

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5950737号
(P5950737)

(45) 発行日 平成28年7月13日 (2016. 7. 13)

(24) 登録日 平成28年6月17日 (2016. 6. 17)

(51) Int. Cl.	F 1
G 0 6 F 17/30 (2006.01)	G 0 6 F 17/30 2 1 0 A
	G 0 6 F 17/30 1 7 0 D

請求項の数 5 (全 11 頁)

(21) 出願番号	特願2012-159901 (P2012-159901)	(73) 特許権者	000004352
(22) 出願日	平成24年7月18日 (2012. 7. 18)		日本放送協会
(65) 公開番号	特開2014-21727 (P2014-21727A)		東京都渋谷区神南2丁目2番1号
(43) 公開日	平成26年2月3日 (2014. 2. 3)	(74) 代理人	100106002
審査請求日	平成27年6月1日 (2015. 6. 1)		弁理士 正林 真之
特許権者において、実施許諾の用意がある。		(74) 代理人	100120891
			弁理士 林 一好
		(72) 発明者	宮崎 勝
			東京都世田谷区砧一丁目10番11号 日
			本放送協会放送技術研究所内
		(72) 発明者	山田 一郎
			東京都世田谷区砧一丁目10番11号 日
			本放送協会放送技術研究所内
		審査官	吉田 誠
		最終頁に続く	

(54) 【発明の名称】 情報抽出装置及びプログラム

(57) 【特許請求の範囲】

【請求項 1】

番組についてのコメントを抽出する第1抽出部と、
前記第1抽出部により抽出されたコメントを解析する第1解析部と、
前記第1解析部で解析された第1解析結果に基づいて、番組を特定する番組特定部と、
前記番組特定部により特定された番組に関する番組情報をデータベースから抽出する第2抽出部と、
前記第2抽出部により抽出された番組情報を解析する第2解析部と、
前記第1解析結果と、前記第2解析部で解析された第2解析結果を比較する比較部と、
前記比較部による比較結果に基づいて、前記第2解析結果に含まれていなかった情報であって、前記第1解析結果に含まれている情報を前記番組特定部で特定した番組の番組情報に追加するか否かを判断する判断部と、
前記判断部で情報を追加すると判断された場合、当該情報を前記番組特定部で特定した番組の番組情報に追加する追加部を備える情報抽出装置。

【請求項 2】

前記第1解析部は、形態素解析により、前記第1抽出部により抽出されたコメントに含まれている全ての単語又は特定の品詞を対象として、特徴的なキーワードを前記第1解析結果として抽出し、

前記第2解析部は、形態素解析により、前記第2抽出部により抽出された番組情報に含まれている全ての単語又は特定の品詞を対象として、特徴的なキーワードを前記第2解析

10

20

結果として抽出する請求項 1 記載の情報抽出装置。

【請求項 3】

前記第 1 解析部は、形態素解析により、前記第 1 抽出部により抽出されたコメントに含まれている語の中から、TF-IDF 値が高い語のみ、又は話題性のある語を前記第 1 解析結果として抽出し、

前記第 2 解析部は、形態素解析により、前記第 2 抽出部により抽出された番組情報に含まれている語の中から、TF-IDF 値が高い語のみ、又は話題性のある語を前記第 2 解析結果として抽出する請求項 1 記載の情報抽出装置。

【請求項 4】

前記第 1 解析部は、形態素解析により、前記第 1 抽出部により抽出されたコメントに含まれている固有表現を前記第 1 解析結果として抽出し、

前記第 2 解析部は、形態素解析により、前記第 2 抽出部により抽出された番組情報に含まれている固有表現を前記第 2 解析結果として抽出する請求項 1 記載の情報抽出装置。

