(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110750988 A
(43)申请公布日 2020.02.04

(21)申请号 201810722238.3

(22)申请日 2018.07.04

(71)申请人 易征宇
地址 101114 北京市通州区张家湾镇三间房806号内1

(72)发明人 不公告发明人

(51)Int.Cl.
G06F 40/289(2020.01)
G06F 16/33(2019.01)
G06K 9/62(2006.01)

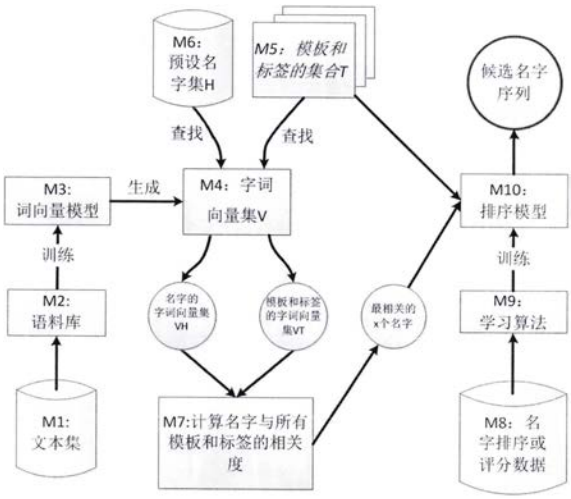
权利要求书2页 说明书7页 附图2页

(54)发明名称

基于人工智能的起名方法

(57)摘要

本发明公开基于人工智能的起名方法，涉及计算机自然语言处理技术和机器学习。主要解决计算机起名过程中筛选规则刚性、可选模式有限、重名率偏高、风格迁移困难、个性化定制不足、排序输出不优等问题。方法如下：训练词向量模型获得字词向量集，训练排序模型获得打分函数；由使用者提供任意名字本身作为模板，或/和对名字的要求作为标签，从字词向量集中查找模板、标签和预提供名字的字词向量，通过计算向量距离，获得相关性最高的多个候选名字，由打分函数计算得分并排序输出。本发明能够帮助使用者获得与模板风格相似、意象关联，与标签要求符合、好名优先的候选名字序列，达到在更高起点上提高计算机起名效果与效率的目的。



CN 110750988 A

1.基于人工智能的起名方法，其特征在于：根据使用者提供的任意名字作为模板，或/和对名字的要求作为标签，基于词向量模型和排序模型，生成候选名字序列。

2.根据权利要求1所述基于人工智能的起名方法，其特征在于：所述模板是使用者希望获得的候选名字范例，所述标签是使用者对名字意蕴的要求；模板和标签数量任意、可以是文字、图片、音频、视频等数据格式；所有模板和标签组成集合T。

3.根据权利要求1所述基于人工智能的起名方法，其特征在于：所述词向量模型，经语料库训练后，可以生成字词向量集V；所述语料库通过对预先搜集的文本集清洗、分词、去停用词和高频词后获得；字词向量集V中包含语料库中字词的特征向量；从V中查找模板和标签集合T中元素的字词向量，得到V的子集VT；同样从V中查找预设名字集H中所有名字对应的字词向量，得到V的另一个子集VH，计算VH中每个向量与VT所有向量在向量空间中距离，获得与模板和标签距离最小、相关度最大的x个名字。

4.根据权利要求3所述的向量在向量空间中距离，用于表征名字与模板或标签的相关度；对于模板和标签的集合T＝{T_1,T_2,...T_i,...T_n)、名字集合H＝{H_1,H_2,...,H_j,...H_m)来说，如果V同时收录了任意T_i和H_j对应的字词向量，那么直接计算字词向量间距离可获得T_i和H_j的相关度；如果V没有同时收录T_i和H_j的字词向量，但收录了组成T_i和H_j的所有单个文字的字词向量，且单个文字数目相等均为s时，分别计算T_i和H_j对应顺位上文字的字词向量距离，获得距离序列D＝{D_1,D_2,...,D_k,...D_s)，其中D_k是T_i与H_j各自第k个文字的字词向量距离，此时T_i和H_j的相关度为$\frac{1}{s}\sum_{k=1}^{s}D_k$。

5.根据权利要求4所述的向量空间中距离，其特征在于距离的类型包括但不限于余弦夹角(Cosine)、欧氏距离(Euclidean Distance)、马氏距离(Mahalanobis Distance)、曼哈顿距离(Manhattan Distance)、切比雪夫距离(Chebyshev Distance)、相关系数(Correlation coefficient)等；本发明的一个优选实施例采用将余弦夹角和欧氏距离归一化后加权求和的方式计算；另一个优选实施例采用将余弦夹角和马氏距离归一化后加权求和的方法。

