期間に亘り記憶される可能性がある。次いで、抽出操作３０６で事実の抽出ツール１２０により返された関連する事実と一緒に、

表示操作３０８にて検索エンジンは、関連する文書、又は関連する文書へのリンクを表示する。

【００２３】

図４は、事実抽出ツール１２０の実施形態が利用する多段アプローチを示す。最初は、事実の抽出ツール１２０は、認識操作４０２にて関係する電子文書からの事実の記述のセットを認識しようと試みる。ここで、図５を参照して下記で詳細に議論する事実－単語表への一致を見つけることに基いて、事実になりそうな記述をテキスト中で見つけることを目標とする。マッチングプロセスを迅速に行うことにより、事実を見つける際に無視すべき電子文書の多くが事実の抽出処理から削除できるし、その際精密度を上げるために使用されるそれに続く段階での効率を向上させることができる。

【００２４】

解析する文書のための事実の記述のセットを識別してから、次いで、抽出操作４０４でこの事実の記述のセットについて事実の抽出が行われる。ここで、文書全体ではなくむしろこの事実の記述のセットについてのみ更に詳細な解析が行われるので、十分な精度が実現される一方、満足できる効率が維持できる。抽出操作の解析は、事実の記述の言語成分の決定に基づく意思決定を含む。このような言語成分は、統語論の成分と、意味論等を具備する。

【００２５】

図５は、認識の詳細と図４の抽出操作の一例を示す。論理操作は、事実の抽出ツール１２０が電子文書をスキャンして、事実－単語表と一致する単語や句を見つけるスキャニング操作５０２で開始される。事実－単語表は、例えばある意見と反対の事実を表現する際に使われやすいと知られている単語や句のリストである。表２に簡単な例を示す。最適な処理性能を提供するためには、タグ操作５０４に関連して下記で議論される最適な会話の部分（ＰＯＳ）タグに、表の単語は関連することに注意が必要である。

【００２６】

【表２】

事実―単語リスト	ＰＯＳタグ
単語／句１	ＰＯＳタグ
単語／句２	ＰＯＳタグ
単語／句Ｎ	ＰＯＳタグ

表２

【００２７】

意見というよりも、事実を示唆している単語を決定するために、調査が行われている。例えば、事実を紹介する単語のクラスは、動詞の分類とその辞書的な機能に関する検索と作用を使用して得ることができる。これを実施する材料として使用することができる関連文献２つを示す。

【００２８】

（１）Ｍｅｌ’ｃｕｋ（１９９６）　辞書機能：辞書での語彙関係を記述するためのツール　Ｉｎ　Ｌ．Ｗａｎｎｅｒ（ｅｄ．）：辞書編集での辞書機能と自然言語処理、アムステルダム／フィラデルフィア：Ｂｅｎｊａｍｉｎｓ，３７－１０２．

【００２９】

（２）Ｆｏｎｔｅｎｅｌｌｅ，Ｔ．（１９９７）：「辞書の見出し語で重要な辞書機能を発見すること」ｉｎ　Ｃｏｗｉｅ，　ＡＰ．（ｅｄ．）　述語：理論、解析、応用、Ｏｘｆｏｒｄ　Ｕｎｉｖｅｒｓｉｔｙ　Ｐｒｅｓｓ，Ｏｘｆｏｒｄ．

【００３０】

こうして、そのような検索を基礎として、非事実ではなくむしろ事実の表現を示唆するこれらの動詞、又は他の単語を含むように表２に示すような事実－単語リストが構築される。例えば、「発明された」又は「雇われた」という用語は、事実の表現を示唆するが、「できる」又は「不平を言う」という用語は示唆しない。事実－単語表の特殊例を、この明細書の最後にある付録Ａに示す。この特殊例は、電子文書内で事実の記述を発見するた