【請求項 5】

番組についてのコメントを抽出する第 1 抽出工程と、

前記第 1 抽出工程により抽出されたコメントを解析する第 1 解析工程と、

前記第 1 解析工程で解析された第 1 解析結果に基づいて、番組を特定する番組特定工程と、

前記番組特定工程により特定された番組に関する番組情報をデータベースから抽出する第 2 抽出工程と、

前記第 2 抽出工程により抽出された番組情報を解析する第 2 解析工程と、

前記第 1 解析結果と、前記第 2 解析工程で解析された第 2 解析結果を比較する比較工程と、

前記比較工程による比較結果に基づいて、前記第 2 解析結果に含まれていなかった情報であって、前記第 1 解析結果に含まれている情報を前記番組特定工程で特定した番組の番組情報に追加するか否かを判断する判断工程と、

前記判断工程で情報を追加すると判断された場合、当該情報を前記番組特定工程で特定した番組の番組情報に追加する追加工程をコンピュータに実行させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、コンテンツに関する新たな情報を抽出する情報抽出装置及びプログラムに関する。

【背景技術】

【0002】

近年、番組、映画といった映像コンテンツや、Web サイト、ブログ等のテキストコンテンツがインターネット上で自由に利用及び閲覧できる環境が整えられている。また、ユーザ自身がそれらのコンテンツを検索するサービスの重要性が高まっている。

【0003】

検索エンジン、インターネットのポータルサイト、又はコンテンツプロバイダーのサイト等において、映像コンテンツ及びテキストコンテンツの中から目的のコンテンツをキーワード検索するための機能が提供されている。しかし、主にキーワードのマッチングによるものが多いため、事前に用意した、コンテンツに関するキーワード情報（メタデータ）が不十分であると、目的のコンテンツに到達できないことも考えられる。

【0004】

また、キーワード情報は、コンテンツプロバイダーが用意する場合が多く、大多数のユーザが共通して想像できるような一般的なものになりがちである。よって、様々な観点から多様なキーワード情報がコンテンツに付与されることはあまりない。

【0005】

そこで、ユーザ自身がコンテンツに関する情報を付与し、それを他のユーザが利用でき

10

20

30

40

50

るサービス等が登場している。これらは「ソーシャル・タギング」と呼ばれる。

例えば、特許文献1では、ソーシャル・ブックマークという、ウェブページにユーザがタグ（キーワード）を付加することでウェブページを分類できるサービスを利用し、各ウェブサイトに自動的に検索のためのキーワードを付与する技術が提案されている。

【0006】

また、特許文献2では、コミュニケーションの場に投稿されたメッセージからデータベースに未登録な用語を検出し、その用語の意味を問う質問メッセージをユーザに送信する技術が提案されている。これは、質問に対するユーザの回答メッセージに基づいて用語の意味をデータベースに登録し、そのコミュニケーションの場に参加するユーザの知識をサポートする技術に関するものである。

10

【先行技術文献】

【特許文献】

【0007】

【特許文献1】特開2009-140089号公報

【特許文献2】特開2007-156833号公報

【発明の概要】

【発明が解決しようとする課題】

【0008】

ところで、特許文献1の技術は、ソーシャル・ブックマークのデータを基本としているため、ユーザが明示的に付与したキーワードの情報しか検索に利用できない。また、特許文献2の技術は、コンテンツの検索に関するものではなく、また、質問文を生成してコミュニケーションの場に関する情報を得る手法をとっているため、質問文に関するユーザからの回答が得られることが前提となってしまふ。つまり、ユーザがタギングや、質問への回答、という明示的な操作を行わなくても、コンテンツに関する情報が蓄積されていくことが望ましい。

20

【0009】

本発明は、ユーザが明示的に指定したキーワードだけではなく、ユーザが重要とは判断してないキーワードを抽出することができる情報抽出装置及びプログラムを提供することを目的とする。

【課題を解決するための手段】

30

【0010】

本発明に係る情報抽出装置は、番組についてのコメントを抽出する第1抽出部と、前記第1抽出部により抽出されたコメントを解析する第1解析部と、前記第1解析部で解析された第1解析結果に基づいて、番組を特定する番組特定部と、前記番組特定部により特定された番組に関する番組情報をデータベースから抽出する第2抽出部と、前記第2抽出部により抽出された番組情報を解析する第2解析部と、前記第1解析結果と、前記第2解析部で解析された第2解析結果を比較する比較部と、前記比較部による比較結果に基づいて、前記第2解析結果に含まれていなかった情報であって、前記第1解析結果に含まれている情報を前記番組特定部で特定した番組の番組情報に追加するか否かを判断する判断部と、前記判断部で情報を追加すると判断された場合、当該情報を前記番組特定部で特定した番組の番組情報に追加する追加部を備える構成である。

