(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 104572624 A
(43)申请公布日 2015.04.29

(21)申请号 201510027487.7

(22)申请日 2015.01.20

(71)申请人 浙江大学
地址 310058 浙江省杭州市西湖区余杭塘路866号

(72)发明人 张引 魏宝刚 庄越挺 黎磊 姚亮

(74)专利代理机构 杭州求是专利事务所有限公司 33200
代理人 邱启旺

(51)Int.Cl.
G06F 17/27(2006.01)
G06F 17/30(2006.01)

权利要求书1页 说明书3页 附图3页

(54)发明名称

一种基于词向量发现单味药与疾病之间的治疗关系的方法

(57)摘要

本发明公开了一种基于词向量的单味药与疾病之间的治疗关系的发现方法。首先需要选择训练集，本发明采用《中华本草》书籍中8980味药作为治疗关系的主体，对于其主治描述提取疾病概念，作为治疗关系的客体，形成“药、治疗、疾病”的三元组，其次采用Google公布的Word2Vec工具作为词向量训练工具，百度百科资源作为训练语料，最后利用训练得到的词向量利用SVM训练出所需要的模型。输入单味药和疾病，该模型可以判断二者是否具有治疗关系。

CN 104572624 A

1.一种基于词向量的单味药与疾病之间的治疗关系的发现方法，其特征在于，包括以下步骤：

（1）对《中华本草》进行 OCR 处理，提取其主治属性；

（2）对主治属性进行三次预处理，第一次预处理按照标点符号分割，得到第一次候选集；第二次预处理将第一次得到的候选集中的所有词汇作为关键字访问百度百科、互动百科以及维基百科，若三者其一包含该关键字的页面，即认为该关键字是某种疾病，加入到疾病集合中，否则加入第二次候选集中；第三次预处理首先利用语法分析器对第二次候选集的词汇进行语法分析，找出结果为形容词 + 名词的形式，将其名词部分作为关键字访问百度百科、互动百科以及维基百科，若三者其一包含该关键字的页面，即认为该形容词 + 名词是某种疾病的具体形式，同样加入到疾病集合中，其余的单词做舍弃处理；经过三次预处理，构造出药与疾病的治疗关系三元组；

（3）将百科数据利用 CRF 模型与最长单词匹配方法相组合进行分词，同时过滤掉停用词、介词和数量词等无用词项，构建词向量的训练集；利用 Word2Vec（google 的开源工具）构造出词向量矩阵，即对每一个单词，用一个向量来表示；

（4）针对步骤 3 得到的三元组，找出药和疾病分别对应的词向量，按照单味药向量减去疾病向量的方式构造治疗关系的词向量；

（5）将步骤 4 构造的治疗关系词向量作为训练元组，其向量维数作为 SVM 的特征空间，利用 SVM 进行训练 ，得到训练模型；

（6）输入单味药和疾病，在步骤 3 构造的词向量矩阵中找到单味药和疾病分别对应的词向量，用单味药的词向量减去疾病的词向量得到关系向量作为步骤 5 训练出的模型的输入，根据训练模型输出结果判断二者是否含有治疗关系。

一种基于词向量发现单味药与疾病之间的治疗关系的方法

技术领域

[0001] 本发明涉及中医药中单味药与疾病之间治疗关系发现领域，是中医药与计算机科学相结合相交叉的产物，尤其涉及一种基于词向量以及 SVM 对于治疗关系发现的方法。

背景技术

[0002] 中医药领域中单味药与疾病之间存在治疗关系这是有据可循的，通过权威书籍和教材都可以查询得到，然而如何发现更多的治疗关系却一直没有一个有效的方法。随着计算机科学的飞速发展，机器学习方法的不断深化与完善为解决中医药领域问题提供了新的思路。特别是词向量的提出，针对每个词都有个向量空间，极大扩展了单词含义，并且单词向量的差值也有一定含义，为关系发现奠定了基础。

发明内容

[0003] 本发明的目的在于针对现有技术的不足，提供一种基于词向量发现单味药与疾病之间的治疗关系的方法，利用机器学习的方式来发现中医药领域中药物与疾病之间的治疗关系。

[0004] 本发明的目的是通过以下技术方案来实现的：一种基于词向量的单味药与疾病之间的治疗关系的发现方法，包括以下步骤：

[0005] (1) 对《中华本草》进行 OCR 处理，提取其主治属性；

[0006] (2) 对主治属性进行三次预处理，第一次预处理按照标点符号分割，得到第一次候选集；第二次预处理将第一次得到的候选集中的所有词汇作为关键字访问百度百科、互动百科以及维基百科，若三者其一包含该关键字的页面，即认为该关键字是某种疾病，加入到疾病集合中，否则加入第二次候选集中；第三次预处理首先利用语法分析器对第二次候选集的词汇进行语法分析，找出结果为形容词 + 名词的形式，将其名词部分作为关键字访问百度百科、互动百科以及维基百科，若三者其一包含该关键字的页面，即认为该形容词 + 名词是某种疾病的具体形式，同样加入到疾病集合中，其余的单词做舍弃处理；经过三次预处理，构造出药与疾病的治疗关系三元组；

