

(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2004-334441

(P2004-334441A)

(43) 公開日 平成16年11月25日(2004.11.25)

(51) Int.Cl.⁷

G06F 17/30

F I

G06F 17/30

350C

テーマコード (参考)

5B075

G06F 17/30

170A

G06F 17/30

370Z

審査請求 未請求 請求項の数 6 O L (全 10 頁)

(21) 出願番号

特願2003-128297 (P2003-128297)

(22) 出願日

平成15年5月6日(2003.5.6)

(71) 出願人 500082702

エヌ・ティ・ティ・データ・テクノロジ株式会社

東京都港区赤坂2丁目2番12号

(74) 代理人 100108730

弁理士 天野 正景

(74) 代理人 100092299

弁理士 貞重 和生

(72) 発明者 中谷 洋一

東京都港区赤坂2丁目2番12号 エヌ・ティ・ティ・データ・テクノロジ株式会社 内

最終頁に続く

(54) 【発明の名称】 類似文書検索方法

(57) 【要約】

【課題】類似文書検索方法において、種文書の意味内容を左右する重要な用語が存在しなくても、高いレベルにランキングされることがあるという欠点に鑑み、種文書中の意味内容において重要な役割を果たしている用語を有する文書が上位にランキング表示できるようにすることを課題とする。

【解決手段】文書集合中の文書と質問文を類似度比較し、類似度順にランキング表示するようにした類似文書検索方法において、ランキング表示される文書群より、表示順位を入れ替えないで、質問文中に含まれる特定の用語を含まない文書のみを排除するようにする。

【選択図】 図6

発明の概念図

順位	類似度	文庫番号	索引用語			
			炊飯	制御	検出	保温
1	183	D4	×	○	○	○
2	140	D6	×	○	○	○
3	135	D5	○	○	○	○
4	72	D1	×	○	○	○
5	67	D2	○	○	○	○
6	60	D3	○	○	○	×

順位を入れ替えないで索引用語「炊飯」を含まない文書を排除する

順位	類似度	文庫番号	索引用語			
			炊飯	制御	検出	保温
1	135	D5	○	○	○	○
2	67	D2	○	○	○	○
3	60	D3	○	○	○	×

【特許請求の範囲】

【請求項 1】

文書集合中の文書と質問文を類似度比較し、類似度順にランキング表示するようにした類似文書検索方法において、
類似度順にランキング付けされた文書群から、その順位を入れ替えないで、質問文中に含まれる特定の用語を含まない文書のみを排除するようにしたこと
を特徴とする類似文書検索方法。

【請求項 2】

請求項 1 に記載された類似文書検索方法において、
上記文書集合は予め分類検索等の検索手段により検索されたものであることを特徴とする 10
類似文書検索方法。

【請求項 3】

請求項 1 に記載された類似文書検索方法において、
上記質問文は用語単位にマークを付けた形態で表示されること
を特徴とする類似文書検索方法。

【請求項 4】

文書集合中の文書と質問文を類似度比較し、類似度順にランキング表示するようにした類似文書検索システムにおいて、
上記類似度比較に基づいてランキングされる文書群から、順位を入れ替えないで、質問文中に含まれる特定の用語を含まない文書のみを排除するようにするフィルタ手段を設け、 20
フィルタ手段を通した文書群をランキング表示するようにしたこと
を特徴とする類似文書検索システム。

【請求項 5】

請求項 4 に記載された類似文書検索システムにおいて、
上記文書集合を作成するための分類検索等の検索手段を設けたこと
を特徴とする類似文書検索方法。

【請求項 6】

請求項 4 に記載された類似文書検索システムにおいて、
用語単位にマークを付けた形態で上記質問文が表示されるようになっている質問文表示手段を設けたこと 30
を特徴とする類似文書検索システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

この発明は類似文書検索方法に関するものである。

【0002】

【従来の技術】

現在情報検索の分野において、ブーリアン検索方式が多く用いられている。ブーリアン検索方式は、用語を AND、OR、NOT 等のブール演算子により組み合わせた論理式を質問式として用い、質問式に一致する用語の組み合わせを有する文書を検索する方式である 40
。ブーリアン検索方式は質問式の用語の論理的組み合わせに厳密に一致する用語群を含む文書のみを検索する点で優れているが、検索質問式や文書中の用語はすべて同じ重みを有しており、用語の重要度を扱えないと言う欠点を有している。

