(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2011-180746
(P2011-180746A)
(43)公開日　平成23年9月15日(2011.9.15)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/27	(2006.01)	G06F	17/27	Z	5B075
G06F 17/30	(2006.01)	G06F	17/30	170A	5B091
G06F 3/048	(2006.01)	G06F	17/30	320D	5E501
		G06F	3/048	656C	

審査請求 未請求　請求項の数 12　ＯＬ　（全 38 頁）

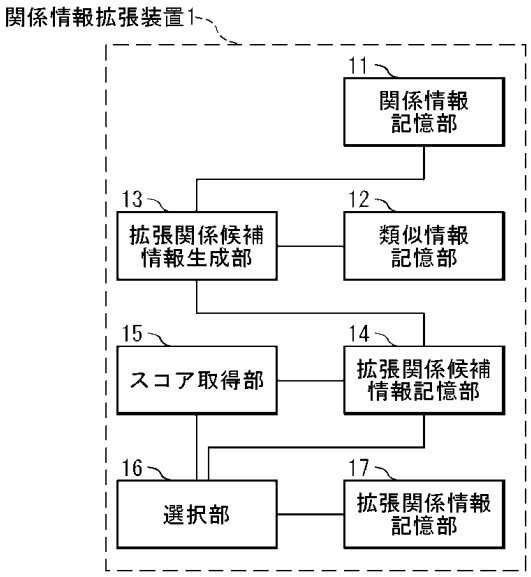
(21)出願番号　特願2010-42938(P2010-42938)
(22)出願日　平成22年2月26日(2010.2.26)

(71)出願人　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100115749
弁理士　谷川　英和
(74)代理人　100121223
弁理士　森本　悟道
(72)発明者　土田　正明
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　ステイン　デ　サーガ
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

最終頁に続く

(54)【発明の名称】関係情報拡張装置、関係情報拡張方法、及びプログラム

(57)【要約】
【課題】意味的関係を有する２以上の言語表現を有する既存の関係情報を元に、新たな関係を獲得することができる関係情報拡張装置を提供する。
【解決手段】関係情報が記憶される関係情報記憶部１１と、類似する２以上の言語表現を有する類似情報が２以上記憶される類似情報記憶部１２と、関係情報に含まれる少なくとも１個の言語表現を、類似情報を用いて、その言語表現に類似する言語表現に置換した拡張関係候補情報を生成して拡張関係候補情報記憶部１４に蓄積する拡張関係候補情報生成部１３と、拡張関係候補情報が意味的関係を有する確からしさを示すスコアを取得するスコア取得部１５と、そのスコアを用いて、拡張関係候補情報のうち、スコアの高い拡張関係候補情報である拡張関係情報を選択する選択部１６と、を備える。
【選択図】図１



【特許請求の範囲】
【請求項１】
意味的関係を有する２以上の言語表現を有する関係情報が記憶される関係情報記憶部と、
類似する２以上の言語表現を有する類似情報が２以上記憶される類似情報記憶部と、
少なくとも１個の言語表現が類似する言語表現に置換された関係情報である拡張関係候補情報が記憶される拡張関係候補情報記憶部と、
前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも１個の言語表現を、前記類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張関係候補情報を生成し、当該拡張関係候補情報を前記拡張関係候補情報記憶部に蓄積する拡張関係候補情報生成部と、
前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的関係を有する確からしさを示すスコアを取得するスコア取得部と、
前記スコア取得部が取得したスコアを用いて、前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報のうち、当該スコアの高い拡張関係候補情報である拡張関係情報を選択する選択部と、を備えた関係情報拡張装置。
【請求項２】
前記スコア取得部は、前記拡張関係候補情報に含まれる２以上の言語表現の共起に関するスコアである共起スコアを取得する、請求項１記載の関係情報拡張装置。
【請求項３】
前記スコア取得部は、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアを取得する、請求項１または請求項２記載の関係情報拡張装置。
【請求項４】
前記スコア取得部は、前記拡張関係候補情報に含まれる２以上の言語表現の共起に関するスコアである共起スコアと、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアとを取得し、
前記選択部は、共起スコアが高く、かつ、経由スコアが高い拡張関係候補情報を選択する、請求項１記載の関係情報拡張装置。
【請求項５】
前記スコア取得部は、前記拡張関係候補情報に含まれる２以上の言語表現と、当該拡張関係候補情報の生成時に用いられた関係情報の意味的関係と同じ種類の意味的関係を有する各関係情報に含まれる２以上の言語表現に対して共起の高い言語表現である共起言語表現とが共起する方が、前記拡張関係候補情報に含まれる２以上の言語表現のみが共起するよりも高い値となる共起スコアを取得する、請求項２または請求項４記載の関係情報拡張装置。
【請求項６】
前記関係情報は、当該関係情報が有する２以上の言語表現の意味的関係の種類を識別する情報である種類識別情報をも有するものであり、
前記拡張関係候補情報生成部は、拡張関係候補情報の生成に用いる関係情報が有する種類識別情報を有する拡張関係候補情報を生成し、
種類識別情報と、当該種類識別情報に対応する、当該種類識別情報で識別される意味的関係の種類に対応する１以上の共起言語表現とを有する対応情報が１以上記憶される対応情報記憶部をさらに備え、
前記スコア取得部は、前記拡張関係候補情報に含まれる２以上の言語表現と、当該拡張関係候補情報が有する種類識別情報に対応する各共起言語表現とが共起する方が、前記拡張関係候補情報に含まれる２以上の言語表現のみが共起するよりも高い値となる共起スコアを取得する、請求項５記載の関係情報拡張装置。
【請求項７】
前記スコア取得部は、２以上の言語表現の組に含まれる当該２以上の言語表現と共起する言語表現を少なくとも素性として用い、当該素性の値及び２以上の言語表現の組に対する

意味的関係の有無を教師データとする機械学習を行い、前記拡張関係候補情報に含まれる２以上の言語表現を入力した場合の出力である確信度に応じた共起スコアを取得する、請求項５記載の関係情報拡張装置。
【請求項８】
前記経由スコアは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となると共に、当該拡張関係候補情報の生成時の置換における置換前の言語表現と置換後の言語表現とが類似しているほど、より高い値となるスコアである、請求項３または請求項４記載の関係情報拡張装置。
【請求項９】
関係情報及び拡張関係候補情報は、第１の言語表現と第２の言語表現とである２個の言語表現を有するものであり、
前記スコア取得部は、ある拡張関係候補情報について、当該拡張関係候補情報と第２の言語表現が一致する各関係情報の第１の言語表現と、当該拡張関係候補情報の第１の言語表現との類似度の和である第１の計算値と、当該拡張関係候補情報と第１の言語表現が一致する各関係情報の第２の言語表現と、当該拡張関係候補情報の第２の言語表現との類似度の和である第２の計算値と、前記関係情報記憶部で記憶されている各関係情報と当該拡張関係候補情報との第１の言語表現同士の類似度及び第２の言語表現同士の類似度の積の和である第３の計算値とのうち、任意の１以上の計算値を引数とする増加関数の値である経由スコアを取得する、請求項８記載の関係情報拡張装置。
【請求項１０】
前記関係情報は、当該関係情報が有する２以上の言語表現の意味的関係の種類を識別する情報である種類識別情報をも有するものであり、
前記類似情報記憶部では、種類識別情報と、当該種類識別情報に対応する類似情報とが記憶されており、
前記拡張関係候補情報生成部は、前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも１個の言語表現を置換する際に、当該関係情報が有する種類識別情報に対応する類似情報を用いて置換を行う、請求項１から請求項９のいずれか記載の関係情報拡張装置。
【請求項１１】
意味的関係を有する２以上の言語表現を有する関係情報が記憶される関係情報記憶部と、類似する２以上の言語表現を有する類似情報が２以上記憶される類似情報記憶部と、少なくとも１個の言語表現が類似する言語表現に置換された関係情報である拡張関係候補情報が記憶される拡張関係候補情報記憶部と、拡張関係候補情報生成部と、スコア取得部と、選択部とを用いて処理される関係情報拡張方法であって、
前記拡張関係候補情報生成部が、前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも１個の言語表現を、前記類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張関係候補情報を生成し、当該拡張関係候補情報を前記拡張関係候補情報記憶部に蓄積する拡張関係候補情報生成ステップと、
前記スコア取得部が、前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的関係を有する確からしさを示すスコアを取得するスコア取得ステップと、
前記選択部が、前記スコア取得ステップで取得したスコアを用いて、前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報のうち、当該スコアの高い拡張関係候補情報である拡張関係情報を選択する選択ステップと、を備えた関係情報拡張方法。
【請求項１２】
コンピュータを、
意味的関係を有する２以上の言語表現を有する関係情報が記憶される関係情報記憶部で記憶されている関係情報に含まれる少なくとも１個の言語表現を、類似する２以上の言語表現を有する類似情報が２以上記憶される類似情報記憶部で記憶されている類似情報を用いて、当該言語表現に類似する言語表現に置換した拡張関係候補情報を生成し、当該拡張関係候補情報を、拡張関係候補情報が記憶される拡張関係候補情報記憶部に蓄積する拡張関

係候補情報生成部、
前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的関係を有する確からしさを示すスコアを取得するスコア取得部、
前記スコア取得部が取得したスコアを用いて、前記拡張関係候補情報記憶部で記憶されている拡張関係候補情報のうち、当該スコアの高い拡張関係候補情報である拡張関係情報を選択する選択部として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、意味的関係を有する２以上の言語表現を有する関係情報を拡張する関係情報拡張装置等に関する。
【背景技術】
【０００２】
従来、意味的関係を有する２以上の言語表現を有する関係情報を取得することが行われていた。ここで、意味的関係とは、例えば、原因や、上位下位等の関係である。したがって、関係情報は、例えば、死亡の原因が心筋梗塞であることを示す「原因＜心筋梗塞、死亡＞」や、頭痛薬の上位語が薬であることを示す「上位＜薬、頭痛薬＞」等である。その関係情報を取得する方法として、例えば、非特許文献１で開示されている方法を用いることができる。その非特許文献１では、例えば、「ＸがＹの原因である」という表現のパターンを手がかりとして、文書からＸ，Ｙの部分を取得することにより、関係情報を取得することができた。しかしながら、そのようなパターンを用いた方法では、その手がかりとなるパターンで書かれた関係しか取得できないという問題があった。
【０００３】
その問題を解決するための方法として、類推を用いて関係情報を取得する方法が、例えば、非特許文献２で開示されている。
【先行技術文献】
【非特許文献】
【０００４】
【非特許文献１】Ｐ．Ｐａｎｔｅｌ，Ｍ．Ｐｅｎｎａｃｃｈｉｏｔｔｉ，「Ｅｓｐｒｅｓｓｏ：Ｌｅｖｅｒａｇｉｎｇ ｇｅｎｅｒｉｃ ｐａｔｔｅｒｎｓ ｆｏｒ ａｕｔｏｍａｔｉｃａｌｌｙ ｈａｒｖｅｓｔｉｎｇ ｓｅｍａｎｔｉｃ ｒｅｌａｔｉｏｎｓ」，Ｉｎ Ｐｒｏｃ．ｏｆ ｔｈｅ ２１ｓｔ ＣＯＬＩＮＧ ａｎｄ ４４ｔｈ ＡＣＬ（ＣＯＬＩＮＧ－ＡＣＬ－０６），ｐ．１１３－１２０，２００６年
【非特許文献２】石川大介、石塚英弘、藤原譲、「特許文献における因果関係を用いた類推による仮説の生成と検証－ライフサイエンス分野を対象として－」情報知識学会誌、Ｖｏｌ．１７、Ｎｏ．３、ｐ．１６４－１８１、２００７年
【発明の概要】
【発明が解決しようとする課題】
【０００５】
しかしながら、非特許文献２では、関係情報に含まれる言語表現を置換する際に、語基が共通する表現にしか置換できないという問題があった。例えば、「ペプチド」を「抗菌ペプチド」に置き換えることはできるが、それ語基を共通にしないものには置換できなかった。また、置換後の関係情報には、不適切なものも含まれてしまうという問題があった。
【０００６】
本発明は、上記問題を解決するためになされたものであり、既存の関係情報を拡張する場合に、関係情報に含まれる言語表現を、語基の共通しないものにも拡張することができ、また、適切な意味的関係を有する関係情報に拡張することができる関係情報拡張装置等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】

【０００７】
上記目的を達成するため、本発明による関係情報拡張装置は、意味的関係を有する２以上の言語表現を有する関係情報が記憶される関係情報記憶部と、類似する２以上の言語表現を有する類似情報が２以上記憶される類似情報記憶部と、少なくとも１個の言語表現が類似する言語表現に置換された関係情報である拡張関係候補情報が記憶される拡張関係候補情報記憶部と、関係情報記憶部で記憶されている関係情報に含まれる少なくとも１個の言語表現を、類似情報を用いて、言語表現に類似する言語表現に置換した拡張関係候補情報を生成し、拡張関係候補情報を拡張関係候補情報記憶部に蓄積する拡張関係候補情報生成部と、拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的関係を有する確からしさを示すスコアを取得するスコア取得部と、スコア取得部が取得したスコアを用いて、拡張関係候補情報記憶部で記憶されている拡張関係候補情報のうち、スコアの高い拡張関係候補情報である拡張関係情報を選択する選択部と、を備えたものである。
【０００８】
このような構成により、言語表現を類似する言語表現に置換するため、語基の共通しないものに言語表現を置換することができるようになる。また、拡張関係候補情報のスコアを取得し、そのスコアの高いものを選択するため、その選択結果である拡張関係情報は、意味的関係を適切に有するものとなりうる。
【０００９】
また、本発明による関係情報拡張装置では、スコア取得部は、拡張関係候補情報に含まれる２以上の言語表現の共起に関するスコアである共起スコアを取得してもよい。
意味的関係を有する言語表現は、共起する可能性が高いと考えられるため、このような構成により、共起スコアの高いものを選択することによって、適切な選択を行うことができうる。
【００１０】
また、本発明による関係情報拡張装置では、スコア取得部は、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアを取得してもよい。
【００１１】
適切な拡張関係情報は、多くの関係情報から得られるものであると考えられるため、このような構成により、経由スコアの高いものを選択することによって、適切な選択を行うことができうる。
【００１２】
また、本発明による関係情報拡張装置では、スコア取得部は、拡張関係候補情報に含まれる２以上の言語表現の共起に関するスコアである共起スコアと、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアとを取得し、選択部は、共起スコアが高く、かつ、経由スコアが高い拡張関係候補情報を選択してもよい。
このような構成により、共有スコアと、経由スコアとの高いものを選択することによって、より適切な選択を行うことができうる。
【００１３】
また、本発明による関係情報拡張装置では、前記スコア取得部は、前記拡張関係候補情報に含まれる２以上の言語表現と、当該拡張関係候補情報の生成時に用いられた関係情報の意味的関係と同じ種類の意味的関係を有する各関係情報に含まれる２以上の言語表現に対して共起の高い言語表現である共起言語表現とが共起する方が、前記拡張関係候補情報に含まれる２以上の言語表現のみが共起するよりも高い値となる共起スコアを取得してもよい。
このような構成により、その共起スコアによって、拡張関係候補情報の生成で用いられた関係情報と同様の関係を有するものほどより高い値となる共起スコアを取得できうる。
【００１４】
また、本発明による関係情報拡張装置では、前記関係情報は、当該関係情報が有する２

以上の言語表現の意味的関係の種類を識別する情報である種類識別情報をも有するものであり、前記拡張関係候補情報生成部は、拡張関係候補情報の生成に用いる関係情報が有する種類識別情報を有する拡張関係候補情報を生成し、種類識別情報と、当該種類識別情報に対応する、当該種類識別情報で識別される意味的関係の種類に対応する１以上の共起言語表現とを有する対応情報が１以上記憶される対応情報記憶部をさらに備え、前記スコア取得部は、前記拡張関係候補情報に含まれる２以上の言語表現と、当該拡張関係候補情報が有する種類識別情報に対応する各共起言語表現とが共起する方が、前記拡張関係候補情報に含まれる２以上の言語表現のみが共起するよりも高い値となる共起スコアを取得してもよい。

このような構成により、対応情報を用いて、前述のような共起スコアを取得することができる。

【００１５】

また、本発明による関係情報拡張装置では、前記スコア取得部は、２以上の言語表現の組に含まれる当該２以上の言語表現と共起する言語表現を少なくとも素性として用い、当該素性の値及び２以上の言語表現の組に対する意味的関係の有無を教師データとする機械学習を行い、前記拡張関係候補情報に含まれる２以上の言語表現を入力した場合の出力である確信度に応じた共起スコアを取得してもよい。

このような構成により、機械学習を用いて、前述のような共起スコアを取得することができる。

【００１６】

また、本発明による関係情報拡張装置では、経由スコアは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となると共に、拡張関係候補情報の生成時の置換における置換前の言語表現と置換後の言語表現とが類似しているほど、より高い値となるスコアであってもよい。