[0007] (3) 将百科数据利用 CRF 模型与最长单词匹配方法相组合进行分词，同时过滤掉停用词、介词和数量词等无用词项，构建词向量的训练集；利用 Word2Vec(google 的开源工具）构造出词向量矩阵，即对每一个单词，用一个向量来表示；

[0008] (4) 针对步骤 3 得到的三元组，找出药和疾病分别对应的词向量，按照单味药向量减去疾病向量的方式构造治疗关系的词向量；

[0009] (5) 将步骤 4 构造的治疗关系词向量作为训练元组，其向量维数作为 SVM 的特征空间，利用 SVM 进行训练，得到训练模型；

[0010] (6) 输入单味药和疾病，在步骤 3 构造的词向量矩阵中找到单味药和疾病分别对应的词向量，用单味药的词向量减去疾病的词向量得到关系向量作为步骤 5 训练出的模型的输入，根据训练模型输出结果判断二者是否含有治疗关系。

[0011] 本发明的有益效果：本发明从权威书籍中经过三遍处理步骤得到标准“单味药、治疗、疾病”三元组，利用百度百科数据和 google 开源工具 word2vec 训练出词向量，将三元组与词向量相结合利用 SVM 分类器进行训练，最终准确的判断了单味药与疾病是否存在治疗关系，并可以有效的揭示一些隐藏的治疗关系，对于中医药转接有很大的参考价值；同时，所阐述的方法具有一般性，只要训练数据集准备充分，可以适用于一般关系的挖掘。

附图说明

[0012] 图 1 为本发明方法整体流程图；

[0013] 图 2 以“紫野苏”为例，第一遍预处理过程图；

[0014] 图 3 以“紫野苏”为例，第二遍预处理过程图；

[0015] 图 4 以“紫野苏”为例，第三遍预处理过程图。

具体实施方式

[0016] 以下结合附图和具体实施例对本发明作进一步详细说明。

[0017] 如图 1 所示，本发明一种基于词向量发现单味药与疾病之间的治疗关系的方法，包括以下步骤：

[0018] (1) 对《中华本草》进行 OCR 处理，提取其主治属性；

[0019] (2) 对主治属性进行三次预处理，第一次预处理按照标点符号分割，得到第一次候选集，如图 2 所示，紫叶苏的主治属性为“主暑天感冒；头痛身重；夫汗恶寒；腹痛吐泻；水肿；疮痈肿毒；蛲虫病；阴道滴虫”，经过第一遍处理之后得到的候选集为“暑天感冒头痛身重夫汗恶寒腹痛吐泻水肿疮痈肿毒蛲虫病阴道滴虫”；第二次预处理将第一次得到的候选集中的所有词汇首先访问本地疾病数据库，若存在，则认为该关键字是疾病概念，否则，作为关键字访问百度百科、互动百科以及维基百科，若三者其一包含该关键字的页面，即认为该关键字是某种疾病，加入到疾病集合中，同时爬取该词条释义加入至本地数据库中，否则加入第二次候选集中，如图 3 所示，百科包括的词汇为“水肿”，“疮痈肿毒”，“蛲虫病”，这些词汇被加入到疾病集合中，剩余词汇加入到第二次候选集中；第三次预处理首先利用语法分析器对第二次候选集的词汇进行语法分析，找出结果为名词 + 动词和名词 + 名词的形式，对名词 + 动词格式的词汇，将其名词部分作为关键字访问百度百科、互动百科以及维基百科，若三者其一包含该关键字的页面，即认为该名词 + 动词格式的词汇是某种疾病的具体形式，加入到疾病集合中，同时爬取该词条释义加入至本地数据库中，对名词 + 名词格式的词汇，对每一个名词部分作为关键字访问百度百科、互动百科以及维基百科，若三者其一包含该关键字的页面，即认为该名词 + 名词格式的词汇是疾病概念的并列形式，也加入到疾病集合中同时爬取该词条释义加入至本地数据库中，其余的单词做舍弃处理，如图 4 所示，“腹痛吐泻”“暑天感冒”，“头痛身重”，“阴道滴虫”被解析为动词 + 名词的形式，分别为“腹痛吐泻”“暑天感冒”“头痛身重”“阴道滴虫”，百科包括的词汇为“吐泻”“感冒”“身重”“滴虫”因此这四个词汇也被加入到疾病候选集中，剩余的“夫汗恶寒”做丢弃处理；经过三次预处理，构造出药与疾病的治疗关系三元组；

