

明 細 書

発明の名称： 計算機及び知識の抽出方法

参照による取り込み

[0001] 本出願は、2023年2月2日に出願された日本特許出願第2023-14586号の優先権を主張し、その内容を参照することにより、本出願に取り込む。

技術分野

[0002] 本発明は、特許及び論文等の文書から有用な知識を抽出する技術に関する。

背景技術

[0003] 公開されている特許及び論文等の文書には、最新の研究成果に基づく知識が記述されている。したがって、文書に記述された知識を用いることによって、研究動向の把握及び最新のデータに基づく分析等が可能となる。例えば、新材料の開発を目的として、文書に記載されている実験データを活用することによって、新材料の特性を予測する統計モデルを生成することができる。

[0004] 文書は、自然文、図、及び表等から構成される非構造データである。そのため、人が特許や論文等の文書から知識を抽出するためには、文書を読み、必要な知識を抽出する必要がある。当該作業は非常に手間がかかり、また、日々、蓄積される文書は増えているため、全ての文書を確認することは困難である。

[0005] そこで、文書から自動的に知識を抽出する技術が求められている。これに対して、非特許文献1等、ルールベース又は機械学習を用いて、文献から自動で有用な知識を抽出する手法が提案されている。

先行技術文献

特許文献

[0006] 特許文献1：米国特許出願効果第2020/0125659号明細書

非特許文献

[0007] 非特許文献1: Swain, M. C., & Cole, J. M.
“ChemDataExtractor: A Toolkit for
Automated Extraction of Chemical I
nformation from the Scientific Lit
erature”, J. Chem. Inf. Model. 201
6, 56 (10), pp 1894-1904 10.1021/ac
s.jcim.6b00207

非特許文献2: Ding, Kaize, et al. “Deep an
omaly detection on attributed netw
orks.”、Proceedings of the 2019 SIA
M International Conference on Data
Mining. Society for Industrial an
d Applied Mathematics, 2019.

発明の概要

発明が解決しようとする課題

[0008] 従来技術を用いることによって、文書から知識を自動的に抽出することが
できる。しかし、従来技術では、抽出された知識の誤り及び抽出された知識
の有用性を判定することが困難である。

[0009] これに対して、特許文献1には、抽出精度を担保する技術が開示されてい
る。具体的には、抽出された業務ルールから、業務データが蓄積されている
DBへの、クエリを作成する。そして、そのクエリにヒットしたデータの数
や割合を指標にして、抽出された業務ルールの妥当性を検証する方法が開示
されている。

[0010] 特許文献1に記載の技術は、抽出された業務ルールに合致する業務デー
タが存在していることを前提としている。したがって、特許文献1は、定まっ
たルールに基づいた業務が行われ、その業務の結果、作成されたデータがD
Bに蓄積されている場合などは有効である。しかし、研究成果のように新規

性が求められるような場合においては、その成果となる特許や論文などの文章から抽出された知識に厳密に合致するデータは、DBには存在しないため、特許文献1の技術を適用することはできない。

[0011] 本発明は、文書から抽出された知識の誤り及び有用性を判定するための情報を提示する知識抽出技術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

[0012] 本願において開示される発明の代表的な一例を示せば以下の通りである。
すなわち、演算装置、及び前記演算装置に接続される記憶装置を備える計算機であって、文字列及び画像を含む文書から抽出された知識を表現する知識グラフを入力とし、前記知識グラフを再現するモデルを保持し、前記知識グラフは、二つ以上のノード及び一つ以上のエッジから構成され、前記計算機は、新規文書を受け付けた場合、前記新規文書から知識を抽出し、抽出された前記知識の前記知識グラフを抽出知識グラフとして生成し、前記抽出知識グラフを前記モデルに入力することによって、再現知識グラフを取得し、前記抽出知識グラフ及び前記再現知識グラフの差異に基づいて、前記モデルを生成するために用いた前記知識グラフに対応する知識に対する、抽出された前記知識の相違の程度を表す相違度を算出し、前記相違度を含む検証情報を出力する。

発明の効果

[0013] 本発明の一形態によれば、文書から抽出された知識の誤り及び有用性を判定するための情報を提示することができる。上記した以外の課題、構成及び効果は、以下の実施例の説明により明らかにされる。

