

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2018年5月11日(11.05.2018)



(10) 国際公開番号

WO 2018/083804 A1

(51) 国際特許分類:
G06F 17/30 (2006.01)

(21) 国際出願番号: PCT/JP2016/083000

(22) 国際出願日: 2016年11月7日(07.11.2016)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(71) 出願人: 富士通株式会社(FUJITSU LIMITED)
[JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 Kanagawa (JP).

(72) 発明者: 溝 渕 裕 司 (MIZOBUCHI, Yuji);
〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内 Kanagawa (JP).

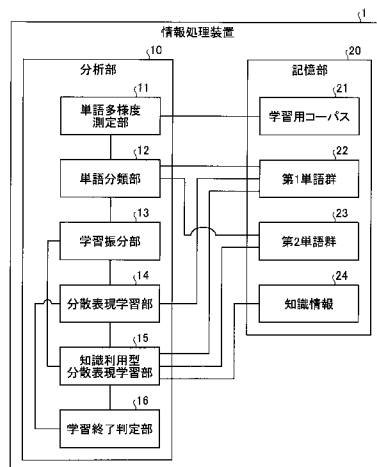
(74) 代理人: 特許業務法人酒井国際特許事務所 (SAKAI INTERNATIONAL PATENT OFFICE); 〒1000013 東京都千代田区霞が関3丁目8番1号 虎の門三井ビルディング Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): 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, DJ, 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, KW, KZ, LA, LC, LK, LR, LS, 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,

(54) Title: ANALYSIS PROGRAM, INFORMATION PROCESSING DEVICE, AND ANALYSIS METHOD

(54) 発明の名称: 分析プログラム、情報処理装置および分析方法

[図1]



- 1 Information processing device
- 10 Analysis unit
- 11 Word degree of variety measurement unit
- 12 Word classification unit
- 13 Learning filtering unit
- 14 Distributed expression learning unit
- 15 Knowledge use-type distributed representation learning unit
- 16 Learning completion assessment unit
- 20 Storage unit
- 21 Corpus for learning
- 22 First word group
- 23 Second word group
- 24 Knowledge information

(57) Abstract: Provided is an information processing device 1 which is capable of learning of distributed representations of words with high precision by: measuring a degree of variety of each word which is included in a document used in learning and classifying each word into a first word group 22 in which the degree of variety is higher than a prescribed regulatory value and a second word group 23 in which the degree of variety is lower than the regulatory value; learning distributed representations of the words which are included in the first word group 22 using a first learning algorithm

SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ,
VC, VN, ZA, ZM, ZW.

- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (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), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

- 一 国際調査報告 (条約第21条(3))

for learning the distributed representations; and, if a word which is used by a second learning algorithm for learning the distributed representations is included in the first word group 22, setting the distributed representation of the included word to the second learning algorithm and learning the distributed representations of the words included in the second word group 23 using the second learning algorithm.

(57) 要約: 情報処理装置 1 が、学習で用いられる文書に含まれる各単語の多様度を測定し、各単語を多様度が所定の規定値より高い第 1 単語群 2 2 および規定値より低い第 2 単語群 2 3 に分類し、第 1 単語群 2 2 に含まれる単語の分散表現を、分散表現を学習する第 1 の学習アルゴリズムを用いて学習し、分散表現を学習する第 2 の学習アルゴリズムにより用いられる単語が第 1 単語群 2 2 に含まれる場合には、該含まれる単語の分散表現を第 2 の学習アルゴリズムに設定し、第 2 の学習アルゴリズムを用いて、第 2 単語群 2 3 に含まれる単語の分散表現を学習することで、単語の分散表現の学習を高精度に行うことができる。

明 細 書

発明の名称： 分析プログラム、情報処理装置および分析方法

技術分野

[0001] 本発明は、分析プログラムなどに関する。

背景技術

[0002] 文章処理において、高精度な単語の表現を如何に獲得するかが、重要であり、従来から多くの研究がなされている。近年では、単語の表現を獲得する技術として、Word2Vecがよく知られている（例えば、非特許文献1参照）。

[0003] Word2Vecは、ニューラルランゲージモデルの一つで「似た単語は同じような文に登場する」という考え方（Distribution hypothesis）に基づいて単語の分散表現を獲得する仕組みである。ニューラルランゲージモデルとは、Bengioらによって考案されたニューラルネットモデルを使って単語の分散表現を学習する方法である。ここでいう単語の分散表現とは、単語間の意味的な関係をベクトルで表現したものである。Word2Vecは、入力層、隠れ層および出力層からなるニューラルネットワークの教師付き学習方法を活用し、文中に現れる単語であって与えられた単語と周辺の単語との関係および与えられた単語の分散表現を学習する。Word2Vecには、Continuous-Bag-of-WordsモデルとSkip-gramモデルが提案されている。ここでいうSkip-gramモデルとは、与えられた単語に対応するベクトルを入力し、周辺の単語を予測するモデルである。

[0004] 図6は、Word2VecのSkip-gramモデルを示す図である。図6に示すように、Skip-gramモデルに対して、Input layer（入力層）、Hidden layer（隠れ層）およびOutput layer（出力層）からなるニューラルネットワークが構築される。

[0005] 入力層には、文中に現れる単語であって与えられた単語に対応するZ次元

の入力ベクトル x が入力される。入力ベクトル x は、 Z 次元の `one-hot` ベクトルである。 Z は、文中の単語数である。`one-hot` ベクトルとは、`one-hot` (ワン・ホット) は 1 つだけ `High` (1) であり、他は `Low` (0) であるようなビット列のことを指す。

[0006] 隠れ層には、与えられた単語の M 次元の単語ベクトル w が生成される。 $W_{Z \times M}$ は、入力層と隠れ層との間の重みであり、 $Z \times M$ の行列により表わされる。出力層には、 C 個のパネルにそれぞれ Z 次元の出力ベクトル y_c が生成される。ここでいうパネルとは、ウィンドウ (window) と同義である。かかる C は、パネル数であり、 y_c は、与えられた単語の前後にある単語に対応する出力ベクトルである。 $W'_{M \times Z}$ は、隠れ層と出力層との間の重みであり、 $M \times Z$ の行列により表わされる。

[0007] `Skip-gram` モデルは、入力層 (Input layer) に入力する入力ベクトルとして Z 次元の `one-hot` ベクトル x_z を入力する。そして、`Skip-gram` モデルは、入力した `one-hot` ベクトル x_z を M 次元の隠れ層 (Hidden layer) に写像した後、 C 個のパネルに対応するそれぞれの Z 次元の出力ベクトルを出力層 (Output layer) に出力する。

[0008] そして、`Skip-gram` モデルは、出力された C 個の出力ベクトルがそれぞれ予め定められた予測ベクトルと異なれば、その差分を学習すべく、隠れ層と出力層との間の重み $W'_{M \times Z}$ 、入力層と隠れ層との間の重み $W_{Z \times M}$ の順にパラメータである重みを更新する。このようなパラメータの更新は、例えば、バックプロパゲーションといわれる。

[0009] そして、繰り返し学習して得られた隠れ層の単語ベクトル w が、与えられた単語 (入力ベクトル x) の分散表現となる。

[0010] 単語の分散表現の精度向上のために、既存知識の活用が多くなされている。かかる既存知識には、例えば、類語、反対語、英語でいう原形、比較級および最上級の関係が挙げられる。例えば、既存知識を活用した分散表現の学習として `RC-NET` の技術が知られている (例えば、非特許文献 2 参照)。`RC-NET` の技術について、図 7 ~ 図 9 を参照して簡単に説明する。

[0011] 図7は、RC-NE Tで用いられる知識の一例を示す図である。図7に示すように、知識グラフは、2つの知識の型を含んでいる。1つの知識は、関係の知識 (Relational Knowledge) であり、もう1つの知識は、カテゴリの知識 (Categorical Knowledge) である。例えば、単語間の関係の知識として、「Elizabeth 2世」は、「UK」との関係について、「UK」の女王 (Queen_of「UK」) という関係を持つ。単語のカテゴリの知識として、「UK」の類義語は、「Britain」「United Kingdom」を含む。

[0012] 図8は、RC-NE Tで用いられる知識の位置関係のイメージを示す図である。図8に示すように、関係の知識 (Relational Knowledge) として、任意の関係にある2つの単語は、距離的に近い関係にある。カテゴリの知識 (Categorical Knowledge) として、任意のカテゴリに含まれる単語同士は、距離的に近い関係にある。単語表現の質を向上させるために、関係の知識およびカテゴリの知識が活用される。

[0013] 図9は、RC-NE Tのアーキテクチャを示す図である。RC-NE Tの目的は、テキストストリーム、関係の知識およびカテゴリの知識に基づいて単語の表現や関係表現を学習することである。図9に示すように、RC-NE Tは、Skip-gramモデルをベースにして、関係の知識およびカテゴリの知識を活用して単語表現の質を向上させる。ここでいう関係の知識がR-NE Tに対応する。ここでいうカテゴリの知識がC-NE Tに対応する。R-NE Tは、学習したい単語の分散表現 w_k と関係の埋め込み (Relation embedding) r を用いて $w'_{k,r}$ を求め、学習したい単語が関係の知識から取得される関係先の単語 $w_{k,r}$ と距離的に近いように学習する。C-NE Tは、同じカテゴリに属する w_k と、 $w_1, w_2, \dots, w_v$ とそれぞれ距離的に近いように学習する。学習には、バックプロパゲーションが用いられる。

先行技術文献

特許文献

[0014] 特許文献1：特表2005-535007号公報

特許文献2：特開平7-334477号公報

特許文献3：特開2005-208782号公報

非特許文献

[0015] 非特許文献1：Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." In Proceedings of Workshop at ICLR, 2013.

非特許文献2：Xu Chang et al. "Rc-net: A general framework for incorporating knowledge into word representations." Proceedings of the 23rd ACM International Conference on Conference on Information and knowledge Management. ACM, 2014.

非特許文献3：Bengio, Yoshua, et al. "A neural probabilistic language model." Journal of machine learning research 3.Feb (2003): 1137-1155.

