(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第3799750号

(P3799750)

(45)発行日　平成18年7月19日(2006.7.19)　(24)登録日　平成18年5月12日(2006.5.12)

(51)Int.Cl.		F I		
G06F 17/30	(2006.01)	G06F	17/30	２２０Ｚ
G06F 19/00	(2006.01)	G06F	19/00	１３０

請求項の数 12　(全 14 頁)

(21)出願番号　特願平9-166516
(22)出願日　平成9年6月9日(1997.6.9)
(65)公開番号　特開平10-340275
(43)公開日　平成10年12月22日(1998.12.22)
審査請求日　平成15年4月16日(2003.4.16)

(73)特許権者　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目１７番２２号
(74)代理人　100098132
弁理士　守山　辰雄
(72)発明者　林　直樹
神奈川県足柄上郡中井町境４３０　グリーンテクなかい　富士ゼロックス株式会社内
(72)発明者　田中　剛
神奈川県足柄上郡中井町境４３０　グリーンテクなかい　富士ゼロックス株式会社内
(72)発明者　宗像　英明
神奈川県足柄上郡中井町境４３０　グリーンテクなかい　富士ゼロックス株式会社内

最終頁に続く

(54)【発明の名称】情報抽出装置

(57)【特許請求の範囲】

【請求項１】

文書を当該文書の更新日時に関する情報とともに保持する文書データベースと、

文書データベースに保持された複数の文書に含まれる特定の語句の時間的な分布を文書の更新日時に基づいて検出して、分布の強度が高い語句を話題語句として抽出する話題語句抽出手段と、

話題語句と同一文書に含まれる他の語句をカテゴリ語句として抽出する話題カテゴリ抽出手段と、

話題語句をカテゴリ語句を用いて分類して管理する話題管理手段と、を備えたことを特徴とする情報抽出装置。

【請求項２】

文書を当該文書の更新日時に関する情報とともに保持する文書データベースと、

文書データベースに保持された複数の文書に含まれる特定の語句の時間的な分布を検出して、分布の強度が高い語句を話題語句として抽出する話題語句抽出手段と、

評価を表すキーワードを保持する評価キーワード保持手段と、

話題語句と同一文書に含まれるキーワードを検出する評価検出手段と、

話題語句と検出されたキーワードとを対応付けて管理する評価管理手段と、を備えたことを特徴とする情報抽出装置。

【請求項３】

文書を当該文書の更新日時に関する情報とともに保持する文書データベースと、

文書データベースに保持された複数の文書に含まれる特定の語句の時間的な分布を検出して、分布の強度が高い語句を話題語句として抽出する話題語句抽出手段と、
話題語句と同一文書に含まれる他の語句をカテゴリ語句として抽出する話題カテゴリ抽出手段と、
話題語句をカテゴリ語句を用いて分類して管理する話題管理手段と、
評価を表すキーワードを保持する評価キーワード保持手段と、
話題語句と同一文書に含まれるキーワードを検出する評価検出手段と、
話題語句と検出されたキーワードとを対応付けて管理する評価管理手段と、を備えたことを特徴とする情報抽出装置。

【請求項４】
請求項１又は請求項３に記載した情報抽出装置において、
前記話題管理手段は、更に、話題語句に対して当該話題語句を含む文書の更新日時に関する情報を対応付けて管理することを特徴とする情報抽出装置。

【請求項５】
請求項２又は請求項３に記載した情報抽出装置において、
前記評価管理手段は、更に、検出されたキーワードを含む文書の更新日時に関する情報を対応付けて管理することを特徴とする情報抽出装置。

【請求項６】
請求項４又は請求項５に記載した情報抽出装置において、
更に、話題管理手段或いは評価管理手段に管理されている情報を時系列に並べて出力する出力手段を備えたことを特徴とする情報抽出装置。

【請求項７】
話題語句抽出手段と話題カテゴリ抽出手段と話題管理手段を有する情報抽出装置により行われる情報抽出方法であって、
　前記話題語句抽出手段により、文書データベースに保持された複数の文書について、これら文書に含まれる特定の語句の時間的な分布を文書の更新日時に基づいて検出して、分布の強度が高い語句を話題語句として抽出し、
　前記話題カテゴリ抽出手段により、話題語句と同一文書に含まれる他の語句をカテゴリ語句として抽出し、
　前記話題管理手段により、話題語句をカテゴリ語句を用いて分類することを特徴とする情報抽出方法。

【請求項８】
話題語句抽出手段と評価検出手段と評価管理手段を有する情報抽出装置により行われる情報抽出方法であって、
　前記話題語句抽出手段により、文書データベースに保持された複数の文書について、これら文書に含まれる特定の語句の時間的な分布を検出して、分布の強度が高い語句を話題語句として抽出し、
　前記評価検出手段により、話題語句と同一文書に含まれる評価を表わすキーワードを検出し、
　前記評価管理手段により、話題語句と検出されたキーワードとを対応付けて管理することを特徴とする情報抽出方法。

【請求項９】
話題語句抽出手段と話題カテゴリ抽出手段と話題管理手段と評価検出手段を有する情報抽出装置により行われる情報抽出方法であって、
　前記話題語句抽出手段により、文書データベースに保持された複数の文書について、これら文書に含まれる特定の語句の時間的な分布を検出して、分布の強度が高い語句を話題語句として抽出し、
　前記話題カテゴリ抽出手段により、話題語句と同一文書に含まれる他の語句をカテゴリ語句として抽出し、
　前記話題管理手段により、話題語句をカテゴリ語句を用いて分類し、