めに使用可能な事実－単語である動詞の網羅的でないリストである。

【００３１】

事実－単語表を電子文書に適用する際か、ＰＯＳタグが既に事実－単語表内の単語と結び付けられているような事実－単語表の適用と平行するかどちらかで、タグ操作５０４にて事実の記述の各々の各単語の会話部分（ＰＯＳ）がタグ付けされている。このタグ付け操作５０４は、スキャン操作５０２と平行して、又は引続いて発生する可能性があるが、名詞句のような統語上の句は、事実のイベント内に含まれるエンティティになると知られていると理解されているので、動詞タグよりも名詞タグの方を支持することによるというように、複数のＰＯＳタグを有する単語を明確に選択させることを含む。未知の、又事前にタグが付いていない単語はどれでも、この理由により名詞に戻ることがある。名詞に関しては、動詞よりも形容詞の方が同様に好まれるので（例えば、動詞としての「計画された」よりも形容詞としての「計画された」）、形容詞は事実のイベント内に含まれるエンティティとして知られる名詞句の一部であるので、形容詞と動詞のタグ両方を有する単語は形容詞に戻る。ＰＯＳタグと事実－単語表の単語の関係を生成する場合、例えばこの表を作成する時に、このような明確化する選択が既に適用されている可能性があるので、例えば「計画された」が結び付けられるのは、表中の形容詞のＰＯＳタグであり、動詞のＰＯＳタグではない。

【００３２】

事実の記述が見つけられ、事実の記述の単語がＰＯＳでタグ付けされると、次いで、事実抽出の精度を向上するために更に徹底した解析が行われるが、文書全体のこのような徹底した処理が必要になるわけではない。識別操作５０６では、名詞句や動詞句のような統語上の句が識別される。統語句は、従来の文法規則や簡単な言語解析を利用して識別できる。近傍にある、即ち文書中で事実の記述のセットに対して非常に局所的であるこれらの統語句が識別され、もし事実記述がそれに関連した統語句を有していなければ、対応する文章がそれ以降の考慮から除外される。こうして、事実の記述の近傍にあるこれらの統語句にのみ焦点を合せることにより、プロセスは文章全体の全ての言語成分は見ないようにしている。

【００３３】

更に、識別操作５０６では、隣り合う統語句を有する事実の記述の言語成分は、事実の記述内で識別されたパターンに基いて、対応する文章内で統語句が果たす役割を評価することにより、さらに決定される。こうして、解析しようとする現在の事実の記述を含む文内で、統語句が主語、又は目的語の役割をするか否かが、事実の記述の単語パターンから決定される。

【００３４】

事実の記述の言語成分が決定されると、即ち統語句とその役割が識別されれば、事実の記述のこれらの名詞句に対して除外規則が適用できて、除外操作５０８で事実の表現にはなりそうにないものが更に除去される。除外規則は、目的語としての統語句、主語としての統語句、又はその役割にはかかわらない統語句に基いて適用することができる。更にこの特定の実施形態では、個々の単語、統語句、又は文章全体に適用される除外規則は同じ結果に至るが、それは文章全体が事実の記述になることから除外することになる。適用可能な除外規則の一例を表３に示す。

【００３５】

【表３】

除外規則	結論
「目的語」は「意見/偏見」修飾語句を持つ	文章の候補を除外する
文章フィルタ： －文章の最初の単語（例　代名詞） －句読点：例”？”	文章の候補を除外する
「主語」は限定詞－固有名詞でなければ	文章の候補を除外する
「目的語」の周囲の「文脈」	周囲の文脈が事実（例　代名詞のあるクラス）を指示していない特定のＰＯＳを有していれば、文章の候補を除外する
文中で停止単語が発生する	文章の候補を除外する
「目的語」の「主語」が代名詞を含む	名詞句を除外する

表３

【００３６】

除外規則を適用するに際してか、又は除外規則を適用するのに平行してかのどちらかに、スコアリング操作５１０でスコアリング規則が適用される。様々な特徴夫々のために主語と目的語の名詞の両方にスコアリング規則は重みを与え、候補となる事実の記述のスコアの合計は、個々の特徴の重みに、一致する事実－単語の確実性スコアを合計したものになる。個々の特徴の重みは、事実を表示する場合は正であり、非事実を表示する場合は負となる。特徴と関連するスコアリング規則の例を下記表４に示す。特徴スコアは、人間の判断を使用して人手により割当てるか、自動的に習得される。