発明の概要

発明が解決しようとする課題

[0016] しかしながら、RC-NETの技術では、単語の分散表現の学習を高精度に行うことができないという問題がある。すなわち、RC-NETでは、関係の知識およびカテゴリの知識に含まれる単語が基準となるため、これら単語の場所が適切な箇所でなければ、学習自体が適切に行われない。ところが、関係の知識に含まれる単語やカテゴリの知識に含まれる単語の分散表現が十分に学習されたかどうかを考慮することなく、学習が行われる場合がある。かかる場合には、単語の分散表現の学習を高精度に行うことができない。

[0017] 本発明は、1つの側面では、単語の分散表現の学習を高精度に行うことを目的とする。

課題を解決するための手段

[0018] 1つの態様では、分析プログラムは、コンピュータに、学習で用いられる文書に含まれる各単語の多様度を測定し、前記各単語を前記多様度が所定の規定値より高い第1の単語群および前記規定値より低い第2の単語群に分類

し、前記第 1 の単語群に含まれる単語の分散表現を、前記分散表現を学習する第 1 の学習アルゴリズムを用いて学習し、前記分散表現を学習する第 2 の学習アルゴリズムにより用いられる単語が前記第 1 の単語群に含まれる場合には、該含まれる単語の分散表現を前記第 2 の学習アルゴリズムに設定し、前記第 2 の学習アルゴリズムを用いて、前記第 2 の単語群に含まれる単語の分散表現を学習する、処理を実行させる。

発明の効果

[0019] 1 実施態様によれば、単語の分散表現の学習を高精度に行うことができる。

図面の簡単な説明

[0020] [図1]図 1 は、実施例に係る情報処理装置の構成を示す機能ブロック図である。

[図2A]図 2 A は、実施例に係る分散表現学習処理の一例を示す図（1）である。

[図2B]図 2 B は、実施例に係る分散表現学習処理の一例を示す図（2）である。

[図2C]図 2 C は、実施例に係る分散表現学習処理の一例を示す図（3）である。

[図3]図 3 は、実施例に係る知識利用型分散表現学習処理の一例を示す図である。

[図4]図 4 は、実施例に係る分析処理のフローチャートの一例を示す図である。

[図5]図 5 は、情報処理装置のハードウェア構成の一例を示す図である。

[図6]図 6 は、Word2VecのSkip-gramモデルを示す図である。

[図7]図 7 は、RC-NETで用いられる知識の一例を示す図である。

[図8]図 8 は、RC-NETで用いられる知識の位置関係のイメージを示す図である。

[図9]図9は、RC-NE Tのアーキテクチャを示す図である。

発明を実施するための形態

[0021] 以下に、本願の開示する分析プログラム、情報処理装置および分析方法の実施例を図面に基づいて詳細に説明する。なお、本発明は、実施例により限定されるものではない。

実施例

[0022] [実施例に係る情報処理装置の構成]

図1は、実施例に係る情報処理装置の構成を示す機能ブロック図である。図1に示す情報処理装置1は、知識を活用した分散表現学習において、学習で用いられる文書に含まれる単語の多様性を考慮して単語の分散表現を学習する。ここでいう単語の分散表現とは、単語間の意味的な関係をベクトルで表現したものをいう。ここでいう単語の多様性とは、単語が異なる表現の文で現れることをいう。異なる表現の文とは、ある単語が現れる文において、其々で共起する単語が異なっていることをいう。例えば、ある単語に注目したとき、その単語の多様性が高い場合、様々な異なる表現の文に現れることをいう。この多様性の度合いを多様度という。

[0023] 情報処理装置1は、分析部10と、記憶部20とを有する。分析部10は、CPU (Central Processing Unit) などの電子回路に対応する。そして、分析部10は、各種の処理手順を規定したプログラムや制御データを格納するための内部メモリを有し、これらによって種々の処理を実行する。分析部10は、単語多様度測定部11、単語分類部12、学習振分部13、分散表現学習部14、知識利用型分散表現学習部15および学習終了判定部16を有する。

[0024] 記憶部20は、例えば、RAM、フラッシュメモリ (Flash Memory) などの半導体メモリ素子、または、ハードディスク、光ディスクなどの記憶装置である。記憶部20は、学習用コーパス21、第1単語群22、第2単語群23および知識情報24を有する。

[0025] 学習用コーパス21は、学習に用いられるコーパスである。なお、コーパ

スとは、文章の集まりのことをいう。

[0026] 第1単語群22は、多様度が規定値より高い単語の集まりであり、偏りが
ない複数の文に存在する単語の集まりである。言い換えれば、第1単語群2
2とは、学習が十分にできる単語の集まりである。

[0027] 第2単語群23は、多様度が規定値より低い単語の集まりであり、偏りが
ある単数または複数の文に存在する単語の集まりである。言い換えれば、第
2単語群23とは、学習が十分にできない単語の集まりである。なお、規定
値は、多様度の高低を判別する際に用いる境界値であれば良く、多様度が規
定値と一致する場合には多様度が高いとしても良いし、低いとしても良い。

[0028] 知識情報24は、学習で活用される知識の情報である。例えば、知識情報
24は、既存知識を活用して単語の分散表現を学習するRC-NETで用い
られる知識の情報であり、関係の知識の情報を示す関係知識情報およびカテ
ゴリの知識の情報を示すカテゴリ知識情報を含む。関係知識情報は、RC-
NETにおけるR-NETで用いられる知識の情報である。カテゴリ知識情
報は、RC-NETにおけるC-NETで用いられる知識の情報である。関
係知識情報には、関係の知識で用いられる単語と分散表現とが予め対応づけ
られていても良い。知識情報24のカテゴリ知識情報には、カテゴリの知識
で用いられる単語と分散表現とが予め対応づけられていても良い。なお、R
C-NETの学習は、「Xu Chang et al. "Rc-net: A general framewor
k for incorporating knowledge into word representations."」に
開示されている技術を使って行われる。

[0029] 単語多様度測定部11は、学習用コーパス21に含まれる単語の多様度を
測定する。

[0030] 第1の例として、単語多様度測定部11は、学習用コーパス21に含まれ
る単語について、当該単語が出現する文における当該単語と周辺の単語との
共起を表す関係（共起関係）を表わす出現ベクトルを生成する。ここでいう
共起関係とは、ある文においてある単語とある単語とが同時に出現する関係
を意味する。単語と出現ベクトルは、文ごとに当該単語の周辺の単語とその

位置とから生成される。当該単語の周辺の単語の数は、Skip-gramモデルの出力層で用いられるパラメータであるパネルのサイズと一致する。そして、単語多様度測定部11は、当該単語について生成された出現ベクトルの種類数に基づいて、当該単語の多様度を測定する。つまり、当該単語を含む単語から生成される出現ベクトルの種類数が多ければ、当該単語が偏っていない複数の文に存在すると推測され、学習が十分にできると推測される。当該単語を含む単語から生成される出現ベクトルの種類数が少なければ、当該単語が偏っている単数または複数の文に存在すると推測され、学習が十分にできないと推測される。

[0031] 具体的には、単語“apple”の多様度を測定することとする。例えば、学習用コーパス21に文1、文2、文3が含まれている。文1は“She cut an apple with a knife”であり、文2は“I eat an apple”であり、文3は“You eat an apple”である。かかる場合には、単語の辞書は、(she, cut, an, apple, with, knife, I, eat, you)であるとする。このような状況の下、パネルの数が2であるとする、文1では、単語“apple”と当該単語の周辺の単語との関係を表す出現ベクトルとして(0, 1, 1, 0, 1, 0, 0, 0, 0)が生成される。文2および文3では、単語“apple”と当該単語の周辺の単語との関係を表す出現ベクトルとして(0, 0, 1, 0, 0, 0, 0, 1, 0)が生成される。したがって、出現ベクトルの種類数として2が得られる。ここでは、出現ベクトルの種類数2が単語“apple”の多様度として測定される。

[0032] 第2の例として、単語多様度測定部11は、学習用コーパス21に含まれる単語について、当該単語が出現する文における当該単語と周辺の単語との関係を表わす出現ベクトルを生成する。出現ベクトルは、文ごとに当該単語の周辺の単語とその位置とから生成される。当該単語の周辺の単語の数は、Skip-gramモデルの出力層で用いられるパラメータであるパネルの数と一致する。そして、単語多様度測定部11は、当該単語について生成された出現ベクトルが複数存在する場合には、生成された出現ベクトル間で出

現ベクトルの類似度の平均や分散を算出する。出現ベクトルの類似度は、例えば、出現ベクトル間の内積で算出される。そして、単語多様度測定部 11 は、当該単語について算出された出現ベクトルの類似度の平均や分散に基づいて、当該単語の多様度を測定する。つまり、当該単語を含む単語から生成される出現ベクトル同士の類似度が低ければ、当該単語の分散は大きくなり、当該単語が偏っていない複数の文に存在すると推測され、学習が十分にできると推測される。当該単語を含む単語から生成される出現ベクトル同士の類似度が高ければ、当該単語の分散は小さくなり、当該単語が偏っている単語または複数の文に存在すると推測され、学習が十分にできないと推測される。

[0033] 具体的には、単語 “apple” の多様度を測定することとする。第 2 の例も、第 1 の例と同様に、例えば、学習用コーパス 21 に文 1、文 2、文 3 が含まれている。文 1 は “She cut an apple with a knife” であり、文 2 は “I eat an apple” であり、文 3 は “You eat an apple” である。かかる場合には、単語の辞書は、(she, cut, an, apple, with, knife, I, eat, you) であるとする。このような状況の下、パネルの数が 2 であるとする。文 1 では、単語 “apple” と当該単語の周辺の単語との関係を表す出現ベクトルとして (0, 1, 1, 0, 1, 0, 0, 0, 0) が生成される。文 2 および文 3 では、単語 “apple” と当該単語の周辺の単語との関係を表す出現ベクトルとして (0, 0, 1, 0, 0, 0, 0, 1, 0) が生成される。そして、文 1 と文 2 との間の出現ベクトルの類似度は、内積により $1/\sqrt{6}$ と算出される。文 1 と文 3 との間の出現ベクトルの類似度も、内積により $1/\sqrt{6}$ と算出される。文 2 と文 3 との間の出現ベクトルの類似度は、内積により 1 と算出される。そして、出現ベクトルの類似度の平均は、 $(1/\sqrt{6} + 1/\sqrt{6} + 1)/3$ の式により約 0.605 と算出される。出現ベクトルの類似度の分散は、約 0.0776 と算出される。ここでは、単語の多様度を平均とする場合には、出現ベクトルの類似度の平均 0.605 が単語 “apple” の多様度として測定される。単語の多様度を分散とする場合には、出現ベ

クトルの類似度の分散 0.0776 が単語 “apple” の多様度として測定される。

[0034] 第3の例として、単語多様度測定部11は、学習用コーパス21に含まれる単語について、当該単語が示す意味が複数存在する場合には、意味の数に基づいて、測定した多様度を補正する。これは、ある単語について同じ表記であっても別の意味を表す多義語が多い場合には、ある単語の多様度が上がりすぎるため、上がりすぎた多様度を補正するためである。

[0035] 具体的には、単語 “apple” の多様度を測定することとする。このような前提の下、単語 “apple” には、果物の「リンゴ」という意味と、米国企業「apple（登録商標）」という意味とがある。したがって、単語に意味が2つ存在すれば、意味が1つである場合よりも多様度が2倍に上がりやすいと推定されるため、上がりすぎた多様度が補正される。例えば、第1の例では、多様度が2であったが、2倍に推定された多様度を単純に $1/2$ にすべく、1の値に補正される。

[0036] このようにして、単語多様度測定部11は、単語の多様度を測定することができる。

[0037] 単語分類部12は、多様度が測定された単語を、第1単語群22および第2単語群23のいずれかに分類する。例えば、単語分類部12は、単語多様度測定部11によって測定された多様度が規定値より高い場合には、多様度が測定された単語を第1単語群22に分類する。すなわち、単語分類部12は、当該単語を、学習が十分にできる単語の集まりである第1単語群22に分類する。そして、単語分類部12は、単語多様度測定部11によって測定された多様度が規定値より低い場合には、多様度が測定された単語を第2単語群23に分類する。すなわち、単語分類部12は、当該単語を、学習が十分にできない単語の集まりである第2単語群23に分類する。

[0038] 学習振分部13は、第1単語群22を、知識を活用しない分散表現学習に振り分ける。例えば、学習振分部13は、第1単語群22を、分散表現学習部14に出力する。また、学習振分部13は、第2単語群23を、知識を活

