

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2015 年 3 月 12 日 (12.03.2015)



(10) 国际公布号
WO 2015/032301 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2014/085732
- (22) 国际申请日: 2014 年 9 月 2 日 (02.09.2014)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201310400244.4 2013 年 9 月 5 日 (05.09.2013) CN
- (71) 申请人: 江苏大学 (JIANGSU UNIVERSITY) [CN/CN]; 中国江苏省镇江市学府路 301 号, Jiangsu 212013 (CN)。
- (72) 发明人: 王秀红 (WANG, Xiuhong); 中国江苏省镇江市学府路 301 号, Jiangsu 212013 (CN)。
- (74) 代理人: 南京正联知识产权代理有限公司 (NANJING ZHENGLIAN INTELLECTUAL PROPERTY AGENCY CO., LTD); 中国江苏省南京市汉中门大街 1 号 A 幢 27 层 A 座, Jiangsu 210029 (CN)。

- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。
- (84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

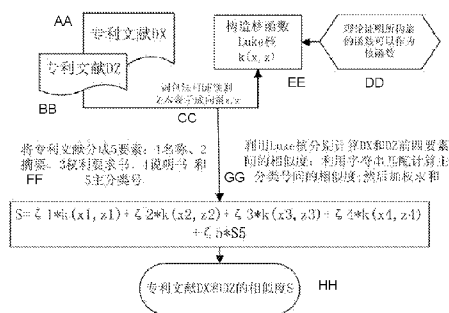
根据细则 4.17 的声明:

— 发明人资格(细则 4.17(iv))

[见续页]

(54) Title: METHOD FOR DETECTING THE SIMILARITY OF THE PATENT DOCUMENTS ON THE BASIS OF NEW KERNEL FUNCTION LUKE KERNEL

(54) 发明名称: 一种基于新核函数 Luke 核的专利文献相似度检测方法



- AA Patent document DX
- BB Patent document DZ
- CC Texts are expressed as a vector x, z on the basis of bags-of-words method and IDF (Inverse Document Frequency) rule
- DD It is proved by the theory that the function constructed can be regarded as a kernel function
- EE A kernel function Luke kernel is constructed
- FF A patent document is divided into five elements: 1 patent name, 2 abstract, 3 claims, 4 description and 5 main classification
- GG The similarity of the first four elements of DX and DZ are calculated respectively by using the Luke kernel, the similarity between the original classification codes is calculated by means of character string matching, and a weighted summation is performed
- HH Similarity S between the patent documents DX and DZ

(57) Abstract: A method for detecting the similarity of the patent documents based on a new kernel function Luke kernel comprises: dividing a patent document into five elements, i.e. patent name, abstract, claims, description and main classification; constructing a new kernel function Luke kernel, calculating the similarity of the first four elements of two patent documents respectively by using the Luke kernel, calculating the similarity between the main classifications of the two patent documents by means of character string matching, and then performing a weighted summation of the similarity of the five elements of the two patent documents to obtain an overall similarity of the patent documents. The method further improves the accuracy and recall rate in the similarity of the patent documents detection, and can be applied to the similarity of the patent documents detection.

(57) 摘要: 一种基于新核函数 Luke 核的专利文献相似度检测方法, 将专利文献分成 5 要素, 即专利名称、摘要、权利要求书、说明书和主分类号; 构造新的核函数 Luke 核, 利用 Luke 核分别计算两篇专利文献前四个要素间的相似度, 再利用字符串匹配计算两专利文献的主分类号之间的相似度, 然后将两专利文献 5 要素间的相似度进行加权求和得专利文献的总体相似度。该方法进一步提高了专利文献的相似度检测的精准率和召回率, 可应用于专利文献的相似度检测。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

