

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第6411800号
(P6411800)

(45) 発行日 平成30年10月24日(2018.10.24)

(24) 登録日 平成30年10月5日(2018.10.5)

(51) Int.Cl.		F I
G06F 17/30	(2006.01)	G06F 17/30 340A
G06F 19/00	(2018.01)	G06F 19/00 130
G06F 13/00	(2006.01)	G06F 17/30 170C
		G06F 17/30 210D
		G06F 13/00 560A

請求項の数 10 (全 16 頁)

(21) 出願番号 特願2014-147085 (P2014-147085)
 (22) 出願日 平成26年7月17日(2014.7.17)
 (65) 公開番号 特開2016-24545 (P2016-24545A)
 (43) 公開日 平成28年2月8日(2016.2.8)
 審査請求日 平成29年2月9日(2017.2.9)

(73) 特許権者 392026693
 株式会社NTTドコモ
 東京都千代田区永田町二丁目11番1号
 (74) 代理人 100125689
 弁理士 大林 章
 (72) 発明者 菊地 悠
 東京都千代田区永田町二丁目11番1号
 株式会社NTTドコモ内
 (72) 発明者 落合 桂一
 東京都千代田区永田町二丁目11番1号
 株式会社NTTドコモ内
 (72) 発明者 鳥居 大祐
 東京都千代田区永田町二丁目11番1号
 株式会社NTTドコモ内

最終頁に続く

(54) 【発明の名称】 情報管理装置、情報管理システム、及び情報管理プログラム

(57) 【特許請求の範囲】

【請求項1】

複数の投稿情報の中から、当該投稿情報に含まれる地理的情報に基づいて、スポットに関連する投稿情報をマッチング処理により選別するスポット投稿選別部と、

前記スポット投稿選別部によって選別された複数の投稿情報の中から、当該投稿情報に含まれる文言に基づいて、イベントに関連する投稿情報を機械学習により選別するイベント投稿選別部と、

前記スポット投稿選別部および前記イベント投稿選別部によって選別された選別済み投稿情報を分析して、前記選別済み投稿情報に含まれるイベント名称と、当該イベント名称に対応するイベントが開催される開催スポット名称とを含むイベント情報を抽出するイベント情報抽出部とを備える

情報管理装置。

【請求項2】

抽出された前記イベント情報のうち同一のイベントに対応する複数のイベント情報を、1つの統合イベント情報に統合するイベント情報統合部をさらに備える

請求項1の情報管理装置。

【請求項3】

前記イベント情報統合部は、

1つの開催スポット名称に対応する複数のイベント名称の類似度に基づいて、複数の前記イベント情報が同一の前記イベントに対応すると判定する

請求項 2 の情報管理装置。

【請求項 4】

前記イベント情報抽出部は、
抽出された前記イベント情報のうち、排除リストに登録された単語に関連するイベント名称を含むイベント情報を排除する

請求項 1 の情報管理装置。

【請求項 5】

前記排除リストは前記スポットの名称を示すスポット名称を含み、
前記イベント情報抽出部は、
抽出された前記イベント情報のうち、前記排除リストに登録された前記スポット名称に一致するイベント名称を含むイベント情報を排除する

請求項 4 の情報管理装置。

【請求項 6】

前記イベント情報抽出部は、
選別された前記選別済み投稿情報のうち、NG ユーザリストに登録されたユーザによって投稿された投稿情報を、前記イベント情報の抽出対象から除外する

請求項 1 の情報管理装置。

【請求項 7】

前記イベント情報抽出部は、
前記選別済み投稿情報を分析して、前記イベント名称に対応する前記イベントが開催される開催日時を更に含む前記イベント情報を抽出する

請求項 1 の情報管理装置。

【請求項 8】

前記イベント情報抽出部は、
前記選別済み投稿情報を分析して、前記選別済み投稿情報に含まれる特徴語を更に含む前記イベント情報を抽出する

請求項 1 の情報管理装置。

【請求項 9】

複数の投稿情報の中から、前記投稿情報に含まれる地理的情報に基づいて、スポットに関連する投稿情報をマッピング処理により選別するスポット投稿選別部と、

前記スポット投稿選別部によって選別された複数の投稿情報の中から、前記投稿情報に含まれる文言に基づいて、イベントに関連する投稿情報を機械学習により選別するイベント投稿選別部と、

前記スポット投稿選別部および前記イベント投稿選別部によって選別された選別済み投稿情報を分析して、前記選別済み投稿情報に含まれるイベント名称と、当該イベント名称に対応するイベントが開催される開催スポット名称とを含むイベント情報を抽出するイベント情報抽出部とを備える

情報管理システム。

【請求項 10】

複数の投稿情報の中から、前記投稿情報に含まれる地理的情報に基づいて、スポットに関連する投稿情報をマッピング処理により選別することと、

前記選別された複数の投稿情報の中から、前記投稿情報に含まれる文言に基づいて、イベントに関連する投稿情報を機械学習により選別することと、

選別された選別済み投稿情報を分析して、前記選別済み投稿情報に含まれるイベント名称と、当該イベント名称に対応するイベントが開催される開催スポット名称とを含むイベント情報を抽出することと

をコンピュータに実行させる情報管理プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

10

20

30

40

50

本発明は、情報管理装置、情報管理システム、及び情報管理方法に関する。

【背景技術】

【0002】

近年、実世界におけるユーザの外出行動を支援するため、スマートフォンやPC等の情報端末を介して、観光スポット、地域に関連したイベント、地域に関連した商品（特産品）等の地域情報を提供する地域情報サービスが提供されている。以上の地域情報サービスの有用性は、最新の（up-to-dateな）地域情報が数多く提供されることにより向上する。

【先行技術文献】

【特許文献】

【0003】

【特許文献1】特開2011-145879号公報

【発明の概要】

【発明が解決しようとする課題】

【0004】

出願人の調査によれば、イベントに関する情報（以下、イベント情報と称する）の提供件数は、観光スポットや商品に関する情報の提供件数と比較して、顕著に少ない。地域情報サービスの有用性の観点からは、イベント情報がより多く提供されると好適である。また、イベントは、日時や期間を限定して開催されるため、頻繁な情報収集によってイベント情報の鮮度を保つことが重要である。しかしながら、多数のイベント情報を高頻度に手動で収集し続けることは、作業負荷及び作業コストの観点から困難である。