【００１７】

適切な拡張関係情報は、多くの関係情報から高い類似度で得られるものであると考えられるため、このような構成により、類似度も考慮して経由スコアを取得することができ、より適切な選択を行うことができうる。

【００１８】

また、本発明による関係情報拡張装置では、関係情報及び拡張関係候補情報は、第１の言語表現と第２の言語表現とである２個の言語表現を有するものであり、スコア取得部は、ある拡張関係候補情報について、拡張関係候補情報と第２の言語表現が一致する各関係情報の第１の言語表現と、拡張関係候補情報の第１の言語表現との類似度の和である第１の計算値と、拡張関係候補情報と第１の言語表現が一致する各関係情報の第２の言語表現と、拡張関係候補情報の第２の言語表現との類似度の和である第２の計算値と、関係情報記憶部で記憶されている各関係情報と拡張関係候補情報との第１の言語表現同士の類似度及び第２の言語表現同士の類似度の積の和である第３の計算値とのうち、任意の１以上の計算値を引数とする増加関数の値である経由スコアを取得してもよい。

このような構成により、関係情報が２個の言語表現を含む場合に、第１から第３の計算値の任意の１以上の計算値を用いて、経由スコアを取得することができる。

【００１９】

また、本発明による関係情報拡張装置では、前記関係情報は、当該関係情報が有する２以上の言語表現の意味的関係の種類を識別する情報である種類識別情報をも有するものであり、前記類似情報記憶部では、種類識別情報と、当該種類識別情報に対応する類似情報とが記憶されており、前記拡張関係候補情報生成部は、前記関係情報記憶部で記憶されている関係情報に含まれる少なくとも１個の言語表現を置換する際に、当該関係情報が有する種類識別情報に対応する類似情報を用いて置換を行ってもよい。

このような構成により、関係情報にふさわしい類似情報を用いて、拡張関係候補情報を生成することができうる。したがって、拡張関係候補情報の精度がより高いものとなり、その結果、拡張関係情報の精度もより高くなりうる。

【００２０】
また、本発明による関係情報拡張装置では、前記関係情報は、当該関係情報が有する２以上の言語表現の意味的関係の種類を識別する情報である種類識別情報をも有するものであり、前記類似情報記憶部では、種類識別情報と、置換対象でない１以上の言語表現と、当該種類識別情報及び当該置換対象でない１以上の言語表現に対応する類似情報とが記憶されており、前記拡張関係候補情報生成部は、前記関係情報記憶部で記憶されている関係情報に含まれる１個の言語表現を置換する際に、当該関係情報が有する種類識別情報及び置換対象でない言語表現に対応する類似情報を用いて置換を行ってもよい。
このような構成により、関係情報や、置換対象でない言語表現にふさわしい類似情報を用いて、拡張関係候補情報を生成することができうる。したがって、拡張関係候補情報の精度がより高いものとなり、その結果、拡張関係情報の精度もより高くなりうる。
【発明の効果】
【００２１】
本発明による関係情報拡張装置等によれば、既存の関係情報を用いて、新たな適切な意味的関係を取得することができる。
【図面の簡単な説明】
【００２２】
【図１】本発明の実施の形態１による関係情報拡張装置の構成を示すブロック図
【図２】同実施の形態による関係情報拡張装置の動作を示すフローチャート
【図３】同実施の形態による関係情報拡張装置の動作を示すフローチャート
【図４】同実施の形態による関係情報拡張装置の動作を示すフローチャート
【図５】同実施の形態おける類似情報の一例を示す図
【図６】同実施の形態における拡張関係候補情報等の一例を示す図
【図７】同実施の形態における拡張関係候補情報等の一例を示す図
【図８】同実施の形態における実験結果を示す図
【図９】同実施の形態による関係情報拡張装置の構成の他の一例を示すブロック図
【図１０】同実施の形態における対応情報の一例を示す図
【図１１】同実施の形態における類似情報記憶部で記憶されている情報の一例を示す図
【図１２】同実施の形態における類似情報記憶部で記憶されている情報の一例を示す図
【図１３】同実施の形態におけるサポートベクトルマシンについて説明するための図
【図１４】同実施の形態におけるコンピュータシステムの外観一例を示す模式図
【図１５】同実施の形態におけるコンピュータシステムの構成の一例を示す図
【発明を実施するための形態】
【００２３】
以下、本発明による関係情報拡張装置について、実施の形態を用いて説明する。なお、以下の実施の形態において、同じ符号を付した構成要素及びステップは同一または相当するものであり、再度の説明を省略することがある。
【００２４】
（実施の形態１）
本発明の実施の形態１による関係情報拡張装置について、図面を参照しながら説明する。本実施の形態による関係情報拡張装置は、既存の関係情報に含まれる少なくとも１個の言語表現を類似する言語表現に置換し、その置換後のものについてスコアを取得し、そのスコアの高いものを選択することによって、関係情報の拡張を行うものである。
【００２５】
図１は、本実施の形態による関係情報拡張装置１の構成を示すブロック図である。本実施の形態による関係情報拡張装置１は、関係情報記憶部１１と、類似情報記憶部１２と、拡張関係候補情報生成部１３と、拡張関係候補情報記憶部１４と、スコア取得部１５と、選択部１６と、拡張関係情報記憶部１７とを備える。
【００２６】
関係情報記憶部１１では、意味的関係を有する２以上の言語表現を有する関係情報が記

憶される。関係情報記憶部１１で記憶されている関係情報の個数は問わないが、後述するスコア取得部１５によって、経由スコアを取得する場合には、複数の関係情報が記憶されていることが好適である。一方、共起スコアのみの取得しか行わない場合には、関係情報記憶部１１において、一個の関係情報が記憶されていてもよく、あるいは、複数の関係情報が記憶されていてもよい。なお、言語表現は、通常、単語（例えば、「健康」等）であるが、連続した単語の並び（例えば、「早朝散歩」等の単語列）であってもよい。また、その単語は、例えば、自立語の単語であってもよく、特に名詞の単語であってもよい。また、単語が名詞である場合には、その名詞を助詞や前置詞でつないだものが言語表現であってもよい（例えば、「私の友人」や「ｆｒｉｅｎｄ　ｏｆ　ｍｉｎｅ」等）。また、関係情報は、通常、２個の言語表現を有するものであるが、３個以上の言語表現を有してもよい。なお、本実施の形態では、関係情報が主に２個の言語表現を有する場合について説明する。また、本実施の形態では、関係情報を＜第１の言語表現、第２の言語表現＞のように表記する。

【００２７】

また、意味的関係は、何らかの意味的な関係であれば、その種類を問わない。例えば、上位と下位の関係（例えば、＜飲み物、コーヒー＞等）であってもよく、原因結果の関係（例えば、＜脳梗塞、死亡＞等）であってもよく、ライバルや対義語の関係（例えば、＜上、下＞や＜高速、低速＞等）であってもよく、製品とメーカーの関係（例えば、＜掃除機、Ａ社＞等）であってもよく、事象と方法の関係（例えば、＜爆発、爆弾＞等）であってもよく、事象とツールの関係（例えば、＜授業、教科書＞等）であってもよく、事象と防ぐものの関係（例えば、＜病気、薬＞）であってもよく、物と材料の関係（例えば、＜缶、アルミニウム＞等）であってもよく、全体と部分の関係（例えば、＜自転車、サドル＞等）であってもよく、事象や物とトラブルの関係（例えば、＜自動車、パンク＞等）であってもよく、事象と対策との関係（例えば、＜雨、傘＞等）であってもよく、事象と必須のものとの関係（例えば、＜記念撮影、写真機＞等）であってもよく、近距離の地点の関係（例えば、＜大阪、神戸＞等）であってもよく、地点と名物や名所の関係（例えば、＜東京、東京タワー＞等）であってもよく、料理と含まれる食材との関係（例えば、＜シチュー、じゃがいも＞等）であってもよく、その他の種類の関係であってもよい。また、意味的関係は、ある言語表現と、他の言語表現とが、例えば、関連あり、ゆかりの人物、ゆかりの寺・神社、食材効能、効く食材、意外な食材、料理効能、効く料理、意外な料理、栄養効能、効く栄養素、成分、旬、旬の食材、代用食品、類似語、仏像・神様、所蔵する寺、祀る神社等の関係を有することであってもよい。なお、意味的関係は、３個以上の言語表現に関する関係であってもよい。例えば、食生活と検査結果と病気の関係（例えば、＜高カロリー、高血糖、糖尿病＞、＜塩分過多、高血圧、脳卒中＞等）であってもよい。

【００２８】

また、関係情報には、その関係情報に応じた意味的関係を識別する情報が含まれてもよく、あるいは、含まれなくてもよい。前者の場合には、例えば、関係情報は、「原因＜脳梗塞、死亡＞」であってもよい。この場合には、「原因」が意味的関係を識別する情報であり、死亡の原因が脳梗塞であることを示している。

【００２９】

また、関係情報記憶部１１で記憶される関係情報は、手作業で作成されたものであってもよく、自動的に取得されたものであってもよい。後者の場合には、例えば、前述の非特許文献１の手法を用いてもよい。

【００３０】

なお、関係情報に含まれる言語表現は、言語表現そのものであってもよく、あるいは、その情報を特定可能な情報であってもよい。後者の場合には、関係情報に含まれる言語表現は、例えば、言語表現が格納されている領域を示すポインタやアドレスであってもよい。このことは、他の情報についても同様であるとする。

【００３１】

類似情報記憶部１２では、２以上の類似情報が記憶される。類似情報は、類似する２以上の言語表現を有する情報である。類似情報は、２個の言語表現を有するものであってもよく、３個以上の言語表現を有するものであってもよい。なお、同一の類似情報に含まれる言語表現は、互いに類似するものであるとする。また、類似情報は、言語表現の類似の程度を示す類似度を含んでいてもよく、あるいは、含んでいなくてもよい。また、類似情報は、手作業で作成されたものであってもよく、自動的に取得されたものであってもよい。

【００３２】

拡張関係候補情報生成部１３は、関係情報記憶部１１で記憶されている関係情報に含まれる少なくとも１個の言語表現を、類似情報を用いて、その言語表現に類似する言語表現に置換した拡張関係候補情報を生成する。そして、拡張関係候補情報生成部１３は、その生成した拡張関係候補情報を拡張関係候補情報記憶部１４に蓄積する。なお、拡張関係候補情報生成部１３は、関係情報に含まれる少なくとも１個の言語表現を類似するものに置換するものである。したがって、拡張関係候補情報生成部１３は、関係情報に含まれる１個の言語表現を類似するものに置換してもよく、あるいは、関係情報に含まれる２個以上の言語表現をそれぞれ類似するものに置換してもよい。また、拡張関係候補情報生成部１３は、関係情報に含まれるすべての言語表現を類似するものに置換してもよい。また、関係情報に含まれる言語表現に類似する言語表現が２以上存在する場合には、拡張関係候補情報生成部１３は、関係情報に含まれる言語表現を、その２以上の各言語表現に置換した拡張関係候補情報をそれぞれ生成してもよい。

【００３３】

なお、拡張関係候補情報生成部１３は、異なる関係情報から、同じ拡張関係候補情報を生成することがありうる。例えば、拡張関係候補情報生成部１３が、異なる関係情報＜脳梗塞、死亡＞、＜心筋梗塞、死亡＞から、同じ拡張関係候補情報＜脳卒中、死亡＞を生成する場合などである。その場合には、例えば、生成後の拡張関係候補情報に対するユニーク処理を行い、拡張関係候補情報の重複を解消してもよく、あるいは、拡張関係候補情報の蓄積時に、すでに蓄積されている拡張関係候補情報を蓄積しないようにしてもよい。

【００３４】

また、拡張関係候補情報生成部１３は、関係情報と同じ情報である拡張関係候補情報を拡張関係候補情報記憶部１４に蓄積してもよく、あるいは、蓄積しなくてもよい。前者の場合には、拡張関係候補情報生成部１３は、さらに積極的に、関係情報記憶部１１で記憶されている各関係情報をそのまま拡張関係候補情報記憶部１４に蓄積するようにしてもよい。

【００３５】

拡張関係候補情報記憶部１４では、拡張関係候補情報が記憶される。拡張関係候補情報は、前述のように、少なくとも１個の言語表現が類似する言語表現に置換された関係情報である。拡張関係候補情報記憶部１４で記憶されている拡張関係候補情報は、前述のように、拡張関係候補情報生成部１３によって生成されたものである。

【００３６】

スコア取得部１５は、拡張関係候補情報記憶部１４で記憶されている拡張関係候補情報が意味的関係を有する確からしさを示すスコアを取得する。このスコアは、拡張関係候補情報に含まれる２以上の言語表現の共起に関するスコアであってもよい。すなわち、スコア取得部１５は、拡張関係候補情報に含まれる２以上の言語表現の共起に関するスコアである共起スコアを取得してもよい。また、このスコアは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアであってもよい。すなわち、スコア取得部１５は、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアを取得してもよい。その経由スコアは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となると共に、その拡張関係候補情報の生成時の置換における置換前の言語表現と置換後の言語表現とが類似しているほど、より高い値となるスコアであってもよい。なお、「置換前

の言語表現」とは、置換の対象となる言語表現のことである。また、スコア取得部１５は、共起スコアと、経由スコアとの一方のみを取得してもよく、あるいは、両方を取得してもよい。

【００３７】

ここで、共起スコアの取得方法と、経由スコアの取得方法とについて説明する。

（１）共起スコアの取得方法

まず、共起について説明する。言語表現Ａと、言語表現Ｂとが共起するとは、決められた長さの範囲内（例えば、一文の範囲内、一段落の範囲内、一ページの範囲内、一の文書の範囲内、一のウェブページの範囲内等）において、言語表現Ａと言語表現Ｂとが同時に出現することである。共起スコアは、その共起を示す尺度のことである。共起を示す尺度としては、例えば、共起頻度や共起率、Ｓｉｍｐｓｏｎ係数、コサイン距離、ダイス係数、相互情報量等が存在する。言語表現Ａと言語表現Ｂとの共起頻度とは、決められた長さの範囲内において、言語表現Ａ，Ｂが同時に出現する数である。共起率とは、共起頻度を、言語表現Ａの出現数（Ｘとする）と言語表現Ｂの出現数（Ｙとする）との和から共起頻度（Ｚとする）を引いたもの（すなわち、Ｘ＋Ｙ－Ｚ）で割った数である。また、Ｓｉｍｐｓｏｎ係数は、共起率の分母を、Ｘ，Ｙの最小値にしたものである。また、コサイン距離は、共起率の分母を、ＸとＹの積の絶対値の自乗根にしたものである。なお、これらの共起を示す尺度についてはすでに公知である。例えば、次の文献を参照されたい。また、上記以外の共起の尺度を用いて共起スコアを取得してもよいことは言うまでもない。

【００３８】

文献：相澤彰子、「共起に基づく類似性尺度」、オペレーションズ・リサーチ、経営の科学５２（１１）、ｐ．７０６－７１２、２００７年１１月

【００３９】

なお、その共起スコアを算出する際には、多くの文書を有するデータベースが必要になる。そのデータベースは、例えば、関係情報拡張装置１が保持していてもよく、あるいは、装置外に存在してもよい。後者の場合には、例えば、そのデータベースは、ウェブサイトであってもよく、あるいは、所定のサーバが有するデータベースであってもよい。また、その共起スコアを算出する際の検索等の処理は、スコア取得部１５がそのデータベースにアクセスすることによって行ってもよく、あるいは、スコア取得部１５は、外部の装置やサーバに対して、２以上の言語表現を渡し、その外部の装置やサーバにおいて生成された共起スコアを受け取るだけであってもよい。

【００４０】

また、あらゆる言語表現の組合せについてあらかじめ共起スコアを算出して保持しておくことによって、ある言語表現Ａ、Ｂの共起スコアを、その保持している情報から取得するようにしてもよい。具体的には、言語表現Ａと言語表現Ｂと、両言語表現の共起スコアとが対応付けられて保持されており、スコア取得部１５は、言語表現Ａ，Ｂの共起ストアの取得方法を取得する際には、その言語表現Ａ，Ｂに対応付けられている共起ストアを読み出してもよい。このように、共起スコアの取得は、共起スコアを算出することであってもよく、共起スコアを読み出すことであってもよい。

【００４１】

また、拡張関係候補情報が３個以上の言語表現を有する場合には、スコア取得部１５は、その３個以上の言語表現の共起スコアを取得する。すなわち、スコア取得部１５は、拡張関係候補情報に含まれるすべての言語表現に対する共起スコアを取得する。その共起スコアが、共起頻度や共起率等のように、３個以上の言語表現についても取得できるものであれば、スコア取得部１５は、その３個以上の言語表現に対応する共起スコアを取得する。一方、共起スコアがダイス係数や相互情報量などのように、２個の言語表現に対してのみ定義されている場合には、スコア取得部１５は、３個以上の言語表現から２個の言語表現のすべてのペアを作り、そのすべてのペアについてダイス係数等の共起の尺度を取得し、そのすべてのペアの共起の尺度を引数とする関数の値を共起スコアとしてもよい。なお、その関数は、各引数の増加関数であるとする。例えば、その関数は、すべてのペアの共

