(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2004-199680
(P2004-199680A)
(43)公開日　平成16年7月15日(2004.7.15)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	220A	5B009
G06F 17/21	G06F 17/30	170A	5B075
	G06F 17/21	550A	

審査請求 未請求　請求項の数 11　OL　（全 25 頁）

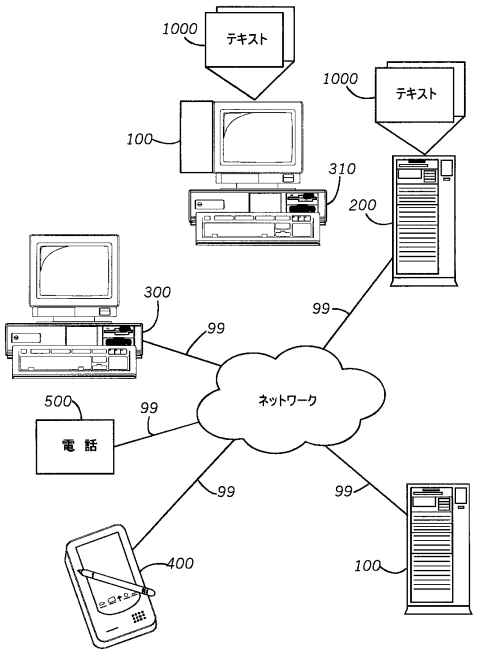
(21)出願番号　特願2003-416163(P2003-416163)
(22)出願日　平成15年12月15日(2003.12.15)
(31)優先権主張番号　10/319,545
(32)優先日　平成14年12月16日(2002.12.16)
(33)優先権主張国　米国(US)

(71)出願人　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　スタンフォード、ロング・リッジ・ロード　８００
(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　フラーンシーヌ　アール　チェン
アメリカ合衆国　カリフォルニア　メンロ　パーク　シャーマン　アベニュー　９７５

最終頁に続く

(54)【発明の名称】トピックベースのインタラクティブなテキストの要約を表示するシステムと方法

(57)【要約】
【課題】従来の要約システムは、導入テキストまたは結論テキストを選択するため、要約は表示できるが、トピックを元にした要約を示すことができないという問題があった。
【解決手段】要約するテキスト（１０００）をセグメント化し、離散キーワード、キーフレーズ、Ｎ－ｇｒａｍ、センテンスや他のセンテンス構成要素ベースの要約を、各テキストセグメントの統計測定に基づいて生成する。トピックベースのインタラクティブな要約は、別の色、フォント、音、触覚要素、またはその他の、省略されたテキストを示すために有用な、人間が感知可能な表示特徴などの、省略されたテキストの人間が感知可能なインジケータと共に表示する。そして、テキスト内のテキストのトピックとサブトピックの生成の概要を提供することで、トピックベースのインタラクティブなテキストの要約と元のテキストへの文脈化されたアクセスを提供する。
【選択図】図１



【特許請求の範囲】
【請求項１】
トピックベースのインタラクティブなテキストの要約を表示する方法であって、
前記方法は、
要約するテキストをダイナミックに決定するステップと、
前記決定されたテキストの少なくとも一部を要約するようにダイナミックに決定するステップと、
少なくとも１つの、センテンス構成要素ベースの要約のタイプを決定するステップと、
前記決定されたテキストの少なくとも一部と、前記少なくとも１つの決定されたセンテンス構成要素ベースの要約タイプとに基づいて、離散センテンス構成要素ベースのトピックベースのテキストの要約をダイナミックに表示するステップとを含む方法。
【請求項２】
前記少なくとも１つのセンテンス構成要素ベースの、トピックベースのインタラクティブなテキストの要約は、中黒、インデント、テキスト配置コンテキストキューのうち少なくとも１つに基づいて表示される請求項１に記載の方法。
【請求項３】
テキストを少なくとも１つのセグメントに再帰的にセグメント化するステップと、
少なくとも１つのセンテンス構成要素ベースの要約タイプをダイナミックに決定するステップと、
前記セグメントの選択、前記要約タイプ、テキストセグメントの特徴とテキストセグメントの区別の統計測定とに基づいて、少なくとも１つのセンテンス構成要素ベースのトピックベースのテキストの要約をダイナミックに決定するステップとをさらに含む請求項１に記載の方法。
【請求項４】
前記センテンス構成要素ベースの、トピックベースのテキストの要約で省略されたテキストと関連付けられた、少なくとも１つの省略されたテキストのインジケータを表示するステップをさらに含む請求項３に記載の方法。
【請求項５】
前記要約するテキストの一部をダイナミックに決定することは、反転表示、境界ボックス、カーソルによる選択、マウスによる選択、反応領域の上でマウスをロールオーバすること、マウスのクリック、スタイラスによる意思表示、音声コマンドのうち少なくとも１つを使用して要約するセグメントを選択することを含む請求項３に記載の方法。
【請求項６】
前記インタラクティブなトピックベースのテキストの要約の選択可能な部分を、前記センテンス構成要素ベースのトピックベースのテキストの要約から省略されたテキストと関連づけるステップをさらに含む請求項３に記載の方法。
【請求項７】
トピックベースのインタラクティブなテキストの要約を表示するシステムであって、
前記システムは、
選択されたセンテンス構成要素ベースの要約タイプ、選択されたテキスト、ダイナミックに決定された要約するテキストの部分を受信する入力／出力回路と、
選択されたテキスト、選択されたセンテンス構成要素ベースの要約タイプ、ダイナミックに決定された要約するテキストの部分を前記入力／出力回路から検索するプロセッサとを備え、
前記プロセッサは前記受信したセンテンス構成要素ベースの要約タイプ、テキストセグメントの特徴とテキストセグメントの区別の決定された分布と統計測定に基づいて、ダイナミックに決定されたテキストの部分に関する離散インタラクティブトピックベースの要約を決定し表示するシステム。
【請求項８】
前記プロセッサはダイナミックに決定されたセンテンス構成要素要約タイプと、要約す

るテキストの部分を受信する請求項７に記載のシステム。
【請求項９】
少なくとも１つの省略されたテキストインジケータが、前記センテンス構成要素ベースのインタラクティブなトピックベースのテキストの要約から省略されたテキストと関連づけられる請求項７に記載のシステム。
【請求項１０】
コンピュータで読み取り可能な記憶媒体であって、
コンピュータで読み取り可能な記憶媒体上に具体化された、コンピュータで読み取り可能なプログラムコードであって、トピックベースのインタラクティブなテキストの要約を表示するようにコンピュータをプログラミングするために使用できるプログラムコードを含み、前記プログラムコードは、
要約するテキストをダイナミックに決定するステップと、
前記決定されたテキストの少なくとも一部を要約するようにダイナミックに決定するステップと、
少なくとも１つの、センテンス構成要素ベースの要約のタイプを決定するステップと、
前記決定されたテキストの少なくとも一部と、前記少なくとも１つの決定されたセンテンス構成要素ベースの要約タイプとに基づいて、離散センテンス構成要素ベースの、トピックベースのテキストの要約をダイナミックに表示するステップとを含む、コンピュータで読み取り可能な記憶媒体。
【請求項１１】
トピックベースのインタラクティブなテキストの要約を表示するために使用可能なコントロールプログラムを、前記プログラムを実行する装置に送信するようにエンコードされた搬送波であって、
前記コントロールプログラムは、
要約するテキストをダイナミックに決定する命令と、
前記決定されたテキストの少なくとも一部を要約するようにダイナミックに決定する命令と、
少なくとも１つのセンテンス構成要素ベースの要約のタイプを決定する命令と、
前記決定されたテキストの少なくとも一部と、前記少なくとも１つの決定されたセンテンス構成要素ベースの要約タイプに基づいて、離散センテンス構成要素ベースの、トピックベースのテキストの要約をダイナミックに表示する命令とを含む搬送波。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、テキスト情報のトピックベースのインタラクティブな要約を表示することに関する。
【背景技術】
【０００２】
一部の要約システムは、統計上重要であるがばらばらなテキストの部分を選択し、表示する。従来の英語テキスト要約機能は、導入テキストまたは結論テキストを選択する。代表的なテキストは、テキストの情報内容を代表するものとして表示される。これらの要約は、単一のテキスト内で論じられたトピックを示すものではない。
【０００３】
【特許文献１】米国特許第５３８４７０３号明細書
【特許文献２】米国特許第５６８９７１６号明細書
【特許文献３】米国特許第５７７８３９７号明細書
【特許文献４】米国特許第５９１８２４０号明細書
【特許文献５】米国特許第６１２８６３４号明細書
【特許文献６】米国特許第５７４５６０２号明細書

【発明の開示】
【発明が解決しようとする課題】
【０００４】
従来の要約システムは、導入テキストまたは結論テキストを選択するため、要約は表示できるが、トピックを元にした要約を示すことができないという問題があった。
【課題を解決するための手段】
【０００５】
本発明によるシステムと方法は、トピックベースでインタラクティブなテキストの要約を表示する。本発明によるシステムと方法は、語、文字列のＮ－ｇｒａｍ（ある文字列の中でＮ個の文字列または単語の組み合わせがどの程度出現するかを調査する言語モデル解析方法）、キーフレーズ、センテンスなどの文の構成要素に基づいて、トピックベースのインタラクティブなテキストの要約を表示する。また、本発明によるシステムと方法はトピックベースのハイアラーキ状の、インタラクティブなテキストの要約を提供し、表示する。
【発明を実施するための最良の形態】
【０００６】
図１は、トピックベースのインタラクティブなテキスト要約システム１００の例の概観図である。トピックベースのインタラクティブな要約システム１００は、通信リンク９９を介して、テキスト情報１０００を含む情報リポジトリ２００、Ｗｅｂ使用可能パーソナルコンピュータ３００、Ｗｅｂ使用可能装置４００、電話５００に接続する。テキスト情報１０００を含むパーソナルコンピュータ３１０は、トピックベースのインタラクティブなテキスト要約システム１００を組み込んでいる。
【０００７】
トピックベースのインタラクティブなテキスト要約システム１００は、情報リポジトリ２００に記憶されたテキスト情報１０００のトピックベースのインタラクティブな要約に対する要求を、Ｗｅｂ使用可能装置４００のユーザから要求を受信する。テキスト情報１０００のトピックベースのインタラクティブなテキスト要約に対する要求は、トピックベースのインタラクティブな要約システム１００が受信し、システム１００はテキスト情報１０００に対する要求を情報リポジトリ２００に転送する。情報リポジトリ２００は要求されたテキスト情報１０００を検索し、これを、トピックベースのインタラクティブな要約システム１００に戻す。トピックベースのインタラクティブなテキストの要約は、ユーザの相互作用、以前に決定されたかおよび／または同時に決定されたユーザの設定に基づいて、トピックベースのインタラクティブな要約システム１００が決定し、通信リンク９９を介してＷｅｂ使用可能装置４００に送信する。
【０００８】
トピックベースのインタラクティブなテキスト要約システム１００および／または情報リポジトリ２００はまた、単一の装置に組み込むこともできる。トピックベースのインタラクティブなテキスト要約システム１００は、ソフトウェア、ルーチン、マネジャーなどとしてパーソナルコンピュータ３１０に組み込まれる。パーソナルコンピュータ３１０内の、トピックベースのインタラクティブな要約システム１００は、ＨＴＭＬファイル、ＷＭＬファイル、ＰＤＦファイル、マイクロソフト社の製品であるワード（登録商標）ファイル、会話認識オーディオテキスト情報など、パーソナルコンピュータ３１０に接続されたコンピュータ読み取り可能媒体および／またはメモリ上でアクセス可能なテキスト情報を記憶する任意のファイルフォーマットで動作する。
【０００９】
Ｗｅｂ使用可能コンピュータ３００のユーザは、情報リポジトリ２００内に含まれるテキスト情報１０００の、トピックベースのインタラクティブなテキスト要約を要求することができる。別法としては、電話５００のユーザは、音声、ＶＭＬおよび／またはＶＸＭＬベースのブラウザなどを介して、情報リポジトリ２００からテキスト情報１０００の、トピックベースのインタラクティブなテキストの要約を要求する。テキスト情報１０００

