

明 細 書

発明の名称：

テキスト情報監視用辞書作成装置、テキスト情報監視用辞書作成方法、及び、テキスト情報監視用辞書作成プログラム

技術分野

[0001] 本発明は、テキスト情報監視用辞書作成装置、テキスト情報監視用辞書作成方法、及び、テキスト情報監視用辞書作成プログラムに関し、特に未知のテキストに対しても高精度なテキスト情報監視用辞書を作成するテキスト情報監視用辞書作成装置、テキスト情報監視用辞書作成方法、及び、テキスト情報監視用辞書作成プログラムに関する。

背景技術

[0002] インターネット上の風評の監視等で大量のテキストの中から監視対象の情報内容の出現を検知するテキスト情報監視技術が重要となっている。本発明で想定するテキスト情報監視システムは辞書ベースでテキスト情報監視を行う。つまり、テキスト情報監視技術の一手法として、検知のための条件をテキスト情報監視用辞書として保持しておき、入力文書中の表現がテキスト情報監視用辞書中の条件と一致するか否かで検知を行う、辞書ベースの手法を用いる。

[0003] 辞書ベースの手法では、高精度の辞書を用いることで、高精度のテキスト情報監視ができる。したがって、高精度の辞書を用いることが重要である。

[0004] 辞書ベースのテキスト情報監視システムにおいて辞書を内省で作るのは時間がかかり、漏れが発生しやすくなるため、困難である。そこで、監視対象の情報内容を含む文書を集めた正例集合と監視対象の情報内容を含まない文書を集めた負例集合を与え、そこから検知条件として登録すべき表現を自動で抽出する手法が望まれる。そうした方法の従来手法として、特徴語抽出手法がある。この特徴語抽出手法とは、正例集合、負例集合を比較し、正例集合に特徴的に出現する単語を特徴語として抽出する手法のことである。

[0005] そのような手法の一例として、特許文献1がある。特許文献1では、テキストマイニングで用いる辞書を構築する際に、分析対象の文書データをグループに分け、各グループに特徴的に出現する表現を辞書候補として用いている。

先行技術文献

特許文献

[0006] 特許文献1：特開2009-015394号公報

発明の概要

発明が解決しようとする課題

[0007] しかし、従来技術の単語レベルや係り受けレベルの短い単位の特徴語抽出手法は、テキスト情報監視システムの性能要件を十分満たすことができない。なぜなら、単語レベルや係り受けレベルの短い単位だけでは、検知の精度が低くなるからである。例えば、コンピュータウィルスに関する記述を検知したい時に、「ウィルス」という1単語をテキスト情報監視用辞書に登録したとしても「風邪のウィルス」のような文書が誤って検知されてしまう。この場合は、「コンピュータ・ウィルス」や「ウィルス・メール」といった1個以上の単語からなるフレーズをテキスト情報監視用辞書に登録する必要がある。

[0008] このように最適なフレーズの長さは何を検知したいかによって変わってくるため、これを事前に一意の値として決めておくことはできない。そこで、可変長のフレーズに対応するために、あらゆる長さのフレーズを候補として抽出し、それぞれに特徴度を計算する必要がある。さらに、互いに重複のある複数のフレーズが同じ特徴度で出力される場合を適切に扱えない。

[0009] 例えば、図3のような正例集合、負例集合が与えられた場合、様々な長さのフレーズを対象に特徴語抽出を行うと、図4のようなフレーズが抽出され、「トロイの木馬」、「トロイ」、「木馬」が同じ特徴度（＝3）として抽出される。しかし、「トロイ」や「木馬」は、この負例集合では出現してい

なかったが、「トロイ遺跡」や「回転木馬」といったウィルスとは関係のない表現が考えられるため、「トロイ」や「木馬」をテキスト情報監視用辞書に登録するのは検知精度を下げることになる。原理的には負例集合には「トロイ遺跡」や「回転木馬」といった表現が出現することで、「トロイ」や「木馬」といった表現の特徴度を小さくし検知精度を下げることも可能であるが、実際には、十分な量の負例集合を得られることは少なく、上記のような問題が頻繁に発生する。

[0010] 特許文献1では、特徴語と共起する単語も含めて辞書登録候補とする手法を開示しているが、辞書登録するか否かの判定は、TF (Term Frequency) と IDF (Inverse Document Frequency) との積といった指標を用いており、互いに重複のある複数のフレーズに対しては上記と同様の課題があると考えられる。

[0011] 以上のように、正例集合、負例集合から計算される特徴度でテキスト情報監視用辞書を構築する従来手法は検知精度が低くなるという課題がある。

[0012] 本発明は上記課題を解決するものであり、従来技術に比べて高精度な検知を実現可能にするテキスト情報監視用辞書作成装置、テキスト情報監視用辞書作成方法、及び、テキスト情報監視用辞書作成プログラムを提供することを目的とする。

