

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5569908号  
(P5569908)

(45) 発行日 平成26年8月13日(2014.8.13)

(24) 登録日 平成26年7月4日(2014.7.4)

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

請求項の数 7 (全 25 頁)

(21) 出願番号 特願2011-18787(P2011-18787)  
 (22) 出願日 平成23年1月31日(2011.1.31)  
 (65) 公開番号 特開2012-159983(P2012-159983A)  
 (43) 公開日 平成24年8月23日(2012.8.23)  
 審査請求日 平成25年11月26日(2013.11.26)

(73) 特許権者 801000027  
 学校法人明治大学  
 東京都千代田区神田駿河台1-1  
 (74) 代理人 100064908  
 弁理士 志賀 正武  
 (74) 代理人 100106909  
 弁理士 棚井 澄雄  
 (74) 代理人 100108578  
 弁理士 高橋 詔男  
 (74) 代理人 100126882  
 弁理士 五十嵐 光永  
 (72) 発明者 高木 友博  
 神奈川県川崎市多摩区東三田1-1-1  
 学校法人明治大学 生田校舎内

最終頁に続く

(54) 【発明の名称】 類推装置、類推方法及びプログラム

(57) 【特許請求の範囲】

【請求項1】

ベース状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース状況部分組合せデータを生成するベース状況部分組合せ生成部と、

ベース結果データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース結果部分組合せデータを生成するベース結果部分組合せ生成部と、

前記ベース状況部分組合せデータのうち1つと前記ベース結果部分組合せデータのうち1つとからなる異なる組合せそれぞれについて、前記ベース状況部分組合せデータが示す単語と前記ベース結果部分組合せデータが示す単語とを関係付ける単語である関連付け単語を、記事記憶装置に記憶されている記事データから抽出する関係集合生成部と、

ターゲット状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるターゲット状況部分組合せデータを生成するターゲット状況部分組合せ生成部と、

前記ターゲット状況部分組合せデータのうち1つと前記関連付け単語のうち1つとからなる異なる組合せそれぞれについて、前記関連付け単語によって前記ターゲット状況部分組合せデータが示す単語と関係付けられる単語を、前記記事記憶装置に記憶されている記事データから類推結果として抽出する類推結果生成部と、

を備えることを特徴とする類推装置。

【請求項2】

前記関係集合生成部が抽出した前記関連付け単語それぞれについて、前記記事記憶装置

に記憶されている前記記事データから得られる当該関連付け単語と前記ベース状況部分組合せデータ及び前記ベース結果部分組合せデータの共起との関連の強さ、当該関連付け単語と前記ベース状況部分組合せデータとの関連の強さ、及び、当該関連付け単語と前記ベース結果部分組合せデータとの関連の強さに基づき、状況と結果の関係付けを行なう単語としての妥当性を定量的に表す関係妥当性スコアを算出する関係妥当性スコア算出部と、

前記類推結果生成部が類推結果として抽出した前記単語のそれぞれについて、前記記事記憶装置に記憶されている前記記事データから得られる当該単語と前記ターゲット状況部分組合せデータ及び前記関連付け単語の共起との関連の強さ、当該単語と前記ターゲット状況部分組合せデータとの関連の強さ、及び、当該単語と前記関連付け単語との関連の強さ、ならびに、前記関連付け単語について算出された前記関係妥当性スコアに基づき、類推結果としての妥当性を定量的に表す類推結果妥当性スコアを算出する類推結果妥当性スコア算出部と、

10

をさらに備えることを特徴とする請求項1に記載の類推装置。

#### 【請求項3】

ベース状況データ及びベース結果データからなる複数の事例データ毎に、前記ベース状況部分組合せ生成部に、前記事例データを構成する前記ベース状況データからベース状況部分組合せデータを生成させ、前記ベース結果部分組合せ生成部に、前記事例データを構成する前記ベース結果データから前記ベース結果部分組合せデータを生成させ、前記関係集合生成部に、前記ベース状況データから生成された前記ベース状況部分組合せデータのうち1つと、前記ベース結果データから生成された前記ベース結果部分組合せデータのう

20

ち1つとからなる異なる組合せそれぞれについて関連付け単語を記事データから抽出させ、前記関係妥当性スコア算出部に、前記関連付け単語それぞれについて関係妥当性スコアを算出させ、前記類推結果生成部に、前記ターゲット状況部分組合せデータのうち1つと前記関連付け単語のうち1つとからなる異なる組合せそれぞれについて記事データから類推結果の単語を抽出させ、前記類推結果妥当性スコア算出部に、類推結果として抽出された前記単語のそれぞれについて類推結果妥当性スコアを算出させる類推処理制御部と、

前記事例データ毎に得られた前記類推結果の単語に含まれる同一の単語を統合するとともに、統合した前記同一の単語について算出された前記類推結果妥当性スコアを積算する類推結果積算部とをさらに備える、

30

ことを特徴とする請求項2に記載の類推装置。

#### 【請求項4】

前記関係集合生成部は、前記ベース状況部分組合せデータが示す単語が主語の名詞かつ前記ベース結果部分組合せデータが示す単語が述部の名詞である前記記事データの文から、述部の動詞を前記関連付け単語として抽出し、

前記類推結果生成部は、前記ターゲット状況部分組合せデータが示す複数の単語が主語の名詞かつ前記関連付け単語が述部の動詞である前記記事データの文から、述部の名詞を類推結果として抽出する、

ことを特徴とする請求項1から請求項3のいずれか1項に記載の類推装置。

#### 【請求項5】

前記記事記憶装置は、さらに、所定の分野に関する単語を含む辞書データを記憶し、前記関係集合生成部は、前記ベース状況部分組合せデータにより示される単語及び前記ベース結果部分組合せデータにより示される単語が共起する前記記事データの文から、前記辞書データに含まれる単語を前記関連付け単語として抽出し、

40

前記類推結果生成部は、前記ターゲット状況部分組合せデータにより示される単語及び前記関連付け単語が共起する前記記事データの文から、前記辞書データに含まれる単語を類推結果として抽出する、

ことを特徴とする請求項1から請求項3のいずれか1項に記載の類推装置。

#### 【請求項6】

類推装置が実行する類推方法であって、ベース状況部分組合せ生成部が、ベース状況データが示す複数の単語から異なる組合せ

50

により選択した所定単語数の前記単語からなるベース状況部分組合せデータを生成するベース状況部分組合せ生成過程と、

ベース結果部分組合せ生成部が、ベース結果データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース結果部分組合せデータを生成するベース結果部分組合せ生成過程と、

関係集合生成部が、前記ベース状況部分組合せデータのうち1つと前記ベース結果部分組合せデータのうち1つとからなる異なる組合せそれぞれについて、前記ベース状況部分組合せデータが示す単語と前記ベース結果部分組合せデータが示す単語とを関係付ける単語である関連付け単語を、記事記憶装置に記憶されている記事データから抽出する関係集合生成過程と、

10

ターゲット状況部分組合せ生成部が、ターゲット状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるターゲット状況部分組合せデータを生成するターゲット状況部分組合せ生成過程と、

類推結果生成部が、前記ターゲット状況部分組合せデータのうち1つと前記関連付け単語のうち1つとからなる異なる組合せそれぞれについて、前記関連付け単語によって前記ターゲット状況部分組合せデータが示す単語と関係付けられる単語を、前記記事記憶装置に記憶されている記事データから類推結果として抽出する類推結果生成過程と、

を有することを特徴とする類推方法。

#### 【請求項7】

類推装置として用いられるコンピュータを、

20

ベース状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース状況部分組合せデータを生成するベース状況部分組合せ生成部、

ベース結果データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース結果部分組合せデータを生成するベース結果部分組合せ生成部、

前記ベース状況部分組合せデータのうち1つと前記ベース結果部分組合せデータのうち1つとからなる異なる組合せそれぞれについて、前記ベース状況部分組合せデータが示す単語と前記ベース結果部分組合せデータが示す単語とを関係付ける単語である関連付け単語を、記事記憶装置に記憶されている記事データから抽出する関係集合生成部、

ターゲット状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるターゲット状況部分組合せデータを生成するターゲット状況部分組合せ生成部、

30

前記ターゲット状況部分組合せデータのうち1つと前記関連付け単語のうち1つとからなる異なる組合せそれぞれについて、前記関連付け単語によって前記ターゲット状況部分組合せデータが示す単語と関係付けられる単語を、前記記事記憶装置に記憶されている記事データから類推結果として抽出する類推結果生成部、

として機能させることを特徴とするプログラム。

#### 【発明の詳細な説明】

#### 【技術分野】

#### 【0001】

本発明は、事例に基づいて類推を行なう類推装置、類推方法及びプログラムに関する。

40

#### 【背景技術】

#### 【0002】

基本的な類推手法の一つとして、四項類推が知られている。四項類推は、一般に以下の式(1)のように表示する。

#### 【0003】

$A : B = C : X ? \quad \dots (1)$

#### 【0004】

上記は、「AならばB、Cならば何であるか？」を意味しており、Xは類推結果を表している。つまり、A、B、CからX?を求めるのが四項類推である。四項類推では、基本構成要素として、規定領域(ベース)と目標領域(ターゲット)をおく。ベースとは類推

50

する際に用いる既存の知識のことであり、ターゲットとは解決をしなければならない未知の問題のことであり、つまり、上記の四項類推の例では、A及びBはベースに属し、C及びXはターゲットに属する。非特許文献1では、この四項類推のアルゴリズムを実現するシステムが提案されている。図13は、そのアルゴリズムの概要を示す図である。

【0005】

図13に示すアルゴリズムでは、まず、ベースにおけるAとBの間の関係集合Rを求める関係抽出(Relation extraction)処理を行なう。続いて、関係抽出処理において求めた関係集合Rをターゲットに移し、Cにその関係集合Rを適用してXを求める関係マッピング(Relation Mapping)処理を行なう。以後、ベースのAを状況、結果Bをその状況での結果、このAとBの組を1つの事例と呼ぶ。

10

【0006】

非特許文献1では、関係抽出処理において得られる関係集合Rを、類似な関係を表す単語 $r_i$ の集合として以下の式(2)のように定義している。

【0007】

$$R = \{ r_i \} \quad (i \text{ は } 1 \text{ 以上の整数}) \quad \dots (2)$$

【0008】

関係マッピング処理においては、関係集合Rに含まれる各単語 $r_i$ を用いてCから複数の類推結果の候補となる単語 $x_j$ を求め、さらに、求めた単語 $x_j$ それぞれについて、尤もらしさを定量的に示す値である $score(x_j)$ を算出する。つまり、類推結果は、単語 $x_j$ と、それに付与された $score(x_j)$ とからなる集合であり、以下の式(3)により表される。

20

【0009】

$$X = \{ x_j \} \quad (j \text{ は } 1 \text{ 以上の整数}) \quad \dots (3)$$

【0010】

一方、非特許文献2では、次のように類推を行っている。まず、DVDタイトルに対応した説明文から抽出された各単語にTF-IDF値のスコアを付与しておき、ユーザの選択操作に従って当該スコアを修正する。そして、その修正したスコアが上位の単語からなるDVDデータと、入力された単語とに基づいて推薦するDVDデータを選択している。

また、非特許文献3では、次のように類推を行なっている。まず、過去の1週間の記事データからTF-IDF値が上位の単語からなるワードベクトルを生成し、生成したワードベクトルに基づいてその1週間の翌日の記事データから候補語を抽出する。このワードベクトルと、候補語の組合せを記事データの時期をずらしながら複数生成しておく。そして、予想する日にちより前の1週間の記事データから同様にワードベクトルを生成して過去の記事データから生成したワードベクトルとのマッチングを行い、マッチするワードベクトルに対応した候補語を予測結果としている。

30

【先行技術文献】

【非特許文献】

【0011】

【非特許文献1】岡田 一宏、外4名、「構造写像理論に基づく類推手法」、第35回ファジィ・ワークショップ講演論文集、2010年3月、p.25-30

40

【非特許文献2】伊達 寅彦、外7名、「CFSを用いたDVD推薦システムの提案」、第35回ファジィ・ワークショップ講演論文集、日本知能情報ファジィ学会、2010年3月、p.33-36