6.根据权利要求1所述的基于人工智能的起名方法，特征在于所述排序模型采用的算法：包括但不限于基于listwise的ListNet、ListMLE等算法，或基于Pairwise的RankNet、LambdaRank等算法；本发明的一个优选实施例采用ListMLE算法；考虑到计算复杂度，另一个优选实施例采用LambdaMART算法；采用上述各类算法的模型通过对训练数据的学习后可得到算分函数$f_{T_i}(VH_j)$，对于任意H_j，存在排序总分SR_{H_j}，$SR_{H_i}=\sum_{i=1}^{n}f_{T_i}(VH_j)$，将$SR_{H_j}$作为降序排序依据获得对应的候选名字序列。

7.根据权利要求6所述的训练数据，特征在于包括以下形式：提供基于特定标签或模板的若干名字后，用户对这些名字进行排序的数据；或是专家对这些名字评分的数据，再或是收集使用者在电子终端对这些名字的选择性点击等操作的记录数据。

8.根据权利要求1所述的基于人工智能的起名方法，特征在于：包括在本发明的精神和原则之内，所作的任何更改、等同替换、改进等，均应包含在本发明的保护范围之内，所述的修改、等同替换和改进包括但不限于如下类型：

a.在本发明的优化实施例中，使用者提供了文字形态的模板和标签，但这种形态也可

以更改为图片、视频、音频等，但只要在起名过程中，这些形态的数据被转换成文字或字词向量(词向量)用于取名计算，这样的更改同样在本发明的权利保护范围之内；

b.在本发明的优化实施例中，起名过程采用了word2vec词向量模型及其改进后的模型、基于pairwise和listwise的若干排序模型，但在实际应用中，也可以使用其它类型的词向量模型和排序模型进行等同替换，或是在模型使用过程中对模型的某些方法进行更改；只要在起名过程中，采用词向量模型获取名字的字词向量，或采用排序模型进行候选名字排序，这样的等同替换和更改同样在本发明的权利保护范围之内；

c.在本发明的优化实施例中，起名过程用到的词向量模型和排序模型，是以采用本发明所提供方法的技术人员为主体，自行编写程序源代码构建，自行获取数据并进行训练的，但使用这两类模型的方式也可以替换为：集成其它机构或个人提供、基于机器学习及神经网络的模型和方法开发后实现的人工智能框架、软件开发包SDK、开源代码、闭源程序、网络接口API等；不管是显式地采用还是隐式地集成，不论是自行开发还是底层实现，只要采用本发明所描述的方法起名，且起名结果中包含了词向量模型和排序模型的方法要素，同样在本发明的权利保护范围之内；

d.在本发明的优化实施例中，字词向量在向量空间中距离的计算方法采用综合计算余弦夹角和欧氏距离、或综合计算余弦夹角和马氏距离并加权求和的方法，但仍然可以替换为单独或综合计算马氏距离、曼哈顿距离、切比雪夫距离、相关系数等方法；只要在起名过程中通过计算字词向量间距离来获取相关名字，同样在本发明的权利保护范围之内；

e.对本发明各步骤的顺序、模块的取舍、优化实施例中步骤和模块中参数的调整，同样在本发明的权利保护范围之内；

f.本发明的优化实施例中，起名过程不包括对名字的类型、偏好等的显式的筛选和排序，但筛选和排序规则仍然可以被增加在各个环节前后，这些规则包括但不限于：加入对姓氏的考虑、对名字中文字进行增减和修改，对名字类型或名字中任意单字进行限定、设置除语义相关度以外的其它指标对候选名字进行评分和排序。

基于人工智能的起名方法

技术领域

[0001] 本申请涉及计算机自然语言处理技术和机器学习,具体涉及基于人工智能的起名方法。

背景技术

[0002] 每个自然人和机构都有起名需求:新生儿出生后需要取名上户口,户籍登记后可能会改名,机构发起或变更时同样需要起名或改名,为了让名字具有良好的意蕴、深刻的内涵、清晰的发音,人们往往会花费大量时间找出候选名字依次斟酌,但由于个体水平差异,往往费时颇多而收效不尽如人意。

[0003] 互联网相关网站和计算机中有关程序已经可以提供计算机自动起名功能,从公开文献和互联网搜索的结果来看,目前的计算机起名主要包括以下模式:模式1.基于民俗设计显式规则对名字库中名字进行筛选;模式2.输入文本形式起名要求后,从一个更大文本集或名字库中采用字符直接命中规则进行摘抄;模式3.是对模式2的扩展,由使用者提供一个名字范例,基于汉语字词的读音、字形、含义规则对该范例进行一轮扩展,扩大字符命中和摘抄的范围,其中基于字词含义的扩展主要靠查字典辞典、预设同义词表或联想词表进行;模式4.预设数目有限的类别,每个类别有固定的常用字和常用名,根据使用者意愿与预设类别的匹配度,将最相似类别的常用字和常用名作为优先用字和优先用名。总的来说基于上述四种模式及其变种的起名方法,存在扩展方法单一、筛选规则刚性、可选模式有限、重名可能较大、风格迁移困难、个性化定制不足等问题。