課題を解決するための手段

[0013] 上記課題を解決する本発明は、テキスト情報監視システムで用いられ、検知条件が登録される辞書を作成するテキスト情報監視用辞書作成装置であって、検知条件候補のフレーズに対して、フレーズが監視対象の情報内容に適合する度合いを表す特徴度を計算する特徴度計算部と、前記特徴度とフレーズによって規定される意味の曖昧さの少なさを表す有用度とに基づいて、フレーズが検知条件として適切であるか否かを判定するフレーズ有用性判定部とを備える。

[0014] 上記課題を解決する本発明は、テキスト情報監視システムで用いられる辞書の作成方法であって、テキスト情報監視用辞書作成装置が、検知条件候補

のフレーズに対して、フレーズが監視対象の情報内容に適合する度合いを表す特徴度を計算し、前記特徴度とフレーズによって規定される意味の曖昧さの少なさを表す有用度とに基づいて、フレーズが検知条件として適切であるか否かを判定し、適切であると判断したフレーズを出力し検知条件として登録する。

[0015] 上記課題を解決する本発明は、検知条件候補のフレーズに対して、フレーズが監視対象の情報内容に適合する度合いを表す特徴度を計算する処理と、前記特徴度とフレーズによって規定される意味の曖昧さの少なさを表す有用度とに基づいて、フレーズが検知条件として適切であるか否かを判定する処理と、適切であると判断したフレーズを出力し検知条件として登録する処理とをテキスト情報監視用辞書作成装置に実行させるテキスト情報監視用辞書作成プログラムである。

発明の効果

[0016] 一般に、フレーズの長さが長いほど意味の曖昧性が少なくなり、検知条件としての適合率は高くなる。本発明では、フレーズの長さに基づき有用度を計算し、有用度と特徴度とに基づいて辞書登録すべきフレーズの抽出を行う。すなわち、長さの長いフレーズを優先する。

[0017] これにより、従来技術に比べて高精度な検知を実現可能にするテキスト情報監視用辞書を作成することができる。

図面の簡単な説明

[0018] [図1]辞書作成装置の機能ブロック図

[図2]辞書作成装置の動作フロー

[図3]正例集合、負例集合の例（従来技術と共通）

[図4]各フレーズの頻度と特徴度の例（従来技術と共通）

[図5]各フレーズの有用度とスコアの例（適用例1）

[図6]各フレーズの有用度とスコアの例（適用例2）

[図7]各フレーズの有用度とスコアの例（適用例3）

[図8]各フレーズの有用度とスコアの例（適用例4）

[図9]各フレーズの有用度とスコアの例（適用例5）

発明を実施するための形態

[0019] ～構成・動作～

次に、本発明の実施の形態の構成及び動作について図面を参照して詳細に説明する。

[0020] 図1は、本実施形態に係る辞書作成装置の機能ブロック図である。本実施形態に係る辞書作成装置は、フレーズ抽出部1と、フレーズ有用性判定部2と、特徴度計算部3と、出力部4から構成される。また、フレーズ有用性判定部2は、有用度計算部21と検知条件判定部22から構成される。

[0021] 各構成の機能について説明する。

[0022] 前提として、監視対象の情報内容を含む文書を集めた正例集合と、監視対象の情報内容を含まない文書を集めた負例集合とが与えられているものとする（図3参照）。

[0023] フレーズ抽出部1は、与えられた正例集合中のテキストに対して言語解析を行い、様々な長さのフレーズを検知条件候補として抽出する。フレーズの抽出は、形態素解析を行い、特定の品詞タグ列となるフレーズを抽出したり、構文解析を行い得られた構文木の部分木をフレーズとしたり、それらの組み合わせを用いて行う。

[0024] フレーズ有用性判定部2は、フレーズ抽出部1で抽出された各フレーズに対して有用度を計算し、さらに、有用度と特徴度計算部3で計算した特徴度とを組み合わせ、そのフレーズが検知条件として適切かどうかを判定する。

[0025] 有用度計算部21は、フレーズ抽出部1で抽出された各フレーズに対して、フレーズの長さ、フレーズの正例集合中の頻度、フレーズ間の包含関係に関する指標を用いて有用度を計算する。ここで、フレーズの有用度とは、そのフレーズによって規定される意味の曖昧さの少なさを表す値のことで、そのフレーズを検知条件としたときの検知精度の良さを表す値のことである。有用度は、フレーズの長さ又はその対数値を用いてもよいし、フレーズの長

さ又はその対数値と正例集合中のフレーズの出現数又はその対数値の積を用いてもよい。もしくは、更にフレーズ間の包含関係に関する指標に基づいて、非特許文献1で提案されているようなC-valueを有用度として用いてもよい。

非特許文献1：Frantzi, K. and Ananiadou, S. (1996). "Extracting Nested Collocations." In Proceedings of the 16th International Conference on Computational Linguistics (COLING 96), pp.41-46.