【0005】

以上の事情を考慮して、本発明は、イベント情報を適切に収集することを目的とする。

【課題を解決するための手段】

【0006】

本発明の情報管理装置は、複数の投稿情報から、当該投稿情報に含まれる地理的情報に基づいて、スポットに関連する投稿情報を選別するスポット投稿選別部と、複数の投稿情報から、当該投稿情報に含まれる文言に基づいて、イベントに関連する投稿情報を選別するイベント投稿選別部と、前記スポット投稿選別部および前記イベント投稿選別部によって選別された選別済み投稿情報を分析して、前記選別済み投稿情報に含まれるイベント名称と、当該イベント名称に対応するイベントが開催される開催スポット名称とを含むイベント情報を抽出するイベント情報抽出部とを備える。

【0007】

本発明の情報管理システムは、複数の投稿情報から、前記投稿情報に含まれる地理的情報に基づいて、スポットに関連する投稿情報を選別するスポット投稿選別部と、複数の投稿情報から、前記投稿情報に含まれる文言に基づいて、イベントに関連する投稿情報を選別するイベント投稿選別部と、前記スポット投稿選別部および前記イベント投稿選別部によって選別された選別済み投稿情報を分析して、前記選別済み投稿情報に含まれるイベント名称と、当該イベント名称に対応するイベントが開催される開催スポット名称とを含むイベント情報を抽出するイベント情報抽出部とを備える。

【0008】

本発明の情報管理方法は、複数の投稿情報から、前記投稿情報に含まれる地理的情報に基づいて、スポットに関連する投稿情報を選別することと、複数の投稿情報から、前記投稿情報に含まれる文言に基づいて、イベントに関連する投稿情報を選別することと、選別された選別済み投稿情報を分析して、前記選別済み投稿情報に含まれるイベント名称と、当該イベント名称に対応するイベントが開催される開催スポット名称とを含むイベント情報を抽出することとを備える。

【発明の効果】

【0009】

本発明によれば、複数の投稿情報からイベント情報が適切に収集される。

【図面の簡単な説明】

10

20

30

40

50

【 0 0 1 0 】

【図 1】実施形態に係る情報管理サーバと、情報管理サーバに関連するコンピュータ装置とを示す図である。

【図 2】実施形態のサーバの物理構成ブロック図である。

【図 3】実施形態のクライアント端末の物理構成ブロック図である。

【図 4】実施形態の情報管理サーバの論理構成ブロック図である。

【図 5】実施形態の投稿情報取得サーバの論理構成ブロック図である。

【図 6】実施形態のイベント情報収集動作を概略的に示すフローチャートである。

【図 7】スポット投稿選別動作を示すフローチャートである。

【図 8】イベント投稿選別動作を示すフローチャートである。

10

【図 9】機械学習（SVM）の流れを示すフローチャートである。

【図 10】学習データの一例を示す図である。

【図 11】イベント情報抽出動作を示すフローチャートである。

【図 12】機械学習（CRF）の流れを示すフローチャートである。

【図 13】学習データの一例を示す図である。

【図 14】イベント得情報統合動作を示すフローチャートである。

【発明を実施するための形態】

【 0 0 1 1 】

以下、図面を参照しながら本発明の実施形態を詳細に説明する。

【 0 0 1 2 】

20

1. 実施形態

1(1). 情報管理サーバの概略

図 1 を参照して、本発明の実施形態に係る情報管理サーバ M S と、情報管理サーバ M S に関係する複数のコンピュータ装置とを概略的に示す。本実施形態のコンピュータ装置は、所定のプロトコル（TCP/IP 等）に従って通信を実行する。

【 0 0 1 3 】

ブログサーバ B S は、ユーザ端末（たとえば、スマートフォン）から投稿される投稿情報 P I を受信して記憶し、外部からの要求に応じて投稿情報 P I を提供する。以下、本実施形態における投稿情報 P I の例を、非限定的に列挙する。

- マイクロブログサービス（例えば、Twitter [登録商標]）における 140 文字以下の短文投稿（例えば、ツイート [登録商標]）

30

- ブログ（ウェブログ）サービスにおける日記や寸評等の投稿

- ソーシャルネットワーキングサービス（SNS）におけるステータス投稿

【 0 0 1 4 】

投稿情報取得サーバ A S は、ブログサーバ B S から提供される投稿情報 P I を取得して蓄積する。投稿情報取得サーバ A S は、ブログサーバ B S に投稿される全ての投稿情報 P I を蓄積してもよいし、一部の投稿情報 P I（例えば、日本語の投稿情報 P I のみ）を蓄積してもよい。

【 0 0 1 5 】

また、投稿情報取得サーバ A S は、情報管理サーバ M S からの要求に応じて投稿情報 P I を情報管理サーバ M S に供給する。投稿情報取得サーバ A S は、蓄積している全ての投稿情報 P I を情報管理サーバ M S に供給してもよいし、一部の投稿情報 P I を供給してもよい。

40

【 0 0 1 6 】

情報管理サーバ M S は、投稿情報取得サーバ A S から供給される投稿情報 P I を解析してイベント情報 E I を収集及び蓄積すると共に、クライアント端末 C L からの要求に応じてイベント情報 E I を提示するコンピュータ装置である。本実施形態においては、日本語の投稿情報 P I が解析対象である。なお、所定の基準に従って限定された投稿情報 P I を解析対象としてもよい。例えば、特定のアカウント（Twitter 上の公式アカウント等）が投稿した投稿情報 P I のみを解析対象としてもよい。

50

【 0 0 1 7 】

投稿情報 P I に基づいて収集されるイベント情報 E I は、祝祭や展覧会、演奏会等のイベントに関する情報である。イベント情報 E I は、イベントの名称及びイベントが開催されるスポットの名称を少なくとも含む。また、イベント情報 E I は、イベントが開催される日時、イベントが開催される場所の緯度・経度、イベントに関するウェブサイトを指定する U R L、イベントに関連する特徴語等を含んでもよい。例えば、あるイベント情報 E I は、「イベント名称：サイエンスフェスティバル」、「開催スポット名称：ディスカバリーパーク焼津」、及び「開催日時：2014年3月23日11時00分から」という情報を含む。