[0004] 机器在自动起名时往往提供多个候选名字,在提供给使用者时需要排序,现行的模式中,普遍采用基于民俗、使用者的显式喜好、文字的音形义等显性规则进行打分,还不能基于对名字好坏评价的大数据来排序,存在和用户交互程度低、排序输出不够智能等问题。

[0005] 人工智能是关于怎样表示知识、获得知识并使用知识的现代科学,机器学习及神经网络(学习)都属于人工智能的范畴。其中,机器学习中的一些传统算法例如梯度提升树(GBDT)等,在排序模型中依然发挥重要作用;神经网络是机器学习的新发展,通过模拟人脑的机制更好地进行特征抽取和表征,能取得比手工设计特征更高效精确的效果,从而提高预测或分类的准确性;

[0006] 基于神经网络较高的特征抽取能力与名字寓意等复杂特征的耦合,也基于传统机器学习算法相对较低的计算复杂度可以预见,基于人工智能的起名方法,先天具备较优的性能和提供大规模服务的可能。

[0007] 以下给出检索的相关文献:

1.石淼,陈议,黄际洲.基于使用者意愿的计算机智能起名,CN 101556574A[P].2009.

2.王强.文字选择方法及装置,CN107391491A[P].2017.

3.靳晓明,何涛.基于深度学习的室内导航方法,CN105444766A[P].2016.

4.Lan Goodfellow,Yoshua Bengio,Aaron Courville.Deep learning[M].赵申剑,等

译.北京:人民邮电出版社,2017:3-196.

5.Xia F,Liu T Y,Wang J,et al.Listwise approach to learning to rank:theory and algorithm[C]//International Conference on Machine Learning.ACM,2008:1192-1199.

6.Burges C J C.From ranknet to lambdarank to lambdamart:An overview[J].Learning,2010,11.

发明内容

[0008] 本发明旨在一定程度上解决上述相关背景技术中的问题之一。

[0009] 为此,本发明提供基于人工智能的起名方法,通过训练词向量模型获得字词向量集,通过训练排序模型获得打分函数;由使用者提供任意名字本身作为模板,或/和对名字的要求作为标签,从字词向量集中查找模板、标签和预提供名字的字词向量,通过计算向量在字词向量空间中的距离,获得相关性最高的多个候选名字,由打分函数计算得分并排序输出。本发明能够帮助使用者获得与模板风格相似、意象关联,与标签要求符合、好名优先的候选名字序列,达到在更高起点上提高计算机起名效果与效率的目的。

[0010] 为实现上述目的,本发明提供基于人工智能的起名方法,包括以下模块:M1:文本集、M2:语料库、M3:词向量模型、M4:字词向量集V、M5:模板和标签的集合T、M6:预设名字集H、M7:计算名字与所有模板和标签的相关度、M8:名字排序或评分数据、M9:学习算法、M10:排序模型。

[0011] 本发明基于人工智能的起名方法,需求输入方式灵活,起名结果相关程度高,可以解决现有计算机取名方法中,规则属性刚、重复起名多、风格迁移难、交互程度低、排序不智能等问题,提高起名需求满足的精准度,减少人工干预需求,降低对人工起名的依赖。

[0012] 此外,根据本发明基于人工智能的起名方法,还可以具有如下附加的技术特征:

[0013] 所述模块M1中文本集,可以是无预设类别的超大规模文本集,也可以是类似于古诗词集这样具备特定范围的适量规模文本集;

[0014] 所述模块M2中语料库,通过对文本集进行分词、清洗,去停用词、高频词等后生成;

[0015] 生成语料库后统计其中出现的所有字词获得字词集W;

[0016] 所述模块M3中词向量模型(Embedding Model),指的是包括但不限于Word2Vec、GloVe、FastText、WordRank等在内,能够基于语料中字词及字词中元素在文档中的共现关系进行训练的神经网络模型,通过训练可以生成字词在向量空间中对应特征向量,从而代替字词本身获得机器计算的便利性;

[0017] 基于硬件计算资源等原因,一种优选的词向量模型是Word2Vec,具体的子模型和相关参数视数据规模和训练过程而定;

[0018] 对于模块M1中特定领域适量规模文本集,另一种优选的词向量模型是基于Word2Vec模型改进的cw2vec模型(Shaosheng Cao等,2018),具体参数视训练过程而定;

[0019] 所述模块M4中字词向量集V的生成过程为:通过语料库训练词向量模型获得,其中,V中向量与字词集W中字词一一对应;

