

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5648890号
(P5648890)

(45) 発行日 平成27年1月7日(2015.1.7)

(24) 登録日 平成26年11月21日(2014.11.21)

(51) Int. Cl. F 1
G 0 6 F 17/27 (2006.01) G 0 6 F 17/27 Z

請求項の数 10 (全 26 頁)

<p>(21) 出願番号 特願2010-36415 (P2010-36415) (22) 出願日 平成22年2月22日 (2010. 2. 22) (65) 公開番号 特開2011-170790 (P2011-170790A) (43) 公開日 平成23年9月1日 (2011. 9. 1) 審査請求日 平成25年1月11日 (2013. 1. 11)</p>	<p>(73) 特許権者 000004237 日本電気株式会社 東京都港区芝五丁目7番1号 (73) 特許権者 899000068 学校法人早稲田大学 東京都新宿区戸塚町1丁目104番地 (74) 代理人 100103090 弁理士 岩壁 冬樹 (74) 代理人 100124501 弁理士 塩川 誠人 (72) 発明者 立石 健二 東京都港区芝五丁目7番1号 日本電気株式会社社内</p>
--	--

最終頁に続く

(54) 【発明の名称】 辞書作成支援装置、辞書作成支援方法及び辞書作成支援プログラム

(57) 【特許請求の範囲】

【請求項1】

辞書作成を支援する辞書作成支援装置であって、

同義語を生成する対象の語として当該辞書作成支援装置に入力される入力語をアンカーテキストに含む資源位置指定子を抽出し、抽出された資源位置指定子で識別されるWebページにリンクする各アンカーテキストと前記入力語との同義性を計算し、計算された同義性が上位のアンカーテキストを同義語候補として生成する同義語候補生成手段と、

予め規定した判定ルールまたは利用者による判定に基づいて入力語の同義語か非同義語かが判定された前記同義語候補を用いて、入力語もしくは当該入力語の同義語と判定された前記同義語候補で表わされるアンカーテキストに設定された、一の資源位置指定子で識別されるWebページへのリンクである第一のリンクと、当該入力語の非同義語と判定された前記同義語候補で表わされるアンカーテキストに設定された、前記一の資源位置指定子で識別されるWebページへのリンクである第二のリンクのうち、少なくとも一方のリンクを削除するリンク削除手段とを備え、

前記リンク削除手段は、一の資源位置指定子について、各アンカーテキストからのリンクの数の、全アンカーテキストからのリンクの数に対する相対的な割合を算出し、一の資源位置指定子に対する第一のリンクと第二のリンクのうち、前記相対的な割合が小さいほうのリンクを削除し、

前記同義語候補生成手段は、前記リンク削除手段が削除する対象から除かれた各資源位置指定子のリンクのアンカーテキストを用いて入力語の同義語候補を生成する

10

20

ことを特徴とする辞書作成支援装置。

【請求項 2】

資源位置指定子で識別される Web ページへのリンクが設定されたアンカーテキストが当該資源位置指定子の実体を表す確率である実体確率を、資源位置指定子とアンカーテキストとのリンクごとに計算する実体確率計算手段を備え、

リンク削除手段は、第一のリンクと第二のリンクのうち、前記実体確率が小さいリンクを削除する

請求項 1 記載の辞書作成支援装置。

【請求項 3】

実体確率計算手段は、各アンカーテキストから一の資源位置指定子で識別される Web ページへのリンクの総数に対する一のアンカーテキストから当該一の資源位置指定子で識別される Web ページへのリンクの数の割合を実体確率として算出する

請求項 2 記載の辞書作成支援装置。

【請求項 4】

実体確率計算手段は、各アンカーテキストから一の資源位置指定子で識別される Web ページへのリンクの総数に対する一のアンカーテキストから当該一の資源位置指定子で識別される Web ページへのリンクの数の割合、及び、前記アンカーテキストを持つ各資源位置指定子へのリンクの総数に対する当該アンカーテキストから前記一の資源位置指定子で識別される Web ページへのリンクの数の割合を用いて実体確率を算出する

請求項 2 記載の辞書作成支援装置。

【請求項 5】

アンカーテキストが同義語候補になっている各資源位置指定子に対するリンクのうち、入力語の同義語と判定された同義語候補がアンカーテキストであるリンクを統合するリンク統合手段を備え、

同義語候補生成手段は、統合された前記リンクのアンカーテキストを用いて入力語の同義語候補を生成する

請求項 1 から請求項 4 のうちのいずれか 1 項に記載の辞書作成支援装置。

【請求項 6】

リンク削除手段は、第一のリンクと第二のリンクの双方が一の資源位置指定子に存在しない場合に、当該一の資源位置指定子に対するアンカーテキストのリンクを削除対象から除く

請求項 1 から請求項 5 のうちのいずれか 1 項に記載の辞書作成支援装置。

【請求項 7】

辞書作成を支援する辞書作成支援方法であって、

辞書作成支援装置の同義語候補生成手段が、同義語を生成する対象の語として入力される入力語をアンカーテキストに含む資源位置指定子を抽出し、

前記同義語候補生成手段が、抽出された資源位置指定子で識別される Web ページにリンクする各アンカーテキストと前記入力語との同義性を計算し、

前記同義語候補生成手段が、計算された同義性が上位のアンカーテキストを同義語候補として生成し、

辞書作成支援装置のリンク削除手段が、予め規定した判定ルールまたは利用者による判定に基づいて入力語の同義語か非同義語かが判定された前記同義語候補を用いて、入力語もしくは当該入力語の同義語と判定された前記同義語候補で表わされるアンカーテキストに設定された、一の資源位置指定子で識別される Web ページへのリンクである第一のリンクと、当該入力語の非同義語と判定された前記同義語候補で表わされるアンカーテキストに設定された、前記一の資源位置指定子で識別される Web ページへのリンクである第二のリンクのうち、少なくとも一方のリンクを削除し、

前記リンク削除手段が、前記リンクを削除する際に、一の資源位置指定子について、各アンカーテキストからのリンクの数の、全アンカーテキストからのリンクの数に対する相対的な割合を算出し、一の資源位置指定子に対する第一のリンクと第二のリンクのうち、

10

20

30

40

50

前記相対的な割合が小さいほうのリンクを削除し、

前記同義語候補生成手段が、同義語候補を生成する際に、削除対象から除かれた各資源位置指定子のリンクのアンカーテキストを用いて入力語の同義語候補を生成することを特徴とする辞書作成支援方法。

【請求項 8】

辞書作成支援装置の実体確率計算手段が、資源位置指定子で識別される Web ページへのリンクが設定されたアンカーテキストが当該資源位置指定子の実体を表す確率である実体確率を、資源位置指定子とアンカーテキストとのリンクごとに計算し、

リンク削除手段が、リンクを削除する際に、第一のリンクと第二のリンクのうち、前記実体確率が小さいリンクを削除する

請求項 7 記載の辞書作成支援方法。

【請求項 9】

辞書作成を支援するコンピュータに搭載される辞書作成支援プログラムであって、前記コンピュータに、

同義語を生成する対象の語として当該コンピュータに入力される入力語をアンカーテキストに含む資源位置指定子を抽出し、抽出された資源位置指定子で識別される Web ページにリンクする各アンカーテキストと前記入力語との同義性を計算し、計算された同義性が上位のアンカーテキストを同義語候補として生成する同義語候補生成処理、および、

予め規定した判定ルールまたは利用者による判定に基づいて入力語の同義語か非同義語かが判定された前記同義語候補を用いて、入力語もしくは当該入力語の同義語と判定された前記同義語候補で表わされるアンカーテキストに設定された、一の資源位置指定子で識別される Web ページへのリンクである第一のリンクと、当該入力語の非同義語と判定された前記同義語候補で表わされるアンカーテキストに設定された、前記一の資源位置指定子で識別される Web ページへのリンクである第二のリンクのうち、少なくとも一方のリンクを削除するリンク削除処理を実行させ、

前記リンク削除処理で、一の資源位置指定子について、各アンカーテキストからのリンクの数の、全アンカーテキストからのリンクの数に対する相対的な割合を算出させ、一の資源位置指定子に対する第一のリンクと第二のリンクのうち、前記相対的な割合が小さいほうのリンクを削除させ、

前記同義語候補生成処理で、前記リンク削除処理で削除する対象から除かれた各資源位置指定子のリンクのアンカーテキストを用いて入力語の同義語候補を生成させる

ことを特徴とする辞書作成支援プログラム。

【請求項 10】

コンピュータに、

資源位置指定子で識別される Web ページへのリンクが設定されたアンカーテキストが当該資源位置指定子の実体を表す確率である実体確率を、資源位置指定子とアンカーテキストとのリンクごとに計算する実体確率計算処理を実行させ、

リンク削除処理で、第一のリンクと第二のリンクのうち、前記実体確率が小さいリンクを削除させる

請求項 9 記載の辞書作成支援プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、Web のアンカーテキストを用いて入力語に対する同義語候補を生成する辞書作成支援装置、辞書作成支援方法及び辞書作成支援プログラムに関する。

【背景技術】

【0002】

同義語辞書は、文書検索、顧客データの名寄せなど、様々なソフトウェアの基本的な資源として使用される。同義語の定義としては様々なものが存在するが、ここでは、表記が異なり、同じ対象物を示す 2 つの語を同義語とする。

10

20

30

40

50

【 0 0 0 3 】

非特許文献 1 には、同義語辞書の作成支援方法として、利用者が入力した語の同義語候補を Web (ウェブ) のアンカーテキストを用いて生成する方法が開示されている。非特許文献 1 に記載された方法は、ある Web ページを示す URL に対する複数のアンカーテキストは、それぞれが同様の表現を含んでいると判断されることから、それらの表現を同義であるとみなすものである。

【 先行技術文献 】

【 非特許文献 】

【 0 0 0 4 】

【 非特許文献 1 】 WEN-HSIANG LU, LEE-FENG CHIEN, HIS-JIAN LEE, Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach, ACM Transactions on Information Systems, Vol. 22, No. 2, pp. 242-269, 2004.