【００３７】

【表４】

特徴	スコアリング規則
一致するパターンの確実性スコア（事実－単語、例　主動詞）	
役割のクラス（即ち、主語又は動詞）例　人、国、組織、等	クラス当たりのスコア
主要な「主語」は固有名詞を含む	普通の重み
「目的語」の長さ	長さのスコア
「主語」の長さ	長さのスコア
文章の長さ	長さのスコア
「主語」は文章の最初に現れる－即ち主語の補正 「目的語」は修飾語（形容詞、副詞）を持つ	位置のスコア 負－基本的重み
「目的語」は限定詞（”the”）	負－基本的 連結詞文で終了する時は除く

表４

【００３８】

事実の記述に対するスコアの合計を予め定義した閾値と比較して、クエリー操作５１２でスコアの合計が閾値を超えるか否かを判定する。閾値を超えなければ、対応する事実の記述は捨ててよい。閾値を超えれば、事実の記述、完全な文章、及び／又は完全なパラグラフ、又は他の文書部分は、表示操作５１４にて事実として表示される。この表示は、事実を表示すること、事実をライブラリに記憶すること等を含む。

【００３９】

スコアリング規則と閾値比較を利用する場合は、特徴及び／又は閾値に割り当てられる重みは、事実の抽出への全体的なアプローチを操作しないで、操作することができる。このようにして、処理ステップは同じままにして、事実の抽出と表示の精度を制御することができる。

【００４０】

図６は、検索の実施により得られるスクリーンショット６００の一例を示す。検索用語を検索フィールド６０２に入力して検索を行なった。この検索用語は、インターネットで利用可能な様々なウェブサイトのリンク６０４に一致させた。ユーザは、普通のやりかたで電子文書にたどりつくことができる。

【００４１】

また、検索用語についての事実６１０，６１２，６１４をセクション６０８に表示する。また、見つけた電子文書のどれかに行かなくても、又事実を読み取って意見と区別する必要なしに、ユーザは検索の主題についての事実を迅速に見極めることができる。この特定の例では、ユーザが事実の出所に関する情報さらに与えること、及び／又はその事実が発見された背景（例えば、関連する事実、他の事実の日付等）を示すことを選択できるように、事実６１０，６１２，６１４はハイパーリンクを含む。

【００４２】

スクリーンショット６００は、事実がユーザに対してどのように提示されるかの一例にすぎないということは認識されよう。図示したように、別の欄で表示するよりも、抽出元の電子文書のサブエレメントとして事実はリスト化されている。又、検索結果のページに事実をリスト化する代わりに、又はそれに加えて、特定の電子文書から抽出された事実が、ユーザがその電子文書自体を見た際に、欄又は他の場所にリスト化される。更に、表示するために事実を文書から分離する代わりに、又はそれに加えて、表示用に選択された時に、検索結果内の文書６０４のリスト内と、完全な電子文書内の両方で、事実を電子文書内で強調表示することができる。更にもう１つの選択肢として、選択可能なリンクだけで事実を表示して元の文書を得るようにして、その場合抽出された事実だけが検索され文書検索を完全に避けるというように検索結果とは別に事実を表示することができる。

【００４３】

又、抽出した事実の提示はスクリーンショット６００内に示すように、ローカルユーザのための検索や事実の抽出を実装しているローカルコンピュータへの表示として提供されることは認識される。あるいは、スクリーンショット６００内に示すように、抽出された事実の提示は、インターネットベースの検索エンジンの場合のように、ローカルコンピュータがリモートコンピュータの代わりに検索や事実の抽出を行うことを要求するリモートコンピュータへの表示として提供される。