[0026] 有用度計算の適用例については後述する（適用例1～4）。

[0027] 検知条件判定部22は、各フレーズに対して、有用度計算部21で計算した有用度と、特徴度計算部3で計算した特徴度とを用いて、そのフレーズが検知条件として適切であるか否かを判定する。例えば、有用度と特徴度の積によって検知条件として適切さを評価し、その値が閾値よりも大きい場合に検知条件として適切であると判定する。また、ここで有用度が閾値よりも小さいフレーズを除外し、特徴度計算するフレーズを減らして計算量を少なく抑えることも可能である（適用例5）。

[0028] 特徴度計算部3は、正例集合と負例集合の統計量を比較し、着目するフレーズが正例集合に出現する度合いを特徴度として計算する。特徴度は、カイ2乗値、相互情報量、ESC (Extended Stochastic Complexity) のようなテキストマイニングで使用される既存の尺度を用いて計算する。ここでの特徴度計算は、フレーズ抽出部1で抽出されたフレーズ全てに対して行ってもよいし、フレーズ有用性判定部2での判定に必要なフレーズに対してのみ行ってもよい。

[0029] 出力部4は、フレーズ有用性判定部2で検知条件として適切だと判定されたフレーズを辞書登録するフレーズとして出力する。出力部4は、辞書登録するフレーズのみを出力するだけでなく、フレーズと有用度、特徴度、検知条件としての適切さを表すスコア等を合わせて出力することで、スコア等を参照しながら人手を用いて辞書登録するフレーズを選別することでテキスト

情報監視用辞書構築作業を軽減することも可能となる。

[0030] 図2は、辞書作成装置の動作フローである。辞書作成プログラムは、動作フローの各処理を辞書作成装置に実行させるものである。プログラムが実行されると、フレーズ抽出部1、フレーズ有用性判定部2、特徴度計算部3、出力部4が機能する。

[0031] まず、フレーズ抽出部1は、与えられた正例集合中のテキストに対して言語解析を行い、様々な長さのフレーズを検知条件候補として抽出する（ステップS1）。

[0032] 次に、有用度計算部21は、フレーズ抽出部1で抽出された各フレーズに対して、有用度を計算する（ステップS2）。

[0033] 一方、特徴度計算部3は、着目するフレーズの特徴度を計算する（ステップS3）。

[0034] 次に、検知条件判定部22は、各フレーズに対して、有用度計算部21が計算した有用度と、特徴度計算部3が計算した特徴度とを用いて、そのフレーズが検知条件として適切であるか否かを判定する（ステップS4）。例えば有用度と特徴度とに基づいてスコアを計算し、スコアに基づいて判定する。

[0035] 最後に、出力部4は、辞書登録するフレーズを出力し（ステップS5）、処理を終える。

[0036] なお、ステップS2とステップ3とはどちらが先でもよいし同時でもよい。

[0037] また、ステップS3およびステップS4において、有用度が閾値以上のフレーズのみ、特徴度を計算し、検知条件として適切であるか否かを判定してもよい。

[0038] ～従来技術の具体例～

従来技術に係る辞書作成装置は、フレーズ抽出部1と、特徴度計算部3と、出力部4から構成される（図示省略）。すなわち、フレーズ有用性判定部2の有無以外は、本願実施形態と共通する。

[0039] 本発明で想定するテキスト情報監視システムは、テキスト情報監視用辞書との文字列一致によってテキスト情報監視をおこない、テキスト情報監視用辞書には検知条件として文字列を登録するものとする。ただし、本発明の対象となるテキスト情報監視システムは上記システムに限定されるわけではなく、品詞タグや構文構造を条件としてテキスト情報監視を行うシステムに対しても本発明は有効である。

[0040] 辞書作成装置は、テキスト情報監視用辞書で用いる辞書を作成するものである。

[0041] 図3は、正例集合、負例集合の例である。このような正例集合、負例集合が与えられていることを前提とする。

[0042] まず、フレーズ抽出部1は、正例集合から検知条件候補の抽出を行う。例えば、図3の正例集合から3文節以下のフレーズを全て抽出すると、「トロイの木馬」、「トロイ」、「木馬」、「トロイの木馬に感染」、「木馬に感染」、「感染」、「メール」といったフレーズが検知条件候補として抽出される。

[0043] 次に、特徴度計算部3は、各検知条件候補に対して特徴度を計算する。図4は各フレーズの頻度と特徴度の例である。例えば、特徴度を、

$$\text{特徴度} = (\text{正例集合での頻度}) - (\text{負例集合での頻度})$$

として計算すると、「トロイの木馬」は特徴度=3、「トロイ」は特徴度=3、「木馬」は特徴度=3、「トロイの木馬に感染」は特徴度=2、「木馬に感染」は特徴度=2、「感染」は特徴度=1、「メール」は特徴度=1と計算される。