10

【 発明の概要 】

【 発明が解決しようとする課題 】

【 0 0 0 5 】

例えば、非特許文献 1 に記載された方法を用いることで、同義語候補を生成することが可能である。以下、非特許文献 1 に記載された方法を利用した同義語候補の生成方法について説明する。

【 0 0 0 6 】

今、URL (Uniform Resource Locator) u に対して 2 種類のアンカーテキスト s 及び t が存在するとする。ここでアンカーテキストとは、URL u へリンクする Web 文書におけるリンク上の文字列を表す。つまり、アンカーテキスト s は、 u に対して s という文字列で別の Web 文書からリンクしていることになる。また、以下の説明では、アンカーテキスト s が、URL u へリンクするリンク上の文字列であることを、URL u がアンカーテキスト s を持つと表現することもある。

20

【 0 0 0 7 】

次に、 $P(x|u)$ を URL u にリンクするアンカーテキスト x が URL u の実体を表す確率と定義する。例えば、URL u が「www.nec.co.jp」であるとする。ここで、アンカーテキスト x が「日本電気」の場合、アンカーテキスト x は、URL u の実体を表すので、 $P(\text{日本電気}|www.nec.co.jp) = 1$ になることが理想である。言い換えると、 $P(x|u)$ は、アンカーテキスト x が URL u によって示す対象をどれだけ正確に表しているかを示す指標である。

30

【 0 0 0 8 】

この $P(x|u)$ を用いて、アンカーテキスト s とアンカーテキスト t の同義性 $Rel(s, t)$ を次の式 1 で定義する。ここで、同義性とは、表記が異なる 2 つの語が同じ対象物を示す尤もらしさのことを言う。

【 0 0 0 9 】

$$Rel(s, t) = E [P(s|u) * P(t|u)]_{u} \quad (\text{式 1})$$

【 0 0 1 0 】

なお、 $P(s|u) * P(t|u)$ は、アンカーテキスト s とアンカーテキスト t の両方が URL u の実体を表す確率を意味する。今、URL の出現確率を全て一様とすれば、同義性 $Rel(s, t)$ は、以下の式 2 により算出される。

40

【 0 0 1 1 】

$$Rel(s, t) = \sum_i [P(s|u_i) * P(t|u_i)] / N \quad (\text{式 2})$$

【 0 0 1 2 】

ここで、 N は、アンカーテキスト s もしくはアンカーテキスト t が出現する URL の数である。式 2 による同義性 $Rel(s, t)$ の算出方法は、アンカーテキスト s 及びアンカーテキスト t が出現する全ての URL でその URL の実体を表していれば、アンカーテキスト s とアンカーテキスト t とは同義語であると判断するという考えに基づく。

【 0 0 1 3 】

50

ところで、 $P(s|u) * P(t|u)$ は、アンカーテキスト s とアンカーテキスト t の両方が URL u の実体を表す確率であり、その確率の上限値は、アンカーテキスト s とアンカーテキスト t のいずれか一方が URL u の実体を表す確率である。つまり、 $P(s|u) * P(t|u)$ について、以下の式 3 の関係が成り立つ

【0014】

$$P(s|u) * P(t|u) = 1 - \{ (1 - P(s|u)) * (1 - P(t|u)) \} \\ = P(s|u) + P(t|u) - P(s|u) * P(t|u) \quad (\text{式 3})$$

【0015】

したがって、同義性 $Rel(s, t)$ を正規化した $NRel(s, t)$ は、下記の式 4 により算出される。

【0016】

【数 1】

$$NRel(s,t) = \frac{\sum_i [P(s|u_i) * P(t|u_i)]}{\sum_i [P(s|u_i) + P(t|u_i) - P(s|u_i) * P(t|u_i)]} \quad (\text{式 4})$$

【0017】

一方、 $P(x|u)$ は、URL u へアンカーテキスト x でリンクする数を用いて推定される。具体的には、URL u に対するリンクの総数を L_u 、URL u に対するアンカーテキスト x によるリンクの数（すなわち、アンカーテキスト x で URL u にリンクする数）を $L_{u,x}$ とする。このとき、 $P(x|u)$ は、以下の式 5 により算出される。

【0018】

$$P(x|u) = L_{u,x} / L_u \quad (\text{式 5})$$

【0019】

なお、式 5 により算出される $P(x|u)$ は、「実体を表すアンカーテキストほど多くの人を用いる」ことを前提とした確率である。

【0020】

図 16 及び図 17 を用いて、通常の辞書作成支援装置が正規化された同義性 $NRel$ を算出する方法について具体的に説明する。図 16 は、URL とアンカーテキストとのリンク情報を示す説明図である。また、図 17 は、図 16 に示すリンク情報をもとに同義性 $NRel$ を算出する過程を示す説明図である。

【0021】

図 16 に示す $u_1 \sim u_4$ は URL を示し、 $s_1 \sim s_4$ はアンカーテキストを示す。また、 $URLu_1 \sim u_4$ とアンカーテキスト $s_1 \sim s_4$ とを結ぶ実線は、アンカーテキスト $s_1 \sim s_4$ がどの URL へリンクしているかを示す。また、表中の「P」は、式 5 により算出される $P(x|u)$ の値を示し、 $URLu_1 \sim u_4$ の下部に記載された値は、アンカーテキスト $s_1 \sim s_4$ による URL u へのリンク数を示す。

【0022】

例えば、利用者から入力語「NEC」を受け付けると、辞書作成支援装置は、アンカーテキスト「NEC」と他のアンカーテキスト間の $NRel$ を計算し、その値 ($NRel$) の上位を同義語候補として利用者に提示する。

【0023】

辞書作成支援装置は、図 16 に示すリンク情報をもとに式 4 を用いて同義性を示す値を算出する。具体的には、図 17 に示すように、「NEC」との同義性をそれぞれ、 $NRel(NEC, \text{日本電気}) = 0.26$ 、 $NRel(NEC, \text{BIGLOBE}) = 0.07$ 、 $NRel(NEC, \text{ビッグロブ}) = 0.03$ と算出する。そして、辞書作成支援装置は、これらを同義語候補としてこの順序で利用者に提示する。利用者は、この結果を順番に閲覧したり、取捨選択したりして、同義語辞書を作成する。なお、NEC、BIGLOBE 及びビッグロブは登録商標である。

【0024】

10

20

30

40

50

しかし、上述の方法を用いて同義語辞書を作成する辞書作成支援装置では、 $P(x|u)$ の値を正しく推定できない場合が存在する。例えば、アンカーテキストが誤って記述されたり、Web文書の作成者がURLの実体を表す文字列の一部にアンカーテキストを設定したりする場合などである。

【0025】

具体例を用いて説明する。例えば、Web文書の記述者が、テキスト「NECが提供するプロバイダー」を記述し、テキスト中の「NEC」の部分のURL「biglobe.ne.jp」へのアンカーテキストとして設定する場合がある。このような場合には、URL「biglobe.ne.jp」にリンクするアンカーテキスト「NEC」が存在することになってしまい、 $P(x|u)$ の値を正しく推定できなくなる。そのため、同義語候補を生成する精度が悪化してしまうという課題がある。

10

【0026】

図16に示すリンク情報を用いて、さらに説明する。図16に示すアンカーテキスト「NEC」は、実際には、URL「biglobe.ne.jp」の実体を表すものではない。そのため、本来は、 $P(\text{NEC}|\text{biglobe.ne.jp})$ の値が0に近くなることが望ましい。しかし、誤った情報が用いられた場合、図17に示すように、 $P(\text{NEC}|\text{biglobe.ne.jp})$ の値は、0.17と算出されてしまうことになる。同様に、アンカーテキスト「NEC」は、実際、URL「nec.jp」の実体を表すものである。そのため、本来は、 $P(\text{NEC}|\text{nec.jp})$ の値が1に近くなることが望ましい。しかし、誤った情報が用いられた場合、図17に示すように、 $P(\text{NEC}|\text{nec.jp})$ の値は、0.4と算出されてしまうことになる。

20

【0027】

このように算出された $P(x|u)$ を用いて同義性を算出すると、同義語と判断されるべき語の同義性 $NRe1$ の値が低く、同義語と判断されるべきでない語の同義性 $NRe1$ の値が高く算出されてしまう。この結果、精度の低い同義語候補が生成されてしまうという問題がある。

【0028】

そこで、本発明は、生成される同義語候補の精度を向上させて辞書作成を支援できる辞書作成支援装置、辞書作成支援システム及び辞書支援作成プログラムを提供することを目的とする。

30

【課題を解決するための手段】

【0029】

本発明による辞書作成支援装置は、辞書作成を支援する辞書作成支援装置であって、同義語を生成する対象の語としてその辞書作成支援装置に入力される入力語をアンカーテキストに含む資源位置指定子を抽出し、抽出された資源位置指定子で識別されるWebページにリンクする各アンカーテキストと入力語との同義性を計算し、計算された同義性が上位のアンカーテキストを同義語候補として生成する同義語候補生成手段と、予め規定した判定ルールまたは利用者による判定に基づいて入力語の同義語か非同義語かが判定された同義語候補を用いて、入力語もしくはその入力語の同義語と判定された同義語候補で表わされるアンカーテキストに設定された、一の資源位置指定子で識別されるWebページへのリンクである第一のリンクと、その入力語の非同義語と判定された同義語候補で表わされるアンカーテキストに設定された、一の資源位置指定子で識別されるWebページへのリンクである第二のリンクのうち、少なくとも一方のリンクを削除するリンク削除手段とを備え、リンク削除手段が、一の資源位置指定子について、各アンカーテキストからのリンクの数の、全アンカーテキストからのリンクの数に対する相対的な割合を算出し、一の資源位置指定子に対する第一のリンクと第二のリンクのうち、相対的な割合が小さいほうのリンクを削除し、同義語候補生成手段が、リンク削除手段により削除する対象から除かれた各資源位置指定子のリンクのアンカーテキストを用いて入力語の同義語候補を生成することを特徴とする。

40

【0030】

50

本発明による辞書作成支援方法は、辞書作成を支援する辞書作成支援方法であって、辞書作成支援装置の同義語候補生成手段が、同義語を生成する対象の語として入力される入力語をアンカーテキストに含む資源位置指定子を抽出し、同義語候補生成手段が、抽出された資源位置指定子で識別されるWebページにリンクする各アンカーテキストと入力語との同義性を計算し、同義語候補生成手段が、計算された同義性が上位のアンカーテキストを同義語候補として生成し、辞書作成支援装置のリンク削除手段が、予め規定した判定ルールまたは利用者による判定に基づいて入力語の同義語か非同義語かが判定された同義語候補を用いて、入力語もしくはその入力語の同義語と判定された同義語候補で表わされるアンカーテキストに設定された、一の資源位置指定子で識別されるWebページへのリンクである第一のリンクと、その入力語の非同義語と判定された同義語候補で表わされるアンカーテキストに設定された、一の資源位置指定子で識別されるWebページへのリンクである第二のリンクのうち、少なくとも一方のリンクを削除し、リンク削除手段が、リンクを削除する際に、一の資源位置指定子について、各アンカーテキストからのリンクの数の、全アンカーテキストからのリンクの数に対する相対的な割合を算出し、一の資源位置指定子に対する第一のリンクと第二のリンクのうち、相対的な割合が小さいほうのリンクを削除し、同義語候補生成手段が、同義語候補を生成する際に、削除対象から除かれた各資源位置指定子のリンクのアンカーテキストを用いて入力語の同義語候補を生成することを特徴とする。