40

【0011】

かかる構成によれば、本発明に係る情報抽出装置は、番組についてのコメントを解析した第1解析結果と、番組情報を解析した第2解析結果を比較し、第2解析結果に含まれていなかった情報であって、第1解析結果に含まれている情報を番組情報に追加するか否かを判断し、判断結果に応じて、番組情報に新たに情報を追加するので、ユーザが明示的に指定したキーワード（情報）だけではなく、ユーザが「重要とは判断してない」キーワードを抽出し、番組情報に追加することができる。

【0012】

また、情報抽出装置では、前記第1解析部は、形態素解析により、前記第1抽出部によ

50

り抽出されたコメントに含まれている全ての単語又は特定の品詞を対象として、特徴的なキーワードを前記第1解析結果として抽出し、前記第2解析部は、形態素解析により、前記第2抽出部により抽出された番組情報に含まれている全ての単語又は特定の品詞を対象として、特徴的なキーワードを前記第2解析結果として抽出する構成でもよい。

【0013】

かかる構成によれば、本発明に係る情報抽出装置は、番組についてのコメントと、番組情報に含まれている全ての単語又は特定の品詞を対象として特徴的なキーワードをそれぞれ抽出して比較する。よって、情報抽出装置は、全単語を対象にする場合には、第2解析結果に含まれていなかった情報であって、第1解析結果に含まれている情報を漏れなく抽出でき、また、特定の品詞（例えば、名詞）を対象にする場合には、特定の品詞についてのみ、第2解析結果に含まれていなかった情報であって、第1解析結果に含まれている情報を漏れなく抽出できる。

10

【0014】

また、情報抽出装置では、前記第1解析部は、形態素解析により、前記第1抽出部により抽出されたコメントに含まれている語の中から、TF-IDF値が高い語のみ、又は話題性のある語を前記第1解析結果として抽出し、前記第2解析部は、形態素解析により、前記第2抽出部により抽出された番組情報に含まれている語の中から、TF-IDF値が高い語のみ、又は話題性のある語を前記第2解析結果として抽出する構成でもよい。

【0015】

かかる構成によれば、本発明に係る情報抽出装置は、番組についてのコメントと、番組情報に含まれている語の中から、TF-IDF値が高い語のみ、又は話題性のある語をそれぞれ抽出して比較する。よって、情報抽出装置は、TF-IDF値が高い語のみを対象にする場合には、第2解析結果に含まれていなかったTF-IDF値が高い語であって、第1解析結果に含まれているTF-IDF値が高い語を漏れなく抽出でき、また、話題性のある語を対象にする場合には、第2解析結果に含まれていなかった話題性のある語であって、第1解析結果に含まれている話題性のある語を漏れなく抽出できる。

20

【0016】

また、情報抽出装置では、前記第1解析部は、形態素解析により、前記第1抽出部により抽出されたコメントに含まれている固有表現を前記第1解析結果として抽出し、前記第2解析部は、形態素解析により、前記第2抽出部により抽出された番組情報に含まれている固有表現を前記第2解析結果として抽出する構成でもよい。

30

【0017】

かかる構成によれば、本発明に係る情報抽出装置は、番組についてのコメントと、番組情報に含まれている語の中から、固有表現のみをそれぞれ抽出して比較する。よって、情報抽出装置は、第2解析結果に含まれていなかった固有表現であって、第1解析結果に含まれている固有表現を漏れなく抽出できる。