図面の簡単な説明

[0014] [図1]実施例1の知識抽出装置の構成の一例を示す図である。

[図2]実施例1の知識抽出装置が実行する処理の流れを示す図である。

[図3]実施例1の知識データベースに格納される知識グラフの一例を示す図である。

[図4]実施例1の参照知識グラフ生成部が生成する参照知識グラフの一例を示

す図である。

[図5]実施例1の知識抽出部が生成する知識グラフの一例を示す図である。

[図6]実施例1の対象知識グラフ生成部が生成する対象知識グラフの一例を示す図である。

[図7]実施例1の知識提示部が提示する画面の一例を示す図である。

[図8]実施例1の知識抽出装置が実行する知識モデル生成処理の一例を説明するフローチャートである。

[図9]実施例1の知識抽出装置が実行する知識検証処理の一例を説明するフローチャートである。

[図10]実施例2の知識抽出装置の構成の一例を示す図である。

[図11]実施例2の知識抽出装置が実行する処理の流れを示す図である。

[図12]実施例3の知識抽出装置の構成の一例を示す図である。

[図13]実施例3の知識抽出装置が実行する処理の流れを示す図である。

発明を実施するための形態

[0015] 以下、本発明の実施例を、図面を用いて説明する。ただし、本発明は以下に示す実施例の記載内容に限定して解釈されるものではない。本発明の思想ないし趣旨から逸脱しない範囲で、その具体的構成を変更し得ることは当業者であれば容易に理解される。

[0016] 以下に説明する発明の構成において、同一又は類似する構成又は機能には同一の符号を付し、重複する説明は省略する。

実施例 1

[0017] 図1は、実施例1の知識抽出装置の構成の一例を示す図である。

[0018] 知識抽出装置100は、入力装置101、出力装置102、演算装置103、及び記憶装置104を有する。

[0019] 入力装置101は、キーボード、マウス、及びタッチパネル等であり、ユーザからのデータ及びコマンド等の入力を受け付ける。

[0020] 出力装置102は、ディスプレイ等であり、知識抽出装置100との対話的な処理を行うための画面を表示する。

- [0021] 演算装置103は、CPU (Central Processing Unit) 等であり、知識抽出装置100の各種情報処理を実行する。演算装置103は、記憶装置104に格納されるプログラムにしたがって処理を実行することによって、特定の機能を実現する機能部（モジュール）として動作する。実施例1の演算装置103は、参照知識グラフ生成部110、知識検証モデル生成部111、知識抽出部112、対象知識グラフ生成部113、知識検証部114、及び知識提示部115として機能する。
- [0022] 記憶装置104は、メモリ等の記憶装置であり、演算装置103が実行するプログラム及びプログラムが使用する情報を格納する。記憶装置104はワークエリアとしても用いられる。なお、知識抽出装置100は、HDD (Hard Disk Drive) 及びSSD (Solid State Drive) 等の大容量の記憶装置を有してもよい。実施例1の記憶装置104は、知識データベース120を格納する。
- [0023] なお、複数の計算機を用いて、知識抽出装置100が有する機能を実現してもよい。
- [0024] 次に、実施例1の知識抽出装置100が実行する処理の流れについて説明する。図2は、実施例1の知識抽出装置100が実行する処理の流れを示す図である。
- [0025] 知識抽出装置100は、知識データベース120を用いて知識検証モデルを生成する。知識抽出装置100は、文書の入力を受け付けた場合、文書から知識を抽出し、知識検証モデルを用いその知識を検証し、検証結果をレビューアに提示する。
- [0026] 知識データベース120には、Resource Description Framework及びProperty Graph等の知識グラフとして知識が蓄積される。
- [0027] 知識グラフは、実体を表すノード、実体間の関係性を表す辺から構成されるグラフである。ノードには、知識を構成する実体の特性を表す情報として、実体の種類等を表すラベル及び実体の属性が付与され、また、辺には、実

体間の関係性の特性を表す情報として、関係性の種類等を表すラベル及び関係性の属性が付与される。実施例1の知識データベース120には、インターネット上に公開されている材料データベース等、人手で作成され、検証が行われた知識グラフが格納されているものとする。

[0028] 図3は、実施例1の知識データベース120に格納される知識グラフの一例を示す図である。

[0029] 図3に示す知識グラフ300は、化学反応に関する知識をProperty Graphの形態で表現した知識グラフである。知識グラフ300は、ノード301、302、303、304、305、306、307、308、309と、辺351、352、353、354、355、356、357、358、359とから構成される。