起の尺度の平均や和、積等であってもよい。

【００４２】

（２）経由スコアの取得方法

次に、経由スコアについて説明する。ここでは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となると共に、その拡張関係候補情報の生成時の置換における置換前の言語表現と置換後の言語表現とが類似しているほど、より高い値となるスコアである経由スコアについて説明する。また、関係情報と、拡張関係候補情報において、２個の言語表現が含まれている場合について説明する。その関係情報と、拡張関係候補情報とにおいて、第１項の言語表現のことを第１の言語表現と呼び、第２項の言語表現のことを第２の言語表現と呼ぶことにする。すなわち、関係情報や、拡張関係候補情報が＜Ｘ，Ｙ＞である場合には、第１の言語表現がＸとなり、第２の言語表現がＹとなる。

【００４３】

ある拡張関係候補情報を＜f_h、s_h＞とする。ここで、f_h、s_hは、言語表現である。意味的関係を有する２個の言語表現を有する与えられた関係情報の集合を、R_{given}とする。$R_{given}=\{r_1=<f_1, s_1>, \cdots, r_n=<f_n, s_n>\}$とする。そして、第１の計算値$S_{FA}$（$f_h$、$s_h$）と、第２の計算値$S_{SA}$（$f_h$、$s_h$）と、第３の計算値$S_{FULL}$（$f_h$、$s_h$）とを、次式のようにして算出する。

【００４４】

なお、第１の計算値S_{FA}（f_h、s_h）は、拡張関係候補情報＜f_h、s_h＞と第２の言語表現s_hが一致する各関係情報の第１の言語表現と、拡張関係候補情報＜f_h、s_h＞の第１の言語表現f_hとの類似度の和である。その和は、次式で示されるように、拡張関係候補情報＜f_h、s_h＞と第２の言語表現s_hが一致する各関係情報の第１の言語表現の集合に関する和である。したがって、その第１の計算値S_{FA}（f_h、s_h）は、拡張関係候補情報＜f_h、s_h＞と第２の言語表現s_hが一致する関係情報に含まれる第１の言語表現の集合を特定し、その集合に含まれる各第１の言語表現と、拡張関係候補情報＜f_h、s_h＞の第１の言語表現f_hとの類似度の、集合の各要素に関する和であるということもできる。

【００４５】

また、第２の計算値S_{SA}（f_h、s_h）は、拡張関係候補情報＜f_h、s_h＞と第１の言語表現f_hが一致する各関係情報の第２の言語表現と、拡張関係候補情報＜f_h、s_h＞の第２の言語表現s_hとの類似度の和である。その和は、次式で示されるように、拡張関係候補情報＜f_h、s_h＞と第１の言語表現f_hが一致する各関係情報の第２の言語表現の集合に関する和である。したがって、その第２の計算値S_{SA}（f_h、s_h）は、拡張関係候補情報＜f_h、s_h＞と第１の言語表現f_hが一致する関係情報に含まれる第２の言語表現の集合を特定し、その集合に含まれる各第２の言語表現と、拡張関係候補情報＜f_h、s_h＞の第２の言語表現s_hとの類似度の、集合の各要素に関する和であるということもできる。

【００４６】

また、第３の計算値S_{FULL}（f_h、s_h）は、関係情報記憶部１１で記憶されている各関係情報と拡張関係候補情報＜f_h、s_h＞との第１の言語表現同士の類似度及び第２の言語表現同士の類似度の積の和である。その和は、次式で示されるように、R_{given}における第１の言語表現と第２の言語表現とのペアの集合に関する和である。したがって、その第３の計算値S_{FULL}（f_h、s_h）は、R_{given}に含まれる各関係情報の言語表現のペアの集合に含まれる一の関係情報について、第１の言語表現と拡張関係候補情報＜f_h、s_h＞の第１の言語表現f_hとの類似度と、第２の言語表現と拡張関係候補情報＜f_h、s_h＞の第２の言語表現s_hとの類似度との積を算出し、その類似度の積の、集合の各要素に関する和であるということもできる。

【数１】

$$S_{FA}(f_h,s_h)=\sum_{f_i\in FA(s_h)} sim(f_h,f_i)$$

$$S_{SA}(f_h,s_h)=\sum_{s_i\in SA(f_h)} sim(s_h,s_i)$$

$$S_{FULL}(f_h,s_h)=\sum_{<f_i,s_i>\in R_{given}} sim(f_h,f_i)sim(s_h,s_i)$$

【００４７】

ここで、ＦＡ（ｓ）は、Ｒｇｉｖｅｎで第２の言語表現がｓである関係情報の第１の言語表現の集合である。また、ＳＡ（ｆ）は、Ｒｇｉｖｅｎで第１の言語表現がｆである関係情報の第２の言語表現の集合である。ｓｉｍは、類似度である。この類似度は、例えば、次の文献の類似度のように自動的に算出されるものを用いてもよく、手作業で設定された類似度を用いてもよく、類義語、同義語など、意味が類似している表現として登録されている２つの言語表現を高い類似度としてもよい。また、言語表現を、意味的な階層構造、木構造で整理している辞書を用いる場合には、２つの言語表現から上の構造を辿って行き、最初の共通の場所までに辿る階層の数が少ないほど類似度が高く、逆に大きいほど類似度が低くなるように設定してもよい。そのような辞書としては、日本語のものとしては、例えば、「分類語彙表」（国立国語研究所）、「日本語語彙大系」（岩波書店）、「角川類語国語辞典」（角川書店）、「日本語大シソーラス」（大修館書店）、「ＥＤＲ概念体系辞書」（ＥＤＲプロジェクト）、「デジタル類語辞典」（ジャングル）、「ＪＳＴ科学技術用語シソーラス」（ＪＳＴ科学技術振興機構）等が存在する。また、英語のものとしては、例えば、「ロジェ類語辞典」、「ＷｏｒｄＮｅｔ」、「ＭｅＳＨ（Ｍｅｄｉｃａｌ　Ｓｕｂｊｅｃｔ　Ｈｅａｄｉｎｇｓ）」等が存在する。

【００４８】

文献：風間淳一、Ｓｔｉｊｎ　Ｄｅ　Ｓａｅｇｅｒ、鳥澤健太郎、村田真樹、「係り受けの確率的クラスタリングを用いた大規模類似語リストの作成」、言語処理学会第１５回年次大会、ｐ．８４－８７、２００９年３月

【００４９】

なお、上記各計算値を取得する際に、類似情報記憶部１２で記憶されている、類似度を含む類似情報を用いる場合には、言語表現Ａ，Ｂが類似するものではなく、類似情報記憶部１２で記憶されている類似情報に含まれないため、ｓｉｍ（Ａ，Ｂ）が分からない場合もありうる。その場合には、ｓｉｍ（Ａ，Ｂ）＝０にするものとする。

【００５０】

スコア取得部１５は、拡張関係候補情報＜ｆｈ、ｓｈ＞の経由スコアを、前述の第１の計算値ＳＦＡ（ｆｈ、ｓｈ）と、第２の計算値ＳＳＡ（ｆｈ、ｓｈ）と、第３の計算値ＳＦＵＬＬ（ｆｈ、ｓｈ）とのうち、任意の１以上の計算値を引数とする増加関数の値を算出することによって取得してもよい。なお、その増加関数は、各引数の増加関数であり、例えば、第１から第３の計算値の和であるＳｓｕｍ（ｆｈ、ｓｈ）であってもよく、第１から第３の計算値の積であるＳｐｒｏｄ（ｆｈ、ｓｈ）であってもよい。Ｓｓｕｍは、第１から第３の計算値のいずれかが高い場合に高くなる。一方、Ｓｐｒｏｄは、第１から第３の計算値のすべてが高い場合に高くなる。つまり、Ｓｐｒｏｄは、バランスよく両方の言語表現に基づき生成される拡張関係候補情報がよい値となる、と考えられる点で、Ｓｓｕｍと異なる。なお、Ｓｐｒｏｄの計算では、０になることを回避するため、各計算値に十分に小さい値を足すものとする。

【００５１】

なお、上記の式において、ｓｉｍ（Ａ，Ｂ）を、言語表現Ａ，Ｂが類似する場合に「１」となり、言語表現Ａ，Ｂが類似しない場合に「０」となる関数であるとすると、経由スコアは、拡張関係候補情報がより多くの関係情報から得られるものであるほど、より高い値となるスコアである経由スコアとなる。すなわち、類似の程度を考慮しないスコアとな

る。したがって、そのようにして前述の計算値を算出するようにしてもよい。例えば、スコア取得部１５は、ｓｉｍ（Ａ，Ｂ）を取得する際に、類似情報記憶部１２で記憶されている類似情報によって、言語表現Ａ，Ｂが類似することが示される場合には、ｓｉｍ（Ａ，Ｂ）＝１として、言語表現Ａ，Ｂが類似しないことが示される場合には、ｓｉｍ（Ａ，Ｂ）＝０としてもよい。

【００５２】

また、関係情報や拡張関係候補情報がｍ個の言語表現を含む場合の経由スコアの算出について簡単に説明する。なお、ｍは２以上の整数である。ここで、ある拡張関係候補情報を＜f_h^1，f_h^2，…，f_h^m＞とする。f_h^1，f_h^2等は、拡張関係候補情報に含まれる言語表現である。また、関係情報の集合R_{given}は次のようであるとする。なお、f_h^jや、f_i^jを第ｊの言語表現や、ｊ番目の言語表現と呼ぶことにする。

R_{given}＝｛r_1＝＜f_1^1，f_1^2，…，f_1^m＞、r_2＝＜f_2^1，f_2^2，…，f_2^m＞、…、r_n＝＜f_n^1，f_n^2，…，f_n^m＞｝

【００５３】

その場合に、計算値$S_{k1k2\cdots kj}(f_h^1,\cdots,f_h^m)$は、次のようになる。

【数２】

$$S_{k1k2\cdots kj}(f_h^1,\cdots,f_h^m) = \sum_{(f_i^{k1},f_i^{k2},\cdots,f_i^{kj})\in M} sim(f_h^{k1},f_i^{k1})\,sim(f_h^{k2},f_i^{k2})\cdots sim(f_h^{kj},f_i^{kj})$$

【００５４】

ここで、計算値$S_{k1k2\cdots kj}(f_h^1,\cdots,f_h^m)$は、拡張関係候補情報が、ｋ１番目の言語表現と、ｋ２番目の言語表現と、…、ｋｊ番目の言語表現とを置換して生成されたものである場合における経由スコアの計算値である。なお、上記式における集合Ｍについて説明する。まず、集合Ｍ１＝｛１，２，３，…，ｍ｝として、集合Ｍ２＝｛ｋ１，ｋ２，ｋ３，…，ｋｊ｝とする。その集合Ｍ２の各要素ｋ１，ｋ２，ｋ３，…，ｋｊは、集合Ｍ１から重複しないように選択した１個以上、ｍ個以下の要素である。また、ｋ１＜ｋ２＜…＜ｋｊであるとする。また、集合Ｍ３を、集合Ｍ１から、集合Ｍ２に含まれる各要素を除去したものであるとする。その集合Ｍ３を、Ｍ３＝｛ｐ１，ｐ２，…，ｐｍ－ｊ｝とする。すると、Ｍは、R_{given}において、第ｐ１項がf_h^{p1}であり、第ｐ２項がf_h^{p2}であり、…、第ｐｍ－ｊ項がf_h^{pm-j}である、第ｋ１項、第ｋ２項、…、第ｋｊ項の組の集合である。ただし、ｊ＝ｍの場合には、Ｍ３は空集合となり、ＭはR_{given}となる。そして、スコア取得部１５は、ｊ＝１からｊ＝ｍまでの各値におけるすべてのＭ２のうち、任意の１以上のＭ２に対する計算値を引数とする増加関数（この増加関数は、各引数の増加関数であるとする）の値を算出することによって、ｍ個の言語表現を含む拡張関係候補情報に対する経由スコアを取得することができる。その増加関数は、例えば、すべてのＭ２について算出した計算値の和であってもよく、すべてのＭ２について算出した計算値の積であってもよい。

【００５５】

また、ここでは、共起スコアと、経由スコアとの２個のスコアについて説明したが、スコア取得部１５は、それら以外の拡張関係候補情報が意味的関係を有する確からしさを示すスコアを取得してもよい。

【００５６】

また、スコア取得部１５は、共起スコアと、経由スコアとの両方を取得する場合に、拡張関係候補情報記憶部１４で記憶されているすべての拡張関係候補情報について、共起スコアと、経由スコアとを取得してもよく、あるいは、一方のスコアを取得し、その取得したスコアの高いものについてのみ、他方のスコアを取得するようにしてもよい。本実施の形態では、後者の場合について主に説明する。

【００５７】

また、スコア取得部１５が取得したスコアは、拡張関係候補情報に対応付けられて蓄積されてもよい。具体的には、スコア取得部１５は、取得したスコアを、拡張関係候補情報記憶部１４で記憶されている、そのスコアに対応する拡張関係候補情報に対応付けて蓄積してもよく、あるいは、取得したスコアを、そのスコアに対応する拡張関係候補情報を識別する情報に対応付けて図示しない記録媒体に蓄積してもよい。

【００５８】

選択部１６は、スコア取得部１５が取得したスコアを用いて、拡張関係候補情報記憶部１４で記憶されている拡張関係候補情報のうち、スコアの高い拡張関係候補情報を選択する。この選択部１６が選択した拡張関係候補情報が拡張関係情報となる。スコア取得部１５が共起スコアと、経由スコアとの両方を取得した場合には、選択部１６は、共起スコアが高く、かつ、経由スコアが高い拡張関係候補情報を選択してもよい。なお、スコアが高い拡張関係候補情報とは、例えば、しきい値以上のスコアである拡張関係候補情報であってもよく、あるいは、スコアの高いものからあらかじめ決められた個数の拡張関係候補情報であってもよい。なお、その２個の条件を合わせて用いてもよい。例えば、スコアの高い拡張関係候補情報は、しきい値以上のスコアである拡張関係候補情報であって、かつ、スコアの高いものからあらかじめ決められた個数内の拡張関係候補情報であってもよい。そのしきい値は、例えば、スコアの最大値にあらかじめ決められた１未満の数（例えば、０．９や０．８など）を掛けた値であってもよく、あらかじめ決められた値であってもよい。

【００５９】

選択部１６は、選択した拡張関係候補情報を後述する拡張関係情報記憶部１７に蓄積してもよく、あるいは、拡張関係候補情報記憶部１４において、選択した拡張関係候補情報に対して、拡張関係情報であることを示すフラグ等の識別情報を設定してもよい。すなわち、拡張関係候補情報を、選択されたものと、選択されていないものとに区別できるのであれば、その選択結果を示す方法は問わない。なお、本実施の形態では、前者の場合、すなわち、選択結果である拡張関係情報が拡張関係情報記憶部１７に蓄積される場合について説明する。

【００６０】

また、拡張関係候補情報記憶部１４に関係情報と同じ情報である拡張関係候補情報が記憶されている場合に、選択部１６は、関係情報と同じ拡張関係情報を選択してもよく、あるいは、しなくてもよい。後者の場合には、選択部１６は、関係情報記憶部１１で記憶されている関係情報を参照し、その関係情報と一致する拡張関係候補情報を選択しないようにしてもよい。

【００６１】

拡張関係情報記憶部１７では、選択部１６によって選択された拡張関係候補情報である拡張関係情報が記憶される。この拡張関係情報は、前述のように、選択部１６によって蓄積されたものである。

【００６２】

なお、関係情報記憶部１１や、類似情報記憶部１２に関係情報や２以上の類似情報が記憶される過程は問わない。例えば、記録媒体を介して関係情報等が関係情報記憶部１１等で記憶されるようになってもよく、通信回線等を介して送信された関係情報等が関係情報記憶部１１等で記憶されるようになってもよく、あるいは、入力デバイスを介して入力された関係情報等が関係情報記憶部１１等で記憶されるようになってもよい。また、関係情報記憶部１１や類似情報記憶部１２、拡張関係候補情報記憶部１４、拡張関係情報記憶部１７での記憶は、ＲＡＭ等における一時的な記憶でもよく、あるいは、長期的な記憶でもよい。また、また、関係情報記憶部１１や類似情報記憶部１２、拡張関係候補情報記憶部１４、拡張関係情報記憶部１７は、所定の記録媒体（例えば、半導体メモリや磁気ディスク、光ディスクなど）によって実現されうる。

【００６３】

また、関係情報記憶部１１と、類似情報記憶部１２と、拡張関係候補情報記憶部１４と

、拡張関係情報記憶部１７とのうち、任意の２以上の記憶部は、同一の記録媒体によって実現されてもよく、あるいは、別々の記録媒体によって実現されてもよい。前者の場合には、例えば、関係情報を記憶している領域が関係情報記憶部１１となり、類似情報を記憶している領域が類似情報記憶部１２となる。