【００４４】

従って、事実は効率的かつ正確に、ユーザに提示するために文書から抽出される。多段アプローチにより、事実の記述が発見された文章全体を詳細に解析しなくてもよいのと同様に、文書全体の詳細な解析をせずに効率を上げることができる。処理の初期段階で文書中に発見された事実の記述に関する更なる解析を利用することにより、正確さが維持される。

【００４５】

様々な実施形態を参照しながら、本発明を特に示して記述しているが、本発明の精神と範囲から外れることなく、形式や詳細について様々な他の変更がなされるということを、当業者は理解している。例えば、後で他の除外規則を適用する期間中よりも、事実の記述のために構文解析する際、文章の句読法に基づいた除外規則のような、事実の記述の言語的な成分に特化しないある除外規則を適用することができる。

【００４６】

付録Ａ－事実単語

abase	advance	appear	avoid
abate	advertise	appease	awake
abort	aerate	apply	award
abrade	afford	argue	back
abridge	aggravate	arouse	bail
absorb	agree	arrange	bank
abstract	aid	arrest	bar
accelerate	aim	arrive	barbarize
accent	air	ask	bare
accept	allay	assemble	base
accredit	alleviate	assert	batter
achieve	alter	asseverate	beach

act	amend	assign	beam
add	amplify	assuage	bear
address	amuse	assure	become
adduce	animate	attach	befog
adjust	announce	attack	befuddle
administer	answer	attenuate	beget
admit	antedate	avert	begin

【００４７】

begrime	buy	compromise	damage
belch	bypass	conceal	damp
belie	canvass	concede	dance
bend	cap	conceive	dangle
benumb	capitalize	conciliate	darken
bequeath	carry	conclude	darn
bestow	cast	conduct	dash
betray	castigate	confess	deaden
better	castrate	confide	deal
bind	catch	confirm	debase
blackleg	chafe	confound	debauch
blanket	change	confuse	debunk
bleach	channel	congeal	decay
blemish	chafe	connect	decide
blend	check	conserve	declare
blight	chill	consolidate	deepen
blister	chime	constitute	deface
block	chip	constrain	defeat
blockade	chock	constrict	defend
blow	choke	continue	deflate
blunder	choose	contort	deflect
blunt	churn	contact	deform
blur	cipher	control	defrost
blurt	circulate	convert	delay
bob	circumvent	convey	delegate
bog	claim	cook	deliver
boil	clash	cool	demise
bolster	clean	cordon	demonstrate
boost	cleanse	correct	dent
bowdlerize	clear	corrode	deny
bowl	climb	corrupt	deplete
brace	clinch	counter	depreciate
brand	clip	countersink	depress
brave	clog	cover	deprive
break	close	crack	depute
brief	clot	crank	derange
brighten	cloud	cash	describe
bring	cockle	craze	desecrate
broadcast	coin	create	design
bruise	collapse	cripple	designate
buckle	collect	crop	desolate
build	colour	cross	despoil

bull	comfort	crumble	destroy
bunch	commission	crush	detail
bundle	commit	cry	detect
bung	communicate	curb	deteriorate
burlesque	compare	curdle	determine
burn	complete	curtail	develop
burst	compound	cushion	die
bury	compress	cut	differentiate

【００４８】

diffuse	earth	exhale	foil
dilute	ease	exhibit	fold
dim	eat	exist	follow
diminish	educate	expand	force
direct	effect	expedite	forge
dirty	elevate	explain	forgive
disable	elicit	expose	form
disappear	elude	expound	foster
discharge	emancipate	express	foul
discipline	embellish	extend	found
disclose	embitter	extinguish	frame
discolor	embody	extort	fray
disconnect	emit	extract	free
discontinue	emphasize	fabricate	freeze
discover	enable	face	frustrate
discuss	encourage	fade	furl
disfigure	end	fail	furnish
disguise	endorse	fake	furrow
dislocate	endow	fall	fuse
dislodge	enforce	falsify	gain
dismantle	engage	familiarize	gallop
dismount	enhance	fasten	garble
disorder	enjoin	father	gash
dispatch	enlarge	fatten	generate
dispense	enliven	feature	gerrymander
disperse	ennoble	feed	get
display	enrich	ferry	give
dispute	enroll	fertilize	gladden
disrupt	enshrine	festoon	glorify
distil	entail	fiddle	gloss
distinguish	entangle	fight	glut
distort	enthrone	fill	go
disturb	entrust	filter	govern
divert	enunciate	finalize	grade
divide	epitomize	find	graduate
dock	equalize	finish	grant
doctor	erect	fire	grate
dodge	escalate	fit	graze
double	establish	fix	ground
douse	evade	flag	group
draft	evaporate	flash	grow