[0030] ノードには、識別子及びノードの概念を表すラベルが付与される。例えば、ノード301には、ノードの識別子「001」と、ノード301に対応する概念を表すラベル「物質」とが付与されている。また、辺にはノード間の関係性を表すラベルが付与される。例えば、辺351にはラベル「原料」が付与され、辺352にはラベル「触媒」が付与される。

[0031] また、ノードには属性を付与できる。例えば、「物質」のラベルが付いたノードであれば、International Chemical Identifier、物質名、その類義語等の物質の名称及び識別子に関する情報、化学組成式及びSimplified molecular input line entry system (SMILES)等の物質の構造を表す情報、並びに融点等、物質の特性を表す情報を付与できる。また、「操作」のラベルが付いたノードであれば、攪拌及び加熱等の操作の種類、温度、圧力、及び時間等の操作の条件に関する情報を属性として付与できる。

[0032] また、辺にも属性を付与できる。例えば、辺351には、ノード301に対応する化学反応において原料として使われたノード302に対応する物質の質量等を属性として付与できる。

[0033] 参照知識グラフ生成部110は、知識データベース120に格納される知

識グラフを参照知識グラフに変換する。参照知識グラフは、後述する知識検証モデルを生成するために用いられるグラフである。参照知識グラフ生成部 110 は、知識グラフから知識検証モデルを生成するために必要な情報を抽出し、参照知識グラフに変換する。

[0034] 図4は、実施例1の参照知識グラフ生成部110が生成する参照知識グラフの一例を示す図である。図4には、図3で説明した知識グラフ300から生成された参照知識グラフ400を示す。

[0035] 参照知識グラフ生成部110は、以下のような手順で、知識グラフ300から参照知識グラフ400を生成する。

[0036] まず、参照知識グラフ生成部110は、ノード301に対応する化学反応に関わるノードを抽出する。具体的には、参照知識グラフ生成部110は、ノード302、303、304、305、306を抽出する。

[0037] 知識グラフには、複数の知識が含まれる場合がある。そこで、参照知識グラフ生成部110は、所定の分割単位で知識グラフを分割し、分割されたグラフから参照知識グラフを生成する。

[0038] 参照知識グラフ生成部110は、参照知識グラフのノードとして、抽出されたノードに対応するノードを定義する。このとき、参照知識グラフ生成部110は、融点等の物質の特性、及び、圧力等の操作の条件等の属性を別のノードとして定義する。例えば、ノード302に対応する物質の融点として120度が付与されている場合、参照知識グラフ生成部110は、「融点、120度」に対応するノードを定義する。最後に、参照知識グラフ生成部110は、各ノードに、ノードの種類に応じた特徴量を付与する。

[0039] 図4に示す参照知識グラフ400において、ノード401、402、403、404、405、406は、知識グラフ300のノード301、302、303、304、305、306に対応する。また、ノード407、408は、ノード302に対応する物質の特性に関する数値を表すノードであり、ノード409は、ノード306に対応する操作の条件に関する数値を表すノードである。

- [0040] 図4で示す参照知識グラフ400には、化学反応、物質、数値、及び操作の4つの種類のノードがあり、ノード401は化学反応のノード、ノード402、403、404は物質のノード、ノード405、406は操作のノード、408、409は数値のノードに対応している。
- [0041] 各ノードは、種類に応じた固定長のベクトルを特徴量が付与されている。具体的には、物質に関するノードの場合、物質の構造に関する情報等から、フィンガープリントと呼ばれる固定長のベクトルを計算し、その物質の化学反応中の役割を表すビット列と共に特徴量とする。操作に関するノードの場合、操作の種類をビット列に変換し、固定長のベクトルとする。数値に関するノードの場合、融点及び温度等、その数値の意味、関連する物質、及び操作との関係をビット列に変換し、数値と共に特徴量とする。化学反応に関するノードの場合、重合反応等化学反応の種類をビット列に変換したものを特徴量とする。なお、ビット列への変換は、例えば、One-hot encodingなどを用いる。
- [0042] 知識検証モデル生成部111は、参照知識グラフを用いて知識検証モデルを生成する。例えば、知識検証モデル生成部111は、参照知識グラフを用いて、非特許文献2に開示されているような参照知識グラフを再構成可能なグラフオートエンコーダを、知識検証モデルとして生成する。なお、様々なグラフオートエンコーダの形態が考えられるが、例えば、参照知識グラフの、隣接行列とノードの特徴量を入力とし、その特徴量を再構成するグラフオートエンコーダが一例として挙げられる。
- [0043] 知識検証モデルは、知識を表す参照知識グラフを再現するモデルである。すなわち、知識検証モデルは、参照知識グラフによって表される知識を、自らの知識として取り込んだモデルである。
- [0044] 知識抽出部112は、自然文、図、及び表から構成された文書から知識を抽出し、抽出された知識を表す知識グラフを生成し、抽出知識グラフとして出力する。具体的には以下のような手順で抽出知識グラフが生成される。
- [0045] 知識抽出部112は、固有表現抽出及び関係性抽出等の公知の自然言語処