は、自動会話認識転写システム、光学文字認識システム出力などからのテキストなど、任意のタイプのテキスト情報を含んでいてもよい。

【００１０】

図２は、本発明による、トピックベースのインタラクティブなテキスト要約に関する例としての方法のフローチャートである。このプロセスはステップＳ１０００で開始し、すぐにオプションのステップＳ２０００に続く。オプションのステップＳ２０００では、練習テキストを決定する。練習テキストは、Ｂｒｏｗｎ　Ｃｏｒｐｕｓのロイター２１５７８データ、または、要約すべき情報を統計的に表す任意の他のテキストまたはテキスト群を含む。コントロールは、オプションのステップＳ３０００に続く。

【００１１】

オプションのステップＳ３０００では、練習テキストの中の潜在クラスの数を決定する。潜在クラスの数は、所与のテキストまたはテキストの集合に関して人間が割り当てたトピックの数に基づくか、または、クラスタ分析に基づくか、テキストまたはテキストの集合内の潜在クラスを決定する任意の方法に基づくかのいずれかによって、あらかじめ決定した数でもよい。コントロールはオプションのステップＳ４０００に続く。

【００１２】

確率的潜在意味解析モデルを、練習テキストと、オプションのステップＳ４０００で決定した潜在クラスの数に基づいて決定する。同時係属出願の米国特許出願第１０／１０３，０５３号（今後、Ｔｓｏｃｈａｎｔａｒｉｄｉｓらと呼ぶ）と、カリフォルニア２０００、バークレーのＰｒｏｃｅｅｄｉｎｇｓ　ｏｆ　ｔｈｅ　ＳＩＧＩＲ－９９の３５ページから４４ページの、Ｔ．ホフマンによる「確率的潜在意味インデックス」では、確率的潜在意味解析モデルを論じている。これらは各々、参照により全体として本明細書に組み込まれている。ステップＳ２０００からＳ４０００はオプションである。すなわち、確率潜在意味解析モデル、潜在意味解析モデル、その他の技術など、分布を決定する任意の方法も使用することができる。コントロールはステップＳ５０００に続く。

【００１３】

ステップＳ５０００では、要約するテキストを選択する。要約テキストは、テキストの一部を反転表示するか、または、全テキストを選択するか、キーボードを使ってテキストを手で入力するか、音声または手書き認識を使ってテキストを入力するか、メモリからテキストを選択するか、すべてのテキストを自動的に選択するか、携帯情報端末識別子などのパラメータに基づくなど、テキストを選択する任意の知られた方法または最近開発された方法に基づいて、選択することができる。

【００１４】

Ｗｅｂ使用可能装置は、限定された大きさの画面ディスプレイを有する。ディスプレイのサイズが限定されているため、これらの装置上で大きなテキストを見ることは難しく間違いを起こしやすい。Ｗｅｂ使用可能装置および／またはＷｅｂ使用可能装置用の媒体通信と関連づけられた、Ｗｅｂ使用可能装置内の、トピックベースのインタラクティブなテキスト要約システムは、Ｗｅｂ使用可能装置に関して決定された設定に基づいて、受信した各テキストに関する、トピックベースのインタラクティブなテキストの要約を自動的に決定する。

【００１５】

これらの装置の限定されたディスプレイをよく利用しながら、テキストの離散トピック部分に関する情報を提供する。テキストに関して決定された、トピックベースのインタラクティブなテキストの要約の一部を選択的に拡大し、テキスト内の離散トピックを限定されたディスプレイ上で容易に見ることができる。コントロールはステップＳ６０００に続く。

【００１６】

ステップＳ６０００では、選択されたテキストを離散セグメントにセグメント化する。選択されたテキストは自動的にセグメント化してもよいし、ＨＴＭＬ、または、パラグラフインジケータ、セクションの見出し、任意の他のテキストのマークアップを使用してセ

グメント化する。テキストは米国特許出願第１０／１０３，０５３号の方法を使用して自動的にセグメント化することができる。

【００１７】

テキストのテキストセグメント境界は、反転表示、マーキング、セグメント境界インジケータの挿入などを使用して手動で示してもよい。しかし、パラグラフ、および／または、セクションの見出しタイトル、他のテキストマークアップに基づいたセグメント化、手によるセグメント化、確率潜在意味解析や統計的な類似性に基づいた自動セグメント化、または他のセグメント化の方法を、単独または組み合わせて使用して、離散テキストセグメントを作成することもできる。コントロールはステップＳ７０００に続く。

【００１８】

ステップＳ７０００では、各テキストセグメントに対して、離散トピックベースのインタラクティブなテキスト要約のタイプを決定する。キーワード、キーフレーズ、センテンス、名詞のフレーズ、文字列のＮ－ｇｒａｍに基づいた、トピックベースのインタラクティブなテキストの要約を、単独または組み合わせて使用することができる。しかし、要約情報を伝えるために有用な任意の知られたセンテンスの構成要素または後から識別されたセンテンスの構成要素を使用してもよい。決定されたセンテンスの構成要素はテキスト内に実際に存在する必要はない。その代わりに、キーワード、キーフレーズ、センテンス、名詞フレーズ、文字列のＮ－ｇｒａｍなどを、要素を変形して決定することができる。要素には、テキストに含まれるセンテンスの構成要素が含まれるが、これに限定されるものではない。ＷＯＲＤＮＥＴルックアップ、同義語の選択、変形などを使用して、センテンスの構成要素に基づいてキーワード、および、キーフレーズなどを決定することができる。選択された要約テキストは、いくつかの異なるテキストの連結でもよい。

【００１９】

例としての離散キーワード、および／または、キーフレーズ、センテンスに基づいた、トピックベースのインタラクティブな要約は、図３から図６のフローチャートに示されたように決定する。キーワード、および／または、キーフレーズ、センテンスベースの要約の選択は、ユーザの環境設定に基づいて決定してもよいし、装置の特性に基づいてあらかじめ決定するかおよび／または自動的に構成してもよいし、ユーザが調整できるようにしてもよいし、および／または、他の任意の技術を使用して決定してもよい。

【００２０】

例としての実施形態は、例としての離散キーワード、および／またはキーフレーズ、センテンスベースの、トピックベースのインタラクティブなテキストの要約の連続的な決定を記述しているが、離散キーワード、および／または、キーフレーズ、センテンスベースの、トピックベースのインタラクティブなテキストの要約は、連続的に決定してもよいしおよび／または同時に決定してもよく、単独で表示してもよいし組み合わせて表示してもよい。

【００２１】

トピックベースのインタラクティブなテキストの要約を決定する方法は、テキスト、テキストのセグメントなどのテキスト領域、および／または、テキストの多数のセグメントまたはテキスト領域に再帰的に適用することができる。テキスト内のトピックおよびサブトピック生成のナビゲーションおよび／または理解が促進される。決定された離散キーワード、および、キーフレーズ、センテンスベースの、トピックベースのインタラクティブなテキストの要約を組み合わせて、テキストのトピックの概要またはインデックスを提供する際に有用な、トピックベースのインタラクティブなテキストの要約を形成する。ユーザがより興味を持つトピックに関連づけられたテキストが、より容易に識別される。コントロールはステップＳ８０００に続く。

【００２２】

ステップＳ８０００では、要約すべき任意の追加のテキストがあるかどうかを決定する。トピックベースのインタラクティブなテキスト要約セッションの終了は、メニューオプションを選択すること、キーボードシーケンスを使用すること、または、音声コマンドを

発すること、マウスまたはスタイラスを使用して終了の意思表示を入力することなど、トピックベースのインタラクティブなテキスト要約セッションを終了させる任意の知られた方法または最近開発された方法を使用することによって指示できる。

【００２３】

ステップＳ８０００で、ユーザが、トピックベースのインタラクティブなテキスト要約セッションをまだ終了させていないと決定された場合、コントロールはすぐにステップＳ５０００にジャンプする。ステップＳ５０００からＳ８０００は、トピックベースのインタラクティブなテキストの要約セッションが終了したという決定が行われるまで繰り返す。次のステップＳ５０００からＳ８０００の繰り返しの間に、離散キーワード、キーフレーズ、文字列のＮ－ｇｒａｍ、センテンスベースの、トピックベースのインタラクティブなテキスト要約の組み合わせの選択を行い、ステップＳ７０００で表示する詳細なレベルを調節する。トピックベースのインタラクティブなテキスト要約セッションが終了すると、コントロールはステップＳ８０００からＳ９０００に続き、プロセスが終了する。

【００２４】

図３は、本発明による、離散キーワードベースの要約を決定する、例としての方法の拡大フローチャートである。図２のステップＳ７０００でユーザが、離散キーワードベースの、トピックベースのインタラクティブなテキストの要約を要求した場合、コントロールは図２のステップＳ７０００から、図３のステップＳ７１００に続く。ステップＳ７１００では、コントロールはすぐにステップＳ７１１０に続き、各テキスト領域ｒの潜在クラスｚの確率Ｐ（ｚ｜ｒ）を決定する。コントロールはステップＳ７１２０に続く。