【 0 0 1 8 】

クライアント端末 C L は、イベント情報 E I を情報管理サーバ M S に要求して取得すると共に、ユーザからの要求に応じてディスプレイに情報を表示する。例えば、クライアント端末 C L は、イベント情報 E I を電子地図上に重ねて表示する。

【 0 0 1 9 】

電子地図の表示に用いられる電子地図データは、ユーザからの要求に応じてクライアント端末 C L が地図サーバ（不図示）から取得してもよいし、予めクライアント端末 C L に記憶されていてもよい。また、イベント情報 E I と共に情報管理サーバ M S から提供されてもよい。その場合、情報管理サーバ M S は予め電子地図データを記憶していてもよいし、地図サーバから都度又は定期的に取り得てもよい。

【 0 0 2 0 】

1 (2) . 物理的構成

1 (2) - 1 . サーバ装置の構成

図 2 は、情報管理サーバ M S を始めとする本実施形態の各サーバ装置の物理的構成を示すブロック図である。各サーバ装置（ A S , B S , M S ）は、ネットワークインタフェース 1 0 と入力部 1 2 と出力部 1 4 と C P U （ Central Processing Unit ） 1 6 と R A M （ Random Access Memory ） 1 8 と R O M （ Read Only Memory ） 2 0 と H D D （ Hard Disk Drive ） 2 2 とを備える。

【 0 0 2 1 】

ネットワークインタフェース 1 0 は、ネットワークを介して他のコンピュータ装置と通信を実行する。入力部 1 2 は、キーボード等の入力装置からの入力信号を受け付ける。出力部 1 4 は、ディスプレイ等の出力装置に対して出力信号を送信する。C P U 1 6 は、主記憶装置である R A M 1 8 及び R O M 2 0 に記憶されているプログラムを実行することにより種々の制御及び演算を行う。H D D 2 2 は、R A M 1 8 上に展開可能なプログラム及びデータを記憶する補助記憶装置である。なお、H D D 2 2 に代えて又は加えて S S D 等の記憶媒体が採用されてもよい。

【 0 0 2 2 】

当業者が当然に理解する通り、1つのサーバが複数のコンピュータ装置によって構成されてもよいし、1つのコンピュータ装置が仮想化された複数のサーバを備えてもよい。

【 0 0 2 3 】

1 (2) - 2 . クライアント端末の構成

図 3 は、本実施形態のクライアント端末 C L の物理的構成を示すブロック図である。クライアント端末 C L は、ネットワークインタフェース 3 0 と入力部 3 2 と出力部 3 4 と C P U 3 6 と R A M 3 8 と R O M 4 0 と H D D 4 2 とを備える。以上の各要素は、各サーバが備える、図 2 を参照して説明された要素と同様の構成を有する。

【 0 0 2 4 】

1 (3) . 論理的構成

図 4 は、情報管理サーバ M S の論理的構成を示すブロック図である。情報管理サーバ M S は、投稿情報選別部 1 0 0 とイベント情報抽出部 1 1 0 とイベント情報統合部 1 2 0 と情報提示部 1 3 0 とを機能ブロックとして備える。投稿情報選別部 1 0 0 は、スポット投稿選別部 1 0 2 とイベント投稿選別部 1 0 4 とを備える。また、情報管理サーバ M S は、

10

20

30

40

50

スポット情報テーブル140とスポット投稿テーブル150とイベント投稿テーブル160とイベント情報テーブル170とを論理テーブルとして備える。

【0025】

各機能ブロックの概略的な動作を説明する。投稿情報選別部100は、投稿情報取得サーバASから供給される投稿情報PIを選別する。より具体的には、スポット投稿選別部102は、投稿情報PIに含まれる地理的情報に基づいて、スポット名称に関連する投稿情報PIを選別する。イベント投稿選別部104は、投稿情報PIに含まれる文言に基づいて、イベントに関連する投稿情報PIを選別する。イベント情報抽出部110は、選別された投稿情報PI(選別済み投稿情報SPI)を分析して、選別済み投稿情報SPIに含まれるイベント名称と、そのイベント名称に対応するイベントが開催される開催スポット名称とを含むイベント情報EIを抽出する。イベント情報統合部120は、抽出されたイベント情報EIのうち同一のイベントに対応する複数のイベント情報EIを、1つの統合イベント情報IEIに統合する。情報提示部130は、統合済みのイベント情報EI(統合イベント情報IEI)をクライアント端末CLに提示する。

10

【0026】

図5は、投稿情報取得サーバASの論理的構成を示すブロック図である。投稿情報取得サーバASは、ブログサーバBSから取得した投稿情報PIを蓄積する投稿情報処理部200を機能ブロックとして備えると共に、投稿情報テーブル210を論理テーブルとして備える。

【0027】

以上の機能ブロックは、各サーバ(AS, MS)の主記憶装置に記憶されているコンピュータプログラムをCPU16が実行することにより実現される。また、以上の論理テーブルは、複数のテーブルが所定の関係(リレーション)に基づいて連結される関係データベースの構成要素である。関係データベースは、不揮発性メモリであるHDD22にデータが記憶されるオンディスクデータベースで実装されてもよいし、揮発性メモリであるRAM18にデータが記憶されるインメモリデータベースで実装されてもよい。

20

【0028】

1(4) . イベント情報収集動作

本実施形態のイベント情報収集動作を以下に説明する。図6は、本実施形態のイベント情報収集動作を概略的に示すフローチャートである。ステップS1010及びS1020にて段階的に選別された投稿情報PIから、イベント情報EIが抽出される(S1030)。1つのイベントに関する複数のイベント情報EIがステップS1040にて統合(名寄せ)される。

30

【0029】

1(4) - 1 . スポット投稿選別動作(ステップS1010)

ステップS1010において、スポット投稿選別部102がスポット投稿選別動作を実行する。以下、図7を参照して具体的な動作を説明する。スポット投稿選別部102は、投稿情報取得サーバASから取得した複数の投稿情報PIの各々に対して形態素解析を実行し、各投稿情報PIに含まれる形態素を取得する(S1012)。

