(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2015-518585
(P2015-518585A)
(43)公表日　平成27年7月2日(2015.7.2)

(51)Int.Cl.	F I	テーマコード（参考）
G06F 17/30 (2006.01)	G06F 17/30	210D

審査請求 有　予備審査請求 未請求　（全 19 頁）

(21)出願番号	特願2014-530448 (P2014-530448)
(86)(22)出願日	平成25年12月13日 (2013.12.13)
(85)翻訳文提出日	平成26年6月25日 (2014.6.25)
(86)国際出願番号	PCT/JP2013/084169
(87)国際公開番号	WO2014/141560
(87)国際公開日	平成26年9月18日 (2014.9.18)
(31)優先権主張番号	13/837,764
(32)優先日	平成25年3月15日 (2013.3.15)
(33)優先権主張国	米国 (US)

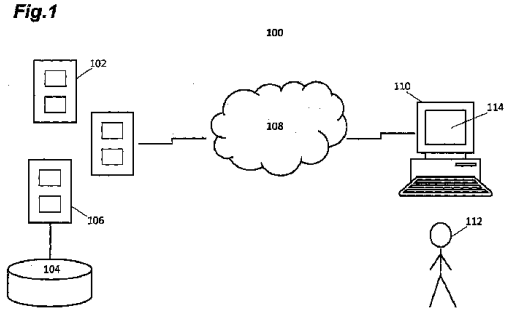
(71)出願人　399037405
楽天株式会社
東京都品川区東品川四丁目１２番３号
(74)代理人　100088155
弁理士　長谷川　芳樹
(74)代理人　100113435
弁理士　黒木　義樹
(74)代理人　100144440
弁理士　保坂　一之
(72)発明者　スタンキエヴィッチ，　ゾフィア
アメリカ合衆国，　１００１３，　ニューヨーク，　ニューヨーク，　２１５　パーク　アヴェニュー　サウス　９ティーエイチ　フロア

最終頁に続く

(54)【発明の名称】半構造化されたデータを解析しカテゴリ分けするための方法

(57)【要約】
ユーザのコミュニティに相互接続されたコンピュータシステムは、料理レシピに関する一又は複数の入力を含む前記ユーザからの通信を受信し、アクセス可能なメモリに前記入力を記憶するようにプログラムされたデータプロセッサ入力モジュールを有する。データプロセッサ決定モジュールは、記憶したデータにアクセスし、レシピに関する統合的なデータベースへと異種のデータ入力を統合し体系化するために前記データにデータ解釈用アルゴリズムを適用するようにプログラムされる。また、検索エントリモジュールは、データベースに適用された検索アルゴリズムをサポートするためにデータベースへのアクセスを許可するためにレシピデータベースに接続される。
【選択図】図１

Fig.1



【特許請求の範囲】
【請求項１】
ユーザのコミュニティに相互接続されたコンピュータシステムであって、
料理レシピに関する一又は複数の入力を含む前記ユーザからの通信を受信し、アクセス可能なメモリに前記入力を記憶するようにプログラムされたデータプロセッサ入力モジュールと、
記憶したデータにアクセスし、レシピに関する統合的なデータベースへと異種のデータ入力を統合し体系化するために前記データにデータ解釈用アルゴリズムを適用するようにプログラムされたデータプロセッサ決定モジュールと、
前記データベースに適用された検索アルゴリズムをサポートするために前記データベースへのアクセスを許可するために前記レシピデータベースに接続された検索エントリモジュールと
を備えるコンピュータシステム。
【請求項２】
複数のユーザにネットワークで接続されたコンピュータシステムであって、
複数の半構造化されたユーザ入力データを記憶するメモリと、
前記半構造化されたユーザ入力データのサブセットにデータ解釈用アルゴリズムを適用するプロセッサと、
前記半構造化されたユーザ入力データの前記サブセットを使用して前記複数の半構造化されたユーザ入力データの残りをカテゴリ分けする第２のプロセッサと、
ユーザが前記カテゴリ分けされた複数の半構造化されたユーザ入力データを検索することを可能にするインターフェースと
を備えるコンピュータシステム。
【請求項３】
前記複数の半構造化されたユーザ入力データが複数のデータフィールドを備える、請求項２に記載のシステム。
【請求項４】
前記複数の半構造化されたユーザ入力データがレシピであり、
前記複数のデータフィールドが、レシピ名称、原材料、命令、タグ、及び画像のうちの少なくとも一つを備える、請求項３に記載のシステム。
【請求項５】
前記データ解釈用アルゴリズムが、混成最大エントロピー及びＬＤＡモデルと、語出現頻度－文献出現頻度の逆数と、コサイン類似度解析とのうちの少なくとも一つを備える、請求項２に記載のシステム。
【請求項６】
半構造化されたデータを解析するための方法であって、
複数の半構造化されたデータエントリをメモリに記憶するステップであり、それぞれの半構造化されたデータエントリが複数のデータフィールドを含む、記憶するステップと、
プロセッサを用いて、それぞれの半構造化されたデータエントリ内の前記半構造化されたデータフィールドをソートするステップと、
プロセッサを用いて、データ解釈用アルゴリズムを使用して半構造化されたデータエントリのサブセットを選択するステップと、
前記半構造化されたデータエントリの前記サブセットの前記半構造化されたデータフィールドからトピックのデータフィールドを選択するステップと、
プロセッサを用いて、前記トピックのデータフィールドを用いて残りの複数の半構造化されたデータエントリを解析するステップと、
プロセッサを用いて、前記解析した残りの複数の半構造化されたデータエントリにデータ解釈用アルゴリズムを使用して、半構造化されたデータエントリの新たなサブセットを選択するステップと、
前記半構造化されたデータエントリの新たなサブセットを前記半構造化されたデータエ