[0044] 出力部4は、例えば、特徴度の高いフレーズ「トロイの木馬」「トロイ」「木馬」を出力し、辞書に登録する。

[0045] ～具体的な適用例1～

フレーズ抽出部1および特徴度計算部3の動作は従来技術と同様である。すなわち、正例集合から検知条件候補の抽出を行い、各検知条件候補に対して特徴度を計算する。

[0046] 更に、有用度計算部 2 1 は各検知条件候補に対して有用度を計算する。図 5 は各フレーズの有用度とスコア（後述）の例である。例えば、有用度を、フレーズの長さ（単語数）と正例集合での頻度との積に基づいて計算する。すなわち、

$$\text{有用度} = (\text{フレーズの長さ}) \times (\text{正例集合での頻度})$$

として計算すると、「トロイの木馬」は有用度=6、「トロイ」は有用度=3、「木馬」は有用度=3、「トロイの木馬に感染」は有用度=6、「木馬に感染」は有用度=4、「感染」は有用度=2、「メール」は有用度=2 と計算される。ここでフレーズの長さは、文節数で計算したが、それ以外にも、形態素数、文字数、バイト長などで長さを計算してもよい。

[0047] 次に、検知条件判定部 2 2 は各検知条件候補の評価を行う（図 5 参照）。例えば、検知条件としての適切さを表すスコアを、有用度と特徴度の積に基づいて計算する。すなわち、

$$\text{スコア} = \text{特徴度} \times \text{有用度}$$

として計算すると、「トロイの木馬」はスコア=18、「トロイ」はスコア=9、「木馬」はスコア=9、「トロイの木馬に感染」はスコア=12、「木馬に感染」はスコア=8、「感染」はスコア=2、「メール」はスコア=2 と計算される。そして、例えば、スコアが 10 以上のフレーズを検知条件として採用すると、「トロイの木馬」と「トロイの木馬に感染」の 2 つが検知条件として適切であると判定する。

[0048] 出力部 4 は、検知条件判定部 2 2 の判定結果に基づいてフレーズ「トロイの木馬」、「トロイの木馬に感染」を出力し、辞書に登録する。

[0049] ～効果～

従来技術と比較することで、本実施形態の効果について説明する。

[0050] 特徴度のみに基づいて検知条件の判定を行う従来技術においては、「トロイの木馬」、「トロイ」、「木馬」が特徴度=3 と最大となり、これらが検知条件となる。しかし、「トロイ」には「トロイ遺跡」、「木馬」には「回転木馬」といった本来検知したくない表現も検知されることになり、検知精度を下げることになる。

[0051] これに対し、本実施の形態では、フレーズ有用性判定部2は、候補となるフレーズの長さを用いて、フレーズを検知条件としたときの検知条件としての良さを表す有用度を計算し、得られた有用度と別途計算する特徴度とを用いて辞書登録すべきフレーズの判定を行う。

[0052] 一般に、フレーズの長さが長いほど意味の曖昧性が少なくなり、検知条件としての適合率は高くなる。そこで、互いに重複があるフレーズが同じ特徴度となった場合は長さが長いフレーズを選択することで特徴度のみを用いる場合より高精度な検知が可能となる。

[0053] 更に、フレーズの長さに加えて、フレーズの文書集合中の頻度を用いて有用度を計算する。フレーズの長さが長いほど適合率は高くなるが、フレーズの出現確率は下がるため再現率は低くなると考えられる、そこでフレーズの長さに加えて頻度を考慮することで、適合率と再現率がバランスした有用度を計算でき、より高精度な検知が可能となる。

[0054] 本実施形態では、「トロイの木馬」、「トロイの木馬に感染」が検知条件となり、「トロイ」、「木馬」は辞書に登録されない。その結果、従来技術に比べて高精度な検知を実現できる。

[0055] ～具体的な適用例2～

上記適用例1では、有用度計算部21は、フレーズの長さと正例集合での頻度との積に基づいて有用度を計算するが、有用度に更なる顕著な差を付けたい場合は、フレーズの長さから補正値を減じてよい。

[0056] 図6は各フレーズの有用度とスコアの別例である。例えば、有用度計算部21はフレーズの長さから補正値を減じた値と正例集合での頻度との積に基づいて有用度を計算する。補正値は経験的に求めてもよい。ここでは補正値を「-0.5」とする。すなわち、

$$\text{有用度} = (\text{フレーズの長さ} - 0.5) \times (\text{正例集合での頻度})$$

として計算すると「トロイの木馬」は有用度=4.5、「トロイ」は有用度=1.5、「木馬」は有用度=1.5、「トロイの木馬に感染」は有用度=5、「木馬に感染」は有用度=3、「感染」は有用度=1、「メール」は有用度=1

と計算される。

[0057] このように、フレーズの長さがより強調されるように補正される。

[0058] 次に、検知条件判定部 22 がスコア＝特徴度×有用度として計算すると、
「トロイの木馬」はスコア＝13.5、「トロイ」はスコア＝4.5、「木馬」はスコア＝4.5、「トロイの木馬に感染」はスコア＝10、「木馬に感染」はスコア＝6、「感染」はスコア＝1、「メール」はスコア＝1 と計算される。
そして、例えば、スコアが 10 以上のフレーズを検知条件として採用すると、「トロイの木馬」と「トロイの木馬に感染」の 2 つが検知条件として適切であると判定する。