[0020] 所述模块M5中,模板和标签都来自使用者的输入,所有模板与标签组成集合T,其中模板是使用者希望获得的候选名字范例,例如“思河”、“枫朗”,标签是使用者对名字意蕴

的要求,比如“阳光”、“清新”;

[0021] 所述模板和标签可以是文字、图片、音频、视频等多种数据形式,但这些数据最终可以转化成向量等数学化表征形式,用于后续起名运算;

[0022] 从V中查找T中任意T_i对应的字词向量VT_i,获得V的子集VT;

[0023] 所述模块M6中名字集H,是在本发明之前预先提供,包含m个名字,通过查找H中任意H_j在V中对应的字词向量VH_j,获得V的子集VH;

[0024] 由于V中难以完全收录T和H中所有模板、标签、名字的字词向量,对于包含不止1个文字的T_i与H_j,除了尝试获取其整体的字词向量,还应获取其包含所有单个文字的字词向量,确保后续计算正常进行;

[0025] 所述模块M7中获得每个名字与所有模板、标签相关度的方法是:计算CH_j与VT中所有向量在向量空间上的距离并求和;

[0026] 如果T_i与H_j任意一方整体的字词向量V中未收录,且T_i与H_j字数相等,则分别计算T_i与H_j对应顺位上文字的字词向量距离并求平均值,视该平均值为T_i与H_j整体的相关度;

[0027] 所述的向量空间上距离的计算方法有很多种,包括但不限于计算余弦夹角(Cosine)、欧氏距离(Euclidean Distance)、马氏距离(Mahalanobis Distance)、曼哈顿距离(Manhattan Distance)、切比雪夫距离(Chebyshev Distance)、相关系数(Correlation coefficient)等,一种优选的方法是将余弦夹角数值和欧氏距离分别归一化后加权求和,另一种优选的方法是将余弦夹角数值和马氏距离分别归一化后加权求和;

[0028] 完成H中所有名字和T中所有模板、标签的相关度计算后,取相关度数值排前x位的名字组成候选名字集N;

[0029] 所述模块M8中名字排序或评分数据用于对排序模型进行训练的标记数据,来自使用者的提交或是专家评分,再或是收集使用者在电子终端上对候选名字的选择性点击等数据;

[0030] 标记数据的格式是:针对特定的查询词(Query),对应多个文档(Doc)的相关度顺序或相关度分值;

[0031] 在本发明中,模板或标签是查询词(Query),候选名字是文档(Doc);

[0032] 所述模块M9中学习算法包括但不限于:基于listwise的ListNet、ListMLE等算法,或是基于Pairwise的RankNet、LambdaRank等算法;

[0033] 其中一种优选的算法的是ListMLE(实现方法参见背景技术部分的相关文献5),考虑到计算复杂度,另一种优选算法的是LambdaMART(实现方法参见背景技术部分的相关文献6);

[0034] 所述模块M10中排序模型提供算分函数$f_{T_i}(VH_j)$,根据特定T_i计算任意H_j的得分,作为排序依据;

[0035] 依次将T中所有元素作为查询词,计算N中所有候选名字的得分,将同一候选名字H_j基于所有查询词下的得分求和得到排序总分SR_{H_j},有$SR_{H_j}=\sum_{i=1}^{n} f_{T_i}(VH_j)$,将N中候选名字按排序总分降序排列输出,获得最终候选名字序列。

[0036] 本发明的附加方面和优点将在下面的描述中部分给出,部分将从下面的描述中变得明显,或通过本发明的实践了解到。

附图说明

[0037] 本发明的上述和/或附加方面和优点可以在结合下面附图对优化实施例的描述中变得明显和容易理解,其中:

[0038] 图1是本发明一个优化实施例的基于人工智能起名方法中各模块的关系图;

[0039] 图2是本发明一个优化实施例的基于人工智能起名方法的流程图;

具体实施方式

[0040] 为使本发明的目的、技术方案和优点更加清楚明了,下面通过附图中示出的具体实施例来描述本发明;但是应该理解,这些描述只是示例性的,仅用于解释本发明,而不能理解为对本发明的限制;此外,在以下说明中,省略了对公知结构和技术的描述,以避免不必要地混淆本发明的概念。

[0041] 以下结合附图描述根据本发明实施例的基于人工智能的起名方法。

[0042] 图1中各模块的功能和特征已经在前述发明内容部分中作了说明;

[0043] 图2中描述了本发明的一个优化实施例,步骤S1:在预设范围内搜集文本生成文本集;步骤S2:生成训练用语料库和对应字词集W;步骤S3:生成字词集W对应的字词向量集V;步骤S4:接收模板和标签等信息后查找对应的字词向量获得V的子集VT;步骤S5:查找预设名字集中所有名字的字词向量获得V的子集VH;步骤S6:计算VT中向量和VH中向量在字词向量空间中的距离;步骤S7:收集名字集排序或评分的标记数据;步骤S8:基于特定学习算法训练排序模型;步骤S9:生成候选名字序列。