前記評価検出手段により、話題語句と同一文書に含まれる評価を表すキーワードを検出し、
前記評価管理手段により、話題語句と検出されたキーワードとを対応付けて管理することを特徴とする情報抽出方法。
【請求項１０】
文書データベースに保持された複数の文書について、これら文書に含まれる特定の語句の時間的な分布を文書の更新日時に基づいて検出して、分布の強度が高い語句を話題語句として抽出する機能と、
話題語句と同一文書に含まれる他の語句をカテゴリ語句として抽出する機能と、
話題語句をカテゴリ語句を用いて分類して管理する機能とを、コンピュータに実現するプログラムを記憶したコンピュータにより読み取り可能な記憶媒体。
【請求項１１】
文書データベースに保持された複数の文書について、これら文書に含まれる特定の語句の時間的な分布を検出して、分布の強度が高い語句を話題語句として抽出する機能と、
話題語句と同一文書に含まれる評価を表すキーワードを検出する機能と、
話題語句と検出されたキーワードとを対応付けて管理する機能とを、コンピュータに実現するプログラムを記憶したコンピュータにより読み取り可能な記憶媒体。
【請求項１２】
文書データベースに保持された複数の文書について、これら文書に含まれる特定の語句の時間的な分布を検出して、分布の強度が高い語句を話題語句として抽出する機能と、
話題語句と同一文書に含まれる他の語句をカテゴリ語句として抽出する機能と、
話題語句をカテゴリ語句を用いて分類して管理する機能と、
話題語句と同一文書に含まれる評価を表すキーワードを検出する機能と、
話題語句と検出されたキーワードとを対応付けて管理する機能とを、コンピュータに実現するプログラムを記憶したコンピュータにより読み取り可能な記憶媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、複数の電子文書から話題となる語句を抽出する情報抽出技術に関する。
【０００２】
【従来の技術】
近年、広域コンピュータネットワークを利用した電子商取引が盛んになってきている。このような電子商取引によると、取り引き相手の所在地や活動地域といった地理的制限に依存せずに、多くの見込み客に対して安価に営業活動を行うことができる。
【０００３】
このような電子商取引では、商品の売り手が商品カタログをネットワーク上に公開したり、直接買い手に商品の提案書を送付している。一方、商品の買い手は、多数の文書に示された複数の購入候補の商品からどの商品が適切なものかを判断しなければならない。
【０００４】
このように買い手が購入候補の商品の中から適切なものを判断する上では、購入候補の商品に関する動向を知ることは有用である。すなわち、今まで商品に関するどのような話題があるか、また、今後どのような商品が予定されているかといった話題を買い手が知ることで、売り手の提案する商品が機能や価格の面で妥当なものであるか否かを買い手が判断できるようになる。
【０００５】
ここで、上記したような話題に関する情報を収集する方法としては、例えば、ネットワーク等に存在する複数の電子文書から収集する方法があり、複数の文書から話題となっている単語を抽出する従来の技術としては、特開平７－１４１３９６号公報に示された技術がある。この技術においては、装置に入力された文書に含まれる単語の出現頻度を記録し、或る文書の入力時にある単語の出現頻度があらかじめ規定された閾値を超えた場合に、こ

の単語を現在の話題として抽出するものである。
【０００６】
また、他にも特開平８－７７１７８号公報に示される技術がある。この技術においては、ある単語の出現頻度が時間的に単調減少するものがあれば、その単語が話題であるとして抽出するものである。
また、他にも特開平７－３２５８３２号公報に示される技術がある。この技術においては、特定分野・期間において話題となった単語を検索したり、特定の単語と同時に話題となった単語を検索したり、特定の単語が話題となった分野・期間を検索したりするものである。
【０００７】
【発明が解決しようとする課題】
しかしながら、従来の技術においては、話題となっている複数の単語は抽出できても、各単語間の関連についての情報は何ら抽出されておらず、例えば、抽出された複数の単語が或る同一の事項（例えば、商品）に関する話題を表す単語である場合であっても、これら抽出された複数の単語が同一の事項についての話題を表しているといったことを認識することはできない。
このため、或る事項についての複数の話題を把握するといったことは不可能であり、また、或る事項についての話題の推移を把握するといったこともできない。
【０００８】
また、従来の技術では、或る話題に関する情報として、或る話題が存在している期間について抽出することができるが、或る商品に人気がある場合であっても、人気がない場合であっても、商品に関する話題が発生する可能性があり、また、或る商品が入手不能となってしまった場合であっても、或る商品が広く定着してしまった場合であっても或る商品に関する話題が消える可能性があるといった現状を考えれば、抽出された話題が存在している期間の情報だけでは、なぜその話題が発生したか、なぜその話題が消えてしまったか等といった話題の発生、消滅等といったことについての要因を把握することができなかった。
【０００９】
本発明は、上記した従来の事情に鑑みてなされたもので、複数の文書から話題を抽出するだけでなく、更に、話題をより有効的に利用するための種々の情報を抽出することのできる情報抽出装置、情報抽出方法、および、これら装置や方法を実現するプログラムを記憶したコンピュータにより読み取り可能な記憶媒体を提供すること目的としている。
【００１０】
【課題を解決するための手段】
上記した目的を達成するために本発明に係る情報抽出装置は、文書を当該文書の更新日時に関する情報とともに保持する文書データベースと、文書データベースに保持された複数の文書に含まれる特定の語句の時間的な分布を文書の更新日時に基づいて検出して、分布の強度が高い語句を話題語句として抽出する話題語句抽出手段と、話題語句と同一文書に含まれる他の語句をカテゴリ語句として抽出する話題カテゴリ抽出手段と、話題語句をカテゴリ語句を用いて分類して管理する話題管理手段と、を備えたことを特徴とする。
【００１１】
ここで、語句とは、一つの単語であってもよく、複数の単語であってもよい。また、分布の強度の高い語句とは、分布している時間範囲に対する分布中のピークの度合いが高い語句のことをいい、例えば、短い時間範囲に分布している語句、分布中のピークが突出している語句等がある。この分布強度が高い語句は、話題性の高い語句とみなすことができる。
【００１２】
上記した情報抽出装置では、話題語句抽出手段が文書データベースに保持された複数の文書に含まれる特定の語句の時間的な分布を文書の更新日時に基づいて検出して、分布の強度が高い語句を話題語句として抽出し、話題カテゴリ抽出手段が話題語句と同一文書に含