[0059] 適用例 1 に比べて、「トロイの木馬」のスコアに対する「トロイ」、「木馬」のスコアの割合が低減している。すなわち、「トロイの木馬」はより確実に辞書に登録され、「トロイ」、「木馬」はより確実に辞書登録から除外される。これにより精度が向上する。

[0060] ～具体的な適用例 3～

上記適用例 1、適用例 2 では、検知条件判定部 22 はスコアが 10 以上のフレーズを検知条件として採用するように設定しているため、「木馬に感染」は辞書に登録されないが、設定によっては登録され得る。「木馬に感染」は「トロイの木馬に感染」に包含され、ほとんどの場合、「トロイの木馬に感染」という言い回し、いわゆる定型フレーズとして用いられる。したがって、「木馬に感染」と「トロイの木馬に感染」の両方を辞書に登録しても意味がない。

[0061] そこで、有用度計算部 21 は、フレーズの長さで正例集合での頻度に加えて、フレーズ間の包含関係を表す指標に基づいて有用度を計算する。例えば、C-value を有用度としてもよい。C-value は以下の式で計算される値のことである。図 7 は各フレーズの有用度 (C-value) とスコアの別例である。

C-value の定義

$$C\text{-value} = (\text{フレーズ長さ}) \times (\text{正例集合での頻度} - T/C) \quad (C > 0 \text{ の場合})$$

$$C\text{-value} = (\text{フレーズ長さ}) \times (\text{正例集合での頻度}) \quad (C = 0 \text{ の場合})$$

場合)

T: 着目フリースを包含し着目フリースよりも長いフリースの出現頻度の合計
C: 着目フリースを包含し着目フリースよりも長いフリースの異なり数 (つ

まり、そのようなフリースがいくつあるか)

[0062] 以下、T, Cについて具体的に説明する (図7参照)。

[0063] 着目フリース: 「トロイの木馬」

着目フリースを包含し着目フリースよりも長いフリース: 「トロイの木馬に

感染」

T=2: 「トロイの木馬に感染」出現頻度2

C=1: 着目フリースを包含し着目フリースよりも長いフリース1

[0064] 着目フリース: 「トロイ」

着目フリースを包含し着目フリースよりも長いフリース: 「トロイの木馬」

「トロイの木馬に感染」

T=3+2=5: 「トロイの木馬」出現頻度3、 「トロイの木馬に感染」出現頻度2

C=2: 着目フリースを包含し着目フリースよりも長いフリース2

[0065] 着目フリース: 「木馬」

着目フリースを包含し着目フリースよりも長いフリース: 「トロイの木馬」

「トロイの木馬に感染」 「木馬に感染」

T=3+2+2=7: 「トロイの木馬」出現頻度3、 「トロイの木馬に感染」出現頻度2

、「木馬に感染」出現頻度2

C=3: 着目フリースを包含し着目フリースよりも長いフリースは3

[0066] 着目フリース: 「トロイの木馬に感染」

着目フリースを包含し着目フリースよりも長いフリース: なし

T=0

C=0

[0067] 着目フリース: 「木馬に感染」

着目フリースを包含し着目フリースよりも長いフリース: 「トロイの木馬に

感染」

T=2: 「トロイの木馬に感染」出現頻度2

C=1: 着目フレーズを包含し着目フレーズよりも長いフレーズ1

[0068] 着目フレーズ: 「感染」

着目フレーズを包含し着目フレーズよりも長いフレーズ: 「トロイの木馬に感染」「木馬に感染」

T=2+2=4: 「トロイの木馬に感染」出現頻度2、「木馬に感染」出現頻度2

C=2: 着目フレーズを包含し着目フレーズよりも長いフレーズ2

[0069] 着目フレーズ: 「メール」

着目フレーズを包含し着目フレーズよりも長いフレーズ: なし

T=0

C=0

[0070] T, Cにより補正することにより、「トロイの木馬」は有用度=2、「トロイ」は有用度=0.5、「木馬」は有用度=0.67、「トロイの木馬に感染」は有用度=6、「木馬に感染」は有用度=0、「感染」は有用度=0、「メール」は有用度=2と計算される。

[0071] 「トロイの木馬に感染」の有用度6に対し「木馬に感染」の有用度0である。この結果は、「木馬に感染」が正例文書集合中で必ず「トロイの木馬に感染」という言い回しで用いられている定型フレーズであるため、「木馬に感染」の用語性は低く、検知条件として「トロイの木馬に感染」があれば、「木馬に感染」を条件として追加する意味がないことを示している。

[0072] 一方、「トロイの木馬」の有用度は2である。「トロイの木馬」は「トロイの木馬に感染」以外にも用例があるため、「木馬に感染」よりも用語性が高く、C-valueも大きくなる。