【００２５】

ステップＳ７１２０では、テキスト領域ｒに発生する語ｗの確率Ｐ（ｗ｜ｒ）を決定する。確率潜在意味解析で練習テキストまたは練習テキストの組を解析することによって、所与の潜在クラスｚの語ｗの条件付確率Ｐ（ｗ｜ｚ）を決定する。条件付確率Ｐ（ｗ｜ｚ）は、確率潜在意味解析モデルが生成されるか、ルックアップテーブルから検索されるか、または、任意の知られた方法や最近開発された方法を使用して決定されると、決定してメモリ内に記憶することができる。これによって、必要な追加の計算の数が減る。Ｐ（ｗ｜ｚ）とＰ（ｚ｜ｒ）に基づいたＰ（ｗ｜ｒ）の決定は例示的なものに過ぎず、領域内のカウントの決定、潜在意味解析の実行などのＰ（ｗ｜ｒ）を決定する任意の方法を使用することもできる。

【００２６】

テキスト領域ｒを畳み込むか、または、テキスト領域の低次元の表示を決定する任意の方法を使用することによって、テキスト領域ｒ上の潜在クラスｚの確率Ｐ（ｚ｜ｒ）を決定する。それから所与のテキスト領域の語ｗの確率は次式に基づいて決定する。

【数１】

$$P(w|r)=\sum_{z}P(w|z)P(z|r) \qquad (1)$$

確率Ｐ（ｗ｜ｒ）を決定した後、コントロールはステップＳ７１２５に続く。ステップＳ７１２５では、テキスト領域の特性と他の領域との区別の統計測定に基づいてキーワードのランクを決める。しかし、テキスト領域特性および／または他の領域との区別の任意の測定を使用することができる。コントロールはステップＳ７１３０に続く。

【００２７】

ステップＳ７３１０では、決定された確率Ｐ（ｗ｜ｒ）に基づいて、表示するためのキーワード要素を選択する。ついで選択されたキーワードをユーザに表示する。離散キーワードベースの、トピックベースのインタラクティブなテキスト要約は、省略されたテキストのインジケータと共に表示することができる。

【００２８】

「＋」文字などの、人間が感知できる表示特徴を使用して、省略されたテキストを示すことができる。省略されたテキストのインジケータは、省略されたテキストのすぐ前また

はすぐ後に続く。省略されたテキストのインジケータは、省略されたテキストをダイナミックに表示および／または拡大するように選択できる。マウスのロールオーバ、または、マウスを一回左クリックすること、その他の選択技術を使用して、省略されたテキストを拡大した表示用に選択する、または、全テキストを拡大するなどを行う。再帰的にセグメント化されたテキストの再帰的な拡大もダイナミックに表示することができる。テキストを代表するトピックのセグメントを再帰的にセグメント化してサブトピック生成を示すことができる。トピックとサブトピックの生成は、インデント、中黒などのサブトピック要素を置いて、トピックベースのインタラクティブなテキストの要約と共に文脈の中で表示できる。サブトピックの省略されたテキストは、上述のマウスのロールオーバ、および、マウスの右クリック、上記に論じた他の選択方法に応答して、ポップアップウィンドウや他のテキスト表示機構を使用して表示される。同様に、マウスのロールオフ、または、続くマウスのクリック、その他の選択方法を使用して、トピックベースのインタラクティブなテキストの要約と関連づけられたテキストを畳み込む。フォントの変化、または、音の変化、ダイナミック点字出力要素、または、人間が感知可能な知られた表示特徴または最近開発された表示特徴が間歇的または可変的に存在することによって、省略されたテキストを示す。ハイアラーキ状に拡大された、トピックベースのインタラクティブなテキストの要約は、任意のレベルで選択的に畳み込む。

【００２９】

文書内のトピック生成ハイアラーキの文脈は、全体のテキストの中で要約されたテキストの位置に対応する、キーワードベースの要約をインデントすることによって示すことができる。しかし、中黒、または、フォントの変化、サムネール画像、ポップアップウィンドウ内のキーワードの反転表示など、他の表示特徴も使用することができる。コントロールはステップＳ７１４０に続き、すぐに図２のステップＳ８０００に戻る。

【００３０】

図４は、本発明による、離散キーフレーズベースの、トピックベースのインタラクティブなテキストの要約を決定する、例としての方法のフローチャートである。図２のステップＳ７０００で、ユーザが、キーフレーズベースの、トピックベースのインタラクティブなテキストの要約を要求した場合、コントロールは図２のステップＳ７０００から、図４のステップＳ７２００に続く。ステップＳ７２００では、コントロールはすぐにステップＳ７２１０に続く。

【００３１】

ステップＳ７２１０では、候補となるキーフレーズの最大長を決定する。ユーザが調節できる、候補となるキーフレーズの最大長は、最も長いキーフレーズ内の語の数を調節する。キーフレーズはＮ－ｇｒａｍであってもよいし、および／または、任意の言語的動機によるフレーズであってもよい。候補となるキーフレーズの最大長を調節することにより、異なるユーザのディスプレイサイズまたは他のパラメータに関してキーフレーズが調節できる。コントロールはステップＳ７２２０に続き、領域ｒ上の潜在クラス確率Ｐ（ｚ｜ｒ）を、畳み込みによって決定する。コントロールはステップＳ７２３０に続く。

【００３２】

ステップＳ７２３０では、セグメントまたはテキスト領域ｒ上のキーフレーズの確率Ｐ（k_i｜ｒ）を、潜在クラス上のキーフレーズの確率と、セグメントまたはテキスト領域ｒ上の潜在クラスの確率に基づいて決定する。しかし、練習テキスト、確率潜在意味解析、潜在意味解析、ルックアップテーブルに基づいた決定など、条件付確率Ｐ（k_i｜ｒ）を決定する任意の方法を使用できる。各キーフレーズは最高でＮワードを含み、ここでＮは候補となるキーフレーズの最大長である。確率潜在意味解析を使用する時、セグメントまたはテキスト領域ｒに関する確率Ｐ（k_i｜ｒ）を、次式に基づいて決定する。

【数２】

$$P(k_i \mid r) = \sum_z P(k_i \mid z)P(z \mid r) \qquad (2)$$

確率Ｐ（k_i｜ｒ）を決定した後、コントロールはステップＳ７２４０に続く。

【００３３】

ステップＳ７２４０では、キーフレーズ確率を調節する。この調節により、キーフレーズが短いと確率が高くなること、キーフレーズが長いと確率が低くなることが補償される。キーフレーズの確率は、次式によって、キーフレーズ長に基づいて正規化することができる。

【数３】

$$P(k_i \mid r)^{1/m_i} \qquad (3)$$

上式で、m_iはキーフレーズk_iの長さである。しかし、長さを補償する任意のキーフレーズ確率正規化を使用することができる。コントロールはステップＳ７２５０に続く。

【００３４】

ステップＳ７２５０では、相互情報量、または、χ^2とＧ^{2}の分布、対数オッズ比、分布を比較するために有用な任意の統計測定などの統計測定に基づいてキーフレーズのランクを決定する。セグメントまたはテキスト領域の組Ｒを伴う相互情報量Ｉ（k_i；Ｒ）は、次式に基づいて決定できる。

【数４】

$$I(k_i;R) = H(k_i) - H(k_i \mid R) \qquad (4)$$

$$= \sum_{f_i \in \{0,1\}} \sum_{r \in R} P(f_i, r) \log \frac{P(f_i, r)}{P(f_i)P(r)}$$

上式で、エントロピＨ（ｘ）は、次式の通りである。

【数５】

$$\mathrm{H(x)} = -\sum_{x \in X} P(x) \log P(x) \qquad (5)$$

上式で、f_iはキーフレーズk_iの存在を示す略語である。確率Ｐ（f_i、ｒ）は、データ内のカウント発生から導出され、Ｐ（f_i）とＰ（ｒ）は、結合分布のそれぞれの限界である。

【００３５】

セグメントまたはテキスト領域を特徴づけるか、および／または、他のセグメントまたはテキスト領域から区別するために、各キーフレーズの全体の寄与度を決定する。相互情報量は、セグメントまたはテキスト領域内のキーフレーズの存在／不在、セグメントまたはテキスト領域外のキーフレーズの存在／不在に基づいて、４つの成分に分割できる。セグメントまたはテキスト領域内のキーフレーズの存在は（ｆ＝１）という項によって示し、不在は（ｆ＝０）という項によって示す。セグメントまたはテキスト領域は項（ｒ）によって示し、現在のセグメントまたはテキスト領域外のセグメントまたはテキスト領域は（－ｒ）によって示す。従って、相互情報量の４つの構成要素または属性は、（ｆ＝１、ｒ）、（ｆ＝０、ｒ）、（ｆ＝１、－ｒ）、（ｆ＝０、－ｒ）に対応する。

【００３６】

セグメントまたはテキスト領域を特徴づけるためには、セグメントまたはテキスト領域内のキーフレーズの存在（ｆ＝１、ｒ）に関連する相互情報量加重が高くなければならない。また、より高い値はより独特のキーフレーズに関連するので、他のセグメントまたはテキスト領域からのキーフレーズの不在（ｆ＝０、－ｒ）に関連する、相互情報量加重も高くなければならない。

【００３７】

対照的に、セグメントまたはテキスト領域（ｒ）内の不在（ｆ＝０）に関する相互情報量加重は、そのキーフレーズがテキストの特徴づけにはあまり役に立たないことを示す。

したがって相互情報量成分（ｆ＝０、ｒ）は、低い値を有する。他のセグメントまたはテキスト領域（－ｒ）内の存在（ｆ＝１）に関する相互情報量成分は、そのキーフレーズが、テキストの他のセグメントまたはテキスト領域に関してそれほど独特ではないことを示す。従って、相互情報量成分（ｆ＝１、－ｒ）は低い値を有する。

【００３８】

相互情報量と相互情報量の成分を使用して、セグメントテキスト領域の特徴と、他のセグメントまたはテキスト領域との区別を決定することができる。しかし、セグメントまたはテキスト領域の特徴を決定するおよび／または他のセグメントまたはテキスト領域と区別する、任意の方法を使用することもできる。セグメントまたはテキスト領域内のキーフレーズの存在に関するランク付けまたはスコアは、次式によって相互情報量ＣＩ（k_i；ｒ）の成分から決定することができる。

【数６】

$$CI(k_i;r) = P(k_i,r)\log\frac{P(k_i,r)}{P(k_i)P(r)} \qquad (6)$$

コントロールはステップＳ７２６０に続く。

【００３９】

ステップＳ７２６０では、離散キーフレーズを、相互情報量および／または統計測定のランク付けまたはスコアに基づいて選択する。各セグメントまたはテキスト領域について、トップのＬランクになった離散キーフレーズを選択することができる。しかし、任意の離散キーフレーズ選択方法を使用することができる。