用する分散表現学習に振り分ける。例えば、学習振分部 1 3 は、第 2 単語群 2 3 を、知識利用型分散表現学習部 1 5 に出力する。

[0039] 分散表現学習部 1 4 は、第 1 単語群 2 2 に含まれる単語の分散表現を、単語の分散表現を学習する技術を用いて学習する。例えば、分散表現学習部 1 4 は、学習用コーパス 2 1 を受け取ると、第 1 単語群 2 2 に含まれる単語の分散表現を、Word2Vec の Skip-gram モデルを用いて学習する。なお、Word2Vec の Skip-gram モデルでの学習は、既存の技術を使って行われる。かかる学習は、「Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.」に開示されている技術を使って行えば良い。また、「Xin Rong. word2vec Parameter Learning Explained」に開示されている技術を使って行えば良い。

[0040] ここで、分散表現学習部 1 4 によって行われる分散表現学習処理を、図 2 A～図 2 C を参照して説明する。図 2 A～図 2 C は、実施例に係る分散表現学習処理の一例を示す図である。図 2 A～図 2 C では、単語の分散表現を獲得する技術を Word2Vec の Skip-gram モデルとして説明する。なお、学習用コーパス 2 1 は、「I drink apple juice. . . .」であるとする。

[0041] 分散表現学習部 1 4 は、Skip-gram モデルに対して、Input layer (入力層)、Hidden layer (隠れ層) および Output layer (出力層) からなるニューラルネットワークを構築する。

[0042] まず、入力層、隠れ層および出力層について説明する。Input layer (入力層) には、与えられた単語に対応する Z 次元の入力ベクトル x が入力される。入力ベクトル x は、one-hot ベクトルである。Z は、学習用コーパス 2 1 に含まれる単語の数である。なお、one-hot ベクトルとは、与えられた単語に対応する要素を 1、他の要素を 0 とするベクトルのことをいう。

[0043] また、Hidden layer（隠れ層）には、最終的には、与えられた単語「apple」の分散表現であるM次元の単語ベクトル w が生成される。 $W_{Z \times M}$ は、入力層と隠れ層との間の重みであり、 $Z \times M$ の行列により表わされる。 $W_{Z \times M}$ の各要素の初期状態として、例えば、ランダムな値が与えられる。

[0044] また、Output layer（出力層）には、図示しないC個のパネルにそれぞれZ次元の出力ベクトル y_c が生成される。かかるCは、予め定められるパネル数であり、 y_c は、与えられた単語の前後にある単語に対応する出力ベクトルである。 $W'_{M \times Z}$ は、隠れ層と出力層との間の重みであり、 $M \times Z$ の行列により表わされる。 $W'_{M \times Z}$ の各要素の初期状態として、例えば、ランダムな値が与えられる。

[0045] 図2Aに示すように、このような入力層、隠れ層および出力層からなるニューラルネットワークを活用したSkip-gramモデルにより、分散表現学習部14は、与えられた単語の分散表現を学習する。例えば、入力ベクトル x が、第1単語群22の単語であって学習用コーパス21内の与えられた単語「apple」に対応する要素を1、他の要素を0とするone-hotベクトルであるとする。すると、分散表現学習部14は、与えられた単語「apple」に対応する入力ベクトル x を受け付けると、入力ベクトル x に $W_{Z \times M}$ の重みをかけて隠れ層の単語ベクトル w を生成する。そして、分散表現学習部14は、単語ベクトル w に $W'_{M \times Z}$ の重みをかけて出力層の出力ベクトル y を生成する。すなわち、分散表現学習部14は、初期状態の $W_{Z \times M}$ 、 $W'_{M \times Z}$ を使って予測する。この結果、単語「drink」である確率が0.1230、単語「juice」である確率が0.1277と予測される。

[0046] 次に、図2Bに示すように、分散表現学習部14は、実際に算出された出力ベクトル y が予め定められた予測ベクトルと異なれば、その差分に基づいて $W'_{M \times Z}$ および $W_{Z \times M}$ の順にパラメータである重みを更新する。このようなパラメータの更新は、例えば、バックプロパゲーションという。そして、分散表現学習部14は、入力ベクトル x に、更新された $W_{Z \times M}$ の重みをかけて隠れ層の単語ベクトル w を生成する。そして、分散表現学習部14は、単語ベ

クトル w に、更新された $W'_{M \times Z}$ の重みをかけて出力層の出力ベクトル y を生成する。すなわち、分散表現学習部14は、更新状態の $W_{Z \times M}$ 、 $W'_{M \times Z}$ を使って予測する。この結果、単語「drink」である確率が0.1236、単語「juice」である確率が0.1289と予測され、前回の予測値より微増している。

[0047] 次に、図2Cに示すように、分散表現学習部14は、バックプロパゲーションのアルゴリズムを用いて、出力ベクトルが最適になるまで繰り返す（イテレーションする）。出力ベクトルが最適であるとは、与えられた単語の周辺の各単語における出現確率の同時確率が最大になることである。すなわち、与えられた単語の周辺の各単語における出現確率が、同時に1.0に近似することである。例えば、分散表現学習部14は、予め定められた回数だけ繰り返すようにすれば良い。予め定められた回数は、一例として、500回であるが、最適となる回数を実験によって求めれば良い。この結果、分散表現学習部14は、隠れ層に現れる単語ベクトル w を、与えられた単語「apple」の分散表現として獲得する。

[0048] 同様に、分散表現学習部14は、第1単語群22の与えられる単語を替えることで、与えられる単語の分散表現を獲得すれば良い。

[0049] 図1に戻って、知識利用型分散表現学習部15は、学習対象が第1単語群22である場合には、第1単語群22に含まれる単語の分散表現を、知識利用型の単語の分散表現を学習する技術を用いて補正する。知識利用型の単語の分散表現を学習する技術は、例えば、RC-NETである。例えば、知識利用型分散表現学習部15は、RC-NETにおけるR-NETに第1単語群22の単語が属している場合には、属している単語について分散表現学習部14によって学習済みの単語の分散表現をR-NETで用いるように設定する。知識利用型分散表現学習部15は、RC-NETにおけるC-NETに第1単語群22の単語が属している場合には、属している単語について分散表現学習部14によって学習済みの単語の分散表現をC-NETで用いるように設定する。そして、知識利用型分散表現学習部15は、学習用コーパ

ス 2 1 を受け取ると、第 1 単語群 2 2 に含まれる単語の分散表現を、R C - N E T を用いて補正する。一例として、知識利用型分散表現学習部 1 5 は、第 1 単語群 2 2 に含まれる単語について、学習済みの単語の分散表現を隠れ層の単語ベクトル w に設定し、学習済みの $W_{Z \times M}$ の重みおよび $W'_{M \times Z}$ の重みを初期状態として設定し、R C - N E T を用いて学習すれば良い。

[0050] また、知識利用型分散表現学習部 1 5 は、学習対象が第 2 単語群 2 3 である場合には、第 2 単語群 2 3 に含まれる単語の分散表現を、知識利用型の単語の分散表現を学習する技術を用いて学習する。知識利用型の単語の分散表現を学習する技術は、例えば、R C - N E T である。例えば、知識利用型分散表現学習部 1 5 は、R C - N E T における R - N E T に第 1 単語群 2 2 の単語が属している場合には、属している単語について分散表現学習部 1 4 によって学習済みの単語の分散表現を R - N E T で用いるように設定する。知識利用型分散表現学習部 1 5 は、R C - N E T における C - N E T に第 1 単語群 2 2 の単語が属している場合には、属している単語について分散表現学習部 1 4 によって学習済みの単語の分散表現を C - N E T で用いるように設定する。そして、知識利用型分散表現学習部 1 5 は、学習用コーパス 2 1 を受け取ると、第 2 単語群 2 3 に含まれる単語の分散表現を、R C - N E T を用いて学習する。一例として、知識利用型分散表現学習部 1 5 は、第 2 単語群 2 3 に含まれる単語について、分散表現学習部 1 4 によって学習済みの単語の分散表現がないため、以下のように学習する。すなわち、知識利用型分散表現学習部 1 5 は、当該単語に対応する要素を 1、他の要素を 0 とする $one-hot$ ベクトルを入力ベクトル x とし、ランダムな $W_{Z \times M}$ の重みおよびランダムな $W'_{M \times Z}$ の重みを初期状態として設定し、R C - N E T を用いて学習すれば良い。

[0051] 学習終了判定部 1 6 は、学習の終了を判定する。ここでいう終了を判定する学習とは、分散表現学習部 1 4 による学習および知識利用型分散表現学習部 1 5 による学習のことをいう。

[0052] 例えば、学習終了判定部 1 6 は、学習するたびに更新される単語ベクトル

wの更新前後の差が閾値より小さくなった場合には、学習を終了する。単語ベクトルwの更新前とは、1回学習した後の単語ベクトルである。単語ベクトルwの更新後とは、1回学習してバックプロパゲーションした後に次回学習した後の単語ベクトルである。すなわち、学習の終了条件は、式(1)で示される。なお、 w_{kold} は、単語ベクトル w_k の更新前である。 w_{knew} は、単語ベクトル w_k の更新後である。kは、隠れ層の単語ベクトルのk番目の要素を指すインデックスである。 ε は、閾値である。閾値は、更新前後の差が十分小さいと判断できる値であれば良い。

$$w_{kold} - w_{knew} < \varepsilon \quad \cdots \text{式(1)}$$

[0053] そして、学習終了判定部16は、単語ベクトルwの更新前後の差が閾値以上である場合には、学習を繰り返す。

[0054] なお、学習終了判定部16は、学習の終了条件として単語ベクトルwの更新前後の差が閾値より小さくなった場合であると説明したが、これに限定されない。学習終了判定部16は、所定の繰り返し(イテレーション)回数分繰り返した場合であっても良い。

[0055] ここで、知識利用型分散表現学習部15によって行われる知識利用型分散表現学習処理を、図3を参照して説明する。図3は、実施例に係る知識利用型分散表現学習処理の一例を示す図である。図3では、知識利用型の単語の分散表現を学習する技術を、RC-NETとして説明する。なお、知識情報24の関係知識情報には、関係の知識で用いられる単語と分散表現とが予め対応づけられているとする。知識情報24のカテゴリ知識情報には、カテゴリの知識で用いられる単語と分散表現とが予め対応づけられているとする。

[0056] 知識利用型分散表現学習部15は、Skip-gramをベースにR-NETおよびC-NETを拡張したニューラルネットワークを構築する。なお、Skip-gramは、分散表現学習部14で説明したので、その説明を省略する。

[0057] まず、知識利用型分散表現学習部15は、知識情報24の関係知識情報に第1単語群22の単語が属している場合には、属している単語に当該単語の

学習済みの分散表現を対応づけて設定する。知識利用型分散表現学習部 15 は、知識情報 24 のカテゴリ知識情報に第 1 単語群 22 の単語が属している場合には、属している単語に当該単語の学習済みの分散表現を対応づけて設定する。

[0058] また、知識利用型分散表現学習部 15 は、Skip-gram に、与えられた単語に対応する入力ベクトル x が入力されたとする。すると、知識利用型分散表現学習部 15 は、与えられた単語に対応する関係の埋め込み (Relation embedding) r と関係先の単語の分散表現 $w_{k, r}$ とを知識情報 24 から取得する。一例として、与えられた単語が「George 6 世」である場合に、関係の埋め込みとして「King__of」、関係先の単語の分散表現として「United Kingdom」の分散表現が知識情報 24 内の関係知識情報から取得される。ここで、関係先の単語が第 1 単語群 22 に属している場合には、関係先の単語の分散表現として学習済みの分散表現が用いられることになる。