10

【0031】

本発明による辞書作成支援プログラムは、辞書作成を支援するコンピュータに搭載される辞書作成支援プログラムであって、コンピュータに、同義語を生成する対象の語としてそのコンピュータに入力される入力語をアンカーテキストに含む資源位置指定子を抽出し、抽出された資源位置指定子で識別されるWebページにリンクする各アンカーテキストと入力語との同義性を計算し、計算された同義性が上位のアンカーテキストを同義語候補として生成する同義語候補生成処理、および、予め規定した判定ルールまたは利用者による判定に基づいて入力語の同義語か非同義語かが判定された同義語候補を用いて、入力語もしくはその入力語の同義語と判定された同義語候補で表わされるアンカーテキストに設定された、一の資源位置指定子で識別されるWebページへのリンクである第一のリンクと、その入力語の非同義語と判定された同義語候補で表わされるアンカーテキストに設定された、一の資源位置指定子で識別されるWebページへのリンクである第二のリンクのうち、少なくとも一方のリンクを削除するリンク削除処理を実行させ、リンク削除処理で、一の資源位置指定子について、各アンカーテキストからのリンクの数の、全アンカーテキストからのリンクの数に対する相対的な割合を算出させ、一の資源位置指定子に対する第一のリンクと第二のリンクのうち、相対的な割合が小さいほうのリンクを削除させ、同義語候補生成処理で、リンク削除処理で削除する対象から除かれた各資源位置指定子のリンクのアンカーテキストを用いて入力語の同義語候補を生成させることを特徴とする。

20

30

【発明の効果】

【0032】

本発明によれば、生成される同義語候補の精度を向上させて辞書作成を支援できる。

【図面の簡単な説明】

40

【0033】

【図1】本発明の第1の実施形態における辞書作成支援装置の例を示すブロック図である。

【図2】第1の実施形態における動作の例を示すフローチャートである。

【図3】リンク情報の例を示す説明図である。

【図4】リンク情報を削除する処理の例を示す説明図である。

【図5】リンク情報が削除された後の状態の例を示す説明図である。

【図6】削除されたリンク情報をもとに同義性の値を算出する過程を示す説明図である。

【図7】本発明の第2の実施形態における辞書作成支援装置の例を示すブロック図である。

50

【図 8】第 2 の実施形態における動作の例を示すフローチャートである。

【図 9】URL とアンカーテキストとのリンク情報を示す説明図である。

【図 10】リンク情報をもとに同義性の値を算出する過程を示す説明図である。

【図 11】本発明の第 3 の実施形態における辞書作成支援装置の例を示すブロック図である。

【図 12】第 3 の実施形態における動作の例を示すフローチャートである。

【図 13】URL とアンカーテキストとのリンク情報を示す説明図である。

【図 14】同義語候補を統合した際の URL とアンカーテキストとのリンク情報の例を示す説明図である。

【図 15】本発明による辞書作成支援装置の最小構成の例を示すブロック図である。

10

【図 16】URL とアンカーテキストとのリンク情報を示す説明図である。

【図 17】リンク情報をもとに同義性の値を算出する過程を示す説明図である。

【発明を実施するための形態】

【0034】

以下、本発明の実施形態を図面を参照して説明する。

【0035】

実施形態 1 .

図 1 は、本発明の第 1 の実施形態における辞書作成支援装置の例を示すブロック図である。本実施形態における辞書作成支援システムは、データ処理部 1 と、記憶部 2 とを備えている。なお、データ処理部 1 と、記憶部 2 とは、それぞれが独立の装置であってもよい。

20

【0036】

記憶部 2 は、リンク情報記憶部 20 と、関連リンク情報記憶部 21 とを備えている。リンク情報記憶部 20 は、リンク情報を記憶する。リンク情報には、アンカーテキストと、そのアンカーテキストでリンクする URL と、URL に対するアンカーテキストを持つリンクの数とを対応付けた情報が含まれる。

【0037】

関連リンク情報記憶部 21 は、リンク情報抽出手段 11 が抽出したリンク情報を格納する。すなわち、関連リンク情報記憶部 21 が記憶するリンク情報は、リンク情報記憶部 20 が記憶するリンク情報のサブセットである。

30

【0038】

リンク情報記憶部 20 及び関連リンク情報記憶部 21 は、記憶部 2 が備える磁気ディスク等によって実現される。

【0039】

データ処理部 1 は、入力語保持手段 10 と、リンク情報抽出手段 11 と、実体確率計算手段 12 と、同義性計算手段 13 と、表示指示手段 14 と、判定結果保持手段 15 と、リンク情報削除手段 16 とを備えている。

【0040】

入力語保持手段 10 は、利用者からキーボードなどの入力装置（図示せず）を介し、同義語を生成する対象の語として入力された入力語を記憶する。

40

【0041】

リンク情報抽出手段 11 は、リンク情報記憶部 20 に記憶されたリンク情報の中から、入力語に関するリンク情報と、入力語がアンカーテキストとして出現する URL に一回以上出現するアンカーテキストに関するリンク情報を抽出し、関連リンク情報記憶部 21 に記憶させる。すなわち、リンク情報抽出手段 11 は、入力語が含まれるアンカーテキストがリンクする URL が存在する場合、その URL にリンクする他のアンカーテキストのリンク情報も抽出し、関連リンク情報記憶部 21 に記憶させる。なお、URL は、Web 上の資源である Web ページを識別する識別子であることから、資源位置指定子と呼ぶことができる。

【0042】

50

実体確率計算手段 1 2 は、ある URL u へリンクするアンカーテキスト x が URL u の実体を表す確率である実体確率 $P(x|u)$ を、関連リンク情報記憶部 2 1 に記憶されたリンク情報ごとに計算する。すなわち、実体確率計算手段 1 2 は、URL とアンカーテキストとのリンクごとに実体確率を計算する。

【 0 0 4 3 】

同義性計算手段 1 3 は、入力語と関連リンク情報記憶部 2 1 に記憶されたアンカーテキスト間の同義性 $N R e l(s, t)$ を実体確率 $P(x|u)$ に基づいて計算する。

【 0 0 4 4 】

表示指示手段 1 4 は、同義性計算手段 1 3 が計算した同義性の上位を同義語候補としてディスプレイ装置などの出力装置（図示せず）に送信し、出力させる。

10

【 0 0 4 5 】

以上のように、リンク情報抽出手段 1 1、実体確率計算手段 1 2、同義性計算手段 1 3 及び表示指示手段 1 4 は、アンカーテキストを用いて入力語の同義語候補を生成する。なお、アンカーテキストを用いて入力語の同義語候補を生成する方法は、非特許文献 1 に記載された方法であってもよい。ただし、アンカーテキストを用いて入力語の同義語候補を生成する方法は、非特許文献 1 に記載された方法に限定されない。アンカーテキストを用いて入力語の同義語候補を生成する具体的な方法については後述する。

【 0 0 4 6 】

判定結果保持手段 1 5 は、同義語候補が入力語の同義語か非同義語かについて判定された結果を記憶する。判定結果保持手段 1 5 は、同義語候補を入力語の同義語もしくは非同義語と判定することを規定したルールに基づいて同義語判定手段（図示せず）が判定した結果を記憶してもよい。もしくは、判定結果保持手段 1 5 は、表示指示手段 1 4 が出力装置（図示せず）に出力させた同義語候補に対して利用者が同義語か非同義語かを判定した結果を記憶してもよい。

20

【 0 0 4 7 】

リンク情報削除手段 1 6 は、関連リンク情報記憶部 2 1 に格納されたリンク情報の中で、ある Web ページの URL に対して入力語および同義語であると判断された同義語候補がそれぞれアンカーテキストであるリンク情報と、その URL に対して非同義語であると判断された同義語候補がアンカーテキストであるリンク情報のうち、いずれかのリンク情報を、上記 URL を指すそれぞれのリンクの数に基づいて削除する。

30

【 0 0 4 8 】

すなわち、リンク情報削除手段 1 6 は、ある URL に対し、アンカーテキストが入力語もしくは同義語候補の中でその入力語の同義語と判定された同義語候補になっているリンクの情報（以下、同義語等に関するリンク情報と記す。）を抽出する。また、リンク情報削除手段 1 6 は、上記 URL に対し、アンカーテキストが同義語候補の中で入力語の非同義語と判定された同義語候補になっているリンクの情報（以下、非同義語に関するリンク情報と記す。）を抽出する。そして、リンク情報削除手段 1 6 は、同義語等に関するリンク情報と非同義語に関するリンク情報の少なくとも一方のリンク情報を削除する。このとき、リンク情報削除手段 1 6 は、上記 URL に対する各アンカーテキストを用いたリンクの数に基づいてリンク情報を削除する。

40

【 0 0 4 9 】

言い換えると、リンク情報削除手段 1 6 は、ある URL に対してあるアンカーテキストを持つリンクの数（ある URL に対してあるアンカーテキストでリンクする回数）に基づいてリンクを削除することができる。そして、リンク情報抽出手段 1 1、実体確率計算手段 1 2、同義性計算手段 1 3 及び表示指示手段 1 4 は、各 URL に対して残された（すなわち、削除対象から除かれた）リンクのアンカーテキストを用いて入力語の同義語候補を生成する。

【 0 0 5 0 】

入力語保持手段 1 0 及び判定結果保持手段 1 5 は、データ処理部 1 が備える磁気ディスク等によって実現される。

50

【 0 0 5 1 】

また、リンク情報抽出手段 1 1 と、実体確率計算手段 1 2 と、同義性計算手段 1 3 と、表示指示手段 1 4 と、リンク情報削除手段 1 6 とは、プログラム（辞書作成支援プログラム）に従って動作するコンピュータの CPU によって実現される。例えば、プログラムは、データ処理部 1 の記憶部（図示せず）に記憶され、CPU は、そのプログラムを読み込み、プログラムに従って、リンク情報抽出手段 1 1、実体確率計算手段 1 2、同義性計算手段 1 3、表示指示手段 1 4 及びリンク情報削除手段 1 6 として動作してもよい。また、リンク情報抽出手段 1 1 と、実体確率計算手段 1 2 と、同義性計算手段 1 3 と、表示指示手段 1 4 と、リンク情報削除手段 1 6 とは、それぞれが専用のハードウェアで実現されていてもよい。

10

【 0 0 5 2 】

次に、動作について説明する。図 2 は、第 1 の実施形態における動作の例を示すフローチャートである。

【 0 0 5 3 】

まず、入力語保持手段 1 0 は、利用者から入力装置（図示せず）を介して入力された入力語を記憶する（ステップ S 1）。本実施形態では、入力語保持手段 1 0 は、入力語として「NEC」を保持しているものとする。

【 0 0 5 4 】

次に、リンク情報抽出手段 1 1 は、リンク情報記憶部 2 0 に記憶されたリンク情報の中で、入力語に関するリンク情報と、入力語がアンカーテキストとして出現する URL に一回以上出現するアンカーテキストに関するリンク情報を抽出し、関連リンク情報記憶部 2 1 に記憶させる（ステップ S 2）。

20

【 0 0 5 5 】

図 3 は、リンク情報の例を示す説明図である。リンク情報とは、どのアンカーテキストがどの URL に何回リンクしたかを示す情報であり、一つのリンク情報には、図 3 に例示するように、「アンカーテキスト」項目、「URL」項目、「リンク回数」項目が含まれる。なお、「アンカーテキスト x に関するリンク情報」とは、「アンカーテキスト」項目が、アンカーテキスト x と一致するリンク情報を表す。