【0018】

本発明に係るプログラムは、番組についてのコメントを抽出する第1抽出工程と、前記第1抽出工程により抽出されたコメントを解析する第1解析工程と、前記第1解析工程で解析された第1解析結果に基づいて、番組を特定する番組特定工程と、前記番組特定工程により特定された番組に関する番組情報をデータベースから抽出する第2抽出工程と、前記第2抽出工程により抽出された番組情報を解析する第2解析工程と、前記第1解析結果と、前記第2解析工程で解析された第2解析結果を比較する比較工程と、前記比較工程による比較結果に基づいて、前記第2解析結果に含まれていなかった情報であって、前記第1解析結果に含まれている情報を前記番組特定工程で特定した番組の番組情報に追加するか否かを判断する判断工程と、前記判断工程で情報を追加すると判断された場合、当該情報を前記番組特定工程で特定した番組の番組情報に追加する追加工程をコンピュータに実行させるためのものである。

40

【0019】

かかる構成によれば、本発明に係るプログラムは、番組についてのコメントを解析した

50

第1解析結果と、番組情報を解析した第2解析結果を比較し、第2解析結果に含まれていなかった情報であって、第1解析結果に含まれている情報を番組情報に追加するか否かを判断し、判断結果に応じて、番組情報に新たに情報を追加するので、ユーザが明示的に指定したキーワード（情報）だけではなく、ユーザが「重要とは判断してない」キーワードを抽出し、番組情報に追加することができる。

【発明の効果】

【0020】

本発明によれば、ユーザが明示的に指定したキーワードだけではなく、ユーザが重要とは判断してないキーワードを抽出することができる。

【図面の簡単な説明】

10

【0021】

【図1】情報抽出装置の構成を示す図である。

【図2】情報抽出装置の動作の流れについての説明に供するフローチャートである。

【発明を実施するための形態】

【0022】

以下、本発明の実施形態の一例について図1を参照しながら説明する。

情報抽出装置1は、図1に示すように、第1抽出部11と、第1解析部12と、番組特定部13と、第2抽出部14と、第2解析部15と、比較部16と、判断部17と、追加部18を備える。

【0023】

20

第1抽出部11は、番組についてのコメントを投稿型のウェブサイト2から抽出する。投稿型のウェブサイト2は、例えば、コメントを投稿して他のユーザとコミュニケーションを図るサービス（ソーシャルネットワークサービス、SNS）により実現される。また、投稿型のウェブサイト2は、SNSに限らず、掲示板（BBS）のような形態で実現されてもよい。

【0024】

また、投稿型のウェブサイト2は、各番組に対するレビューページを備える構成であってもよい。各ページは、ユーザが番組に対して自由に意見及び感想を書き込むことができる。このような構成によれば、第1抽出部11は、投稿型のウェブサイト2に含まれている各番組に対するレビューページにアクセスして、投稿されているコメントを抽出する。

30

なお、第1抽出部11は、投稿型のウェブサイト2に投稿されているコメントに限らず、番組についてのコメントを抽出できればよい。

【0025】

例えば、情報抽出装置1とPCが有線又は無線で接続されているものとする。また、PCは、番組についてのコメントの書き込み及び閲覧が可能なVODアプリケーションを起動しているものとする。第1抽出部11は、VODアプリケーションから番組についてのコメントを抽出する構成でもよい。

【0026】

また、情報抽出装置1とスマートフォン等の携帯無線機器が有線又は無線で接続されているものとする。また、携帯無線機器は、番組についてのコメントの書き込み及び閲覧が可能な番組レビューアプリケーションを起動しているものとする。第1抽出部11は、番組レビューアプリケーションから番組についてのコメントを抽出する構成でもよい。

40

【0027】

また、情報抽出装置1とスマートテレビが有線又は無線で接続されているものとする。また、スマートテレビは、例えば、ハイブリッドキャスト（登録商標）対応テレビであって、番組についてのコメントの書き込み及び閲覧が可能な番組レビューアプリケーションを起動しているものとする。第1抽出部11は、番組レビューアプリケーションから番組についてのコメントを抽出する構成でもよい。