【0030】

次いで、スポット投稿選別部102は、各投稿情報PIについて、ステップS1012にて取得された形態素と、スポット情報テーブル140に含まれるスポット名称とのマッチングを実行する(S1014)。スポット情報テーブル140は、複数のスポット情報エントリを含む。各スポット情報エントリは、スポット名称及びスポットの緯度・経度を少なくとも含む他、スポットに関する説明文等の付加情報を含む場合がある。なお、同一のスポットに対応する複数のスポット名称(例えば、「金閣寺」と「鹿苑寺」)がスポット情報テーブル140に含まれると、投稿情報PIに含まれるスポット名称の揺らぎを吸収できるので好適である。

40

【0031】

ステップS1014のマッチング処理の結果、スポット名称が含まれると判定された投稿情報PIに対して、スポット投稿選別部102がフィルタリング処理を実行する(S1016)

50

。任意のフィルタリング処理が適用され得る。例えば、スポット名称でも人名でもあり得る語句（例えば、「松島」や「川崎」）が、スポット名称ではなく人名として使用されている投稿情報 P I を排除する人名フィルタリングや、複数のスポットに対応する 1 つのスポット名称（例えば、京都府の「円山公園」と北海道の「円山公園」）がどのスポットを示すかを判定する同名フィルタリングが実行される。複数のフィルタリング処理が並列的に実行されてもよい。なお、人名として使用され得るスポット名称に対応するスポット情報エントリに、その旨を示すフラグが付与されていると好適である。重複するスポット名称に対応するスポット情報エントリについても同様である。

【 0 0 3 2 】

以上のフィルタリング処理には任意のアルゴリズムが採用され得る。例えば、スポットに関する説明文から逆文書頻度（Inverse Document Frequency, I D F）に基づいて取得される語句が投稿情報 P I に含まれるか否かに応じて以上のフィルタリング（人名フィルタリング、同名フィルタリング）が実行されてもよい。また、人名フィルタリングに関しては、同一の語句が人名として使用されているかスポット名称として使用されているかを判定するための機械学習をスポット投稿選別部 1 0 2 に実行させた上で、その機械学習により構築されたモデルに基づいて人名かスポット名称かが判定されてもよい。以上の機械学習は、例えば、条件付き確率場（Conditional Random Field, C R F）に基づいて実行される。人名フィルタリングに関して、I D F によるフィルタリングと C R F によるフィルタリングとが並列して実行されてもよい。

【 0 0 3 3 】

以上のステップ S1010（ステップ S1012 から S1016）により、投稿情報 P I の全体からスポット名称に関連する投稿情報 P I が選別され、スポット投稿テーブル 1 5 0 に記憶される。なお、選別された投稿情報 P I がイベント投稿選別部 1 0 4 に直接的に供給されてもよい。

【 0 0 3 4 】

1 (4) - 2 . イベント投稿選別動作（ステップ S1020）

ステップ S1020 において、イベント投稿選別部 1 0 4 がイベント投稿選別動作を実行する。以下、図 8 を参照して具体的な動作を説明する。イベント投稿選別部 1 0 4 は、スポット投稿テーブル 1 5 0（又はスポット投稿選別部 1 0 2）から取得した複数の投稿情報 P I の各々から素性（feature）を抽出する（S1022）。

【 0 0 3 5 】

抽出される素性は、投稿情報 P I に含まれる文言に関する任意の指標である。例えば、投稿情報 P I に含まれる単語を示す単語ベクトルが素性の例として挙げられる。以上の単語ベクトルは、典型的には単語の出現頻度をも示すベクトルであるが、単語の出現頻度を示さない（単語の有無のみを示す）単語ベクトルが採用されてもよい。また、逆文書頻度に基づく単語の珍しさを示す指標が素性として採用されてもよい。

【 0 0 3 6 】

イベント投稿選別部 1 0 4 は、抽出された素性に基づいて、イベントに関連する投稿情報 P I を選別する（S1024）。投稿情報 P I がイベントに関連するか否かは、任意の基準によって判定される。例えば、投稿情報 P I とイベントとの関連性に関する機械学習の結果に基づいて、投稿情報 P I がイベントに関連するか否かをイベント投稿選別部 1 0 4 が判定すると好適である。

【 0 0 3 7 】

図 9 及び図 1 0 を参照して、機械学習の具体例を説明する。図 9 はステップ S1024 の推定フェーズに先立って実行される学習フェーズの動作フローであり、図 1 0 は学習フェーズで用いられる学習データの例である。本実施形態では、教師有り機械学習の一種であるサポートベクタマシン（Support Vector Machine, S V M）が採用される。S V M のカーネル関数として線形カーネルが採用されると処理負荷が軽いため好適であるが、他のカーネル関数が採用されてもよい。

【 0 0 3 8 】

学習フェーズにおいては、まず、イベントに関連するか否かを示すラベルが投稿情報 P I に付与される (S800)。イベント名称 (イベントの固有名称) を含む投稿情報 P I には「関連有り」を示すラベルが付与され、イベント名称を含まない投稿情報 P I には「関連無し」を示すラベルが付与される。図 10 には、ラベルが付与された後の学習データが示されている。

【0039】

ラベルが付与された投稿情報 P I の各々について、ステップ S1022 と同様の素性抽出処理が実行される (S810)。その後、「関連有り」ラベルが付与された投稿情報 P I を正例とし、「関連無し」ラベルが付与された投稿情報 P I を負例として、イベント投稿選別部 104 に機械学習を行わせる (S820)。

10

【0040】

ステップ S1024 では、以上の機械学習によって構築されたモデルに基づいて、イベントに関連すると推定される投稿情報 P I が選別される。選別された投稿情報 P I はイベント投稿テーブル 160 に記憶される。なお、選別された投稿情報 P I がイベント情報抽出部 110 に直接的に供給されてもよい。

【0041】

選別されなかった投稿情報 P I (イベントに関連しないと推定された投稿情報 P I) は破棄されてもよいし、何らかの目的のために (例えば、機械学習の負例として) 蓄積されてもよい。

【0042】

1(4) - 3. イベント情報抽出動作 (ステップ S1030)

ステップ S1030 において、イベント情報抽出部 110 がイベント情報抽出動作を実行する。以下、図 11 を参照して具体的な動作を説明する。イベント情報抽出部 110 は、イベント投稿テーブル 160 (又はイベント投稿選別部 104) から取得した複数の投稿情報 P I の各々から素性を抽出する (S1032)。