【００４０】

第１のトップにランクされたフレーズを選択することができる。キーフレーズは、離散マルチキーフレーズベースの、トピックベースのインタラクティブなテキストの要約に追加する。選択されたキーフレーズの要素を、セグメントまたはテキスト領域ｒから除去し、r_{new}を作成し、新しい潜在クラスの組を決定する。新しい確率Ｐ（k_i｜r_{new}）、新しい相互情報量および／または統計測定スコアと、新しいキーフレーズのランク付けを決定し、最もランクの高いキーフレーズの新しい組を決定し選択する。セグメントまたはテキスト領域の、離散マルチキーフレーズのトピックに基づくインタラクティブなテキスト要約のために必要な数のキーフレーズを選択するまで、このプロセスを続ける。

【００４１】

現在のキーフレーズベースの、トピックベースのインタラクティブな要約をユーザに表示する。キーフレーズベースの、トピックベースのインタラクティブなテキストの要約は、省略したテキストの前または後の特殊文字、または、フォント、色、その他の、省略されたテキストを示すために有用な、人間が感知可能な任意の知られた表示特徴などの、省略されたテキストのインジケータを使用して、省略したテキストを示し、コントロールはステップＳ７２７０に続く。コントロールの流れは図４のステップＳ７２７０から図２のステップＳ８０００に戻り、ここで、処理は図２に関して上述したように続く。

【００４２】

図５は、本発明による、離散センテンスベースの、トピックベースのインタラクティブなテキストの要約を決定する、センテンスレベルの、第１の例としての方法である。図２のステップＳ７００でユーザが、離散センテンスベースの、トピックベースのインタラクティブなテキストの要約を要求した場合、コントロールは図２のステップＳ７０００から、図５のステップＳ７５００に続く。ステップＳ７５００では、コントロールはすぐにステップＳ７５１０に続き、所与のセグメントまたはテキスト領域ｒの潜在クラスの条件付確率を、畳み込みを使用して決定する。コントロールはステップＳ７５１０に続く。ステップＳ７５１０では、セグメントまたはテキスト領域ｒ内の項w_iの確率を、確率潜在意味解析を使用して、所与の潜在クラスの項w_iの条件付確率Ｐ（w_i｜ｚ）と、セグメントまたはテキスト領域上の潜在クラスの確率とに基づいて決定する。先に確率潜在意味解析モデルを生成する間に決定され記憶された、Ｐ（w_i｜ｚ）などの条件付確率を検索し、

使用することができる。ついでコントロールはステップＳ７５２０に続く。

【００４３】

ステップＳ７５２０では、項w_iの存在を示す関数であるf_iと、セグメントまたはテキスト領域（ｒ）の間のポイントごとの相互情報量Ｉ（f_i、ｒ）を決定する。ポイントごとの相互情報量は、次式に従って決定することができる。

【数７】

$$I(f_i,r)=\log\frac{P(f_i\mid r)}{P(f_i)}=\log\frac{P(f_i,r)}{P(f_i)P(r)} \qquad (7)$$

上式でＰ（f_i）Ｐ（ｒ）とＰ（f_i、ｒ）は、センテンス上を合計することによって計算し、Ｐ（f_i、ｒ）は項w_iを含むトピックセグメントまたは領域ｒ内のセンテンスの数である。しかし、任意の統計測定を使用することができる。コントロールはステップＳ７５３０に続く。

【００４４】

ステップＳ７５３０では、セグメントまたはテキスト領域（ｒ）内の項の存在（f_i＝１）と、他のセグメントまたはテキスト領域（－ｒ）の項の不在（f_i＝０）に基づいて、項領域スコアを決定する。決定は次式に基づく。

【数８】

$$Score_{w_i}=\lambda P(f_i=1\mid r)I(f_i=1,r)+(1-\lambda)P(f_i=0\mid\neg r)I(f_i=0,\neg r) \qquad (8)$$

パラメータλのデフォルトの値は０．５であるが、セグメントまたはテキスト領域内の語の存在または不在の相対的な重要性を表すために調節することができる。

【００４５】

対数オッズ比は次式によって定義する。

【数９】

$$Score(w_i)=\log\frac{P(f_i=1\mid r)/P(f_i=0\mid r)}{P(f_i=1\mid\neg r)/P(f_i=0\mid\neg r)} \qquad (9)$$

対数オッズ比は、セグメントまたはテキスト領域内の、テキストに対する項の類似性と、他のセグメントまたはテキスト領域内のテキストからの非類似性を示す、代替の測定として使用する。しかし、相互情報量、χ^2とG^2の分布、対数オッズ比、または、セグメントまたはテキスト領域に対する項の分布の類似性および／または、他のセグメントまたはテキスト領域に対する非類似性を比較するために有用な任意の測定を使用することができる。コントロールはステップＳ７５４０に続く。

【００４６】

ステップＳ７５４０では、センテンスのスコアを、項の加重平均に基づいて決定する。項のサブセットは、ストップリストなどに現れる語を無視するように考慮することができる。たとえば、Ｐ（w_i）はストップリスト上で語を無視する、センテンス内のw_iの相対頻度である。ついで次式に基づいてセンテンススコアを決定する。

【数１０】

$$Score(S)=\sum_{w_i\in S}P(w_i)Score(w_i) \qquad (10)$$

上式で、w_iはセンテンスＳ内の項に対応する。各センテンスのスコアを比較する。セグメントまたはテキスト領域に関するセンテンスレベルの、トピックベースのインタラクティブな要約を、高いスコアに基づいて選択する。

【００４７】

領域内の候補センテンスの位置など追加の特徴を使用して、選択するセンテンスのラン

クを決定するために使用するスコアをさらに調節することもできる。所定のテキストでは、重要な情報はテキストの最初に置かれるか最後に置かれる場合がある。スコアの調節は、センテンスの位置、または、センテンスの配置、インジケータフレーズ、キューフレーズ、重要な情報を識別する際に有用な任意のテキスト特性など、テキストの属性に基づいて行うことができる。コントロールはステップＳ７５５０に続く。

【００４８】

ステップＳ７５５０では、ランクまたはスコアがもっとも高いセンテンスを選択し、トピックベースのインタラクティブなテキストの要約に追加する。選択したセンテンスは、非本質的な要素をトピックベースのインタラクティブなテキストの要約に追加する前に除去して圧縮してもよい。除去する要素は、統語解析、および／または、意味解析、統計フィルタリングなど、センテンス内の非本質的な要素を決定する任意の方法を使用して決定する。コントロールはステップＳ７５６０に続き、センテンスベースの、トピックベースのインタラクティブなテキストの要約のために追加のセンテンスが必要かどうかを決定する。追加のセンテンスが必要な場合、コントロールはステップＳ７５７０に続く。

【００４９】

ステップＳ７５７０では、選択されたセンテンスに関連づけられた項のカウントを差し引く。これによって、マルチセンテンスベースの、トピックベースのインタラクティブなテキストの要約内の同じ項の繰り返しが減る。コントロールはステップＳ７５８０に続き、確率を調節して差し引いた項のカウントを補償する。このようにして、項のカウントをゼロにすることができる。コントロールはステップＳ７５２０に飛ぶ。ステップＳ７５６０で、離散センテンスベースの、トピックベースのインタラクティブなテキストの要約のために追加のセンテンスが必要ではないと決定されるまで、ステップＳ７５２０からＳ７５６０を繰り返す。コントロールはステップＳ７５９０に続く。

【００５０】

ステップＳ７５９０では、離散センテンスベースの、トピックベースのインタラクティブなテキストの要約を表示する。上記のように、センテンスベースの、トピックベースのインタラクティブなテキストの要約は、人間が感知可能な表示特徴を使用して、離散トピックベースのインタラクティブなテキストの要約内で省略されたテキストを示すことができる。コントロールはオプションのステップＳ７５９５に続く。オプションのステップＳ７５９５では、選択されたセンテンスの非本質的な要素を除去する。選択されたセンテンスの非本質的な要素は、統語解析、または、意味解析、統計フィルタリングなど、センテンスの非本質的な要素を除去する際に有用な任意の知られた方法に基づいて除去する。別のオプションの動作では、テキストを調節して読みやすさを向上させる。充填語を追加する、または、前方照応的指示を解決するなど、テキストの読みやすさを向上させる任意の方法を使用する。コントロールはステップＳ７６００に続く。コントロールの流れは、図５のステップＳ７６００から図２のステップＳ８０００に戻り、図２に関して上述したように処理を継続する。

【００５１】

図６は、本発明による、離散センテンスベースの、トピックベースのインタラクティブなテキストの要約を決定する第２の例としての方法のフローチャートである。図２のステップＳ７０００でセンテンスベースの、トピックベースのインタラクティブなテキストの要約が要求されると、コントロールは図２のステップＳ７０００から図６のステップＳ７３００に続く。ステップＳ７３００では、コントロールはすぐにステップＳ７３１０に続く。

【００５２】

ステップＳ７３１０では、候補となるキーフレーズの最大長を決定する。上述のように、候補となるキーフレーズの最大長は、ユーザが調節できる。候補となるキーフレーズの最大長は、ユーザが調節可能なデフォルトの数、または、以前に保存したユーザのセッション、装置の表示特徴、累積的および／または集合的使用など、候補となるキーフレーズの最大長を決定する任意の方法を使用して決定する。コントロールはステップＳ７３２０

に続き、領域ｒ上の潜在クラス確率Ｐ（ｚ｜ｒ）を畳み込みによって決定する。コントロールはステップＳ７３３０に続く。

【００５３】

セグメントまたはテキスト領域（ｒ）内のキーフレーズの確率Ｐ（k_i｜ｒ）は、確率潜在意味解析を介して、所与の潜在クラスのキーフレーズの条件付確率Ｐ（k_i｜ｚ）と、セグメントまたはテキスト領域上（ｒ）の潜在クラスの確率に基づいて決定する。Ｐ（k_i｜ｚ）などの条件付確率は、練習テキスト、および／または、確率潜在意味分析、モデル生成、ルックアップテーブルからも決定できる。さらに、Ｐ（k_i｜ｚ）とＰ（ｚ｜ｒ）に基づいたＰ（k_i｜ｒ）の決定は例示的なものであり、テキストの潜在意味解析などの任意の方法も使用することができる。各キーフレーズは最高でＮキーフレーズ語からなり、Ｎは決定されたキーフレーズの最大長である。セグメントまたはテキスト領域ｒの確率Ｐ（k_i｜ｒ）は次式に基づいて決定する。

【数１１】

$$P(k_i \mid r) = \sum_z P(k_i \mid z)P(z \mid r) \qquad (11)$$

確率Ｐ（k_i｜ｒ）を決定した後、コントロールはステップＳ７３４０に続く。

【００５４】

ステップＳ７３４０では、キーフレーズ長に基づいて確率を正規化することによって、長いキーフレーズと関連づけられた低い確率を補償する調節を行う。キーフレーズは次式に基づいて正規化することができる。