说明书

一种基于新核函数 Luke 核的专利文献相似度检测方法

技术领域

本发明属于信息检索技术领域，具体涉及专利文献的文本相似度计算技术。

背景技术

专利的相似度是旨专利间技术内容上的相似性，现有的计算方法大体分成两类：一是基于专利引文的分析，二是基于专利内容的分析。利用引文分析法来分析文献间的相似性的研究已久。在专利相似检测方面，Stuart 用专利的共引关系测量日本 10 家半导体企业的技术相似度。Lai 利用共引分析法来测量专利的相似度。McGill 和 Mowery 等在分析专利联盟内企业之间的关系时，采用互引率测量企业的专利相似度。利用引文分析法来测量专利的相似度有许多不足：只能体现有引用有关系的专利间的相似，不能表明所有真正相关的专利间的相似关系，如中国专利大部分没有引文，这样的专利文献相似度计算通过引文分析法无法很好地解决。基于专利内容来分析专利内容上的相似性的目前研究主要有：Bergmann, Moehrle 等提出专利语义分析方法；Gerken 于 2012 年提出一种基于语义专利分析的方法来测量专利的新颖性。Cascini 提出发明功能树方法，通过比较该树中组件以及组件的功能和层次关系来确定专利的相似度，反映的是专利概念上的相似度而非专利内容上的相似度。Magerman 等验证了文本挖掘技术测量专利相似度的准确性和可能性，Yoon 等利用文本挖掘技术对专利文献进行预处理，构建专利的关键词向量、利用传统的方法，通过计算向量间的欧氏距离来计算专利的相似度，相似检测的精准率和召回率有待进一步提高。陈茂熙等依据专利文献特征构建专利模型树和节点，基于现有的向量空间模型进行相似计算，以专利的名称和摘要信息加权相似度作为分类的依据。彭继东和谭宗颖提出一种基于文本挖掘技术，以专利名称、摘要、权利要求和说明书 4 个文本要素的加权相似度作为专利相似度的计算方法^[1]。Kim 等 2012 年提出使用奇异值方法来计算给定的节点对节点相似矩阵的贡献，从而检测有影响力的专利。Moehrle 于 2012 年提出基于设计决策和结果的文本专利相似测量方法。基于内容的专利相似度计算方法，比起引文分析方法来具有更准确和全面的优势。现有的研究中，大部分是通过分析专利文献的特征，利用现有的向量空间模型计算方法或文本挖掘技术来计算同类类间或同一特征内的相似度；本课题组提出的 S_Wang 核^[2]（专利号 ZL201210105942.7）在分布式信息检索结果融合中有较好的表现。

专利文献的相似度检测中最本质的问题是计算两个专利文献间的相似度。现有技术中

用于计算专利文献的相似度的数学模型往往采用传统的现有向量相似计算数学模型，缺少针对性；在专利文献的结构要素方面只考虑到名称、摘要、权利要求和说明书，忽视了国际专利分类号在专利文献相似计算中的重要作用；现有的方法导致在对专利文献进行相似度计算时的精准率和召回率均有待进一步提高。

[1] 彭继东; 谭宗颖一种基于文本挖掘的专利相似度测量方法及其应用, 情报理论与实践, 2012 (12): 114-118.

[2] 王秀红. 一种基于核函数的文档相似检测方法, 专利号 ZL201210105942.7.

发明内容

本发明的目的在于提供一种基于新核函数 Luke 核的专利文献相似度检测方法，进一步提高专利相似计算的精准率和召回率。

为了解决以上技术问题，本发明构造新的适合专利文献相似度计算的核函数，并结合考虑国际专利分类号在专利文献相似度计算中的重要作用。具体技术方案如下：

一种基于新核函数 Luke 核的专利文献相似度检测方法，其特征在于包括以下步骤：

步骤 1，将待比对的两篇专利文献 DX 和 DZ 的文本分别表示成向量 x 和 z 的步骤；

步骤 2，专利文献结构化表示的步骤：将专利文献分成专利名称、摘要、权利要求、说明书以及主分类号即 IPC 主分类号 5 个要素；所述待比对的两篇专利文献 DX 和 DZ 的所述前 4 个要素分别依次据步骤 1 所述的方法表示成向量为 x_1 、 x_2 、 x_3 、 x_4 和 z_1 、 z_2 、 z_3 、 z_4 ；