ントリの前記サブセットと統合するステップと
を含む方法。
【請求項７】
前記複数の半構造化されたデータエントリがレシピであり、
前記複数のデータフィールドが、レシピ名称、原材料、命令、タグ、及び画像のうちの少なくとも一つを備える、請求項６に記載の方法。
【請求項８】
前記データ解釈用アルゴリズムが、混成最大エントロピー及びＬＤＡモデルを備える、請求項６に記載の方法。
【請求項９】
前記データ解釈用アルゴリズムが、語出現頻度－文献出現頻度の逆数とコサイン類似度解析とを備える、請求項６に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、生の及び／又は構造化されていないデータに調和のとれたフィールド割り当てを用いて、十分に構造化された入力に及ぶ複数の異なるフォーマットに体系化されたデータのセットを解析することに関する。特に、本発明は、違ったふうに構造化されたデータの全スペクトルをカテゴリ分けするために、一又は複数のデータ解釈モデルに従った様々な入力のデータ処理に関する。
【背景技術】
【０００２】
今日の現代社会では、数え切れない様々な理由で複数のユーザによってデータが入力される。いったん、データが入力されると、他のユーザは、データを解釈し、関連する結果を迅速に見つけるために、データを検索しかつソートする能力を望む。しかしながら、データが、様々なフォーマット、場所及び言語で様々なユーザによって入力されるので、データの入力方法に一貫性がないことがある。したがって、例えば、一貫性のない登録のために、特定の単語又はトピックの検索において関連する情報を見落とすことがある。
【０００３】
この一例は、ウェブサイト上に掲示される、ユーザが入力したレシピである。レシピの構造化された部分は、レシピの名前、原材料、料理法、及びイベントに関するフィールドであり得る。しかしながら、いったん、ユーザがそのフィールドに情報を入力し始めると、情報がどのようにして実際に提示されるかにおいて、大きな差異があり得る。任意の構造化されたフィールドでは、書き間違いは普通であり得る。また、レシピ及び原材料に関する異名又は代替名もよくあることである。例えば、何人かのユーザは「ポテトスープ」又は「冷製ポテトスープ」とレシピを入力することがあるが、他のユーザ、「ビシソワーズ」という名前を入力することがある。しかしながら、「ビシソワーズ」を検索するユーザは、すべての入手可能な冷製ポテトスープのレシピを見ることをおそらく望むであろう。
【０００４】
これは、別のタイプのユーザ入力データについても当てはまる。このデータの多くは、半構造化されることがあるし、何らかの構造とか何かが欠如することがあり得る。本発明は、構造化されていない又は半構造化されたデータから有用な知識を抽出し得る。
【発明の概要】
【０００５】
本明細書における教示は、多種多様なフォーマットで受信したデータ、又は同じフォーマットであるが記録する際に異なるレベルの構造、精度及び忠実度で受信したデータを統合及び体系化することに伴う上記の問題の一又は複数を軽減することで、データが一つ若しくは複数の評価に適用され得るようにすることである。例示的な実施形態では、コンピュータシステムは、複数のユーザがデータ／通信ネットワークを介して中央サーバ／ネッ

トワークに接続することを可能にする。サーバは、システム稼働プログラミングによって管理されたプラットフォームへのゲートウェイを形成する。選択データは、プログラムされた入力構造に従って集められ、一又は複数のデータ構造化アルゴリズムによって解釈及び／又は体系化される。プロセスデータは、次に、例えば、検索照会等などの様々な機能をサポートするために使用される。

【０００６】

別の例では、コンピュータシステムは複数のユーザにネットワークで接続される。システムは少なくとも、複数の半構造化されたユーザ入力データを記憶するメモリと、半構造化されたユーザ入力データのサブセットにデータ解釈用アルゴリズムを適用するプロセッサとを有する。同じ又は第２のプロセッサのいずれかは、半構造化されたユーザ入力データのサブセットを用いて、複数の半構造化されたユーザ入力データの残りをカテゴリ分けし得る。システムは、カテゴリ分けされた複数の半構造化されたユーザ入力データをユーザが検索することを可能にするインターフェースも有する。このインターフェースは、検索及び結果の取得を可能にするウェブページ、アプリケーション、又は他のポータルであってもよい。

【０００７】

システムの別の一例では、複数の半構造化されたユーザ入力データはデータフィールドを有する。さらに、複数の半構造化されたユーザ入力データはレシピである。そして、複数のデータフィールドは、レシピ名称、原材料、命令、タグ、及び画像のうちの少なくとも一つであってもよい。

【０００８】

データ解釈用アルゴリズムの例として、混成（ｈｙｂｒｉｄ）最大エントロピー及びＬＤＡモデルと、語出現頻度－文献出現頻度の逆数と、コサイン類似度解析とがあり得る。

【０００９】

半構造化されたデータを解析するための方法の例は、複数の半構造化されたデータエントリをメモリに記憶するステップを含む。それぞれの半構造化されたデータエントリは、複数のデータフィールドを含む得る。プロセッサは、それぞれの半構造化されたデータエントリ内の半構造化されたデータフィールドをソートし、データ解釈用アルゴリズムを使用して、半構造化されたデータエントリのサブセットを選択し得る。続いて、トピックのデータフィールドが、半構造化されたデータエントリのサブセットの半構造化されたデータフィールドから選択され得る。続いて、残りの複数の半構造化されたデータエントリがトピックのデータフィールドを用いて解析され得る。そして、半構造化されたデータエントリの新たなサブセットが、解析された残りの複数の半構造化されたデータエントリについてデータ解釈用アルゴリズムを使用して選択され、半構造化されたデータエントリのサブセットと統合され得る。

【００１０】

別の一例では、複数の半構造化されたデータエントリはレシピであり、複数のデータフィールドは、レシピ名称、原材料、命令、タグ、及び画像のうちの少なくとも一つを備える。データ解釈用アルゴリズムは、混成最大エントロピー及びＬＤＡモデルと、語出現頻度－文献出現頻度の逆数と、コサイン類似度解析とを含んでもよい。