【0003】

そのため、検索質問式、文書中に出現する用語の重要度を反映し、検索結果に順序付けを行うことができないため、利用者は全検索結果を精査して所望の文書を探し出す作業をしなければならなかった。

【0004】

この問題点を解決する手法として、類似文書検索が提案されている。類似文書検索は、検索質問として種文書を用い、この種文書とデータベース中の文書の類似度を算出して、類 50

似度順に順序付けして、検索結果を提示する手法である。

【0005】

類似度算出には、用語の重要度を算出することが不可欠であり、通常、用語の重要度の算出には $TF-IDF$ 法という手法が用いられる。この方法には次の二つの値が基礎となっている。

【0006】

第一の値は、文書 D_i において、ある用語 T_j が何回出現したかを表わす値であって、語頻度と呼ばれており、

$TF_j^i = \text{語頻度} = \text{文書 } D_i \text{ における用語 } T_j \text{ の出現回数}$
のように示される。

10

【0007】

第二の値は、ある用語 T_j が、データベースの中でいくつぐらいの文書に現れるかと言うことを表わす値であって、文書頻度と呼ばれ、

$DF_j = \text{文書頻度} = \text{用語 } T_j \text{ を含む文書数}$
のように示される。

【0008】

$TF-IDF$ 法は上述の二つの値を以下のように組み合わせることにより、文書 D_i における用語 T_j の重要度 W が、

$$W_j^i = TF_j^i \times \log N / DF_j$$

によって算出される。ここで N はデータベースの全文書数を表わす。

20

【0009】

以上の式から判るように、ある用語 T_j の重要度は、文書 D_i に多く出現すればするほど、また、データベースにおいて、ある用語 T_j が出現する文書数が少なければ少ないほど大きくなることを表している。

【0010】

類似文書検索における類似度算出に際しては、データベース中の全ての用語（索引語となる全ての用語 T_j ）について文書頻度 DF_j 及び全文書数（ N ）を基に $\log N / DF_j$ （以下重みと言う）が算出され、合わせて、データベース中の文書 D_i 毎に、当該文書中出现する全ての用語の重要度 $W_j^i = TF_j^i \times \log N / DF_j$ が算出される。また、種文書 Q_s についても、上述の $\log N / DF_j$ （重み）と種文書中の各用語 T_j の出現頻度即ち語頻度 TF_j^s を基に種文書中出现する全ての用語の重要度 $W_j^s = TF_j^s \times \log N / DF_j$ の算出が行われる。

30

【0011】

種文書と各文書間の類似度の算出は、一般的にはベクトル空間法により行われる。ベクトル空間法では、夫々の用語 T_j にベクトルを対応付けて考えられる。即ち、いま t 個の用語があり、全ての用語に夫々のベクトルが対応付けられた場合、これ等のベクトルが線形独立であれば、 t 次元のベクトル空間が定義されることになり、この空間における全てのベクトルは、 t 個の用語に対応する t 個のベクトル V_i の線形結合として表現できる。このようなベクトル空間において文書 D_i は、

$$D_i = \sum_{j=1}^t W_j^i V_j$$

のように表すことができる。

40

【0012】

検索質問文即ち種文書 Q_s もまた同様に、

$$Q_s = \sum_{j=1}^t W_j^s V_j$$

のように表わすことができる。

【0013】

ベクトル空間において、ベクトルの類似度は、

$$X \cdot Y = |X| \times |Y| \cos \theta$$

のようにベクトルの内積により算出することができる。なお、ここで、 θ は二つのベクトル X 、 Y のなす角度を表す。

50

【0014】

通常、用語をベクトルに対応付ける場合には、同義語を除いて言語は個々に異なった意味を有するものと仮定できるので、 t 個のベクトルはそれぞれ直交していると看做することができる。