20

【0043】

抽出される素性は、各投稿情報 P I に含まれる文言に関する任意の情報である。本実施形態においては、投稿情報 P I の形態素解析により取得される単語の表記、品詞、原形、読み、及びグループ、並びに文字数及び文字種を含む情報が、素性として抽出される。

【0044】

イベント情報抽出部 110 は、抽出された素性に基づいてイベント情報 E I を抽出する (S1034)。前述の通り、イベント情報 E I はイベント名称と開催スポット名称とを含む。イベント名称は、任意の手法によって抽出される。例えば、投稿情報 P I に含まれるイベント名称に関する機械学習の結果に基づいた固有表現抽出によって、イベント情報抽出部 110 がイベント名称を抽出すると好適である。なお、以上の固有表現抽出によって開催日時も抽出されると更に好適である。

30

【0045】

図 12 及び図 13 を参照して、機械学習の具体例を説明する。図 12 はステップ S1034 の推定フェーズに先立って実行される学習フェーズの動作フローであり、図 13 は学習フェーズで用いられる学習データの例である。本実施形態では、教師有り機械学習の一種である条件付き確率場 (Conditional Random Field, C R F) が採用される。ただし、他の機械学習アルゴリズムが採用されてもよい。

40

【0046】

学習フェーズにおいては、まず、固有表現に関するラベルが投稿情報 P I に含まれる語句に付与される (S900)。図 13 に示される通り、以下の 5 種類のラベルが語句に付与される (B は Begin を、I は Inside を、O は Outside を、それぞれ意味する)。

- ・ B-Event : イベント名称に相当する固有表現の開始を示す。
- ・ I-Event : イベント名称に相当する固有表現の一部であることを示す。
- ・ B-Time : 開催日時に相当する固有表現の開始を示す。
- ・ I-Time : 開催日時に相当する固有表現の一部であることを示す。

50

・0：その他の要素であることを示す。

【0047】

ラベルが付与された投稿情報PIの各々について、ステップS1032と同様の素性抽出処理が実行される(S910)。その後、付与されたラベルと抽出された素性とに基づいて、イベント情報抽出部110に機械学習を行わせる(S920)。

【0048】

ステップS1034では、以上の機械学習によって構築されたモデルに基づいて、イベント情報EIに含まれるべきイベント名称が抽出される。より具体的には、イベント情報抽出部110は、1つのB-Eventラベルと1つ以上のI-Eventラベルを付与した一連の形態素を、イベント名称として抽出する。好適には、イベント情報EIに含まれるべき開催日時も同様に抽出される。

10

【0049】

イベント情報EIに含まれるべき開催スポット名称については、前述のステップS1014にて特定されたスポット名称がそのまま開催スポット名称に採用されてもよいし、機械学習(例えば、上述のCRFを用いた機械学習)によって開催スポット名称が抽出されてもよい。CRFを用いた機械学習により抽出される場合、前述の学習フェーズにおいては、開催スポット名称についてのラベル(B-Spot及びI-Spot)も語句に付与される。

【0050】

抽出されたイベント情報EIは、イベント情報テーブル170に記憶される。なお、抽出されたイベント情報EIがイベント情報統合部120に直接的に供給されてもよい。

20

【0051】

後述するイベント情報統合動作が実行されずに、ステップS1030(ステップS1032, S1034)にてイベント情報収集動作が終了してもよい。しかしながら、少なくとも従来イベント情報収集によれば、同一のイベントに対して複数のイベント名称が抽出されるケースが多い。本実施形態では、以下のイベント情報統合動作によって、以上の課題を解決する。

【0052】

1(4) - 4. イベント情報統合動作(ステップS1040)

ステップS1040において、イベント情報統合部120がイベント情報統合動作(名寄せ)を実行する。以下、図14を参照して具体的な動作を説明する。イベント情報統合部120は、イベント情報テーブル170(又はイベント情報抽出部110)から取得した複数のイベント情報EIを、開催スポット名称ごとにグループ化する(S1042)。本例では、ある1つの開催スポット名称について、[21世紀の未来展]というイベント名称(ブラケットはイベント名称の文字列に含まれない。以下同様とする)を有するイベント情報E Iaと、[21世紀のみらい展]というイベント名称を有するイベント情報E Ibとがグループ化されたと想定する。

30

【0053】

1つの開催スポット名称について複数のイベント情報EIが存在する場合、イベント情報統合部120は、これらのイベント情報EIに含まれるイベント名称のペア(スポット同一イベントペア)を全通り作成する(S1044)。本例では、[21世紀の未来展] - [21世紀のみらい展]という1つのスポット同一イベントペアが作成される。1つの開催スポット名称についてのイベント情報EIの数がn個である場合、スポット同一イベントペアの数は $n C_2$ 個であることが当然に理解される。

40

【0054】

なお、1つの開催スポット名称について1つのイベント情報EIのみが存在する場合、その開催スポット名称についてのイベント情報統合動作は終了する。

【0055】

次いで、イベント情報統合部120は、各スポット同一イベントペアについて、最長共通部分列長を取得した後に標準化を行い、標準化最長共通部分列長を算出する(S1046)。標準化共通部分列長は、スポット同一イベントペアに対応する2つのイベント名称の類

50

似度を示す指標の一例である。最長共通部分列長とは、2つの文字列に共通して表れる部分文字列のうち最長の文字列の長さを意味する。本例（[21世紀の未来展] - [21世紀のみらい展]）では、[21世紀の展]が最長共通部分列であり、その列長は「6」である。また、ステップS1046における標準化は、文字列の長さの影響を排除するための処理であり、スポット同一イベントペアのうちより長い方の文字列長で最長共通部分列長を除算する処理である。本例の標準化では、より長い[21世紀のみらい展]の文字列長「9」で最長共通部分列長「6」が除算される。したがって、本例における標準化最長共通部分列長は「0.67」（ $= 6 / 9$ 。小数点以下3桁目を四捨五入）である。

【0056】

イベント情報統合部120は、標準化最長共通部分列長に基づいてイベント情報E Iを統合すべきか否かを判定する（S1048）。標準化最長共通部分列長が閾値Th以上である場合、イベント情報統合部120は、スポット同一イベントペアに対応する2つのイベント情報E Iを1つの統合イベント情報I E Iに統合するようにイベント情報テーブル170を更新する（S1050）。一方、標準化最長共通部分列長が閾値Th未満である場合、イベント情報E Iは統合されない（S1052）。