[0073] 用語性とはひとかたまりのフレーズとして使われやすさを表す指標であり、用語性が高いとは、ひとかたまりのフレーズとして使われやすいという意味である。

[0074] このように、有用度としてC-valueを用いることで他のより長いフレーズに

包含されるフレーズは値が小さくなり、冗長な検知条件が追加されることをなくし、辞書精度の向上を図ることができる。

[0075] 次に、検知条件判定部 22 がスコア＝特徴度×有用度として計算すると、「トロイの木馬」はスコア＝6、「トロイ」はスコア＝1.5、「木馬」はスコア＝2、「トロイの木馬に感染」はスコア＝1.2、「木馬に感染」はスコア＝0、「感染」はスコア＝0、「メール」はスコア＝2 と計算される。そして、例えば、スコアが 5 以上のフレーズを検知条件として採用すると、「トロイの木馬」と「トロイの木馬に感染」の 2 つが検知条件として適切であると判定する。

[0076] ～具体的な適用例 4～

適用例 3 において、適用例 2 で説明した補正值を用いてもよい。ここでは補正值を「-1」とする。図 8 は各フレーズの有用度 (C-value) とスコアの別例である。

C-value の定義

$$C\text{-value} = (\text{フレーズ長さ} - 1) \times (\text{正例集合での頻度} - T/C)$$

($C > 0$ の場合)

$$C\text{-value} = (\text{フレーズ長さ} - 1) \times (\text{正例集合での頻度})$$

($C = 0$ の場合)

T: 着目フレーズを包含し着目フレーズよりも長いフレーズの出現頻度の合計

C: 着目フレーズを包含し着目フレーズよりも長いフレーズの異なり数 (つまり、そのようなフレーズがいくつあるか)

フレーズ長さの項にある「-1」は、適用例 2 で記載した補正值「-0.5」と同種の値である。すなわち、フレーズの長さをより強調する補正值である。

[0077] これにより、有用度の差が更に顕著なる。

[0078] ～具体的な適用例 5～

有用度が閾値以上のフレーズのみ、特徴度計算部 3 は特徴度を計算し、検

知条件判定部 22 は検知条件として適切であるか否かを判定する。

[0079] 具体的に適用例 2 と比較して説明する。図 8 は各フレーズの有用度とスコアの別例である。

[0080] 適用例 2 と同様に、有用度計算部 21 は「トロイの木馬」は有用度=4.5、「トロイ」は有用度=1.5、「木馬」は有用度=1.5、「トロイの木馬に感染」は有用度=5、「木馬に感染」は有用度=3、「感染」は有用度=1、「メール」は有用度=1 と計算する。

[0081] 特徴度計算部 3 は、例えば有用度が 3 以上のフレーズ「トロイの木馬」「トロイの木馬に感染」「木馬に感染」のみ特徴度を計算する。次に、検知条件判定部 22 がスコア=特徴度×有用度として計算すると、「トロイの木馬」はスコア=13.5、「トロイの木馬に感染」はスコア=10、「木馬に感染」はスコア=6 と計算される。そして、例えば、スコアが 10 以上のフレーズを検知条件として採用すると、「トロイの木馬」と「トロイの木馬に感染」の 2 つが検知条件として適切であると判定する。

[0082] 適用例 2 では、全フレーズ（7 フレーズ）に対し特徴度計算および判定を行うのに対し、適用例 5 では「トロイの木馬」「トロイの木馬に感染」「木馬に感染」の 3 フレーズのみ特徴度計算および判定を行う。一方、判定結果は、適用例 2 も適用例 5 も同じであり精度は同じである。

[0083] これにより、精度を維持しながら、計算量を少なく抑えることができる。

[0084] ～補足～

適用例 1 は、主に請求項 4 および請求項 7 の詳細について説明している。適用例 2 は、主に請求項 4 を除く請求項 3 について説明している。適用例 3・4 は、主に請求項 5 および請求項 6 について説明している。適用例 5 は、主に請求項 8 について説明している。

[0085] 本発明は、テキスト情報監視システムで用いられる辞書を作成する装置であるが、インターネットを対象とした風評監視システムや評判抽出システム等にも適用できる。

[0086] ～付記～

上記実施形態において、各部をハードウェアで構成してもよいし、コンピュータプログラムにより実現してもよい。この場合、プログラムメモリに格納されているプログラムで動作するプロセッサによって、上述と同様の機能、動作を実現させる。また、一部の機能のみをコンピュータプログラムにより実現してもよい。

[0087] また、上記の実施形態の一部又は全部は、以下の付記のようにも記載されるが、以下には限られない。

[0088] 本発明は、

テキスト情報監視システムで用いられ、検知条件が登録される辞書を作成するテキスト情報監視用辞書作成装置であって、

検知条件候補のフレーズに対して、フレーズが監視対象の情報内容に適合する度合いを表す特徴度を計算する特徴度計算部と、

前記特徴度とフレーズによって規定される意味の曖昧さの少なさを表す有用度とに基づいて、フレーズが検知条件として適切であるか否かを判定するフレーズ有用性判定部と

を備えることを特徴とする。

[0089] 本発明のテキスト情報監視用辞書作成装置において、好ましくは、前記フレーズ有用性判定部は、

フレーズの長さに基づいて前記有用度を計算する有用度計算部と、

前記有用度計算部で計算された有用度と前記特徴度とに基づいて、フレーズが検知条件として適切であるか否かを判定する検知条件判定部とを有する。

[0090] 本発明のテキスト情報監視用辞書作成装置において、より好ましくは、

前記有用度計算部は、前記フレーズの長さと文書集合中の頻度とに基づいて有用度を計算する。

[0091] 一般に、フレーズの長さが長いほど意味の曖昧性が少なくなり、検知条件としての適合率は高くなる。本発明では上記構成により、長さの長いフレーズを優先する。その結果、従来技術に比べて高精度な検知が実現可能となる