まれる他の語句をカテゴリ語句として抽出し、話題管理手段が話題語句をカテゴリ語句を用いて分類して管理すると言ったように、本発明に係る情報抽出方法を実施することができる。

【００１３】

したがって、文書から話題性の高い話題語句を抽出できるとともに、話題語句が関係するカテゴリ語句によって複数の話題語句を関連付けて抽出することができる。

更に、話題語句に当該話題語句を含む文書の更新日時に関する情報を対応付けておくことにより、同一カテゴリ語句に分類される話題語句を時系列に表示させることができ、同一カテゴリ語句に関する話題語句の推移を示すことができる。

【００１４】

また、本発明に係る情報抽出装置は、文書を当該文書の更新日時に関する情報とともに保持する文書データベースと、文書データベースに保持された複数の文書に含まれる特定の語句の時間的な分布を検出して、分布の強度が高い語句を話題語句として抽出する話題語句抽出手段と、評価を表すキーワードを保持する評価キーワード保持手段と、話題語句と同一文書に含まれるキーワードを検出する評価検出手段と、話題語句と検出されたキーワードとを対応付けて管理する評価管理手段と、を備えたことを特徴とする。

【００１５】

上記した情報抽出装置では、話題語句抽出手段が文書データベースに保持された複数の文書に含まれる特定の語句の時間的な分布を検出して、分布の強度が高い語句を話題語句として抽出し、評価検出手段が話題語句と同一文書に含まれる評価キーワード保持手段に保持されたキーワードと同一のキーワードを検出し、評価管理手段が話題語句と検出されたキーワードとを対応付けて管理すると言ったように、本発明に係る情報抽出方法を実施することができる。

【００１６】

したがって、文書から話題性の高い話題語句が抽出できるとともに、話題語句に対する評価を表しているキーワードを抽出することができる。

更に、検出されたキーワードに、当該キーワードを含む文書の更新日時に関する情報を対応付けておくことにより、或る話題語句に対する評価を表しているキーワードを時系列に表示させることができ、話題語句に対する評価の推移を示すことができる。

なお、本発明に係る情報抽出装置や方法は、上記のカテゴリ語句と評価キーワードとに話題語句を対応付けて管理するものでもあり、更には、本発明に係るコンピュータにより読み取り可能な記憶媒体は、これらの情報抽出装置や情報抽出方法をコンピュータにより実現するプログラムを記憶したものである。

【００１７】

【発明の実施の形態】

本発明の一実施例に係る情報抽出装置を図１を参照して説明する。

この情報抽出装置は、文書データベース１と、単語解析部２と、単語記憶部３と、話題解析部４と、話題記憶部５と、話題カテゴリ解析部６と、話題カテゴリ記憶部７と、評価キーワード記憶部８と、話題評価解析部９と、話題評価記憶部１０と、提示選択部１１と、情報提示部１２と、を備えている。

【００１８】

文書データベース１は、複数の電子文書を記憶する機能を持ち、各電子文書を、文書の内容が最後に更新された更新日時の情報と、各文書を一意に識別するための文書ＩＤとともに記憶している。

本実施例では、文書フォーマットとしては、ＨＴＭＬ（HyperText Mark-Up Language）を用い、文書ＩＤとしては、文書が格納されている場所を示すＵＲＬ（uniform resource locator）を用いている。

【００１９】

また、文書データベース１は、他の機能部から電子文書に対する問い合わせ命令を受け付け、当該問い合わせ命令に従って処理を行い、当該処理の結果を前記他の機能部へ返答す

る機能を有している。
図２には、本文書データベース１が受け付ける問い合わせ命令及び当該問い合わせ命令の引数と、問い合わせ命令に対する処理の結果を示してある。
例えば、文書データベース１が問い合わせ命令”全ＩＤ”を受け付けた場合には、文書データベース１に蓄積された全文書の文書ＩＤを返答し、また、或る文書ＩＤを引数とする問い合わせ命令”更新日時”を受け付けた場合には、引数の文書ＩＤで示される文書の更新日時を返答する。
なお、この文書データベース１は、ネットワークによって接続された複数のコンピュータシステムによって実現することもでき、単一のコンピュータシステムによって実現することもできる。

【００２０】
単語解析部２は、文書データベース１に対して問い合わせを行い、文書データベース１に記憶されたすべての文書の文書ＩＤを入手し、入手した文書ＩＤが示す文書の内容を文書データベース１に対して問い合わせを行って順次入手し、入手した内容に含まれる単語を抽出し、抽出した単語と、当該単語を含んでいた文書の文書ＩＤとを単語記憶部３に通知する。
なお、文書の内容から単語を抽出する方法としては、形態素解析等といった一般に知られている方法を用いることができる。

【００２１】
単語記憶部３は、単語解析部２から通知された単語と文書ＩＤとを対応付けて記憶する機能を有している。
本実施例では、辞書形式のデータ構造によって単語と文書ＩＤとを対応付けて記憶している。すなわち、辞書の項目見出しに当たるデータを単語とし、項目内容に当たるデータを文書ＩＤとして記憶している。
なお、単語記憶部３は、記憶した単語が重複しないように、データを更新している。

【００２２】
話題解析部４は、単語記憶部３に記憶された各単語について、各単語を含んでいる文書の更新日時の情報に基づいて時間的な分布を検出し、短期間に分布している単語又はピークを持つ分布をしている単語、すなわち、話題性が高いと考えられる単語を話題語句として抽出し、抽出した話題語句を話題記憶部５に通知する機能を有している。なお、話題解析部４の機能の詳細な説明については後述する動作説明において行う。
話題記憶部５は、話題解析部４から通知された話題語句を記憶する機能を有している。