【００６４】

次に、本実施の形態による関係情報拡張装置１の動作について、図２のフローチャートを用いて説明する。

（ステップＳ１０１）拡張関係候補情報生成部１３は、関係情報記憶部１１で記憶されている関係情報の少なくとも１個の言語表現を、類似情報記憶部１２で記憶されている類似情報を用いて類似する言語表現に置換した拡張関係候補情報を生成し、拡張関係候補情報記憶部１４に蓄積する。なお、この処理の詳細については、図３のフローチャートを用いて説明する。

【００６５】

（ステップＳ１０２）スコア取得部１５は、拡張関係候補情報記憶部１４で記憶されている各拡張関係候補情報について、スコアを取得する。なお、この処理の詳細については、図４のフローチャートを用いて後述する。

【００６６】

（ステップＳ１０３）選択部１６は、スコア取得部１５が取得したスコアの高い拡張関係候補情報である拡張関係情報を選択する。この選択は、例えば、前述のように、しきい値よりも大きいスコアの拡張関係候補情報を選択することによって行われてもよく、スコアの高いものからあらかじめ決められた個数の拡張関係候補情報を選択することによって行われてもよい。そして、選択部１６は、その選択結果の拡張関係情報を、拡張関係情報記憶部１７に蓄積する。このようにして、関係情報を拡張する一連の処理が終了する。

【００６７】

図３は、図２のフローチャートにおける拡張関係候補情報の生成の処理（ステップＳ１０１）の詳細を示すフローチャートである。

（ステップＳ２０１）拡張関係候補情報生成部１３は、カウンタｉを１に設定する。

【００６８】

（ステップＳ２０２）拡張関係候補情報生成部１３は、カウンタｊを１に設定する。

【００６９】

（ステップＳ２０３）拡張関係候補情報生成部１３は、関係情報記憶部１１で記憶されているｉ番目の関係情報に含まれるｊ番目の言語表現に類似する言語表現が、類似情報記憶部１２で記憶されているかどうか判断する。そして、ｉ番目の関係情報に含まれるｊ番目の言語表現に類似する言語表現が類似情報記憶部１２で記憶されている場合には、ステップＳ２０４に進み、そうでない場合には、ステップＳ２０６に進む。なお、この判断は、例えば、そのｊ番目の言語表現を検索キーとして、類似情報記憶部１２を検索することによって行ってもよい。そして、その検索でヒットした場合には、ｊ番目の言語表現に類似する言語表現が記憶されていることになり、ヒットしなかった場合には、ｊ番目の言語表現に類似する言語表現が記憶されていないことになる。

【００７０】

（ステップＳ２０４）拡張関係候補情報生成部１３は、ｉ番目の関係情報に含まれるｊ番目の言語表現に類似する言語表現を類似情報記憶部１２から読み出す。具体的には、ステップＳ２０３の検索でヒットした類似情報から、ｉ番目の関係情報に含まれるｊ番目の言語表現以外の言語表現を読み出すことによって行われてもよい。その読み出した言語表現は、図示しない記録媒体で一時的に記憶されてもよい。そして、拡張関係候補情報生成部１３は、ｉ番目の関係情報に含まれるｊ番目の言語表現を、類似情報記憶部１２から読み出した言語表現に置換した拡張関係候補情報を生成する。例えば、類似情報記憶部１２から２以上の言語表現を読み出した場合には、拡張関係候補情報生成部１３は、ｉ番目の関係情報に含まれるｊ番目の言語表現を、その２以上の言語表現のそれぞれに置換した２以上の拡張関係候補情報を生成することになる。

【００７１】
なお、このステップＳ２０４において、拡張関係候補情報生成部１３は、ｉ番目の関係情報に関して、（ｊ－１）番目までの少なくともいずれかの言語表現を置換した拡張関係候補情報についても、ｊ番目の言語表現の置換を行ってもよい。例えば、関係情報が３個の言語表現を含んでおり、ｊ＝３である場合に、ｉ番目の関係情報に関して、１番目の言語表現のみが置換された拡張関係候補情報と、２番目の言語表現のみが置換された拡張関係候補情報と、１番目の２番目の言語表現の両方が置換された拡張関係候補情報とが存在する場合には、拡張関係候補情報生成部１３は、その３個の拡張関係候補情報について、ｊ番目（＝３番目）の言語表現を類似する言語表現に置換する処理を行ってもよい。
【００７２】
（ステップＳ２０５）拡張関係候補情報生成部１３は、置換後の拡張関係候補情報を拡張関係候補情報記憶部１４に蓄積する。なお、ステップＳ２０４において、ｉ番目の関係情報に関して、それまでに置換の行われた拡張関係候補情報についても置換を行う場合には、このステップＳ２０５の蓄積時に、蓄積対象の拡張関係候補情報が、ｉ番目の関係情報に関するものであることが分かるように蓄積することが好適である。拡張関係候補情報生成部１３は、例えば、カウンタｉの値に対応付けて拡張関係候補情報を蓄積してもよい。
【００７３】
また、拡張関係候補情報生成部１３は、ステップＳ２０４と、ステップＳ２０５との処理を、１個の拡張関係候補情報を生成するごとに繰り返して実行してもよい。例えば、ｉ番目の関係情報に含まれるｊ番目の言語表現に類似する言語表現が２以上あった場合には、拡張関係候補情報生成部１３は、ｊ番目の言語表現を各言語表現に置換するごとに、置換後の拡張関係候補情報を拡張関係候補情報記憶部１４に蓄積してもよい。
【００７４】
（ステップＳ２０６）拡張関係候補情報生成部１３は、カウンタｊを１だけインクリメントする。
【００７５】
（ステップＳ２０７）拡張関係候補情報生成部１３は、ｉ番目の関係情報にｊ番目の言語表現が存在するかどうか判断する。そして、存在する場合には、ステップＳ２０３に戻り、そうでない場合には、ステップＳ２０８に進む。なお、関係情報に含まれる言語表現の個数はあらかじめ決まっているため、例えば、その個数を図示しない記録媒体で記憶しておき、拡張関係候補情報生成部１３は、その記録媒体で記憶されている言語表現の個数と、カウンタｊの値とを比較することによって、このステップＳ２０７の処理を行ってもよい。その場合には、ｊ＞（記憶されている言語表現の個数）である場合には、ステップＳ２０８に進むことになる。
【００７６】
（ステップＳ２０８）拡張関係候補情報生成部１３は、カウンタｉを１だけインクリメントする。
【００７７】
（ステップＳ２０９）拡張関係候補情報生成部１３は、関係情報記憶部１１にｉ番目の関係情報が存在するかどうか判断する。そして、存在する場合には、ステップＳ２０２に戻り、そうでない場合には、図２のフローチャートに戻る。
【００７８】
図４は、図２のフローチャートにおけるスコアの取得の処理（ステップＳ１０２）の詳細を示すフローチャートである。なお、図４のフローチャートでは、共起スコアを取得し、その後に共起スコアの高いものを暫定的に選択して、その選択された拡張関係候補情報についてのみ、経由スコアを取得する場合について説明する。
【００７９】
（ステップＳ３０１）スコア取得部１５は、拡張関係候補情報記憶部１４で記憶されているすべての拡張関係候補情報について、共起スコアを取得する。

【００８０】
（ステップＳ３０２）選択部１６は、ステップＳ３０１で取得された共起スコアの高い拡張関係候補情報を暫定的に選択する。選択部１６は、例えば、その選択後の拡張関係候補情報を拡張関係候補情報記憶部１４や、図示しない記録媒体に蓄積してもよく、あるいは、暫定的に選択した、拡張関係候補情報記憶部１４で記憶されている拡張関係候補情報に対して、暫定的に選択されたことを示すフラグ等の識別情報を設定してもよい。
【００８１】
（ステップＳ３０３）スコア取得部１５は、ステップＳ３０２で暫定的に選択されたすべての拡張関係候補情報について、経由スコアを取得する。
【００８２】
このように、共起スコアを取得し、共起スコアの高いものを暫定的に選択し、その暫定的に選択された拡張関係候補情報について経由スコアを取得することによって、経由スコアを取得する拡張関係候補情報の個数を減らすことができる。通常、共起スコアの取得よりも、経由スコアの取得の方が負荷の高い処理であるため、この順序でスコアの取得を行うことによって、スコアの取得の処理を軽減することができうる。この場合には、ステップＳ１０３の選択の処理において、経由スコアの高い拡張関係候補情報を選択すればよいことになる。
【００８３】
なお、図４のフローチャートにおけるスコアの取得の方法は一例であり、拡張関係候補情報記憶部１４で記憶されているすべての拡張関係候補情報について、共起スコアと経由スコアとの両方を取得してもよい。この場合には、ステップＳ１０３の選択の処理において、共起スコアが高く、かつ、経由スコアが高い拡張関係候補情報を選択すればよいことになる。また、共起スコアと、経由スコアとの一方のみを用いた選択を行う場合には、図４のフローチャートにおいて、拡張関係候補情報記憶部１４で記憶されているすべての拡張関係候補情報について、その選択で用いるスコアのみの取得を行ってもよい。この場合には、ステップＳ１０３の選択の処理において、選択で用いるスコア（共起スコア、または、経由スコア）の高い拡張関係候補情報を選択すればよいことになる。また、図４のフローチャートとは逆に、まず経由スコアを取得し、その後に経由スコアの高いものを暫定的に選択して、その選択された拡張関係候補情報についてのみ、共起スコアを取得してもよい。
【００８４】
次に、本実施の形態による関係情報拡張装置１の動作について、簡単な具体例を用いて説明する。この具体例において、共起スコアのみを用いて選択を行うものとする。また、この具体例では、関係情報記憶部１１において、死亡の原因が心筋梗塞であることを示す関係情報＜心筋梗塞、死亡＞のみが記憶されているものとする。また、類似情報記憶部１２では、図５で示される類似情報が記憶されているものとする。図５において、一つのレコードが、一つの類似情報である。また、一つの類似情報に含まれる各言語表現は、互いに類似するものである。例えば、心筋梗塞、脳梗塞、脳卒中、うつ病は、互いに類似する言語表現である。
【００８５】
関係情報を拡張する処理が開始されると、まず、拡張関係候補情報生成部１３が、関係情報記憶部１１で記憶されている関係情報＜心筋梗塞、死亡＞から拡張関係候補情報を生成する処理を行う（ステップＳ１０１）。具体的には、拡張関係候補情報生成部１３は、その関係情報の１番目の言語表現「心筋梗塞」を検索キーとして類似情報記憶部１２を検索する。その結果、１番目のレコードに含まれる「心筋梗塞」がヒットするため、拡張関係候補情報生成部１３は、その１番目のレコードから、検索キー以外の言語表現「脳梗塞」「脳卒中」「うつ病」を読み出して図示しない記録媒体に蓄積すると共に、１番目の関係情報の１番目の言語表現「心筋梗塞」に類似する言語表現が存在すると判断する（ステップＳ２０１～Ｓ２０３）。そして、拡張関係候補情報生成部１３は、関係情報＜心筋梗塞、死亡＞の１番目の言語表現「心筋梗塞」を、それに類似する言語表現「脳梗塞」「脳

卒中」「うつ病」に置換した拡張関係候補情報をそれぞれ生成し、それらの拡張関係候補情報を、その時点のカウンタｉの値に対応付けて拡張関係候補情報記憶部１４に蓄積する（ステップＳ２０４，Ｓ２０５）。図６の１番目から３番目のレコードは、そのようにして蓄積された拡張関係候補情報を含んでいる。なお、図６において、関係情報ＩＤは、カウンタｉの値である。共起スコアは、後にスコア取得部１５によって取得されるものであるため、現段階では空欄である。その後、拡張関係候補情報生成部１３は、関係情報＜心筋梗塞、死亡＞の２番目の言語表現「死亡」に類似する言語表現「病死」「急死」を類似情報記憶部１２から読み出して蓄積すると共に、１番目の関係情報の２番目の言語表現「死亡」に類似する言語表現が存在すると判断する（ステップＳ２０６，Ｓ２０７，Ｓ２０３）。そして、拡張関係候補情報生成部１３は、関係情報＜心筋梗塞、死亡＞と、それまでに蓄積された関係情報ＩＤ「１」に対応する拡張関係候補情報＜脳梗塞、死亡＞、＜脳卒中、死亡＞、＜うつ病、死亡＞との２番目の言語表現「死亡」を、それに類似する言語表現「病死」「急死」に置換した拡張関係候補情報をそれぞれ生成し、それらの拡張関係候補情報を、その時点のカウンタｉの値に対応付けて拡張関係候補情報記憶部１４に蓄積する（ステップＳ２０４，Ｓ２０５）。その結果、拡張関係候補情報記憶部１４で記憶されている情報は、図６で示されるようになる。

【００８６】

次に、スコア取得部１５は、図６で示される各拡張関係候補情報に含まれる２個の言語表現の共起スコアをそれぞれ取得し、その拡張関係候補情報に対応付けて拡張関係候補情報記憶部１４に蓄積する（ステップＳ１０２，Ｓ３０１）。その結果、拡張関係候補情報記憶部１４で記憶されている情報は、図７で示されるようになったとする。なお、この共起スコアは共起頻度であるとする。その後、選択部１６は、共起スコアを用いた選択を行う。この選択では、しきい値が５０に設定されており、そのしきい値以上の共起スコアの拡張関係候補情報が、拡張関係情報として選択されるものとする。すると、選択部１６は、図７でしめされる拡張関係情報のうち、＜うつ病、病死＞と＜うつ病、急死＞以外の拡張関係情報を選択して拡張関係情報記憶部１７に蓄積する（ステップＳ１０３）。その結果、拡張関係情報記憶部１７では、図７の１番目から９番目までの拡張関係候補情報である拡張関係情報が記憶されることになる。このようにして、関係情報＜心筋梗塞、死亡＞を、拡張関係情報＜脳梗塞、死亡＞等に拡張することができる。なお、この具体例で示した共起スコア等は、本実施の形態による関係情報拡張装置１の動作の詳細を説明するために示したものであり、実際の文書等を用いて取得したデータではない。

【００８７】

なお、この具体例では、互いに類似する２以上の言語表現が一の類似情報に含まれる場合について説明したが、そうでなくてもよい。類似情報は、例えば、類似する２個の言語表現を有する情報であってもよい。その場合には、例えば、図５の１番目のレコードは、心筋梗塞と脳梗塞のペア、心筋梗塞と脳卒中のペア、心筋梗塞とうつ病のペア、脳梗塞と脳卒中のペア、脳梗塞とうつ病のペア、脳卒中とうつ病のペアというように、６個の類似情報に分かれることになる。

【００８８】

次に、本実施の形態による関係情報拡張装置１の実験例について説明する。この実験例では、「ＸはＹの原因となる」の関係を有する関係情報＜Ｘ，Ｙ＞から得られた拡張関係情報の精度の評価と、従来のパターンベースの方法で取得することが困難であった関係を取得できているかどうかの評価とを行う。なお、この実験例において、言語表現は、名詞または連続する名詞である。

【００８９】

まず、評価方法について説明する。評価は、３人の評価者によって行った。そして、（１）常識的に正解と判断された場合、あるいは、（２）常識的に正解と判断されなくても、ウェブに正しいと支持するエビデンスが１つ以上見つかった場合を正解とした。なお、正解であるとは、「ＸはＹの原因となる」という関係が成り立つことである。また、３人の評価者のうち、２名以上一致（ｌｅｎｉｅｎｔ）、３名一致（ｓｔｒｉｃｔ）で精度を

測定した。

【００９０】

なお、（２）では、１つの関係に関して、ＹａｈｏｏＡＰＩを用いて「Ｘ、Ｙ、原因」のＡＮＤ検索で１０ページを獲得し、各ページから「Ｘ、Ｙ、原因」が２００文字以内に存在するテキストセグメントを最大３つ抽出して、最大３０個（＝１０×３）のセグメントを評価者に提示することによって行った。その最大３０個のセグメントのうち、少なくとも１つが評価者によって妥当であると判断された場合には、正解であることになる。

【００９１】

本実験では、各評価者で合計４００個の評価を行ったが、評価者間のｋａｐｐａ値は、平均で、０．６２９であった。一般的に、ｋａｐｐａ値が０．６以上ならば「かなりよい一致率」と言われていることから、評価者間の判定の一致率は概ねよいといえる。

【００９２】

次に、この実験例で用いた類似情報について説明する。類似語の獲得には、前述の風間らの文献の方法で作成された約５０万名詞に対する類似度付き類似語リスト（ＡＬＡＧＩＮフォーラムで公開されている文脈類似語データベースＶｅｒｓｉｏｎ１のｏｌｄ．５００ｋ－２ｋ．ｄａｔａ．）を用いた。

【００９３】

風間らの方法では、大量コーパスから各名詞ｎの（助詞、動詞）、（の、名詞）の大きく２種類の係り受け関係ｄｅｐを収集し、Ｔｏｒｉｓａｗａの文献の手法（次の文献参照）、