步骤 3，构造适于专利文献相似度计算的新核函数 $k(x, z)$ ，并对所述函数 $k(x, z)$ 是否可以作为相似度计算的核函数给予理论证明；

步骤 4，首先利用所述核函数 $k(x, z)$ ，先计算所述待比对的两篇专利文献 DX 和 DZ 前四个各对应要素间的相似度 S_j ， $S_j = k(x_j, z_j)$ ， $j = 1, 2, 3, 4$ ；

然后，对于所述待比对的两篇专利文献 DX 和 DZ 的主分类号，直接进行字符串匹配比对计算两篇专利文献 DX 和 DZ 的主分类号之间的相似度 S_5 ，具体算法过程为：依部、大类、小类、大组、小组顺序从前往后比较主分类号，如果两个专利的主分类号完全相同即小组号相同，则 $S_5 = 1$ ；如果小组号不同，但大组号相同，则 $S_5 = 0.75$ ；如果大组号不同，但

小类号相同, 则 $S_5=0.5$; 如果小类号不同, 但大类号相同, 则 $S_5=0.25$; 如果大类号不同, 但部号相同, 则 $S_5=0.1$; 如果完全不同, 即部号不同, 则 $S_5=0$;

最后加权求和得所述待比对的两篇专利文献 DX 和 DZ 的相似度 S , 具有如下形式

$$S = \sum_{j=1}^5 \zeta_j S_j; \text{ 此处, } \sum_{j=1}^5 \zeta_j = 1, \quad 0 \leq \zeta_j \leq 1, \quad j = 1, 2, \dots, 5。$$

所述的新核函数 $k(x, z)$ 具有形式 $k(x, z) = \log_2^{(x^T z + 1)}$ 。

所述的新核函数可以作为核函数的理论证明过程如下:

令 $\mathbf{X}$ 是 R^n 上的一个紧集, $k(x, z)$ 是 $X \times X$ 上连续实值对称函数, 则有:

$$\iint_{X \times X} k(x, z) f(x) f(z) dx dz \geq 0, \forall f \in L_2(X) \quad (1)$$

称此为 Mercer 条件;

(1)式等价于 $k(x, z)$ 是一个核函数即 $k(x, z) = (\phi(x) \cdot \phi(z))$, $x, z \in X$ 其中 ϕ 为某个从 X 到 Hilbert 空间 H 的映射 $\phi: X \rightarrow \phi(X) \in H$, $(\cdot)$ 是 Hilbert 空间 L_2 上的内积。

下面证明所构建的函数 $k(x, z) = \log_2^{(x^T z + 1)}$ 可以作为核函数, 满足 Mercer 条件;

1) 令 $k_1(x, z) = x^T z$, 所述新核函数可以改写为

$$k(x, z) = \log_2^{(x^T z + 1)} = \log_2^{(k_1(x, z) + 1)} \quad (2)$$

2) 显然 $k_1(x, z) = x^T z$ 是线性核函数, 它满足当 $\mathbf{X}$ 是 R^n 上的一个紧集时, $k_1(x, z)$ 是 $X \times X$ 上为连续实值对称函数, 因文档向量 $\mathbf{x}$ 和 $\mathbf{z}$ 所有元素值均为非负, 所以 $k_1(x, z)$ 为非负;

3) 当两篇专利文献 DX 和 DZ 完全相同时, $k_1(x, z) = x^T z = 1$, 而此时必然有 $k(x, z) = \log_2^{(k_1(x, z) + 1)} = \log_2^2 = 1$; 当两篇文档完全不同时, $k_1(x, z) = 0$, 而此时必然有 $k(x, z) = \log_2^{(k_1(x, z) + 1)} = \log_2^1 = 0$;

综上所述, 当 $\mathbf{X}$ 是 R^n 上的一个紧集时, $k(x, z) = \log_2^{(x^T z + 1)}$ 是 $X \times X$ 上为连续实值对称函数,

且为非负；则由 Mercer 定理可推出 $\iint_{X \times X} k(x, z) f(x) f(z) dx dz \geq 0, \forall f \in L_2$ 。于是有所构造