理の手法を活用し、文書に記述された単語の種類、単語間の関係を抽出する。また、知識抽出部 112 は、公知の画像処理の手法を用いて、文書の図及び表等から数値を抽出する。そして、知識抽出部 112 は、抽出した単語及び数値、並びに、関係に基づいて、知識データベース 120 に格納される知識グラフと同じデータ形式の知識グラフを生成する。ここで、同じデータ形式の知識グラフとは、ノード及び辺に付与されるラベル及び属性の種類が同一である知識グラフを意味する。

[0046] ここで、化学反応に関する文章「物質 A と物質 B から、物質 C を合成した。具体的には、攪拌機を備えた耐熱反応容器に、物質 B (XX g)、物質 D (YY g) を入れ、十分に物質 E で置換した後、容器内を WW MPa に保ちながら攪拌下 ZZ 度まで昇温した。」が入力された場合の抽出知識グラフの具体的な生成方法を説明する。図 5 は、実施例 1 の知識抽出部 112 が生成する知識グラフの一例を示す図である。

[0047] 知識抽出部 112 は、文章から、物質を表す単語として「物質 A」、「物質 B」、「物質 C」、「物質 D」、「物質 E」を抽出し、操作を表す単語として「置換」、「攪拌」を抽出し、質量を表す単語として「XX g」、「YY g」を抽出し、圧力を表す単語として「WW MPa」を抽出し、温度を表す単語として「ZZ 度」を抽出する。また、知識抽出部 112 は、抽出された単語間の関係性を抽出する。例えば、「XX g」は「物質 B」の質量、「YY g」は「物質 D」の質量、「置換」に関連する物質は「物質 B」、「物質 D」、「物質 E」、攪拌の条件は「WW MPa」、「ZZ 度」等を単語間の関係性として抽出する。知識抽出部 112 は、抽出された情報を用いて、図 5 に示すような、文章に含まれる知識を表現する抽出知識グラフ 500 を生成する。抽出知識グラフ 500 は、ノード 501、502、503、504、506、507、508 と、辺 551、552、553、554、555、556、557、558、559、560 とから構成される。

[0048] 抽出知識グラフ 500 の各ノードには、ノードの識別子、ノードのラベル

、対応する文章中の単語等が付与される。なお、上記の文章には記述されていないが、化学反応という実体があると捉えることもできる。したがって、図5では、上記の文章に記述されている化学反応に対応するノード501を追加し、そのノードと、文章中の単語から抽出されたノードとの関連も追加している。例えば、「物質A」は化学反応中で、「原料」として使われているため、ノード501と、「物質A」を表すノード502との間の辺551にはラベル「原料」が付与されている。また、化学反応中で、「物質B」は原料としてXXg使われ、物質Dは触媒としてYYg使われているので、ノード501と、「物質B」を表すノード503との間の辺552には、「原料」というラベルと「質量：XXg」という属性が付与され、ノード501と、物質Dを表すノード505との間の辺554には、「触媒」というラベルと、「質量：YYg」という属性がされている。

[0049] 対象知識グラフ生成部113は、必要に応じて、知識データベース120に格納される知識グラフや他のDBなどを参照して、抽出知識グラフの各ノードに属性等を追加する。また、対象知識グラフ生成部113は、抽出知識グラフを変換して、参照知識グラフと同じデータ形式の対象知識グラフに変換する。対象知識グラフの生成方法は、参照知識グラフの生成方法と同様の方法である。