【0057】

ステップS1048にて使用される閾値Thは任意に設定される。例えば、出願人によるテストデータを用いた試行によれば、閾値Thの好適な一例は「0.44」である。閾値Thが好適であるか否かの評価手法も任意である。

【0058】

ステップS1050において、スポット同一イベントペアのうちいずれのイベント名称を統合イベント情報I E Iのイベント名称に採用するかの基準は任意である。例えば、より短い方のイベント名称が採用されてもよいし、より多くの投稿情報P Iに含まれるイベント名称が採用されてもよい。また、投稿情報P Iがツイート（登録商標）である場合には、より多くツイートされた投稿情報P Iに含まれるイベント名称が採用されてもよい。その他、検索エンジンに入力した場合にヒット数がより多いイベント名称が採用されてもよいし、URL等の付加的情報を含む投稿情報P Iに含まれるイベント名称が採用されてもよい。

【0059】

1(5) . 実施例（実験結果）

実施形態についての1つの実施例を以下に説明する。本実施例の実験では、2013年11月に投稿された日本語ツイート（約16億9200万件）について、本実施形態のイベント情報収集動作（S1010～S1040）が実行された。SVMの機械学習（S800～S820）には2012年に投稿された2000件のツイート（イベント名称を含む正例200件、イベント名称を含まない負例1800件）が使用され、CRFの機械学習（S900～S920）には2012年に投稿されたイベント関連情報を含む254件のツイートが使用された。結果として、9781件のイベント情報E Iが収集された。なお、ステップS1040のイベント情報統合動作が実行される前のイベント情報E Iは、約13万7800件であった。

【0060】

収集された9781件のイベント情報E Iから400件を無作為に抽出して、収集されたイベント名称の精度を評価した。400件のうち、実際のイベント名称と完全に一致したイベント名称が276件（69%）、実際のイベント名称と部分的に一致したイベント名称が55件（13%）、誤検出であったイベント名称が69件（17%）であった。

【0061】

1(6) . 本実施形態の効果

以上の本実施形態の構成によれば、段階的な投稿情報P Iの選別が行われた後に、選別された投稿情報P Iからイベント情報E Iの抽出が実行される。したがって、イベント情報E Iが適切に収集される。また、同一のイベントに対応する複数のイベント情報E Iが1つのイベント情報E I（統合イベント情報I E I）に統合される。そのため、収集されたイベント情報E Iがより適切に整理される。

10

20

30

40

50

【 0 0 6 2 】

2 . 変形例

以上の実施形態は多様に变形される。具体的な变形の態様を以下に例示する。以上の実施の形態および以下の例示から任意に選択された2以上の態様は、相互に矛盾しない限り適宜に併合され得る。

【 0 0 6 3 】

2 (1) . 変形例 1

以上の実施形態では、スポット投稿選別動作 (S1010) において形態素解析による文言解析が実行される。しかし、形態素解析は必須では無く、他の手法による文言解析、例えば、Nグラムによる解析が実行されてもよい。

10

【 0 0 6 4 】

また、以上の実施形態では、文言的なマッチング処理によって投稿情報 P I が選別されるが、他の選別手法、例えば、投稿情報 P I が有する緯度経度情報 (ジオタグ) によって選別が実行されてもよい。例えば、スポット情報テーブル 1 4 0 に含まれるスポットの緯度経度から所定距離以内の緯度経度を有する投稿情報 P I が選別されてもよい。

【 0 0 6 5 】

本例におけるイベント情報 E I は、イベントの名称及びそのイベントが開催される場所の緯度・経度を少なくとも含む。「イベントが開催される場所の緯度・経度」は任意に取得される。例えば、イベントに関連すると推定された投稿情報 P I が投稿された位置の緯度経度情報 (ジオタグ) がイベント情報 E I に含まれる。

20

【 0 0 6 6 】

以上から理解されるように、ステップ S1010 においては、投稿情報 P I に含まれる地理的情報 (スポット名称、ジオタグ等) に基づいて、スポットに関連する投稿情報 P I (スポット名称を含む投稿情報 P I、スポットの近傍で投稿された投稿情報 P I 等) が選別される。

【 0 0 6 7 】

2 (2) . 変形例 2

以上の実施形態では、イベント投稿選別動作 (S1020) における機械学習アルゴリズムとして SVM が採用されるが、他の機械学習アルゴリズムが採用されてもよい。また、機械学習以外の手法によってイベントに関連する投稿情報 P I が選別されてもよい。例えば、「開催」や「参加」等のイベントに関連する単語を有するか否かに基づいて投稿情報 P I が選別 (フィルタリング) されてもよい。

30

【 0 0 6 8 】

2 (3) . 変形例 3

以上の実施形態では、スポット投稿選別動作の後にイベント投稿選別動作が実行されるが、逆順に実行されてもよい (すなわち、S1020 の後に S1010 が実行されてもよい) 。また、両動作が並列的に実行されてもよい。ただし、実施形態のように、ステップ S1010 にてマッチング等の比較的単純な処理を実行し、ステップ S1020 にて SVM 等の比較的複雑な処理を実行する場合には、ステップ S1010 にて投稿情報 P I を選別した上でステップ S1020 を実行することにより、ステップ S1020 における処理負荷をより低減することができる。

40

【 0 0 6 9 】

2 (4) . 変形例 4

イベント情報抽出動作 (ステップ S1030) において、イベント情報抽出部 1 1 0 が、NG ユーザリストに登録されている NG ユーザによって投稿された投稿情報 P I を、イベント情報 E I を抽出する対象から除外してもよい。NG ユーザリストには任意のユーザが登録され得る。例えば、ログサーバ B S にて凍結された (サスペンドされた) アカントに対応するユーザが、NG ユーザとして登録されてもよい。