【数３】

$$P(n,dep)=\sum_{c_i\in C}P(c_i)P(n\mid c_i)P(dep\mid c_i)$$

に基づき、ＥＭアルゴリズムでＰ（ｃ）、Ｐ（ｎ｜ｃ）、Ｐ（ｄｅｐ｜ｃ）を推定する（確率モデルとしてはＰＬＳＩと等価である）。これによって、ｄｅｐをそのまま素性とする場合と比べてスムージング効果が期待できる。次に、上記パラメータからＰ（ｃ｜ｎ）を計算し、名詞ｎ１、ｎ２の類似度をＰ（ｃ｜ｎ１）、Ｐ（ｃ｜ｎ２）のＪｅｎｓｅｎ－Ｓｈａｎｎｏｎ（ＪＳ）ダイバージェンスとして求める。ＪＳダイバージェンスは確率分布間の距離の一種で、以下の式で計算する。

【数４】

$$JS(P1\parallel P2)=\frac{1}{2}(KL(P1\parallel P_{mean})+KL(P2\parallel P_{mean}))$$

【００９４】

ここで、Ｐ１、Ｐ２は確率分布、ＫＬ（Ｐ１｜｜Ｐ２）はＫＬダイバージェンス、P_{mean}はＰ１、Ｐ２をベクトルとしてみた場合の平均である。ＪＳダーバージェンスは０から１を取り、小さいほど類似していることになる。そのため、単語ｎ１、ｎ２の類似度は次のようにする。

【数５】

$$sim(n1,n2)=1-JS(P(c\mid n1)\parallel P(c\mid n2))$$

【００９５】

最終的に、可能な単語集合の中の全ペアについて、（Ａ）ｓｉｍ（ｎ１、ｎ２）がしきい値T_{sim}以上である、（Ｂ）互いの類似度のトップＭ単語に含まれる、の２つの条件を満たす単語ペアを類似情報として獲得した。この実験例では、しきい値T_{sim}＝０．７、Ｍ＝２０として類似情報を生成した。なお、実験例で用いた各類似情報には、類似している単語のペアと、その単語のペアの類似度とが含まれている。

【００９６】

文献：Ｋ．Ｔｏｒｉｓａｗａ，「Ａｎ　Ｕｎｓｕｐｅｒｖｉｅｓｅｄ　Ｍｅｔｈｏｄ

ｆｏｒ Ｃａｎｏｎｉｃａｌｉｚａｔｉｏｎ ｏｆ Ｊｐａｎｅｓｅ Ｐｏｓｔｐｏｓｉｔｉｏｎｓ」，Ｉｎ Ｐｒｏｃ．ｏｆ ｔｈｅ ６ｔｈ ＮＬＰＲＳ，ｐ．２１１－２１８，２００１年

【００９７】

また、この実験例において、単語共起頻度である共起スコアを用いた。その単語共起頻度は、約１億文書で上記と同じ約５０万名詞の全ペアに対して、近接４文内で共起する文書頻度を計算したデータ（ＡＬＡＧＩＮフォーラムで公開されている単語共起頻度データベースＶｅｒｓｉｏｎ１の５００ｋ－５００ｋ．１００ｍ－ｄｏｃｓ．ｗ４．ｄａｔａ）を用いた。その共起頻度である共起スコアのしきい値は、T_{cooc}＝２０とした。これらのしきい値等のパラメータは、共起スコアを用いた選択を行った場合に、経験的に、関係情報の約１０倍の関係が生成されることを目安に設定した。

【００９８】

次に、この実験例で用いた関係情報について説明する。関係情報は、Ｄｅ Ｓａｅｇｅｒらの文献の方法（次の文献参照）で獲得した関係から、明らかに不適切な関係をクリーニングした上で、トップ１万個を用いた。そのＤｅ Ｓａｅｇｅｒらの文献の方法による関係獲得でパターン学習に用いたデータは約５千万文書で、対象の単語集合は、上記と同じ約５０万名詞である。その方法の詳細は文献に譲るが、シードパターンを入力し、それらシードパターンと同じ２語を抽出できる全パターンを用いて関係を再獲得してランキングするため、パターンベースの方法では、最高レベルの網羅性と考えられる。前述の評価法と同様の基準で関係情報の精度を測定したところ、ｌｅｎｉｅｎｔで０．８０、ｓｔｒｉｃｔで０．７０であった。つまり、ノイズが含まれる関係情報からの類推となる。ただし、本評価は、Ｄｅ Ｓａｅｇｅｒらの文献と方法が異なり、Ｄｅ Ｓａｅｇｅｒらの文献と比較するとやや低めの値となる傾向にある点に注意されたい。

【００９９】

文献：Ｓ．Ｄｅ Ｓａｅｇｅｒ，Ｋ．Ｔｏｒｉｓａｗａ，Ｊ．Ｋａｚａｍａ，Ｋ．Ｋｕｒｏｄａ，Ｍ．Ｍｕｒａｔａ，「Ｌａｒｇｅ Ｓｃａｌｅ Ｒｅｌａｔｉｏｎ Ａｃｑｕｉｓｉｔｉｏｎ Ｕｓｉｎｇ Ｃｌａｓｓ Ｄｅｐｅｎｄｅｎｔ Ｐａｔｔｅｒｎｓ」，Ｉｎ Ｐｒｏｃ．ｏｆ ｔｈｅ ９ｔｈ ＩＣＤＭ，ｐ．７６４－７６９，２００９年

【０１００】

これらの類似情報、関係情報、単語共起頻度を用いて、拡張関係候補情報を生成し、共起スコアを用いた選択を行ったところ、１万の関係情報から１０２２９０個の新しい関係（拡張関係候補情報）が生成された。

【０１０１】

次に、この実験例で用いた経由スコアについて説明する。この実験例では、経由スコアとして、前述のS^{sum}とS^{prod}との２種類を用い、それぞれでランキングした結果を評価した。精度は、関係情報の関係を除いた上で、各経由スコアのトップ１万から１００個、１万から３万の１００個の２００個を評価した。結果を図８に示す。図８の１５０００位以降の精度は、トップ１万までの精度と、トップ１万から３万までの精度を用いて補間したものである。ｌｅｎｉｅｎｔを正解とすると、ｐｒｏｄ（S^{prod}）は、トップ１万の精度が０．６３、ｓｕｍ（S^{sum}）は、０．５３であった。関係情報の精度（ｌｅｎｉｅｎｔで０．８０）と比較するとやや精度が低下しているが、文中での書かれ方を用いずに、これだけの精度を達成できた。なお、トップ１万は、１万個の関係情報から１万個の拡張関係情報（これには関係情報は含まれない）を取得したことになる。したがって、少しの精度の低下によって、関係の個数を倍に拡張できることが分かる。また、図８の通り、ｓｕｍとｐｒｏｄを比較すると、ｐｒｏｄの方が上位の精度が高いため、よいスコアとなっていると考えられる。

【０１０２】

次に、パターンベースで獲得困難である関係を取得できたかどうかを調べた。具体的には、トップ１万のｌｅｎｉｅｎｔを正解と考え、Ｄｅ Ｓａｅｇｅｒらの文献の方法で、順位が１００万位以下である関係の数の割合を１００個のサンプリングを用いて調査し、

その割合を用いてトップ１万に含まれる正解の関係数を推定した。ただし、Ｄｅ　Ｓａｅｇｅｒらの文献の方法では、５千万文書を用いていることに対して、この実験例では１億文書での共起頻度を用いているので、フェアな比較とはいえない。厳密には、文書集合を揃えた上で比較すべきで、Ｄｅ　Ｓａｅｇｅｒらとの比較は、参考的なものである。結果は、次のようになった。

【０１０３】

	Ｄｅ　Ｓａｅｇｅｒら方法で１００万位以下の関係数
ｓｕｍ	約３１００
ｐｒｏｄ	約３３００

【０１０４】

上記の結果より、この実験では、パターンベースの従来法では獲得困難であった関係が実際に獲得されていることが確認できた。例えば、関係情報に含まれなかった関係として、＜ミネラル不足、花粉症＞や、＜食習慣、ニキビ＞を取得することができた。なお、前者は、関係情報＜カルシウム不足、アトピー＞等から両方の単語を置換することによって生成されたものであり、後者は、関係情報＜生活習慣、ニキビ＞等から一方の単語を置換することによって生成されたものである。このように、本実験例では、関係情報に含まれる両方の単語を置換した場合であっても、スコアを用いた選択を行うことによって、精度を維持できていると考えられる。

【０１０５】

以上のように、本実施の形態による関係情報拡張装置１によれば、類似情報を用いて関係情報を拡張するため、従来のパターンベースの手法では獲得することのできなかった関係を生成することができる。また、言語表現を類似する言語表現に置換することによって新たな関係を生成するため、語基の共通しないものに言語表現を置換することができ、前述の非特許文献２の場合よりも、より広範囲な拡張が可能となる。また、拡張した関係に対して、スコアを取得し、そのスコアを用いた選択を行うことによって、不適切な関係を除去することができうる。したがって、選択された拡張関係情報は、意味的関係を適切に有するものになりうる。

【０１０６】

ここで、このようにして生成された拡張関係情報によって示される関係の用い方について簡単に説明する。例えば、本実施の形態による関係情報拡張装置１によって、拡張関係情報＜心筋梗塞、急死＞、＜脳梗塞、急死＞、＜脳卒中、急死＞を得ることができていた場合には、情報検索システムのキーワード推薦として、ユーザが「急死」を入力した場合に、「"急死"の"原因"には、"心筋梗塞"、"脳梗塞"、"脳卒中"などがあります」等のように、意味的関係で整理された推薦を行うことができるようになる。また、拡張関係情報に、上位＜薬、抗ウィルス薬剤＞や、効果＜抗ウィルス薬剤、インフルエンザ＞が存在する場合には、「インフルエンザに効く薬は？」という質問に対して、適切な推論によって「抗ウィルス薬剤」と答えることができる。また、拡張関係情報の示す関係を、他の種々の用途で用いることもできうる。なお、その際に、拡張関係情報のみを用いてもよく、拡張関係情報と関係情報とをマージしたものを用いてもよい。

【０１０７】

なお、本実施の形態による関係情報拡張装置１において、スコアを取得する方法は、前述のものに問われないことは言うまでもない。例えば、スコア取得部１５が取得するスコアは、共起スコアと、経由スコアとを２個の引数とする各引数に関する増加関数の値であってもよい。そして、増加関数の値であるスコアを用いて、選択が行われてもよい。その増加関数は、例えば、Ｃ１×共起スコア＋Ｃ２×経由スコアであってもよい。ここで、Ｃ１，Ｃ２は、正の係数である。

【０１０８】

また、本実施の形態による関係情報拡張装置１において、共起スコアを取得する方法は、前述のものに問われないことは言うまでもない。例えば、スコア取得部１５は、拡張関係候補情報に含まれる２以上の言語表現と、その拡張関係候補情報に対応する共起言語表

現とが共起する方が、その拡張関係候補情報に含まれる２以上の言語表現のみが共起するよりも高い値となる共起スコアを取得してもよい。ここで、拡張関係候補情報に対応する共起言語表現とは、その拡張関係候補情報の生成時に用いられた関係情報の意味的関係と同じ種類の意味的関係を有する各関係情報（この関係情報は、関係情報記憶部１１で記憶されている関係情報であってもよく、あるいは、そうでなくてもよい。）に含まれる２以上の言語表現に対して共起の高い言語表現である。なお、共起が高い言語表現とは、前述のスコアが高い場合と同様に、例えば、しきい値以上の共起の頻度である言語表現であってもよく、あるいは、共起の頻度の高いものからあらかじめ決められた個数の言語表現であってもよい。具体例を用いて説明すると次のようになる。例えば、拡張関係候補情報＜心筋梗塞、急死＞が、関係情報＜心筋梗塞、死亡＞を用いて生成されたとする。そして、その関係情報＜心筋梗塞、死亡＞の意味的関係の種類が「原因」であったとする。また、意味的関係の種類が「原因」である２以上（多数であることが好適である）の各関係情報に含まれる２以上の言語表現（すべての言語表現）と共起の高い言語表現として、「原因」「理由」「要因」…があったとする。すると、「原因」「理由」「要因」…が、共起言語表現になる。また、拡張関係候補情報＜心筋梗塞、急死＞に含まれる２個の言語表現が、共起言語表現「原因」「理由」「要因」…のいずれかと共起する場合には、拡張関係候補情報＜心筋梗塞、急死＞に含まれる２個の言語表現が、共起言語表現「原因」「理由」「要因」…のいずれとも共起しない場合に比べて、共起スコアは高くなる。その場合に、（１）保持している共起言語表現を用いて共起スコアを取得する方法と、（２）機械学習を用いて共起スコアを取得する方法とがある。以下、その各方法について説明する。なお、（１）（２）以外の方法によって、上述のように共起スコアを取得してもよいことは言うまでもない。

【０１０９】

（１）保持している共起言語表現を用いて共起スコアを取得する方法

この方法では、関係情報拡張装置は、図９で示されるように、１以上の対応情報が記憶される対応情報記憶部２１をさらに備えている。ここで、対応情報は、種類識別情報と、その種類識別情報に対応する、その種類識別情報で識別される意味的関係の種類に対応する１以上の共起言語表現とを有する情報である。ここで、種類識別情報は、関係情報の意味的関係の種類を識別する情報である。本方法の場合には、関係情報記憶部１１で記憶されている各関係情報は、その関係情報が有する２以上の言語表現の意味的関係の種類を識別する情報である種類識別情報をも有するものであるとする。ここでは、関係情報が、「種類識別情報＜第１の言語表現、第２の言語表現＞」の形式で示されるものとする。例えば、「原因＜心筋梗塞、死亡＞」となる。また、拡張関係候補情報生成部１３は、拡張関係候補情報の生成に用いる関係情報が有する種類識別情報を有する拡張関係候補情報を生成するものとする。したがって、関係情報「原因＜心筋梗塞、死亡＞」を用いて生成された拡張関係候補情報は、例えば、「原因＜脳梗塞、死亡＞」となる。そして、スコア取得部１５は、拡張関係候補情報に含まれる２以上の言語表現と、その拡張関係候補情報が有する種類識別情報に対応する各共起言語表現とが共起する方が、その拡張関係候補情報に含まれる２以上の言語表現のみが共起するよりも高い値となる共起スコアを取得するものとする。ここで、「拡張関係候補情報に含まれる２以上の言語表現のみが共起する」とは、その拡張関係候補情報に含まれる２以上の言語表現が、共起言語表現のいずれとも共起しないことである。

【０１１０】

なお、共起スコアが、拡張関係候補情報に含まれる２以上の言語表現と、その拡張関係候補情報が有する種類識別情報に対応する各共起言語表現とが共起する方が、その拡張関係候補情報に含まれる２以上の言語表現のみが共起するよりも高い値となることは、結果としてそのようになればよいのであって、その方法は問わない。例えば、スコア取得部１５は、拡張関係候補情報に含まれる２以上の言語表現と、その拡張関係候補情報が有する種類識別情報に対応するいずれかの共起言語表現とが共起する場合には、前述のように、拡張関係候補情報に含まれる２以上の言語表現に対する共起頻度などを用いて取得した共

起の尺度に対して、１を超える数（例えば、１．２や１．５、２など）を掛けた値を共起スコアとしてもよい。また、共起する共起言語表現の数が多いほど、共起スコアが高くなるようにしてもよい。例えば、ある拡張関係候補情報について、その拡張関係候補情報が有する種類識別情報に対応する共起言語表現の数がＡであるとする。また、そのＡ個の共起言語表現のうち、その拡張関係候補情報に含まれる２以上の言語表現と共起する共起言語表現の数がＢであったとする。また、Ｒ＝Ｂ／Ａとする。そして、スコア取得部１５は、Ｒを引数とする増加関数の値である共起スコアを取得してもよい。具体的には、前述のように、共起頻度などを用いて取得した共起の尺度に対して、（１＋Ｃ×Ｒ）を掛けた値を共起スコアにしてもよい。なお、Ｃは、正の係数である。さらに、共起言語表現との共起の程度が高いほど、共起スコアがより高くなるようにしてもよい。例えば、前述のＢの値を、拡張関係候補情報に含まれる２以上の言語表現と共起言語表現との共起頻度の和などにしてもよい。