[0050] 図5に示す抽出知識グラフ500への属性の追加は、例えば、以下のように行われる。

[0051] 物質に関するノードに対する属性の追加では、対象知識グラフ生成部113は、物質A、物質B等、文章中における物質の表現に基づいて知識データベース120を参照し、同一の物質に対応するノードを検索する。同一の物質に対応するノードの検索では、知識データベース120の知識グラフの物質名称や識別子に関する情報が用いられる。同一の物質に対応するノードが存在する場合、対象知識グラフ生成部113は、検索されたノードの物質の構造に関する情報等を、抽出知識グラフの対応するノードに属性として追加する。

- [0052] なお、International Union of Pure and Applied Chemistryが定義する化合物の命名方法等、物質の構造が、その文章中の記述から一意に決まる場合、文章中の表現から構造を推定し、それを抽出知識グラフの対応するノードの属性として追加してもよい。
- [0053] 操作に関するノードに対する属性の追加では、例えば、対象知識グラフ生成部113は、ガス置換、攪拌など、予め定義された操作の種類に関する情報を参照し、置換、攪拌等、文章中における操作の表現に基づいて、操作の種類を特定する。この特定には、類語辞典、予め学習された機械学習モデル等を用いることもできる。対象知識グラフ生成部113は、特性した操作の種類を属性として、操作に関するノードに追加する。
- [0054] 抽出知識グラフのノードへの属性の追加が実行された後は、対象知識グラフ生成部113は、参照知識グラフ生成部110と同様の処理方法で、抽出知識グラフからノードを抽出し、抽出された各ノードに、種類に応じた固定長のベクトルで表現される特徴量を付与することによって、対象知識グラフを生成する。
- [0055] なお、知識データベース120や外部DB及び文章中の表現から、操作種類が特定できない操作や構造が特定できなかった物質に関しては、操作や構造が特定できなかったことが判別できるような識別子を、対応するノードに付与する。
- [0056] 図6は、実施例1の対象知識グラフ生成部113が生成する対象知識グラフの一例を示す図である。図6には、図5で説明した抽出知識グラフ500から生成された対象知識グラフ600を示す。対象知識グラフ600は、図4で示す参照知識グラフ400と同じデータ形式のグラフである。
- [0057] 対象知識グラフ600は、ノード601、602、603、604、606、607、608、609、610、611、612と、辺651、652、653、654、655、656、657、658、659、660、661、662、663とから構成される。

- [0058] 対象知識グラフ600のノード601、602、603、604、605、606、607、608は、抽出知識グラフ500のノード501、502、503、504、505、506、507、508に対応する。また、ノード609はノード503に対応する物質Bの質量、ノード610はノード505に対応する物質Cの質量を表すノードである。ノード611、612は、ノード508に対応する攪拌の条件である圧力及び温度を表すノードである。各ノードには種類に応じた固定長のベクトルが特徴量として付与されている。
- [0059] 知識検証部114は、知識検証モデルを用いて、入力された対象知識グラフの各ノードの特徴量の再構成を行い、入力された対象知識グラフの各ノードの特徴量と、再構成された各ノードの特徴量との再構成誤差に基づいて、各ノードの特徴量の相違度を計算する。知識検証部114は、対象知識グラフ及び相違度を知識提示部115に出力する。
- [0060] 知識検証モデルは、知識を表す参照知識グラフを再現するモデルである。すなわち、知識検証モデルは、参照知識グラフによって表される知識を、自らの知識として取り込んだモデルである。入力された対象知識グラフに類似している知識が知識検証モデルに取り込まれている場合、入力された知識グラフと、出力された知識グラフとの差異は小さい。入力された対象知識グラフに対応する知識が知識検証モデルに取り込まれている知識と異なる情報を含む場合、入力された知識グラフと、出力された知識グラフとの差異は大きくなる。
- [0061] したがって、前述の差異の大きさを表す相違度は、既存の知識に対する、文書から抽出された知識の相違の程度を表す値として利用できる。
- [0062] 例えば、参照知識グラフの隣接行列及びノードの特徴量ベクトルを入力とし、その特徴量を再構成するオートエンコーダが知識検証モデルとして生成された場合を考える。この場合、知識検証部114は、知識検証モデルに、対象知識グラフの隣接行列及び各ノードの特徴量ベクトルを入力する。知識検証モデルからは、再構成した各ノードの特徴量ベクトルが出力される。知

識検証部 114 は、入力した特徴量ベクトルと、再構成された特徴量ベクトルとの差のノルム等を相違度として算出する。なお、特徴量が付与できなかったノードに関しては、そのノードを無視するか、周辺のノードなどから、そのノードの特徴量を推定する。