【 0 0 7 0 】

2 (5) . 変形例 5

以上の実施形態では、文字列の類似度を示す標準化共通部分列長に基づいてイベント情

50

報統合動作（S1040）が実行されるが、類似度を示す他の指標（例えば、編集距離）に基づいてイベント情報 E I の統合が実行されてもよい。例えば、スポット同一イベントペアに対応する 2 つのイベント名称の編集距離が閾値 $T_h 2$ 以下である場合に、イベント情報 E I が統合されてもよい。編集距離は、ある文字列を他の文字列に変形するのに必要な編集操作（挿入、削除、及び置換）の回数を意味する。以上の変形例において、標準化された編集距離が用いられてもよい。標準化の手法は任意であるが、例えば、実施形態と同様に、スポット同一イベントペアのうちより長い方の文字列長で最長共通部分列長を除算する手法が採用され得る。

【 0 0 7 1 】

また、イベント情報 E I に含まれる緯度・経度に基づいてイベント情報統合動作（S1040）が実行されてもよい。例えば、ステップ S1042 において、イベント情報統合部 1 2 0 が、複数のイベント情報 E I を、緯度及び経度に基づいて分割された網目状の領域（メッシュ）ごとにグループ化してもよい。以上のグループ化は、緯度経度によって投稿情報 P I を選別する変形例 1 の構成（投稿情報 P I がスポット名称を含まない可能性がある構成）において、特に好適である。なお、以上のメッシュの具体例としては、日本国の総務省統計局が提供する標準地域メッシュ（例えば、4 分の 1 地域メッシュ）が挙げられる。

10

【 0 0 7 2 】

2 (6) . 変形例 6

イベント情報抽出部 1 1 0 が、抽出されたイベント情報 E I のうち、不要なイベント情報 E I を排除してもよい。例えば、イベント情報抽出部 1 1 0 が、排除リストに登録された単語（例えば、NG ワード等）を含むイベント情報 E I を排除してもよい。

20

【 0 0 7 3 】

また、スポット情報テーブル 1 4 0 に含まれるスポット名称が以上の排除リストに含まれてもよい。スポットの名称自体は、イベント名称では無いと考えられる。したがって、イベント情報抽出部 1 1 0 が、排除リストに登録されたスポット名称に一致するイベント名称を含むイベント情報 E I を排除すると好適である。

【 0 0 7 4 】

2 (7) . 変形例 7

イベント情報抽出部 1 1 0 が、抽出されたイベント名称を整形してもよい。例えば、カギ括弧等の区切り文字（デリミタ）がイベント名称に含まれる場合、イベント情報抽出部 1 1 0 は、その区切り文字をイベント名称から削除すると好適である。区切り文字自体はイベント名称に含まれない可能性が高いからである。

30

【 0 0 7 5 】

2 (8) . 変形例 8

1 つの投稿情報 P I に、複数のイベント名称と複数の開催スポット名称とが含まれる場合がある。例えば、投稿情報 P I が『 [イベント名称 1] は [開催スポット名称 1] で開催されます。 [イベント名称 2] は [開催スポット名称 2] で開催されます。』という本文を含むような場合である。以上の場合、ステップ S1030 において、間違っただ対応付けがなされたイベント情報 E I が抽出される場合がある。例えば、 [イベント名称 1] と [開催スポット名称 2] とが対応付けられる場合である。

40

【 0 0 7 6 】

以上のような間違っただ対応付けを抑制するため、イベント情報抽出部 1 1 0 が、イベント情報 E I を抽出する際に、イベント名称と開催スポット名称との距離（間に存在する文字数）が最も小さくなるように対応付けを行うと好適である。また、単純な文字数ではなく、文字の種別ごとに重み付け（例えば、句読点の重みを他の文字の重みよりも重くする）をした重み付け距離に基づいてイベント名称と開催スポット名称との対応付けが実行されてもよい。

【 0 0 7 7 】

2 (9) . 変形例 9

イベント情報抽出動作において、イベント名称、開催スポット名称、及び開催日時以外

50

の様々な情報が抽出されてもよい。例えば、イベント情報抽出部 110 が投稿情報 P I を解析し、その投稿情報 P I に含まれる外部リンク情報 (U R L 等) を抽出してイベント情報 E I に含めてもよい。また、イベント情報 E I が投稿情報 P I を解析し、その投稿情報 P I に含まれる特徴語を抽出してイベント情報 E I に含めてもよい。特徴語を抽出するための手法は任意である。例えば、イベント情報抽出部 110 が、投稿情報 P I に含まれる各語句について tf-idf (term frequency - inverse document frequency) を算定し、tf-idf の値が最も大きい語を特徴語として抽出してもよい。また、tf-idf の値が所定の閾値より高い 1 以上の語を特徴語として抽出してもよい。

【 0 0 7 8 】

2 (1 0) . 変形例 1 0

以上の実施形態では、1つの装置である情報管理サーバ M S が、イベント情報収集動作を実行する。しかしながら、複数の装置によって構成される情報管理システムが、イベント情報収集動作を実行してもよい。例えば、スポット投稿選別部 102、イベント投稿選別部 104、イベント情報抽出部 110、イベント情報統合部 120、及び情報提示部 130 が 2 以上のサーバ装置を含む情報管理システムによって実装されてもよい。

【 0 0 7 9 】

投稿情報取得サーバ A S と情報管理サーバ M S とが一体である構成も採用可能である。

【 0 0 8 0 】

2 (1 1) . 変形例 1 1

コンピュータ装置、特に、情報管理サーバ M S 及び投稿情報取得サーバ A S において C P U が実行する各機能は、F P G A (Field Programmable Gate Array) または D S P (Digital Signal Processor) 等のプログラマブルロジックデバイスで実行されてもよい。

【 符号の説明 】

【 0 0 8 1 】

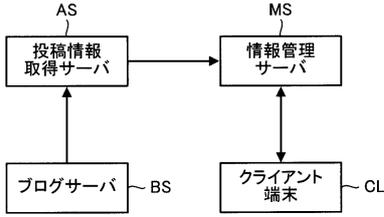
A S …… 投稿情報取得サーバ、 B S …… ブログサーバ、 C L …… クライアント端末、 E I …… イベント情報、 I E I …… 統合イベント情報、 M S …… 情報管理サーバ、 P I …… 投稿情報、 1 0 0 …… 投稿情報選別部、 1 0 2 …… スポット投稿選別部、 1 0 4 …… イベント投稿選別部、 1 1 0 …… イベント情報抽出部、 1 2 0 …… イベント情報統合部、 1 3 0 …… 情報提示部、 1 4 0 …… スポット情報テーブル、 1 5 0 …… スポット投稿テーブル、 1 6 0 …… イベント投稿テーブル、 1 7 0 …… イベント情報テーブル、 2 0 0 …… 投稿情報処理部、 2 1 0 …… 投稿情報テーブル。