【非特許文献3】伊藤 慎一郎、外1名、「言語による経済動向の予測」、第35回ファジィ・ワークショップ講演論文集、2010年3月、p.37-40

【発明の概要】

【発明が解決しようとする課題】

【0012】

一般に四項類推では、ベースの状況であるA、ターゲットの状況であるCとも、1つの単語で構成されるいわば1次元の表現であり、これは、非特許文献1においても同様であ

50

る。一方、非特許文献2及び非特許文献3では、複数の単語で表される状況を事例として類推を行うことができるが、ターゲットもベースも同じ概念に属していなければならない。例えば、非特許文献1では、「魚ならばうるこ、鳥ならば何であるか?」というように、ターゲットが魚に関する概念に属し、ベースが鳥に関する概念に属していても類推を行うことができる。しかし、非特許文献2では、ターゲットもベースもDVDの概念に属し、非特許文献3では、ターゲットもベースも経済の概念に属する。

【0013】

本発明は、このような事情を考慮してなされたもので、その目的は、ベースにおける状況及び結果とからなる事例と、ベースが属する概念とは異なる概念に属するターゲットにおける状況が与えられたときに、ベース及びターゲットの状況がそれぞれ複数の単語で構成される場合においても、ターゲットにおける類推結果を求めることができる類推装置、類推方法、及び、プログラムを提供することにある。

10

【課題を解決するための手段】

【0014】

この発明は、上記の課題を解決すべくなされたもので、ベース状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース状況部分組合せデータを生成するベース状況部分組合せ生成部と、ベース結果データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース結果部分組合せデータを生成するベース結果部分組合せ生成部と、前記ベース状況部分組合せデータのうち1つと前記ベース結果部分組合せデータのうち1つとからなる異なる組合せそれぞれについて、前記ベース状況部分組合せデータが示す単語と前記ベース結果部分組合せデータが示す単語とを関係付ける単語である関連付け単語を、記事記憶装置に記憶されている記事データから抽出する関係集合生成部と、ターゲット状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるターゲット状況部分組合せデータを生成するターゲット状況部分組合せ生成部と、前記ターゲット状況部分組合せデータのうち1つと前記関連付け単語のうち1つとからなる異なる組合せそれぞれについて、前記関連付け単語によって前記ターゲット状況部分組合せデータが示す単語と関係付けられる単語を、前記記事記憶装置に記憶されている記事データから類推結果として抽出する類推結果生成部と、を備えることを特徴とする類推装置である。

20

【0015】

また本発明は、上述した類推装置であって、前記関係集合生成部が抽出した前記関連付け単語それぞれについて、前記記事記憶装置に記憶されている前記記事データから得られる当該関連付け単語と前記ベース状況部分組合せデータ及び前記ベース結果部分組合せデータの共起との関連の強さ、当該関連付け単語と前記ベース状況部分組合せデータとの関連の強さ、及び、当該関連付け単語と前記ベース結果部分組合せデータとの関連の強さに基づき、状況と結果の関係付けを行なう単語としての妥当性を定量的に表す関係妥当性スコアを算出する関係妥当性スコア算出部と、前記類推結果生成部が類推結果として抽出した前記単語のそれぞれについて、前記記事記憶装置に記憶されている前記記事データから得られる当該単語と前記ターゲット状況部分組合せデータ及び前記関連付け単語の共起との関連の強さ、当該単語と前記ターゲット状況部分組合せデータとの関連の強さ、及び、当該単語と前記関連付け単語との関連の強さ、ならびに、前記関連付け単語について算出された前記関係妥当性スコアに基づき、類推結果としての妥当性を定量的に表す類推結果妥当性スコアを算出する類推結果妥当性スコア算出部と、をさらに備えることを特徴とする。

30

40

【0016】

また本発明は、上述した類推装置であって、ベース状況データ及びベース結果データからなる複数の事例データ毎に、前記ベース状況部分組合せ生成部に、前記事例データを構成する前記ベース状況データからベース状況部分組合せデータを生成させ、前記ベース結果部分組合せ生成部に、前記事例データを構成する前記ベース結果データから前記ベース結果部分組合せデータを生成させ、前記関係集合生成部に、前記ベース状況データから生

50

成された前記ベース状況部分組合せデータのうち1つと、前記ベース結果データから生成された前記ベース結果部分組合せデータのうち1つとからなる異なる組合せそれぞれについて関連付け単語を記事データから抽出させ、前記関係妥当性スコア算出部に、前記関連付け単語それぞれについて関係妥当性スコアを算出させ、前記類推結果生成部に、前記ターゲット状況部分組合せデータのうち1つと前記関連付け単語のうち1つとからなる異なる組合せそれぞれについて記事データから類推結果の単語を抽出させ、前記類推結果妥当性スコア算出部に、類推結果として抽出された前記単語のそれぞれについて類推結果妥当性スコアを算出させる類推処理制御部と、前記事例データ毎に得られた前記類推結果の単語に含まれる同一の単語を統合するとともに、統合した前記同一の単語について算出された前記類推結果妥当性スコアを積算する類推結果積算部とをさらに備える、ことを特徴とする。

10

## 【0017】

また本発明は、上述した類推装置であって、前記関係集合生成部は、前記ベース状況部分組合せデータが示す単語が主語の名詞かつ前記ベース結果部分組合せデータが示す単語が述部の名詞である前記記事データの文から、述部の動詞を前記関連付け単語として抽出し、前記類推結果生成部は、前記ターゲット状況部分組合せデータが示す複数の単語が主語の名詞かつ前記関連付け単語が述部の動詞である前記記事データの文から、述部の名詞を類推結果として抽出する、ことを特徴とする。

## 【0018】

また本発明は、上述した類推装置であって、前記記事記憶装置は、さらに、所定の分野に関する単語を含む辞書データを記憶し、前記関係集合生成部は、前記ベース状況部分組合せデータにより示される単語及び前記ベース結果部分組合せデータにより示される単語が共起する前記記事データの文から、前記辞書データに含まれる単語を前記関連付け単語として抽出し、前記類推結果生成部は、前記ターゲット状況部分組合せデータにより示される単語及び前記関連付け単語が共起する前記記事データの文から、前記辞書データに含まれる単語を類推結果として抽出する、ことを特徴とする。

20

## 【0019】

また本発明は、類推装置が実行する類推方法であって、ベース状況部分組合せ生成部が、ベース状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース状況部分組合せデータを生成するベース状況部分組合せ生成過程と、ベース結果部分組合せ生成部が、ベース結果データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース結果部分組合せデータを生成するベース結果部分組合せ生成過程と、関係集合生成部が、前記ベース状況部分組合せデータのうち1つと前記ベース結果部分組合せデータのうち1つとからなる異なる組合せそれぞれについて、前記ベース状況部分組合せデータが示す単語と前記ベース結果部分組合せデータが示す単語とを関係付ける単語である関連付け単語を、記事記憶装置に記憶されている記事データから抽出する関係集合生成過程と、ターゲット状況部分組合せ生成部が、ターゲット状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるターゲット状況部分組合せデータを生成するターゲット状況部分組合せ生成過程と、類推結果生成部が、前記ターゲット状況部分組合せデータのうち1つと前記関連付け単語のうち1つとからなる異なる組合せそれぞれについて、前記関連付け単語によって前記ターゲット状況部分組合せデータが示す単語と関係付けられる単語を、前記記事記憶装置に記憶されている記事データから類推結果として抽出する類推結果生成過程と、を有することを特徴とする類推方法である。

30

40

## 【0020】

また本発明は、類推装置として用いられるコンピュータを、ベース状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース状況部分組合せデータを生成するベース状況部分組合せ生成部、ベース結果データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるベース結果部分組合せデータを生成するベース結果部分組合せ生成部、前記ベース状況部分組合せデータのうち

50

1つと前記ベース結果部分組合せデータのうち1つとからなる異なる組合せそれぞれについて、前記ベース状況部分組合せデータが示す単語と前記ベース結果部分組合せデータが示す単語とを関係付ける単語である関連付け単語を、記事記憶装置に記憶されている記事データから抽出する関係集合生成部、ターゲット状況データが示す複数の単語から異なる組合せにより選択した所定単語数の前記単語からなるターゲット状況部分組合せデータを生成するターゲット状況部分組合せ生成部、前記ターゲット状況部分組合せデータのうち1つと前記関連付け単語のうち1つとからなる異なる組合せそれぞれについて、前記関連付け単語によって前記ターゲット状況部分組合せデータが示す単語と関係付けられる単語を、前記記事記憶装置に記憶されている記事データから類推結果として抽出する類推結果生成部、として機能させることを特徴とするプログラムである。

10

## 【発明の効果】

## 【0021】

本実施形態によれば、ベースにおける状況及び結果とからなる事例と、ベースが属する概念とは異なる概念に属するターゲットにおける状況が与えられたときに、ベース及びターゲットの状況がそれぞれ複数の単語で構成される場合においても、ターゲットにおける類推結果を求めることができる。

## 【図面の簡単な説明】

## 【0022】

【図1】本発明の第1の実施形態による類推装置の関係抽出処理の概要を示す図である。

【図2】同実施形態による類推装置の関係マッピング処理の概要を示す図である。

20

【図3】同実施形態による類推装置の構成を示すブロック図である。

【図4】同実施形態による類推装置の類推処理フローを示す図である。

【図5】同実施形態による類推装置の類推処理フローを示す図である。

【図6】同実施形態による単語rの抽出処理を説明するための図である。

【図7】同実施形態による単語xの抽出処理を説明するための図である。

【図8】同実施形態による単語r及び単語xの抽出処理を説明するための図である。

【図9】本発明の第2の実施形態による類推装置の処理概要を示す図である。

【図10】同実施形態による類推装置の構成を示すブロック図である。

【図11】同実施形態による類推装置の処理フローを示す図である。

【図12】同実施形態による類推装置の処理フローを示す図である。

30

【図13】従来技術の類推アルゴリズムを示す図である。

## 【発明を実施するための形態】

## 【0023】

以下、図面を参照して本発明の実施形態を説明する。

## 【0024】

## [第1の実施形態]

一般に、四項類推における規定領域(ベース)の状況Aと目標領域(ターゲット)の状況Cは、1つの単語で構成されるいわば1次元の表現である。本実施形態では、この状況A及び状況Cを、複数の単語で構成される多次元の構成に拡張する。これを、式(1)で示したmodus ponensと同様の表現方法で記述すると、以下の式(4)のようになる。

40

## 【0025】

## 【数1】

$$\frac{A(t_{a1}, t_{a2}, \dots, t_{am}) : B(w_{b1}, w_{b2}, \dots, w_{bg})}{C(t_{c1}, t_{c2}, \dots, t_{cm}) : X(w_{x1}, w_{x2}, \dots, w_{xg})} ? \quad \dots (4)$$

## 【0026】

上記のように、状況Aは、単語 $t_{a1}, t_{a2}, \dots, t_{am}$ で構成され、状況Cは、単語 $t_{c1}, t_{c2}, \dots, t_{cm}$ で構成される。つまり、状況A及び状況Cは、m個の単語

50

を要素とする  $m$  次元ワードベクトルである。一方、ベースの状況  $A$  での結果  $B$  は、単語  $w_{b_1}, w_{b_2}, \dots, w_{b_g}$  で構成され、ターゲットの状況  $C$  での結果  $X$  は、単語  $w_{x_1}, w_{x_2}, \dots, w_{x_g}$  で構成される。つまり、結果  $B$  及び結果  $X$  は、 $g$  個の単語を要素とする  $g$  次元ワードベクトルである。

本実施形態の類推装置は、式 (4) に示すような多次元ベクトルであるベースの状況  $A$ 、ベースの状況  $A$  での結果  $B$ 、及び、ターゲットの状況  $C$  から、ターゲットの状況  $C$  での結果  $X$  を求める処理を行なう。