[0063] 知識提示部 115 は、対象知識グラフ及び相違度を含む検証情報を生成し、出力装置 102 を介してレビューに提示する。このとき、知識提示部 115 は、相違度が大きいノードを抽出し、当該ノードを強調表示してもよい。レビューは、対象知識グラフとともに、相違度を参照することによって、抽出された知識と既存の知識との間の相違点を把握できる。例えば、抽出された知識の誤り、新規性等を把握することができる。

[0064] 図 7 は、実施例 1 の知識提示部 115 が提示する画面の一例を示す図である。画面 700 は、文書表示領域 701 及び知識表示領域 702 を含む。

[0065] 文書表示領域 701 は、入力された文書に関する情報を表示する領域である。図 7 の文書表示領域 701 には、対象となった文章を表示する表示欄 710 が含まれる。

[0066] 知識表示領域 702 は、文書から抽出された知識に関する情報を表示する領域である。図 7 の知識表示領域 702 には、表示欄 710 に表示される文章から抽出された知識が表示される。

[0067] 入力された文書（文献）に基づいて表示欄 710 への表示が行われる。抽出知識グラフ、対象知識グラフ、及び相違度に基づいて知識表示領域 702 への表示が行われる。画面 700 に表示される文章及び知識は、図 5 及び図 6 で示した抽出知識グラフ 500 及び対象知識グラフ 600 と同一のものである。

[0068] 画面 700 は、レビューが必要に応じて抽出結果を変更できるようになっている。例えば、文章中の単語「物質 A」は、「物質」と認識されているが、レビューはプルダウン 720 を操作して、操作等、別の種類に修正できる。同様に、「物質 A」の化学反応中の役割は「触媒」と認識されているが、プルダウン 721 を操作して、原料等、他の役割に修正できる。また、「X

X g」は物質Bの質量と認識されているが、削除して、代わりに物質Cの質量に変更することもできる。

[0069] 図7では、知識提示部115は、相違度が大きいノードに対応する項目の確認及び修正を促す表示を行う。例えば、物質Aの役割は「触媒」と認識されたが、相違度が大きい場合、物質Aの役割の認識結果を表示する欄の近くにメッセージを表示する。また、相違度が小さくなる役割を示すメッセージ722が表示されてもよい。メッセージ722には、相違度が小さくなる物質Aの役割が原料であることを示唆している。

[0070] 次に、知識抽出装置100が実行する処理の詳細について説明する。

[0071] 図8は、実施例1の知識抽出装置100が実行する知識モデル生成処理の一例を説明するフローチャートである。

[0072] 参照知識グラフ生成部110は、知識データベース120に格納される知識グラフを用いて参照知識グラフを生成する（ステップS801）。

[0073] 知識検証モデル生成部111は、参照知識グラフを用いて知識検証モデルを生成する（ステップS802）。知識検証モデルは公知の機械学習の手法を用いて生成される。

[0074] 図9は、実施例1の知識抽出装置100が実行する知識検証処理の一例を説明するフローチャートである。知識抽出装置100は、文書の入力を受け付けた場合、知識抽出処理を開始する。

[0075] 知識抽出部112は、受け付けた文書から知識を抽出し、抽出知識グラフを生成する（ステップS901）。知識抽出部112は、抽出知識グラフを対象知識グラフ生成部113及び知識提示部115に出力する。

[0076] 対象知識グラフ生成部113は、抽出知識グラフを用いて、対象知識グラフを生成する（ステップS902）。対象知識グラフ生成部113は、必要に応じて知識データベースを参照する。

[0077] 知識検証部114は、知識検証モデルを用いて、対象知識グラフにおける相違度を算出し、知識提示部115に対象知識グラフ及び相違度を出力する（ステップS903）。なお、抽出知識グラフから複数の対象知識グラフが

生成された場合、各対象知識グラフに対し、ステップS903の処理が実行される。なお、対象知識グラフの相違度が一定基準以下の場合には提示しなくてもよい。

[0078] 知識提示部115は、抽出知識グラフ、対象知識グラフ、及び相違度に基づいて、抽出された知識に関する情報を提示する（ステップS904）。

[0079] 実施例1によれば、知識抽出装置100は、文書から抽出された知識が既存の知識とどの程度相違するかを定量的に評価し、評価結果を含む情報をユーザに提示できる。ユーザは、情報を参照することによって、抽出された知識の誤り及び有用性を把握することができる。