【0015】

したがって、文書 D_i と種文書 Q_s との類似度 $\text{sim}(D_i, Q_s)$ は、

$$\text{sim}(D_i, Q_s) = \sum_{j=1}^t W_j^i \times W_j^s \cdots (1)$$

のようになる（非特許文献 1、2 参照）。

【0016】

図 1 乃至図 5 は、種文書と文書間の類似度算出を単純化した例によって具体的に数字を当てはめることにより上記処理の概要を示したものである。ここで、図 1 は種文書中の用語重要度を表にしたもの、図 2 は種文書ベクトルを表にしたもの、図 3 は文書ベクトルの表、図 4 は文書—種文書ベクトル類似度計算値の表、及び図 5 はランキング表示の表である。

【0017】

1. 種文書中の用語重要度（重み×語頻度）には、種文書中に「炊飯」、「制御」、「検出」および「保温」なる用語が用いられており、「炊飯」にはベクトル V_1 が対応付けられており、種文書中の出現頻度は 2 回であり、また、データベース中において「炊飯」が出現する文書頻度より導き出される重みが 2 であることを表わしている。また同様に「制御」、「検出」、「保温」には夫々ベクトル V_2 、 V_3 、 V_4 が対応付けられており、出現頻度が夫々 2 回、3 回、3 回であり、重みが 1、1、3 であることをそれぞれ表わしている。なお、重みの算出は上述の式に基づいて算出される。

【0018】

2. 種文書ベクトルは種文書をベクトル表現したもので各ベクトルに対応する用語の重要度を付して示されている。

【0019】

3. 文書ベクトルは、種文書ベクトルと同様に表現したもので、 D_1 、 D_2 、 D_3 、 D_4 、 D_5 及び D_6 の 6 件の文書についてベクトル表現されている。なお、 D_1 、 D_4 及び D_6 のベクトル V_1 に 0 が付されている、これは文書中に V_1 即ち「炊飯」なる語が存在しないことを意味している。

【0020】

4. 文書—種文書の類似度計算値は、式 (1) を基に類似度を算出したもので、具体的な数字から見られるように、種文書、文書中に同一の用語が存在し、かつ、種文書、文書中において一致する用語の重要度が共に高い場合、その用語に対応するベクトルの内積は高くなり、一致する用語がない場合、その用語に対応するベクトルの内積は 0 となる。以上を総合すると種文書と文書に重要度の高い同一の用語が多く存在すればするほど種文書と文書の類似度が高くなることが判る。

【0021】

5. 類似度計算の結果を基にランキング表示され、図 5 のようになる。

【0022】

以上のように、類似文書検索では、種文書と文書に重要度の高い同一の用語が多く存在すれば種文書と文書の類似度が高くなる仕組みとなっているので、仮令、種文書の意味内容を左右する重要な用語が存在しなくても、文書によっては高いレベルにランキングされることがある。

【0023】

このような例は、図 4 の表に見ることができる。すなわち、文書 D_4 及び D_6 には種文書中の比較的重要度の高い「炊飯」（図 1 中での V_1 で示されている）が存在しないにもかかわらず、図 5 のように上位にランキング表示されている。また、存在したとしても出現頻度が少ないけれども「重要」なる意味概念を有する用語などで強調しているような場合、結果として、類似検索における重要度は低く計算され、文書 D_4 及び D_6 より低くラン

クされるようになる。つまり、類似文書検索において用語の重要度は用語の出現頻度に基づいて算出されるので、意味的に重要とされる用語と必ずしも重要性の点で一致しないケースがある。そのため、種文書中の必須とされる用語が含まれており、本来ならば類似しているものとして、上位にランクされるべき文書であっても、当該用語の文書中における出現頻度が少なければ、当該文書が上位にランク付けされないケースもある。逆に言えば、種文書中の必須とされる用語が含まれていなくても、種文書中の必須とされない用語と一致する用語が多く含まれているために上位にランクされる文書もある。

【0024】

【非特許文献1】

長尾 真 外4名 著 「言語情報処理」岩波講座 言語の科学9 p65～67

10

【非特許文献2】

徳永健伸 著 「情報検索と言語処理」財団法人東京大学出版会 言語と計算5
p11～p43

【0025】

【発明が解決しようとする課題】

この発明の類似文書検索方法は、上述のような従来の類似文書検索システムが有している欠点、すなわち、種文書の意味内容を左右する重要な用語が存在しなくても、高いレベルにランキングされることがあるという欠点、に鑑みなされたものであり、類似文書検索において、種文書中の意味内容において重要な役割を果たしている用語、言い換えれば種文書中の必須の構成要件事項に相当する用語を有する文書が上位にランキング表示できるようにした類似文書検索方法を提供することを課題とする。