【０１１１】

ここで、具体例を用いて説明する。対応情報記憶部２１において、図１０で示される対応情報が記憶されていたとする。図１０の対応情報において、種類識別情報と、共起言語表現とが対応付けられている。例えば、種類識別情報「食材」に、共起言語表現「材料」「レシピ」…が対応している。したがって、種類識別情報「食材」を有する関係情報（例えば、食材＜シチュー、じゃがいも＞のように、シチューの食材がじゃがいもであることを示す関係情報等）に含まれるすべての言語表現と共起の高い言語表現が、「材料」や「レシピ」等であることが示されていることになる。また、拡張関係候補情報記憶部１４において、拡張関係候補情報「食材＜シチュー、サツマイモ＞」が記憶されていたとする。すると、スコア取得部１５は、前述のように、２個の言語表現「シチュー」「サツマイモ」の共起の尺度を算出する。また、スコア取得部１５は、その拡張関係候補情報「食材＜シチュー、サツマイモ＞」に含まれる種類識別情報「食材」に対応する共起言語表現「材料」「レシピ」…を、図１０の対応情報を用いて取得する。そして、共起言語表現を順番に変えながら、３個の言語表現「シチュー」「サツマイモ」「共起言語表現」が共起するかどうか判断する。ここで、例えば、種類識別情報「食材」に対応する共起言語表現の総数が２０個であり、そのうち、「シチュー」「サツマイモ」と共起した共起言語表現の個数が５個であったとする。すると、スコア取得部１５は、前述のように、共起スコア＝共起の尺度×（１＋Ｃ×５／２０）を取得してもよい。一方、例えば、拡張関係候補情報「食材＜シチュー、デンプン＞」に対しては、２個の言語表現「シチュー」「サツマイモ」と共起する共起言語表現が存在しなかったとする。すると、その拡張関係候補情報「食材＜シチュー、デンプン＞」に対しては、スコア取得部１５は、２個の言語表現「シチュー」「サツマイモ」の共起の尺度そのものを共起スコアにする。

【０１１２】

（２）機械学習を用いて共起スコアを取得する方法

この方法では、スコア取得部１５は、機械学習を用いて、共起スコアを取得する。すなわち、スコア取得部１５は、２以上の言語表現の組に含まれるその２以上の言語表現と共起する言語表現を少なくとも素性として用い、その素性の値及び２以上の言語表現の組に対する意味的関係の有無（なお、この意味的関係の有無は、その言語表現の組に含まれる２以上の言語表現の意味的関係の有無である）を教師データとする機械学習を行い、拡張関係候補情報に含まれる２以上の言語表現を入力した場合の出力である確信度に応じた共起スコアを取得する。その機械学習について、以下、説明する。

【０１１３】

この機械学習の問題（入力）は、共起スコアを取得する対象となる、言語表現の組（その言語表現の組は、２以上の言語表現を有している。また、その言語表現の組に含まれる言語表現の数は、関係情報に含まれる言語表現の数と同じであるとする。）である。具体的には、拡張関係候補情報である。また、その機械学習の解（出力）は、問題（入力）である２以上の言語表現の組（拡張関係候補情報）に含まれる２以上の言語表現が、その拡張関係候補情報に対応する意味的関係の種類と同じ意味的関係の種類を有する関係情報と

同様の共起であるかどうかの確信度である。拡張関係候補情報に対応する意味的関係の種類とは、その拡張関係候補情報の生成で用いられた関係情報の意味的関係の種類である。なお、その解（出力）には、問題（入力）の拡張関係候補情報に含まれる２以上の言語表現が、その拡張関係候補情報に応じた意味的関係の種類と同じ意味的関係の種類を有する関係情報と同様の意味的関係を有するかどうかの情報が含まれてもよい。また、その機械学習の素性には、問題（入力）である２以上の言語表現の組に含まれる２以上の言語表現（すべての言語表現）と共起する言語表現のリストが含まれるものとする。そのリストは、問題（入力）である２以上の言語表現の組に含まれる２以上の言語表現と共起するすべての言語表現のリストであってもよく、あるいは、問題（入力）である２以上の言語表現の組に含まれる２以上の言語表現と共起の高い言語表現のリストであってもよい。このリストを作成するためには、スコア取得部１５は、例えば、問題（入力）である２以上の言語表現の組に含まれる２以上の言語表現と、あらゆる言語表現とが共起するかどうかを判断してもよい。ここで、その処理で用いられるあらゆる言語表現は、例えば、あらかじめ図示しない記録媒体で記憶されている言語表現群であってもよい。そして、共起すると判断された言語表現を、そのリストに含めるようにしてもよい。また、共起の高いもののみをリストに含める場合には、スコア取得部１５は、共起の尺度（例えば、共起頻度や共起率等である。なお、ここで用いられる共起の尺度は、例えば、ダイス係数や相互情報量などのように、２個の言語表現に対してのみ定義されている共起の尺度を用いたものではなく、３個以上の言語表現に対しても定義されているものであることが好適である）も算出し、それに応じて共起の高いものを選択してもよい。なお、共起が高いものとは、前述のスコアが高い場合と同様に、例えば、しきい値以上の共起の頻度である言語表現であってもよく、あるいは、共起の頻度の高いものからあらかじめ決められた個数の言語表現であってもよい。また、その機械学習の素性には、問題（入力）である２以上の言語表現の組に含まれる２以上の言語表現、問題（入力）である２以上の言語表現の組に含まれる２以上の言語表現の共起の尺度、問題（入力）である２以上の言語表現の組に含まれる２以上の言語表現の属性（例えば、言語表現の品詞や、言語表現の上位語等）、問題（入力）である２以上の言語表現の組としての拡張関係候補情報に対応する意味的関係の種類、問題（入力）である２以上の言語表現の組に含まれる２以上の言語表現と共起する言語表現のリストに含まれる各言語表現に関する、問題（入力）である２以上の言語表現の組に含まれる２以上の言語表現との共起の尺度（例えば、共起頻度や共起率等であり、前述のように、３個以上の言語表現に対しても定義されている共起の尺度であることが好適である）のうち、任意の１以上のものが素性に含まれてもよい。なお、意味的関係の種類を素性に用いる場合には、例えば、問題（入力）である２以上の言語表現の組に種類識別情報が含まれており、その種類識別情報を素性に用いてもよい。また、意味的関係の種類を素性に用いずに、意味的関係の種類ごとに学習を行い、その意味的関係の種類ごとの学習結果を用いて、共起スコアを取得してもよい。例えば、種類識別情報「原因」に対応する拡張関係候補情報に対する共起スコアを取得する際には、種類識別情報「原因」に対応して学習された学習結果を用いて、共起スコアを取得してもよい。

【０１１４】

また、その機械学習で用いられる教師データ（訓練データ）は、２以上の言語表現の組に対する意味的関係の有無と、その２以上の言語表現の組に対応する、前述の素性の各値とである。例えば、教師データの正例（すなわち、意味的関係のあるもの）である、種類識別情報「原因」に対応する２以上の言語表現の組としては、種類識別情報「原因」で識別される意味的関係の種類である関係情報を用いてもよい。その関係情報は、関係情報記憶部１１で記憶されているものであってもよく、あるいは、そうでないものであってもよい。また、教師データの負例（すなわち、意味的関係のないもの）である、種類識別情報「原因」に対応する２以上の言語表現の組としては、任意のコーパスからランダムに取得した２以上の言語表現の組を用いてもよい。

【０１１５】

教師データを用いた学習の後に、判断の対象となる、拡張関係候補情報を入力すると、

その拡張関係候補情報に関する素性の各値が取得され、その拡張関係候補情報に含まれる２以上の言語表現の意味的関係に関する確信度が出力される。例えば、その確信度は、－１から１までの範囲の値であってもよい。また、先述のように、意味的関係を有するかどうかの結果も出力されてもよい。例えば、確信度が－１から０までであれば意味的関係を有しないという結果になり、確信度が０を超えて１までであれば意味的関係を有するという結果になる。スコア取得部１５は、その確信度に応じた共起スコアを取得する。具体的には、確信度が－１から１までの値である場合には、共起スコアは、その確信度を引数とする増加関数の値であってもよい。具体的には、共起スコア＝Ｃ×（確信度＋１）であってもよい。なお、Ｃは、任意の係数である。また、共起スコアが正の値になるように、確信度に１を足している。なお、拡張関係候補情報に含まれる２以上の言語表現が意味的関係を有する場合にも、有しない場合にも、確信度が０から１までの範囲の値であるのであれば、意味的関係を有する場合には、共起スコア＝Ｃ×（１＋確信度）として、意味的関係を有しない場合には、共起スコア＝Ｃ×（１－確信度）としてもよい。また、共起スコアは、拡張関係候補情報に含まれる２以上の言語表現が意味的関係を有するという結果の場合の確信度のみを用いてもよい。その場合には、例えば、共起スコア＝Ｃ×確信度であってもよい。このように、機械学習を用いて共起スコアを取得することによって、結果として、拡張関係候補情報に含まれる２以上の言語表現と、その拡張関係候補情報に対応する共起言語表現とが共起する方が、その拡張関係候補情報に含まれる２以上の言語表現のみが共起するよりも高い値となる共起スコアを取得できることになる。

【０１１６】

また、本実施の形態では、類似情報記憶部１２で記憶されているすべての類似情報を用いて拡張関係候補情報を生成する場合について説明したが、そうでなくてもよい。すなわち、拡張関係候補情報生成部１３は、類似情報記憶部１２で記憶されている一部の類似情報を用いて、拡張関係候補情報の生成を行ってもよい。そのため、例えば、関係情報記憶部１１で記憶されている関係情報は、その関係情報が有する２以上の言語表現の意味的関係の種類を識別する情報である種類識別情報をも有するものであってもよい。また、類似情報記憶部１２では、種類識別情報と、その種類識別情報に対応する類似情報とが記憶されていてもよい。そして、拡張関係候補情報生成部１３は、関係情報記憶部１１で記憶されている関係情報に含まれる少なくとも１個の言語表現を置換する際に、その関係情報が有する種類識別情報に対応する類似情報を用いて置換を行ってもよい。

【０１１７】

具体的には、類似情報記憶部１２で、図１１で示される類似情報が記憶されていたとする。図１１において、類似情報と、種類識別情報とが対応付けられている。なお、図１１の類似情報は、図５の類似情報とは異なり、互いに類似する２個の言語表現のみを対応付ける情報である。そして、関係情報記憶部１１において、関係情報「名産＜愛媛、みかん＞」が記憶されていたとする。この関係情報は、愛媛の名産がみかんであることを示すものである。この関係情報を用いて拡張関係候補情報を生成する場合には、拡張関係候補情報生成部１３は、その関係情報から種類識別情報「名産」を取得し、その種類識別情報「名産」に対応付けられている類似情報を特定する。そして、拡張関係候補情報生成部１３は、その特定した類似情報を用いて、関係情報「名産＜愛媛、みかん＞」に含まれる各言語表現「愛媛」「みかん」の少なくとも１個を置換した拡張関係候補情報を生成する。具体的には、愛媛が香川に置換された拡張関係候補情報「名産＜香川、みかん＞」等が生成され、拡張関係候補情報記憶部１４に蓄積されることになる。なお、拡張関係候補情報には、種類識別情報が含まれていてもよく、あるいは、含まれていなくてもよい。なお、図１１で示される種類識別情報と、類似情報との対応は、手作業で生成されたものであってもよく、あるいは、その他の方法によって生成されてものであってもよい。

【０１１８】

なお、ここでは、種類識別情報に対応する類似情報を用いて拡張関係候補情報を生成する場合について説明したが、さらに、種類識別情報と、置換対象でない言語表現とに対応する類似情報を用いて、拡張関係候補情報を生成してもよい。その場合にも、関係情報は

種類識別情報を有するものであるとする。また、類似情報記憶部１２では、種類識別情報と、置換対象でない言語表現と、それらに対応する類似情報とが記憶されているものとする。そして、拡張関係候補情報生成部１３は、関係情報記憶部１１で記憶されている関係情報に含まれる１個の言語表現を置換する際に、その関係情報が有する種類識別情報と、その関係情報に含まれる置換対象ではない言語表現とに対応する類似情報を用いて置換を行うものとする。ここで、関係情報にＮ個（Ｎは２以上の整数）の言語表現が含まれている場合には、置換対象でない言語表現は、（Ｎ－１）個となる。したがって、その場合には、類似情報記憶部１２において、類似情報は、種類識別情報と、（Ｎ－１）個の置換対象でない言語表現とに対応付けられていることになる。例えば、Ｎ＝２の場合に、類似情報記憶部１２において、図１２で示される情報が記憶されていたとする。図１２において、種類識別情報と、１個の置換対象でない言語表現と、類似情報とが対応付けられている。また、図１２の類似情報は、図１１の場合と同様に、２個の言語表現を対応付ける類似情報である。そして、例えば、拡張関係候補情報生成部１３が、関係情報「原因＜心筋梗塞、死亡＞」を用いて拡張関係候補情報を生成する処理について説明する。その処理において、関係情報の第１の言語表現「心筋梗塞」の置換を行う場合には、拡張関係候補情報生成部１３は、その関係情報の種類識別情報「原因」と、置換対象でない言語表現「死亡」とを取得する。そして、図１２の情報を参照し、それらに対応する類似情報を特定する。そして、その特定した類似情報を用いて、第１の言語表現「心筋梗塞」を、「脳梗塞」等に置換した拡張関係候補情報「原因＜脳梗塞、死亡＞」「原因＜脳卒中、死亡＞」等を生成して、拡張関係候補情報記憶部１４に蓄積する。なお、図１２で示される種類識別情報と、類似情報との対応は、手作業で生成されたものであってもよく、あるいは、その他の方法によって生成されてものであってもよい。後者の場合には、例えば、置換対象でない言語表現を含む文書のみから、類似情報を生成してもよい。または、例えば、図１１の各レコードにおいて、置換対象でない言語表現と、類似情報に含まれるすべての言語表現との共起が高い場合に、そのレコードに、その置換対象でない言語表現を含めたレコードを作成し、図１２のレコードとしてもよい。例えば、図１１の１番目のレコードにおいて、置換対象でない言語表現「みかん」と、類似情報に含まれるすべての言語表現「愛媛」「香川」との共起が高いとすると、その図１１の１番目のレコードに置換対象でない言語表現「みかん」を追加したレコードを生成し、図１２の情報に追加してもよい。なお、図１２の情報は、拡張関係候補情報の生成時に、一時的に生成されて類似情報記憶部１２で記憶されるものであってもよい。例えば、種類識別情報が「原因」である場合に、置換対象でない言語表現が「死亡」であれば、それらに対応する類似情報を生成して類似情報記憶部１２に蓄積し、次に、置換対象でない言語表現が「急死」になれば、種類識別情報「原因」、置換対象でない言語表現「急死」に対応する類似情報を生成して類似情報記憶部１２に蓄積するようにしてもよい。その類似情報等の生成は、例えば、拡張関係候補情報生成部１３が行ってもよく、他の構成要素が行ってもよい。

【０１１９】

また、本実施の形態による関係情報拡張装置１は、拡張関係情報記憶部１７で記憶された拡張関係情報や、あるいは、拡張関係候補情報記憶部１４においてフラグ等によって拡張関係情報であることが示された拡張関係候補情報を出力する出力部を備えてもよい。その出力部による出力は、例えば、表示デバイス（例えば、ＣＲＴや液晶ディスプレイなど）への表示でもよく、所定の機器への通信回線を介した送信でもよく、プリンタによる印刷でもよく、記録媒体への蓄積でもよい。なお、その出力部は、出力を行うデバイス（例えば、表示デバイスやプリンタなど）を含んでもよく、あるいは含まなくてもよい。また、その出力部は、ハードウェアによって実現されてもよく、あるいは、それらのデバイスを駆動するドライバ等のソフトウェアによって実現されてもよい。

【０１２０】

また、本実施の形態による関係情報拡張装置１が処理を行う関係情報や拡張関係候補情報等に含まれる言語表現の言語は問わない。言語表現は、例えば、日本語や英語、ドイツ語、フランス語、ロシア語、中国語、スペイン語等で記述されたものであってもよい。た

だし、関係情報記憶部１１、類似情報記憶部１２、拡張関係候補情報記憶部１４、拡張関係情報記憶部１７で記憶される関係情報等の言語は、すべて共通しているものとする。

【０１２１】

［機械学習に関する説明］

ここで、上記実施の形態で用いられる機械学習について説明する。機械学習の手法は、問題－解の組のセットを多く用意し、そのセットを用いて学習を行なうことによって、どういう問題のときにどういう解になるかを学習し、その学習結果を利用して、新しい問題のときも解を推測できるようにする方法である。例えば、次の文献を参照されたい。

【０１２２】

文献：村田真樹、「機械学習に基づく言語処理」，龍谷大学理工学部．招待講演、２００４年（ｈｔｔｐ：／／ｗｗｗ２．ｎｉｃｔ．ｇｏ．ｊｐ／ｊｔ／ａ１３２／ｍｅｍｂｅｒｓ／ｍｕｒａｔａ／ｐｓ／ｒｋ１－ｓｉｒｙｏｕ．ｐｄｆ）

文献：村田真樹，馬青，内元清貴，井佐原均、「サポートベクトルマシンを用いたテンス・アスペクト・モダリティの日英翻訳」，電子情報通信学会言語理解とコミュニケーション研究会　ＮＬＣ２０００－７８，２００１年

文献：村田真樹，内山将夫，内元清貴，馬青，井佐原均、「ＮＳＥＶＡＬ２Ｊ辞書タスクでのＣＲＬの取り組み」、電子情報通信学会言語理解とコミュニケーション研究会　ＮＬＣ２００１－４０，２００１年