【数１２】

$$P(k_i \mid r)^{1/m_i} \qquad (12)$$

上式でm_iはキーフレーズk_iの長さである。コントロールはステップＳ７３５０に続き、ここで、各センテンスを分析して、キーフレーズk_iが現れるかどうかを決定する。コントロールはステップＳ７３６０に続く。

【００５５】

ステップＳ７３６０では、キーフレーズk_iの存在を示す関数f_iと、セグメントまたはテキスト領域（ｒ）の間のポイントごとの相互情報量Ｉ（f_i、ｒ）を決定する。ポイントごとの相互情報量は、次式に基づいて決定する。

【数１３】

$$I(f_i, r) = \log\frac{P(f_i \mid r)}{P(f_i)} = \log\frac{P(f_i, r)}{P(f_i)P(r)} \qquad (13)$$

上式で、Ｐ（f_i）Ｐ（ｒ）とＰ（f_i、ｒ）は、センテンス上を合計することによって計算し、Ｐ（f_i、ｒ）は、所与のｒに関してキーフレーズk_iを含むトピックセグメントまたはテキスト領域ｒ内のセンテンスの数を文書内のセンテンスの数によって除算した、キーフレーズの複合確率f_iである。しかし、テキスト特徴および／または特異性を決定する際に有用な任意の統計測定もまた使用できる。コントロールはステップＳ７３７０に続く。

【００５６】

ステップＳ７３７０では、セグメントまたはテキスト領域（ｒ）内のキーフレーズの存在（f_i＝１）と、他のセグメントまたはテキスト領域（－ｒ）からのキーフレーズの不在（f_i＝０）に基づいて、キーフレーズ領域スコアを決定する。この決定は次式に基づいて行う。

【数１４】

$$Score_{k_i} = \lambda P(f_i = 1 \mid r)I(f_i = 1, r) + (1-\lambda)P(f_i = 0 \mid \neg r)I(f_i = 0, \neg r) \qquad (14)$$

パラメータλのデフォルトの値は０．５であるが、セグメントまたはテキスト領域内のキーフレーズの存在または不在の相対的な重要性を加重するために調節してもよい。

【００５７】

対数オッズ比は次式に基づく。

【数１５】

$$Score(k_i) = \log \frac{P(f_i = 1 \mid r) / P(f_i = 0 \mid r)}{P(f_i = 1 \mid \neg r) / P(f_i = 0 \mid \neg r)} \qquad (15)$$

対数オッズ比は、セグメントまたはテキスト領域内のテキストに対するキーフレーズの類似性または他のセグメントまたはテキスト領域からの非類似性を示す代替測定として使用する。しかし、キーフレーズのセグメントまたはテキスト領域に対する類似性または他のセグメントまたはテキスト領域に対する非類似性を決定する任意の方法もまた使用できる。コントロールはステップＳ７３８０に続く。

【００５８】

ステップＳ７３８０では、次の公式に基づいてセンテンス内のキーフレーズ上の加重平均に基づいてセンテンススコアを決定する。

【数１６】

$$Score(S) = \sum_{k_i \in S} P(k_i) Score(k_i) \qquad (16)$$

上式で、k_iはセンテンスＳ内のキーフレーズに対応し、Ｐ（k_i）はキーフレーズk_iを見る確率である。たとえば、Ｐ（k_i）はセンテンス内のk_iの相対的な頻度である。各センテンスのスコアを比較し、センテンスベースの、トピックベースのインタラクティブな要約を選択するための基礎として高いスコアを使用する。

【００５９】

領域内で候補となるセンテンスの場所など追加の特徴を使用して、選択するセンテンスのランクを決定するために使用するスコアをさらに調節する。所定のテキストでは、重要な情報はテキストの開始部または終了部に置かれる傾向がある。従って、センテンスの場所、センテンスの配置、インジケータフレーズ、キューフレーズなど、本質的な情報を識別する際に有用な他のテキスト上の特徴に基づいて、センテンススコアをさらに調節する。コントロールはステップＳ７３９０に続く。

【００６０】

ステップＳ７３９０では、センテンスランクまたはスコアが最も高いセンテンスを選択し、トピックベースのインタラクティブな要約に追加する。コントロールはステップＳ７４００に続き、ここで、離散センテンスベースの、トピックベースのインタラクティブなテキストの要約のために追加のセンテンスが必要かどうかを決定する。追加のセンテンスが必要な場合、コントロールはステップＳ７４１０に続く。

【００６１】

ステップＳ７４１０では、選択されたセンテンスに関連づけられたキーフレーズのカウントを差し引く。これによって、マルチセンテンスのトピックベースのインタラクティブなテキストの要約内で同じキーフレーズが繰り返される可能性を防ぐか減らす。コントロールはステップＳ７４２０に続き、差し引かれたキーフレーズカウントを補償するために確率を調節する。コントロールはステップＳ７３６０にジャンプする。ステップＳ７４００で、離散マルチセンテンスベースの、トピックベースのインタラクティブなテキストの要約のために追加のセンテンスがこれ以上必要でないことが決定されるまで、ステップＳ７３６０からＳ７４００を繰り返す。コントロールはステップＳ７４３０に続く。

【００６２】

ステップＳ７４３０では、離散センテンスベースの、トピックベースのインタラクティブなテキストの要約を表示する。上述のように、センテンスベースの、トピックベースの

インタラクティブなテキストの要約は、人間が感知可能な表示特徴を使用して要約内で省略されたテキストを示してもよい。コントロールはオプションのステップＳ７４３５に続く。オプションのステップＳ７４３５では、選択されたセンテンスから非本質的な要素を除去する。選択されたセンテンスの非本質的な要素は、統語解析、または、意味解析、統計フィルタリングなど、非本質的なセンテンス要素を除去する際に有用な任意の方法に基づいて除去できる。さらにオプションの動作では、読みやすさを改善するためにテキストを調節する。充填語を追加する、および／または、前方照応的指示を解決するなど、任意の方法を使用して読みやすさを改善することができる。コントロールはステップＳ７４４０に続く。コントロールの流れは、図６のステップＳ７４４０から図２のステップＳ８００に戻り、図２に関して上述されたように処理が続く。

【００６３】

種々の例示的なフローチャートは、他のすべてのトピックセグメントまたはテキスト領域、もしくは、隣接するトピックセグメントまたはテキスト領域のいずれかに関して、単一のトピックセグメントまたはテキスト領域の要約を選択するために必要な決定を示す。トピックセグメントまたはテキスト領域は各々別に考慮する。他のトピック領域（－Ｒ内のr_j）の各々における項の発生の合成スコアを平均して、合成を決定する。これは次式に基づく。

【数１７】

$$Score(f_i, \neg R) = \sum_{r_j \in \neg R} P(f_i, r_j) \log \frac{P(f_i \mid r_j)}{P(f_i)} \qquad (17)$$

項および／またはセンテンススコアは上述のように決定する。

【００６４】

多数のセンテンス構成要素ベースの、トピックベースのインタラクティブなテキストの要約を、単独または組み合わせて表示し、多数の異なるタイプのセンテンス構成要素に基づいた、トピックベースのハイアラーキ状のインタラクティブなテキストの要約を形成することができる。

【００６５】

図７は、本発明による、トピックベースのインタラクティブなテキスト要約システムの例を示す。トピックベースのインタラクティブな要約システム１００は、プロセッサ２０、メモリ３０、確率と潜在クラス決定回路４０、統計測定決定回路５０、確率正規化回路６０、センテンススコア決定回路７０、表示決定回路８０を備え、これらの各々は、入力／出力回路１０を介して通信リンク９９に接続する。トピックベースのインタラクティブな要約システム１００は、通信リンク９９を介して、情報リポジトリ２００、および、Ｗｅｂ使用可能パーソナルコンピュータ３００、Ｗｅｂ使用可能携帯機器４００、電話５００に接続する。

【００６６】

トピックベースのインタラクティブな要約システム１００は、通信リンク９９上で、Ｗｅｂ使用可能携帯機器４００および／またはＷｅｂ使用可能パーソナルコンピュータ３００のユーザから、トピックベースのインタラクティブなテキストの要約１０００に対する要求を受信する。トピックベースのインタラクティブなテキストの要約システム１００はテキスト１０００を情報リポジトリ２００から要求する。情報リポジトリ２００は通信リンク９９を介してテキスト１０００を、トピックベースのインタラクティブな要約システム１００の入力／出力回路１０に転送する。

【００６７】

トピックベースのインタラクティブな要約システム１００のプロセッサ２０は、センテンスの構成要素要約タイプ、候補となるキーフレーズの最大長、人間が感知可能なデフォルトの表示特徴などを指定する、以前に記憶および／または選択されたユーザの環境設定またはデフォルトのパラメータを検索する。記憶された環境設定および／または選択され

た環境設定に基づいて、センテンス構成要素要約タイプを決定する。センテンス構成要素要約タイプは、キーワード、および、キーフレーズ、文字列のＮ－ｇｒａｍ、センテンスベースであってよい。離散トピックを決定する際に有用な任意のタイプのセンテンス構成要素を使用できる。

【００６８】

プロセッサ２０は、確率と潜在クラス決定回路４０を作動させ、セグメントまたはテキスト領域ｒ内のキーフレーズＰ（k_i｜ｒ）の確率を決定する。たとえば、Ｐ（ｋｉ｜ｒ）は、確率潜在意味解析または同様な技術を使って、所与の潜在クラス内のキーフレーズの条件付確率Ｐ（k_i｜ｚ）とセグメントまたはテキスト領域ｒ上の潜在クラスの確率に基づいて決定することができる。前のステップで確率潜在意味解析モデルが生成されると、キーフレーズＰ（k_i｜ｚ）の条件付確率を決定および／またはメモリ内に記憶することができる。これによって必要な計算が減る。Ｐ（k_i｜ｚ）とＰ（ｚ｜ｒ）に基づいたＰ（k_i｜ｒ）の決定は例示的なものに過ぎず、Ｐ（k_i｜ｒ）を決定する任意の方法を使用できる。

【００６９】

プロセッサ２０は確率正規化回路６０を作動させ、キーフレーズ長に基づいてキーフレーズの確率を正規化することによって、短いキーフレーズの高い確率と、長いキーフレーズの低い確率を補償する。

【００７０】

プロセッサ２０は統計測定決定回路５０を作動させ、キーフレーズの分布を決定する。相互情報量、χ^2、G^2分布を使用して、キーフレーズ分布を決定することができる。しかし、キーフレーズがセグメントまたはテキスト領域を特徴づけるかおよび／または他の領域からセグメントまたはテキスト領域を差別化する測定を決定する際に有用な、任意の統計測定をまた使用することができる。

