【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】平成23年6月30日(2011.6.30)

【公開番号】特開2006-48683(P2006-48683A)
【公開日】平成18年2月16日(2006.2.16)
【年通号数】公開・登録公報2006-007
【出願番号】特願2005-216527(P2005-216527)
【国際特許分類】
Ｇ０６Ｆ　17/30　　(2006.01)
【ＦＩ】
Ｇ０６Ｆ　17/30　　３４０Ａ
Ｇ０６Ｆ　17/30　　３５０Ｃ
Ｇ０６Ｆ　17/30　　２１０Ｄ

【誤訳訂正書】
【提出日】平成23年5月9日(2011.5.9)
【誤訳訂正１】
【訂正対象書類名】明細書
【訂正対象項目名】０００９
【訂正方法】変更
【訂正の内容】
【０００９】
本システムは、更に、フレーズが文書内の他のフレーズの存在を予測する能力に基づいて、相互に関連するフレーズを識別するようになっている。より詳細には、２フレーズの実際の共出現率を、その２フレーズの期待される共出現率と関連させる予測指標を用いる。実際の共出現率と期待される共出現率との比としての情報ゲインは、このような予測指標の１つである。予測指標が所定の閾値を超える場合に２フレーズを関連付ける。その場合、第２フレーズは、第１フレーズに対して大きな情報ゲインをもつ。意味的に、関連するフレーズは、「President of the United States」と「White House」等の所与の主題または概念を検討または説明するために普通に用いるものである。所与のフレーズに対して、それぞれの予測指標に基づく関連性または重要性に従って、関連フレーズをランク付けできる。
【誤訳訂正２】
【訂正対象書類名】明細書
【訂正対象項目名】００６６
【訂正方法】変更
【訂正の内容】
【００６６】
文書コレクションを最終的に通過させることにより、見込フレーズリストは、大規模コーパスのフレーズ使用の期待される分布により、比較的短くなる。従って、例えば１０回の通過（例えば、１０，０００，０００文書）の場合、フレーズは、まさに最初の回だけに出現し、その時点で良好フレーズとなる可能性はほとんどない。そのフレーズは使用されるようになったばかりの新規フレーズかもしれず、従って、後続のクロールをしている間に次第に共通となる。この場合、それぞれのカウントは増大し、最終的に良好フレーズとなる閾値を満たすことになる。
【誤訳訂正３】
【訂正対象書類名】明細書
【訂正対象項目名】００６８
【訂正方法】変更

【訂正の内容】
【００６８】
上記のように、共出現マトリックス２１２は、良好フレーズと関係付けられる格納データのｍ・ｍマトリックスである。マトリックスの各行ｊは、良好フレーズg_jを表し、各列ｋは良好フレーズg_kを表す。各良好フレーズg_jについて期待値Ｅ（g_j）が計算される。期待値Ｅは、g_jを含むと期待されるコレクション内の文書の割合である。例えば、これは、g_jを含む文書の数と、クロールされているコレクション内の文書の総数Ｔとの比、Ｐ（ｊ）／Ｔとして計算される。
【誤訳訂正４】
【訂正対象書類名】明細書
【訂正対象項目名】００６９
【訂正方法】変更
【訂正の内容】
【００６９】
上記のように、g_jを含む文書の数は、g_jが１文書に出現する毎に更新される。Ｅ（g_j）の値は、g_jのカウントが１つ増加する毎に、またはこの第３段階の間に更新できる。
【誤訳訂正５】
【訂正対象書類名】明細書
【訂正対象項目名】００７１
【訂正方法】変更
【訂正の内容】
【００７１】
ｉ）期待値Ｅ（g_k）を計算する。g_jおよびg_kの期待される共出現率Ｅ（ｊ、ｋ）は、両者が関連性のないフレーズの場合、Ｅ（g_j）＊Ｅ（g_k）である；
【誤訳訂正６】
【訂正対象書類名】明細書
【訂正対象項目名】００７２
【訂正方法】変更
【訂正の内容】
【００７２】
ｉｉ）g_jおよびg_kの実際の共出現率Ａ（ｊ、ｋ）を計算する。これは文書総数Ｔで除した生の共出現カウントＲ（ｊ、ｋ）である；
【誤訳訂正７】
【訂正対象書類名】明細書
【訂正対象項目名】００７３
【訂正方法】変更
【訂正の内容】
【００７３】
ｉｉｉ）実際の共出現率Ａ（ｊ、ｋ）が或る閾値量だけ前記期待される共出現率Ｅ（ｊ、ｋ）を超える場合、g_jはg_kを予測するといえる。
【誤訳訂正８】
【訂正対象書類名】明細書
【訂正対象項目名】００９０
【訂正方法】変更
【訂正の内容】
【００９０】
最初に、思い出すべきは、共出現マトリックス２１２が良好フレーズg_jを含み、そのそれぞれは、情報ゲイン閾値より大きな情報ゲインをもつ少なくとも１つの他の良好フレーズg_kを予測するということである。次いで、関連フレーズを識別する（４００）ために、良好フレーズの各対（g_j、g_k）毎に、その情報ゲインを関連フレーズ閾値、例えば１