【００２３】
話題カテゴリ解析部６は、話題の推移を分類するための項目を表す語句（以下、話題カテゴリ語句という。）を、文書データベース１に記憶された文書から抽出し、抽出した話題カテゴリ語句と、当該話題カテゴリ語句に関わる複数の話題語句とを話題カテゴリ記憶部７に通知する。

【００２４】
本実施例では、話題カテゴリ語句としては、複数の話題語句とともに文書に出現する割合の高い単語としている。すなわち、単語記憶部３に記憶された或る単語に対応づけられている複数の文書ＩＤが、話題記憶部５に記憶された複数の話題語句に単語記憶部３において対応付けられている複数の文書ＩＤに含まれる割合が高い（例えば、９割以上）場合に、前記単語を前記複数の話題語句に対する話題カテゴリ語句とする。なお、話題カテゴリ語句は、話題語句でない語句が好ましく、話題語句である場合にあっては、より広い時期に渡って文書に出現するものが好ましい。

【００２５】
例えば、”２００MHz”という話題語句が”文書１”、”文書２”に対応付けられ、”３００MHz”という話題語句が”文書２”、”文書３”、”文書４”に対応付けられ、”ＣＰＵ”という単語が”文書１”、”文書２”、”文書３”、”文書４”に対応付けられている場合には、”ＣＰＵ”が話題語句”２００MHz”及び”３００MHz”の話題カテゴリ語

句となる。
【００２６】
話題カテゴリ記憶部７は、話題カテゴリ解析部６から通知された、話題カテゴリ語句と複数の話題語句とを対応付けて記憶する機能を有している。
【００２７】
評価キーワード記憶部８は、話題に対する評価を表すキーワードを記憶する機能を有している。
本実施例では、品質に関する評価、コストに関する評価、デリバリーに関する評価、総合的な評価という４種類の評価を表すキーワードを記憶しており、品質に関する評価のキーワードとしては、例えば、「最新」「最先端」「ハイエンド」「エントリー」「多機能」「高速」「高？画質」「機能？不足」「練り込み？不足」「完成度」「個性的」等があり、コストに関する評価のキーワードとしては、例えば、「コストパフォーマンス」「ランニングコスト」「高コスト」「高価」「低コスト」「安価」等があり、デリバリーに関する評価のキーワードとしては、例えば、「発売？開始」「販売？開始」「リリース？遅れ」等があり、総合的な評価のキーワードとしては、例えば、「おすすめ」「チョイス」「様子？見」「待ち」「見送り」「模様眺め」等がある。
【００２８】
なお、上記したキーワード中の”？”は、０又は１個の文字が間に入っているといったことを示しており、例えば、「高？画質」といったキーワードは、「高画質」、「高い画質」等といったものをすべて含んだものとなっている。
上記したように、複数の観点の違うキーワードを用意しておくことによって、或る話題に対する種々の観点からの評価を文書から抽出することができる。
【００２９】
話題評価解析部９は、評価キーワード記憶部８に記憶されているキーワードに基づいて、文書から話題語句に対する評価のキーワードを検出し、話題語句と、当該話題語句に対する評価のキーワードと、当該キーワードを抽出した文書ＩＤと、当該文書ＩＤの文書の更新日時の情報と、を組としたデータを話題評価記憶部１０に通知する機能を有している。なお、話題評価解析部９の機能の詳細な説明については後述する動作説明において行う。
【００３０】
話題評価記憶部１０は、話題評価解析部９から通知されたデータを記憶する機能を有しており、新たなデータが通知された場合には、現在までのデータに追加して記憶する。
【００３１】
提示選択部１１は、情報提示部１２に提示させる情報に対するユーザからの選択入力を受け付ける機能を有している。
本実施例では、すべての話題語句の一覧、すべての話題カテゴリの一覧、特定の話題カテゴリに分類される話題語句の一覧、特定の話題語句と同一の話題カテゴリに分類される話題語句の一覧、特定の話題語句に対する評価一覧、特定話題語句に対する評価の推移、特定話題カテゴリにおける話題語句の推移、特定の話題カテゴリにおける話題語句の推移と評価の推移、特定の話題語句が含まれる文書の一覧、特定の文書内容等といった選択肢がある。なお、上記した選択肢を選択した後に、更に複数の選択肢がある場合についても、提示選択部１１は、選択入力を受け付ける。
【００３２】
情報提示部１２は、提示選択部１１での選択入力に応じた情報を収集してユーザに提示する機能と、提示選択部１１による選択入力に際して、取りうる選択肢（例えば、特定の話題語句と同一の話題カテゴリに分類される話題語句の一覧を表示させる選択入力を提示選択部１１が受け付けた場合には、特定の話題語句として選ぶことができる話題語句の候補）を提示する機能とを有している。なお、情報提示部１２の機能の詳細な説明は後述する動作説明において行う。
【００３３】
ここで、特許請求の範囲にいう話題語句抽出手段は、主に、上記した話題解析部４によっ

て構成され、話題カテゴリ抽出手段は、主に、上記した話題カテゴリ解析部６によって構成され、話題管理手段は、主に、上記した話題カテゴリ記憶部７によって構成され、評価検出手段は、主に、話題評価解析部９によって構成され、評価管理手段は、主に、話題評価記憶部１０によって構成され、出力手段は、主に、情報提示部１２によって構成されている。

【００３４】

次に、上記した情報抽出装置の処理動作を図３を参照して説明する。なお、図４及び図５を途中取り混ぜて説明を行う。

まず、単語解析部２が文書データベース１の文書から単語を抽出し、当該単語と文書ＩＤとを単語記憶部３に通知し、通知を受けた単語記憶部３が単語と文書ＩＤとを対応付けて記憶する（ステップＳ１）。