【００１１】

図面は、限定ではなく単に例として、本技術に従った一又は複数の実装形態を示す。図面では、類似の参照番号は同じ又は類似の要素を指す。

【図面の簡単な説明】

【００１２】

【図１】本発明を実装するためのネットワークの一例の図である。

【図２Ａ】ウェブサイト及び半構造化されたデータフィールドを示す図である。

【図２Ｂ】半構造化されたデータエントリの一例の図である。

【図３Ａ】システムによって解析されたままの半構造化されたデータの図である。

【図３Ｂ】システムによって解析されたままの半構造化されたデータのもう一つの例の図

である。
【図４】料理を決定するために訓練データを使用する方法の一例の流れ図である。
【図５】レシピ類似度を決定するための方法の一例の流れ図である。
【図６】レシピ類似度を決定するための方法のもう一つの例の流れ図である。
【実施形態の説明】
【００１３】
下記の詳細な説明では、数多くの具体的な詳細が、関連する教示の十分な理解を提供するために例として記述される。しかしながら、本教示がこのような詳細がなくとも実行し得ることは当業者には明らかであるはずである。別の事例では、良く知られた方法、手順、構成要素、及び／又は回路は、本教示の態様を不必要に不明瞭にすることを避けるために、詳細なしに比較的高次のレベルで記述されてきている。
【００１４】
本発明は、半構造化されたデータから、構造化された「知識」を抽出するためのシステム及び方法を提供する。具体的な例としては、ウェブサイトでユーザによって入力されたレシピをソートしカテゴリ分けすることである。
【００１５】
図１に転じて、システム１００は、記憶装置１０４及びプロセッサ１０６を有する一又は複数のサーバ１０２を含んでもよい。サーバ１０２は、下記のプロセスのすべてを制御し得るし、本技術において知られたように多くのサーバに分散され得る。サーバ１０２は、サーバファーム内の多くのサーバのうちの一つであってもよいし、数多くのサーバが地理的に分散されて各々が下記のタスクのすべてを実行してもよいし、複数のタスクが複数のサーバ１０２の間で分割されてもよい。各サーバ１０２は、機能に必要なプログラミング及び下記に説明するデータを記憶するための記憶装置若しくはメモリ１０４を有してもよいし、中央記憶装置にリンクされてもよい。一例では、記憶装置１０４は非一時的なメモリである。さらに、プロセッサ１０６は、下記のタスクを実行するために使用されてもよいし、特定のタスクが、複数のプロセッサの間で分割されてもよいし、複数のプロセッサが一つのタスクを完了させるために必要とされてもよい。
【００１６】
サーバ１０２は一又は複数のユーザ装置１１０にネットワーク１０８で接続される。ネットワーク１０８は、インターネットプロトコル（ＩＰ）に基づくネットワーク、ローカルエリアネットワーク（ＬＡＮ）、ワイドエリアネットワーク（ＷＡＮ）、パーソナルエリアネットワーク（ＰＡＮ）、イントラネット、インターネット、セルラネットワーク（例えば、ＧＳＭ（グローバルシステムフォーモバイルコミュニケーションズ）、ＣＤＭＡ（符号分割多重アクセス）、ＷＣＤＭＡ（広帯域ＣＤＭＡ）、ＬＴＥ（ロングタームエボリューション）、ＩＥＥＥ８０２．１１ｘ、等）、光ファイバネットワーク、又はデータを送信することができる別のタイプのネットワークなどの、一又は複数のパケット交換ネットワークを含んでもよい。ネットワーク１０８は、旧来型の電話機に対する電話サービスを提供するための公衆交換電話ネットワーク（ＰＳＴＮ）などの、回路交換ネットワークを含んでもよい。
【００１７】
ユーザ装置１１０は、インターネット１０８と通信する一又は複数の装置を含んでもよい。例えば、ユーザ装置１１０は、インターネット１０８に接続するためのアプリケーション（例えば、インターネットエクスプローラ（登録商標）、クローム（登録商標）等）及び通信インターフェース（例えば、有線又は無線通信インターフェース）を含むテレビを含んでもよい。ユーザ装置１１０は、インターネットサービスを提供するためにインターネット１０８と通信する一又は複数の装置も含んでもよい。例えば、ユーザ装置１１０は、デスクトップコンピュータ、ラップトップコンピュータ、パームトップコンピュータ、ノートブック、タブレット、スマートフォン、等、又は別のタイプの通信装置を含んでもよい。ユーザ１１２は、ネットワーク１０８を介してサーバ１０２にアクセスするためにユーザ装置１１０を利用し得る。

【００１８】
サーバ１０２は、ユーザ装置１１０を使用してユーザ１１２によってアクセスされることが可能なウェブページ１１４をホスティングし得る。図２Ａでは、ウェブページ１１４は、半構造化されたデータフィールド２００の要素を有する。この例では、半構造化されたデータフィールド２００は、レシピを入力するためのユーザ１１２用のデータフィールドを表示する。しかしながら、半構造化されたデータフィールド２００は、ホテル、レストラン、又は旅行行き先情報を含む任意の別のタイプの半構造化されたデータを表示し得る。

【００１９】
本明細書において使用するように、「半構造化されたデータ」は、多くの制約なしにユーザによって入力されることが可能な情報である。半構造化されたデータは、データフィールド２００の名称から伝わる構造を有することができ、ユーザがこれらのフィールドに任意の情報を入力できるようには構造化されていない。半構造化されたデータ及びデータフィールドの例は以下に示す。比較すると、「構造化されたデータ」は、ユーザが登録するために多くの制約を有するデータである。構造化されたデータの例は、２値の又は固定されたデータ選択肢、例えば、「はい／いいえ」形式の質問、１～１０までの尺度へのランク付け、又はプルダウン選択肢を含む。「構造化されていないデータ」は、登録のために制約がなく、基本的にユーザ１１２は任意のタイプのデータを入力するために空白のページを与えられる。

【００２０】
図２Ａの例に関して、半構造化されたデータフィールド２００はレシピ名称２０２、原材料２０４、命令又はステップ２０６、タグ２０８、及び画像２１０を含んでもよい。レシピ名称２０２は、後に続くレシピの簡単な記述語である。レシピ名称は、冷製ポテトスープ若しくはミートソース中のパスタのような通称、または、ビシソワーズ若しくはスパゲッティボロネーゼなどのような正式名称を含んでもよい。半構造化されたという用語は、ユーザ１１２がこのフィールドに任意の単語を入力できるという理由でレシピ名称２０２に適用されるが、構造は、入力されたものがレシピの名前であるという事実から来る。

【００２１】
原材料２０４は、原材料量２０４Ａ、原材料名２０４Ｂ、及び原材料修飾語２０４Ｃに関するデータフィールドがそこにあり得るという点で、半構造化され、体系化され、及び違ったふうに入力され得る。これらのフィールドのエントリは、「２切れ」、「ニンニク」及び「ミンチした」、又は「３カップ」、「小麦粉」及び「ふるいにかけた」を含み得る。一実施形態では、原材料フィールド２０４Ａ、２０４Ｂ、２０４Ｃのすべてが、すべて半構造化され、３つのフィールドに任意の数又は値を可能にする。或いは、フィールド２０４Ａ、２０４Ｂ、２０４Ｃのうちの一又は複数は、例えば、量２０４Ａ又は原材料２０４Ｂに関するプルダウンリストを含んで、構造化され得る。プルダウンメニューは、「カップ」、「小さじ」、「大さじ」、「スティック」等の最も一般的なデータ入力リストを含んでもよい。原材料名２０４Ｂ選択は、「小麦粉」、「砂糖」、及び「バター」のような簡単な記述語であってもよく、修飾語２０４Ｃは、「多用途の」、「粉末化した」及び「無塩の」から、「皮をむいた」、「刻んだ」又は「ふるいにかけた」までのいずれであってもよい。