[0044] 图2中各步骤包括以下实施细节:

[0045] 步骤S1:文本集采用用UTF-8编码;

[0046] 步骤S2:采用第三方分词工具切分文本集,分词后除了基本的数据清洗外,还需要去除不包含人名和机构名在内的停用词,获得语料库C,C中包含的所有不同字词组成字词集W;

[0047] 对于W中任意字词w_t,定义一个函数P(w_t),如果P(w_t)≥0,则w_t在文本集中属于高频词,并从C和W中删除w_t;

[0048] P(w_t)的数学表达式为:

$$P(w_i)=1-\sqrt{\frac{t}{f(w_i)}}$$,其中f(w_t)为字词w_t在语料所有字词中出现的概率,t为阈值,介于e^{-3}到e^{-5}之间;

[0049] 步骤S3:将语料库C输入词向量模型进行训练,生成字词集W对应的字词向量集V;

[0050] 向量集V中向量的具有d维,基于当前小型工作站的计算资源条件下,一种优选的方案是取d=300;

[0051] 步骤S4:接收使用者提供的模板和标签共n个,得到模板和标签的集合T:

$T=\{T_1,T_2,...T_i,...T_n\}$;

[0052] 从V中查找集合T对应的字词向量子集VT:

$VT=\{VT_1,VT_2,...VT_i,...VT_n\}$;

[0053] 由于V中不一定包含所有模板、标签的字词向量,对于组成文字数目为s1(s1≥2)

的模板或标签T_i，如果V没有收录T_i的字词向量，此时需要将T_i视作文字组合，将其拆分成单个文字分别从V中查找；

[0054] 具体设$T_i=\{t_1^i,t_2^i,...,t_k^i,...t_{s1}^i\}$，其中$t_k^i$为$T_i$第k个文字，分别从V中查找$t_k^i,(k=1,2,...,s1)$的字词向量，获得V的字词向量子集$VT_Sub_i$：

$VT\_Sub_i=\{VT\_Sub_1^i,VT\_Sub_2^i,...,VT\_Sub_k^i,...VT\_Sub_{s1}^i\}$，其中$VT_Sub_k^i$为第$T_i$中第k个文字的字词向量；

[0055] 如果V中收录了T_i整体的字词向量，记为VT_Whl_i；

[0056] 步骤S5：预设名字集$H=\{H_1,H_2,...,H_j,...H_m\}$，在V中查找H对应的字词向量VH：$VH=\{VH_1,VH_2,...VH_i,...VH_m\}$；

[0057] 如果V中收录了H_i整体的字词向量，记为VH_Whl_j；

[0058] 对于组成文字数目为s2(s2≥2)的名字H_j，需要将H_j视作文字组合，拆分成单个文字，分别从V中查找字词向量；

[0059] 具体的，设$H_j=\{h_1^j,h_2^j,...,h_k^j,...h_{s2}^j\}$，其中$h_k^j$为第j个名字$H_j$的第k个文字，分别从V中查找$h_k^j,(k=1,2,...,s2)$的字词向量，获得$VH_Sub_j$；

$VH\_Sub_j=\{VH\_Sub_1^j,VH\_Sub_2^j,...,VH\_Sub_k^j,...VH\_Sub_{s2}^j\}$，其中$VH_Sub_k^j$为第j个名字$H_j$第k个文字的字词向量；

[0060] 步骤S6：一种优选的距离计算方式是将余弦夹角和欧氏距离归一化后，加权求和；

[0061] 对于任意T_i与H_j，所述余弦夹角的计算方法为：

$$Cosine_{(T_i,H_j)}=\cos(VT_i,VH_j)=\frac{VT_i\times VH_j}{|VT_i|\times|VH_j|}$$

其中，$|VT_i|$和$|VH_j|$分别是VT_i和VH_j的模，$VT_i\times VH_j$是VT_i和VH_j的内积；

[0062] 如果VT_Whl_i和VH_Whl_j都可从V中查找到，则：

$$Cosine_{(T_i,H_j)}=Cosine\_Whl_{(T_i,H_j)}=\cos(VT\_Whl_i,VH\_Whl_j)=\frac{VT\_Whl_i\times VH\_Whl_j}{|VT\_Whl_i|\times|VH\_Whl_j|}$$

[0063] 如果VT_Whl_i或VH_Whl_j任意一方无法获得，且s1＝s2＝s，有：

$$Cosine_{(T_i,H_j)}=Cosine\_Sub_{(T_i,H_j)}=\cos(VT\_Sub_i,VH\_Sub_j)=\frac{1}{s}\sum_{k=1}^{s}\frac{VT\_Sub_k^i\times VH\_Sub_k^j}{|VT\_Sub_k^i|\times|VH\_Sub_k^j|}$$