的 $k(x, z)$ 可以作为核函数，即 $k(x, z) = (\phi(x) \cdot \phi(z))$ ， $x, z \in X$ 。

所述的步骤 1 具体为：

Step1，词包表达：将所有待比对的专利文献的整个集合称为文集，将出现在文集集中的实词的集合称为词典；分别将待比对的两篇专利文献 DX 和 DZ 视为两个词包；

$$\phi: DZ \rightarrow zz = \phi_1(Z) = (tf(t_1, z), tf(t_2, z), \dots, tf(t_N, z)) \in R^N,$$

$$\phi: DX \rightarrow xx = \phi_1(X) = (tf(t_1, x), tf(t_2, x), \dots, tf(t_N, x)) \in R^N,$$

ϕ 为词包法映射关系，N 为所有待比对的专利文献中的实词构成的词典中词的个数； t_i 为词典中的实词； $f(t_i, z)$ 表示实词 t_i 在专利文献 DZ 中出现的频率， $f(t_i, x)$ 表示实词 t_i 在专利文献 DX 中出现的频率； $i = 1, 2, \dots, N$ ；

Step2，语义表示：由于词包表示未考虑词的语义信息，为此在包表示法的基础上构建语义核；不同的词对主题的重要程度不同，采用一个词在文档中出现的频率来量化这个词所带的信息重要程度，即逆文档频率 IDF（Inverse Document Frequency）规则，具体为

$$w(t) = \ln \left(\frac{l}{df(t)} \right) \quad (3)$$

其中 l 为所述文集中存在的专利文献的个数， $df(t)$ 是包含实词 t 的专利文献的个数， $w(t)$ 为逆文档频率 IDF 规则定义的衡量实词 t 的权重的绝对尺度；

所述待比对的专利文献的带语义的向量表示形式为：

$$z_0 = (\omega(t_1)tf(t_1, z), \omega(t_2)tf(t_2, z), \dots, \omega(t_N)tf(t_N, z)) \in R^N$$

$$x_0 = (\omega(t_1)tf(t_1, x), \omega(t_2)tf(t_2, x), \dots, \omega(t_N)tf(t_N, x)) \in R^N$$

再对向量 z_0 和 x_0 分别进行归一化处理，得所述向量 x 和 z 。

本发明具有有益效果。一方面，将本发明构造的新的核函数 Luke 核应用到专利文献的相似度计算，进一步提高了专利文献相似度计算的精准率和召回率。另一方面，本发明通

过将专利文献分成 5 个要素，考虑到国际专利分类号在相似度计算方面的作用，通过先分别计算两个待比对的专利文献的对应要素间的相似度然后再加权求和得两篇专利文献的总相似度，提高了相似度计算的精准率和召回率的同时，减少了计算开销，提高了计算效率。

本发明由以下项目资助完成：

[1]国家自然科学基金委员会，青年科学基金，项目编号：71403107，“专利文献的要素组合拓扑结构及向量空间语义表示与相似度计算研究”；

[2]中国博士后科学基金第七批特别资助，项目编号：2014T70491，“综合位置和语义的专利文献核函数构造及相似度计算研究” 2014.7-2016.6；

[3] 教育部人文社科基金，项目编号：13YJC870026 “基于新核函数的相似专利文献检索研究”。

附图说明

图 1 为本发明方法流程图。

具体实施方式

下面结合附图，对本发明的技术方案作进一步详细说明。

如图 1 所示为本发明的思路图。为了方便描述，将本发明的新核函数 $k(x, z) = \log_2^{(x^T z + 1)}$ 简称为 Luke 核。

步骤 1，利用词包法和逆文档频率 IDF 规则将专利文献的专利名称、摘要、权利要求、说明书四个要素分别表示成对应的向量 x_1 、 x_2 、 x_3 、 x_4 和 z_1 、 z_2 、 z_3 、 z_4 ；