【００２２】
ステップ２０６は、ユーザ１１２がレシピを準備し料理するための命令を入力することが可能な一つのフィールド又は多重フィールド２０６Ａ．．．２０６ｎであり得る。多重フィールドの例では、命令は、個々のステップ、すなわち、「ミキサ内のクリームバター」又は「複数の卵を１個ずつ加え、追加する度にボールの側部をこする」へと分解され得る。ステップ２０６は、原材料フィールド２０４よりも構造化されず、一般的には構造化されていないデータフィールドである。別の一例では、画像をステップ２０６のうちの一又は複数に対して追加することが可能であり、そのステップについて必要な技術を図説する。画像は、バッター濃度、形成後の形状、又は縛り目でさえも含んでもよい。

【００２３】
１以上のタグ２０８Ａ～２０８Ｃが含まれてもよい。タグ２０８は、いくつかの例では、料理法、料理内容、食事、及びイベントを表示し得る。このように、タグ２０８は「日本料理」、「中国料理」、「イタリア料理」、又は「韓国料理」のような料理法についてであってもよい。料理内容タグ２０８は、「スープ」、「サラダ」、「魚介類」、又は「肉類」であってもよい。食事タグ２０８は、「朝食」、「昼食」、「夕食」、「前菜」等であってもよく、イベントタグは、「休日」、「ピクニック」、又は「感謝祭」であってもよい。タグ２０８を用いると、ユーザ１１２は、レシピをカテゴリ分けするために使用可能な追加の短い識別子を追加し得る。上記のように、これらのタグ２０８のうちのいくつかは構造化されたデータとして提示されることが可能であり、「日本料理、中国料理、イタリア料理、韓国料理、フランス料理、ロシア料理、ドイツ料理、タイ料理、インド料理、ドイツ料理、及びメキシコ料理」のような料理法に関するプルダウンメニューをユーザ１１２に提供する。

【００２４】
追加例では、他のデータ２１０が入力されてもよい。他のデータ２１０は、レシピの画像であってもよい。本明細書において使用するように、「画像」は、完成したレシピ又はレシピのステップの任意の表示であることに留意すべきである。画像という用語は、静止画、動画、及び又はオーディオファイルを識別するために使用され得る。さらに、他のデータ２１０は、レシピを特別にさせる若しくはユーザ１１２がレシピ、すなわち料理本の名前を入手する場合に役立つユーザ１１２によって共有される特別な秘密、又はレシピを伝えた家族の名前であってもよい。

【００２５】
一つのレシピについてのこれらのデータフィールド２００の集合が、半構造化されたデータエントリ２１２である。半構造化されたデータエントリ２１２は、上記の半構造化されたデータフィールド２００を含む。一例では、データエントリ２１２は、一つの特定のレシピに入力されたデータフィールド２００のすべてを考慮し得る。図３Ａは、簡単なレシピを使用するデータエントリ２１２を一例として示す。データエントリ２１２はホイップクリームに関するものであり、データフィールド２００の各々は一つのエントリ２１２にリンクされ、ユーザ１１２がレシピをたどることを可能にする。

【００２６】
半構造化されたデータフィールド２００が半構造化されたものであるという理由だけで、これらのフィールドに入力されたデータは一般的には整合性がない。ユーザは単語を書き誤るか、原材料に対して別の名前を使用するし、情報（最も一般的にはタグ２０８）を省略することがある。この整合性のないデータエントリは、カテゴリ分けすること、検索すること、及び別のユーザにデータを送り返すことに関係する問題を提起する。本発明の例では、データから構造化された知識を抽出し、よりデータを簡単にカテゴリ分けすること、検索すること、及び送り返すことを可能にするために、半構造化されたデータを解析できる。

【００２７】
レシピ例では、構造化された知識は、類似の名前を付けられたレシピに関係させることが可能であり、そのため一つの検索語がすべてを返すことが可能である。このように、ユーザ１１２が「ポテトスープ」、「冷製ポテトスープ」、又は「ビシソワーズ」を入力するかに拘わらず、これらの語のいずれか一つに関する検索が入力されると、レシピのすべてが返され得る。さらに、与えられた原材料に関する典型的な料理が返されてもよい。例えば、ユーザが「ナス」を検索する場合には、システム１００は、「焼きナス」、「麻婆ナス」、「ナスのから揚げ」、「ナスのパルミジャーナ」、「ムサカ」、及び「ラタトゥイユ」に関するレシピを返すことが可能である。関係するレシピを用いて、料理名の類義語、重要な原材料、及び国料理の知識は、関連する結果を返す際に全面的に支援し得る。

【００２８】
図３Ａ、図３Ｂ、及び図４は、半構造化されたデータから構造化された知識を抽出する

方法の例を示す。レシピ例を続ける際に、システム１００は既に、メモリ１０４に記憶された半構造化されたデータのレポジトリ３００を有する。半構造化されたデータは、半構造化されたデータエントリ２１２として記憶される（ステップ４００）。数多くの半構造化されたデータエントリ２１２がある。レシピ例では、５００，０００個のユーザ入力レシピであってもよい。次に、半構造化されたデータエントリ２１２のレポジトリは各エントリ内の半構造化されたデータフィールド２００によってソートされ得る（ステップ４０２）。例えば、タグ２０８は、どのレシピが国料理タグでタグ付けされるかによってソートされ得る。次に、データエントリ２１２のサブセット３０２がそれらのトピック３０４によって選択され得る（ステップ４０４）。このサブセットは、訓練データ３０２と呼ばれることがある。

【００２９】

例では、トピックは、国料理タグ２０８において識別した一又は複数の国である。より具体的には、５００，０００個のレシピのうちの２０，０００個だけが国料理タグ２０８を有する。タグ２０８は、日本料理、中国料理、イタリア料理、等であってもよい。いったんトピックが選択されると、そのデータフィールド２００はトピックデータフィールド３０４になり、したがって、国料理タグフィールドはトピックデータフィールドになる。よって、データサブセット３０２又は訓練データは、国料理タグを有する２０，０００個のレシピであり得る。