そして、話題解析部４が単語記憶部３に記憶された単語から話題語句を抽出する処理を行い、話題記憶部５が話題解析部４から通知された話題語句を記憶する（ステップＳ２）。

【００３５】

ここで、上記した話題解析部４による話題語句を抽出する処理を図４を参照して詳細に説明する。

まず、話題解析部４は、分布を調べる対象となる時間軸上の区間を設定する（ステップＳ７）。本実施例では、図６（ａ）に示すように、文書データベース１の最古の文書の更新日時Ｔoを時間軸上の始点とし、最新の文書の更新日時Ｔnを時間軸上の終点とした区間（以下、設定区間という）を設定する。

そして、単語記憶部３に記憶されているすべての単語について処理を行ったか否かを判断し（ステップＳ８）、未処理の単語がある場合には、未処理の単語について処理（ステップＳ９～Ｓ１４）を行う一方、未処理の単語がない場合には、処理を終了する。

【００３６】

すなわち、未処理の単語がある場合には、単語記憶部３から未処理の単語を一つ選択し（ステップＳ９）、選択した単語と対応付けられた文書ＩＤが表す文書の更新日時を文書データベース１に問い合わせ、問い合わせた結果の更新日時を図６（ｂ）に示すように時間軸上にマップする（ステップＳ１０）。

そして、時間軸上にマップされた分布の幅が、設定区間に較べて十分短いか否かを判断する（ステップＳ１１）。なお、本実施例では、前記分布の幅が（Ｔn－Ｔo）／４ より短いか否かによって判断している。

【００３７】

そして、設定区間に比べて十分に短くないと判断した場合には、設定区間をｎ個（例えば、２０個）の小区間に分割し、各小区間ごとに、前記マップした更新日時の頻度の分布を採取し（ステップＳ１２）、当該採取した分布がピークを持つ山なりの分布（例えば、図６（ｃ）に示す分布）に当てはまるか否かを判定する（ステップＳ１３）。なお、本実施例では、分布が山なりであるか否かの判定は、カイ２乗分布にあてはまるか否かによって判定している。

そして、この判定の結果、山なり分布ではない、例えば、図６（ｄ）に示すような一様に近い分布の場合には、前記選択した単語が常に一様に出現することを意味している、すなわち、話題性が高くないことを意味しているので、前記選択した単語についての処理を終了する一方、山なりの分布である場合には、前記選択した単語が一時的にかなりの頻度で出現したことを意味している、すなわち、話題性が高いことを意味しているので、前記選択した単語を話題語句として話題記憶部５に通知し（ステップＳ１４）、前記選択した単語に対する処理を終了する。

【００３８】

一方、設定区間に比べて十分に短いと判断した場合には、前記選択した単語が或る短い時間に出現していることを意味しているので、前記選択した単語を話題性の高い話題語句として話題記憶部５に通知し（ステップＳ１４）、前記選択した単語に対する処理を終了する。

【００３９】
上記したステップＳ２の処理が終了すると、話題カテゴリ解析部６が、話題記憶部５に記憶されている話題語句の中の複数の話題語句に対応付けられている文書ＩＤを単語記憶部３から把握し、把握した文書ＩＤと同一の文書ＩＤが高い確率で対応付けられている単語を単語記憶部３から話題カテゴリ語句として抽出し、当該抽出した話題カテゴリ語句及び前記話題語句中の複数の話題語句を話題カテゴリ記憶部７に通知する。
そして、話題カテゴリ記憶部７が、話題カテゴリ解析部６から通知された、話題カテゴリ語句と複数の話題語句とを対応付けて記憶する（ステップＳ３）。
これにより、関係を有する複数の話題語句同士を関連付けて管理することができる。
【００４０】
次いで、話題評価解析部９が話題語句に対する評価を文書から解析する処理を行い、話題評価記憶部１０が話題評価解析部９から通知されたデータを記憶する（ステップＳ４）。
【００４１】
ここで、上記した話題語句に対する評価を解析する処理を図５を参照して詳細に説明する。
まず、話題評価解析部９は、話題記憶部５に記憶されているすべての話題語句について処理を行ったか否かを判断し（ステップＳ１５）、未処理の話題語句がある場合には、未処理の話題語句について処理（ステップＳ１６～Ｓ２３）を行う一方、未処理の話題語句がない場合には、処理を終了する。
【００４２】
すなわち、未処理の話題語句がある場合には、話題記憶部５から未処理の話題語句を一つ選択し（ステップＳ１６）、当該話題語句に対応付けられている文書ＩＤの一つを単語記憶部３から選択し（ステップＳ１７、Ｓ１８）、選択した文書ＩＤの文書の内容を文書データベース１から取り出して、当該文書の内容から評価キーワード記憶部８に記憶された評価キーワードと同一の評価キーワードを検出する（ステップＳ１９）。
【００４３】
そして、文書の内容に評価キーワードが検出できない場合には、他の文書ＩＤの文書についての処理（ステップＳ１７～Ｓ２３）を行う。
一方、文書の内容に評価キーワードが検出できた場合には、当該文書の内容に話題記憶部５に記憶されている他の話題語句が含まれているか否かを検出し（ステップＳ２０）、また、文書の内容において、評価キーワードが前記話題語句と同一の文中に含まれるか否かを検出し（ステップＳ２１）、また、文書の内容において、評価キーワード及び前記話題語句が含まれている段落に、他の話題語句が含まれるているか否かを検出する（ステップＳ２２）。
【００４４】
検出の結果、他の話題語句が含まれていないと検出した場合、同一の文中に評価キーワード及び前記話題語句が含まれていると検出した場合、又は、評価キーワード及び前記話題語句が含まれている段落に他の話題語句が含まれていないと検出した場合には、前記検出された評価キーワードは、すべて前記話題語句に対する評価であると考えられるので、前記文書の更新日時を文書データベース１から取得し、当該更新日時と、検出した評価キーワードと、前記文書のＩＤと、前記話題語句とを組にしたデータを話題評価記憶部１０に通知して（ステップＳ２３）、他の文書ＩＤの文書に対する処理（ステップＳ１７～Ｓ２３）を行う。
一方、上記したいずれにも該当しない場合には、当該文書は前記話題語句に対する評価を有さないこと表しているので、他の文書についての処理（ステップＳ１７～Ｓ２３）を行う。
【００４５】
上記したステップＳ４が終了した後に、提示選択部１１が提示可能な情報を表す選択肢を情報提示部１２に提示させて、ユーザからの選択入力を受け付け（ステップＳ５）、情報提示部１２が提示選択部１１によって受け付けられた選択入力に応じて情報を提示する（