步骤 2，利用构造的新核函数 Luke 核 $k(x, z) = \log_2^{(x^T z + 1)}$ 分别计算专利名称、摘要、权利要求、说明书各要素对应的文本相似度 $S_j = k(x_j, z_j) = \log_2^{(x_j^T z_j + 1)}$ ， $j = 1, 2, 3, 4$ 。

步骤 3，利用字符串比较算法计算不同专利文献主分类号之间的相似度 S_5 ，具体算法过程为：从前往后比较，依部、大类、小类、大组、小组顺序比较。如果两个专利的主分类号相同即到小组号均相同，则 $S_5 = 1$ ；如果小组号不同，但大组号相同，则 $S_5 = 0.75$ ；如果大组号不同，但小类号相同，则 $S_5 = 0.5$ ；如果小类号不同，但大类号相同，则 $S_5 = 0.25$ ；如果大类号不同，但部号相同，则 $S_5 = 0.1$ ；如果部号也不同，则 $S_5 = 0$ 。

步骤 4，计算两篇专利文献的总体相似度 $S = \sum_{j=1}^4 \log_2^{(x_j^T z_j + 1)} + S_5$ 。

实验采用的评价指标分别为精准率 (Precision)、召回率 (Recall) 和综合评价指标 F。

评价指标的具体算法为：

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (4)$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (5)$$

$$F_{\beta} - measure = \frac{(1 + \beta^2) * precision * recall}{\beta^2 precision + recall} \quad (6)$$

将专利文献相似度计算中的召回率和精准率视为同等重要，本实施例中综合评价指标中的参数 β 取 1，得 F_1 指标。

实验数据取 DEWENT 专利数据库中 2000 个美国专利，则文集中专利文献的个数 $l=2000$ ，训练/测试的比例是 3:1。使用的软件有 MATLAB7.0。信息检索工具箱选用卡内基-梅隆大学信息检索及语言模型工作组研发的 Lemur 工具箱。该 Lemur 工具箱支持对大规模文本数据库的索引，以及对文档、提问或文档子集构建简单的语言模型，除此之外，它还支持传统的检索模型，如向量空间模型 VSM 等。实验中线性学习器采用 LibSVM。

现有研究中专利号为 ZL201210105942.7 的“一种基于核函数的文档相似检测方法”中的 S-Wang 核与其它现有的核函数比较起来在文本相似度计算方面有更好的精准率和召回率表现。在此基础上，本实施例将 Luke 核与 S-Wang 核函数和线性核在专利文献相似度检测中的效果进行了比较最终得不同核函数的相似度计算表现。实验还对比了将专利文献作为整体、依前四个要素即专利名称、摘要、权利要求书和说明书先分别进行相似度计算再加权求和、考虑主分类号在内的 5 个要素进行相似度计算再加权求和，实验结果分别如表 1、表 2 和表 3 所示。表中，P 表示相似度计算精准率分值，R 表示相似度计算召回率分值， F_1 为综合评价指标分值。

表 1 专利文献作为一个整体，直接利用核函数计算相似度

	线性核	S_wang 核	Luke 核
P	0.21	0.36	0.43
R	0.87	0.91	0.93
F_1	0.34	0.52	0.59

表 2 不考虑 IPC，只考虑前 4 个要素间的相似度，然后再加权求和

	线性核	S_wang 核	Luke 核
P	0.25	0.39	0.50
R	0.88	0.93	0.95
F_1	0.39	0.55	0.66

表 3 考虑 5 个要素间的相似度，然后再加权求和

	线性核	S_wang 核	Luke 核
P	0.29	0.41	0.58
R	0.90	0.94	0.96
F ₁	0.44	0.57	0.72

*本实施例中，专利名称、摘要、权利要求、说明书以及主分类号五个要素的相似度权重系数分别依次取 $\zeta_1=0.1$, $\zeta_2=0.1$, $\zeta_3=0.25$, $\zeta_4=0.25$, $\zeta_5=0.3$ 。