【００３０】

各トピック３０４について、トピック３０４及びトピックデータフィールドに関係する１以上の他の半構造化されたデータフィールド２００が選択される（ステップ４０６）。これらの他の半構造化されたデータフィールド２００はデータサブセット３０２から選択されるだけである。選択された半構造化されたデータフィールドは特徴データフィールド３０６であり得る。特徴データフィールド３０６の情報はトピックデータフィールドの１以上のトピックに多少なりとも関係する。この例では、レシピ名２０２及び鍵となる原材料２０４の両者が、国料理の特徴と考えられる。このように、「ミソ」、「カツ」、「スシ」、「炒め物」、「ロウミン」、「フーヤン」、「パルミジャーナ」、「スパゲッティ」、及び「ウォッカソース」のような単語を含むレシピ名２０２は、日本料理、中国料理、及びイタリア料理のトピックに関連し得る。これらは、このトピック３０４の特徴３０６のうちの一つである。さらに、ノリ、マグロ、練りミソ、海鮮醤、チンゲン菜、ビーフン、トマトソース、モツァレラ、及びブロッコリラーブは、同じ料理の特徴である原材料２０４である。

【００３１】

一例では、レシピの最大部分である又はレシピ中に最も多く存在する原材料が、特定の国料理の特徴として典型的に選択される。しかしながら、コメは日本料理と中国料理との間でおそらく同じように共通であり、ニンニクは中国料理とイタリア料理との間で共通であるので、何らかの注意を払わなければならない。

【００３２】

いったん、トピック３０４及び特徴３０６が識別されると、メモリ１０４に記憶された半構造化されたデータの残りのレポジトリ３０８の解析が実行される（ステップ４０８）。解析は、選択したデータフィールド２００により残りのデータエントリ２１２を分類するために実行される。レシピの残りのレポジトリ３０８はそれらの特定の特徴３０６について解析され、レポジトリがどのトピック３０４に属するかを決定する。より簡単に、レシピ名及び原材料が数学モデルを使用して解析され、これらがどの国料理に属する可能性があるかを決定する。

【００３３】

解析は、最大エントロピー分類法（Ｍａｘｉｍｕｍ　Ｅｎｔｒｏｐｙ　Ｃｌａｓｓｉｆｉｅｒ）及びレイテントディリクレアロケーション（Ｌａｔｅｎｔ　Ｄｉｒｉｃｈｌｅｔ　Ａｌｌｏｃａｔｉｏｎ）モデル（以降「ＬＤＡ」）の混成を使用して実行され得る。両者は、自然言語処理及び機械学習において使用される数学的技術である。最大エントロピ

ーの理論はベイズ統計に基づく。この技術は、ある種の語が文書の文脈内に存在する確率を推定するためにデータの検証可能なサブセットを必要とする。ＬＤＡも、データの類似度を判断するための統計モデルである。ＬＤＡは、特定のトピック（例では料理法）によって各文書（又はこの例ではレシピ）を特徴付ける。一般的な例として、文書が「子犬」又は「吠え声」のような単語を含む場合には、文書は「犬」トピックを有すると判断され得る。例では、レシピ名が「パルミジャーナ」という単語を含み、鍵となる原材料としてトマトを使用する場合には、レシピは「イタリア料理」であると判断され得る。

【００３４】

ＬＤＡ構成要素は、トピック確率のベクトルを用いて各データエントリ（レシピ）を記述し得る。このベクトルはレシピの「スコア」と考えられ得る。この方法は、解析する次元が少ないので、「バッグオブワーズ（Ｂａｇ　ｏｆ　Ｗｏｒｄｓ）」モデルを使用するよりも単純である。バッグオブワーズモデルは、文書内の単語の出現頻度を主に検査する。これは、ほとんどすべての単語が散在して使用され通常は共通の繰返し数を有するレシピで利用するのは困難である。

【００３５】

ＬＤＡモデルは、各料理についてＬＤＡ重心を計算する（ステップ４０８）。ＬＤＡ重心は、一例では、所定の国からのすべてのレシピからのトピック全体に渡る合計であり、所定の料理に関する「重心ベクトル」を計算する。初期の訓練データ３０２又は更新された訓練データ３１０（図３Ｂ参照）は、料理重心に対するコサイン類似度に基づいて選択され得る（ステップ４１０）。続いて、更新された訓練データ３１０は以前の訓練データ３０２と統合される（ステップ４１２）。トピック及び特徴を決定する上記のプロセス（ステップ４０６）が繰り返され、次に、残りのレポジトリ３０８ａが再び解析される。一例では、更新された訓練データ３１０が、正確に決定された料理を有するレシピ（データエントリ）の上位１０％であるという理由で、選択される。このプロセスは、いったん更新された訓練データ３１０が解析されると、重心ベクトルを変更することを許容する。プロセスが繰り返されるので、レシピ名は、訓練データ更新、及び最大エントロピー分類法によって高い信頼度で分類されたレシピから、訓練データ中の初期の国がタグ付けされたレシピに基づいて抽出され得る。

【００３６】

特徴が２値であるので、上記の混成最大エントロピー及びＬＤＡモデルは相対的に単純化される、すなわち、料理名若しくは原材料名がレシピ中に存在するか又は第１の事例には存在しない。これが上記のスコア化を単純化する。一例として、レシピ名及び料理法に基づくレシピのスコアは、

【数１】

$$z-score(dish,cuisine)=\frac{f(dish,cuisine)-expected}{st.dev}=$$

$$\frac{f(dish,cuisine)-p(dish,all\_cuisine)*f(all\_dish,cuisine)}{\sqrt{f(all\_dish,all\_cuisine)*p(dish,all\_cuisine)*(1-p(dish,all\_cuisine))}}$$

であり得、ここでは、「ｚスコア」は標準スコアであり、関数ｆはレシピの生のスコアを計算するために使用され、ｐは確率である。

【００３７】

中国料理、イタリア料理及び韓国料理の料理について実行したときに、繰返し反復は有望な結果をもたらした。結果は、

【表１】

反復	P	R
反復０	91.67%	70.21%
反復１	92.50%	78.72%
反復２	92.85%	82.98%

であり、ここでは、「Ｐ」は精度であり、「Ｒ」は再現度又は相関性である。スコアがすべての一般的でない料理について検査されたときの結果は

【表２】

反復	P	R
反復０	63.51%	40.87%
反復１	68.29%	48.69%
反復２	67.40%	50.43%

である。

【００３８】

サンプル料理名の評価例を約４００，０００個までのエントリレシピデータベースについて実行した。結果は驚くべきものであった。料理が正しいかどうかを判断するために、１よりも大きなｚスコアを有するレシピを各反復の後で解析した。