20

【0026】

【課題を解決するための手段】

上記課題は、以下の手段により解決される。すなわち、第1番目の発明の解決手段は、文書集合中の文書と質問文を類似度比較し、類似度順にランキング表示するようにした類似文書検索方法において、類似度順にランキング付けされた文書群から、その順位を入れ替えないで、質問文中に含まれる特定の用語を含まない文書のみを排除するようにしたことを特徴とする類似文書検索方法である。

【0027】

第2番目の発明の解決手段は、第1番目の発明の類似文書検索方法において、上記文書集合が予め分類検索等の検索手段により検索されたものであることを特徴とする類似文書検索方法である。

30

【0028】

第3番目の発明の解決手段は、第1番目の発明の類似文書検索方法において、上記質問文が用語単位にマークを付けた形態で表示されることを特徴とする類似文書検索方法である。

【0029】

第4番目の発明の解決手段は、文書集合中の文書と質問文を類似度比較し、類似度順にランキング表示するようにした類似文書検索システムにおいて、上記類似度比較に基づいてランキングされる文書群から、順位を入れ替えないで、質問文中に含まれる特定の用語を含まない文書のみを排除するようにするフィルタ手段を設け、フィルタ手段を通した文書群をランキング表示するようにしたことを特徴とする類似文書検索システムである。

40

【0030】

第5番目の発明の解決手段は、第4番目の発明の類似文書検索システムにおいて、上記文書集合を作成するための分類検索等の検索手段が設けられていることを特徴とする類似文書検索方法である。

【0031】

第6番目の発明の解決手段は、第4番目の発明の類似文書検索システムにおいて、用語単位にマークを付けた形態で上記質問文が表示されるようになっている質問文表示手段が設けられていることを特徴とする類似文書検索システムである。

50

【0032】

【発明の実施の形態】

図6は本発明に従った処理の概念の説明図である。この図において、上の図の左側の3項目は図5と同じであり、右側4項目は、各文書中において、用語「炊飯」、「制御」、「検出」及び「保温」の存在／不存在を示している。

【0033】

今仮に、種文書即ち質問文中の「炊飯」なる用語が重要な意味内容を有しており、意図する検索目標として不可欠の要件、つまり必須の構成要件であると仮定する。その場合、本発明では、図5のランキング表示の表示順位を入れ替えることなく、「炊飯」なる用語を含んでいない文書、つまり文書D4、D6及びD1、が排除される。

10

【0034】

この結果が、図6下の表に示される。用語「炊飯」は、検索の目的からしてなくてはならない用語（そのように前提をおいた）であるから、ランキング表示に続くスクリーニングではその用語を含まない文書は結局は不要な文書である。このような不要な文書がランキング表示から排除され、しかも残る文書の表示順序はランキングのままでありながら、スクリーニング対象となる文書数が減少する。つまり、文書集合中の文書と質問文を類似度比較し、類似度順にランキング表示するにあたり、類似度順にランキング付けされた文書群から、その順位を入れ替えないで、質問文中に含まれる特定の用語を含まない文書のみが排除される。このためスクリーニング対象となる文書数が減少し、効率的に目的とする文書にたどり着くことができる。

20

【0035】

図7はこの方法発明を実施するシステムの概要図である。図7中1は第一検索質問入力部であり、第一検索質問入力部1に入力された分類記号よりなる質問式は第一検索部3において分類索引部2のデータと照合され、質問式と一致するデータを有する文書が第一検索集合記憶部4に記憶される。つまり、文書集合は予め分類検索等の検索手段により検索されることになる。

【0036】

第一検索DB5には文書の記事データが蓄積されており第一検索DB5より取り出された記事データは形態素解析部6により用語単位に切り出され、索引用語・頻度情報蓄積部7に索引用語とその出現頻度が文書単位に蓄積される。第一索引生成部8は、第一検索集合記憶部4に記憶されている文書集合に対応する文書の索引用語及びその頻度を索引用語・頻度情報蓄積部7より文書毎に抽出し記憶する。