【0028】

第1解析部12は、第1抽出部11により抽出されたコメントを解析する。

50

番組特定部 13 は、第 1 解析部 12 で解析された第 1 解析結果に基づいて、番組を特定する。

【0029】

第 2 抽出部 14 は、番組特定部 13 により特定された番組に関する番組情報をデータベース 3 から抽出する。データベース 3 は、例えば、放送事業者により提供されるものであり、番組ごとに番組に関する情報が番組情報として格納されている。番組情報は、例えば、番組のタイトル、出演者、番組概要等により構成されている。

【0030】

第 2 解析部 15 は、第 2 抽出部 14 により抽出された番組情報を解析する。

比較部 16 は、第 1 解析結果と、第 2 解析部 15 で解析された第 2 解析結果を比較する。
10

【0031】

判断部 17 は、比較部 16 による比較結果に基づいて、第 2 解析結果に含まれていなかった情報であって、第 1 解析結果に含まれている情報を番組特定部 13 で特定した番組の番組情報に追加するか否かを判断する。

追加部 18 は、判断部 17 で情報を追加すると判断された場合、当該情報を番組特定部 13 で特定した番組の番組情報に追加する。

【0032】

このように構成されることにより、情報抽出装置 1 は、投稿型のウェブサイト 2 に投稿されているコメントを解析した第 1 解析結果と、番組情報を解析した第 2 解析結果を比較し、第 2 解析結果に含まれていなかった情報であって、第 1 解析結果に含まれている情報を番組情報に追加するか否かを判断し、判断結果に応じて、番組情報に新たに情報を追加する。
20

【0033】

よって、情報抽出装置 1 は、ユーザによるキーワードのタギングや、ユーザとシステムのインタラクションによってコンテンツ検索のためのキーワードを付加するのではなく、コンテンツに関するユーザの意見、感想といったテキスト情報、又は、コンテンツに関してユーザ同士で行われたテキストコミュニケーションの情報からコンテンツに関する重要なキーワードを抽出し、それをそのコンテンツのメタデータとして蓄積するものである。これにより、情報抽出装置 1 は、ユーザが明示的に指定したキーワード（情報）だけではなく、ユーザが「重要とは判断してない」キーワードを抽出し、番組情報に追加することができる。
30

また、情報抽出装置 1 により抽出したキーワードを利用して、他のユーザにとって意外なコンテンツの検索や推薦に役立てることができる。

【0034】

また、第 1 解析部 12 は、形態素解析により、第 1 抽出部 11 により抽出されたコメントに含まれている全ての単語又は特定の品詞を対象として、特徴的なキーワードを第 1 解析結果として抽出する構成でもよい。第 2 解析部 15 は、形態素解析により、第 2 抽出部 14 により抽出された番組情報に含まれている全ての単語又は特定の品詞を対象として、特徴的なキーワードを第 2 解析結果として抽出する構成でもよい。
40

【0035】

かかる構成によれば、情報抽出装置 1 は、投稿型のウェブサイト 2 に投稿されているコメントと、番組情報に含まれている全ての単語又は特定の品詞を対象として特徴的なキーワードをそれぞれ抽出して比較する。

【0036】

よって、情報抽出装置 1 は、全単語を対象にする場合には、第 2 解析結果に含まれていなかった情報であって、第 1 解析結果に含まれている情報を漏れなく抽出でき、また、特定の品詞（例えば、名詞）を対象にする場合には、特定の品詞についてのみ、第 2 解析結果に含まれていなかった情報であって、第 1 解析結果に含まれている情報を漏れなく抽出できる。
50

【0037】

また、第1解析部12は、形態素解析により、第1抽出部11により抽出されたコメントに含まれている語の中から、TF-IDF値が高い語のみ、又は話題性のある語を第1解析結果として抽出する構成でもよい。第2解析部15は、形態素解析により、第2抽出部14により抽出された番組情報に含まれている語の中から、TF-IDF値が高い語のみ、又は話題性のある語を第2解析結果として抽出する構成でもよい。