从表 1、表 2 和表 3 中可以看出，本发明的 Luke 核具有很好的相似度计算表现。从表 2 和表 3 的比较中可以看出，本发明将主分类号考虑进去将专利文献分成 5 个要素，先计算各要素间的相似度然后再加权求和得专利文献的相似度的技术方案，进一步提高了相似度计算的表现。

实验结果表明，本发明采用的专利文献的相似度计算技术方案，提高了专利文献相似度计算的精准率和召回率。

权利要求书

1. 一种基于新核函数 Luke 核的专利文献相似度检测方法，其特征在于包括以下步骤：

步骤 1，将待比对的两篇专利文献 DX 和 DZ 的文本分别表示成向量 x 和 z 的步骤；

步骤 2，专利文献结构化表示的步骤：将专利文献分成专利名称、摘要、权利要求、说明书以及主分类号 5 个要素；所述待比对的两篇专利文献 DX 和 DZ 的所述前 4 个要素分别依次据步骤 1 所述的方法表示成向量为 x_1 、 x_2 、 x_3 、 x_4 和 z_1 、 z_2 、 z_3 、 z_4 ；

步骤 3，构造适于专利文献相似度计算的新核函数 $k(x, z)$ ，并对所述函数 $k(x, z)$ 是否可以作为相似度计算的核函数给予理论证明；

步骤 4，首先利用所述核函数 $k(x, z)$ ，先计算所述待比对的两篇专利文献 DX 和 DZ 前四个各对应要素间的相似度 S_j ， $S_j = k(x_j, z_j)$ ， $j = 1, 2, 3, 4$ ；

然后，对于所述待比对的两篇专利文献 DX 和 DZ 的主分类号要素，直接进行字符串匹配对比计算两篇专利文献 DX 和 DZ 的主分类号之间的相似度 S_5 ，具体算法过程为：依部、大类、小类、大组、小组顺序从前往后比较主分类号，如果两个专利的主分类号完全相同即小组号相同，则 $S_5 = 1$ ；如果小组号不同，但大组号相同，则 $S_5 = 0.75$ ；如果大组号不同，但小类号相同，则 $S_5 = 0.5$ ；如果小类号不同，但大类号相同，则 $S_5 = 0.25$ ；如果大类号不同，但部号相同，则 $S_5 = 0.1$ ；如果完全不同，即部号不同，则 $S_5 = 0$ ；

最后加权求和得所述待比对的两篇专利文献 DX 和 DZ 的相似度 S

$$S = \sum_{j=1}^5 \zeta_j S_j ; \text{ 此处, } \sum_{j=1}^5 \zeta_j = 1, \quad 0 \leq \zeta_j \leq 1, \quad j = 1, 2, \dots, 5。$$

2. 如权利要求 1 所述的一种基于新核函数 Luke 核的专利文献相似度检测方法，其特征在于：所述的新核函数 $k(x, z)$ 具有形式 $k(x, z) = \log_2^{(x^T z + 1)}$ 。

3. 如权利要求 2 所述的一种基于新核函数 Luke 核的专利文献相似度检测方法，其特征在于所述的新核函数可以作为核函数的理论证明过程如下：

令 $\mathbf{X}$ 是 R^n 上的一个紧集, $k(x, z)$ 是 $X \times X$ 上连续实值对称函数, 则有:

$$\iint_{X \times X} k(x, z) f(x) f(z) dx dz \geq 0, \forall f \in L_2(x) \quad (1)$$

称此为 Mercer 条件;

(1)式等价于 $k(x, z)$ 是一个核函数即 $k(x, z) = (\phi(x) \cdot \phi(z))$, $x, z \in X$ 其中 ϕ 为某个从 X 到 Hilbert 空间 H 的映射 $\phi: | \rightarrow \phi(x) \in H$, $(\cdot)$ 是 Hilbert 空间 L_2 上的内积。

下面证明所构建的函数 $k(x, z) = \log_2^{(x^T z + 1)}$ 可以作为核函数, 满足 Mercer 条件;