[0059] そして、知識利用型分散表現学習部 15 は、与えられた単語の分散表現 w_k と関係の埋め込み r を用いて $w'_{k, r}$ を求め、求められた $w'_{k, r}$ が関係先の単語の分散表現 $w_{k, r}$ と距離的に近いように学習する。すなわち、知識利用型分散表現学習部 15 は、与えられた単語が関係先の単語と距離的に近いように学習する。そして、知識利用型分散表現学習部 15 は、 $w'_{k, r}$ と、関係先の単語の分散表現 $w_{k, r}$ との距離 $d(w_{k, r}, w'_{k, r})$ が予め定められた距離以上であれば、収束していないので、パラメータである重み W' および W を更新する。パラメータの更新は、例えば、バックプロパゲーションを用いれば良い。

[0060] 加えて、知識利用型分散表現学習部 15 は、Skip-gram に、与えられた単語に対応する入力ベクトル x が入力されたとする。すると、知識利用型分散表現学習部 15 は、与えられた単語と同じカテゴリに属する単語の分散表現を知識情報 24 から取得する。一例として、与えられた単語が「George 6 世」である場合に、当該単語と同じカテゴリ「Male」に属

する単語「Prince Philip」「Charles」のそれぞれの分散表現が知識情報24内のカテゴリ知識情報から取得される。ここで、同じカテゴリに属する単語が第1単語群22に属している場合には、同じカテゴリに属する単語の分散表現として学習済みの分散表現が用いられることになる。

[0061] そして、知識利用型分散表現学習部15は、与えられた単語の分散表現 w_k と、与えられた単語と同じカテゴリに属する単語の分散表現 $w_1, w_2, \dots, w_v$ とそれぞれ距離的に近いように学習する。そして、知識利用型分散表現学習部15は、 w_k と、 $w_1, w_2, \dots, w_v$ とのそれぞれの距離 $d(w_k, w_1), d(w_k, w_2), \dots, d(w_k, w_v)$ のいずれかが予め定められた距離以上であれば、収束していないので、パラメータである重み W' および W を更新する。パラメータの更新は、例えば、バックプロパゲーションを用いれば良い。

[0062] なお、学習対象の単語（与えられた単語）が第1単語群22である場合には、知識利用型分散表現学習部15は、RC-NETのSkip-gramに初期状態として、以下を設定すれば良い。すなわち、知識利用型分散表現学習部15は、与えられた単語に対応する要素を1、他の要素を0とするone-hotベクトルを入力ベクトル x として設定する。知識利用型分散表現学習部15は、隠れ層に分散表現学習部14により学習された学習済みの分散表現 w_k を設定する。知識利用型分散表現学習部15は、重みに学習済みの $W_{Z \times M}$ および $W'_{M \times Z}$ を設定すれば良い。

[0063] 学習対象の単語（与えられた単語）が第2単語群23である場合には、知識利用型分散表現学習部15は、RC-NETのSkip-gramに初期状態として、以下を設定すれば良い。すなわち、知識利用型分散表現学習部15は、与えられた単語に対応する要素を1、他の要素を0とするone-hotベクトルを入力ベクトル x として設定する。知識利用型分散表現学習部15は、重みにランダムな $W_{Z \times M}$ および $W'_{M \times Z}$ を設定すれば良い。

[0064] [分析処理のフローチャート]

図4は、実施例に係る分析処理のフローチャートの一例を示す図である。

[0065] 図4に示すように、分析部10は、分析処理要求を受け付けたか否かを判定する（ステップS10）。分析処理要求を受け付けていないと判定した場合には（ステップS10；No）、分析部10は、分析処理要求を受け付けるまで、判定処理を繰り返す。

[0066] 一方、分析処理要求を受け付けたと判定した場合には（ステップS10；Yes）、単語多様度測定部11は、学習用コーパス21に含まれる単語の多様度を測定する（ステップS20）。例えば、単語多様度測定部11は、学習用コーパス21に含まれる単語について、当該単語と、当該単語の前方のN個の単語および当該単語の後方のN個の単語との関係を示す出現ベクトルを生成する。そして、単語多様度測定部11は、生成した出現ベクトルの種類数に基づいて、当該単語の多様度を測定する。ここでいうNは、Skip-gramモデルの出力層で用いられるパラメータであるパネルのサイズである。なお、単語多様度測定部11は、単語の多様度を当該単語の出現ベクトルの種類数と説明したが、これに限定されず、当該単語の多様度を当該単語の出現ベクトルの類似度の平均や分散の値としても良い。

[0067] 単語分類部12は、測定された多様度に基づいて、各単語を第1単語群22および第2の単語群のいずれかに分類する（ステップS30）。例えば、単語分類部12は、測定された多様度が規定値より高い場合には、この多様度の単語を第1単語群22に分類する。単語分類部12は、測定された多様度が規定値より低い場合には、この多様度の単語を第2単語群23に分類する。

[0068] そして、分散表現学習部14は、第1単語群22に含まれる各単語について、それぞれ分散表現を学習する（ステップS40）。例えば、分散表現学習部14は、第1単語群22に含まれる各単語の分散表現を、Word2VecのSkip-gramモデルを用いて学習する。

[0069] そして、学習終了判定部16は、学習終了条件を満たすか否かを判定する（ステップS50）。例えば、学習終了判定部16は、単語ベクトルwの更

新前後の差が閾値より小さくなったか否かを判定する。ここでいう単語ベクトル w の更新前とは、1回学習した後の単語ベクトルである。単語ベクトル w の更新後とは、1回学習してバックプロパゲーションした後に次回学習した後の単語ベクトルである。学習終了条件を満たさないと判定した場合には（ステップS50；No）、学習終了判定部16は、さらに学習すべく、ステップS40に移行する。

[0070] 一方、学習終了条件を満たすと判定した場合には（ステップS50；Yes）、学習終了判定部16は、知識を活用した学習に遷移すべく、ステップS60に移行する。

[0071] ステップS60において、知識利用型分散表現学習部15は、RC-NE TにおけるR-NE TおよびC-NE Tに第1単語群の分散表現を設定する（ステップS60）。例えば、知識利用型分散表現学習部15は、RC-NE TにおけるR-NE Tに第1単語群22の単語が属している場合には、属している単語について学習済みの単語の分散表現をR-NE Tで用いるように設定する。知識利用型分散表現学習部15は、RC-NE TにおけるC-NE Tに第1単語群22の単語が属している場合には、属している単語について学習済みの単語の分散表現をC-NE Tで用いるように設定する。

[0072] そして、知識利用型分散表現学習部15は、第1単語群22に含まれる各単語について、それぞれ知識を利用して分散表現を学習する（ステップS70）。例えば、知識利用型分散表現学習部15は、第1単語群22に含まれる各単語の学習済みの分散表現を、さらにRC-NE Tを用いて補正する。

[0073] そして、学習終了判定部16は、学習終了条件を満たすか否かを判定する（ステップS80）。学習終了条件を満たさないと判定した場合には（ステップS80；No）、学習終了判定部16は、さらに学習すべく、ステップS70に移行する。

[0074] 一方、学習終了条件を満たすと判定した場合には（ステップS80；Yes）、学習終了判定部16は、第2単語群23の学習に遷移すべく、ステップS90に移行する。

- [0075] ステップS90において、知識利用型分散表現学習部15は、RC-NE TにおけるR-NE TおよびC-NE Tに第1単語群22の分散表現を設定する（ステップS90）。例えば、知識利用型分散表現学習部15は、RC-NE TにおけるR-NE Tに第1単語群22の単語が属している場合には、属している単語について学習済みの単語の分散表現をR-NE Tで用いるように設定する。知識利用型分散表現学習部15は、RC-NE TにおけるC-NE Tに第1単語群22の単語が属している場合には、属している単語について学習済みの単語の分散表現をC-NE Tで用いるように設定する。
- [0076] そして、知識利用型分散表現学習部15は、第2単語群23に含まれる各単語について、それぞれ知識を利用して分散表現を学習する（ステップS100）。例えば、知識利用型分散表現学習部15は、第2単語群23に含まれる各単語の分散表現を、RC-NE Tを用いて学習する。
- [0077] そして、学習終了判定部16は、学習終了条件を満たすか否かを判定する（ステップS110）。学習終了条件を満たさないと判定した場合には（ステップS110；No）、学習終了判定部16は、さらに学習すべく、ステップS100に移行する。
- [0078] 一方、学習終了条件を満たすと判定した場合には（ステップS110；Yes）、学習終了判定部16は、分析処理を終了する。これにより、分析部10は、十分に学習した第1単語群22の単語と単語間の関係などの知識とを活用することで、十分に学習できない第2単語群23の単語を高精度に学習することが可能となる。
- [0079] なお、実施例に係る分析処理は、単語の分散表現を獲得することで、単語の意味を解釈できるようになり、文書検索や文の類似関係の検索を、単語の分散表現を用いることで高精度に行うことができるようになる。また、実施例に係る分析処理は、Q/A（Question/Answer）サイトに適用されることで、過去の膨大な問合せ（Question）の中から適切な回答（Answer）を、単語の分散表現を用いて高精度に検索することができるようになる。
- [0080] [実施例の効果]

上記実施例によれば、情報処理装置 1 は、学習で用いられる学習用コーパス 2 1 に含まれる各単語の多様度を測定し、各単語を多様度が規定値より高い第 1 単語群 2 2 および規定値より低い第 2 単語群 2 3 に分類する。情報処理装置 1 は、第 1 単語群 2 2 に含まれる単語の分散表現を、分散表現を学習する第 1 の学習アルゴリズムを用いて学習する。情報処理装置 1 は、分散表現を学習する第 2 の学習アルゴリズムにより用いられる単語が第 1 単語群 2 2 に含まれる場合には、該含まれる単語の分散表現を第 2 の学習アルゴリズムに設定する。そして、情報処理装置 1 は、第 2 の学習アルゴリズムを用いて、第 1 単語群 2 2 に含まれる単語の分散表現を補正する。かかる構成によれば、情報処理装置 1 は、多様度が規定値より高い単語の分散表現を第 2 の学習アルゴリズムで活用することで、多様度が規定値より低い単語であっても分散表現の学習を高精度に行うことができる。すなわち、情報処理装置 1 は、十分に学習した単語の分散表現を第 2 の学習アルゴリズムである例えば R C - N E T で活用することで、十分に学習できない単語の分散表現の学習を高精度に行うことができる。

[0081] また、上記実施例によれば、情報処理装置 1 は、第 1 の学習アルゴリズムを用いて学習された単語の分散表現を、第 2 の学習アルゴリズムを用いて補正する。かかる構成によれば、情報処理装置 1 は、既に学習された単語の分散表現を、多様度が規定値より高い単語の分散表現を活用した第 2 の学習アルゴリズムを用いてさらに学習することで、既に学習された単語の分散表現の学習を高精度に行うことができる。

[0082] また、上記実施例によれば、情報処理装置 1 は、学習用コーパス 2 1 に含まれる単語について、当該単語が出現する文における当該単語と周辺の単語との共起関係を表わす出現ベクトルを算出する。情報処理装置 1 は、算出した出現ベクトルの数に基づいて、当該単語の多様度を測定する。そして、情報処理装置 1 は、測定した多様度に基づいて、当該単語を第 1 単語群 2 2 および第 2 単語群 2 3 に分類する。かかる構成によれば、情報処理装置 1 は、学習用コーパス 2 1 に含まれる単語を、出現ベクトルの数に基づいた多様度