００、と比較する。すなわち、g_jおよびg_kは、次の場合、関連フレーズである：
【誤訳訂正９】
【訂正対象書類名】明細書
【訂正対象項目名】００９２
【訂正方法】変更
【訂正の内容】
【００９２】
この高い閾値を用いて、統計的期待率を十分超える良好フレーズの共出現を識別する。統計的に、これはフレーズg_jおよびg_kが、期待される共出現率より１００倍多く共出現する、ということを意味する。例えば、文書や「Monica Lewinsky（モニカ・ルインスキー）」というフレーズを考えてみると、フレーズ「Bill Clinton（ビル・クリントン）」は同一文書に１００倍以上出現する可能性があり、ひいては、フレーズ「Bill Clinton」は、任意のランダムに選択される文書にも出現する可能性がある。言いかえると、出現率が１００：１なので、予測精度は９９．９９９％である。
【誤訳訂正１０】
【訂正対象書類名】明細書
【訂正対象項目名】０１００
【訂正方法】変更
【訂正の内容】
【０１００】
一実施の形態では、クラスタ番号は、フレーズ間の直交関係をも示すクラスタービットベクトルにより決定される。クラスタービットベクトルは、良好フレーズリスト２０８内の良好フレーズ数ｎの長さのビットのシーケンスである。所与の良好フレーズg_jについて、ビット位置はソートしたg_jの関連フレーズＲに対応する。ビットは、Ｒの関連フレーズg_kがフレーズg_jと同一のクラスタにある場合に設定される。より一般的には、これは、g_jとg_kとの間のいずれかの方向の情報ゲインがある場合に、クラスタービットベクトルの対応するビットが設定される、ということを意味している。
【誤訳訂正１１】
【訂正対象書類名】明細書
【訂正対象項目名】０１０１
【訂正方法】変更
【訂正の内容】
【０１０１】
次いで、クラスタ番号は、結果的に得られるビット列の値となる。この実装は、多くの方向、または一方向の情報ゲインをもつ関連フレーズが同一クラスタに出現するという特性をもつ。
【誤訳訂正１２】
【訂正対象書類名】明細書
【訂正対象項目名】０２２３
【訂正方法】変更
【訂正の内容】
【０２２３】
「blue merle」::「Australian Shepherd」,「red merle」,「tricolor」,「aussie」;
【誤訳訂正１３】
【訂正対象書類名】明細書
【訂正対象項目名】０２２４
【訂正方法】変更
【訂正の内容】
【０２２４】

「agility training」::「weave poles(編んだ柱)」,「teeter（シーソー）」,「tunnel（トンネル）」,「obstacle(障害)」,「border collie(ボーダーコリー）」。