ステップＳ６）。
【００４６】
例えば、提示選択部１１によって、すべての話題語句の一覧についての選択入力が受け付けられた場合には、情報提示部１２は話題記憶部５からすべての話題語句を取り出して、リスト形式の一覧表示を行う。
また、提示選択部１１によってすべての話題カテゴリの一覧についての選択入力が受け付けられた場合には、情報提示部１２は話題カテゴリ記憶部７からすべての話題カテゴリ語句を取り出して、リスト形式の一覧表示を行う。
【００４７】
また、提示選択部１１によって特定の話題カテゴリに分類される話題語句の一覧についての選択入力が受け付けられ、更に、対象の話題カテゴリ語句の入力が受け付けられた場合には、情報提示部１２は、話題カテゴリ記憶部７から該当する話題カテゴリ語句に対応する複数の話題語句を取り出して、リスト形式の一覧表示を行う。これにより、ユーザに複数の関連する話題語句を把握させることができる。
また、提示選択部１１によって特定の話題語句に対する評価一覧についての選択入力が受け付けられ、更に、対象の話題語句の入力が受け付けられた場合には、情報提示部１２が、話題評価記憶部１０から話題語句に対応付けられている評価を取り出して、リスト形式の一覧表示を行う。これにより、ユーザに話題語句についてどのような評価がなされているのかを容易に把握させることができる。
【００４８】
また、提示選択部１１によって特定の話題カテゴリにおける話題語句の推移と評価の推移についての選択入力が受け付けられ、更に、対象の話題カテゴリ語句の入力が受け付けられた場合には、情報提示部１２は、話題カテゴリ記憶部７から話題カテゴリ語句に対応付けられている話題語句を抽出し、話題評価記憶部１０から当該抽出した話題語句に対応するデータを取り出して、これらデータに基づいて図７（ａ）に示すチャート形式で表示を行う。なお、図７（ａ）には、話題カテゴリ語句”ＣＰＵ”についての話題語句の推移と評価の推移を表示したものを示してある。
【００４９】
同図では、一つの横棒２１が一つの話題語句と対応付けられており、話題語句が文書に出現する期間を示している。そして、横棒には、吹き出し２２が併設されており、一つの吹き出し２２は、横棒２１の表す話題語句に対する評価を表す評価キーワード及び当該評価がされた日時を示している。なお、評価キーワードと当該評価がされた時間軸状の位置は、話題評価記憶部１０に記憶されたデータによって一意に決定できる。
上記したチャートによると、ユーザは、或る話題カテゴリ語句にまとめられる複数の話題語句の推移を容易に把握することができるとともに、各話題語句に対する評価の推移についても容易に把握することができる。
【００５０】
図７（ｂ）には、上記した図７（ａ）と吹き出しを表示する方法を変えたものを示してある。
この図７（ｂ）では、カーソル２３が接した位置に評価がある場合に当該評価の吹き出し２２を表示するものである。なお、吹き出しを表示する方法は、提示選択部１１により選択することができるようになっている。
【００５１】
なお、上記したチャートでは、話題語句の出現した期間を表す横棒２１が単色により描画されているが、当該横棒に対して、話題語句の出現する頻度によって、例えば、濃淡（頻度が高い場合に濃くする）或いは色相（頻度が高い場合に暖色を用い、頻度が少ない場合に寒色を用いる）を変えることにより、対応する話題語句がよく取り上げられているか否かをユーザに容易に把握させることができる。
なお、特定の話題カテゴリにおける話題語句の推移について、或いは、特定の話題語句に対する評価の推移については、上記した特定の話題カテゴリにおける話題語句の推移と評

価の推移についての表示動作の一部の動作によって、表示させることができる。

【００５２】

なお、上記した実施例では、話題カテゴリ語句として、複数の話題語句とともに文書に出現する割合の高い単語を用いていたが、これに限らず、例えば、話題語句の分類に適すると考えられる複数の話題カテゴリ語句の候補を予め保持しておき、当該保持している候補の中のいずれかが話題語句と同一文書に出現する場合に、当該出現する候補を前記話題語句の話題カテゴリ語句としてもよく、また、文書から話題語句の上位概念を表す語句を解析して、当該語句を話題カテゴリ語句としてもよく、要は、話題語句と同一の文書に含まれる他の語句であり、話題語句の分類に適した語句であればよい。

【００５３】

また、上記した実施例では、話題語句を分類するための話題カテゴリ語句、及び、話題語句に対する評価を文書から抽出して管理するようにしていたが、例えば、話題カテゴリ語句だけを抽出して管理するようにしてもよく、この場合には、複数の関連する話題語句を結び付けて提供することができる。また、話題語句に対する評価だけを抽出して管理するようにしてもよく、この場合には、話題語句に対する評価を容易に提供することができる。

【００５４】

なお、上記した各機能手段はプロセッサが（ＲＯＭ、ＲＡＭ等に予め格納された）制御プログラムを実行することにより構成されるが、本発明ではこれら機能手段を独立したハードウエア回路として構成してもよい。