10

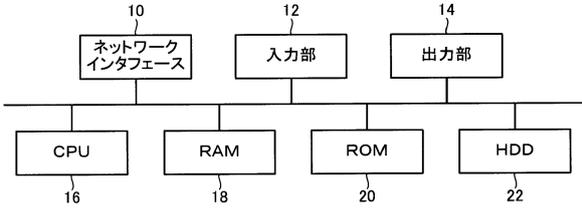
20

30

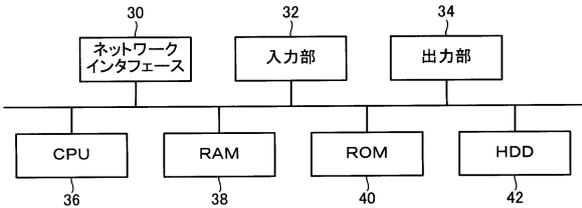
【図1】



【図2】



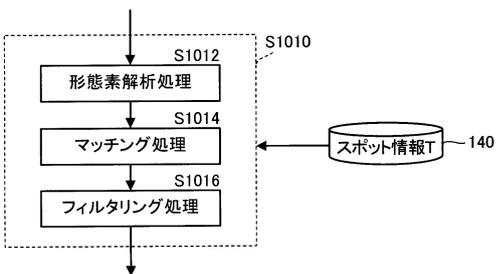
【図3】



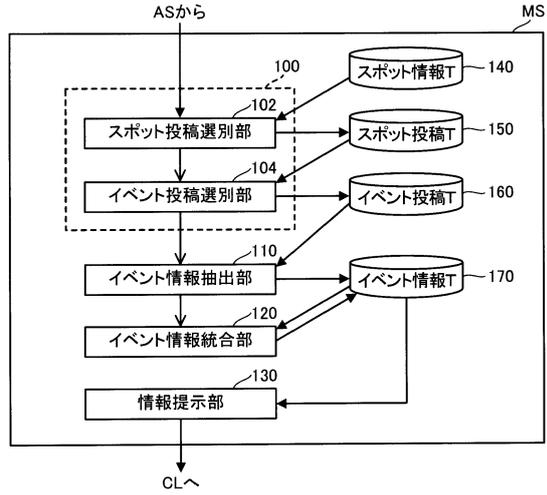
【図6】



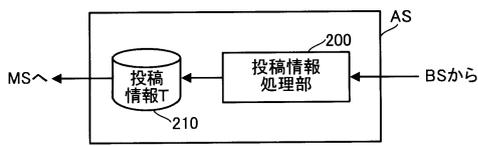
【図7】



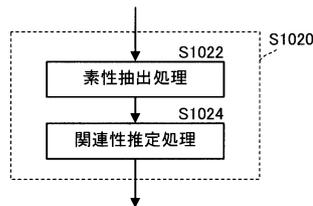
【図4】



【図5】



【図8】



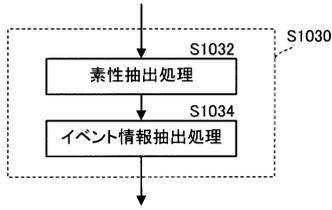
【図9】



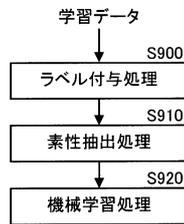
【図10】

ラベル	投稿情報の本文
関連有り	〇〇未来館で「21世紀の未来展」を12月3日から開催します。
関連有り	〇〇未来館の21世紀の未来展に参加したいなあ。
関連無し	〇〇未来館はバリアフリー化活動を推進しています。
関連無し	××で18時から飲み会を開催します。
⋮	⋮

【図 1 1】



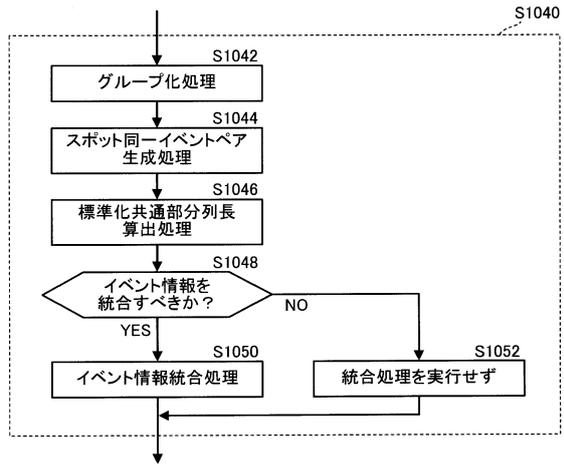
【図 1 2】



【図 1 3】

表記	品詞	ラベル
未来	名詞	B-EVENT
展	名詞接尾辞	I-EVENT
は	格助詞:連用	O
21	名詞:日時	B-TIME
日	名詞:日時	I-TIME
から	格助詞:連用	O
⋮	⋮	⋮

【図 1 4】



フロントページの続き

- (72)発明者 太田 賢
東京都千代田区永田町二丁目11番1号 株式会社NTTドコモ内
- (72)発明者 山田 渉
東京都千代田区永田町二丁目11番1号 株式会社NTTドコモ内

審査官 松尾 真人

- (56)参考文献 特開2014-010513(JP,A)
特開2013-168021(JP,A)
特開2014-049052(JP,A)
特開2014-006737(JP,A)
吉田 将人, 外2名, “ ブログ記事とWebページを用いたイベント情報抽出手法の提案 ”, 情報処理学会研究報告, 社団法人情報処理学会, 2009年 3月18日, 第2009巻, 第35号, p. 37-44
中澤 昌美, 外3名, “ Twitterからのイベント検出および関連情報収集の高精度化 ”, FIT2012 第11回情報科学技術フォーラム 講演論文集, 一般社団法人電子情報通信学会, 一般社団法人情報処理学会, 2012年 8月21日, 第2分冊, p. 93-94

(58)調査した分野(Int.Cl., DB名)

G06F 13/00
G06F 17/30
G06F 19/00
G06Q 10/00-99/00