【0038】

かかる構成によれば、情報抽出装置1は、投稿型のウェブサイト2に投稿されているコメントと、番組情報に含まれている語の中から、TF-IDF値が高い語のみ、又は話題性のある語をそれぞれ抽出して比較する。

10

【0039】

よって、情報抽出装置1は、TF-IDF値が高い語のみを対象にする場合には、第2解析結果に含まれていなかったTF-IDF値が高い語であって、第1解析結果に含まれているTF-IDF値が高い語を漏れなく抽出でき、また、話題性のある語を対象にする場合には、第2解析結果に含まれていなかった話題性のある語であって、第1解析結果に含まれている話題性のある語を漏れなく抽出できる。

【0040】

また、第1解析部12は、形態素解析により、第1抽出部11により抽出されたコメントに含まれている固有表現を第1解析結果として抽出する構成でもよい。第2解析部15は、形態素解析により、第2抽出部14により抽出された番組情報に含まれている固有表現を第2解析結果として抽出する構成でもよい。

20

【0041】

かかる構成によれば、情報抽出装置1は、投稿型のウェブサイト2に投稿されているコメントと、番組情報に含まれている語の中から、固有表現のみをそれぞれ抽出して比較する。よって、情報抽出装置1は、第2解析結果に含まれていなかった固有表現であって、第1解析結果に含まれている固有表現を漏れなく抽出できる。

【0042】

つぎに、情報抽出装置1の動作の流れについて、図2に示すフローチャートを参照しながら説明する。

ステップST1において、第1抽出部11は、投稿型のウェブサイト2に投稿されているコメントを抽出する。

30

【0043】

ステップST2において、第1解析部12は、ステップST1の工程により抽出されたコメントを解析する。

ステップST3において、番組特定部13は、ステップST2の工程により解析された第1解析結果に基づいて、番組を特定する。

【0044】

ステップST4において、第2抽出部14は、ステップST3の工程により特定された番組に関する番組情報をデータベース3から抽出する。

ステップST5において、第2解析部15は、ステップST4の工程により抽出された番組情報を解析する。

40

【0045】

ステップST6において、比較部16は、ステップST2の工程で解析された第1解析結果と、ステップST5の工程により解析された第2解析結果を比較する。

ステップST7において、判断部17は、ステップST6の工程による比較結果に基づいて、第2解析結果に含まれていなかった情報であって、第1解析結果に含まれている情報を番組特定工程で特定した番組の番組情報に追加するか否かを判断する。追加すると判断した場合には、ステップST8に進み、追加しないと判断した場合には、一連の処理を終了する。

【0046】

50

ステップS T 8において、追加部1 8は、ステップS T 7の工程で情報を追加すると判断された場合、当該情報をステップS T 3の工程で特定した番組の番組情報に追加する。

【0047】

このようにして、情報抽出装置1は、投稿型のウェブサイト2に投稿されているコメントを解析した第1解析結果と、番組情報を解析した第2解析結果を比較し、第2解析結果に含まれていなかった情報であって、第1解析結果に含まれている情報を番組情報に追加するか否かを判断し、判断結果に応じて、番組情報に新たに情報を追加するので、ユーザが明示的に指定したキーワード（情報）だけではなく、ユーザが重要とは判断していないキーワードを抽出し、番組情報に追加することができる。

【0048】

つぎに、情報抽出装置1の具体的な実施例について説明する。

上述したように、情報抽出装置1は、番組ごとにレビュー投稿が可能なSNS（投稿型のウェブサイト2）を利用するものであり、ユーザの番組に対する自由な書き込みから、番組を検索するためのキーワードを自動抽出して、登録するものである。

【0049】

情報抽出装置1は、ユーザの書き込みテキストを解析し、特徴的なキーワードを抽出する。このテキスト解析には、一般的な形態素解析処理、重要度計算処理、又は固有表現抽出処理を用いることができる。

【0050】

重要度計算処理には、TF-IDFといった指標や、トピック抽出の技術等を用いてもよい。また、重要度の計算をせずに、形態素解析の結果得られた単語群をそのまま利用してもよい。