#### 【0027】

図1及び2を用いて、本実施形態の類推装置の処理概要を説明する。

図1は、本実施形態の類推装置における関係抽出処理の概要を示す図である。類推装置は、状況  $A$  を表す  $m$  個 ( $m$  は2以上の整数) の単語の列である  $m$  次元のワードベクトルから  $n$  個 ( $n$  は1以上  $m$  以下の整数) の単語を選択したときの全ての組合せを生成する。さらに、類推装置は、結果  $B$  を表す  $g$  個 ( $g$  は2以上の整数) の単語の列である  $g$  次元のワードベクトルの中から  $h$  ( $h$  は1以上  $g$  以下の整数) 個の単語を選択したときの全ての組合せを生成する。類推装置は、状況  $A$  から生成した組合せを構成する  $n$  個の単語と、結果  $B$  から生成した組合せを構成する  $h$  個の単語とが共起する記事を検索する。記事には複数の単語が含まれており、それら複数の単語からなるワードベクトルとみなすことができる。このワードベクトルを構成する単語の中には、他の単語同士を関連付ける単語も含まれる。類推装置は、状況  $A$  から生成した組合せを構成する  $n$  個の単語と、結果  $B$  から生成した組合せを構成する  $h$  個の単語とを関係付ける単語  $r$  を記事から抽出する。抽出された単語  $r$  をそれぞれ  $r_1, r_2, \dots$  とすると、関係集合  $R = \{r_i\}$  と表すことができる ( $i$  は1以上の整数)。

#### 【0028】

図2は、関係マッピング処理の概要を示す図である。類推装置は、状況  $C$  を構成する  $m$  個の単語の列である  $m$  次元のワードベクトルから  $n$  個の単語を選択したときの全ての組合せを生成する。類推装置は、関係集合  $R$  を構成する各単語  $r_i$  によって、状況  $C$  から生成した組合せと関連付けられる単語を記事から抽出し、抽出した単語群を類推結果  $X$  とする。

#### 【0029】

図3は、本発明の第一の実施形態による類推装置1の構成を示すブロック図である。類推装置1は、例えば、1台または複数台のコンピュータ装置で実現することができ、記事データを記憶する記事記憶装置5とネットワークを介して接続される。記事データは、例えば、ニュースのテキストデータ、雑誌のテキストデータ、知識データベースの内容などである。記事データは、複数の単語からなるが、その中には、他の単語間を関係付ける単語が含まれる。例えば、自然言語の文の場合、主語の名詞(単語)と、述部にある名詞(単語)とを、述部の動詞(単語)が関連付けている。

#### 【0030】

同図に示すように、類推装置1は、事例記憶部11、入力部12、関係抽出部13、処理結果記憶部14、関係マッピング部15及び出力部16を備えて構成される。

事例記憶部11は、状況  $A$  のワードベクトルを示す状況  $A$  データ(ベース状況データ)と、結果  $B$  のワードベクトルを示す結果  $B$  データ(ベース結果データ)を記憶する。状況  $A$  データは、 $m$  個の単語を要素とする  $m$  次元ワードベクトルを示し、結果  $B$  データは、 $g$  個の単語を要素とする  $g$  次元ワードベクトルを示す。入力部12は、キーボードなどによって、状況  $C$  のワードベクトルを示す状況  $C$  データ(ターゲット状況データ)の入力を受ける。状況  $C$  データは、 $m$  個の単語を要素とする  $m$  次元ワードベクトルである。なお、入力部12は、ネットワークを介して接続される他のコンピュータ装置から状況  $C$  データを受信したり、コンピュータ読み取り可能な記録媒体から状況  $C$  データを読み出したりしてもよい。処理結果記憶部14は、関係抽出部13及び関係マッピング部15の各部による処理結果を記憶する。

#### 【0031】

10

20

30

40

50



関係抽出部 1 3 は、ベース状況部分組合せ生成部 2 1、ベース結果部分組合せ生成部 2 2、ベース共起記事検索部 2 3、関係集合生成部 2 4 及び関係妥当性スコア算出部 2 5 を備える。

ベース状況部分組合せ生成部 2 1 は、事例記憶部 1 1 から読み出した状況 A データが示す  $m$  個の単語から  $n$  個の単語を選択したときの全ての組合せ ( $n = 1$  の場合も説明の便宜上、組合せと記載する。) を生成し、これらの組合せそれぞれを示すデータである部分組合せ  $A_1$  (ベース状況部分組合せデータ) を生成する ( ${}^1_1 C_n$ 、 $1$  は整数)。つまり、部分組合せ  $A_1$  は、 $n$  個の単語を要素とする  $n$  次元のワードベクトルを示す。

【0032】

ベース結果部分組合せ生成部 2 2 は、事例記憶部 1 1 から読み出した結果 B データが示す  $g$  個の単語から  $h$  個の単語を選択したときの全ての組合せ ( $h = 1$  の場合も説明の便宜上、組合せと記載する。) を生成し、これらの組合せそれぞれを示すデータである部分組合せ  $B_k$  (ベース結果部分組合せデータ) を生成する ( ${}^1_k C_h$ 、 $k$  は整数)。つまり、部分組合せ  $B_k$  は、 $h$  個の単語を要素とする  $h$  次元のワードベクトルを示す。

【0033】

ベース共起記事検索部 2 3 は、1 つの部分組合せ  $A_1$  と、1 つの部分組合せ  $B_k$  とからなる全ての組合せそれぞれについて記事記憶装置 5 に記憶されている記事データを検索し、部分組合せ  $A_1$  及び部分組合せ  $B_k$  が示す全ての単語が共起する記事を示す記事データを特定する。ベース共起記事検索部 2 3 は、特定した記事データの集合からなるデータである記事集合  $D$  を生成する。関係集合生成部 2 4 は、1 つの部分組合せ  $A_1$  と 1 つの部分組合せ  $B_k$  とからなる全ての組合せそれぞれについて、部分組合せ  $A_1$  が示す  $n$  個の単語と、部分組合せ  $B_k$  が示す  $h$  個の単語とを関係付ける単語  $r$  (関連付け単語) を記事集合  $D$  に含まれる各記事データから抽出し、抽出した単語  $r$  の集合を示すデータである関係集合  $R$  を生成する。抽出された各単語  $r$  を、 $r_1$ 、 $r_2$ 、... とする。関係妥当性スコア算出部 2 5 は、関係集合生成部 2 4 が生成した関係集合  $R$  に含まれる各単語  $r_i$  ( $i$  は 1 以上の整数) の関係妥当性スコア  $score(r_i)$  を算出する。関係妥当性スコア  $score(r_i)$  は、単語  $r_i$  が、部分組合せ  $A_1$  と部分組合せ  $B_k$  とを関連付ける単語として妥当であるかの尤もらしさを定量的に表す値である。

【0034】

関係マッピング部 1 5 は、ターゲット状況部分組合せ生成部 2 6、ターゲット共起記事検索部 2 7、類推結果生成部 2 8 及び類推結果妥当性スコア算出部 2 9 を備える。

ターゲット状況部分組合せ生成部 2 6 は、入力部 1 2 により入力された状況 C データが示す  $m$  個の単語から  $n$  個の単語を選択したときの全ての組合せを生成し、これらの組合せそれぞれを示すデータである部分組合せ  $C_f$  (ターゲット状況部分組合せデータ) を生成する ( ${}^1_f C_n$ 、 $f$  は整数)。つまり、部分組合せ  $C_f$  は、 $n$  個の単語を要素とする  $n$  次元のワードベクトルを示す。

【0035】

ターゲット共起記事検索部 2 7 は、1 つの部分組合せ  $C_f$  と、関係集合  $R$  に含まれる 1 つの単語  $r_i$  とからなる全ての組合せそれぞれについて記事記憶装置 5 を検索し、部分組合せ  $C_f$  が示す全ての単語と、単語  $r_i$  とが共起する記事を示す記事データを特定する。ターゲット共起記事検索部 2 7 は、特定した記事データの集合からなるデータである記事集合  $E$  を生成する。類推結果生成部 2 8 は、単語  $r_i$  によって部分組合せ  $C_f$  が示す  $n$  個の単語と関係付けられる単語  $x$  を記事集合  $E$  に含まれる各記事データから抽出し、抽出された単語  $x$  の集合を示すデータである類推結果集合  $X$  を生成する。抽出された各単語  $x$  を、 $x_1$ 、 $x_2$ 、... とする。類推結果妥当性スコア算出部 2 9 は、類推結果生成部 2 8 により生成された類推結果集合  $X$  に含まれる各単語  $x_j$  ( $j$  は 1 以上の整数) の類推結果妥当性スコア  $score(x_j)$  を算出する。類推結果妥当性スコア  $score(x_j)$  は、単語  $x_j$  が、類推結果として妥当であるかを定量的に表す値である。

【0036】

出力部 1 6 は、類推結果生成部 2 8 により生成された類推結果集合  $X$  が示す各単語  $x_j$

10

20

30

40

50

と、類推結果妥当性スコア算出部 29 により算出された当該単語  $x_j$  の類推結果妥当性スコア  $score(x_j)$  とからなる類推結果データをディスプレイに表示させる。あるいは、出力部 16 は、類推結果データをプリンタなどにより印刷してもよく、情報記録媒体へ書き込んでよく、ネットワークを介して接続されるコンピュータ装置へ送信してもよい。この類推結果データは、結果 X を表す  $g$  次元ワードベクトルの要素である単語  $w_{x_1}, w_{x_2}, \dots, w_{x_g}$  のいずれかであると類推される単語とその類推結果妥当性スコアの集合である。つまり、類推結果は、単語  $x_j$  からなるファジィ集合によって表現される。

#### 【0037】

図 4 及び図 5 は、図 3 に示す類推装置 1 の類推処理フローを示す図である。

我々が日常使用している言語では 3 万語程度であるが、ここから各記事の特徴を表す重要語を抽出し、この抽出した重要語を用いて生成した状況 A データ及び結果 B データを類推装置 1 の事例記憶部 11 に記憶させておく。一般的に、状況 A のワードベクトルの次元数  $m$  や結果 B のワードベクトルの次元数  $g$  として 20 ~ 50 を用いるが、それ以外の次元数でもよい。また、処理結果記憶部 14 は、初期値 N U L L の記事集合 D、記事集合 E、関係集合 R 及び類推結果集合 X を記憶する。

#### 【0038】

図 4 において、類推装置 1 の入力部 12 は、状況 C データの入力を受ける（ステップ S100）。続いて、ベース状況部分組合せ生成部 21 は、事例記憶部 11 から状況 A データを読み出し、状況 A データが示すワードベクトルの要素である  $m$  個の単語から（ $m$  は 2 以上の整数）、 $n$  個（ $n$  は 1 以上  $m$  以下の整数）の単語を選択したときの組合せを全て生成する。組合せの数は、 ${}_m C_n$  となる。ベース状況部分組合せ生成部 21 は、生成した単語の組合せを要素とする  $n$  次元ワードベクトルを示す部分組合せ  $A_1$ （ $1 \leq 1 \leq {}_m C_n$ 、 $1$  は整数）を生成し、処理結果記憶部 14 に書き込む（ステップ S105）。抽出単語数  $n$  は、2 ~ 5 程度を用いるが、これ以外の値でもよい。

#### 【0039】

続いて、ベース結果部分組合せ生成部 22 は、事例記憶部 11 から結果 B データを読み出し、結果 B データが示すワードベクトルの要素である  $g$  個の単語から、 $h$  個（ $h$  は 1 以上  $g$  以下の整数）の単語を選択したときの全ての組合せを生成する。組合せの数は、 ${}_g C_h$  となる。ベース結果部分組合せ生成部 22 は、生成した単語の組合せを要素とする  $h$  次元ワードベクトルを示す部分組合せ  $B_k$ （ $1 \leq k \leq {}_g C_h$ 、 $k$  は整数）を生成し、処理結果記憶部 14 に書き込む（ステップ S110）。抽出単語数  $h$  は、2 ~ 5 程度を用いるが、これ以外の値でもよい。