【 0 0 5 6 】

本実施形態では、リンク情報抽出手段 1 1 がリンク情報記憶部 2 0 に記憶されたリンク情報の中から、入力語「NEC」に関するリンク情報及び入力語がアンカーテキストとして出現する URL に一回以上出現するアンカーテキストに関するリンク情報として、図 3 に例示する情報を抽出し、関連リンク情報記憶部 2 1 に記憶させたものとする。また、図 3 に例示する表の内容をグラフ化したものが、図 1 6 に例示するグラフである。以下の説明では、図 1 6 に例示する内容をリンク情報として用いるものとする。

30

【 0 0 5 7 】

次に、実体確率計算手段 1 2 は、ある URL u にリンクするアンカーテキスト x が URL u の実体を表す確率である実体確率 $P(x|u)$ を、関連リンク情報記憶部 2 1 に格納されたリンク情報ごとに計算する（ステップ S 3）。各アンカーテキストから URL u に対するリンクの総数を L_{u} 、URL u に対するアンカーテキスト x を持つリンクの数（すなわち、アンカーテキスト x で URL u にリンクする数）を $L_{u, x}$ としたとき、実体確率 $P(x|u)$ は、以下の式 6 によって算出される。

40

【 0 0 5 8 】

$$P(x|u) = L_{u, x} / L_{u} \quad (\text{式 6})$$

【 0 0 5 9 】

ここで、 $P(x|u)$ は、「実体を表すアンカーテキストほど多くの人が用いる」ことを前提とした確率である。図 1 6 に示す例では、「P」の下部に示す数値が実体確率の計算結果である。このように、実体確率計算手段 1 2 は、各アンカーテキストから URL u へのリンクの総数（ L_{u} ）に対するアンカーテキスト x から URL u へのリンクの数（ $L_{u, x}$ ）の割合を実体確率として算出する。

50

【 0 0 6 0 】

次に、同義性計算手段 1 3 は、入力語と関連リンク情報記憶部 2 1 に格納されたアンカーテキスト間の同義性 $NRel(s, t)$ を実体確率に基づいて計算する (ステップ S 4)。アンカーテキスト s とアンカーテキスト t との間の同義性 $NRel(s, t)$ は下記の式 7 により算出される。なお、 i により総和を求める範囲 (すなわち、 i の取りうる範囲) は、アンカーテキスト s あるいはアンカーテキスト t がアンカーテキストとして出現する URL の数である。

【 0 0 6 1 】

【 数 2 】

$$NRel(s, t) = \frac{\sum_i [P(s|u_i) * P(t|u_i)]}{\sum_i [P(s|u_i) + P(t|u_i) - P(s|u_i) * P(t|u_i)]} \quad (\text{式 7})$$

10

【 0 0 6 2 】

図 1 6 に例示するグラフ (すなわち、図 3 に例示する関連リンク情報記憶部 2 1 のリンク情報) の内容に対し、同義性計算手段 1 3 が式 7 に基づいて同義性 $NRel(s, t)$ を計算した結果が図 1 7 に例示する内容である。図 1 7 に示す例では、同義性計算手段 1 3 が式 7 を用いて、「NEC」と「日本電気」、「NEC」と「BIGLOBE」及び「NEC」と「ビッグロブ」の同義性を算出していることを示す。

【 0 0 6 3 】

次に、表示指示手段 1 4 は、同義性計算手段 1 3 が計算した同義性を示す値の上位を同義語候補として出力装置 (図示せず) に送信し、表示させる (ステップ S 5)。例えば、アンカーテキスト s とアンカーテキスト t とが同義であると判断する時における同義性 $NRel(s, t)$ の値の閾値を 0.01 に設定したとする。この場合、表示指示手段 1 4 は、図 1 7 に例示する同義性の計算結果に対して、「日本電気」、「BIGLOBE」及び「ビッグロブ」を同義語候補として出力装置 (図示せず) に送信する。

20

【 0 0 6 4 】

上記説明では、表示指示手段 1 4 は、同義性 $NRel(s, t)$ が示す値と比較する閾値を設定して、同義語候補を出力する場合について説明した。他にも、表示指示手段 1 4 は、同義性を示す値の上位 n 件もしくは同義性を示す値の上位 $n\%$ などの指標を用いて同義語候補を出力してもよい。

30

【 0 0 6 5 】

次に、判定結果保持手段 1 5 は、同義語候補が同義語か非同義語かを利用者が判断した判断結果を入力装置 (図示せず) を介して受け取り、その判断結果を記憶する (ステップ S 6)。

【 0 0 6 6 】

本実施形態では、利用者に提示した同義語候補のうち、入力語「NEC」に対する同義語として「日本電気」が、非同義語として「BIGLOBE」がそれぞれ、判断結果として入力装置 (図示せず) に入力されたものとする。

【 0 0 6 7 】

次に、リンク情報削除手段 1 6 は、関連リンク情報記憶部 2 1 に記憶されたリンク情報の中で、以下のリンク情報のうちのいずれかを、URL を指すそれぞれのリンクの数に基づいて削除する。一つは、ある Web ページの URL に対して、アンカーテキストが入力語もしくはその入力語の同義語であると判断された同義語候補になっているリンク情報 (すなわち、同義語等に関するリンク情報) であり、もう一つは、その URL に対して、入力語の非同義語であると判断された同義語候補がアンカーテキストであるリンク情報 (すなわち、非同義語に関するリンク情報) である。

40

【 0 0 6 8 】

例えば、図 1 6 に例示するリンク情報には、URL 「biglobe.ne.jp」に対するアンカーテキストとして、入力語である「NEC」と、同義語候補である「日本電

50

気」、「BIGLOBE」及び「ビッグロブ」が存在する。このうち、「NEC」は入力語であり、「日本電気」は同義語、「BIGLOBE」は非同義語と判断されている。

【0069】

このとき、URL「biglobe.ne.jp」に対する「NEC」および「日本電気」のアンカーテキストと、「BIGLOBE」のアンカーテキストの少なくともいずれか一方は、URL「biglobe.ne.jp」の実体を表していないと言える。「NEC」と「BIGLOBE」とは非同義語の関係にあり、両者は別の対象物を示すはずだからである。

【0070】

リンク情報削除手段16は、削除の判断基準に実体確率を用いる。具体的には、リンク情報削除手段16は、実体確率が小さいアンカーテキストに関するリンク情報を削除する。上記の例では、URL「biglobe.ne.jp」に着目した場合、実体確率は、それぞれ、 $P(\text{BIGLOBE} | \text{biglobe.ne.jp}) = 0.33$ 、 $P(\text{日本電気} | \text{biglobe.ne.jp}) = 0.17$ 、 $P(\text{NEC} | \text{biglobe.ne.jp}) = 0.17$ と算出される。よって、リンク情報削除手段16は、実体確率が小さいURL「biglobe.ne.jp」に対する「日本電気」「NEC」に関するリンク情報を削除する。

【0071】

同様に、リンク情報削除手段16は、URL「nec.jp」に対する「NEC」に関するリンク情報と、「BIGLOBE」に関するリンク情報のうち、実体確率の小さい「BIGLOBE」に関するリンク情報を削除する。

【0072】

なお、リンク情報削除手段16がリンク情報を削除するのは、同じURLに対して、上記の2種類のリンク情報（すなわち、同義語等に関するリンク情報と非同義語に関するリンク情報）の双方が存在する場合のみである。例えば、「nec.co.jp」に対しては入力語「NEC」と同義語「日本電気」のリンクは存在するが、非同義語「BIGLOBE」のリンクは存在しない。そのため、リンク情報削除手段16は、「nec.co.jp」に関するリンク情報を削除の対象としない。

【0073】

また、リンク情報削除手段16が、アンカーテキストが入力語であるリンク情報を削除する場合、そのリンク情報のURLは指定された入力語に関する同義語候補抽出では利用されなくなる。これは、入力語に関する同義語候補を抽出するタスクにおいて実質的にURLを削除したことと同じ意味を持つ。

【0074】

図4は、リンク情報削除手段16がリンク情報を削除する処理の例を示す説明図である。また、図5は、リンク情報が削除された後の状態の例を示す説明図である。図4に例示する点線部が、リンク情報削除手段16により削除される対象になるリンク情報である。その他の内容は、図16に記載した内容と同様である。以上の処理により、図4に例示する点線部のリンク情報が削除され、その結果、関連リンク情報記憶部21には、図5に例示するリンク情報が残ることになる。

【0075】

この処理の後、ステップS3に戻り、以降の処理を繰り返す。リンク情報削除手段16がリンク情報を削除した結果をもとに実体確率計算手段12が算出した実体確率が、図5に例示する「P」の下部に示す数値である。その他の内容は、図16に記載した内容と同様である。図5に示す例では、 $P(\text{NEC} | \text{biglobe.ne.jp}) = 0$ になり、図4に例示する1回目に算出された実体確率の値である0.17よりも減少していることが分かる。同様に、図5に示す例では、 $P(\text{NEC} | \text{nec.co.jp}) = 0.5$ になり、図4に例示する1回目に算出された実体確率の値である0.4よりも増加していることが分かる。この例からも、リンク情報の削除によって、実体確率をより正確に推定できるようになったと言える。

10

20

30

40

50

【 0 0 7 6 】

なお、本実施形態において算出した実体確率は、一つのURLに対して実体を表す語が一つであることを前提としている。したがって、同義語が存在する場合は、実体を表すアンカーテキストであってもその実体確率は1にならない。例えば、 $P(NEC | nec.co.jp)$ の理論上の最大値は0.5になる。したがって、リンク削除手段により実体確率を理論上の最大値にまで上昇できたことになる。

【 0 0 7 7 】

図6は、削除されたリンク情報をもとに同義性の値を算出する過程を示す説明図である。図6に示す例では、同義性計算手段13が式7を用いて同義性の値を算出する2回目の過程を示している。図17に例示する1回目の計算結果と比較すると、「NEC」と「日本電気」の同義性が、1回目の0.26から0.33に増加していることがわかる。また、「NEC」と「BIGLOBE」及び「NEC」と「ビッグロブ」の同義性が、1回目の0.07及び0.03から、それぞれ0に減少していることがわかる。

10

【 0 0 7 8 】

以上のことから、実体確率計算手段12が実体確率を正確に推定できたことにより、同義性計算手段13が同義性の値をより正確に算出できたと言える。

【 0 0 7 9 】

また、同義語候補を表示する際の閾値を1回目と同様に0.01に設定したとする。この場合、表示指示手段14は、2回目の判断結果として、同義語候補「日本電気」のみを表示することになる。このことから、提示する同義語候補の精度がリンク情報の削除によって向上することがわかる。

20

【 0 0 8 0 】

以上、第1の実施形態の動作を説明した。なお、上記説明では、判定結果保持手段15が、入力装置を通じて利用者から入力された同義語候補に対する判断結果を記憶する場合について説明した。それ以外の方法として、判定結果保持手段15が、あらかじめ部分的な同義語辞書と非同義語辞書を記憶しておき、リンク情報削除手段16は、同義語候補をそれらの辞書にあてはめて、同義語及び非同義語を判断してもよい。