1) 令 $k_1(x, z) = x^T z$, 所述新核函数可以改写为

$$k(x, z) = \log_2^{(x^T z + 1)} = \log_2^{(k_1(x, z) + 1)} \quad (2)$$

2) 显然 $k_1(x, z) = x^T z$ 是线性核函数, 它满足当 $\mathbf{X}$ 是 R^n 上的一个紧集时, $k_1(x, z)$ 是 $X \times X$

上为连续实值对称函数, 因文档向量 X 和 Z 所有元素值均为非负, 所以 $k_1(x, z)$ 为非负;

3) 当两篇专利文献 DX 和 DZ 完全相同时, $k_1(x, z) = x^T z = 1$, 而此时必然有

$k(x, z) = \log_2^{(k_1(x, z) + 1)} = \log_2^2 = 1$; 当两篇文档完全不同时, $k_1(x, z) = 0$, 而此时必然有

$k(x, z) = \log_2^{(k_1(x, z) + 1)} = \log_2^1 = 0$;

综上所述, 当 $\mathbf{X}$ 是 R^n 上的一个紧集时, $k(x, z) = \log_2^{(x^T z + 1)}$ 是 $X \times X$ 上为连续实值对称函数,

且为非负; 则由 Mercer 定理可推出 $\iint_{X \times X} k(x, z) f(x) f(z) dx dz \geq 0, \forall f \in L_2$, 于是有所构造

的 $k(x, z)$ 可以作为核函数, 即 $k(x, z) = (\phi(x) \cdot \phi(z))$, $x, z \in X$ 。

4. 如权利要求 1 所述的一种基于新核函数 Luke 核的专利文献相似度检测方法, 其特征在于所述的步骤 1 具体为:

Step1, 词包表示: 将所有待比对的专利文献的整个集合称为文集, 将出现在文集其中的实词的集合称为词典; 分别将待比对的两篇专利文献 DX 和 DZ 视为两个词包,

$$\phi: DZ \rightarrow zz = \phi_1(Z) = (tf(t_1, z), tf(t_2, z), \dots, tf(t_N, z)) \in R^N,$$

$$\phi: DX \rightarrow xx = \phi_1(X) = (tf(t_1, x), tf(t_2, x), \dots, tf(t_N, x)) \in R^N,$$

ϕ 为词包法映射关系, N 为所有待比对的专利文献中的实词构成的词典中实词的个数;
 t_i 为词典中的实词; $f(t_i, z)$ 表示实词 t_i 在专利文献 DZ 中出现的频率, $f(t_i, x)$ 表示实词 t_i 在专利文献 DX 中出现的频率; $i = 1, 2, \dots, N$;

Step2, 语义表达: 由于词包表示未考虑词的语义信息, 为此在包表示法的基础上构建语义核; 不同的词对主题的重要程度不同, 采用一个词在文档中出现的频率来量化这个词所带的信息重要程度, 即逆文档频率 IDF 规则, 具体为

$$w(t) = \ln \left(\frac{l}{df(t)} \right) \quad (3)$$

其中 l 为所述文集中存在的专利文献的个数, $df(t)$ 是包含实词 t 的专利文献的个数, $w(t)$ 为逆文档频率 IDF 规则定义的衡量实词 t 的权重的绝对尺度;

进一步地, 所述待比对的专利文献 DX 和 DZ 的带语义的向量表示形式为:

$$z_0 = (\omega(t_1)tf(t_1, z), \omega(t_2)tf(t_2, z), \dots, \omega(t_N)tf(t_N, z)) \in R^N$$

$$x_0 = (\omega(t_1)tf(t_1, x), \omega(t_2)tf(t_2, x), \dots, \omega(t_N)tf(t_N, x)) \in R^N$$