#### 【0040】

ベース共起記事検索部 23 は、ステップ S105 において処理結果記憶部 14 に書き込まれた部分組合せ  $A_1$  の 1 つと、ステップ S110 において処理結果記憶部 14 に書き込まれた部分組合せ  $B_k$  の 1 つとからなる全ての組合せを生成する（ステップ S115）。つまり、生成される組合せは  $A_1 - B_1, A_1 - B_2, \dots, A_1 - B_{({}_g C_h)}, A_2 - B_1, A_2 - B_2, \dots, A_{({}_m C_n - 1)} - B_{({}_g C_h)}, A_{({}_m C_n)} - B_1, A_{({}_m C_n)} - B_2, \dots, A_{({}_m C_n)} - B_{({}_g C_h)}$  である。ベース共起記事検索部 23 は、ステップ S115 において生成した組合せのうち、まだステップ S125 の処理対象としていない組合せ  $A_1 - B_k$  を選択する（ステップ S120）。

#### 【0041】

ベース共起記事検索部 23 は、ステップ S120 において選択した組合せ  $A_1 - B_k$  を構成する部分組合せ  $A_1$  及び部分組合せ  $B_k$  を処理結果記憶部 14 から読み出す。ベース共起記事検索部 23 は、記事記憶装置 5 が記憶する記事データを検索し、読み出した部分組合せ  $A_1$  が示す  $n$  個の単語と、読み出した部分組合せ  $B_k$  が示す  $h$  個の単語とが全て含まれる記事を示す記事データを特定する。ベース共起記事検索部 23 は、特定した記事データを記事記憶装置 5 から読み出し、読み出した記事データを処理結果記憶部 14 に記憶されている記事集合 D に追加する（ステップ S125）。ただし、ベース共起記事検索部 23 は、抽出した記事データがすでに記事集合 D に含まれている場合は追加しない。関係

10

20

30

40

50

集合Dに含まれる記事データをそれぞれ記事データ $d_1$ 、 $d_2$ 、...とする。

【0042】

ベース共起記事検索部23は、ステップS115において生成した全ての組合せをステップS125の処理対象としたかを判断する(ステップS130)。まだステップS125の処理対象としていない組合せがあると判断した場合(ステップS130:NO)、ベース共起記事検索部23は、ステップS120からの処理を繰り返す。ベース共起記事検索部23が全ての組合せを処理対象としたと判断した場合(ステップS130:YES)、関係集合生成部24は、ステップS140の処理を実行する。

【0043】

関係集合生成部24は、ステップS115と同様に、部分組合せ $A_1$ の1つと部分組合せ $B_k$ の1つとからなる全ての組合せを生成する(ステップS135)。関係集合生成部24は、ステップS135において生成した組合せのうち、まだステップS145の処理対象としていない組合せ $A_1 - B_k$ を選択する(ステップS140)。

10

【0044】

関係集合生成部24は、ベース共起記事検索部23により処理結果記憶部14に書き込まれた記事集合Dに含まれる記事データ $d_1$ 、 $d_2$ 、...が示す記事それぞれから、選択した組合せ $A_1 - B_k$ の部分組合せ $A_1$ が示す $n$ 個の単語と、部分組合せ $B_k$ が示す $h$ 個の単語とを関係付ける単語 $r$ を抽出する(ステップS145)。関係集合生成部24は、抽出した単語を処理結果記憶部14に記憶されている関係集合Rに追加する。ただし、関係集合生成部24は、抽出した単語 $r$ がすでに関係集合Rに含まれている場合は追加しない。関係集合Rに含まれる単語 $r$ をそれぞれ単語 $r_1$ 、 $r_2$ 、...とする。次元数 $m$ 、 $g$ が20~50であり、抽出単語数 $n$ 、 $h$ が2~5個である場合、関係集合Rに含まれる単語数は、20~50語程度となる。

20

【0045】

関係集合生成部24は、ステップS135において生成した全ての組合せをステップS145の処理対象としたかを判断する(ステップS150)。まだステップS145の処理対象としていない組合せがあると判断した場合(ステップS150:NO)、関係集合生成部24は、ステップS140からの処理を繰り返す。関係集合生成部24が全ての組合せをステップS145の処理対象としたと判断した場合(ステップS150:YES)、関係妥当性スコア算出部25は、ステップS155の処理を実行する。

30

【0046】

関係妥当性スコア算出部25は、関係集合生成部24により処理結果記憶部14に書き込まれた関係集合Rに含まれる各単語 $r_i$ ( $i$ は1以上の整数)について、以下の式(5)により関係妥当性スコア $score(r_i)$ を算出し、処理結果記憶部14に書き込む(ステップS155)。但し、 $MI(A_1, r_i)$ は、部分組合せ $A_1$ と単語 $r_i$ との相互情報量、 $MI(B_k, r_i)$ は、部分組合せ $B_k$ と単語 $r_i$ との相互情報量、 $MI(A_1 B_k, r_i)$ は、部分組合せ $A_1$ 及び部分組合せ $B_k$ の共起と単語 $r_i$ との相互情報量である。相互情報量とは、2つの確率変数が相互に依存する尺度を表す量である。

【0047】

$score(r_i) = \frac{1}{k} MI(A_1, r_i) * MI(B_k, r_i) * MI(A_1 B_k, r_i) \dots (5)$

40

【0048】

上記のように、式(5)は、相互情報量 $MI(A_1, r_i)$ 、相互情報量 $MI(B_k, r_i)$ 、及び、相互情報量 $MI(A_1 B_k, r_i)$ を乗算した値を、全ての部分組合せ $A_1$ 、全ての部分組合せ $B_k$ について累積加算した値である。従って、相互情報量 $MI(A_1, r_i)$ 、相互情報量 $MI(B_k, r_i)$ 、相互情報量 $MI(A_1 B_k, r_i)$ が大きいほど、つまり、部分組合せ $A_1$ と単語 $r_i$ の結びつきが強いほど、部分組合せ $B_k$ と単語 $r_i$ の結びつきが強いほど、また、部分組合せ $A_1$ 及び部分組合せ $B_k$ の共起と単語 $r_i$ の結びつきが強いほど、関係妥当性スコア $score(r_i)$ も大きくなる。

【0049】

50

なお、関係妥当性スコア算出部 25 は、相互情報量  $MI(A_1, r_i)$ 、相互情報量  $MI(B_k, r_i)$ 、及び、相互情報量  $MI(A_1 B_k, r_i)$  をそれぞれ、式 (6) ~ 式 (8) により算出する。

【 0 0 5 0 】

【 数 2 】

$$MI(A_1, r_i) = \log\left(1 + \frac{p(A_1, r_i)}{p(A_1) * p(r_i)}\right) \quad \dots (6)$$

$$MI(B_k, r_i) = \log\left(1 + \frac{p(B_k, r_i)}{p(B_k) * p(r_i)}\right) \quad \dots (7)$$

$$MI(A_1 B_k, r_i) = \log\left(1 + \frac{p(A_1 B_k, r_i)}{p(A_1 B_k) * p(r_i)}\right) \quad \dots (8)$$

10

【 0 0 5 1 】

但し、 $p(x)$  は、 $x$  が記事記憶装置 5 に記憶されている記事データに出現する確率であり、 $p(x, y)$  は、 $x$  と  $y$  が同時に記事記憶装置 5 に記憶されている記事データに出現する確率である。つまり、 $p(A_1)$  は、部分組合せ  $A_1$  が示す全ての単語が出現する記事データの確率、 $p(B_k)$  は、部分組合せ  $B_k$  が示す全ての単語が出現する記事データの確率、 $p(r_i)$  は、単語  $r_i$  が出現する記事データの確率である。また、 $p(A_1 B_k)$  は、部分組合せ  $A_1$  が示す全ての単語と部分組合せ  $B_k$  が示す全ての単語とが出現する記事データの確率、 $p(A_1, r_i)$  は、部分組合せ  $A_1$  が示す全ての単語と単語  $r_i$  とが出現する記事データの確率、 $p(B_k, r_i)$  は、部分組合せ  $B_k$  が示す全ての単語と単語  $r_i$  とが出現する記事データの確率、 $p(A_1 B_k, r_i)$  は、部分組合せ  $A_1$  が示す全ての単語及び部分組合せ  $B_k$  が示す全ての単語と単語  $r_i$  とが出現する記事データの確率である。

20

【 0 0 5 2 】

関係妥当性スコア算出部 25 は、各確率を以下のように算出する。関係妥当性スコア算出部 25 は、記事記憶装置 5 に記憶されている記事データの数である合計記事数をカウントする。続いて、関係妥当性スコア算出部 25 は、記事記憶装置 5 に記憶されている記事データのうち、部分組合せ  $A_1$  が示す全ての単語が出現する記事データの数、部分組合せ  $B_k$  が示す全ての単語が出現する記事データの数、単語  $r_i$  が出現する記事データの数をカウントし、これらのカウント数をそれぞれ合計記事数で除算することにより  $p(A_1)$ 、 $p(B_k)$ 、 $p(r_i)$  を算出する。

30

【 0 0 5 3 】

さらに、関係妥当性スコア算出部 25 は、記事記憶部 5 に記憶されている記事データのうち部分組合せ  $A_1$  が示す全ての単語と部分組合せ  $B_k$  が示す全ての単語とが出現する記事データの数、部分組合せ  $A_1$  が示す全ての単語と単語  $r_i$  とが出現する記事データの数、部分組合せ  $B_k$  が示す全ての単語と単語  $r_i$  とが出現する記事データの数、部分組合せ  $A_1$  が示す全ての単語及び部分組合せ  $B_k$  が示す全ての単語と単語  $r_i$  とが出現する記事データの数をカウントし、これらのカウント数をそれぞれ合計記事数で除算することにより、 $p(A_1 B_k)$ 、 $p(A_1, r_i)$ 、 $p(B_k, r_i)$ 、 $p(A_1 B_k, r_i)$  を算出する。

40

【 0 0 5 4 】

なお、式 (6) ~ 式 (8) に用いられる  $p(A_1)$ 、 $p(B_k)$ 、 $p(r_i)$ 、 $p(A_1 B_k)$ 、 $p(A_1, r_i)$ 、 $p(B_k, r_i)$ 、 $p(A_1 B_k, r_i)$  を、記事記憶装置 5 に記憶される記事データに基づいて算出した確率としているが、記事集合 D に含まれ

50

る記事データに出現する確率としてもよい。この場合、関係妥当性スコア算出部25は、記事記憶部5に記憶されている記事データに代えて、記事集合Dに含まれる記事データを用い、上記と同様に $p(A_1)$ 、 $p(B_k)$ 、 $p(r_i)$ 、 $p(A_1 B_k)$ 、 $p(A_1, r_i)$ 、 $p(B_k, r_i)$ 、 $p(A_1 B_k, r_i)$ を算出する。

【0055】

図5において、ターゲット状況部分組合せ生成部26は、図4のステップS100において入力された状況Cデータが示すワードベクトルの要素であるm個の単語からn個の単語を選択したときの全ての組合せを生成する。組合せの数は、 ${}_m C_n$ となる。ターゲット状況部分組合せ生成部26は、生成した単語の組合せを要素とするn次元ワードベクトルを示す部分組合せ $C_f$  ( $1 \leq f \leq {}_m C_n$ 、fは整数)を生成し、処理結果記憶部14に書き込む(ステップS200)。

10

【0056】

ターゲット共起記事検索部27は、処理結果記憶部14に書き込まれた部分組合せ $C_f$ の1つと、関係集合Rに含まれる単語 $r_i$ の1つとからなる全ての組合せを生成する(ステップS205)。つまり、生成される組合せは $C_1 - r_1$ 、 $C_1 - r_2$ 、...、 $C_2 - r_1$ 、 $C_2 - r_2$ 、...、 $C_{m C_n} - r_1$ 、 $C_{m C_n} - r_2$ 、...である。なお、ターゲット共起記事検索部27は、関係妥当性スコア $score(r_i)$ が閾値以上の単語 $r_i$ のみ、あるいは、関係妥当性スコア $score(r_i)$ が高いものから所定数の単語 $r_i$ のみを組合せを生成する対象としてもよい。ターゲット共起記事検索部27は、ステップS205において生成した全ての組合せのうち、まだステップS215の処理対象としていない