dramatize	evince	flaunt	guide
draw	evoke	flay	halt
dredge	exacerbate	float	halve
dress	exact	flood	hamper
drive	exaggerate	floodlight	handle
drop	examine	flourish	happen
drown	exasperate	flush	harass
duff	exceed	fly	harbour
dull	excite	fog	harden

【００４９】

harm	instigate	link	navigate
harmonize	instill	listen	neaten
harry	institute	litter	nick
hasten	integrate	live	nip
hatch	intend	liven	notch
head	intensify	load	notice
heal	interpolate	lock	nourish
hear	interrupt	loose	nurse
heat	intimate	loosen	obfuscate
heighten	introduce	lose	obscure
help	invert	lower	obstruct
hide	invigorate	lump	obtain
hit	invite	magnify	occupy
hoard	invoke	maintain	occur
hoist	involve	make	offend
hold	issue	manage	offer
hope	jab	mangle	open
hound	jam	manipulate	operate
hurt	jettison	manufacture	oppose
identify	jingle	mark	order
illuminate	join	marshal	originate
imagine	jumble	mask	outline
impair	jump	match	overcharge
impart	justify	matter	overdo
impeach	keep	maul	overflow
impede	kick	measure	overturn
imperil	kill	meet	overwork
implant	kindle	mellow	pacify
improve	knock	melt	pack
inaugurate	lacerate	mend	pad
increase	ladder	mention	panic
indent	lance	mildew	paralyze
indenture	land	mind	pare
indicate	laugh	misrepresent	parley
induce	launch	miss	parole
induct	lay	mist	parry
infect	layer	mitigate	part
infiltrate	lead	modify	partition
infix	leave	mollify	pass
inflame	lend	moot	patch

inflate	lengthen	mould	pay
inflict	lessen	move	peal
influence	let	muddle	peddle
inform	level	muddy	peg
infuse	liberate	muffle	penalize
initial	lie	muss	perform
initiate	light	muster	perish
injure	lighten	mute	persecute
insert	limit	mutilate	pervert
inspire	line	narrow	phrase

【００５０】

pick	prove	refuse	rock
pillow	provide	regard	roll
pique	provoke	register	rotate
pit	prune	regulate	rouse
placard	publicize	rehabilitate	row
place	publish	rehearse	ruffle
plan	pull	reinforce	ruin
plant	pulp	reissue	rumple
play	punch	reject	run
pluck	puncture	rekindle	rush
plug	punish	relate	rustle
plunge	punt	relax	sail
point	purge	release	salvage
poison	push	relieve	sap
pole	put	reline	save
polish	qualify	remould	scald
poll	quarter	remove	scold
pool	quench	rend	score
pop	question	renew	scotch
pose	quicken	renovate	scratch
position	quieten	reopen	scream
post	quilt	repair	scuff
pound	race	replace	scupper
preach	raise	report	scuttle
precipitate	ransack	republish	seal
predate	rap	require	sear
prefer	rationalize	rerun	seat
prejudice	rattle	reseat	secure
preoccupy	re-engage	resist	see
prepare	re-establish	rest	sell
present	re-form	restart	send
preserve	read	restore	serve
prettify	rear	restrain	set
prevent	reawaken	result	settle
prick	recall	resurrect	server
prime	receive	retail	shake
proclaim	reclaim	retain	shame
procure	recline	retire	sharpen
produce	recognize	retract	shatter

profess recommend retrench sheathe
programme reconcile retrieve shed
promote reconsider return shelter
promulgate record reveal shield
prop recruit reverse shift
propagandize reduce revive shine
propel refer rewind shingle
profound refine right shirk
prosecute reflect ring shoot
protect refloat rise shorten
protest reform roast shout