【０１２３】

機械学習アルゴリズムを動作させるために、問題の状況を機械に伝える際に、素性（解析に用いる情報で問題を構成する各要素）というものが必要になる。問題を素性によって表現するのである。例えば、日本語文末表現の時制の推定の問題において、問題：「彼が話す。」－－－解「現在」が与えられた場合に、素性の一例は、「彼が話す。」「が話す。」「話す。」「す」「。」となる。

【０１２４】

すなわち、機械学習の手法は、素性の集合－解の組のセットを多く用意し、そのセットを用いて学習を行なうことによって、どういう素性の集合のときにどういう解になるかを学習し、その学習結果を利用して、新しい問題のときも、その問題から素性の集合を取り出して、その素性に対応する解を推測する方法である。なお、ここで、「解」とは、例えば、前述の回答情報であるかどうかや、分類情報などである。

【０１２５】

機械学習の手法として、例えば、ｋ近傍法、シンプルベイズ法、決定リスト法、最大エントロピー法、サポートベクトルマシン法などの手法を用いることができる。なお、以下の説明では、文書を分類する場合（問題－解のセットが、文－分類である場合）の機械学習について主に説明するが、それ以外の機械学習についても、同様に適用可能であることは言うまでもない。

【０１２６】

ｋ近傍法は、最も類似する一つの事例のかわりに、最も類似するｋ個の事例を用いて、このｋ個の事例での多数決によって解（分類）を求める手法である。ｋは、あらかじめ定める整数の数字であって、一般的に、１から９の間の奇数を用いる。

【０１２７】

シンプルベイズ法は、ベイズの定理にもとづいて各解（分類）の確率を推定し、その確率値が最も大きい解を、求める解とする方法である。

【０１２８】

シンプルベイズ法において、文脈ｂで分類ａを出力する確率は、次式で与えられる。

【数６】

$$p(a|b)=\frac{p(a)}{p(b)}p(b|a)$$
$$\cong\frac{\tilde{p}(a)}{p(b)}\prod_i\tilde{p}(f_i|a)$$

【０１２９】

ただし、ここで文脈ｂは、あらかじめ設定しておいた素性ｆ$_j$（∈Ｆ，１≦ｊ≦ｋ）の集合である。ｐ（ｂ）は、文脈ｂの出現確率である。ここで、分類ａに非依存であって定数のために計算しない。Ｐ（ａ）（ここでＰはｐの上部にチルダ）とＰ（ｆ$_i$｜ａ）は、それぞれ教師データから推定された確率であって、分類ａの出現確率、分類ａのときに素性ｆ$_i$を持つ確率を意味する。Ｐ（ｆ$_i$｜ａ）として最尤推定を行って求めた値を用いると、しばしば値がゼロとなり、上記の２行目の式の値がゼロで分類先を決定することが困難な場合が生じる。そのため、スムージングを行う。ここでは、次式を用いてスムージングを行ったものを用いる。

【数７】

$$p(f_i|a)=\frac{freq(f_i,a)+0.01*freq(a)}{freq(a)+0.01*freq(a)}$$

【０１３０】

ただし、ｆｒｅｑ（ｆ$_i$，ａ）は、素性ｆ$_i$を持ち、かつ分類がａである事例の個数、ｆｒｅｑ（ａ）は、分類がａである事例の個数を意味する。

なお、スムージングは、上記式を用いた方法に限られるものではなく、その他の方法を用いてもよいことは言うまでもない。

【０１３１】

決定リスト法は、素性と分類先の組とを規則とし、それらをあらかじめ定めた優先順序でリストに蓄えおき、検出する対象となる入力が与えられたときに、リストで優先順位の高いところから入力のデータと規則の素性とを比較し、素性が一致した規則の分類先をその入力の分類先とする方法である。

【０１３２】

決定リスト方法では、あらかじめ設定しておいた素性ｆ$_j$（∈Ｆ，１≦ｊ≦ｋ）のうち、いずれか一つの素性のみを文脈として各分類の確率値を求める。ある文脈ｂで分類ａを出力する確率は、次式によって与えられる。

【数８】

$$p(a|b)=p(a|f_{\max})$$

【０１３３】

ただし、ｆ$_{max}$は、次式によって与えられる。

【数９】

$$f_{\max}=\arg\max_{f_j\in F}\max_{a_i\in A}\tilde{p}(a_i|f_j)$$

【０１３４】

また、Ｐ（ａ$_i$｜ｆ$_j$）（ここでＰはｐの上部にチルダ）は、素性ｆ$_j$を文脈に持つ場合の分類ａ$_i$の出現の割合である。

【０１３５】

最大エントロピー法は、あらかじめ設定しておいた素性ｆ$_j$（１≦ｊ≦ｋ）の集合をＦとするとき、以下の所定の条件式を満足しながらエントロピーを意味する式を最大にするときの確率分布ｐ（ａ，ｂ）を求め、その確率分布にしたがって求まる各分類の確率のうち、最も大きい確率値を持つ分類を求める分類先とする方法である。

【０１３６】
所定の条件式は、次式で与えられる。
【数１０】

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b)$$
$$for\ \forall f_j\ (1\le j\le k)$$

【０１３７】
また、エントロピーを意味する式は、次式で与えられる。
【数１１】

$$H(p) = -\sum_{a\in A,b\in B} p(a,b)\log(p(a,b))$$

【０１３８】
ただし、Ａ、Ｂは分類と文脈の集合を意味する。また、ｇ$_{j}$（ａ，ｂ）は文脈ｂに素性ｆ$_{j}$があって、なおかつ分類がａの場合１となり、それ以外で０となる関数を意味する。また、Ｐ（ａ$_{i}$｜ｆ$_{j}$）（ここでＰはｐの上部にチルダ）は、既知データでの（ａ，ｂ）の出現の割合を意味する。
【０１３９】
上記の条件式は、確率ｐと出力と素性の組の出現を意味する関数ｇをかけることで出力と素性の組の頻度の期待値を求めることになっており、右辺の既知データにおける期待値と、左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として、エントロピー最大化（確率分布の平滑化）を行なって、出力と文脈の確率分布を求めるものとなっている。最大エントロピー法の詳細については、以下の文献を参照されたい。
【０１４０】
文献：Ｅｒｉｃ Ｓｖｅｎ Ｒｉｓｔａｄ，「Ｍａｘｉｍｕｍ Ｅｎｔｒｏｐｙ Ｍｏｄｅｌｉｎｇ ｆｏｒ Ｎａｔｕｒａｌ Ｌａｎｇｕａｇｅ」，（ＡＣＬ／ＥＡＣＬ Ｔｕｔｏｒｉａｌ Ｐｒｏｇｒａｍ，Ｍａｄｒｉｄ，１９９７年
文献：Ｅｒｉｃ Ｓｖｅｎ Ｒｉｓｔａｄ，「Ｍａｘｉｍｕｍ Ｅｎｔｒｏｐｙ Ｍｏｄｅｌｉｎｇ Ｔｏｏｌｋｉｔ，Ｒｅｌｅａｓｅ１．６ｂｅｔａ」，（ｈｔｔｐ：／／ｗｗｗ．ｍｎｅｍｏｎｉｃ．ｃｏｍ／ｓｏｆｔｗａｒｅ／ｍｅｍｔ），１９９８年
【０１４１】
サポートベクトルマシン法は、空間を超平面で分割することにより、二つの分類からなるデータを分類する手法である。
【０１４２】
図１３にサポートベクトルマシン法のマージン最大化の概念を示す。図１３において、白丸は正例、黒丸は負例を意味し、実線は空間を分割する超平面を意味し、破線はマージン領域の境界を表す面を意味する。図１３（Ａ）は、正例と負例の間隔が狭い場合（スモールマージン）の概念図、図１３（Ｂ）は、正例と負例の間隔が広い場合（ラージマージン）の概念図である。
【０１４３】
このとき、二つの分類が正例と負例からなるものとすると、学習データにおける正例と負例の間隔（マージン）が大きいものほどオープンデータで誤った分類をする可能性が低いと考えられ、図１３（Ｂ）に示すように、このマージンを最大にする超平面を求めそれを用いて分類を行なう。
【０１４４】
基本的には上記のとおりであるが、通常、学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や、超平面の線形の部分を非線型にする拡張（カーネル関数の導入）がなされたものが用いられる。
【０１４５】

この拡張された方法は、以下の識別関数（ｆ（ｘ））を用いて分類することと等価であり、その識別関数の出力値が正か負かによって二つの分類を判別することができる。

【数１２】

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(x_i,x)+b\right) \qquad \text{(M1)}$$

$$b=-\frac{\max_{i,y_i=-1} b_i+\min_{i,y_i=1} b_i}{2}$$

$$b_i=\sum_{j=1}^{l}\alpha_j y_j K(x_j,x_i)$$

【０１４６】

ただし、ｘは識別したい事例の文脈（素性の集合）を、x_iとy_j（ｉ＝１，…，ｌ，$y_j\in\{1,-1\}$）は学習データの文脈と分類先を意味し、関数ｓｇｎは、

ｓｇｎ（ｘ）＝１（ｘ≧０）

－１（ｏｔｈｅｒｗｉｓｅ）

であり、また、各α_iは、式（Ｍ３）と式（Ｍ４）の制約のもと、式（Ｍ２）を最大にする場合のものである。

【数１３】

$$L(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \qquad \text{(M2)}$$

$$0\le\alpha_i\le C \qquad (i=1,\ldots,l) \qquad \text{(M3)}$$

$$\sum_{i=1}^{l}\alpha_i y_i=0 \qquad \text{(M4)}$$

【０１４７】

また、関数Ｋはカーネル関数と呼ばれ、様々なものが用いられるが、本形態では、例えば、以下の多項式のものを用いる。

Ｋ（ｘ，ｙ）＝（ｘ・ｙ＋１）ｄ

【０１４８】

ここで、Ｃ、ｄは実験的に設定される定数である。例えば、Ｃはすべての処理を通して１に固定した。また、ｄは、１と２の二種類を試している。ここで、$\alpha_i>0$となるx_iは、サポートベクトルと呼ばれ、通常、式（Ｍ１）の和をとっている部分は、この事例のみを用いて計算される。つまり、実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない。

【０１４９】

なお、拡張されたサポートベクトルマシン法の詳細については、次の文献を参照されたい。

文献：Ｎｅｌｌｏ Ｃｒｉｓｔｉａｎｉｎｉ，Ｊｏｈｎ Ｓｈａｗｅ－Ｔａｙｌｏｒ，「Ａｎ Ｉｎｔｒｏｄｕｃｔｉｏｎ ｔｏ Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ Ｍａｃｈｉｎｅｓ ａｎｄ ｏｔｈｅｒ ｋｅｒｎｅｌ－ｂａｓｅｄ ｌｅａｒｎｉｎｇ ｍｅｔｈｏｄｓ」，Ｃａｍｂｒｉｄｇｅ Ｕｎｉｖｅｒｓｉｔｙ Ｐｒｅｓｓ，２０００年

文献：Ｔａｋｕ Ｋｕｄｏｈ，「Ｔｉｎｙｓｖｍ：Ｓｕｐｐｏｒｔ Ｖｅｃｔｏｒ ｍａｃｈｉｎｅｓ」，（ｈｔｔｐ：／／ｃｌ．ａｉｓｔｎａｒａ．ａｃ．ｊｐ／ｔａｋｕ－ｋｕ／／ｓｏｆｔｗａｒｅ／Ｔｉｎｙ ＳＶＭ／ｉｎｄｅｘ．ｈｔｍｌ），２０００年

【０１５０】

サポートベクトルマシン法は、分類の数が２個のデータを扱うものである。したがって

、分類の数が３個以上の事例を扱う場合には、通常、これにペアワイズ法またはワンＶＳレスト法などの手法を組み合わせて用いることになる。

【０１５１】

ペアワイズ法は、ｎ個の分類を持つデータの場合に、異なる二つの分類先のあらゆるペア（ｎ（ｎ－１）／２個）を生成し、ペアごとにどちらがよいかを二値分類器、すなわちサポートベクトルマシン法処理モジュールで求めて、最終的に、ｎ（ｎ－１）／２個の二値分類による分類先の多数決によって、分類先を求める方法である。

【０１５２】

ワンＶＳレスト法は、例えば、ａ、ｂ、ｃという三つの分類先があるときは、分類先ａとその他、分類先ｂとその他、分類先ｃとその他、という三つの組を生成し、それぞれの組についてサポートベクトルマシン法で学習処理する。そして、学習結果による推定処理において、その三つの組のサポートベクトルマシンの学習結果を利用する。推定するべき問題が、その三つのサポートベクトルマシンではどのように推定されるかを見て、その三つのサポートベクトルマシンのうち、その他でないほうの分類先であって、かつサポートベクトルマシンの分離平面から最も離れた場合のものの分類先を求める解とする方法である。例えば、ある解くべき問題が、「分類先ａとその他」の組の学習処理で作成したサポートベクトルマシンにおいて分離平面から最も離れた場合には、その解くべき問題の分類先は、ａと推定する。

【０１５３】

図示しない解推定手段が推定する、解くべき問題についての、どのような解（分類先）になりやすいかの度合いの求め方は、図示しない機械学習手段が機械学習の手法として用いる様々な方法によって異なる。

【０１５４】

例えば、機械学習手段が、機械学習の手法としてｋ近傍法を用いる場合、機械学習手段は、教師データの事例同士で、その事例から抽出された素性の集合のうち重複する素性の割合（同じ素性をいくつ持っているかの割合）にもとづく事例同士の類似度を定義して、前記定義した類似度と事例とを学習結果情報として学習結果記憶手段に記憶しておく。

【０１５５】

そして、解推定手段は、解くべき問題の素性（文章群属性情報）が抽出されたときに、学習結果記憶手段において定義された類似度と事例を参照して、抽出された解くべき問題の素性について、その解くべき問題の素性の類似度が高い順にｋ個の事例を学習結果記憶手段の事例から選択し、選択したｋ個の事例での多数決によって決まった分類先を、解くべき問題の分類先（解）として推定する。すなわち、解推定手段では、解くべき問題についての、どのような解（分類先）になりやすいかの度合いを、選択したｋ個の事例での多数決の票数とする。

【０１５６】

また、機械学習手法として、シンプルベイズ法を用いる場合には、図示しない機械学習手段は、教師データの事例について、前記事例の解と素性の集合との組を学習結果情報として学習結果記憶手段に記憶する。そして、解推定手段は、解くべき問題の素性が抽出されたときに、学習結果記憶手段の学習結果情報の解と素性の集合との組をもとに、ベイズの定理にもとづいて、解くべき問題の素性の集合の場合の各分類になる確率を算出して、その確率の値が最も大きい分類を、その解くべき問題の素性の分類（解）と推定する。すなわち、解推定手段では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、各分類になる確率とする。

【０１５７】

また、機械学習手法として決定リスト法を用いる場合には、図示しない機械学習手段は、教師データの事例について、素性と分類先との規則を所定の優先順序で並べたリストを、予め、何らかの手段により、学習結果記憶手段に記憶させる。そして、解くべき問題の素性が抽出されたときに、解推定手段は、学習結果記憶手段のリストの優先順位の高い順に、抽出された解くべき問題の素性と規則の素性とを比較し、素性が一致した規則の分類

先をその解くべき問題の分類先（解）として推定する。

【０１５８】

また、機械学習手法として最大エントロピー法を使用する場合には、図示しない機械学習手段は、教師データの事例から解となりうる分類を特定し、所定の条件式を満足し、かつエントロピーを示す式を最大にするときの素性の集合と解となりうる分類の二項からなる確率分布を求めて、学習結果記憶手段に記憶する。そして、解くべき問題の素性が抽出されたときに、解推定手段は、学習結果記憶手段の確率分布を利用して、抽出された解くべき問題の素性の集合についてその解となりうる分類の確率を求めて、最も大きい確率値を持つ解となりうる分類を特定し、その特定した分類をその解くべき問題の解と推定する。すなわち、解推定手段では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、各分類になる確率とする。

【０１５９】

また、機械学習手法としてサポートベクトルマシン法を使用する場合には、図示しない機械学習手段は、教師データの事例から解となりうる分類を特定し、分類を正例と負例に分割して、カーネル関数を用いた所定の実行関数にしたがって事例の素性の集合を次元とする空間上で、その事例の正例と負例の間隔を最大にし、かつ正例と負例を超平面で分割する超平面を求めて学習結果記憶手段に記憶する。そして、解くべき問題の素性が抽出されたときに、解推定手段は、学習結果記憶手段の超平面を利用して、解くべき問題の素性の集合が超平面で分割された空間において正例側か負例側のどちらにあるかを特定し、その特定された結果にもとづいて定まる分類を、その解くべき問題の解と推定する。すなわち、解推定手段では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、分離平面からのその解くべき問題の事例への距離の大きさとする。