に応じて分類することができる。

[0083] また、上記実施例によれば、情報処理装置 1 は、学習用コーパス 2 1 に含まれる単語について、当該単語が出現する文における当該単語と周辺の単語との共起関係を表わす出現ベクトルを算出する。情報処理装置 1 は、算出した出現ベクトルが複数存在する場合には、算出した出現ベクトルを用いて出現ベクトルの類似度を算出する。情報処理装置 1 は、算出した出現ベクトルの類似度に基づいて、当該単語の多様度を測定する。そして、情報処理装置 1 は、測定した多様度に基づいて、当該単語を第 1 単語群 2 2 および第 2 単語群 2 3 に分類する。かかる構成によれば、情報処理装置 1 は、学習用コーパス 2 1 に含まれる単語を、文の類似度に基づいた多様度に応じて分類することができる。

[0084] また、上記実施例によれば、情報処理装置 1 は、学習用コーパス 2 1 に含まれる単語について、当該単語が示す意味が予め定められた辞書に複数存在する場合には、意味の数に基づいて、該測定した多様度を補正する。かかる構成によれば、情報処理装置 1 は、意味の数に基づいて多様度を補正することで、多様度に応じた分類を精度良く行うことができる。

[0085] [その他]

なお、上記実施例では、情報処理装置 1 は、RC-NET や Skip-gram を用いて、知識を活用した分散表現学習について、学習で用いられる文書に含まれる単語の多様性を考慮した単語の分散表現を学習することについて説明した。しかしながら、RC-NET や Skip-gram は、一例であって、これに限定されない。例えば、情報処理装置 1 は、Skip-gram に代えて Continuous-Bag-of-Words モデルや、非特許文献 3 を発端とするニューラルランゲージモデルの派生形に適用することも可能である。また、例えば、情報処理装置 1 は、RC-NET における R-NET の知識と C-NET の知識を使って単語ベクトル間の距離を小さくなるように処理するが、これに限定されず、単語ベクトルの内積を -1 になるように処理することも可能である。

[0086] また、図示した装置の各構成要素は、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、装置の分散・統合の具体的態様は図示のものに限られず、その全部または一部を、各種の負荷や使用状況などに応じて、任意の単位で機能的または物理的に分散・統合して構成することができる。例えば、知識利用型分散表現学習部 15 は、第 1 単語群 22 の分散表現を補正する補正学習部と、第 2 単語群 23 の分散表現を学習する学習部とに分散しても良い。また、学習終了判定部 16 を分散表現学習部 14 と併合しても良いし、学習終了判定部 16 を知識利用型分散表現学習部 15 と併合しても良い。また、記憶部 20 を情報処理装置 1 の外部装置としてネットワーク経由で接続するようにしても良い。

[0087] 図 5 は、情報処理装置のハードウェア構成の一例を示す図である。図 5 に示すように、コンピュータ 500 は、各種演算処理を実行する CPU 501 と、ユーザからのデータ入力を受け付ける入力装置 502 と、モニタ 503 とを有する。また、コンピュータ 500 は、記憶媒体からプログラムなどを読み取る媒体読取装置 504 と、他の装置と接続するためのインターフェース装置 505 と、他の装置と無線により接続するための無線通信装置 506 とを有する。また、コンピュータ 500 は、各種情報を一時記憶する RAM (Random Access Memory) 507 と、ハードディスク装置 508 とを有する。また、各装置 501 ~ 508 は、バス 509 に接続される。

[0088] ハードディスク装置 508 には、図 1 に示した分析部 10 と同様の機能を有する分析プログラムが記憶される。また、ハードディスク装置 508 には、分析プログラムを実現するための各種データが記憶される。各種データには、図 1 に示した記憶部 20 内のデータが含まれる。

[0089] CPU 501 は、ハードディスク装置 508 に記憶された各プログラムを読み出して、RAM 507 に展開して実行することで、各種の処理を行う。これらのプログラムは、コンピュータ 500 を図 1 に示した各機能部として機能させることができる。

[0090] なお、上記の分析プログラムは、必ずしもハードディスク装置 508 に記

憶されている必要はない。例えば、コンピュータ 500 が読み取り可能な記憶媒体に記憶されたプログラムを、コンピュータ 500 が読み出して実行するようにしてもよい。コンピュータ 500 が読み取り可能な記憶媒体は、例えば、CD-ROM や DVD ディスク、USB (Universal Serial Bus) メモリなどの可搬型記録媒体、フラッシュメモリなどの半導体メモリ、ハードディスクドライブなどが対応する。また、公衆回線、インターネット、LAN (Local Area Network) などに接続された装置にこれらのプログラムを記憶させておき、コンピュータ 500 がこれらのプログラムを読み出して実行するようにしても良い。

符号の説明

- [0091] 1 情報処理装置
 - 10 分析部
 - 11 単語多様度測定部
 - 12 単語分類部
 - 13 学習振分部
 - 14 分散表現学習部
 - 15 知識利用型分散表現学習部
 - 16 学習終了判定部
 - 20 記憶部
 - 21 学習用コーパス
 - 22 第1単語群
 - 23 第2単語群
 - 24 知識情報

請求の範囲

[請求項1]

コンピュータに、

学習で用いられる文書に含まれる各単語の多様度を測定し、前記各単語を前記多様度が所定の規定値より高い第1の単語群および前記規定値より低い第2の単語群に分類し、

前記第1の単語群に含まれる単語の分散表現を、前記分散表現を学習する第1の学習アルゴリズムを用いて学習し、

前記分散表現を学習する第2の学習アルゴリズムにより用いられる単語が前記第1の単語群に含まれる場合には、該含まれる単語の分散表現を前記第2の学習アルゴリズムに設定し、前記第2の学習アルゴリズムを用いて、前記第2の単語群に含まれる単語の分散表現を学習する

処理を実行させることを特徴とする分析プログラム。

[請求項2]

前記第1の学習アルゴリズムを用いて学習された単語の分散表現を、前記第2の学習アルゴリズムを用いて補正する

ことを特徴とする請求項1に記載の分析プログラム。

[請求項3]

前記分類する処理は、

前記文書に含まれる単語について、当該単語が出現する文における当該単語と周辺の単語との共起関係を表わす出現ベクトルを算出し、

算出した出現ベクトルの数に基づいて、当該単語の多様度を測定し、

測定した多様度に基づいて、当該単語を前記第1の単語群および前記第2の単語群に分類する

ことを特徴とする請求項1に記載の分析プログラム。

[請求項4]

前記分類する処理は、

前記文書に含まれる単語について、当該単語が出現する文における当該単語と周辺の単語との共起関係を表わす出現ベクトルを算出し、

算出した出現ベクトルが複数存在する場合には、算出した出現ベク

トルを用いて出現ベクトルの類似度を算出し、

算出した出現ベクトルの類似度に基づいて、当該単語の多様度を測定し、

測定した多様度に基づいて、当該単語を前記第 1 の単語群および前記第 2 の単語群に分類する

ことを特徴とする請求項 1 に記載の分析プログラム。

[請求項5]

前記分類する処理は、

前記文書に含まれる単語について、当該単語が示す意味が予め定められた辞書に複数存在する場合には、意味の数に基づいて、該測定した多様度を補正する

ことを特徴とする請求項 3 または請求項 4 に記載の分析プログラム

。

[請求項6]

学習で用いられる文書に含まれる各単語の多様度を測定し、前記各単語を前記多様度が所定の規定値より高い第 1 の単語群および前記規定値より低い第 2 の単語群に分類する分類部と、

前記第 1 の単語群に含まれる単語の分散表現を、前記分散表現を学習する第 1 の学習アルゴリズムを用いて学習する第 1 の学習部と、

前記分散表現を学習する第 2 の学習アルゴリズムにより用いられる単語が前記第 1 の単語群に含まれる場合には、該含まれる単語の分散表現を前記第 2 の学習アルゴリズムに設定し、前記第 2 の学習アルゴリズムを用いて、前記第 2 の単語群に含まれる単語の分散表現を学習する第 2 の学習部と、

を有することを特徴とする情報処理装置。

[請求項7]

コンピュータが、

学習で用いられる文書に含まれる各単語の多様度を測定し、前記各単語を前記多様度が所定の規定値より高い第 1 の単語群および前記規定値より低い第 2 の単語群に分類し、

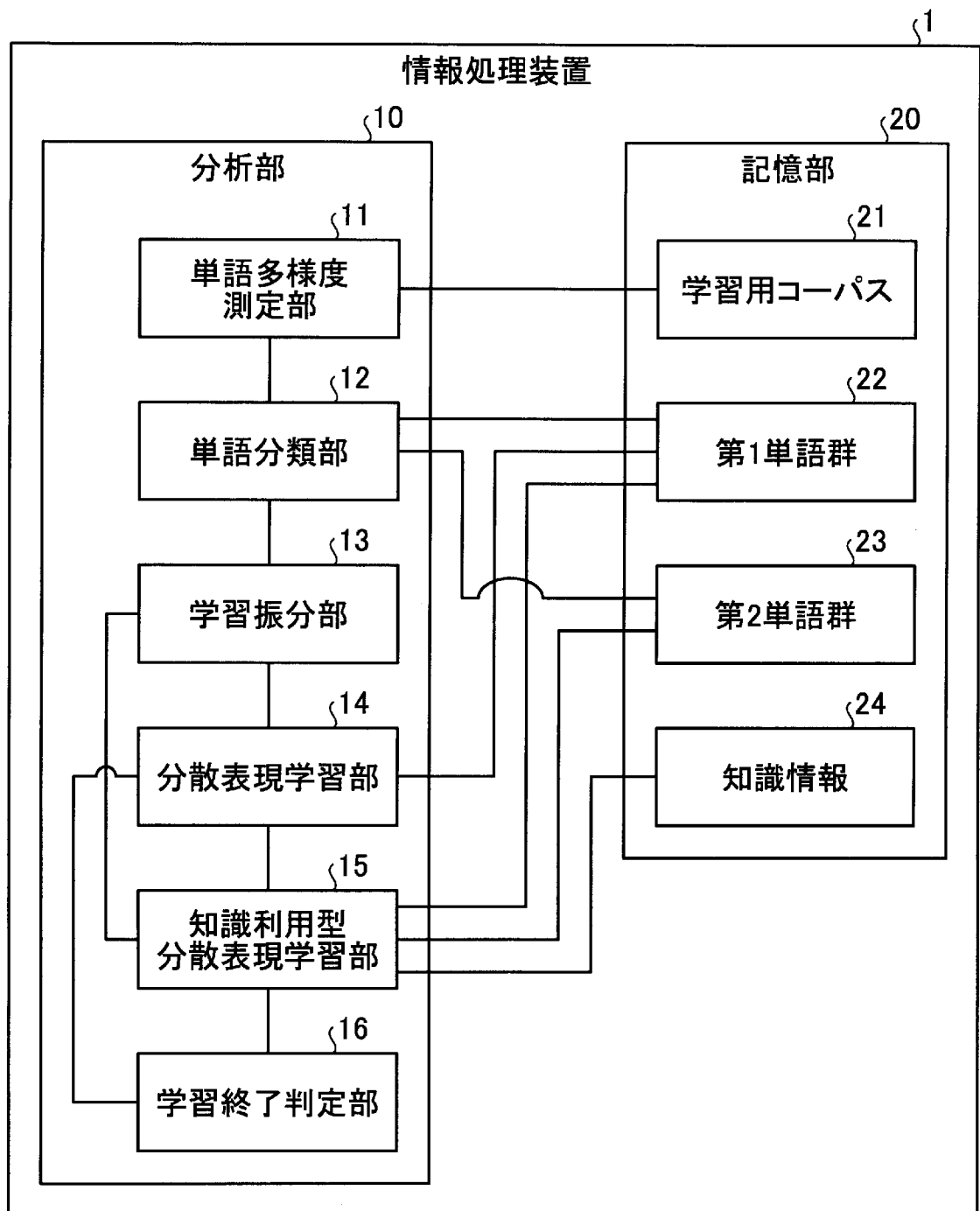
前記第 1 の単語群に含まれる単語の分散表現を、前記分散表現を学

習する第 1 の学習アルゴリズムを用いて学習し、

前記分散表現を学習する第 2 の学習アルゴリズムにより用いられる単語が前記第 1 の単語群に含まれる場合には、該含まれる単語の分散表現を前記第 2 の学習アルゴリズムに設定し、前記第 2 の学習アルゴリズムを用いて、前記第 2 の単語群に含まれる単語の分散表現を学習する