[0064] 对于任意T_i与H_j，所述欧氏距离的计算方法为：

$Euclidean_{(T_i,H_j)}=dist(VT_i,VH_j,)=\sqrt{\sum_{p=1}^{d}(vt_p^i-vh_p^j)^2}$，其中$vt_p^i$为$VT_i$第p维取值，其中$vh_p^j$为$VH_j$第p维取值；

[0065] 如果VT_Whl_i和VH_Whl_j都可从V中查找到，则：

$$Euclidean_{(T_i,H_j)}=Euclidean\_Whl_{(T_i,H_j)}=dist(VT\text{-}Whl_i,VH\text{-}Whl_j,)=\sqrt{\sum_{p=1}^{d}(vt\_whl_p^i-vh\_whl_p^j)}$$

其中$vt_whl_p^i$是$VT\text{-}Whl_i$第p维取值，$vh_whl_p^j$是$VH\text{-}Whl_j$第p维取值；

[0066] 如果VT_Whl_i或VH_Whl_j任意一方无法获得，且s1＝s2＝s，有：

$$\mathrm{Euclidean}_{(T_i,H_j)}=\mathrm{Euclidean\_Sub}_{(T_i,H_j)}=dist(\mathrm{VT\_Sub}_i,\mathrm{VH\_Sub}_j,)=\frac{1}{s}\sum_{k=1}^{s}\sqrt{\sum_{p=1}^{d}(vt\_sub_p^i-vh\_sub_p^j)}$$

其中，$vt_sub_p^i$是VT_Sub$_i$第p维取值，$vh_sub_p^j$是VH_Sub$_j$第p维取值；

[0067] 余弦夹角归一化计算方法为：

$$\mathrm{Normalization\_Cosine}_{(T_i,H_j)}=0.5+0.5\times\mathrm{Cosine}_{(T_i,H_j)}$$

[0068] 欧氏距离归一化计算方法为：

$$\mathrm{Normalization\_}Euclidean_{(T_i,H_j)}=\frac{1}{1+Euclidean_{(T_i,H_j)}}$$

[0069] T_i与H_j相关度$Simi_{(T_i,H_j)}$的计算方法为：

$Simi_{(T_i,H_j)}=a\times\mathrm{Normalization\_Cosine}_{(T_i,H_j)}+b\times\mathrm{Normalization\_}Euclidean_{(T_i,H_j)}$，其中，a和b是本优化实施例中可调试的常数，其中a∈[0,+∞)，b∈[0,+∞)；

[0070] 将相关度$Simi_{(T_i,H_j)}$数值较大的前x个字词归纳为集合N：N＝{N_1,N_2,...N_r,...N_x}；

[0071] 步骤S7：预设一个模板库MK和标签库TK，收集基于MK和TK中特定模板或标签的候选名字排序或评分数据形成数据集，优选的有3种方法，下文举例说明：给定一个标签“清新”和若干候选名字，由使用者对这些候选名字的“清新”程度进行人工排序，将排序结果返回给本优化实施例，基于所有模板和标签的排序数据组合起来形成数据集A；类似的，由起名专家对这若干个候选名字的“清新”程度进行人工评分，将评分结果返回给本优化实施例，基于所有模板和标签的评分数据组合起来形成数据集B；同样类似的，向使用者提供基于“清新”标签的的若干候选名字，由电子终端记录下使用者进行点选等操作的顺序，跟据特定转化规则计算每个候选名字的得分，基于所有模板和标签的得分数据组合起来形成数据集C；

[0072] 步骤S8：本发明一种优选的排序模型采用LambdaMART算法：把任意模板或标签作为查询词，把数据集A中基于该查询词的候选名字序列作为算法中的文档及文档顺序，LambdaMART算法基于上述形式数据学习，学习结束后获得MC和TC中每个模板或标签的算分函数$f_{T_i}(VH_j)$，学习结束条件视具体情形而定；

[0073] 本发明的另一种优选的排序模型采用基于神经网络的ListMLE算法或ListNet算法，在步骤S8中，优选的，把任意模板或标签作为查询词，把数据集B或C中基于该查询词的候选名字及得分作为算法中的文档及文档得分进行学习，学习结束后获得MC和TC中每个模板或标签的算分函数$f_{T_i}(VH_j)$，学习结束条件视具体情形而定；

[0074] 步骤S9：对于用户输入的任意模板与标签T_i，如果T_i∈MK或T_i∈TK，则生成基于查询词T_i，N的得分集S_i：$S_i=\{s_1^i,s_2^i,...s_r^i,...,s_x^i\}$，其中$s_r^i$是前述算分函数$f_{T_i}(VH_j)$基于查询词$T_i$计算得出的候选名字$N_r$的分值；