【表３】

	正解	恐らく	不正解	合計	%正解
中国料理０	73	4	8	85	85.88%
中国料理１	96	12	9	117	82.05%
中国料理２	98	11	11	120	81.67%
イタリア料理０	46	5	2	53	86.79%
イタリア料理１	81	16	13	110	73.64%
イタリア料理２	82	16	12	110	74.55%
韓国料理０	30	0	0	30	100.00%
韓国料理１	43	2	0	45	95.56%
韓国料理２	45	2	0	47	95.74%

【００３９】

レシピ名２０２を使用することの利点は、オペレータが容易で迅速な評価を行うことが可能であることであり、レシピ毎にチェックする必要がない。さらに、解析は能動的学習方式で用いることが可能であり、すなわち、レシピ名の人間の評価結果は、訓練データの更新のために使用され得る。

【００４０】

図５に図示したもう一つの例は、どのレシピが相互に類似しているかを判断するために、レポジトリ３００を解析する。解析の一つのレベルは、レシピ名２０２を一致させることである。もう一つは、原材料２０４に基づいて類似度をチェックすることである。レシピに関するもう一つのベクトルが、「語出現頻度－文献出現頻度の逆数」（本明細書では「ＴＦ－ＩＤＦ」）として知られる方法を使用して計算され得る（ステップ５００）。ＴＦ－ＩＤＦは、文献集内のある文献に対して単語がどれだけ重要であるかを反映する数値統計である。レポジトリ内の一つのレシピに対して原材料がどれだけ重要であるかが、例

に対して上手く記述された。この文脈における「語」は原材料であり、「文献出現頻度」はその原材料がその中に現れるレシピタイプの数である。いったんベクトルがレポジトリ３００内のすべてのレシピについて計算されると、コサイン類似度が計算され得る（ステップ５０２）。コサイン類似度は、２つのベクトルがどれだけ類似しているかを評価する。ベクトルが相互に良く類似しているほど、レシピが相互に良く類似している可能性がある。

【００４１】

図６は、さらなる精度でレシピ類似度を判定することが可能なより進んだ例を示す。原材料データフィールド２０４を使用することに加えて、もう一つの次元又は変数が使用され得る。上の例を有する文脈では、もう一つのデータフィールド２００が使用され得る。例えば、料理データフィールド２０８、別のタグ情報２０８、又はステップ／準備方法２０６が解析において含まれ得る。レポジトリ３００内のレシピ（データエントリ２１２）に関するベクトル又はスコアは、概念に対して原材料をマッピングすることによって又は概念に対する原材料のレシピから計算されることが可能であり、これらの類似度を比較する。このように、原材料のベクトルとして各レシピ／料理内容を表示すること及びこれらのベクトルを使用してレシピ類似度を比較することの代わりに、まず、各原材料は、レシピカテゴリのＴＦ－ＩＤＦ値にしたベクトルとして表示され得る。レシピ名２０２及び原材料２０４以外のもう一つのデータフィールド２０４が選択されてもよい。次に、各レシピ名についてのベクトルが、その原材料ベクトルの重心（又は単純に平均）として計算され得る。

【００４２】

この方法は明示的意味解析（Ｅｘｐｌｉｃｉｔ　Ｓｅｍａｎｔｉｃ　Ａｎａｌｙｓｉｓ）（「ＥＳＡ」）に匹敵する。ＥＳＡは、自然言語処理の形態であり、テキスト（個々の単語又は全体の文書）のベクトル表現である情報検索である。具体的に、ＥＳＡでは、単語はテキストのＴＦ－ＩＤＦ行列の列ベクトルとして表示され、文書（一連の単語）はその単語を表すベクトルの重心として表示される。

【００４３】

本方法では、ユーザは、レシピ名２０２及び原材料２０４フィールドとともに含むように追加のデータフィールド２０４を選択する（ステップ６００）。次に、フィールドに基づいて各原材料についてのＴＦ－ＩＤＦ値にしたベクトルを計算する（ステップ６０２）。原材料ベクトルを用いて、原材料ベクトルの重心としてレシピ名ベクトルを計算する（ステップ６０４）。

【００４４】

任意選択で、ラグランジアンカーネル（Ｌａｇｒａｎｇｉａｎ　ｋｅｒｎｅｌ）が、カテゴリ毎の類似度を考慮しスコアを平準化することが可能なレシピ－カテゴリ行列に適用されてもよい。ラグランジアンの方法はレシピ及びカテゴリの力学をまとめることが可能である。

【００４５】

論じてきている構成要素、ステップ、機能、目的、利益及び利点は、単に例示的である。これらのいずれも又はこれらに関係する議論は、多少なりとも保護の範囲を限定しないものとする。数多くの他の実施形態もやはり想定される。これらは、より数少ない、追加の及び／又は異なる構成要素、ステップ、機能、目的、利益及び利点を有する実施形態を含む。これらは、構成要素及び／又はステップが別なふうに配置される及び／又は順番にされる実施形態をやはり含む。

【００４６】

別なように述べない限り、別記の特許請求の範囲を含む本明細書において記述するすべての測定値、値、格付け、位置、強度、サイズ、及び他の仕様は、厳密ではなくおおよそである。これらは、これらが関係する機能と矛盾せず及びこれらが属する技術において慣用的である妥当な範囲を有するものとする。

【００４７】

この開示において引用してきているすべての論文、特許、特許出願、及びその他の刊行物は、参照によって本明細書に組み込まれる。

【００４８】

特許請求の範囲において使用されるときに「～のための手段」という句は、説明してきている対応する構造及び材料並びにこれらの等価物を包含するものであり、包含するように解釈すべきである。同様に、特許請求の範囲において使用されるときに「～のためのステップ」という句は、説明してきている対応する行為及びこれらの等価物を包含するものであり、包含するように解釈すべきである。特許請求の範囲においてこれらの句がないことは、特許請求の範囲が対応する構造、材料、若しくは行為のいずれかに又はこれらの等価物に限定されないものであり、限定されるように解釈すべきではない。

【００４９】

記述してきた又は図示してきたものは、特許請求の範囲において述べられているかどうかに拘わらず、一般に公開された任意の構成要素、ステップ、機能、目的、利益、利点、又は等価物に限ることを意図するものではなく、限定するように解釈すべきではない。