【０１６０】

また、上記各実施の形態において、ある構成要素が機械学習を用いて処理を行う場合に、その所望の処理が実行されるまでに学習が行われるのであれば、その学習のタイミングは問わない。

【０１６１】

また、上記実施の形態では、関係情報拡張装置１がスタンドアロンである場合について説明したが、関係情報拡張装置１は、スタンドアロンの装置であってもよく、サーバ・クライアントシステムにおけるサーバ装置であってもよい。後者の場合には、拡張関係情報等が、通信回線を介して出力されてもよい。

【０１６２】

また、上記実施の形態において、各処理または各機能は、単一の装置または単一のシステムによって集中処理されることによって実現されてもよく、あるいは、複数の装置または複数のシステムによって分散処理されることによって実現されてもよい。

【０１６３】

また、上記実施の形態において、各構成要素が実行する処理に関係する情報、例えば、各構成要素が受け付けたり、取得したり、選択したり、生成したり、送信したり、受信したりした情報や、各構成要素が処理で用いるしきい値や数式、アドレス等の情報等は、上記説明で明記していない場合であっても、図示しない記録媒体において、一時的に、あるいは長期にわたって保持されていてもよい。また、その図示しない記録媒体への情報の蓄積を、各構成要素、あるいは、図示しない蓄積部が行ってもよい。また、その図示しない記録媒体からの情報の読み出しを、各構成要素、あるいは、図示しない読み出し部が行ってもよい。

【０１６４】

また、上記実施の形態において、各構成要素等で用いられる情報、例えば、各構成要素が処理で用いるしきい値やアドレス、各種の設定値等の情報がユーザによって変更されてもよい場合には、上記説明で明記していない場合であっても、ユーザが適宜、それらの情報を変更できるようにしてもよく、あるいは、そうでなくてもよい。それらの情報をユーザが変更可能な場合には、その変更は、例えば、ユーザからの変更指示を受け付ける図示

しない受付部と、その変更指示に応じて情報を変更する図示しない変更部とによって実現されてもよい。その図示しない受付部による変更指示の受け付けは、例えば、入力デバイスからの受け付けでもよく、通信回線を介して送信された情報の受信でもよく、所定の記録媒体から読み出された情報の受け付けでもよい。

【０１６５】

また、上記実施の形態において、関係情報拡張装置１に含まれる２以上の構成要素が通信デバイスや入力デバイス等を有する場合に、２以上の構成要素が物理的に単一のデバイスを有してもよく、あるいは、別々のデバイスを有してもよい。

【０１６６】

また、上記実施の形態において、各構成要素は専用のハードウェアにより構成されてもよく、あるいは、ソフトウェアにより実現可能な構成要素については、プログラムを実行することによって実現されてもよい。例えば、ハードディスクや半導体メモリ等の記録媒体に記録されたソフトウェア・プログラムをＣＰＵ等のプログラム実行部が読み出して実行することによって、各構成要素が実現され得る。なお、上記実施の形態における関係情報拡張装置１を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、コンピュータを、意味的関係を有する２以上の言語表現を有する関係情報が記憶される関係情報記憶部で記憶されている関係情報に含まれる少なくとも１個の言語表現を、類似する２以上の言語表現を有する類似情報が２以上記憶される類似情報記憶部で記憶されている類似情報を用いて、言語表現に類似する言語表現に置換した拡張関係候補情報を生成し、拡張関係候補情報を、拡張関係候補情報が記憶される拡張関係候補情報記憶部に蓄積する拡張関係候補情報生成部、拡張関係候補情報記憶部で記憶されている拡張関係候補情報が意味的関係を有する確からしさを示すスコアを取得するスコア取得部、スコア取得部が取得したスコアを用いて、拡張関係候補情報記憶部で記憶されている拡張関係候補情報のうち、スコアの高い拡張関係候補情報である拡張関係情報を選択する選択部として機能させるためのプログラムである。

【０１６７】

なお、上記プログラムにおいて、情報を送信する送信ステップや、情報を受信する受信ステップなどでは、ハードウェアでしか行われない処理、例えば、送信ステップにおけるモデムやインターフェースカードなどで行われる処理は少なくとも含まれない。

【０１６８】

また、このプログラムは、サーバなどからダウンロードされることによって実行されてもよく、所定の記録媒体（例えば、ＣＤ－ＲＯＭなどの光ディスクや磁気ディスク、半導体メモリなど）に記録されたプログラムが読み出されることによって実行されてもよい。また、このプログラムは、プログラムプロダクトを構成するプログラムとして用いられてもよい。

【０１６９】

また、このプログラムを実行するコンピュータは、単数であってもよく、複数であってもよい。すなわち、集中処理を行ってもよく、あるいは分散処理を行ってもよい。

【０１７０】

図１４は、上記プログラムを実行して、上記実施の形態による関係情報拡張装置１を実現するコンピュータの外観の一例を示す模式図である。上記実施の形態は、コンピュータハードウェア及びその上で実行されるコンピュータプログラムによって実現されうる。

【０１７１】

図１４において、コンピュータシステム９００は、ＣＤ－ＲＯＭ（Ｃｏｍｐａｃｔ Ｄｉｓｋ Ｒｅａｄ Ｏｎｌｙ Ｍｅｍｏｒｙ）ドライブ９０５、ＦＤ（Ｆｌｏｐｐｙ（登録商標） Ｄｉｓｋ）ドライブ９０６を含むコンピュータ９０１と、キーボード９０２と、マウス９０３と、モニタ９０４とを備える。

【０１７２】

図１５は、コンピュータシステム９００の内部構成を示す図である。図１５において、コンピュータ９０１は、ＣＤ－ＲＯＭドライブ９０５、ＦＤドライブ９０６に加えて、Ｍ

ＰＵ（Ｍｉｃｒｏ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）９１１と、ブートアッププログラム等のプログラムを記憶するためのＲＯＭ９１２と、ＭＰＵ９１１に接続され、アプリケーションプログラムの命令を一時的に記憶すると共に、一時記憶空間を提供するＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）９１３と、アプリケーションプログラム、システムプログラム、及びデータを記憶するハードディスク９１４と、ＭＰＵ９１１、ＲＯＭ９１２等を相互に接続するバス９１５とを備える。なお、コンピュータ９０１は、ＬＡＮへの接続を提供する図示しないネットワークカードを含んでいてもよい。

【０１７３】

コンピュータシステム９００に、上記実施の形態による関係情報拡張装置１の機能を実行させるプログラムは、ＣＤ－ＲＯＭ９２１、またはＦＤ９２２に記憶されて、ＣＤ－ＲＯＭドライブ９０５、またはＦＤドライブ９０６に挿入され、ハードディスク９１４に転送されてもよい。これに代えて、そのプログラムは、図示しないネットワークを介してコンピュータ９０１に送信され、ハードディスク９１４に記憶されてもよい。プログラムは実行の際にＲＡＭ９１３にロードされる。なお、プログラムは、ＣＤ－ＲＯＭ９２１やＦＤ９２２、またはネットワークから直接、ロードされてもよい。

【０１７４】

プログラムは、コンピュータ９０１に、上記実施の形態による関係情報拡張装置１の機能を実行させるオペレーティングシステム（ＯＳ）、またはサードパーティプログラム等を必ずしも含んでいなくてもよい。プログラムは、制御された態様で適切な機能（モジュール）を呼び出し、所望の結果が得られるようにする命令の部分のみを含んでいてもよい。コンピュータシステム９００がどのように動作するのかについては周知であり、詳細な説明は省略する。

【０１７５】

また、本発明は、以上の実施の形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。

【産業上の利用可能性】

【０１７６】

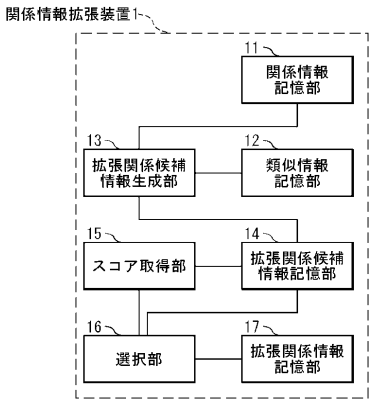
以上より、本発明による関係情報拡張装置等によれば、関係情報を適切に拡張することができるという効果が得られ、新たな関係を取得する装置等として有用である。

【符号の説明】

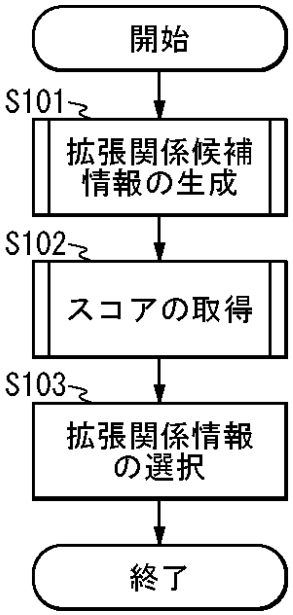
【０１７７】

１　関係情報拡張装置
１１　関係情報記憶部
１２　類似情報記憶部
１３　拡張関係候補情報生成部
１４　拡張関係候補情報記憶部
１５　スコア取得部
１６　選択部
１７　拡張関係情報記憶部
２１　対応情報記憶部

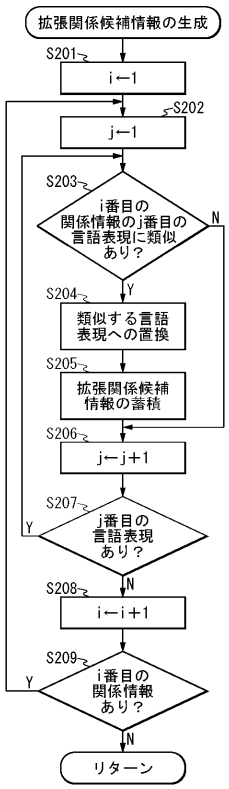
【図１】


関係情報拡張装置1
11
関係情報記憶部
13
拡張関係候補情報生成部
12
類似情報記憶部
15
スコア取得部
14
拡張関係候補情報記憶部
16
選択部
17
拡張関係情報記憶部

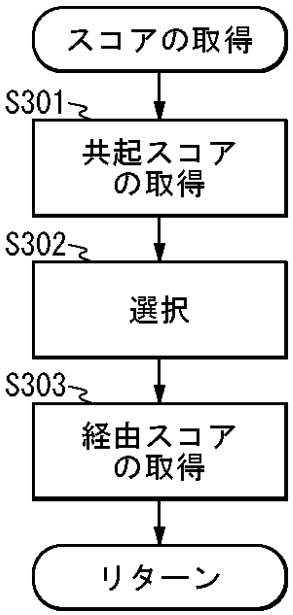
【図２】


開始
S101
拡張関係候補情報の生成
S102
スコアの取得
S103
拡張関係情報の選択
終了

【図３】


拡張関係候補情報の生成
S201
i←1
S202
j←1
S203
i番目の関係情報のj番目の言語表現に類似あり？
N
Y
S204
類似する言語表現への置換
S205
拡張関係候補情報の蓄積
S206
j←j+1
S207
j番目の言語表現あり？
Y
N
S208
i←i+1
S209
i番目の関係情報あり？
Y
N
リターン

【図４】


スコアの取得
S301
共起スコアの取得
S302
選択
S303
経由スコアの取得
リターン

【図５】

類似情報
心筋梗塞,脳梗塞,脳卒中,うつ病
死亡,病死,急死
⋮

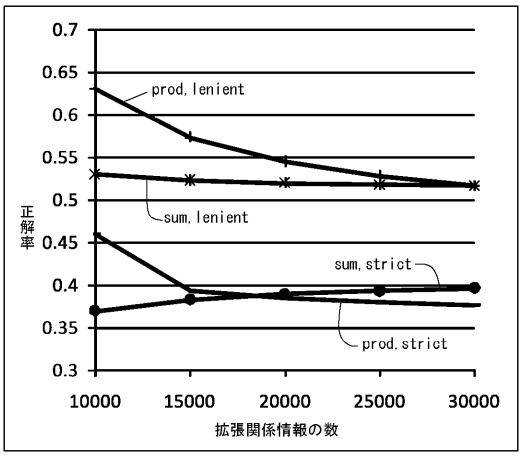
【図６】

関係情報ＩＤ	拡張関係候補情報	共起スコア
1	＜脳梗塞,死亡＞	—
1	＜脳卒中,死亡＞	—
1	＜うつ病,死亡＞	—
1	＜心筋梗塞,病死＞	—
1	＜心筋梗塞,急死＞	—
1	＜脳梗塞,病死＞	—
1	＜脳梗塞,急死＞	—
1	＜脳卒中,病死＞	—
1	＜脳卒中,急死＞	—
1	＜うつ病,病死＞	—
1	＜うつ病,急死＞	—

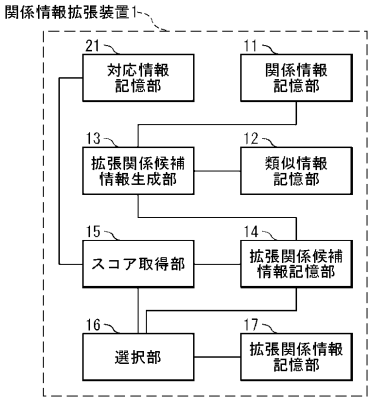
【図７】

関係情報ＩＤ	拡張関係候補情報	共起スコア
1	＜脳梗塞,死亡＞	540
1	＜脳卒中,死亡＞	190
1	＜うつ病,死亡＞	280
1	＜心筋梗塞,病死＞	220
1	＜心筋梗塞,急死＞	360
1	＜脳梗塞,病死＞	410
1	＜脳梗塞,急死＞	470
1	＜脳卒中,病死＞	120
1	＜脳卒中,急死＞	150
1	＜うつ病,病死＞	20
1	＜うつ病,急死＞	16

【図８】



【図９】



【図１０】

対応情報

種類識別情報	共起言語表現
原因	原因,理由,要因,・・・
食材	材料,レシピ,・・・
⋮	⋮

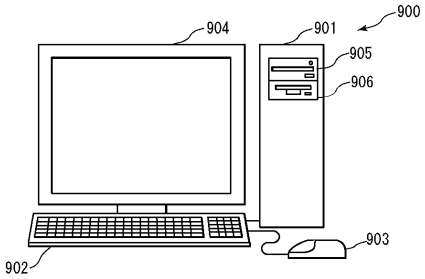
【図１１】

種類識別情報	類似情報
名産	愛媛,香川
名産	香川,徳島
名産	徳島,高知
⋮	⋮
原因	死亡,急死
⋮	⋮

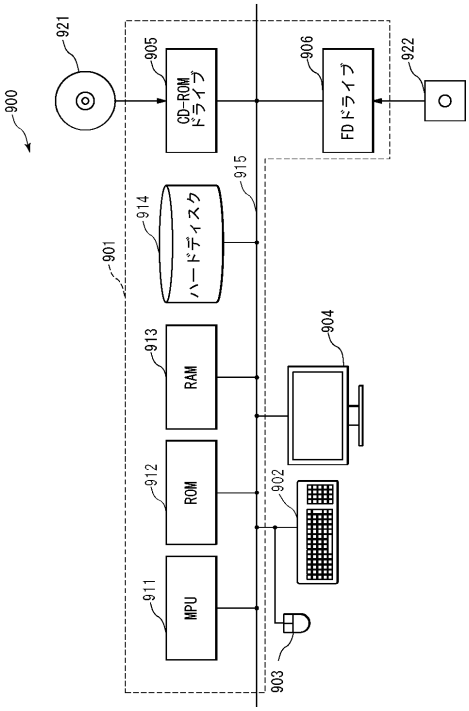
【図１２】

種類識別情報	置換対象でない言語表現	類似情報
原因	死亡	心筋梗塞,脳梗塞
原因	死亡	心筋梗塞,脳卒中
⋮	⋮	⋮
名産	みかん	愛媛,香川
⋮	⋮	⋮

【図１４】

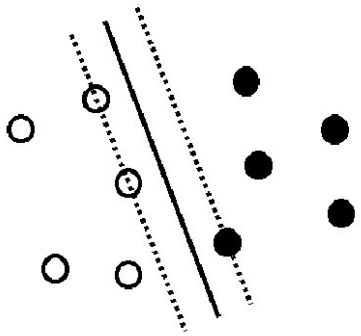


【図１５】

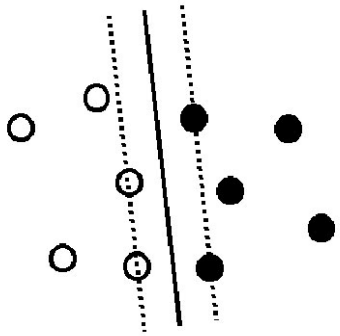


【図１３】

（Ａ）スモールマージン



（Ｂ）ラージマージン



フロントページの続き

(72)発明者　鳥澤　健太郎
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　風間　淳一
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　黒田　航
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

Ｆターム(参考) 5B075 ND03　NK35
5B091 AA15　AB17　CA12　CC05　CC16
5E501 AA01　AC01　AC34　DA07　EA31　FA47