30

【0037】

種文書入力部9より質問文が入力されると、前述の記事データから用語を切り出した形態素解析部6により質問用語が切り出され、質問用語及びその出現頻度が質問用語・頻度情報生成部10に登録される。類似度算出部11において、質問用語及びその出現頻度と第一索引生成部8において蓄積されている各文書の索引用語及びその出現頻度情報を基に類似度が計算され、類似度順に整理されてランキング表示部17にランキング表示される。

【0038】

一方、必須用語入力部14には種文書が切り出された用語単位にマークを付けた形態で表示されている。種文書中に検索目的を達成する上で、必要不可欠な用語がある場合、用語を指定入力すると、第二索引生成部12に第一検索集合記憶部4に蓄積されている検索集合に対応する文書の、索引用語が文書毎に抽出され記憶される。

40

【0039】

続いて、第二検索部13において必須用語入力部14に入力された用語と索引用語が比較され、一致する用語を含む文書が第二検索集合記憶部15に蓄積されフィルタ部16に送られる。フィルタ部16は類似度算出部11において類似度順に整理された類似文書検索結果の順位を入れ替えることなく上述の一致する用語を含む文書だけを通過させて、ランキング表示部17に表示するようになっている。

【0040】

50

なお、この発明のシステムの概要において、第一検索部は分類検索によるものとして説明したが、分類以外の言語による全文検索、キーワード検索又は分類とこれ等の組み合わせ検索を用いることもできる。また、種文書中に検索目的を達成する上で、必要不可欠な用語は一語に限るものではなく、何語でも入力できるようにする。

【0041】

更に、この発明の本質と関係がないため上の説明において、敢えて説明を省略したが、第一、第二索引生成部において、同義語辞書等を用いて、索引語の異表記同義語を統制して検索モレを少なくすればなお効果的である。なお、この発明のシステムの概要説明において、第一検索質問入力部1、種文書入力部9及び必須用語入力部14は夫々別のものとして説明したが、実際にはコンピュータ端末のディスプレイ装置に夫々の用途に応じて、呼び出される画面を指しており、ハードウェアとしては同一である。 10

【0042】

【発明の効果】

この発明によれば、類似文書検索法によって文書調査するとき、意図する検索目標を達成する上で必要不可欠な用語がある場合、質問文においてその用語を指定入力することにより、その用語を含まない文書が排除されてランキング表示されるので、上位から順次精査していくことにより、効率的に検索目標に合致する文書に到達することができるという効果を奏する。

【0043】

特に、特許文献の先行技術調査などにおいて、発明の目的、請求の範囲などの発明を構成する必須の要件などの中に、特許性の可否に影響を与える技術的事項が有る場合が想定できる。そのような場合、本発明ではその技術的事項を示す用語を入力すれば、その用語を確実に含む文献が、しかも審査案件と類似度の高い順にランキング表示されることになるので、事後の審査判断をイメージしながら、効率的に目的とする文書にたどりつくことが出来ることになる。 20

【0044】

また、本発明によれば、必須用語入力部において、種文書が切り出された用語単位にマークを付けた形態で表示されているので、類似文書検索における索引用語切り出しルールと全く同じルールにより切り出された用語を指定入力して検索できるので、用語の入力ミスが防止できるという効果を奏する。 30

【図面の簡単な説明】

【図1】種文書中の用語重要度を表にした表図である。

【図2】種文書ベクトルを表にした表図である。

【図3】文書ベクトルの表図である。

【図4】文書－種文書ベクトル類似度計算値の表図である。

【図5】ランキング表示の表図である。

【図6】本発明に従った処理の概念を示した説明図である。

【図7】本方法発明を実施するシステムの概要図である。

【符号の説明】

- | | | |
|----|--------------|----|
| 1 | 第一検索質問入力部 | 40 |
| 2 | 分類索引部 | |
| 3 | 第一検索部 | |
| 4 | 第一検索集合記憶部 | |
| 5 | 第一検索DB | |
| 6 | 形態素解析部 | |
| 7 | 索引用語・頻度情報蓄積部 | |
| 8 | 第一索引生成部 | |
| 9 | 種文書入力部 | |
| 10 | 質問用語・頻度情報生成部 | |
| 11 | 類似度算出部 | 50 |