[0080] 例えば、知識の抽出精度が低い場合でも、抽出された知識の誤りを把握し、修正することによって、質の高い知識を取得することができる。

実施例 2

[0081] 実施例2の知識抽出装置100は、外部から知識グラフを取得し、知識データベース120に格納する。以下、実施例1との差異を中心に実施例2について説明する。

[0082] 図10は、実施例2の知識抽出装置の構成の一例を示す図である。図11は、実施例2の知識抽出装置100が実行する処理の流れを示す図である。

[0083] 実施例2の知識抽出装置100は、新たに、通信装置105を有する。通信装置105は、インターネット等のネットワークを介して外部と接続し、演算装置103の処理内容を送信し、また、外部からの情報を受信する。

[0084] 実施例2の演算装置103は、新たに、知識収集部116として機能する。知識収集部116は、通信装置105を介し、インターネットに公開されているデータベース等から、知識グラフを収集し、知識データベース120に格納する。なお、知識グラフの形態で情報が存在しない場合、知識抽出部112等が取得した知識を知識グラフに変換し、知識データベース120に格納してもよい。

[0085] なお、外部からの知識（知識グラフ）の収集は、知識検証モデルの生成時、又は知識検証処理の実行時に行われてもよいし、また、前述の処理とは独

立した実行タイミングで実行されてもよい。

実施例 3

- [0086] 実施例3の知識抽出装置100は、レビューが確認した対象知識グラフを、知識データベース120に格納可能なデータ形式の知識グラフに変換し、格納する。以下、実施例1との差異を中心に実施例3について説明する。
- [0087] 図12は、実施例3の知識抽出装置の構成の一例を示す図である。図13は、実施例3の知識抽出装置100が実行する処理の流れを示す図である。
- [0088] 実施例3の知識抽出装置100のハードウェア構成は実施例1と同一である。実施例3の演算装置103は、新たに、知識追加部117として機能する。
- [0089] 知識追加部117は、レビューが確認及び修正した対象知識グラフを、知識データベース120に格納可能なデータ形式の知識グラフに変換し、格納する。
- [0090] なお、本発明は上記した実施例に限定されるものではなく、様々な変形例が含まれる。また、例えば、上記した実施例は本発明を分かりやすく説明するために構成を詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。また、各実施例の構成の一部について、他の構成に追加、削除、置換することが可能である。
- [0091] また、上記の各構成、機能、処理部、処理手段等は、それらの一部又は全部を、例えば集積回路で設計する等によりハードウェアで実現してもよい。また、本発明は、実施例の機能を実現するソフトウェアのプログラムコードによっても実現できる。この場合、プログラムコードを記録した記憶媒体をコンピュータに提供し、そのコンピュータが備えるプロセッサが記憶媒体に格納されたプログラムコードを読み出す。この場合、記憶媒体から読み出されたプログラムコード自体が前述した実施例の機能を実現することになり、そのプログラムコード自体、及びそれを記憶した記憶媒体は本発明を構成することになる。このようなプログラムコードを供給するための記憶媒体としては、例えば、フレキシブルディスク、CD-ROM、DVD-ROM、ハ

ードディスク、SSD (Solid State Drive)、光ディスク、光磁気ディスク、CD-R、磁気テープ、不揮発性のメモ리카ード、ROMなどが用いられる。

[0092] また、本実施例に記載の機能を実現するプログラムコードは、例えば、アセンブラ、C/C++、perl、Shell、PHP、Python、Java等の広範囲のプログラム又はスクリプト言語で実装できる。

[0093] さらに、実施例の機能を実現するソフトウェアのプログラムコードを、ネットワークを介して配信することによって、それをコンピュータのハードディスクやメモリ等の記憶手段又はCD-RW、CD-R等の記憶媒体に格納し、コンピュータが備えるプロセッサが当該記憶手段や当該記憶媒体に格納されたプログラムコードを読み出して実行するようにしてもよい。

[0094] 上述の実施例において、制御線や情報線は、説明上必要と考えられるものを示しており、製品上必ずしも全ての制御線や情報線を示しているとは限らない。全ての構成が相互に接続されていてもよい。