20

【0057】

ターゲット共起記事検索部27は、ステップS210において選択した組合せ $C_f - r_i$ を構成する部分組合せ $C_f$ 及び単語 $r_i$ を処理結果記憶部14から読み出す。ターゲット共起記事検索部27は、記事記憶装置5が記憶する記事データを検索し、読み出した部分組合せ $C_f$ が示すn個の単語と、単語 $r_i$ とが全て含まれる記事を示す記事データを特定する。ターゲット共起記事検索部27は、特定した記事データを記事記憶装置5から読み出し、読み出した記事データを処理結果記憶部14に記憶されている記事集合Eに書き込む(ステップS215)。ただし、ターゲット共起記事検索部27は、抽出した記事データがすでに記事集合Eに含まれている場合は追加しない。関係集合Eに含まれる記事データをそれぞれ記事データ $e_1$ 、 $e_2$ 、...とする。

30

【0058】

ターゲット共起記事検索部27は、ステップS205において生成した全ての組合せをステップS215の処理対象としたかを判断する(ステップS220)。まだステップS215の処理対象としていない組合せがあると判断した場合(ステップS220:NO)、ターゲット共起記事検索部27は、ステップS210からの処理を繰り返す。ターゲット共起記事検索部27が全ての組合せを処理対象としたと判断した場合(ステップS220:YES)、類推結果生成部28は、ステップS225の処理を実行する。

【0059】

類推結果生成部28は、ステップS205と同様に、部分組合せ $C_f$ の1つと関係集合Rに含まれる単語 $r_i$ の1つとからなる全ての組合せを生成する(ステップS225)。類推結果生成部28は、ステップS225において生成した組合せのうち、まだステップS235の処理対象としていない組合せ $C_f - r_i$ を選択する(ステップS230)。

40

【0060】

類推結果生成部28は、ターゲット共起記事検索部27が処理結果記憶部14に書き込んだ記事集合Eに含まれる各記事データ $e_1$ 、 $e_2$ 、...が示す記事それぞれから、選択した組合せの単語 $r_i$ によって部分組合せ $C_f$ が示すn個の単語と関係付けられる単語xを抽出する(ステップS235)。この時、ターゲットにおける $C_f - r_i - x$ の構造が、ベースで単語 $r_i$ が抽出されたときの $A_1 - r_i - B_k$ の構造と同じになるようにxを選択する。類推結果生成部28は、抽出した単語xを処理結果記憶部14に記憶されている

50

類推結果集合 X に追加する。ただし、類推結果生成部 28 は、抽出した単語 x がすでに類推結果集合 X に含まれている場合は追加しない。類推結果集合 X に含まれる単語 x をそれぞれ、 $x_1$ 、 $x_2$ 、... とする。

【0061】

類推結果生成部 28 は、ステップ S 225 において生成した全ての組合せをステップ S 235 の処理対象としたかを判断する(ステップ S 240)。類推結果生成部 28 は、まだ処理対象としていない組合せがあると判断した場合(ステップ S 240: NO)、ステップ S 230 からの処理を繰り返す。類推結果生成部 28 が全ての組合せをステップ S 235 の処理対象としたと判断した場合(ステップ S 240: YES)、類推結果妥当性スコア算出部 29 は、ステップ S 245 の処理を実行する。

10

【0062】

類推結果妥当性スコア算出部 29 は、処理結果記憶部 14 に記憶されている類推結果集合 X で示される単語  $x_j$  ( $j = 1, 2, \dots$ ) について、以下の式(9)により類推結果妥当性スコア  $score(x_j)$  を算出し、処理結果記憶部 14 に書き込む(ステップ S 245)。但し、 $MI(C_f, x_j)$  は、部分組合せ  $C_f$  と単語  $x_j$  との相互情報量、 $MI(r_i, x_j)$  は、単語  $r_i$  と単語  $x_j$  との相互情報量、 $MI(C_f r_i, x_j)$  は、部分組合せ  $C_f$  及び単語  $r_i$  の共起と単語  $x_j$  との相互情報量である。また、類推結果妥当性スコア算出部 29 は、関係妥当性スコア  $score(r_i)$  を処理結果記憶部 14 から読み出す。

20

【0063】

$$score(x_j) = \prod_i MI(C_f, x_j) * MI(r_i, x_j) * MI(C_f r_i, x_j) \dots (9)$$

【0064】

上記のように、式(9)は、相互情報量  $MI(C_f, x_j)$ 、相互情報量  $MI(r_i, x_j)$ 、相互情報量  $MI(C_f r_i, x_j)$  及び関係妥当性スコア  $score(r_i)$  を乗算した値を、全ての部分組合せ  $C_f$ 、全ての単語  $r_i$  について累積加算した値である。従って、相互情報量  $MI(C_f, x_j)$ 、相互情報量  $MI(r_i, x_j)$ 、相互情報量  $MI(C_f r_i, x_j)$  が大きいほど、つまり、部分組合せ  $C_f$  と単語  $x_j$  の結びつきが強いほど、単語  $r_i$  と単語  $x_j$  の結びつきが強いほど、部分組合せ  $C_f$  及び単語  $r_i$  の共起と単語  $x_j$  の結びつきが強いほど、また、関係妥当性スコア  $score(r_i)$  が大きいほど、類推結果妥当性スコア  $score(x_j)$  も大きくなる。

30

【0065】

なお、類推結果妥当性スコア算出部 29 は、相互情報量  $MI(C_f, x_j)$ 、相互情報量  $MI(r_i, x_j)$ 、及び、相互情報量  $MI(C_f r_i, x_j)$  をそれぞれ、式(10)~式(12)により算出する。

【0066】

【数3】

$$MI(C_f, x_j) = \log\left(1 + \frac{p(C_f, x_j)}{p(C_f) * p(x_j)}\right) \dots (10)$$

40

$$MI(r_i, x_j) = \log\left(1 + \frac{p(r_i, x_j)}{p(r_i) * p(x_j)}\right) \dots (11)$$

$$MI(C_f r_i, x_j) = \log\left(1 + \frac{p(C_f r_i, x_j)}{p(C_f, r_i) * p(x_j)}\right) \dots (12)$$

【0067】

50

但し、 $p(C_f)$ は、部分組合せ $C_f$ で示される全ての単語が出現する記事データの確率、 $p(x_j)$ は、単語 $x_j$ が出現する記事データの確率である。また、 $p(C_f, x_j)$ は、部分組合せ $C_f$ が示す全ての単語と単語 $x_j$ とが出現する記事データの確率、 $p(r_i, x_j)$ は、単語 $r_i$ と単語 $x_j$ が出現する記事データの確率、 $p(C_f, r_i)$ は、部分組合せ $C_f$ が示す全ての単語と単語 $r_i$ とが出現する記事データの確率、 $p(C_f, r_i, x_j)$ は、部分組合せ $C_f$ が示す全ての単語及び単語 $r_i$ と単語 $x_j$ とが出現する記事データの確率である。

【0068】

類推結果妥当性スコア算出部29は、各確率を以下のように算出する。類推結果妥当性スコア算出部29は、記事記憶装置5に記憶されている記事データの数である合計記事数をカウントする。続いて、類推結果妥当性スコア算出部29は、記事記憶装置5に記憶されている記事データのうち、部分組合せ $C_f$ が示す全ての単語が出現する記事データの数、単語 $x_j$ が出現する記事データの数、単語 $r_i$ が出現する記事データの数をカウントし、これらのカウント数をそれぞれ合計記事数で除算することにより $p(C_f)$ 、 $p(x_j)$ 、 $p(r_i)$ を算出する。

10

【0069】

さらに、類推結果妥当性スコア算出部29は、記事記憶部5に記憶されている記事データのうち部分組合せ $C_f$ が示す全ての単語と単語 $x_j$ とが出現する記事データの数、単語 $r_i$ と単語 $x_j$ とが出現する記事データの数、部分組合せ $C_f$ が示す全ての単語と単語 $r_i$ とが出現する記事データの数、部分組合せ $C_f$ が示す全ての単語及び単語 $r_i$ と単語 $x_j$ とが出現する記事データの数をカウントし、これらのカウント数をそれぞれ合計記事数で除算することにより、 $p(C_f, x_j)$ 、 $p(r_i, x_j)$ 、 $p(C_f, r_i)$ 、 $p(C_f, r_i, x_j)$ を算出する。

20

【0070】

なお、式(10)～式(12)に用いられる $p(C_f)$ 、 $p(x_j)$ 、 $p(r_i)$ 、 $p(C_f, x_j)$ 、 $p(r_i, x_j)$ 、 $p(C_f, r_i)$ 、 $p(C_f, r_i, x_j)$ を、記事記憶装置5に記憶される記事データに基づいて算出した確率としているが、記事集合Eに含まれる記事データに出現する確率としてもよい。この場合、類推結果妥当性スコア算出部29は、記事記憶部5に記憶されている記事データに代えて、記事集合Eに含まれる記事データを用い、上記と同様に $p(C_f)$ 、 $p(x_j)$ 、 $p(r_i)$ 、 $p(C_f, x_j)$ 、 $p(r_i, x_j)$ 、 $p(C_f, r_i)$ 、 $p(C_f, r_i, x_j)$ を算出する。

30

【0071】

出力部16は、類推結果生成部28が処理結果記憶部14に書き込んだ類推結果集合Xが示す各単語 $x_j$ と、類推結果妥当性スコア算出部29が書き込んだ当該単語 $x_j$ の類推結果妥当性スコア $score(x_j)$ とからなる類推結果データをディスプレイに表示させるなどして出力する(ステップS245)。このとき、出力部16は、類推結果妥当性スコアが閾値以上の単語 $x_j$ とその類推結果妥当性スコア $score(x_j)$ のみを出力するようにしてもよく、類推結果妥当性スコアが高い順に所定数の単語 $x_j$ とその類推結果妥当性スコア $score(x_j)$ のみを出力してもよい。

40

【0072】

続いて、図4に示すステップS145における単語 $r$ の抽出処理、図5に示すステップS235における単語 $x$ の抽出処理の詳細な処理手順を説明する。以下では、2つの抽出処理について示しているが、事例や類推の対象に応じていずれを用いてもよい。ここでは、簡単のため、状況A、結果B、状況Cとも3次元のワードベクトルであり、抽出単語数 $n$ 、 $h$ が2である場合を例にして説明する。

【0073】

この場合、図4のステップS105において、状況Aのワードベクトル $(t_{a1}, t_{a2}, t_{a3})$ から部分組合せ $A_1(t_{a1}, t_{a2})$ 、 $A_2(t_{a1}, t_{a3})$ 、 $A_3(t_{a2}, t_{a3})$ が生成され、ステップS110において、結果Bのワードベクトルから部分組合せ $B_1(w_{b1}, w_{b2})$ 、 $B_2(w_{b1}, w_{b3})$ 、 $B_3(w_{b2}, w_{b3})$

50

が生成される。従って、ステップS 1 3 5においては、組合せ $A_1 - B_1$ 、 $A_1 - B_2$ 、 $A_1 - B_3$ 、 $A_2 - B_1$ 、 $A_2 - B_2$ 、 $A_2 - B_3$ 、 $A_3 - B_1$ 、 $A_3 - B_2$ 、 $A_3 - B_3$ が生成される。

【0074】

(抽出処理1)：記事データは、自然言語で記述された文書のテキストデータである。ステップS 1 4 5において、関係集合生成部2 4は、記事集合Dに含まれる記事データ $d_1$ 、 $d_2$ 、...が示す文書の形態素解析を行う。関係集合生成部2 4は、形態素解析の結果を参照し、部分組合せ $A_1$ が示す $n$ 個の単語が主語に含まれ、かつ、部分組合せ $B_k$ が示す $h$ 個の単語が述部にある名詞として含まれている文から、部分組合せ $A_1$ が示す $n$ 個の単語と部分組合せ $B_k$ が示す $h$ 個の単語を関係づける述部の動詞を単語 $r$ として抽出する。

10

また、ステップS 2 3 5において、類推結果生成部2 8は、記事集合Eに含まれる記事データ $e_1$ 、 $e_2$ 、...が示す文書の形態素解析を行う。類推結果生成部2 8は、形態素解析の結果を参照し、部分組合せ $C_f$ が示す $n$ 個の単語が主語に含まれ、かつ、単語 $r_i$ が述部の動詞として含まれている文から、述部にある名詞を単語 $x$ として抽出する。