各処理を実行することを特徴とする分析方法。

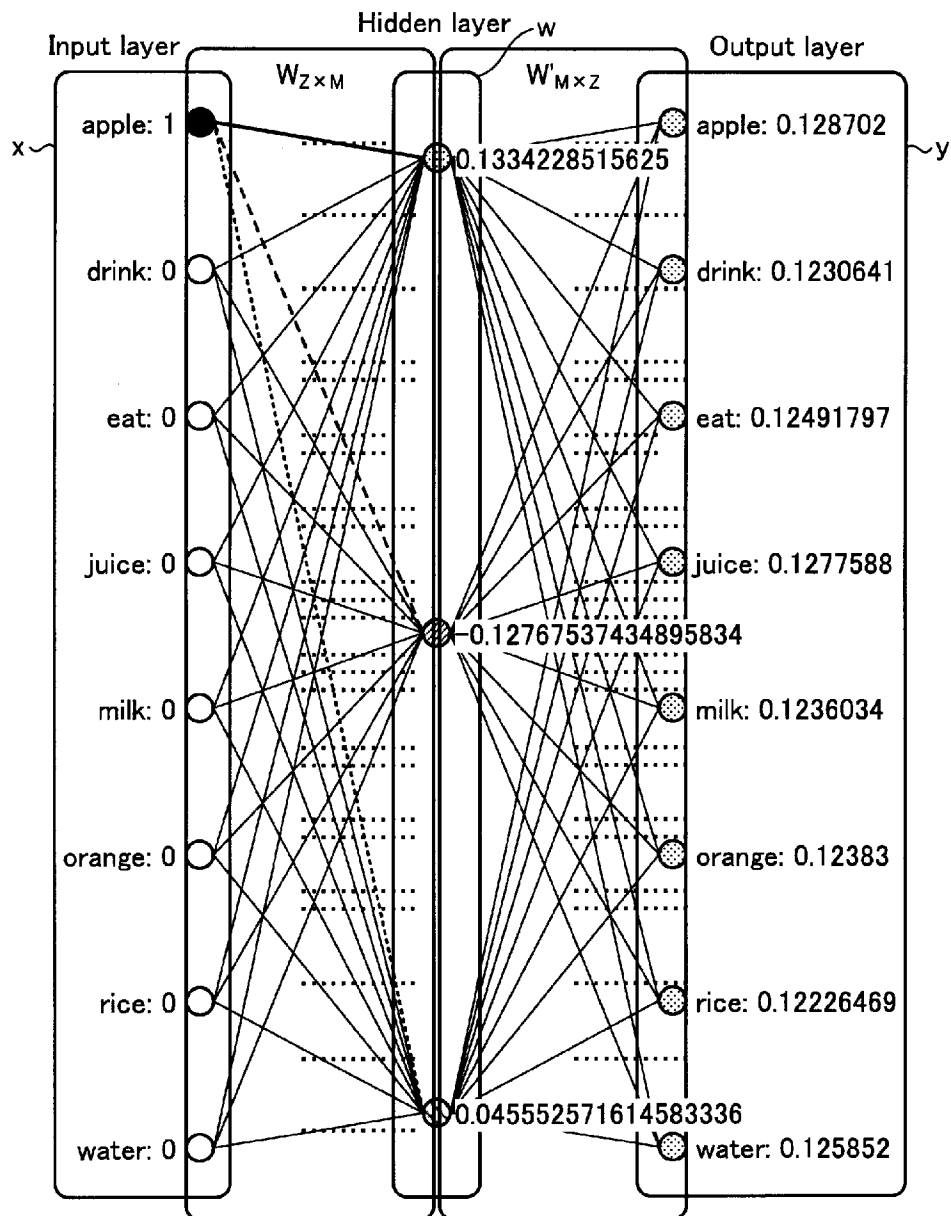
[図1]



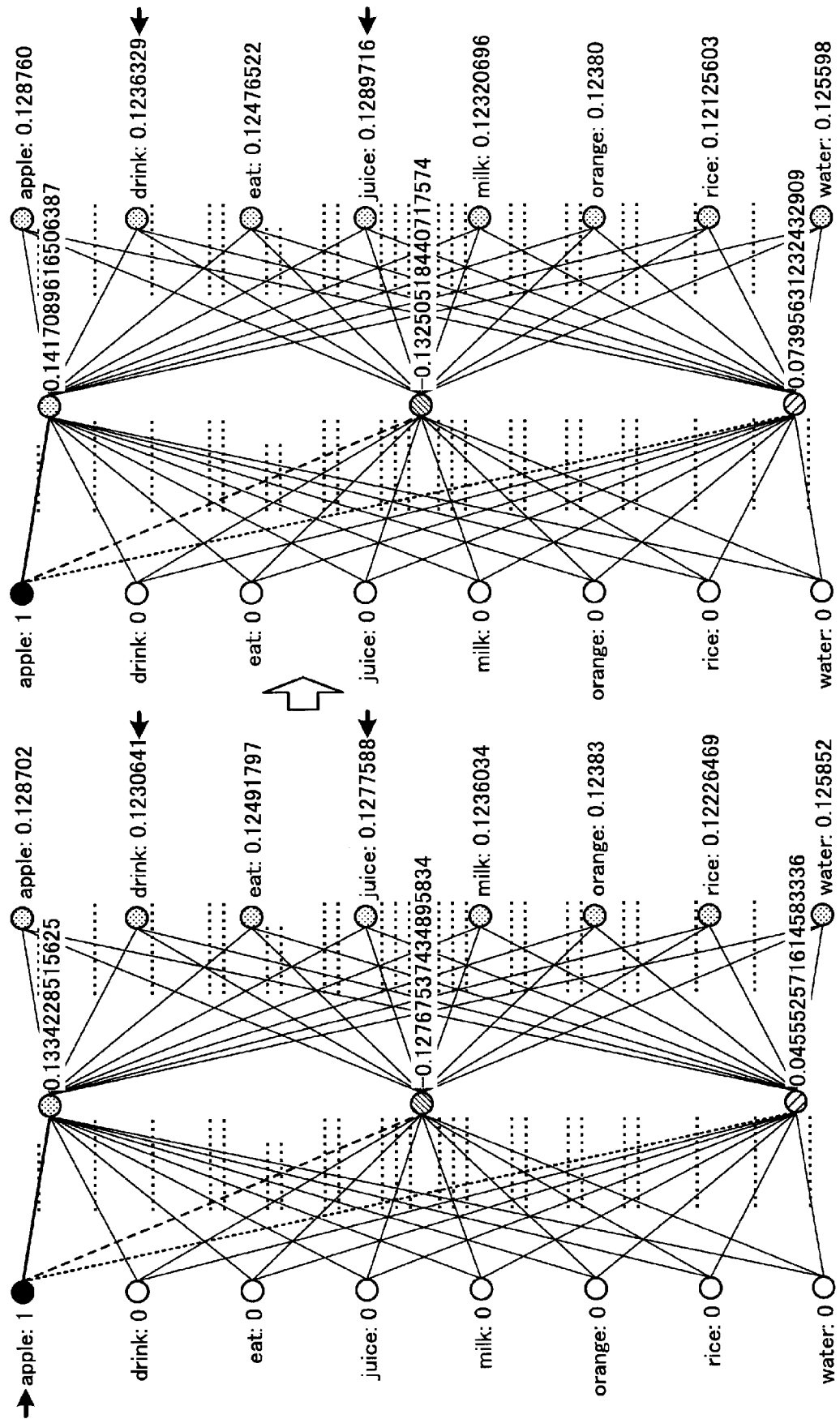
[図2A]

学習用コーパス:

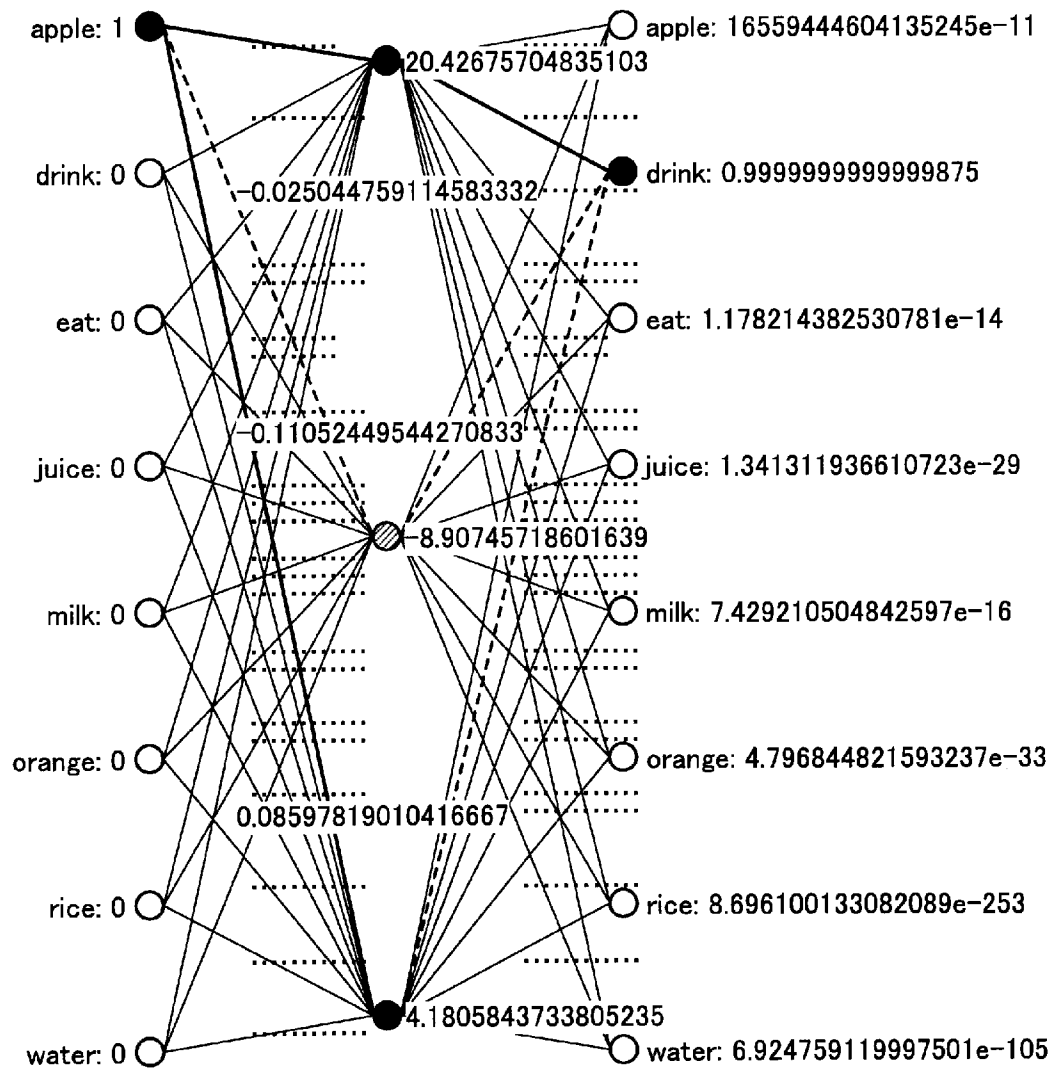
... I drink apple juice. ...