また、本発明は上記の制御プログラムをフロッピーディスクやＣＤ－ＲＯＭ等といったコンピュータが読み取り可能な記憶媒体として把握させておき、当該制御プログラムを記憶媒体から読み取ってプロセッサに実行させることにより、本発明に係る処理を実行させることもできる。

【００５５】

【発明の効果】

以上説明したように、本発明では、複数の文書に含まれる特定の語句の時間的な分布を文書の更新日時に基づいて検出して、分布の強度が高い語句を話題語句として抽出し、話題語句と同一文書に含まれる他の語句をカテゴリ語句として抽出し、話題語句をカテゴリ語句を用いて分類して管理するようにしたために、関係を有する複数の話題語句を関連付けて管理することができる。したがって、ユーザに複数の関連する話題語句を提供することができる。

【００５６】

また、本発明では、複数の文書に含まれる特定の語句の時間的な分布を検出して、分布の強度が高い語句を話題語句として抽出し、話題語句と同一文書に含まれる話題語句に対する評価を表すキーワードを検出し、話題語句と検出されたキーワードとを対応付けて管理するようにしたために、話題語句に対する評価を管理することができる。したがって、ユーザに話題語句に対する評価を提供することができる。

更に、本発明では、文書から抽出して管理する情報に対して、情報を抽出した文書の更新日時を対応付けるようにしたために、情報を時系列に表示出力することができ、ユーザに関連する話題語句の推移を提供することや、話題語句に対する評価の推移を提供することができる。

【図面の簡単な説明】

【図１】　本発明の一実施例に係る情報抽出装置の構成図である。

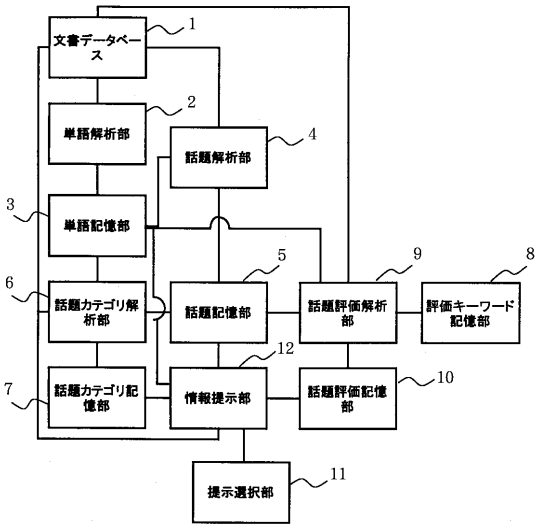
【図２】　本発明の一実施例に係る文書データベースの問い合わせ命令を説明する図である。

【図３】　本発明の一実施例に係る情報抽出装置の全体の処理動作を説明するフローチャートである。

【図４】　本発明の一実施例に係る話題解析部の処理動作を説明するフローチャートである。

【図５】　本発明の一実施例に係る話題評価解析部の処理動作を説明するフローチャートである。
【図６】　分布を説明する図である。
【図７】　本発明の一実施例に係る情報提示部による情報提示の一例を示す図である。
【符号の説明】
１・・文書データベース、　２・・単語解析部、
３・・単語記憶部、　４・・話題解析部、
５・・話題記憶部、　６・・話題カテゴリ解析部、
７・・話題カテゴリ記憶部、　８・・評価キーワード記憶部、
９・・話題評価解析部、　１０・・話題評価記憶部、
１１・・提示選択部、　１２・・情報提示部。

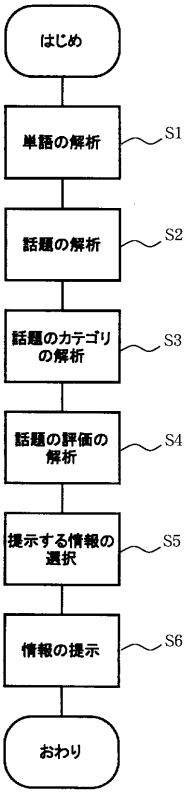
【図１】



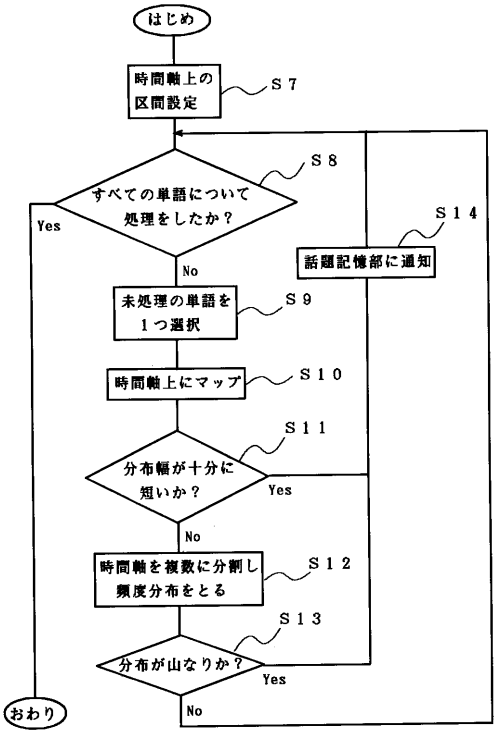
【図２】

問い合わせ命令	引数	問い合わせ結果
全ID	なし	文書データベースに記憶された文書すべての文書ID
文書内容	文書ID	引数の文書IDで示される文書の内容
更新日時	文書ID	引数の文書IDで示される文書の更新日時
最新の文書	なし	更新日時が最も新しい文書の文書ID
最古の文書	なし	更新日時が最も古い文書の文書ID