【0075】

図6は、ステップS 1 4 0における単語 $r$ の抽出処理を説明するための図である。同図においては、ステップS 1 3 5において部分組合せ $A_1$ (サメ, マグロ)及び部分組合せ $B_1$ (ひれ, 尾)の組合せが選択されている例を示している。関係集合生成部2 4は、記事データ $d_1$ が示す文「サメやマグロのような魚類は、ひれや尾を使って高速に泳ぐことができる。」の主語は部分組合せ $A_1$ が示す単語「サメ」及び「マグロ」であり、述部にある名詞は部分組合せ $B_1$ が示す単語「ひれ」及び「尾」であるため、当該文の述部の動詞「泳ぐ」を単語 $r_1$ として抽出する。また、関係集合生成部2 4は、記事データ $d_2$ が示す文「サメやマグロのような魚類は、ひれや尾を使って広範囲を移動することができる。」の主語は部分組合せ $A_1$ が示す単語「サメ」及び「マグロ」であり、述部にある名詞は部分組合せ $B_1$ が示す単語「ひれ」及び「尾」であるため、当該文の述部の動詞「移動する」を単語 $r_2$ として抽出する。

20

【0076】

このように、2以上の単語からなる部分組合せを利用することによって、1つの単語を用いる場合よりも、目的とする関係を高精度に抽出することができる。部分組合せに含まれる単語数が多いほどより正確な関係を示す単語を抽出することが可能となるが、関係の抽出対象となる文は減少する。

30

【0077】

図7は、ステップS 2 3 5における単語 $x$ の抽出処理を説明するための図である。同図は、ステップS 2 3 0において組合せの一方として部分組合せ $C_1$ (ライオン, オオカミ)が選択された場合について示している。類推結果生成部2 8は、記事集合Eに含まれるいずれの記事データからも、部分組合せ $C_1$ が示す単語「ライオン」及び「オオカミ」が主語に含まれ、かつ、単語 $r_1$ 「泳ぐ」が述部の動詞である文を含む文は検出されなかったとする。一方、類推結果生成部2 8は、記事集合Eに含まれる記事データ $e_1$ が示す文「ライオンやオオカミなどは足を使って広範囲を移動する必要があり...」の主語は部分組合せ $C_1$ が示す単語「ライオン」及び「オオカミ」であり、述部の動詞は単語 $r_2$ 「移動する」であるため、当該文の述部にある名詞「足」を単語 $x_1$ として抽出する。

40

【0078】

(抽出処理2)：記事データは、自然言語で記述された文書のテキストデータであり、記事記憶装置5は、分野別の辞書データ(コーパス)をさらに記憶している。また、状況A、結果B、関係集合Rは、特定の分野の用語とする。さらに、図4のステップS 1 0 0において、類推装置1の入力部1 2は、さらに、分野を示す情報の入力を受ける。

【0079】

図4のステップS 1 4 5において、関係集合生成部2 4は、入力された分野の情報に対応した辞書データを特定し、部分組合せ $A_1$ が示す $n$ 個の単語と、部分組合せ $B_k$ が示す

50



h個の単語が含まれる文から、特定された辞書データに登録されている単語を単語rの候補r'として抽出する。抽出された候補r'をそれぞれ、候補r<sub>1</sub>'、r<sub>2</sub>'...とする。関係集合生成部24は、抽出した候補r<sub>y</sub>'(y=1、2、...)と、部分組合せA<sub>1</sub>、部分組合せB<sub>k</sub>それぞれとの関連の強さに基づいて関係を表す単語としての妥当性を示す値を以下の式(13)により算出する。但し、MI(A<sub>1</sub>, r<sub>y</sub>')は、部分組合せA<sub>1</sub>と候補r<sub>y</sub>'との相互情報量、MI(B<sub>k</sub>, r<sub>y</sub>')は、部分組合せB<sub>k</sub>と候補r<sub>y</sub>'との相互情報量である。なお、相互情報量MI(A<sub>1</sub>, r<sub>y</sub>')、相互情報量MI(B<sub>k</sub>, r<sub>y</sub>')は、単語r<sub>i</sub>の代わりに候補r<sub>y</sub>'を用いることにより、式(6)、式(7)と同様に算出される。

【0080】

$$\text{スコア}(r_y') = \text{MI}(A_1, r_y') + \text{MI}(B_k, r_y') \quad \dots (13)$$

【0081】

関係集合生成部24は、算出した値が、関連が強いと判断する所定の条件以上である候補r'を、単語rとして選択する。

また、ステップS230において、類推結果生成部28は、部分組合せC<sub>f</sub>が示すn個の単語と、単語r<sub>i</sub>が含まれる文から、入力された分野の情報に対応した辞書データに登録されている単語を単語xとして抽出する。

【0082】

図8は、単語r及び単語xの抽出処理を説明するための図である。同図においては、分野の情報がコンピュータであり、ステップS140において部分組合せA<sub>1</sub>(画像、写真)及び部分組合せB<sub>1</sub>(GIF、JPEG)の組合せが選択されている場合の例について示している。記事データd<sub>1</sub>が示す文「画像や写真の圧縮には、GIF、JPEGなどのファイル形式が使えます。」には、部分組合せA<sub>1</sub>が示す単語「画像」及び「写真」と、部分組合せB<sub>1</sub>が示す単語「GIF」及び単語「JPEG」が含まれている。関係集合生成部24は、分野の情報からコンピュータ辞書データを特定し、この文に含まれる単語のうち、コンピュータ辞書データに登録されている「圧縮」、「ファイル形式」を候補r<sub>1</sub>'、r<sub>2</sub>'として抽出する。関係集合生成部24は、これらの抽出した候補r<sub>1</sub>'、r<sub>2</sub>'について、上記の式(13)によりスコアを算出した結果、候補r<sub>1</sub>'「圧縮」は単語rとして選択せず、候補r<sub>2</sub>'「ファイル形式」を単語r<sub>1</sub>として選択する。

【0083】

類推結果生成部28は、ステップS230において部分組合せC<sub>1</sub>(音楽、会話)と単語r<sub>1</sub>'「ファイル形式」の組合せが選択されている場合、部分組合せC<sub>1</sub>が示す単語「音楽」及び「会話」、ならびに、単語r<sub>1</sub>'「ファイル形式」が含まれる文から、記事記憶装置5が記憶しているコンピュータ辞書データに登録されている「MP3」を単語x<sub>1</sub>として抽出する。

【0084】

[第2の実施形態]

続いて、本発明の他の実施形態を説明する。

第1の実施形態では、式(4)に示したように、ベースとなる事例である「A:B」は一对のみ存在し、この事例によりターゲットの状況Cから結果Xを求めていた。つまり、第1の実施形態において、類推装置1は、多次元一事例における四項類推の処理をおこなっていた。一方、本実施形態では、「A:B」で示されるベースの事例が複数存在する場合にターゲットの状況Cから結果Xを求める、多次元多事例の四項類推の処理を考える。これを、式(4)で示したmodus ponensと同じ表現方法で記述すると、以下の式(14)のようになる。

【0085】

10

20

30

40

【数4】

$$\begin{aligned}
 &A(1) (t_{a11}, t_{a12}, \dots, t_{a1m}) : B(1) (w_{b11}, w_{b12}, \dots, w_{b1g}) \\
 &A(2) (t_{a21}, t_{a22}, \dots, t_{a2m}) : B(2) (w_{b21}, w_{b22}, \dots, w_{b2g}) \\
 &\quad \vdots \\
 &A(N) (t_{aN1}, t_{aN2}, \dots, t_{aNm}) : B(N) (w_{bN1}, w_{bN2}, \dots, w_{bNg}) \\
 &\hline
 &C (t_{c1}, t_{c2}, \dots, t_{cm}) : X (w_{x1}, w_{x2}, \dots, w_{xg}) ?
 \end{aligned}$$

... (14)

10

【0086】

上記のように、本実施形態では、事例がN個（Nは2以上の整数）あり、各事例を事例（I）とする（Iは2以上N以下の整数）。事例（I）は、ベースの状況A（I）と、ベースの状況A（I）での結果B（I）とからなる。本実施形態の類推装置は、A（I）：B（I）を用いて、状況Cに対応する結果Xを求める。

【0087】

なお、状況A（I）は、m個（mは2以上の整数）の単語  $t_{aI1}, t_{aI2}, \dots, t_{aIm}$  を要素とするm次元ワードベクトルであり、ターゲットの状況Cも、第1の実施形態と同様のm次元ワードベクトルである。また、結果B（I）は、g個（gは2以上の整数）の単語  $w_{bI1}, w_{bI2}, \dots, w_{bIg}$  を要素とするg次元ワードベクトルであり、ターゲットの結果Xも、第1の実施形態と同様のg次元ワードベクトルである。

20

【0088】

図9は、本実施形態による類推装置の動作概要を示す図である。

本実施形態では、事例が多事例であるため、以下の式（15）のように表わすことができる。

【0089】

$$A(I) : B(I) = C : X(I) ? \quad (I \text{ は } 2 \text{ 以上 } N \text{ 以下の整数}) \quad \dots (15)$$

【0090】

同図に示すように、本実施形態の類推装置は、各事例（I）について第1の実施形態の類推装置1と同様の関係抽出処理を行なうことによって、状況A（I）の部分組合せと結果B（I）の部分組合せとの関係を示す単語の集合である関係集合R（I）を生成する。本実施形態の類推装置は、各関係集合R（I）について第1の実施形態の類推装置1と同様に関係マッピング処理を行なうことによって、類推結果集合X（I）を生成し、類推結果妥当性スコアを算出する。本実施形態の類推装置は、類推結果集合X（1）～X（N）を統合し、重複する単語xがある場合は、重複を削除する。重複削除後の類推結果妥当性スコアは、重複する単語xについての類推結果妥当性スコアを合計した値である。

30

【0091】

図10は、本発明の第2の実施形態による類推装置1aの構成を示すブロック図である。同図において、図3に示す第1の実施形態による類推装置1と同一の部分には同一の符号を付し、その説明を省略する。図10に示す類推装置1aが、図3に示す第1の実施形態の類推装置1と異なる点は、事例記憶部11に代えて事例記憶部11aを備える点、類推処理制御部17及び類推結果積算部18を備える点である。

40

【0092】

事例記憶部11aは、状況A（I）のワードベクトルを示す状況A（I）データ（ベース状況データ）と、結果B（I）のワードベクトルを示す結果B（I）データ（ベース結果データ）とからなる事例（I）を記憶する。類推処理制御部17は、第1の実施形態と同様の処理を事例記憶部11aに記憶されている事例（1）～事例（N）について処理を行なうよう関係抽出部13及び関係マッピング部15に指示する。類推結果積算部18は、各事例（I）について得られた類推結果の単語xの集合を示すデータである類推結果X

50

( I ) を統合する。

【 0 0 9 3 】

図 1 1 及び図 1 2 は、図 1 0 に示す類推装置 1 a の処理フローを示す図である。

類推装置 1 a の事例記憶部 1 1 a は、状況 A ( I ) データと結果 B ( I ) データとからなる事例 ( I ) を記憶している ( I は 2 以上 N 以下の整数)。状況 A ( I ) データは、m 個 ( m は 2 以上の整数) の単語  $t_{a I 1}, t_{a I 2}, \dots, t_{a I m}$  を要素とする m 次元ワードベクトルを示す ( m は 2 以上の整数)。結果 B ( I ) データは、g 個 ( g は 2 以上の整数) の単語  $w_{b I 1}, w_{b I 2}, \dots, w_{b I g}$  を要素とする g 次元ワードベクトルを示す。また、処理結果記憶部 1 4 は、初期値 N U L L の記事集合 D、記事集合 E、関係集合 R ( I )、類推結果集合 X ( I ) を記憶する。