- 1 2 第二索引生成部
- 1 3 第二検索部
- 1 4 必須用語入力部
- 1 5 第二検索集合記憶部
- 1 6 フィルタ部
- 1 7 ランキング表示部

【図 1】

1. 種文書中の用語重要度 (重み×語頻度)

V1=炊飯	2×2
V2=制御	1×2
V3=検出	1×3
V4=保温	3×3

【図 2】

2. 種文書ベクトル

$Q = 4 * V1 + 2 * V2 + 3 * V3 + 9 * V4$

【図 3】

3. 文書ベクトル

D1=0×V1+3×V2+4×V3+6×V4+3×V5+・・・
D2=6×V1+2×V2+4×V3+3×V4+2×V5+・・・
D3=6×V1+6×V2+8×V3+0×V4+4×V5+・・・
D4=0×V1+6×V2+3×V3+18×V4+2×V5+・・・
D5=1×V1+10×V2+10×V3+9×V4+1×V5+・・・
D6=0×V1+1×V2+1×V3+15×V4+6×V5+・・・

【図 4】

4. 文書—種文書ベクトル類似度計算値

	V1	V2	V3	V4	V5
sim (D1, Q)	0×4+3×2+4×3+6×9+3×0=7.2				
sim (D2, Q)	6×4+2×2+4×3+3×9+2×0=6.7				
sim (D3, Q)	6×4+6×2+8×3+0×9+4×0=6.0				
sim (D4, Q)	0×4+6×2+3×3+18×9+2×0=18.3				
sim (D5, Q)	1×4+10×2+10×3+9×9+1×0=13.5				
sim (D6, Q)	0×4+1×2+1×3+15×9+6×0=14.0				

【図 5】

5. ランキング表示

順位	類似度	文献番号
1	18.3	D4
2	14.0	D6
3	13.5	D5
4	7.2	D1
5	6.7	D2
6	6.0	D3

【図 6】

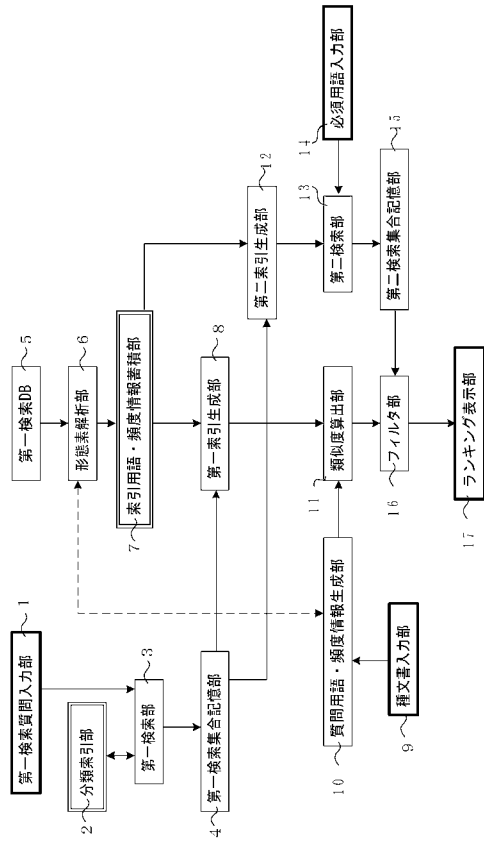
発明の概念図

順位	類似度	文献番号	索引用語			
			炊飯	制御	検出	保温
1	18.3	D4	×	○	○	○
2	14.0	D6	×	○	○	○
3	13.5	D5	○	○	○	○
4	7.2	D1	×	○	○	○
5	6.7	D2	○	○	○	○
6	6.0	D3	○	○	○	×

順位を入れ替えないで索引用語「炊飯」を含まない文献を排除する

順位	類似度	文献番号	索引用語			
			炊飯	制御	検出	保温
1	13.5	D5	○	○	○	○
2	6.7	D2	○	○	○	○
3	6.0	D3	○	○	○	×

【図 7】



フロントページの続き

(72)発明者 須藤 了

東京都港区赤坂2丁目2番12号 エヌ・ティ・ティ・データ・テクノロジ株式会社内

Fターム(参考) 5B075 ND03 NR12 PP23 PQ12 PQ74 PR01 PR06 QM05 UU06