[0075] 将T中n个元素依次作为查询词，由$f_{T_i}(VH_j)$得到N的总体得分集S：

$$S=\{\sum_{i=1}^{n}s_1^i,\sum_{i=1}^{n}s_2^i,...,\sum_{i=1}^{n}s_r^i,....,\sum_{i=1}^{n}s_x^i\};$$

[0076] 将N中候选名字按照S中对应得分降序排列后，生成候选名字序列提供给使用者。

[0077] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内；因此，本发明的保护范围应该以权利要求书的保护范围为准。

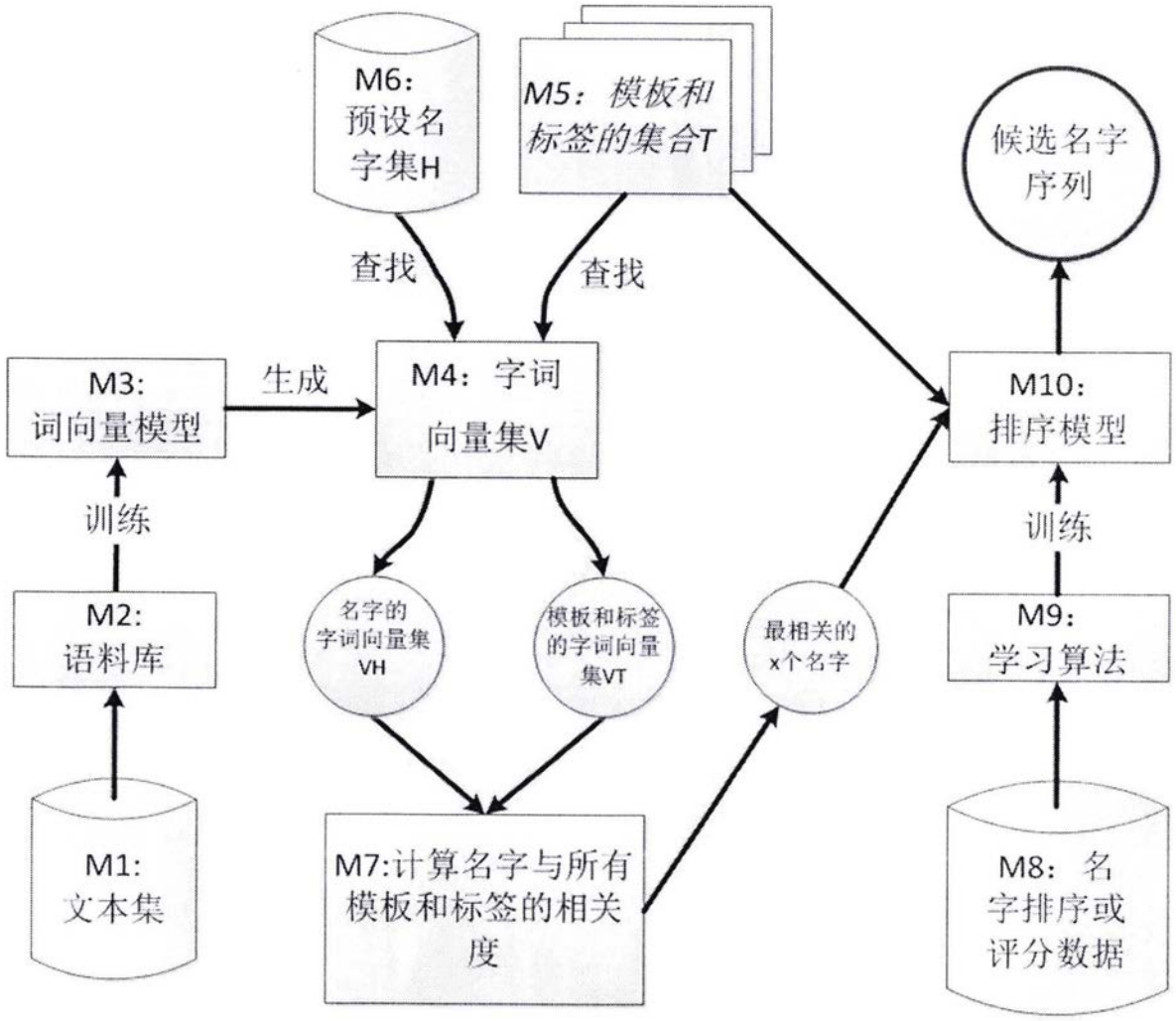

M6：
预设名
字集H
M5：模板和
标签的集合T
候选名字
序列
查找
查找
M3:
词向量模型
生成
M4：字词
向量集V
M10：
排序模型
训练
训练
M2:
语料库
名字的
字词向量集
VH
模板和标签
的字词向量
集VT
最相关的
x个名字
M9：
学习算法
M1:
文本集
M7:计算名字与所有
模板和标签的相关
度
M8：名
字排序或
评分数据