10

【 0 0 9 4 】

図 1 1 において、類推装置 1 a の入力部 1 2 は、図 4 に示す第 1 の実施形態のステップ S 1 0 0 と同様に、状況 C データの入力を受ける ( ステップ S 3 0 0 )。類推処理制御部 1 7 は、事例記憶部 1 1 a に記憶されている事例 ( I ) のうち、まだ処理対象としていない事例 ( I ) を選択する ( ステップ S 3 0 5 )。類推処理制御部 1 7 は、選択した事例 ( I ) についての処理を実行するよう、関係抽出部 1 3 に指示する。これにより、関係抽出部 1 3 は、類推処理制御部 1 7 により選択された事例 ( I ) の状況 A ( I ) データ、結果 B ( I ) データを事例記憶部 1 1 から読み出し、第 1 の実施形態における状況 A データ、結果 B データの代わりに用いて、図 4 に示す第 1 の実施形態におけるステップ S 1 0 5 ~ S 1 5 5 と同様の処理を実行する ( ステップ S 3 1 0 ~ S 3 6 0 )。これにより、事例 ( I ) についての関係集合 R と、関係集合 R に含まれる各単語  $r_i$  の関係妥当性スコア  $score(r_i)$  が得られる。事例 ( I ) を用いて得られた関係集合 R を関係集合 R ( I ) とし、関係集合 R に含まれる単語  $r_1, r_2, \dots$  をそれぞれ単語  $r_{I 1}, r_{I 2}, \dots$  とし、単語  $r_i$  について算出された関係妥当性スコア  $score(r_i)$  を、関係妥当性スコア  $score(r_{I i})$  とする。これらのデータは、事例 ( I ) の識別情報と対応づけて処理結果記憶部 1 4 に書き込まれる。

20

【 0 0 9 5 】

類推処理制御部 1 7 は、事例記憶部 1 1 a に記憶されている事例 ( I ) に、まだ処理対象としていない事例 ( I ) があるかを判断する。まだ処理対象としていない事例 ( I ) があると判断した場合 ( ステップ S 3 6 5 : N O )、類推処理制御部 1 7 は、ステップ S 3 0 5 からの処理を繰り返す。類推処理制御部 1 7 が全ての事例 ( I ) を処理対象としたと判断した場合 ( ステップ S 3 6 5 : Y E S )、関係マッピング部 1 5 を起動する。

30

【 0 0 9 6 】

図 1 2 において、ターゲット状況部分組合せ生成部 2 6 は、図 5 に示す第 1 の実施形態のステップ S 2 0 0 と同様に、状況 C データが示すワードベクトルから部分組合せ  $C_f ( 1 \leq f \leq m, C_n, f \text{ は整数} )$  を生成し、処理結果記憶部 1 4 に書き込む ( ステップ S 3 0 0 )。

【 0 0 9 7 】

類推処理制御部 1 7 は、まだ関係マッピング部 1 5 における処理対象としていない事例 ( I ) を特定する ( ステップ S 3 0 5 )。類推処理制御部 1 7 は、特定した事例 ( I ) についての処理を実行するよう、関係マッピング部 1 5 に指示する。これにより、関係マッピング部 1 5 は、類推処理制御部 1 7 により選択された事例 ( I ) の識別情報と対応付けて処理結果記憶部 1 4 に記憶されている関係集合 R ( I ) 及び関係妥当性スコア  $score(r_{I i})$  を読み出し、第 1 の実施形態における関係集合 R、関係妥当性スコア  $score(r_i)$  の代わりに用いて、図 5 に示す第 1 の実施形態におけるステップ S 2 1 0 ~ S 2 4 5 の処理を実行する ( ステップ S 4 1 5 ~ S 4 4 0 )。これにより類推結果集合 X と、類推結果集合 X に含まれる各単語  $x_j$  の類推結果妥当性スコア  $score(x_j)$  が得られる。事例 ( I ) に対応した関係集合 R ( I ) 及び関係妥当性スコア  $score(r_{I j})$  を用いて得られたこの類推結果集合 X を類推結果集合 X ( I ) とし、類推結果集合 X に含まれる単語  $x_1, x_2, \dots$  をそれぞれ単語  $x_{I 1}, x_{I 2}, \dots$  とし、単語  $x_j$  につ

40

50

いて算出された類推結果妥当性スコア  $score(x_j)$  を、関係妥当性スコア  $score_I(x_{Ij})$  とする。これらのデータは、事例 (I) の識別情報と対応づけて処理結果記憶部 14 に書き込まれる。

【0098】

類推処理制御部 17 は、まだ関係マッピング部 15 における処理対象としていない事例 (I) があるかを判断する。まだ処理対象としていない事例 (I) があると判断した場合 (ステップ S455: NO)、類推処理制御部 17 は、ステップ S405 からの処理を繰り返す。類推処理制御部 17 が全ての事例 (I) を処理対象としたと判断した場合 (ステップ S455: YES)、類推結果積算部 18 を起動する。

【0099】

類推結果積算部 18 は、以下に示す式 (16) のように、処理結果記憶部 14 から読み出した類推結果集合  $X(1) \sim (N)$  を統合したデータである統合類推結果集合  $X$  を生成する。

【0100】

$$X = \bigcup_I X(I) \quad (I = 1 \sim N) \quad \dots (16)$$

【0101】

類推結果積算部 18 は、類推結果集合  $X(1) \sim (N)$  に含まれる単語  $x_{Ij}$  に重複があれば一つの単語のみを残して重複をなくし、重複をなくした単語  $x_{Ij}$  の集合を示すデータである統合類推結果集合  $X$  を生成する。統合類推結果集合  $X$  に含まれる単語を  $x_1$ 、 $x_2$ 、... とする。

【0102】

続いて、類推結果積算部 18 は、以下の式 (17) により、統合類推結果集合  $X$  に含まれる各単語  $x_j$  の類推結果妥当性スコアを算出する。ただし、 $score_I(x_j)$  は、事例 (I) について得られた単語  $x_j$  の類推結果妥当性スコアである。

【0103】

$$score(x_j) = \bigcup_I score_I(x_j) \quad \dots (17)$$

【0104】

つまり、統合類推結果集合  $X$  に含まれる単語  $x_j$  の類推結果妥当性スコアは、当該単語  $x_j$  に統合された単語  $x_{Ij}$  の類推結果妥当性スコア  $score_I(x_{Ij})$  を積算した値である。類推結果積算部 18 は、統合類推結果集合  $X$  に含まれる単語  $x_j$  及び当該単語  $x_j$  の類推結果妥当性スコア  $score(x_j)$  を処理結果記憶部 14 に書き込む (ステップ S460)。

【0105】

出力部 16 は、ステップ S445 において類推結果積算部 18 が処理結果記憶部 14 に書き込んだ統合類推結果集合  $X$  が示す各単語  $x_j$  及び当該単語  $x_j$  の類推結果妥当性スコア  $score(x_j)$  とからなる類推結果データをディスプレイに表示させるなどして出力する (ステップ S465)。このとき、出力部 16 は、類推結果妥当性スコアが閾値以上の単語  $x_j$  のみを出力するようにしてもよく、類推結果妥当性スコアが高い順に所定数の単語  $x_j$  のみを出力してもよい。

【0106】

上述した実施形態では、記事記憶装置 5 を第 1 の実施形態の類推装置 1、第 2 の実施形態の類推装置 1a とネットワークを介して接続される外部の装置としているが、類推装置 1、類推装置 1a が記事記憶装置 5 を内部に備える構成としてもよい。

また、上述した実施形態では、記事集合 D、E を、記事データの集合としているが、記事データが記憶されている URL (Universal Resource Locator) や記事データのデータ名など、記事データの識別情報を示すデータでもよい。この場合、類推装置 1 及び類推装置 1a は、記事集合 D、E に含まれる記事データの識別情報で特定される記事記憶装置 5 内の記事データを参照し、上述した処理を実行する。

また、上述した実施形態では、状況 A、状況 A(1) ~ 状況 A(N) と状況 C のワードベクトルの次元数が同じ場合について説明したが、状況 A、状況 A(1) ~ 状況 A(N)

10

20

30

40

50

の次元数と状況Cのワードベクトルの次元数とが異なってもよい。また、状況A(1)～状況A(N)のワードベクトルの次元数は全て同一でなくともよい。また、結果B(1)～結果B(N)のワードベクトルの次元数は全て同一でなくともよい。

【0107】

上述した類推装置1及び類推装置1aは、内部にコンピュータシステムを有している。そして、類推装置1の関係抽出部13、関係マッピング部15及び出力部16、ならびに、類推装置1aの関係抽出部13、関係マッピング部15、出力部16、類推処理制御部17及び類推結果積算部18の動作の過程は、プログラムの形式でコンピュータ読み取り可能な記録媒体に記憶されており、このプログラムをコンピュータシステムが読み出して実行することによって、上記処理が行われる。ここでいうコンピュータシステムとは、CPU及び各種メモリやOS、周辺機器等のハードウェアを含むものである。

10

【0108】

また、「コンピュータシステム」は、WWWシステムを利用している場合であれば、ホームページ提供環境(あるいは表示環境)も含むものとする。

また、「コンピュータ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。さらに「コンピュータ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、短時間の間、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、一定時間プログラムを保持しているものも含むものとする。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良く、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合せで実現できるものであっても良い。

20

【符号の説明】

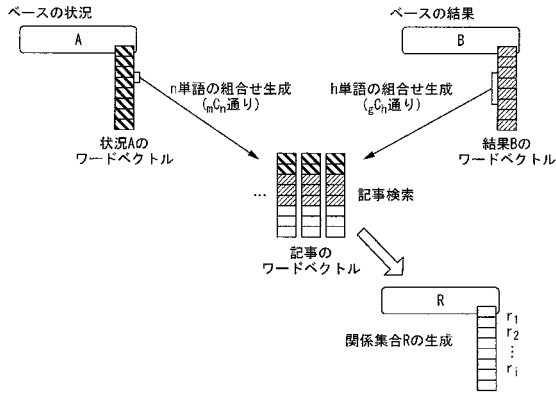
【0109】

- 1、1a...類推装置
- 11、11a...事例記憶部
- 12...入力部
- 13...関係抽出部
- 14...処理結果記憶部
- 15...関係マッピング部
- 16...出力部
- 17...類推処理制御部
- 18...類推結果積算部
- 21...ベース状況部分組合せ生成部
- 22...ベース結果部分組合せ生成部
- 23...ベース共起記事検索部
- 24...関係集合生成部
- 25...妥当性スコア算出部
- 26...ターゲット状況部分組合せ生成部
- 27...ターゲット共起記事検索部
- 28...類推結果生成部
- 29...類推結果妥当性スコア算出部
- 5...記事記憶装置