【００５１】

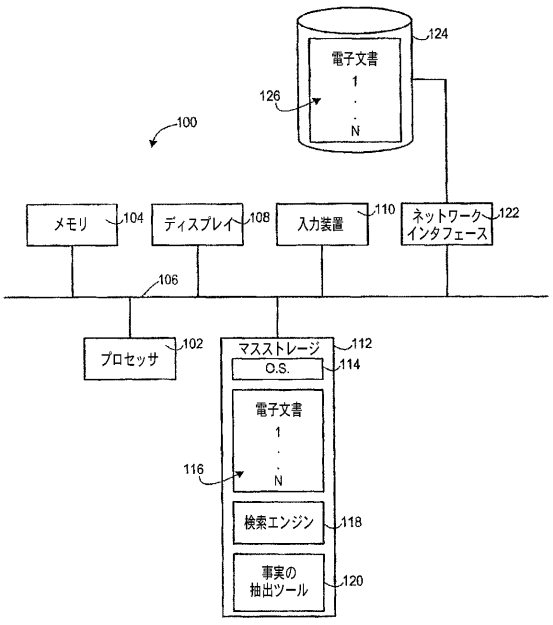
show spoil subvert trample
shrink sponsor succeed transfer
shut sport suffer transplant
sift spot suggest trap
sign spout suit travel
signal sprain summarize treat
signalize spray supplement trigger
signify spread supply trim
simmer spring support truss
sing square suppose try
singe squash suppress tumble
sink squeeze surface turn
sit stack surrender twang
site staff survive twiddle
situate stain suspend twirl
skirt stalemate sustain twist
slacken stall sweep unblock
slake stamp sweeten unburden
slash stand swell unclog
sleep star swing undo
slice starch swish unfasten
slip start taint unfix
slow staunch tarnish unfold
smear stay task unhinge
smile steady teach unhitch
smudge steer tear unite
snag stem telephone unloose
snap step temper unravel
snarl stick tend unsaddle
snuff stiffen thank unseat
sober still thaw unsex
soften stir thin unstop
soil stoke thrill untangle
solace stop throw untwist
solidify store thrust uphold
soothe straighten thump upset
sort strain thwart urge
sound strand tidy use

sour	strengthen	tighten	validate
sow	stress	toll	vandalize
spare	stretch	tootle	veer
spark	strike	topple	veil
speak	strip	torment	ventilate
speck	strum	torture	vocalize
speed	study	total	voice
spill	stuff	touch	vote
spin	stultify	toughen	vulgarize
splinter	stunt	tousle	waft
spilt	subdue	tow	waggle
splodge	subscribe	train	wake

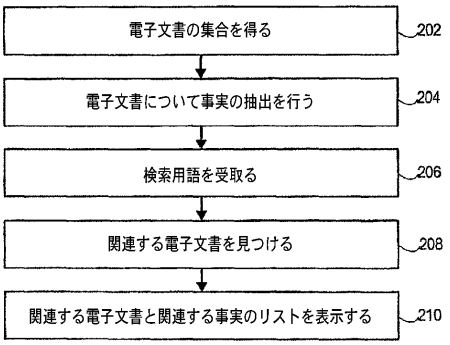
【００５２】

walk	wear	wilt	work
wangle	weave	win	worry
warm	weep	wind	wreak
warn	weld	wing	wreck
warp	whet	wipe	wrest
warrant	whirl	wire	wring
wash	whitewash	wish	wrinkle
watch	widen	withdraw	write
weaken	wield	wither	yield
wean	wiggle	withhold	