。

[0092] 例えば、

前記有用度計算部は、フレーズの長さ又はその対数値と文書集合中の頻度又はその対数値との積によって有用度を計算する。

[0093] 本発明のテキスト情報監視用辞書作成装置において、好ましくは、

前記有用度計算部は、前記フレーズの長さ、文書集合中の頻度とフレーズ間の包含関係を表す指標とに基づいて有用度を計算する。

[0094] より好ましくは、

前記フレーズ間の包含関係を表す指標は、

着目フレーズより長い他フレーズが着目フレーズを包含する場合、

他フレーズの出現頻度の合計と他フレーズの数との比である。

[0095] 包含関係を考慮することにより、他のより長いフレーズに包含されるフレーズは値が小さくなり、冗長な検知条件が追加されることをなくし、辞書精度の向上を図ることができる。

[0096] 本発明のテキスト情報監視用辞書作成装置において、好ましくは、

前記検知条件判定部は、前記有用度又はその対数値と前記特徴度又はその対数値との積によってフレーズが検知条件として適切であるか否かを判定する。

[0097] これにより、有用度を考慮した検知ができる。

[0098] 本発明のテキスト情報監視用辞書作成装置において、より好ましくは、

前記有用度計算部で計算された有用度が閾値以上のフレーズに対し、

前記特徴度計算部は特徴度を計算し、

前記検知条件判定部はフレーズが検知条件として適切であるか否かを判定する。

[0099] これにより、精度を維持しながら、計算量を少なく抑えることができる。

[0100] 本発明は、

テキスト情報監視システムで用いられる辞書の作成方法であって、

テキスト情報監視用辞書作成装置が、

検知条件候補のフレーズに対して、フレーズが監視対象の情報内容に適合する度合いを表す特徴度を計算し、

前記特徴度とフレーズによって規定される意味の曖昧さの少なさを表す有用度とに基づいて、フレーズが検知条件として適切であるか否かを判定し、適切であると判断したフレーズを出力し検知条件として登録することを特徴とする。

[0101] 本発明のテキスト情報監視用辞書作成方法において、好ましくは、フレーズの長さに基づいて前記有用度を計算し、前記有用度と前記特徴度とに基づいて、フレーズが検知条件として適切であるか否かを判定する。

[0102] より好ましくは、前記フレーズの長さと文書集合中の頻度とに基づいて有用度を計算する。

[0103] 例えば、フレーズの長さ又はその対数値と文書集合中の頻度又はその対数値との積によって有用度を計算する。

[0104] 本発明のテキスト情報監視用辞書作成方法において、好ましくは、前記フレーズの長さと文書集合中の頻度とフレーズ間の包含関係を表す指標とに基づいて有用度を計算する。

[0105] より好ましくは、前記フレーズ間の包含関係を表す指標は、着目フレーズより長い他フレーズが着目フレーズを包含する場合、他フレーズの出現頻度の合計と他フレーズの数との比である。

[0106] 本発明のテキスト情報監視用辞書作成方法において、好ましくは、前記有用度又はその対数値と前記特徴度又はその対数値との積によってフレーズが検知条件として適切であるか否かを判定する。