【００７１】

センテンス要約が選択されたと決定された場合、プロセッサ２０はセンテンススコア決定回路７０を作動させる。センテンススコア決定回路７０は、センテンス上でキーフレーズの加重平均を決定する。プロセッサ２０はテキストのセグメントまたはテキスト領域に関して、ランク／スコアが最も高いセンテンスを選択し、このセンテンスを、センテンスベースの、トピックベースのインタラクティブなテキスト要約に追加する。マルチセンテンスの、トピックベースのインタラクティブなテキスト要約を選択した場合、追加したセンテンスと関連づけられたキーフレーズカウントを差し引いて、繰り返すセンテンスを除去する。必要な数のセンテンスを選択するまで、統計測定とセンテンススコアを再決定し、ランク／スコアが最も高い新しいセンテンスを決定する。

【００７２】

離散インタラクティブトピックベースのテキストの要約を決定した後、表示回路８０を作動させてテキストを表示する。離散キーワードと離散センテンスベースの、トピックベースのインタラクティブな要約は、トピック生成の指示を提供するテキストに隣接して同時に表示してもよい。トピックベースのインタラクティブなテキストの要約の中のテキストインデント、または、スペース、配置、テキストの色、中黒、サムネールテキスト画像など、人間が感知可能な任意の表示特徴を使用して、テキスト内の要約されたトピックの配置および／または生成を同時に示すことができる。

【００７３】

図８は、本発明によって要約されたテキストの一部を示す。テキスト１０００は、ディスプレイの１６行に渡って４つのセンテンスを含む。テキストは稠密に描かれ、トピックとサブトピックの生成を決定することは難しい。

【００７４】

図９は、本発明によるトピックベースのインタラクティブなテキストの要約の例を示す。テキスト１０００の、トピックベースのインタラクティブなテキストの要約７００は、センテンスベースの要約７２１から７２３と、キーフレーズベースの要約７１１から７１

３の両方に基づく。テキストの要約された行の数は、行番号７３１から７３３により示される。省略されたテキストは、省略されたテキストインジケータとして、キーフレーズベースの要約７１１の下線によって示す。しかし、人間が感知可能な任意の知られた表示特徴かまたは後から開発された表示特徴を、本発明によって省略されたテキストのインジケータとして使用することができる。下線の付いた省略されたテキストのインジケータは、参考のための追加の関連テキストを表示する。省略されたテキストの人間が感知可能なインジケータを１回クリックすると、インタラクティブに拡大し、キーワードベースの、トピックベースのインタラクティブなテキストの要約と関連づけられたテキストセグメントおよび／または全テキストを表示する。

【００７５】

同様に、人間が感知可能なイタリックと太字の表示特徴は、キーフレーズベースの要約７１２から７１３の省略されたテキストインジケータに関連づけられている。キーワードとセンテンスベースの要約は共の位置にあり配置されて、テキスト１０００内のトピックの場所を示す文脈内の構成と位置情報を提供する。

【００７６】

図１０は、本発明による、トピックベースのインタラクティブなテキスト要約表示の例としてのフローチャートである。このプロセスはステップＳ１００００で開始し、すぐにステップＳ１００１０に続く。

【００７７】

ステップＳ１００１０では、要約するテキストを選択する。テキストは、ダイアログボックスを使用してファイル名を入力するか、マウスまたはコンピュータが認識可能な意思表示を使用してファイル名を選択するか、音声コマンドか、またはテキストを選択する任意の知られた方法または最近開発された方法を使用して選択する。テキストを選択した後、コントロールはステップＳ１００１５に続き、要約のためにテキストの一部をダイナミックに選択する。テキストは、反転表示、境界ボックスなどの知られた選択方法または最近開発された選択方法を使用して選択する。コントロールはステップＳ１００２０に続く。

【００７８】

ステップＳ１００２０では、選択したテキストをセグメント化する。上記のように、テキストは既存のセグメントインジケータを使用してセグメント化するか、自動的にセグメント化するか、または任意の知られたセグメント化技術または最近開発されたセグメント化技術を使用してセグメント化する。テキストを反転表示させるか、セグメント境界ボックスでテキストの境界を定めるなどによって、セグメントを手で作成することができる。選択したテキストをセグメント化した後、コントロールはステップＳ１００３０に続く。

【００７９】

ステップＳ１００３０では、要約するセグメントを決定する。セグメントは自動的に決定してもよい。しかし、ユーザがセグメントを決定することもできる。要約するテキストの選択された部分内のセグメントは、境界ボックス、マウス、または、ポインティングデバイス使用したテキストの反転表示、または任意の他の技術などを使用することによって示すことができる。セグメントは、テキストを反転表示する、セグメント境界ボックスでテキストの境界を定めるなどで、手動で作成することもできる。境界ボックスを使って手で作成したセグメントは、要約するセグメントとして暗に選択される。しかし、本発明の範囲から離れることなく、これらのデフォルト条件のうち任意の条件を変えることができる。テキストのセグメントを再帰的に選択して要約し、さらに、決定したセグメントをより小さな単位にセグメント化することもできる。セグメントのサブセグメントまたは部分は、インデントなどを使用して文脈内で表示し、関連づけられたハイアラーキレベルを示すことができる。要約するセグメントを選択した後、コントロールはステップＳ１００４０に続く。

【００８０】

ステップＳ１００４０では、センテンス構成要素ベースの要約タイプを決定する。全セ

ンテンス、および／または、キーフレーズ、キーワードなど、任意のセンテンス構成要素を、トピックベースのインタラクティブなテキストの要約を決定する基礎として選択できる。異なるタイプのセンテンス構成要素ベースの、トピックベースのインタラクティブなテキスト要約も、単一または組み合わせて、ハイアラーキ状の、センテンス構成要素ベースの、トピックベースのインタラクティブなテキストの要約を形成することができる。コントロールはステップＳ１００５０に続く。

【００８１】

ステップＳ１００５０では、センテンス構成要素に基づいた、トピックベースのインタラクティブなテキスト要約を、選択されたテキスト、および、決定されたセグメント、テキストの決定された部分、決定されたセンテンス構成要素ベースの要約タイプに基づいて決定する。キーワード、および、キーフレーズ、センテンスベースの、トピックベースのインタラクティブなテキストの要約は、テキストのダイナミックに選択された部分について決定することができる。コントロールはステップＳ１００６０に続き、トピックベースのインタラクティブなテキストの要約を表示する。

【００８２】

「＋」文字など、人間が感知可能な表示特徴を使用して、省略されたテキストを示すことができる。省略されたテキストインジケータは、省略されたテキストの前または後に挿入することができる。省略したテキストインジケータは、省略したテキストをダイナミックに表示および／または拡大するために選択することができる。マウスのロールオーバ、マウスを１回右クリックすること、または、任意の知られた選択技術または最近の選択技術などを使用して、別のポップアップウィンドウに拡大して表示する、省略されたテキストを選択し、現在のウィンドウ内で、関連する、センテンス構成要素ベースの、トピックベースのインタラクティブなテキスト要約を、アコーディオン状に全テキスト拡大として表示するなどができる。同様に、マウスのロールオフ、または、続くマウスのクリック、他の選択方法を使用して、センテンス構成要素ベースの、トピックベースのインタラクティブなテキスト要約に関連づけられた、省略されたテキストを畳み込むことができる。音声合成装置、または、ダイナミック点字装置、その他の出力表示技術を介して出力する要約されたテキストは、フォントの変化、または、音の変化、ダイナミック点字出力要素、人間が感知可能な任意のその他の表示特徴が間歇的または可変的に存在することによって、省略されたテキストを示すことができる。コントロールはステップＳ１００７０に続く。

【００８３】

ステップＳ１００７０では、ユーザがセッションを終了したかどうかを決定する。セッションが終了していない場合、コントロールはすぐにステップＳ１００１５にジャンプし、ステップＳ１００７０でセッションが終了するまで、ステップＳ１００１５からＳ１００７０を繰り返す。セッションがステップＳ１００７０で終了すると、コントロールはステップＳ１００８０に続き、プロセスは終了する。

【００８４】

図１１は、本発明による、トピックベースのインタラクティブなテキストの要約内で省略されたテキストの拡大と畳み込みを拡大して表示したフローチャートを示す。このプロセスはステップＳ１０１００で始まり、すぐにステップＳ１０１１０に続く。ステップＳ１０１１０では、例としての、トピックベースのインタラクティブなテキストの要約を表示する。マルチセンテンスベースの、トピックベースのインタラクティブなテキストの要約は、ユーザが選択して表示させることができる。

【００８５】

ステップＳ１０１２０では、表示された、マルチセンテンスの、トピックベースのインタラクティブなテキストの要約の中で、省略されたテキストのインジケータが選択されたかどうかについて決定する。省略されたテキストのインジケータは、文字、下線など、人間が感知可能な任意の表示特徴を含んでいてもよい。省略されたテキストのインジケータは、省略されたテキストのインジケータの上でマウスカーソルを転がすか、マウスを右ク

リックするか、任意の他の選択方法を使用して選択する。省略されたテキストのインジケータが選択されたと決定された場合、コントロールはステップＳ１０１３０に続く。その他の場合はコントロールはステップＳ１０１４０に続く。

【００８６】

ステップＳ１０１３０では、省略されたテキストインジケータと関連づけられたテキストを拡大し表示する。省略されたテキストインジケータと関連づけられたテキストは、ポップアップウィンドウ内に表示してもよいし、スタティックなウィンドウ内に表示してもよいし、トピックベースのインタラクティブなテキストの要約の中に省略されたテキストとして拡大して表示してもよい。小さな画面の装置でトピックベースのインタラクティブなテキストの要約を見るとき、トピックに基づいたインタラクティブなテキスト要約内の省略されたテキストの拡大表示は、テキスト内の文脈を維持しながら、テキスト内のトピックを効率的にナビゲーションし選択的に見ることを可能にする。対照的に、省略されたテキストのインジケータ上でカーソルをロールオーバさせたことに応答するポップアップウィンドウは、センテンス構成要素ベースのトピックベースのインタラクティブなテキストの要約を変更せずに、センテンス構成要素ベースの、トピックベースのテキストの要約を表示する。コントロールはステップＳ１０１４０に続く。

【００８７】

ステップＳ１０１４０では、省略されたテキストインジケータに関連づけられた省略されたテキストが畳み込むように選択されたかどうかを決定する。省略されたテキストインジケータに関連づけられたテキストは、省略されたテキストのインジケータの上からマウスまたはカーソルを動かす、または、インジケータ上でマウスを右クリックするなどの任意の選択方法を使用して畳み込む。ステップＳ１０１４０で、省略されたテキストを畳み込むように選択したことが決定されると、コントロールはステップＳ１０１５０に続く。その他の場合はコントロールはＳ１０１６０に続く。