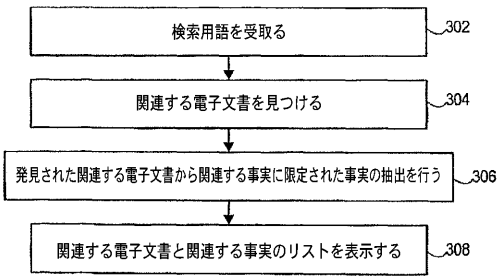
【図１】



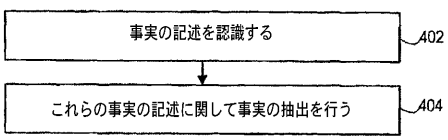
【図２】



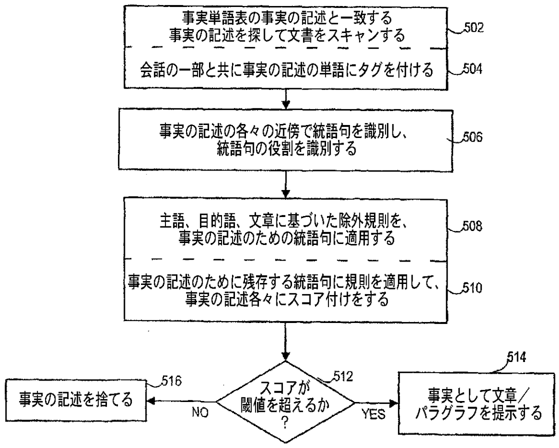
【図３】



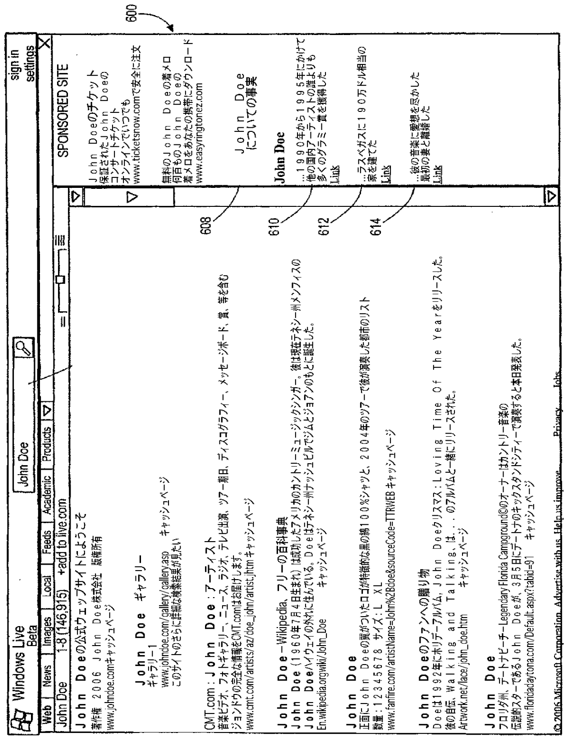
【図４】



【図５】


事実単語表の事実の記述と一致する
事実の記述を探して文書をスキャンする
502
会話の一部と共に事実の記述の単語にタグを付ける
504
事実の記述の各々の近傍で統語句を識別し、
統語句の役割を識別する
506
主語、目的語、文章に基づいた除外規則を、
事実の記述のための統語句に適用する
508
事実の記述のために残存する統語句に規則を適用して、
事実の記述各々にスコア付けをする
510
スコアが
閾値を超えるか
？
512
NO
YES
事実の記述を捨てる
516
事実として文章／
パラグラフを提示する
514

【図６】


600
608
610
612
614
SPONSORED SITE
sign in
settings
Windows Live
Beta
Web
News
Images
Local
Feeds
Academic
Products
John Doe

フロントページの続き

(72)発明者　ケビン　ウィリアム　ハンフリーズ
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション　インターナショナル　パテンツ内

審査官　早川　学

(56)参考文献　米国特許出願公開第２００５／０１０８６３０（ＵＳ，Ａ１）
特開２００１－３５７０６４（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F　17/30
G06F　17/21