【0051】

例えば、投稿型のウェブサイト2において、番組PR1に対するユーザU1の書き込みP1が「この番組を見て、〇〇（映画名）という映画のことを思い出しました。大きなアンテナが・・・」というものだったとする。また、番組PR1のメタデータM1は、「タイトル：〇〇」と、「出演者：〇〇、△△」と、「番組概要：アメリカにある天文台では、巨大なアンテナ太陽系外からの電波を・・・」が含まれているものとする。

【0052】

情報抽出装置1は、投稿型のウェブサイト2にアクセスして、ユーザU1の書き込みP1を形態素解析して単語に分割する。また、情報抽出装置1は、データベース3にアクセスして、番組PR1に関するメタデータM1（番組情報）を取得し、解析する。

【0053】

情報抽出装置1は、書き込みP1を解析して得られた単語リストWL1と、メタデータM1を解析して得られた単語リストWL2とを比較し、WL1に含まれているがWL2に含まれていない単語を抽出する。

【0054】

具体的には、単語リストWL1には、「〇〇〇（映画タイトル）」、「アンテナ」が含まれている。また、単語リストWL2には、「〇〇」（タイトル）、「〇〇」（出演者）、「△△」（出演者）、「アメリカ」、「天文台」、「太陽系外」、「電波」、「アンテナ」が含まれている。

【0055】

情報抽出装置1は、WL1に含まれる「〇〇〇（映画タイトル）」という単語W1が単語リストWL2に含まれていないため、新たに単語W1を番組PR1のメタデータM1に追加する処理を行う。例えば、情報抽出装置1は、オリジナルのメタデータM1には直接追加せず、追加メタデータM2という形式で新たな単語W1を追加する。

【0056】

つまり、番組PR1のメタデータM1を作成した業者は、番組PR1と単語W1（映画タイトル）に関連があることを想定しておらず、そのような情報をメタデータに関連付けていなかった。

10

20

30

40

50

【0057】

情報抽出装置1は、上述したように、あるユーザU1がこの番組を見て投稿した書き込みP1から番組PR1と単語W1（映画タイトル）との関係を抽出し、単語W1を番組PR1の新たなメタデータとして蓄積する。

このようにして、新たなメタデータをデータベース3に随時蓄積しておくことにより、例えば、ある別のユーザU2が「○○○（映画名）」というキーワードで番組を検索した場合に、番組PR1を「関連番組」として提供することが可能となり、ユーザU2に対して意外性のある番組提供を実現することができる。

【0058】

また、書き込みP1から単語リストWL1を作成する処理には、様々な手法を利用することができる。例えば、形態素解析を用いて書き込みP1から単語群WL0を生成し、生成した全ての単語群WL0を単語リストWL1に含ませることができる。

【0059】

また、TF-IDF等の手法を用いて単語群WL0から重要な単語だけを抽出し、それを単語リストWL1に含ませてもよい。

【0060】

また、投稿群の時間的推移から現在話題になっている単語のみを取り出し、それを単語リストWL1に含ませてもよい。

さらに、単語群WL0から、人名、地名、番組タイトル、といった固有名詞（固有表現）のみを抽出し、それを単語リストWL1に含ませてもよい。

【0061】

また、メタデータM1から単語リストWL2を作成する手順も同様である。メタデータM1に含まれる全てを単語リストWL2としてもよいし、メタデータM1の特定の属性（例えば、「番組名」、「出演者名」、又は「番組概要文に含まれる人物名」といった細かい設定も可能）のみを抽出し、単語リストWL2としてもよい。

【0062】

なお、本実施例では、主に情報抽出装置の構成と動作について説明したが、これに限られず、各構成要素を備え、ユーザが明示的に指定したキーワードだけではなく、ユーザが重要とは判断していないキーワードを抽出するための方法、及びプログラムとして構成されてもよい。

【0063】

さらに、情報抽出装置の機能を実現するためのプログラムをコンピュータで読み取り可能な記録媒体に記録して、この記録媒体に記録されたプログラムをコンピュータシステムに読み込ませ、実行することによって実現してもよい。