再对向量 z_0 和 x_0 分别进行归一化处理, 得所述向量 x 和 z 。

说明书附图

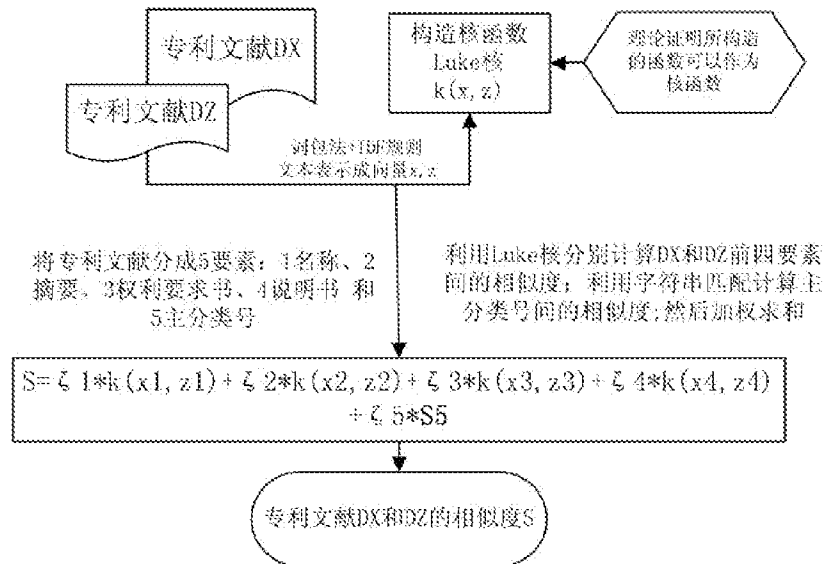


图 1

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2014/085732

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, EPODOC, WPI, CNKI, IEEE, GOOGLE: weight, patent, similar+, detect+, match+, accura+, effic+, document, element, taxonomy

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 103455609 A (JIANGSU UNIVERSITY), 18 December 2013 (18.12.2013), claims 1-4	1-4
Y	CN 101625680 A (NORTHEASTERN UNIVERSITY), 13 January 2010 (13.01.2010), claims 1-7, description, page 6, line 1 to page 7, 4 th line from the bottom, and figure 1	1-2
Y	CN 102651034 A (JIANGSU UNIVERSITY), 29 August 2012 (29.08.2012), claims 1-6	1-2
A	US 2013138665 A1 (THE UNIVERSITY OF MEMPHIS RESEARCH FOUNDATION), 30 May 2013 (30.05.2013), the whole document	1-4

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 24 October 2014 (24.10.2014)	Date of mailing of the international search report 02 December 2014 (02.12.2014)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer LIANG, Jing Telephone No.: (86-10) 62414423

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2014/085732

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103455609 A	18 December 2013	None	
CN 101625680 A	13 January 2010	None	
CN 102651034 A	29 August 2012	None	
US 2013138665 A1	30 May 2013	None	

A. 主题的分类 G06F 17/30 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类		
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNPAT, EPDOC, WPI, CNKI, IEEE, GOOGLE: 加权, 专利, 相似度, 检测, 匹配, 精准, 效率, 文献, 要素, 分类法, weight, patent, similar+, detect+, match+, accura+, effic+, document, element, taxonomy		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 103455609 A (江苏大学) 2013年 12月 18日 (2013 - 12 - 18) 权利要求1-4	1-4
Y	CN 101625680 A (东北大学) 2010年 1月 13日 (2010 - 01 - 13) 权利要求1-7, 说明书第6页第1行-第7页倒数第4行, 附图1	1-2
Y	CN 102651034 A (江苏大学) 2012年 8月 29日 (2012 - 08 - 29) 权利要求1-6	1-2
A	US 2013138665 A1 (THE UNIVERSITY OF MEMPHIS RESEARCH FOUNDATION) 2013年 5月 30日 (2013 - 05 - 30) 全文	1-4
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。		
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件		
国际检索实际完成的日期		国际检索报告邮寄日期
2014年 10月 24日		2014年 12月 02日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 北京市海淀区蓟门桥西土城路6号 100088 中国 传真号 (86-10)62019451		授权官员 梁静 电话号码 (86-10)62414423

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2014/085732

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	103455609	A	2013年 12月 18日	无	
CN	101625680	A	2010年 1月 13日	无	
CN	102651034	A	2012年 8月 29日	无	
US	2013138665	A1	2013年 5月 30日	无	