【００８８】

ステップＳ１０１５０では、選択された、省略されたテキストインジケータに関連づけられた選択されたテキストを畳み込むかおよび／または表示から除去する。選択された、省略されたテキストインジケータは、カーソルを省略されたテキストインジケータから動かしたときに省略されたテキストのポップアップウィンドウを除去するか、トピックベースのインタラクティブなテキストの要約内の拡大された省略されたテキストを畳み込むことによって、表示から畳み込むおよび／または除去することができる。コントロールはステップＳ１０１６０に続く。

【００８９】

ステップＳ１０１６０では、ユーザがセッションの終了を要求したかどうかを決定する。ユーザがセッションの終了を要求したと決定されると、コントロールはステップＳ１０１７０に続く。その他の場合、コントロールはステップＳ１０１２０に続き、セッションの終わりが要求されたと決定されるまで、ステップＳ１０１２０からＳ１０１６０を繰り返す。コントロールはステップＳ１０１７０に続き、プロセスは終了する。

【００９０】

図１２は、本発明による、ハイアラーキ状の、トピックベースのインタラクティブなテキストの要約の例を示す。ハイアラーキ状の、トピックベースのテキストの要約は、テキストに関する、トピックベースの要約のハイアラーキ状の表示を示す。ハイアラーキ状の、トピックベースのインタラクティブなテキストの要約の要素をインデントして、ハイアラーキを文脈内で表示することができる。サブトピックは、ハイアラーキ内のレベルに基づいてインデントして示す。要約または要約されたサブセグメントに基づいた要約は、（０）という行番号インジケータまたは任意の他のインジケータで示すことができる。

【００９１】

テキスト１０００の、ハイアラーキ状のトピックベースのインタラクティブなテキストの要約７００は、センテンスベースの要約７２１から７２５と、キーフレーズベースの要約７１１から７１５の両方に基づく。入手可能な省略されたテキスト７６０は、関連付け

られたキーフレーズベースの要約７１３に省略されたテキストインジケータとして下線を引くことによって示す。しかし、人間が感知可能な任意の表示特徴を、省略されたテキストのインジケータとして使用できる。省略されたテキストインジケータは、追加の関連するテキストを見ることができることを示す。キーフレーズ要約７１３上でカーソルまたはマウスポインタ７４０を動かすことに応答して、ポップアップウィンドウ７５０はキーフレーズ７１３と関連づけられた省略されたテキスト７６０を表示する。表示された要約を含む任意の要素または要素の組み合わせが、省略されたテキストインジケータとしての役割を果たすことができる。省略されたテキストの、人間が感知可能なインジケータの上を一回クリックすると、インタラクティブに拡大し、キーワードベースのトピックに基づいたインタラクティブなテキストの要約に関連づけられたテキストのセグメントおよび／または全テキストを表示する。

【００９２】

イタリックと太字も同様に、キーフレーズベースの要約７１２から７１３に関する省略されたテキストインジケータと関連づけられた、人間が感知可能な表示特徴である。キーワードとセンテンスベースの要約は、同じ場所に配置して、テキスト１０００内のトピックの近似の位置を示す、文脈上の構成と位置情報を提供する。

【００９３】

本発明の範囲から離れることなく、多数の異なるセンテンス構成要素ベースの、トピックベースのインタラクティブなテキストの要約も、表示し、ダイナミックに拡大および／または畳み込むことができる。第３と第４の要約の表示に関連づけられたインデントは、ハイアラーキ状の関連を文脈内で示す。

【００９４】

トピックベースのインタラクティブなテキストの要約システム１００の回路１０から２０と４０から８０の各々は、適切にプログラミングされた汎用コンピュータの一部として実装できる。別法としては、トピックベースのインタラクティブなテキストの要約システム１００の１０から２０と４０から８０は、ＡＳＩＣ内の物理的に別のハードウェア回路として実装してもよいし、または、ＦＰＧＡ、ＰＤＬ、ＰＬＡ、ＰＡＬを使用してもよいし、離散論理要素または離散回路要素を使用してもよい。上述の、トピックベースのインタラクティブなテキストの要約システム１００の回路１０から２０と４０から８０の各々がとる特定の形態は、設計上の選択であり、当業者であれば明らかで断定可能であろう。

【００９５】

さらに、上述の、トピックベースのインタラクティブなテキストの要約システム１００および／または種々の回路は各々、プログラミングされた汎用コンピュータ、特殊目的コンピュータ、マイクロプロセッサなどの上で実行するソフトウェアルーチン、または、マネジャー、オブジェクトとして実装できる。この場合、上述の、トピックベースのインタラクティブなテキストの要約システム１００および／または種々の回路は各々、通信ネットワーク内に組み込まれた１つまたは複数のルーチンとして実装してもよいし、または、サーバ上に常駐するリソースとしても実装できる。上述の、トピックベースのインタラクティブなテキストの要約システム１００および／または種々の回路は各々、通信ネットワークに埋め込まれた１つまたは複数のルーチン、サーバ上に常駐するリソースなどとしても実装できる。上述の、トピックベースのインタラクティブなテキストの要約システム１００と種々の回路は、トピックに基づいたインタラクティブなテキストの要約システム１００を、Ｗｅｂサーバまたはクライアント装置のハードウェアシステムとソフトウェアシステムなど、ソフトウェアシステムおよび／またはハードウェアシステム内に物理的に組み込んで実装することもできる。

【００９６】

図７に示すように、メモリ３０は可変性の揮発性メモリまたは不揮発性メモリ、または、非可変性メモリまたは固定メモリの任意の適切な組み合わせを使用して実装できる。

【００９７】

図１と図７に示された通信リンク９９は、装置を接続し通信を促進するために使用可能

な、任意の知られた接続システムまたは構造、または、最近開発された接続システムまたは構造であってよい。

【図面の簡単な説明】

【００９８】

【図１】本発明による、例としての、トピックベースのインタラクティブな要約システムの概略図である。

【図２】本発明による、トピックベースのインタラクティブなテキスト要約に関する例としての方法のフローチャートである。

【図３】本発明による、離散キーワードベースの、トピックベースのインタラクティブな要約を決定する、例としての方法の拡大フローチャートである。

【図４】本発明による、離散キーフレーズベースの、トピックベースのインタラクティブな要約を決定する、例としての方法の拡大フローチャートである。

【図５】本発明による、離散センテンスベースの、トピックベースのインタラクティブな要約を決定する、第１の例としての方法の拡大フローチャートである。

【図６】本発明による、離散センテンスベースの、トピックベースのインタラクティブな要約を決定する、第２の例としての方法の拡大フローチャートである。

【図７】本発明による、トピックベースのインタラクティブなテキストの要約の例を示す図である。

【図８】本発明によって要約されるテキストの一部を示す図である。

【図９】本発明による、トピックベースのインタラクティブなテキストの要約の例を示す図である。

【図１０】本発明による、トピックベースのインタラクティブなテキスト要約を表示する例としての方法のフローチャートである。

【図１１】本発明による、トピックベースのインタラクティブなテキストの要約内で省略されたテキストの拡大と畳み込みの拡大表示フローチャートである。

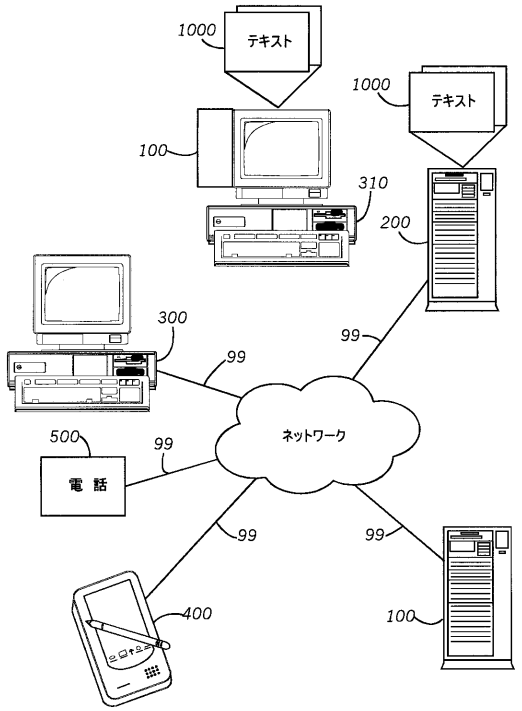
【図１２】本発明による、ハイアラーキ状のトピックベースのインタラクティブなテキストの要約の例を示す図である。

【符号の説明】

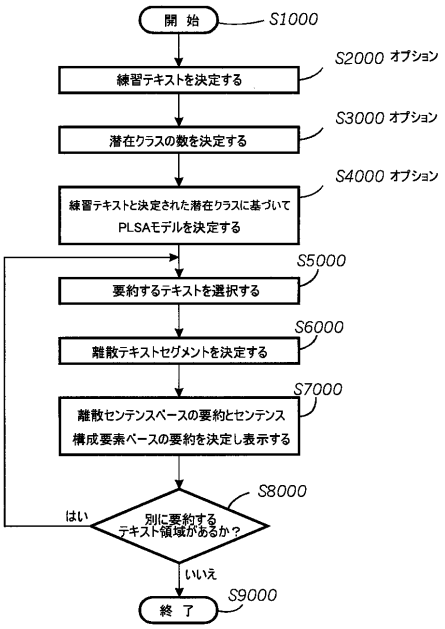
【００９９】

９９　通信リンク、１００　テキスト要約システム、２００　リポジトリ、３００　Ｗｅｂ使用可能パーソナルコンピュータ、３１０　パーソナルコンピュータ、４００　Ｗｅｂ使用可能装置、５００　電話、１０００　テキスト情報。

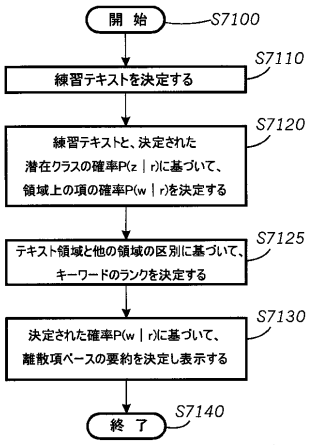
【図１】


1000
テキスト
1000
テキスト
100
310
200
300
99
99
99
ネットワーク
500
電 話
99
99
400
100

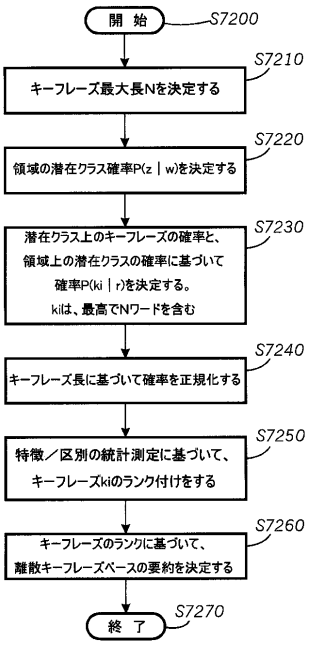
【図２】


開 始
S1000
練習テキストを決定する
S2000 オプション
潜在クラスの数を決定する
S3000 オプション
練習テキストと決定された潜在クラスに基づいて
PLSAモデルを決定する
S4000 オプション
要約するテキストを選択する
S5000
離散テキストセグメントを決定する
S6000
離散センテンスベースの要約とセンテンス
構成要素ベースの要約を決定し表示する
S7000
別に要約する
テキスト領域があるか？
S8000
はい
いいえ
終 了
S9000