【００５０】

保護の範囲は、別記の特許請求の範囲によってのみ限定される。その範囲は、本明細書及び後に続く法的処置履歴を考慮して解釈するときに、特許請求の範囲において使用される言語の通常の意味と整合するようにできるだけ広いものであり、かつ広くなるように解釈すべきであり、そしてすべての構造的及び機能的等価物を包含するものであり、かつ包含するように解釈すべきである。

【図１】

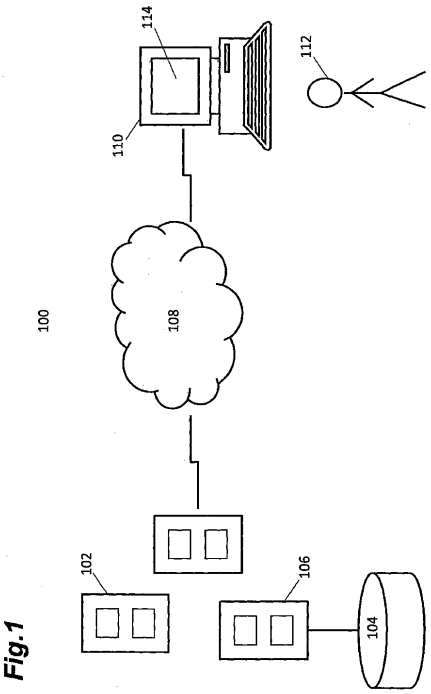
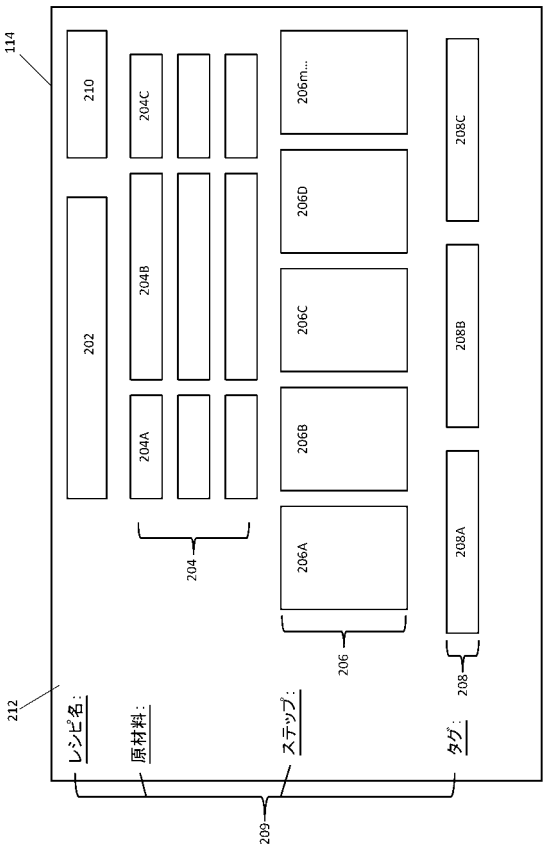
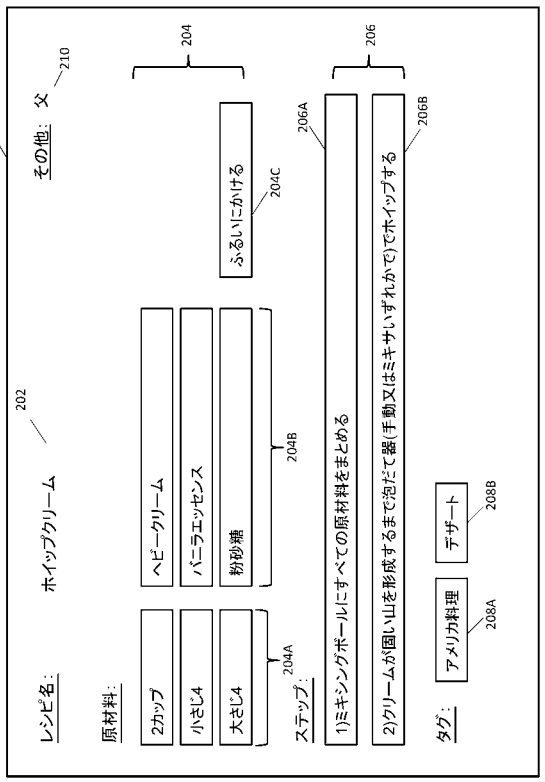