【0064】

ここでいう「コンピュータシステム」とは、OSや周辺機器等のハードウェアを含むものとする。また、「コンピュータで読み取り可能な記録媒体」とは、フレキシブルディスク、光磁気ディスク、ROM、CD-ROM等の可搬媒体、コンピュータシステムに内蔵されるハードディスク等の記憶装置のことをいう。

【0065】

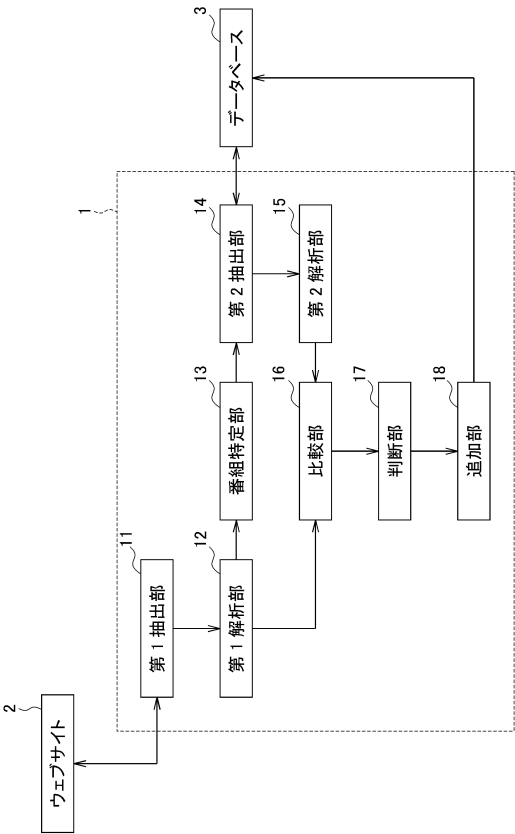
さらに「コンピュータで読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線等の通信回線を介してプログラムを送信する場合の通信線のように、短時刻の間、動的にプログラムを保持するもの、その場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリのように、一定時刻プログラムを保持しているものも含んでもよい。また、上記プログラムは、前述した機能の一部を実現するためのものであってもよく、さらに前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるものであってもよい。

【符号の説明】

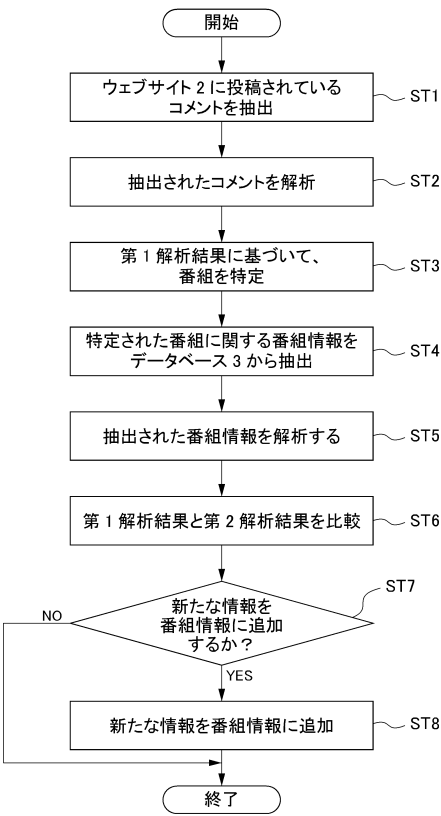
【0066】

- 1 1 第 1 抽出部
- 1 2 第 1 解析部
- 1 3 番組特定部
- 1 4 第 2 抽出部
- 1 5 第 2 解析部
- 1 6 比較部
- 1 7 判断部
- 1 8 追加部

【図 1】



【図 2】



フロントページの続き

- (56)参考文献 特開2011-234198 (JP, A)
特開2005-209020 (JP, A)
国際公開第2012/070182 (WO, A1)
特開2007-274605 (JP, A)

- (58)調査した分野(Int.Cl., DB名)
G06F 17/30