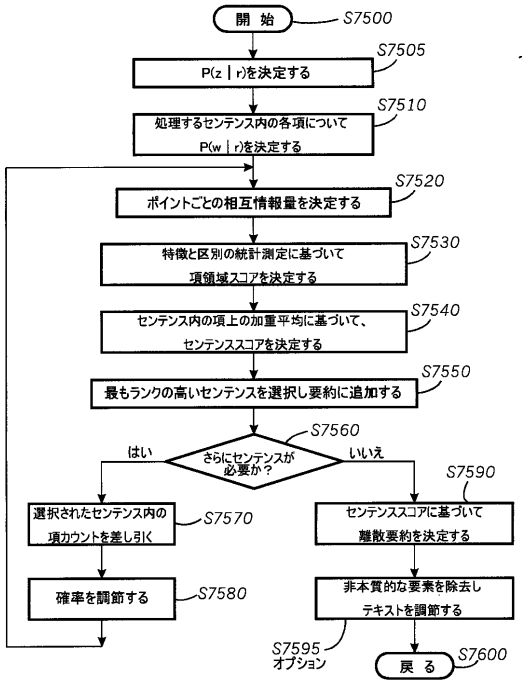
【図３】


開 始
S7100
練習テキストを決定する
S7110
練習テキストと、決定された
潜在クラスの確率P(z｜r)に基づいて、
領域上の項の確率P(w｜r)を決定する
S7120
テキスト領域と他の領域の区別に基づいて、
キーワードのランクを決定する
S7125
決定された確率P(w｜r)に基づいて、
離散項ベースの要約を決定し表示する
S7130
終 了
S7140

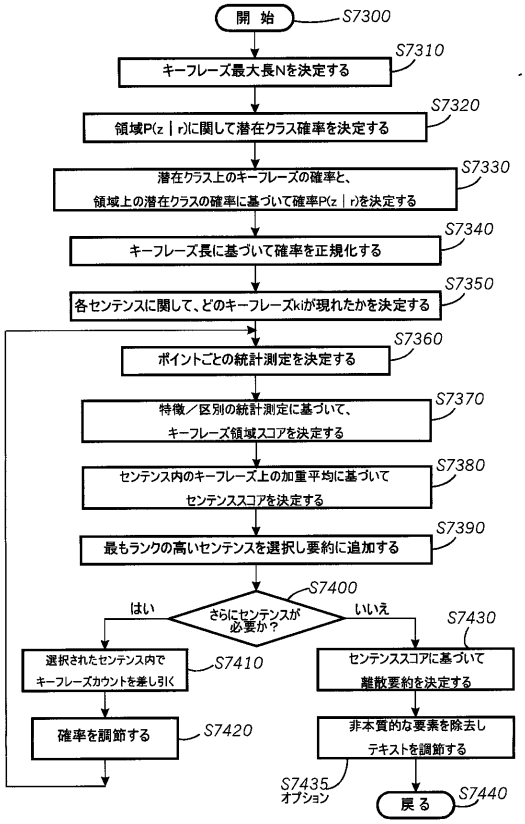
【図４】


開 始
S7200
キーフレーズ最大長Nを決定する
S7210
領域の潜在クラス確率P(z｜w)を決定する
S7220
潜在クラス上のキーフレーズの確率と、
領域上の潜在クラスの確率に基づいて
確率P(ki｜r)を決定する。
kiは、最高でNワードを含む
S7230
キーフレーズ長に基づいて確率を正規化する
S7240
特徴／区別の統計測定に基づいて、
キーフレーズkiのランク付けをする
S7250
キーフレーズのランクに基づいて、
離散キーフレーズベースの要約を決定する
S7260
終 了
S7270

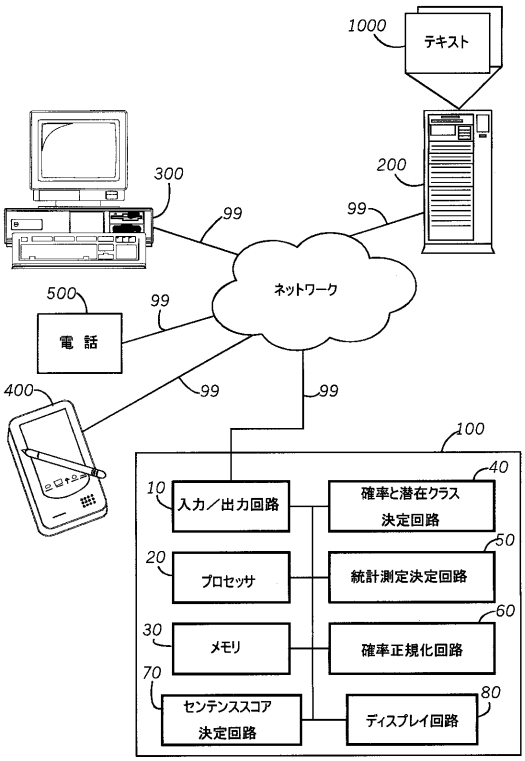
【図５】



【図６】



【図７】

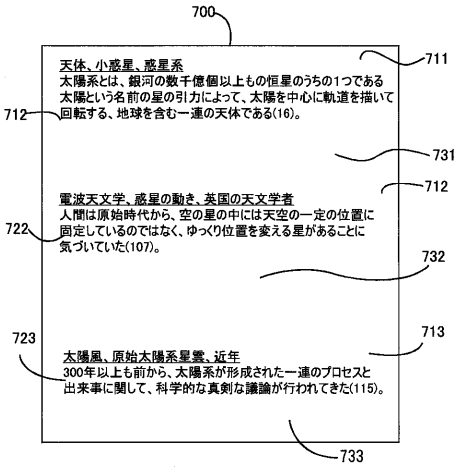


【図８】

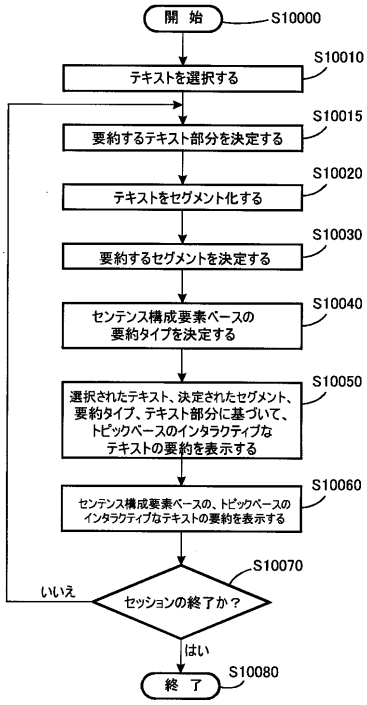
太陽系とは、銀河の数千億個以上もの恒星のうちの1つである太陽という名前の星の引力によって、太陽を中心に軌道を描いて回転する、地球を含む一連の天体である。太陽系内には、9つの惑星、54以上の衛星、観察された限りで1000以上の彗星、そして小惑星として知られている何千もの小さな天体、流星（流星と隕石の項を参照）などの天体が含まれる。これらの天体はすべて、壊れやすい岩石質の惑星間塵粒子でできたやっかいな海を漂っている。
惑星間塵の起源は明らかではないが、彗星が内惑星付近を通過したときに彗星から噴出された塵か、または、小惑星同士が衝突した結果できた塵と考えられている。太陽は、直接観察によって惑星群を伴うことが分かった唯一の恒星であるが、銀河系には他にも惑星系を持つ恒星がたくさんあるのではないか、さらにそういった惑星の中には技術文明を発達させている惑星があるのではないか、と推測されている。

1000

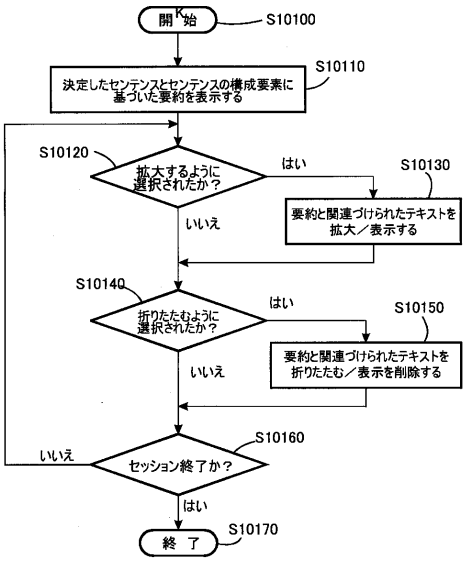
【図９】



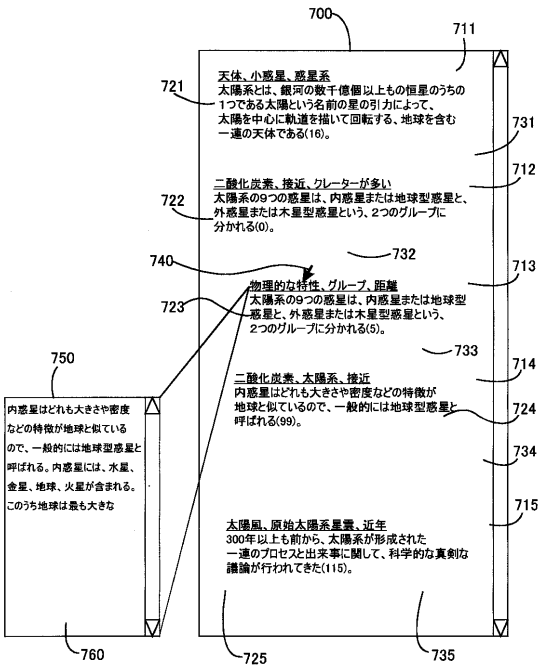
【図１０】



【図１１】



【図１２】



フロントページの続き

(72)発明者　トーステン　エイチ　ブランツ
アメリカ合衆国　カリフォルニア　パロ　アルト　モンロー　ドライブ　４３２
(72)発明者　アニー　イー　ザエネン
アメリカ合衆国　カリフォルニア　レッドウッド　シティ　ジェファーソン　アベニュー　３９５０

Ｆターム(参考) 5B009 QA05　RB23　RB24
5B075 ND03　NS01　PQ02

【要約の続き】