图1

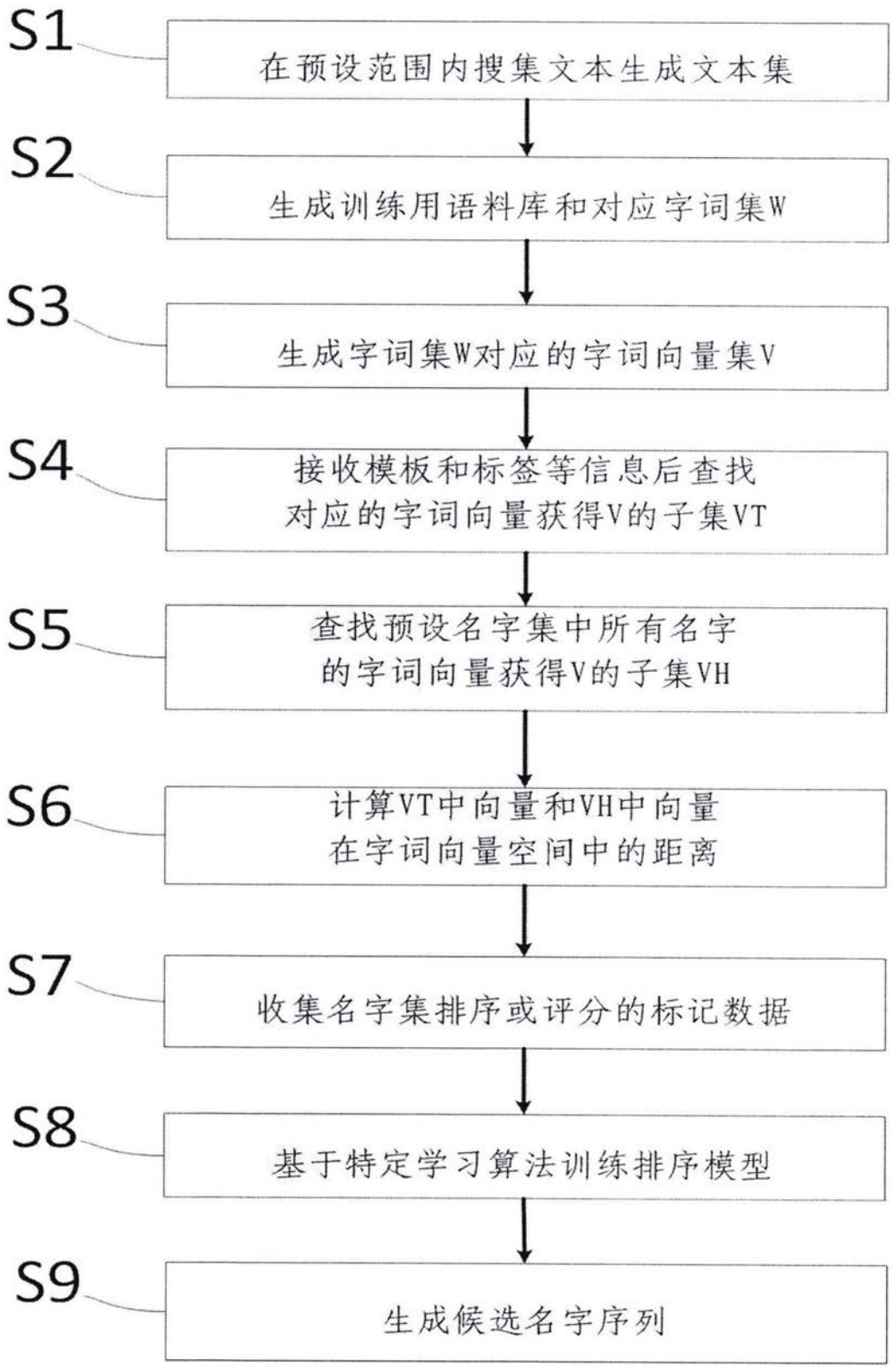

S1
在预设范围内搜集文本生成文本集
S2
生成训练用语料库和对应字词集W
S3
生成字词集W对应的字词向量集V
S4
接收模板和标签等信息后查找
对应的字词向量获得V的子集VT
S5
查找预设名字集中所有名字
的字词向量获得V的子集VH
S6
计算VT中向量和VH中向量
在字词向量空间中的距离
S7
收集名字集排序或评分的标记数据
S8
基于特定学习算法训练排序模型
S9
生成候选名字序列

图2