請求の範囲

- [請求項1] 演算装置、及び前記演算装置に接続される記憶装置を備える計算機であって、
- 文字列及び画像を含む文書から抽出された知識を表現する知識グラフを入力とし、前記知識グラフを再現するモデルを保持し、
- 前記計算機は、
- 新規文書を受け付けた場合、前記新規文書から知識を抽出し、
- 抽出された前記知識の前記知識グラフを抽出知識グラフとして生成し、
- 前記抽出知識グラフを前記モデルに入力することによって、再現知識グラフを取得し、
- 前記抽出知識グラフ及び前記再現知識グラフの差異に基づいて、前記モデルを生成するために用いた前記知識グラフに対応する知識に対する、抽出された前記知識の相違の程度を表す相違度を算出し、
- 前記相違度を含む検証情報を出力することを特徴とする計算機。
- [請求項2] 請求項1に記載の計算機であって、
- 前記知識グラフは、知識を構成する実体を表すノード及び実体間の関係性を表すエッジから構成され、
- 前記ノードは特徴量を有し、
- 前記計算機は、前記抽出知識グラフの前記ノードの特徴量と、前記再現知識グラフの前記ノードの特徴量との差に基づいて、前記抽出知識グラフの前記ノードごとに前記相違度を算出することを特徴とする計算機。
- [請求項3] 請求項2に記載の計算機であって、
- 前記抽出知識グラフが、前記実体の特性に関する特性情報が付与された前記ノードから構成される場合、前記特性情報を前記特徴量に変換し、付与することを特徴とする計算機。
- [請求項4] 請求項1に記載の計算機であって、

信頼性が高い前記知識グラフが格納されるデータベースにアクセス可能であって、

前記データベースに格納される前記知識グラフを用いて、前記モデルを生成することを特徴とする計算機。

[請求項5]

請求項4に記載の計算機であって、

前記検証情報を参照したユーザによって修正された前記抽出知識グラフを前記データベースに格納することを特徴とする計算機。

[請求項6]

計算機が実行する知識の抽出方法であって、

前記計算機は、

演算装置、及び前記演算装置に接続される記憶装置を有し、

文字列及び画像を含む文書から抽出された知識を表現する知識グラフを入力とし、前記知識グラフを再現するモデルを保持し、

前記知識の抽出方法は、

前記計算機が、新規文書を受け付けた場合、前記新規文書から知識を抽出する第1のステップと、

前記計算機が、抽出された前記知識の前記知識グラフを抽出知識グラフとして生成する第2のステップと、

前記計算機が、前記抽出知識グラフを前記モデルに入力することによって、再現知識グラフを取得する第3のステップと、

前記計算機が、前記抽出知識グラフ及び前記再現知識グラフの差異に基づいて、前記モデルを生成するために用いた前記知識グラフに対応する知識に対する、抽出された前記知識の相違の程度を表す相違度を算出する第4のステップと、

前記計算機が、前記相違度を含む検証情報を出力する第5のステップと、を含むことを特徴とする知識の抽出方法。

[請求項7]

請求項6に記載の知識の抽出方法であって、

前記知識グラフは、知識を構成する実体を表すノード及び実体間の関係性を表すエッジから構成され、

前記ノードは特徴量を有し、

前記第4のステップは、前記計算機が、前記抽出知識グラフの前記ノードの特徴量と、前記再現知識グラフの前記ノードの特徴量との差に基づいて、前記抽出知識グラフの前記ノードごとに前記相違度を算出するステップを含むことを特徴とする知識の抽出方法。

[請求項8]

請求項7に記載の知識の抽出方法であって、

前記第3のステップは、前記抽出知識グラフが、前記実体の特性に関する特性情報が付与された前記ノードから構成される場合、前記計算機が、前記特性情報を前記特徴量に変換し、付与するステップを含むことを特徴とする知識の抽出方法。

[請求項9]

請求項6に記載の知識の抽出方法であって、

前記計算機は、信頼性が高い前記知識グラフが格納されるデータベースにアクセス可能であって、

前記知識の抽出方法は、前記計算機が、前記データベースに格納される前記知識グラフを用いて、前記モデルを生成するステップを含むことを特徴とする知識の抽出方法。

[請求項10]

請求項9に記載の知識の抽出方法であって、

前記計算機が、前記検証情報を参照したユーザによって修正された前記抽出知識グラフを前記データベースに格納するステップを含むことを特徴とする知識の抽出方法。