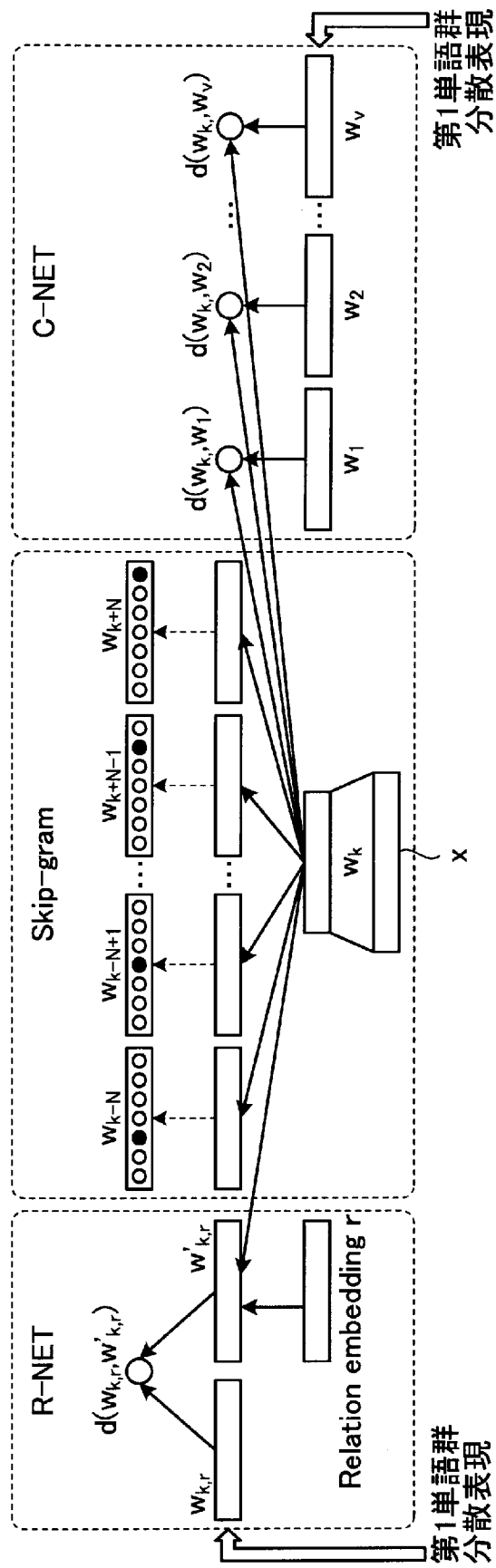
[図2B]



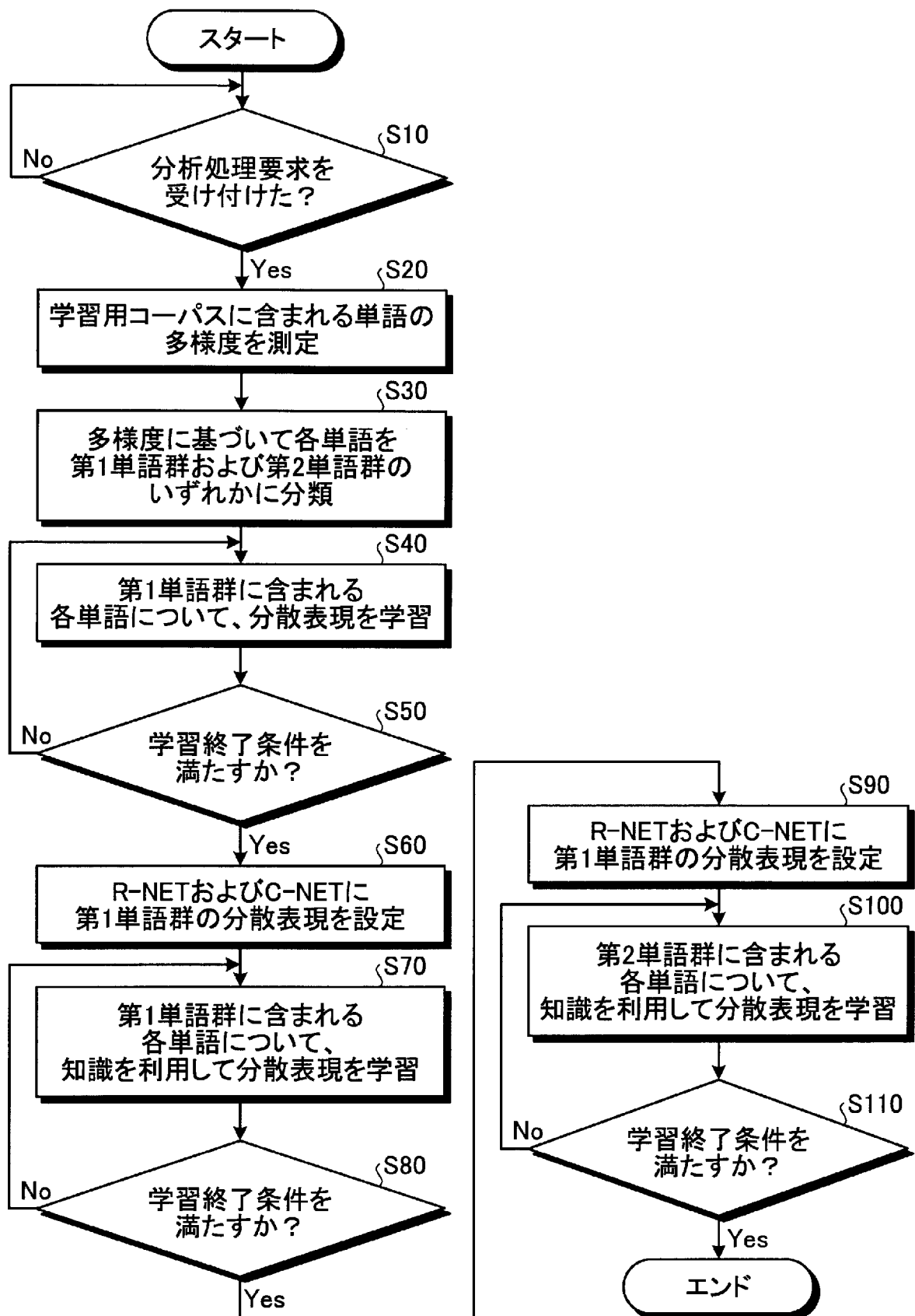
[図2C]



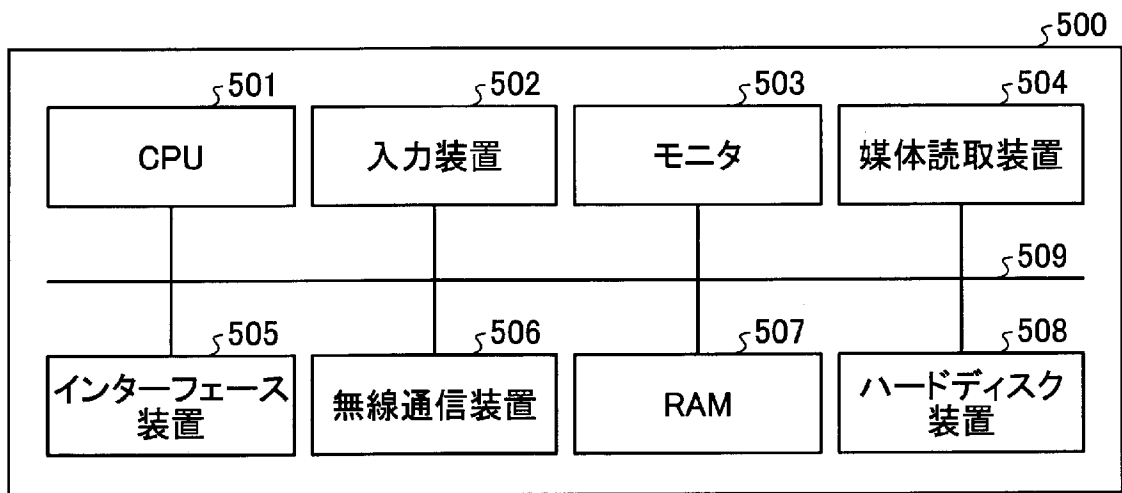
[図3]



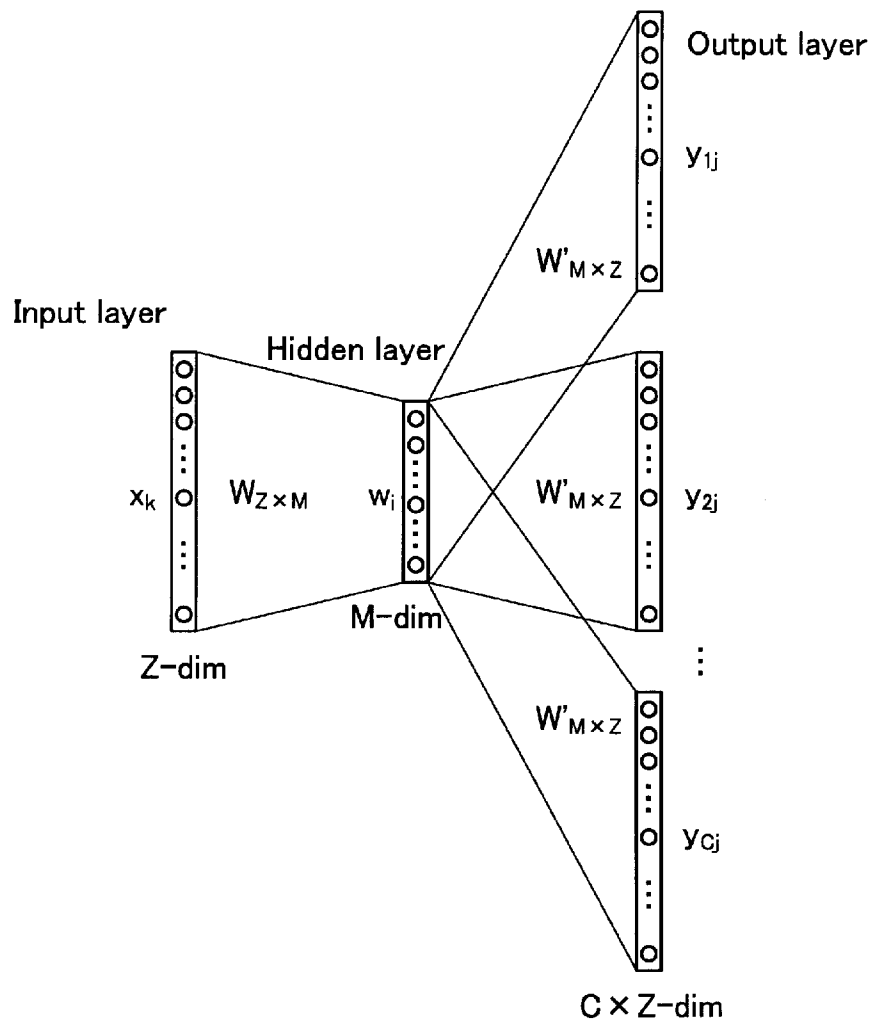
[図4]