Fig.1

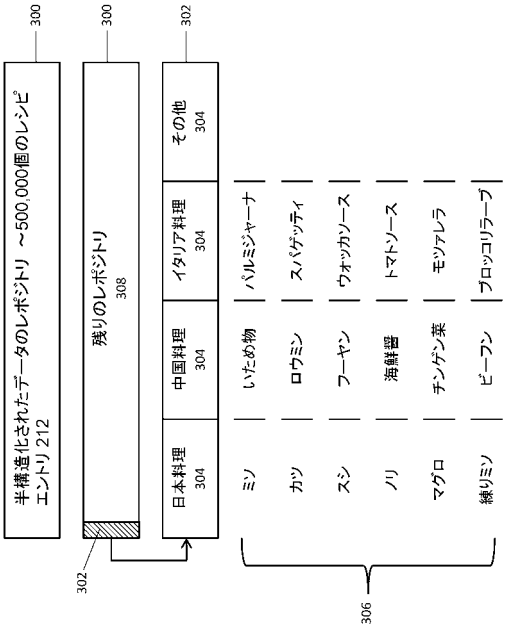
【図２Ａ】



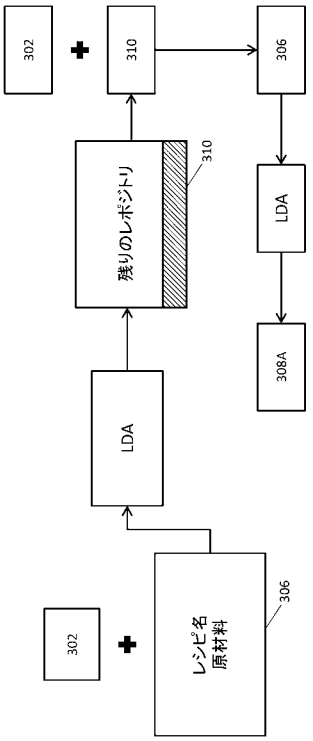
【図２Ｂ】



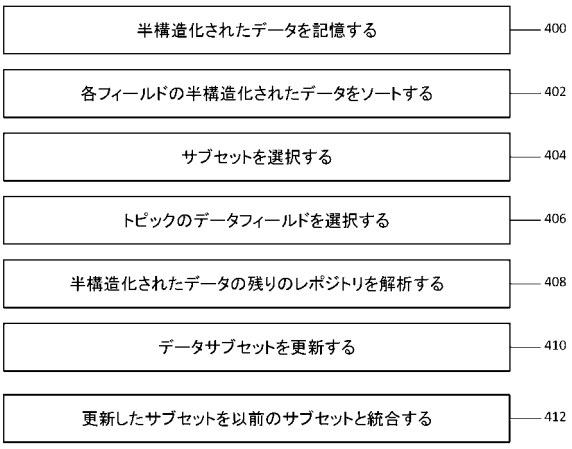
【図３Ａ】



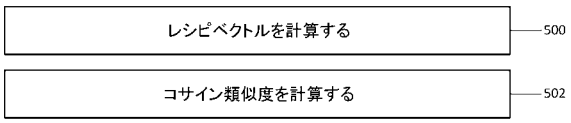
【図３Ｂ】



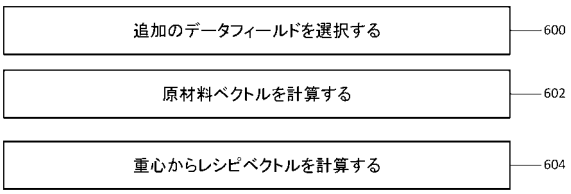
【図４】



【図５】



【図６】



【手続補正書】
【提出日】平成26年6月25日(2014.6.25)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
所定のネットワークを介して複数のユーザにトピック、特徴、および前記トピックの属性を提供するコンピュータシステムであって、
複数の半構造化されたデータエントリを記憶するメモリであって、それぞれの半構造化されたデータエントリが所定のトピックおよびトピックデータフィールドを提供する、該メモリと、
前記半構造化されたデータエントリのサブセットを選択し、前記トピックに関連する半構造化されたデータフィールドと前記トピックデータフィールドとのうちの少なくとも一つを選択し、選択された前記半構造化されたデータフィールドの前記トピックおよび前記特徴にデータ解釈用アルゴリズムを適用することで前記トピックの属性を決定するプロセッサと
を備えるコンピュータシステム。
【請求項２】
前記トピック、前記特徴、および前記トピックの属性がそれぞれ、レシピ名称、原材料、および国料理に対応する、
請求項１に記載のコンピュータシステム。
【請求項３】

前記データ解釈用アルゴリズムが、混成最大エントロピー及びＬＤＡモデルと、語出現頻度－文献出現頻度の逆数と、コサイン類似度解析とのうちの少なくとも一つを備える、請求項１に記載のシステム。
【請求項４】
前記半構造化されたデータエントリのサブセットが訓練データであり、
前記プロセッサが、更新された訓練データを以前の訓練データと統合し、前記トピックおよび前記特徴の決定と前記データ解釈用アルゴリズムの適用とを繰り返す、
請求項１に記載のシステム。
【請求項５】
所定のネットワークを介して複数のユーザにトピック、特徴、および前記トピックの属性を提供するための方法であって、
複数の半構造化されたデータエントリをメモリに記憶するステップであって、それぞれの半構造化されたデータエントリが所定のトピックおよびトピックデータフィールドを提供する、該ステップと、
プロセッサが、前記半構造化されたデータエントリのサブセットをそれらのトピックに基づいて選択するステップと、
前記プロセッサが、前記トピックに関連する半構造化されたデータフィールドと前記トピックデータフィールドとのうちの少なくとも一つを選択するステップと、
前記プロセッサが、前記トピックの属性を決定するために、選択された前記半構造化されたデータフィールドの前記トピックおよび前記特徴にデータ解釈用アルゴリズムを適用するステップと、
を含む方法。
【請求項６】
前記トピック、前記特徴、および前記トピックの属性がそれぞれ、レシピ名称、原材料、および国料理に対応する、
請求項５に記載の方法。
【請求項７】
前記データ解釈用アルゴリズムが、混成最大エントロピー及びＬＤＡモデルと、語出現頻度－文献出現頻度の逆数と、コサイン類似度解析とのうちの少なくとも一つを備える、請求項５に記載の方法。
【請求項８】
前記半構造化されたデータエントリのサブセットが訓練データであり、
前記方法が、
前記プロセッサが、更新された訓練データを以前の訓練データと統合し、前記トピックおよび前記特徴の決定と前記データ解釈用アルゴリズムの適用とを繰り返すステップを更に含む、
請求項５に記載の方法。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/JP2013/084169

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/30
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2009/009815 A1 (KARASIK GREGORY [US] ET AL) 8 January 2009 (2009-01-08) the whole document	1-9
Y	KUNLUN LI ET AL: "Multi-class text categorization based on LDA and SVM", PROCEDIA ENGINEERING, vol. 15, 31 December 2011 (2011-12-31), pages 1963-1967, XP028337688, ISSN: 1877-7058, DOI: 10.1016/J.PROENG.2011.08.366 [retrieved on 2011-12-06] the whole document	1-9
X	US 5 960 440 A (BRENNER RICHARD K [US] ET AL) 28 September 1999 (1999-09-28) the whole document	1-9

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 February 2014	07/03/2014

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	de Castro Palomares

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/JP2013/084169

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	SERAFETTIN TASCI ET AL: "LDA-based keyword selection in text categorization", COMPUTER AND INFORMATION SCIENCES, 2009. ISCIS 2009. 24TH INTERNATIONAL SYMPOSIUM ON, IEEE, PISCATAWAY, NJ, USA, 14 September 2009 (2009-09-14), pages 230-235, XP031549461, ISBN: 978-1-4244-5021-3 the whole document	1-9
A	US 6 970 881 B1 (MOHAN RENGASWAMY [US] ET AL) 29 November 2005 (2005-11-29) the whole document	1-9

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/JP2013/084169

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2009009815 A1	08-01-2009	NONE	
US 5960440 A	28-09-1999	NONE	
US 6970881 B1	29-11-2005	NONE	

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JP,KE,KG,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LT,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US

（特許庁注：以下のものは登録商標）
１．ＧＳＭ
２．ＷＣＤＭＡ