【 0 0 8 1 】

また、判定結果保持手段15が、利用者から入力された判定結果を記憶する代わりに、同義性計算手段13が、計算した同義性の値の上位を同義語、同義性の値の下位を非同義語とする判断結果を判定結果保持手段15に記憶させてもよい。なお、同義性計算手段13は、この場合の上位及び下位の判断を、予め定められた閾値や範囲に基づいて行えばよい。この場合、同義性計算手段13の初期の計算結果がある程度正しければ、利用者には負担をかけずに、実体確率の推定精度を向上させることが可能になる。

30

【 0 0 8 2 】

また、リンク情報削除手段16は、2種類のリンク情報の中で実体確率が小さい方のアンカーテキストに関するリンク情報を削除する代わりに、単純に $L_{u,x}$ (u に対する x を持つリンクの数)が小さい方のアンカーテキストに関するリンク情報を削除してもよい。

【 0 0 8 3 】

式6に例示する実体確率の算出式「実体確率 $P(x|u) = L_{u,x} / L_u$ 」における L_u の値は、比較する両者で同一となる。すなわち、実体確率は、リンクの数に基づいて算出される値であり、実質的には $L_{u,x}$ で両者を比較しているのと同じだからである。

40

【 0 0 8 4 】

また、上記説明では、リンク情報削除手段16が、2種類のリンク情報(すなわち、同義語等に関するリンク情報と非同義語に関するリンク情報)のうち、実体確率が小さい方のアンカーテキストに関するリンク情報を削除する場合について説明した。その代わりに、リンク情報削除手段16は、実体確率が閾値以下のアンカーテキストを削除するようにしても良い。この場合、実体確率が閾値以下であれば、両方のアンカーテキストに関する

50

リンク情報が削除される場合も、両方が削除されない場合もありうることになる。

【 0 0 8 5 】

また、同義性計算手段 1 3 が同義性を計算する方法は、式 7 に例示した計算式を利用する場合に限定されない。同義性計算手段 1 3 は、例えば、正規化部分を用いない式 2 に例示する $Rel(s, t)$ の計算式を利用して同義性を計算してもよい。また、同義性計算手段 1 3 は、以下の式 8 に示す計算式を用いて同義性を計算してもよい。

【 0 0 8 6 】

$$Rel2(s, t) = Ave(i(L_{u_i}, s) / L_s, i(L_{u_i}, t) / L_t) \quad (式 8)$$

10

【 0 0 8 7 】

ここで、 $L_{u_i, x}$ は u_i に対するアンカーテキスト x を持つリンクの数 (u_i に対して x でリンクする数) を表し、 L_x はアンカーテキスト x を持つリンクの総数を表す。 i により総和を求める範囲 (すなわち、 i の取りうる範囲) は、アンカーテキスト s とアンカーテキスト t の両方がアンカーテキストとして出現する URL の数である。また、 Ave は 2 つの値の平均を表す。この平均は、相加平均、相乗平均及び調和平均のいずれであってもよい。

【 0 0 8 8 】

また、本実施形態では、リンク情報削除手段 1 6 が、リンク情報を削除すると判断した場合に関連リンク情報記憶部 2 1 のリンク情報を削除する場合について説明した。加えて、リンク情報削除手段 1 6 は、対応するリンク情報記憶部 2 0 のリンク情報を削除してもよい。このようにすることで、リンク情報を削除したことを、他の入力語に対しても反映することが可能になる。

20

【 0 0 8 9 】

また、上記説明では、リンク情報抽出手段 1 1 が、リンク情報記憶部 2 0 に記憶されたリンク情報の中から、入力語に関するリンク情報と、入力語がアンカーテキストとして出現する URL に一回以上出現するアンカーテキストに関するリンク情報を抽出し、関連リンク情報記憶部 2 1 に記憶させる場合について説明した。ただし、リンク情報抽出手段 1 1 は、上記リンク情報の中で、入力語または同義語と判断されたアンカーテキストが最も多く出現する URL (あるいは最も多いものから N 個の URL) に関するリンク情報のみを抽出し、関連リンク情報記憶部 2 1 に記憶させてもよい。この方法では、各アンカーテキストに関する公式ページに近い URL のみを採用するため、同義語候補の精度が向上する可能性がある。

30

【 0 0 9 0 】

以上のように、本実施形態によれば、リンク情報抽出手段 1 1、実体確率計算手段 1 2、同義性計算手段 1 3 及び表示指示手段 1 4 が、アンカーテキストを用いて入力語の同義語候補を生成する。リンク情報削除手段 1 6 は、ある URL に対して、アンカーテキストが入力語もしくは同義語候補の中でその入力語の同義語と判定された同義語候補になっているリンクの情報 (すなわち、同義語等に関するリンク情報) を抽出する。また、リンク情報削除手段 1 6 は、上記 URL に対して、アンカーテキストが同義語候補の中で入力語の非同義語と判定された同義語候補になっているリンクの情報 (すなわち、非同義語に関するリンク情報) を抽出する。そして、リンク情報削除手段 1 6 は、同義語等に関するリンク情報と非同義語に関するリンク情報のうちの少なくとも一方のリンクを削除する。このとき、リンク情報削除手段 1 6 は、上記 URL に対するアンカーテキストごとのリンクの数に基づいて、その URL とアンカーテキストとのリンクの情報を削除する。そして、リンク情報抽出手段 1 1、実体確率計算手段 1 2、同義性計算手段 1 3 及び表示指示手段 1 4 が、削除対象から除かれたリンクのアンカーテキストを用いて入力語の同義語候補を生成する。このような構成により、生成される同義語候補の精度を向上させて辞書作成を支援できる。

40

【 0 0 9 1 】

50

具体的には、本実施形態では、推定された実体確率をもとに同義性を判断することで、同義性候補の精度を向上させている。したがって、同義性を判断するためには、実体確率を正しく推定することが必要になる。本実施形態では、リンク情報削除手段16が、例えば、ユーザの判断結果に基づいてURLの実体を表さないアンカーテキストに関するリンク情報を削除することで、実体確率を0に減少させている。また、URLの実体を表さないアンカーテキストに関するリンク情報を削除することで、URLの実体を表す正しいアンカーテキストに関する実体確率を1に近づけることが可能になる。このように、不要なリンク情報を削除して実体確率の推定精度を向上させることで、同義語候補の精度を向上させることができる。

【0092】

実施形態2.

図7は、本発明の第2の実施形態における辞書作成支援装置の例を示すブロック図である。なお、第1の実施形態と同様の構成については、図1と同一の符号を付し、説明を省略する。本実施形態における辞書作成支援システムも、データ処理部1と、記憶部2とを備えている。データ処理部1と、記憶部2とは、それぞれが独立の装置であってもよい。

【0093】

記憶部2は、リンク情報記憶部20と、関連リンク情報記憶部21とを備えている。リンク情報記憶部20及び関連リンク情報記憶部21の構成は、第1の実施形態と同様である。

【0094】

データ処理部1は、入力語保持手段10と、リンク情報抽出手段11と、改良実体確率計算手段17と、同義性計算手段13と、表示指示手段14と、判定結果保持手段15と、リンク情報削除手段16とを備えている。すなわち、実体確率計算手段12の代わりに改良実体確率計算手段17を備えている点で、第1の実施形態と異なる。それ以外の構成については、第1の実施形態と同様である。

【0095】

改良実体確率計算手段17は、あるURL u にリンクするあるアンカーテキスト x がURL u の実体を表す確率である実体確率を関連リンク情報記憶部21に格納されたリンク情報ごとに計算する。このとき、改良実体確率計算手段17は、あるURL u へのリンクの総数に対するそのURL u へのアンカーテキスト x を持つリンクの数の割合に加え、そのアンカーテキスト x を持つリンクの総数に対するそのURL u へのそのアンカーテキスト x を持つリンクの数の割合を用いて、実体確率を計算する。

【0096】

すなわち、改良実体確率計算手段17は、まず、あるURL u へのリンクの総数に対する、あるアンカーテキスト x から上記URL u へのリンクの数の割合を計算する。さらに、改良実体確率計算手段17は、上記アンカーテキスト x を持つ各URLへのリンクの総数に対する、上記アンカーテキストから上記URL u へのリンクの数の割合を計算する。改良実体確率計算手段17は、このように算出した2つの割合を用いて実体確率を計算する。例えば、改良実体確率計算手段17は、この2つの割合を乗じた値を実体確率として計算してもよい。

【0097】

なお、リンク情報抽出手段11と、改良実体確率計算手段17と、同義性計算手段13と、表示指示手段14と、リンク情報削除手段16とは、プログラム(辞書作成支援プログラム)に従って動作するコンピュータのCPUによって実現される。また、リンク情報抽出手段11と、改良実体確率計算手段17と、同義性計算手段13と、表示指示手段14と、リンク情報削除手段16とは、それぞれが専用のハードウェアで実現されていてもよい。

【0098】

次に、動作について説明する。図8は、第2の実施形態における動作の例を示すフローチャートである。リンク情報抽出手段11がリンク情報を抽出して、関連リンク情報記憶

10

20

30

40

50

部 2 1 に記憶させるステップ S 1 ~ ステップ S 2 までの処理、及び、同義性計算手段 1 3 が実体確率に基づいて同義性を示す値を算出してから、リンク情報削除手段 1 6 がリンク情報を削除するステップ S 4 ~ ステップ S 7 までの処理は、図 2 に例示する第 1 の実施形態における処理と同様である。

【 0 0 9 9 】

改良実体確率計算手段 1 7 は、ある URL u にリンクするあるアンカーテキスト x が URL u の実体を表す確率である実体確率 $P(x | u)$ を、関連リンク情報記憶部 2 1 に記憶されたリンク情報ごとに求める (ステップ S 3 a)。具体的には、改良実体確率計算手段 1 7 は、上記 URL u に対するリンクの総数におけるその URL u に対してアンカーテキスト x を持つリンクの数の割合に加え、そのアンカーテキスト x を持つリンクの総数における上記 URL に対して上記アンカーテキスト x を持つリンクの数の割合を用いて実体確率を計算する。

10

【 0 1 0 0 】

ここで、各アンカーテキストから URL u に対するリンクの総数を $L_{u, x}$ 、アンカーテキスト x を持つ各 URL へのリンクの総数を $L_{u, x}$ 、URL u に対するアンカーテキスト x を持つリンクの数 (アンカーテキスト x で URL u にリンクする数) を $L_{u, x}$ としたとき、改良実体確率計算手段 1 7 は、以下の式 9 を用いて実体確率 $P(x | u)$ を算出する。

【 0 1 0 1 】

$$P(x | u) = (L_{u, x} / L_{u, x}) * (L_{u, x} / L_{u, x}) \quad (\text{式 9})$$

20

【 0 1 0 2 】

図 9 は、URL とアンカーテキストとのリンク情報を示す説明図である。図 9 に例示する「P」の下部に示された値は、図 3 に例示する関連リンク情報記憶部 2 1 のリンク情報をもとに式 9 を用いて計算されたリンクごとの実体確率である。それ以外については、図 1 7 に示す内容と同様である。