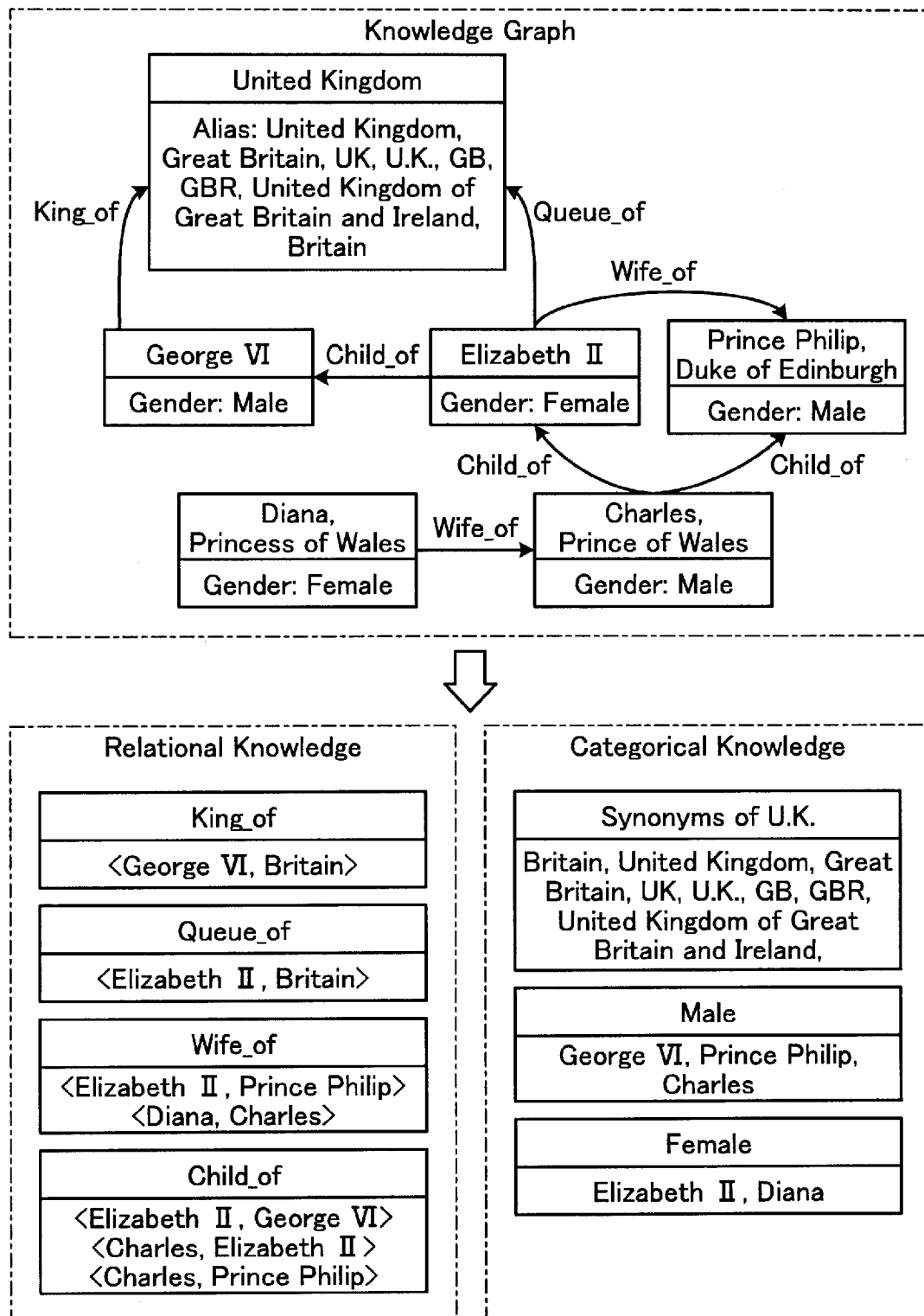
[図5]



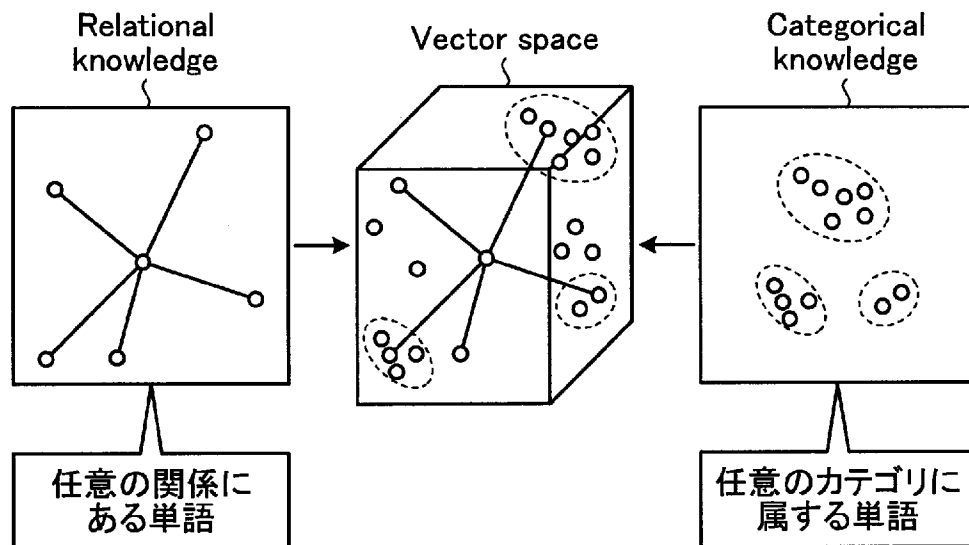
[図6]



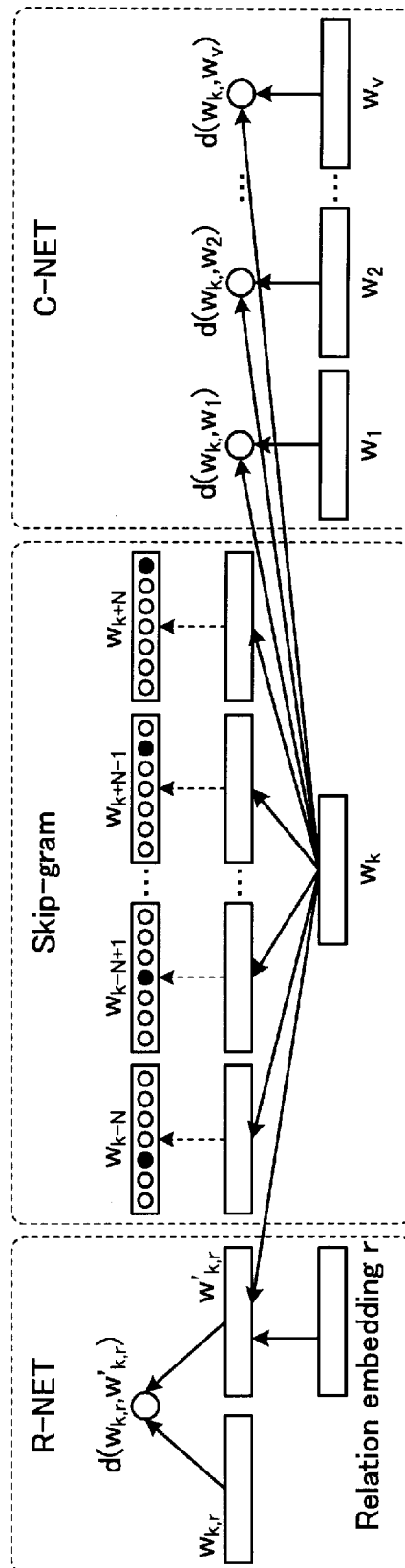
[図7]



[図8]



[図9]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2016/083000

A. CLASSIFICATION OF SUBJECT MATTER G06F17/30(2006.01) i		
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) G06F17/30		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2016 Kokai Jitsuyo Shinan Koho 1971-2016 Toroku Jitsuyo Shinan Koho 1994-2016		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) Google Scholar		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	LI, Qi, et al., Learning Word Sense Embeddings from Word Sense Definitions, arXiv preprint arXiv:1606.04835 [online], 2016, [retrieval date 2016.12.05], Internet: <URL: http://arxiv.org/abs/1606.04835>	1-7
A	FARUQUI, Manaal, et al., Retrofitting Word Vectors to Semantic Lexicons, arXiv preprint arXiv:1411.4166 [online], 2014, [retrieval date 2016.12.05], Internet: <URL: http://arxiv.org/abs/1411.4166>	1-7
☒ Further documents are listed in the continuation of Box C. ☐ 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 08 December 2016 (08.12.16)		Date of mailing of the international search report 20 December 2016 (20.12.16)
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan		Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2016/083000

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	BOLLEGALA, Danushka, et al., Joint Word Representation Learning using a Corpus and a Semantic Lexicon, arXiv preprint arXiv:1511.06438 [online], 2015, [retrieval date 2016.12.05], Internet: <URL: http://arxiv.org/abs/1511.06438 >	1-7

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06F17/30(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（I P C））

Int.Cl. G06F17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2016年
日本国実用新案登録公報	1996-2016年
日本国登録実用新案公報	1994-2016年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Google Scholar

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	LI, Qi, et al., Learning Word Sense Embeddings from Word Sense Definitions, arXiv preprint arXiv:1606.04835 [オンライン], 2016, [検索日 2016.12.05], インターネット: <URL: http://arxiv.org/abs/1606.04835>	1-7
A	FARUQUI, Manaal, et al., Retrofitting Word Vectors to Semantic Lexicons, arXiv preprint arXiv:1411.4166 [オンライン], 2014, [検索日 2016.12.05], インターネット: <URL: http://arxiv.org/abs/1411.4166>	1-7

☒ C欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日

08.12.2016

国際調査報告の発送日

20.12.2016

国際調査機関の名称及びあて先

日本国特許庁（I S A/J P）

郵便番号100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

吉田 誠

電話番号 03-3581-1101 内線 3599

5M

6291

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	BOLLEGALA, Danushka, et al., Joint Word Representation Learning using a Corpus and a Semantic Lexicon, arXiv preprint arXiv:1511.06438 [オンライン], 2015, [検索日 2016.12.05], インターネット: <URL: http://arxiv.org/abs/1511.06438 >	1-7