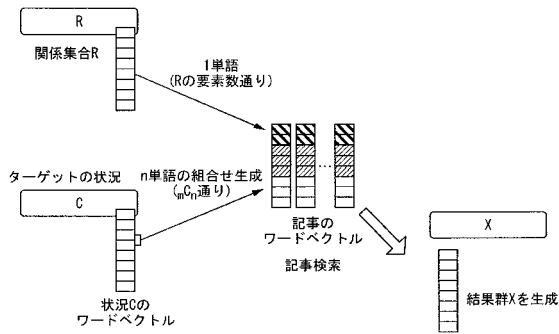
30

40

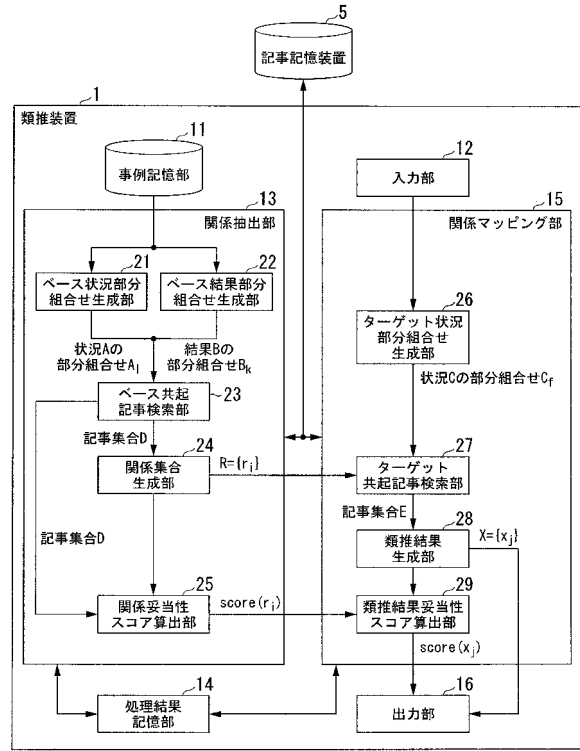
【図1】



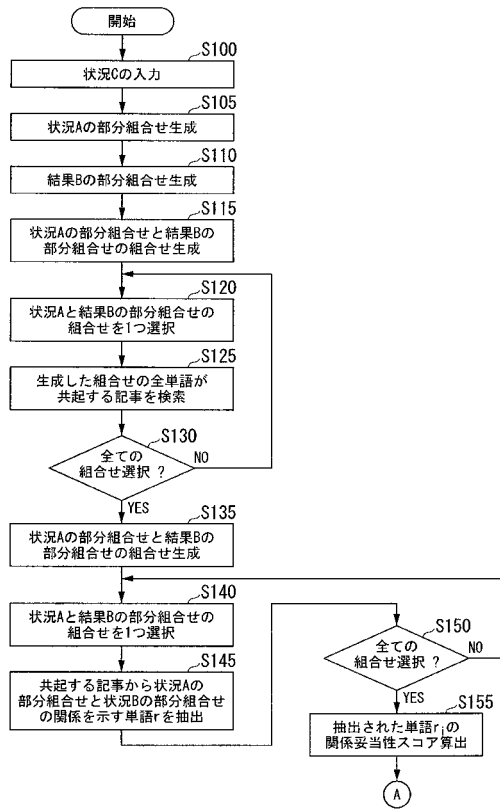
【図2】



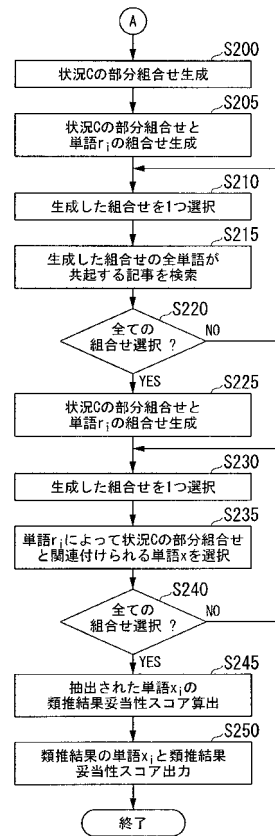
【図3】



【図4】



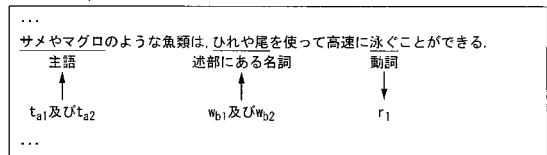
【図5】



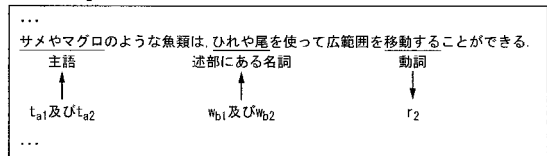
【図 6】

$A_1(t_{a1}, t_{a2})=(サメ, マグロ), B_1(w_{b1}, w_{b2})=(ひれ, 尾)$

記事データ $d_1$



記事データ $d_2$

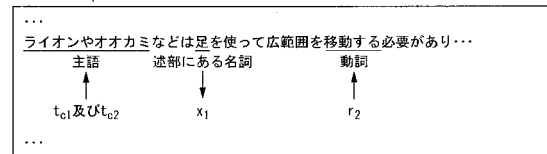


【図 7】

$C_1(t_{c1}, t_{c2})=(ライオン, オオカミ), r_1=(泳ぐ)$   
→ヒットする記事データなし

$C_1(t_{c1}, t_{c2})=(ライオン, オオカミ), r_2=(移動する)$

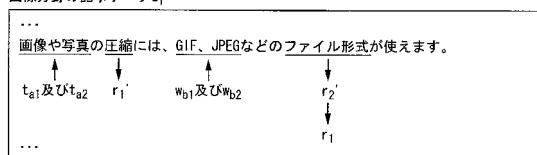
記事データ $e_1$



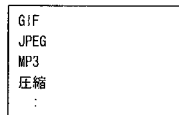
【図 8】

$A_1(t_{a1}, t_{a2})=(画像, 写真), B_1(w_{b1}, w_{b2})=(GIF, JPEG)$

画像分野の記事データ $d_1$

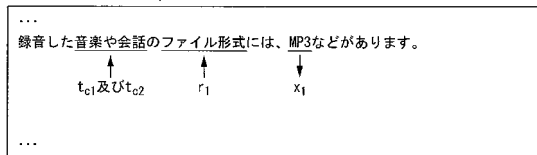


コンピュータ辞書

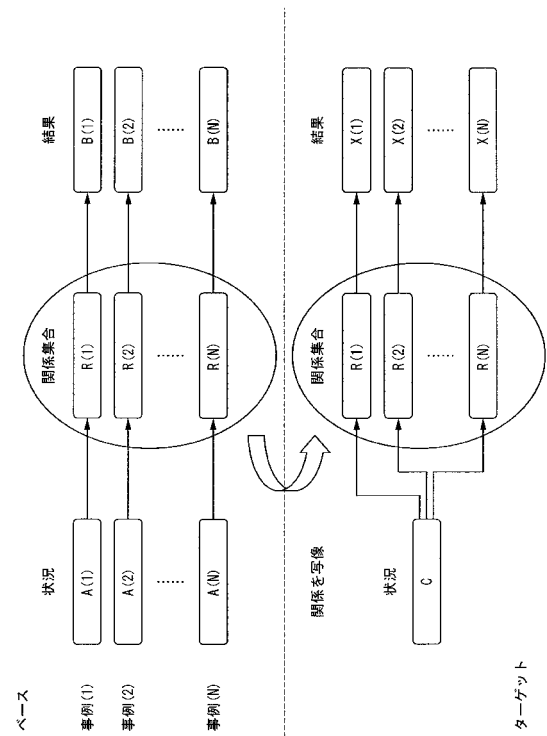


$C_1(t_{c1}, t_{c2})=(音楽, 会話), r_1=(ファイル形式)$

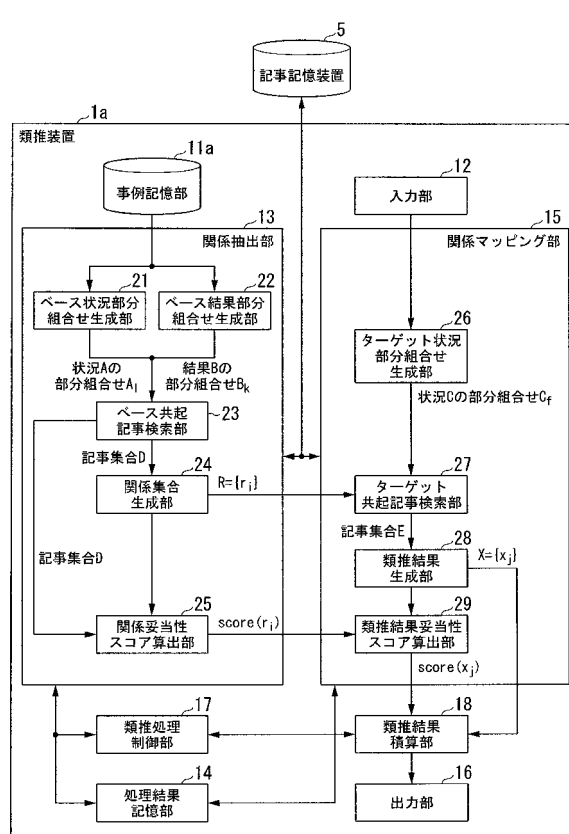
音響分野の記事データ $e_1$



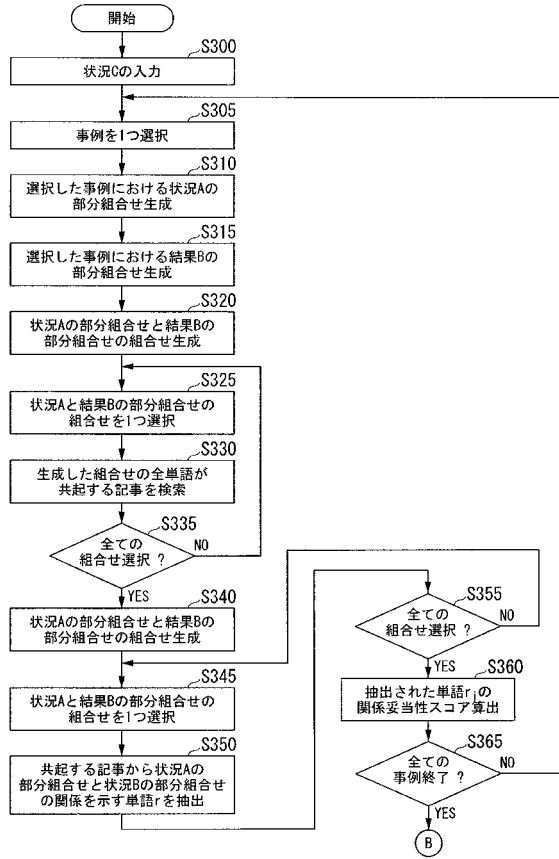
【図 9】



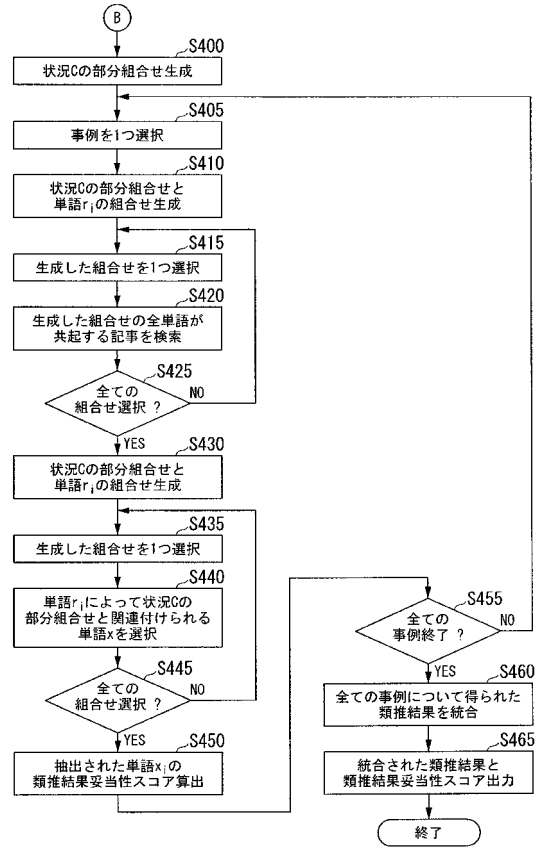
【図 10】



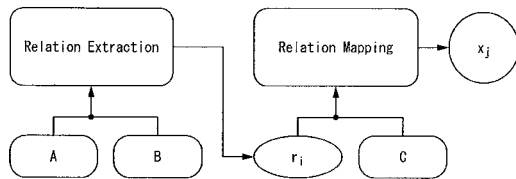
【図11】



【図12】



【図13】





---

フロントページの続き

審査官 久々宇 篤志

- (56)参考文献 特開2010-134844(JP,A)  
米国特許第06101490(US,A)  
Nguyen Tuan Duc 他,単語ペア間の潜在的関係を利用する関係検索エンジン,第72回(平成22年)全国大会講演論文集(5),社団法人情報処理学会,2010年3月8日,5-255~5-256  
加藤 誠 他,関係の類似性に基づくWebからのオブジェクト名検索,情報処理学会論文誌データベース Vol.2 No.2[CD-ROM],日本,社団法人情報処理学会,2009年11月15日,pp.110-125

- (58)調査した分野(Int.Cl.,DB名)  
G06F 17/30  
Cini  
JSTPlus(JDreamIII)