【 0 1 0 3 】

例えば、図 9 に示す例では、 $P(NEC | biglobe.ne.jp) = 0.03$ 、 $P(NEC | nec.jp) = 0.2$ であり、両者の比は 6.7 である。一方、第 1 の実施形態における方法で実体確率を算出した場合、上記値は、図 1 6 に例示するように、 $P(NEC | biglobe.ne.jp) = 0.17$ 、 $P(NEC | nec.jp) = 0.4$ であり両者の比は 2.35 である。

30

【 0 1 0 4 】

第 1 の実施形態における方法で算出した実体確率に比べ、第 2 の実施形態における方法で算出した実体確率は、 $P(NEC | biglobe.ne.jp)$ の値と $P(NEC | nec.jp)$ の値の両方とも減少している。しかし、 $P(NEC | biglobe.ne.jp)$ と $P(NEC | nec.jp)$ との比の値を比較した場合、第 2 の実施形態における方法で算出した実体確率の比の値がより大きくなっている。このことから、上記方法によれば、URL の実体を表わさないアンカーテキストの実体確率をより減少できることが分かる。

【 0 1 0 5 】

40

図 1 0 は、図 9 に例示する実体確率を用いて同義性を示す値を計算する過程を示す説明図である。図 1 0 に示す例では、改良実体確率計算手段 1 7 が算出した実体確率を用いて、同義性計算手段 1 3 が、「NEC」と「日本電気」、「NEC」と「BIGLOBE」及び「NEC」と「ビッグロブ」の同義性を式 7 を用いて算出していることを示す。

【 0 1 0 6 】

ここで、「NEC」と「日本電気」は同義語であり、「NEC」と「BIGLOBE」及び「NEC」と「ビッグロブ」は非同義語であるとする。このとき、同義性計算手段 1 3 は、同義語の同義性 $NRel(NEC, 日本電気) = 0.10$ 、非同義語の同義性 $NRel(NEC, BIGLOBE) = 0.01$ 、 $NRel(NEC, ビッグロブ) = 0.01$ と算出する。この場合、同義語と非同義語との間の同義性の値の比の平均は 10 で

50

ある。

【 0 1 0 7 】

一方、図 1 7 に例示する算出結果によれば、第 1 の実施形態において算出される同義語と非同義語との間の同義性の値の比の平均は 6 . 2 である。以上の具体例からも、第 2 の実施形態における方法で算出される実体確率を用いることで、非同義語の同義性を相対的に減少させられることがわかる。

【 0 1 0 8 】

以上、第 2 の実施形態の動作を説明した。なお、上記説明では、改良実体確率計算手段 1 7 が、式 9 に例示する算出式を用いて実体確率を算出する方法について説明した。他にも、改良実体確率計算手段 1 7 は、式 9 に例示する算出式の一部を用いた下記の式 1 0 を用いて実体確率を計算してもよい。

【 0 1 0 9 】

$$P(x|u) = L_{u,x} / L_x \quad (\text{式 1 0})$$

【 0 1 1 0 】

以上のように、本実施形態によれば、改良実体確率計算手段 1 7 が、各アンカーテキストから URL u へのリンクの総数に対するアンカーテキスト x からその URL u へのリンクの数の割合を算出する。さらに、改良実体確率計算手段 1 7 が、アンカーテキスト x を持つ各 URL へのリンクの総数に対するそのアンカーテキスト x から上記 URL u へのリンクの数の割合を用いて実体確率を算出する。このような構成により、生成される同義語候補の精度を向上させて辞書作成を支援できる。

【 0 1 1 1 】

具体的には、第 1 の実施形態と同様、推定された実体確率 $P(x|u)$ をもとに同義性を判断することで、同義性候補の精度を向上させている。したがって、同義性を判断するためには、実体確率を正しく推定することが必要になる。本実施形態では、改良実体確率計算手段 1 7 が、アンカーテキスト x を持つリンクが URL u に対するリンクの中でどの程度多数派をしめるかを表す指標（例えば、第 1 の実施形態における式 6 で算出される実体確率）を算出する。さらに、改良実体確率計算手段 1 7 が、アンカーテキスト x が URL u のみにどの程度多くリンクするかを表す指標を算出して実体確率を計算する。したがって、アンカーテキストのリンク状況をより反映させた実体確率の値を推定できるため、同義語候補の精度を向上させることができる。

【 0 1 1 2 】

実施形態 3 .

図 1 1 は、本発明の第 3 の実施形態における辞書作成支援装置の例を示すブロック図である。なお、第 1 の実施形態と同様の構成については、図 1 と同一の符号を付し、説明を省略する。本実施形態における辞書作成支援システムも、データ処理部 1 と、記憶部 2 とを備えている。データ処理部 1 と、記憶部 2 とは、それぞれが独立の装置であってもよい。

【 0 1 1 3 】

記憶部 2 は、リンク情報記憶部 2 0 と、関連リンク情報記憶部 2 1 とを備えている。リンク情報記憶部 2 0 及び関連リンク情報記憶部 2 1 の構成は、第 1 の実施形態と同様である。

【 0 1 1 4 】

データ処理部 1 は、入力語保持手段 1 0 と、リンク情報抽出手段 1 1 と、実体確率計算手段 1 2 と、同義性計算手段 1 3 と、表示指示手段 1 4 と、判定結果保持手段 1 5 と、リンク情報削除手段 1 6 と、リンク情報統合手段 1 8 を備えている。すなわち、リンク情報統合手段 1 8 をさらに備えている点で、第 1 の実施形態と異なる。それ以外の構成については、第 1 の実施形態と同様である。

【 0 1 1 5 】

リンク情報統合手段 1 8 は、ある Web ページの URL に対して入力語および同義語であると判断された同義語候補がそれぞれアンカーテキストであるリンク情報を統合する。

ここで、リンク情報の統合とは、各同義語候補のアンカーテキストを同一のアンカーテキスト（以下、統合アンカーテキストと記す。）とみなし、各同義語候補のアンカーテキストでリンクしていたURLを、統合アンカーテキストでリンクするURLとみなしたリンク情報を生成することである。このとき、リンク情報におけるリンクの数も、統合アンカーテキストを持つリンクの数として集約する。

【0116】

すなわち、リンク情報統合手段18は、アンカーテキストが同義語候補になっている各URLに対するリンクのうち、入力語の同義語と判定された同義語候補がアンカーテキストであるリンク情報を統合する。したがって、リンク情報が統合された後、リンク情報抽出手段11、実体確率計算手段12、同義性計算手段13及び表示指示手段14は、統合されたリンクのアンカーテキストを用いて入力語の同義語候補を生成する。

10

【0117】

リンク情報抽出手段11と、実体確率計算手段12と、同義性計算手段13と、表示指示手段14と、リンク情報削除手段16と、リンク情報統合手段18とは、プログラム（辞書作成支援プログラム）に従って動作するコンピュータのCPUによって実現される。また、リンク情報抽出手段11と、実体確率計算手段12と、同義性計算手段13と、表示指示手段14と、リンク情報削除手段16と、リンク情報統合手段18とは、それぞれが専用のハードウェアで実現されていてもよい。

【0118】

次に、動作について説明する。図12は、第3の実施形態における動作の例を示すフローチャートである。リンク情報抽出手段11がリンク情報を抽出してから、リンク情報削除手段16がリンク情報を削除するステップS1～ステップS7までの処理は、図2に例示する第1の実施形態における処理と同様である。

20

【0119】

リンク情報統合手段18は、あるWebページのURLに対して、アンカーテキストが入力語もしくは入力語の同義語であると判断された同義語候補になっているリンク情報を統合する（ステップS8）。

【0120】

図13は、URLとアンカーテキストとのリンク情報を示す説明図である。以下、関連リンク情報記憶部21が、図13に例示するリンク情報を記憶している場合について説明する。図13に例示する「P」の下部に示された値は、図13に例示するリンク情報を用いて実体確率計算手段12が計算した実体確率である。それ以外については、図17に示す内容と同様である。

30

【0121】

図13に例示するように、アンカーテキスト「NEC」とアンカーテキスト「日電」とは、同一のURLにリンクしていない。よって、アンカーテキスト「日電」がリンクするURLに対する「NEC」の実体確率は0になるため（ $P(\text{NEC} | \text{nec.com.cn}) = 0$ ）、「NEC」と「日電」の同義性は0になる。したがって、このままでは「日電」を「NEC」の同義語候補にすることができない。

【0122】

ここで、表示指示手段14が「NEC」と「日本電気」とを同義語候補として出力し、利用者が、両者を同義語と判断したとする。このとき、リンク情報統合手段18は、「NEC」と「日本電気」に関するリンク情報を統合する。同義語候補を統合した際のURLとアンカーテキストとのリンク情報を図14に示す。

40

【0123】

この処理の後、ステップS3に戻り、以降の処理を繰り返す。リンク情報統合手段18がリンク情報を統合した結果をもとに実体確率計算手段12が算出した実体確率が、図14に例示する「P」の下部に示す数値である。図14に示す例では、「日電」がリンクするURLに対する「NEC」の実体確率は、 $P(\text{NEC} | \text{nec.com.cn}) = 0.23$ と算出される。このことから、図13に例示する1回目に算出された実体確率の値

50

0よりも増加していることが分かる。この実体確率を用いて同義性を示す値を計算することで、「NEC」と「日電」との同義性も向上する。よって、「日電」を「NEC」の同義語候補とすることができる。

【0124】

以上、第3の実施形態の動作を説明した。なお、上記説明では、リンク情報削除手段16がリンク情報を削除した後、リンク情報統合手段18がリンク情報を統合する場合について説明した。ただし、リンク情報を削除する処理と、リンク情報を統合する処理とが行われる順番は、上記順番に限定されない。リンク情報統合手段18がリンク情報を統合した後で、リンク情報削除手段16がリンク情報を削除してもよい。

【0125】

以上のように、本実施形態によれば、リンク情報統合手段18が、アンカーテキストが同義語候補になっている各URLに対するリンクのうち、入力語の同義語と判定された同義語候補がアンカーテキストであるリンクを統合することで、生成される同義語候補の精度を向上させて辞書作成を支援する。

【0126】

すなわち、本実施形態でも、推定された実体確率 $P(x|u)$ をもとに同義性を判断して、同義語候補の精度を向上させる。したがって、同義性を判断するためには、実体確率を正しく推定することが必要になる。本実施形態では、リンク情報統合手段18が、同義語と判断されたリンク情報を統合する。リンク情報を統合することにより、実体確率の値を推定するためのリンク情報を増加させることができる。このように、他のアンカーテキストのリンク情報も反映させた実体確率の値を推定できるため、同義語候補の精度をより向上させることができる。