[0107] 本発明のテキスト情報監視用辞書作成方法において、より好ましくは、前記有用度計算部で計算された有用度が閾値以上のフレーズに対し、特徴度を計算し、

フレーズが検知条件として適切であるか否かを判定する。

[0108] 本発明は、

テキスト情報監視用辞書作成プログラムであって、

検知条件候補のフレーズに対して、フレーズが監視対象の情報内容に適合する度合いを表す特徴度を計算する処理と、

前記特徴度とフレーズによって規定される意味の曖昧さの少なさを表す有用度とに基づいて、フレーズが検知条件として適切であるか否かを判定する処理と、

適切であると判断したフレーズを出力し検知条件として登録する処理とをテキスト情報監視用辞書作成装置に実行させることを特徴とする。

[0109] 本発明のテキスト情報監視用辞書作成プログラムにおいて、好ましくは、フレーズの長さに基づいて前記有用度を計算する処理と、

前記有用度と前記特徴度とに基づいて、フレーズが検知条件として適切であるか否かを判定する処理と

を実行させる。

[0110] 本発明のテキスト情報監視用辞書作成プログラムにおいて、より好ましくは、

前記有用度計算処理では、前記フレーズの長さと文書集合中の頻度とに基づいて有用度を計算する。

[0111] 例えば、

前記有用度計算処理では、フレーズの長さ又はその対数値と文書集合中の頻度又はその対数値との積によって有用度を計算する。

[0112] 本発明のテキスト情報監視用辞書作成プログラムにおいて、好ましくは、

前記有用度計算処理では、前記フレーズの長さと文書集合中の頻度とフレーズ間の包含関係を表す指標とに基づいて有用度を計算する。

[0113] より好ましくは、

前記フレーズ間の包含関係を表す指標は、

着目フレーズより長い他フレーズが着目フレーズを包含する場合、

他フリースの出現頻度の合計と他フリースの数との比である。

[0114] 本発明のテキスト情報監視用辞書作成プログラムにおいて、好ましくは、

前記検知条件判定処理では、前記有用度又はその対数値と前記特徴度又は

その対数値との積によってフリースが検知条件として適切であるか否かを判

定する。

[0115] 本発明のテキスト情報監視用辞書作成プログラムにおいて、より好ましく

は、

前記有用度計算処理で計算された有用度が閾値以上のフリースに対し、

前記特徴度計算処理では、特徴度を計算し、

前記検知条件判定処理では、フリースが検知条件として適切であるか否か

を判定する。

[0116] 本出願は、2012年9月27日に出願された日本出願特願2012-21

3536号を基礎とする優先権を主張し、その開示の全てをここに取り込む

符号の説明

[0117] 1 フリース抽出部

2 フリース有用性判定部

3 特徴度計算部

4 出力部

21 有用度計算部

22 検知条件判定部

請求の範囲

- [請求項1] テキスト情報監視システムで用いられ、検知条件が登録される辞書を作成するテキスト情報監視用辞書作成装置であって、
- 検知条件候補のフレーズに対して、フレーズが監視対象の情報内容に適合する度合いを表す特徴度を計算する特徴度計算部と、
- 前記特徴度とフレーズによって規定される意味の曖昧さの少なさを表す有用度とに基づいて、フレーズが検知条件として適切であるか否かを判定するフレーズ有用性判定部と
- を備えることを特徴とするテキスト情報監視用辞書作成装置。
- [請求項2] 前記フレーズ有用性判定部は、
- フレーズの長さに基づいて前記有用度を計算する有用度計算部と、
- 前記有用度計算部で計算された有用度と前記特徴度とに基づいて、フレーズが検知条件として適切であるか否かを判定する検知条件判定部と
- を有することを特徴とする請求項1に記載のテキスト情報監視用辞書作成装置。
- [請求項3] 前記有用度計算部は、前記フレーズの長さと文書集合中の頻度とに基づいて有用度を計算する
- ことを特徴とする請求項2に記載のテキスト情報監視用辞書作成装置。
- [請求項4] 前記有用度計算部は、フレーズの長さ又はその対数値と文書集合中の頻度又はその対数値との積によって有用度を計算する
- ことを特徴とする請求項3に記載のテキスト情報監視用辞書作成装置。
- [請求項5] 前記有用度計算部は、前記フレーズの長さと文書集合中の頻度とフレーズ間の包含関係を表す指標とに基づいて有用度を計算する
- ことを特徴とする請求項2に記載のテキスト情報監視用辞書作成装置。

置。

[請求項6]

前記フレーズ間の包含関係を表す指標は、
着目フレーズより長い他フレーズが着目フレーズを包含する場合、
他フレーズの出現頻度の合計と他フレーズの数との比である
ことを特徴とする請求項5に記載のテキスト情報監視用辞書作成装置。

[請求項7]

前記検知条件判定部は、
前記有用度又はその対数値と前記特徴度又はその対数値との積によってフレーズが検知条件として適切であるか否かを判定する
ことを特徴とする請求項2～6に記載のテキスト情報監視用辞書作成装置。

[請求項8]

前記有用度計算部で計算された有用度が閾値以上のフレーズに対し、
前記特徴度計算部は特徴度を計算し、
前記検知条件判定部はフレーズが検知条件として適切であるか否かを判定する
ことを特徴とする請求項2～7に記載のテキスト情報監視用辞書作成装置。

[請求項9]

テキスト情報監視システムで用いられる辞書の作成方法であって、
テキスト情報監視用辞書作成装置が、
検知条件候補のフレーズに対して、フレーズが監視対象の情報内容に適合する度合いを表す特徴度を計算し、
前記特徴度とフレーズによって規定される意味の曖昧さの少なさを表す有用度とに基づいて、フレーズが検知条件として適切であるか否かを判定し、
適切であると判断したフレーズを出力し検知条件として登録する
ことを特徴とするテキスト情報監視用辞書作成方法。

[請求項10]

検知条件候補のフレーズに対して、フレーズが監視対象の情報内容

に適合する度合いを表す特徴度を計算する処理と、

前記特徴度とフレーズによって規定される意味の曖昧さの少なさを表す有用度とに基づいて、フレーズが検知条件として適切であるか否かを判定する処理と、

適切であると判断したフレーズを出力し検知条件として登録する処理と

をテキスト情報監視用辞書作成装置に実行させることを特徴とするテキスト情報監視用辞書作成プログラム。