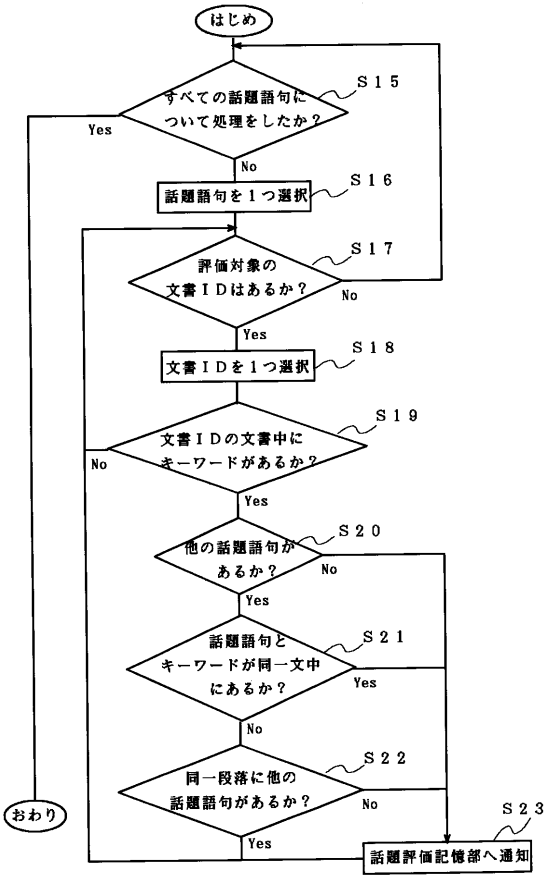
【図３】



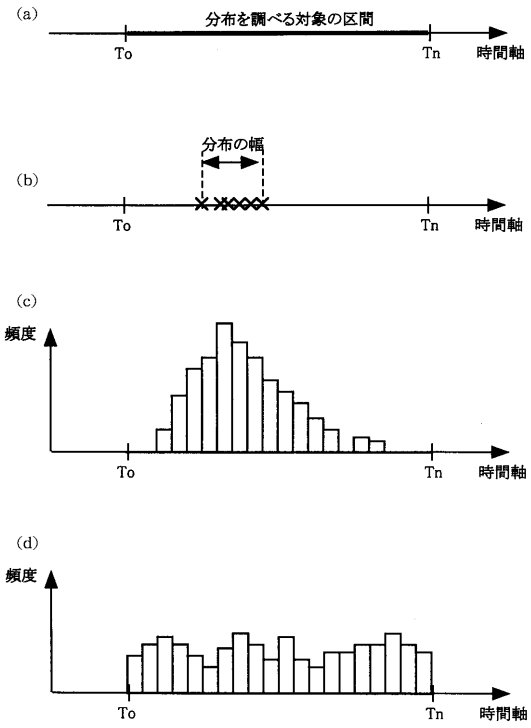
【図４】


はじめ
時間軸上の区間設定
S７
すべての単語について処理をしたか？
S８
Yes
No
S１４
話題記憶部に通知
未処理の単語を１つ選択
S９
時間軸上にマップ
S１０
分布幅が十分に短いか？
S１１
Yes
No
時間軸を複数に分割し頻度分布をとる
S１２
分布が山なりか？
S１３
Yes
No
おわり

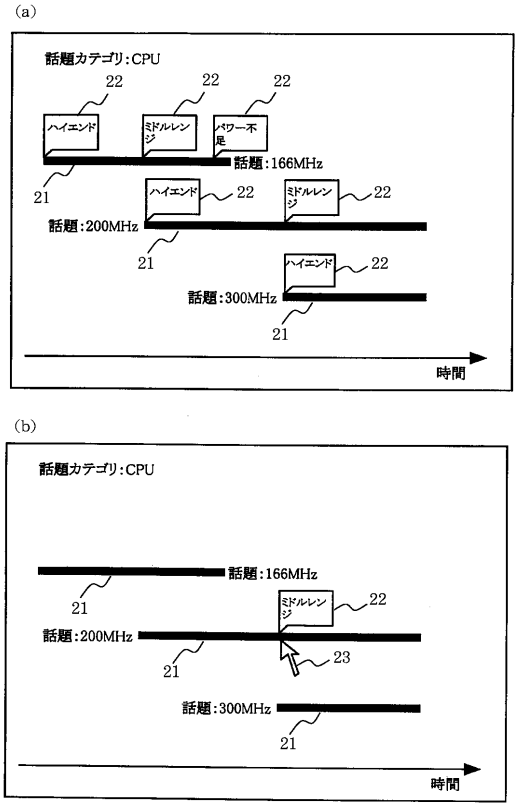
【図５】


はじめ
すべての話題語句について処理をしたか？
S１５
Yes
No
話題語句を１つ選択
S１６
評価対象の文書ＩＤはあるか？
S１７
No
Yes
文書ＩＤを１つ選択
S１８
文書ＩＤの文書中にキーワードがあるか？
S１９
No
Yes
他の話題語句があるか？
S２０
No
Yes
話題語句とキーワードが同一文中にあるか？
S２１
Yes
No
同一段落に他の話題語句があるか？
S２２
No
Yes
S２３
話題評価記憶部へ通知
おわり

【図６】


(a)
分布を調べる対象の区間
To
Tn
時間軸
(b)
分布の幅
To
Tn
時間軸
(c)
頻度
To
Tn
時間軸
(d)
頻度
To
Tn
時間軸

【図７】


(a)
話題カテゴリ:CPU
22
ハイエンド
ミドルレンジ
パワー不足
話題:166MHz
21
話題:200MHz
話題:300MHz
時間
(b)
話題カテゴリ:CPU
話題:166MHz
話題:200MHz
話題:300MHz
23
時間

フロントページの続き

審査官　深津　始

(56)参考文献　特開平８－７７１７８（ＪＰ，Ａ）
特開平７－３２５８３２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/30
G06F 19/00