【0127】

次に、本発明による辞書作成支援装置の最小構成の例を説明する。図15は、本発明による辞書作成支援装置の最小構成の例を示すブロック図である。本発明による辞書作成支援装置は、辞書作成を支援する辞書作成支援装置であって、Webページを識別する識別子である各資源位置指定子（例えば、URL）に対してリンクする文字列を表すアンカーテキストを用いて、同義語を生成する対象の語として辞書作成支援装置に入力される入力語の同義語候補を生成する同義語候補生成手段81（例えば、リンク情報抽出手段11、実体確率計算手段12、同義性計算手段13及び表示指示手段14）と、一の資源位置指定子に対して、アンカーテキストが入力語もしくは同義語候補の中でその入力語の同義語と判定された同義語候補になっているリンクである第一のリンク（例えば、同義語等に関するリンク情報）と、アンカーテキストが同義語候補の中で入力語の非同義語と判定された同義語候補になっているリンクである第二のリンク（例えば、非同義語に関するリンク情報）のうち、少なくとも一方のリンクを削除するリンク削除手段82（例えば、リンク情報削除手段16）とを備えている。

【0128】

リンク削除手段82は、一の資源位置指定子に対するアンカーテキストごとのリンクの数に基づいて、上記一の資源位置指定子とアンカーテキストとのリンクを削除し、同義語候補決定手段81は、リンク削除手段82が削除する対象から除かれたリンクのアンカーテキストを用いて入力語の同義語候補を生成する。

【0129】

そのような構成により、生成される同義語候補の精度を向上させて辞書作成を支援できる。

【0130】

なお、少なくとも以下に示すような辞書作成支援装置も、上記に示すいずれかの実施形態に開示されている。

【0131】

(1) 辞書作成を支援する辞書作成支援装置であって、Webページを識別する識別子である各資源位置指定子（例えば、URL）に対してリンクする文字列を表すアンカーテキ

10

20

30

40

50

ストを用いて、同義語を生成する対象の語として辞書作成支援装置に入力される入力語の同義語候補を生成する同義語候補生成手段（例えば、リンク情報抽出手段 1 1、実体確率計算手段 1 2、同義性計算手段 1 3 及び表示指示手段 1 4）と、一の資源位置指定子に対して、アンカーテキストが入力語もしくは同義語候補の中でその入力語の同義語と判定された同義語候補になっているリンクである第一のリンク（例えば、同義語等に関するリンク情報）と、アンカーテキストが同義語候補の中で入力語の非同義語と判定された同義語候補になっているリンクである第二のリンク（例えば、非同義語に関するリンク情報）のうち、少なくとも一方のリンクを削除するリンク削除手段（例えば、リンク情報削除手段 1 6）とを備え、リンク削除手段が、一の資源位置指定子に対するアンカーテキストごとのリンクの数に基づいて、上記一の資源位置指定子とアンカーテキストとのリンクを削除し、同義語候補決定手段が、リンク削除手段が削除する対象から除かれたリンクのアンカーテキストを用いて入力語の同義語候補を生成する辞書作成支援装置。

10

【 0 1 3 2 】

（ 2 ）資源位置指定子に対してリンクするアンカーテキストがその資源位置指定子の実体を表す確率である実体確率（例えば、 $P(x|u)$ ）を、資源位置指定子とアンカーテキストとのリンクごとに計算する実体確率計算手段（例えば、実体確率計算手段 1 2）を備え、リンク削除手段が、第一のリンクと第二のリンクのうち、実体確率が小さいリンクを削除する辞書作成支援装置。

【 0 1 3 3 】

（ 3 ）実体確率計算手段（例えば、実体確率計算手段 1 2）が、各アンカーテキストから一の資源位置指定子へのリンクの総数に対する一のアンカーテキストからその一の資源位置指定子へのリンクの数の割合を実体確率として算出する（例えば、式 6 を用いて算出する）辞書作成支援装置。

20

【 0 1 3 4 】

（ 4 ）実体確率計算手段（例えば、改良実体確率計算手段 1 7）は、各アンカーテキストから一の資源位置指定子へのリンクの総数に対する一のアンカーテキストからその一の資源位置指定子へのリンクの数の割合、及び、アンカーテキストを持つ各資源位置指定子へのリンクの総数に対するそのアンカーテキストから一の資源位置指定子へのリンクの数の割合を用いて実体確率を算出する（例えば、式 9 を用いて算出する）辞書作成支援装置。

30

【 0 1 3 5 】

（ 5 ）アンカーテキストが同義語候補になっている各資源位置指定子に対するリンクのうち、入力語の同義語と判定された同義語候補がアンカーテキストであるリンクを統合するリンク統合手段（例えば、リンク情報統合手段 1 8）を備え、同義語候補決定手段が、統合されたリンクのアンカーテキストを用いて入力語の同義語候補を生成する辞書作成支援装置。

【 0 1 3 6 】

（ 6 ）リンク削除手段が、第一のリンクと第二のリンクの双方が一の資源位置指定子に存在しない場合に、その一の資源位置指定子に対するアンカーテキストのリンクを削除対象から除く辞書作成支援装置。

【 産業上の利用可能性 】

40

【 0 1 3 7 】

本発明は、Webのアンカーテキストを用いて入力語に対する同義語候補を生成する辞書作成支援装置に好適に適用される。また、本発明による辞書作成支援装置で作成された同義語辞書は、文書検索、顧客データの名寄せなど、様々なソフトウェアの基本的な資源として使用可能である。

【 符号の説明 】

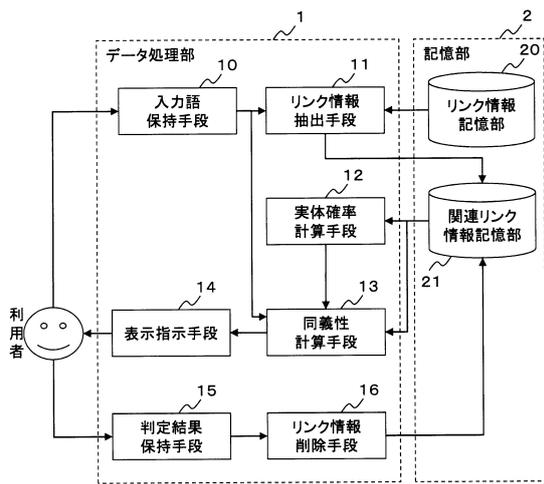
【 0 1 3 8 】

- 1 データ処理部
- 2 記憶部
- 1 0 入力語保持手段

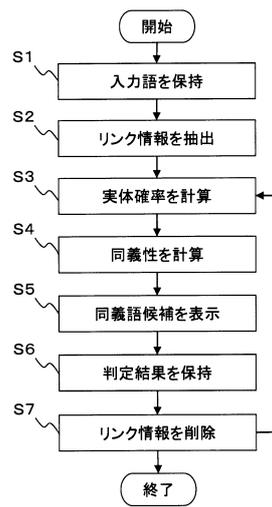
50

- 1 1 リンク情報抽出手段
- 1 2 実体確率計算手段
- 1 3 同義性計算手段
- 1 4 表示指示手段
- 1 5 判定結果保持手段
- 1 6 リンク情報削除手段
- 1 7 改良実体確率計算手段
- 1 8 リンク情報統合手段
- 2 0 リンク情報記憶部
- 2 1 関連リンク情報記憶部

【図 1】



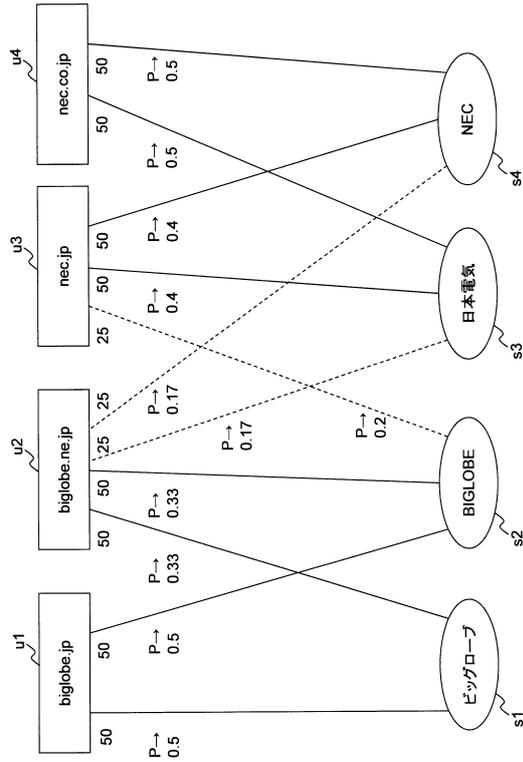
【図 2】



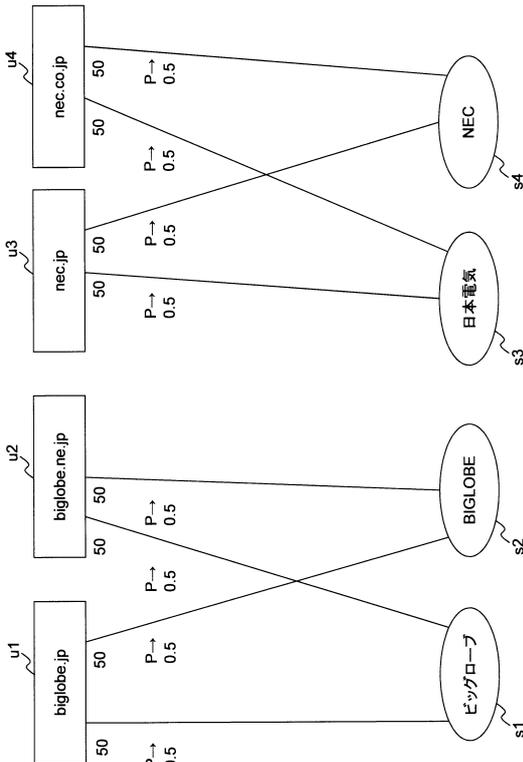
【図3】

アンカーテキスト	URL	リンク回数
ビッグロブ	biglobe.jp	50
ビッグロブ	biglobe.ne.jp	50
BIGLOBE	biglobe.jp	50
BIGLOBE	biglobe.ne.jp	50
BIGLOBE	nec.jp	25
日本電気	biglobe.ne.jp	25
日本電気	nec.jp	50
日本電気	nec.co.jp	50
NEC	biglobe.ne.jp	25
NEC	nec.jp	50
NEC	nec.co.jp	50

【図4】



【図5】



【図6】

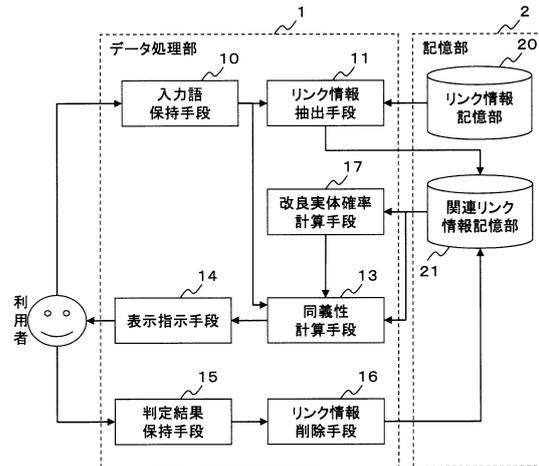
入力語: NEC 2回目

$$NRel(NEC, 日本電気) = \frac{0.5 \cdot 0.5 + 0.5 \cdot 0.5}{0.5 + 0.5 - 0.5 \cdot 0.5 + 0.5 + 0.5 - 0.5 \cdot 0.5} = 0.5 / 1.5 = 0.33$$

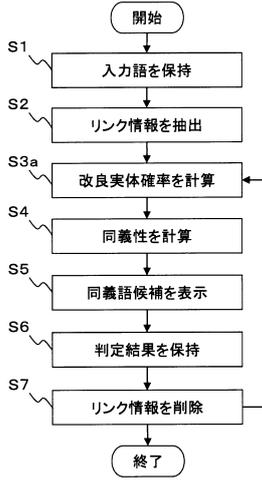
$$NRel(NEC, BIGLOBE) = 0$$

$$NRel(NEC, ビッグロブ) = 0$$

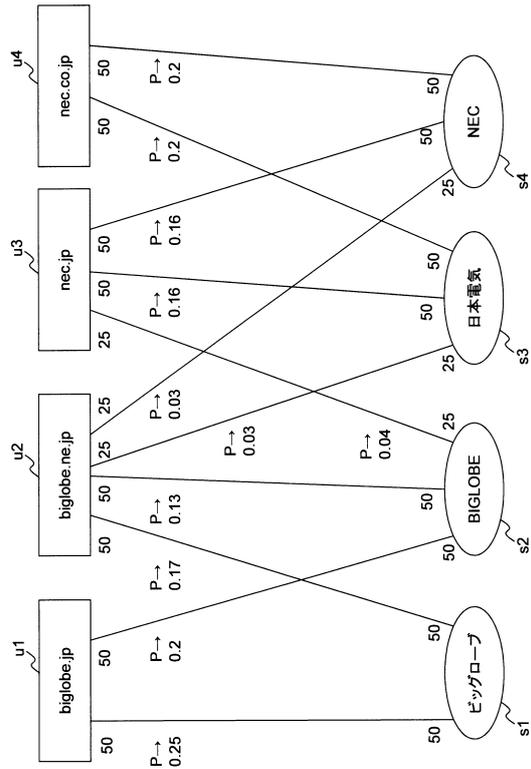
【図7】



【図8】



【図9】



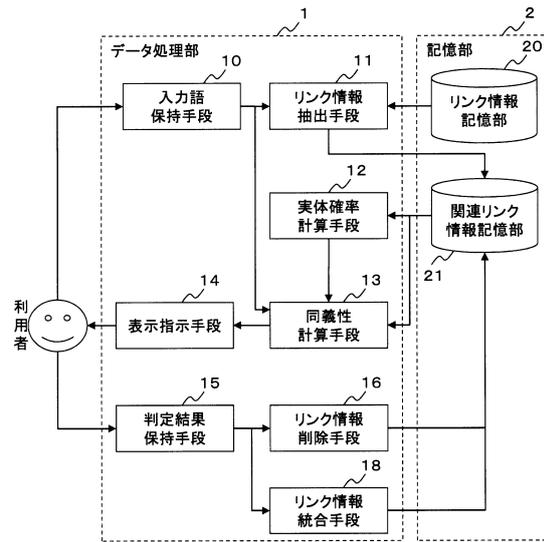
【図10】

入力語: NEC 1回目

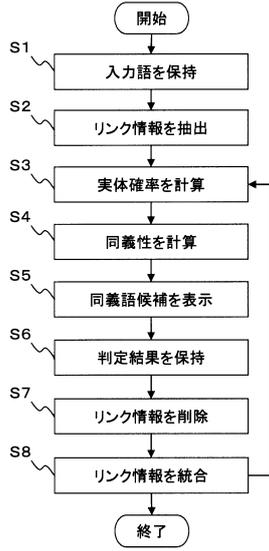
$$\begin{aligned}
 NRe(NEC, \text{日本電気}) &= \frac{0.03 \cdot 0.03 + 0.16 \cdot 0.16 + 0.2 \cdot 0.2}{0.03 + 0.03 - 0.03 \cdot 0.03 + 0.16 + 0.16 - 0.16 \cdot 0.16 + 0.2 + 0.2 - 0.2 \cdot 0.2} \\
 &= \frac{0.00 + 0.03 + 0.04}{0.03 + 0.03 - 0.00 + 0.16 + 0.16 - 0.16 \cdot 0.16 + 0.2 + 0.2 - 0.04} \\
 &= 0.07 / 0.71 \\
 &= 0.10 \\
 NRe(NEC, \text{BIGLOBE}) &= \frac{0.2 \cdot 0 + 0.13 \cdot 0.03 + 0.04 \cdot 0.16 + 0 \cdot 0.2}{0.2 + 0 - 0.2 \cdot 0 + 0.13 + 0.03 - 0.13 \cdot 0.03 + 0.04 + 0.16 - 0.04 \cdot 0.16 + 0.2 + 0 - 0.2 \cdot 0} \\
 &= \frac{0.00 + 0.01}{0.2 + 0.13 + 0.03 - 0.00 + 0.04 + 0.16 - 0.01 + 0.2} \\
 &= 0.01 / 0.75 \\
 &= 0.01
 \end{aligned}$$

$$\begin{aligned}
 NRe(NEC, \text{ビッグローブ}) &= \frac{0.25 \cdot 0 + 0.17 \cdot 0.03 + 0 \cdot 0.16 + 0 \cdot 0.2}{0.25 + 0 - 0.25 \cdot 0 + 0.17 + 0.03 - 0.17 \cdot 0.03 + 0 + 0.16 - 0 \cdot 0.16 + 0 + 0.2 - 0 \cdot 0.2} \\
 &= \frac{0.01}{0.25 + 0.17 + 0.03 - 0.01 + 0.16 + 0.2} \\
 &= 0.01 / 0.8 \\
 &= 0.01
 \end{aligned}$$

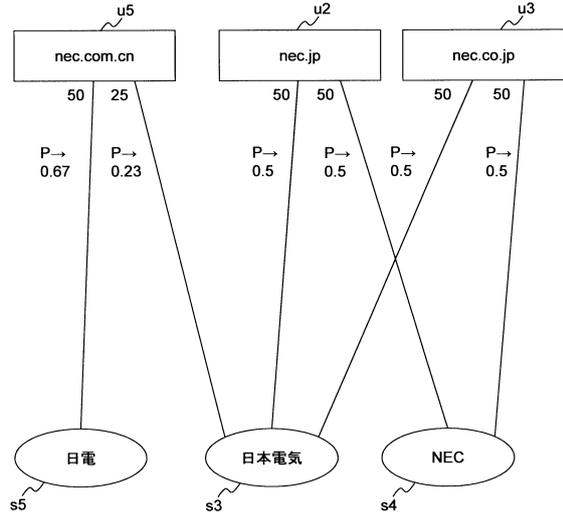
【図11】



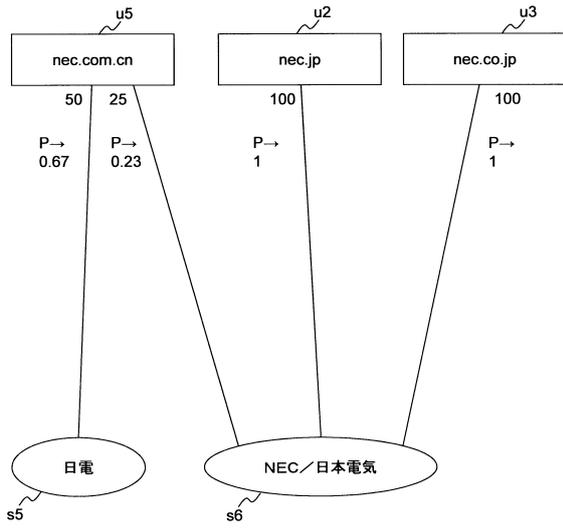
【図12】



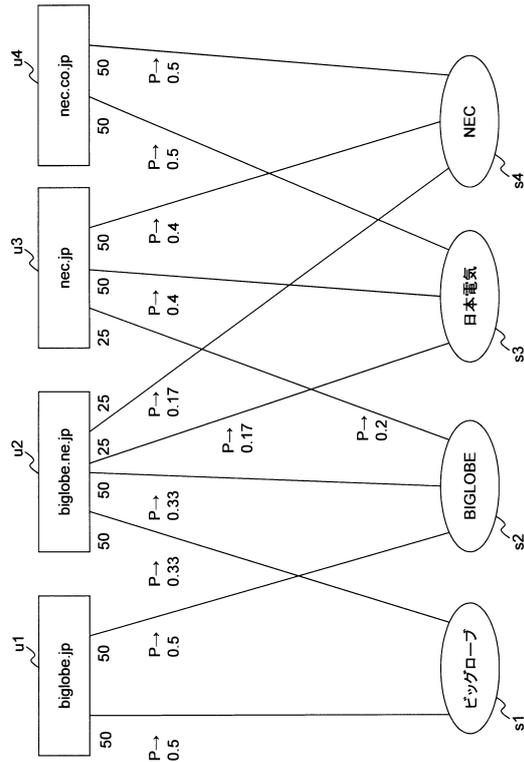
【図13】



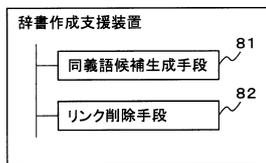
【図14】



【図16】



【図15】



【 17 】

入力語: NEC 1回目

$$\begin{aligned} \text{NRel(NEC,日本電気)} &= \frac{0.17*0.17+0.4*0.4+0.5*0.5}{0.17+0.17-0.17*0.17+0.4+0.4-0.4*0.4+0.5+0.5-0.5*0.5} \\ &= 0.03+0.16+0.25 / 0.34-0.03+0.8-0.16+1-0.25 \\ &= 0.44 / 1.7 \\ &= 0.26 \end{aligned}$$

$$\begin{aligned} \text{NRel(NEC,BIGLOBE)} &= \frac{0.5*0+0.33*0.17+0.2*0.4+0.5*0}{0.5+0-0.5*0+0.33+0.17-0.33*0.17+0.2+0.4-0.2*0.4+0.5+0.5*0} \\ &= 0+0.06+0.08+0.05+0.33+0.17-0.06+0.2+0.4-0.08+0.5 \\ &= 0.14 / 1.96 \\ &= 0.07 \end{aligned}$$

$$\begin{aligned} \text{NRel(NEC,ビッグロブ)} &= \frac{0.5*0+0.33*0.17+0*0.4+0*0.5}{0.5+0-0.5*0+0.33+0.17-0.33*0.17+0+0.4-0*0.4+0+0.5-0*0.5} \\ &= 0+0.06+0+0+0.05+0.33+0.17-0.06+0.4+0.5 \\ &= 0.06 / 1.84 \\ &= 0.03 \end{aligned}$$

フロントページの続き

(72)発明者 細見 格

東京都港区芝五丁目7番1号 日本電気株式会社内

(72)発明者 山名 早人

東京都新宿区戸塚町1丁目104番地 学校法人早稲田大学内

審査官 長 由紀子

(56)参考文献 特開2007-219580(JP,A)

特開2007-164635(JP,A)

特開2009-086979(JP,A)

(58)調査した分野(Int.Cl., DB名)

G06F 17/20-28