[0020] (3) 将百科数据利用 CRF 模型与最长单词匹配方法相组合进行分词，同时过滤掉停用词、介词和数量词等无用词项，构建词向量的训练集；利用 Word2Vec(google 的开源工

具）构造出词向量矩阵，即对每一个单词，用一个向量来表示；

[0021] (4) 针对步骤 3 得到的三元组，找出药和疾病分别对应的词向量，按照单味药向量减去疾病向量的方式构造治疗关系的词向量；

[0022] (5) 将步骤 4 构造的治疗关系词向量作为训练元组，其向量维数作为 SVM 的特征空间，利用 SVM 进行训练，得到训练模型；

[0023] (6) 输入单味药和疾病，在步骤 3 构造的词向量矩阵中找到单味药和疾病分别对应的词向量，用单味药的词向量减去疾病的词向量得到关系向量作为步骤 5 训练出的模型的输入，根据训练模型输出结果判断二者是否含有治疗关系。也可以输入单味药，模型输出其可能治疗的疾病，举例如下表：

[0024]

药	标准集中已存在治疗关系的疾病	通过模型新发现的疾病
生姜	风寒感冒，恶寒发热，头痛，鼻塞，呕吐，痰饮，咳喘，泄泻	咳嗽，哮喘，胃痛
白术	小便不利，水肿，痰饮，气虚自汗	腹胀泄泻，脾虚
独活	风寒湿痹腰膝疼痛头痛	风寒
谷芽	胀满泄泻脾虚脚气	宿食不化

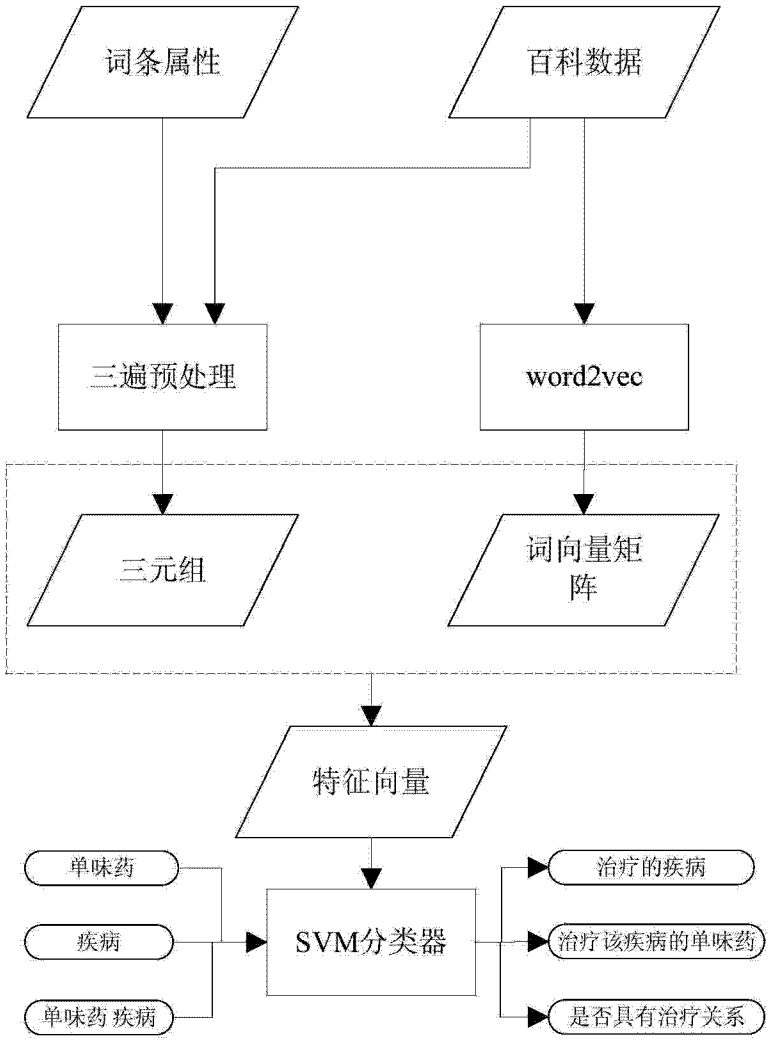
词条属性
百科数据
三遍预处理
word2vec
三元组
词向量矩阵
特征向量
单味药
疾病
单味药 疾病
SVM分类器
治疗的疾病
治疗该疾病的单味药
是否具有治疗关系

图 1

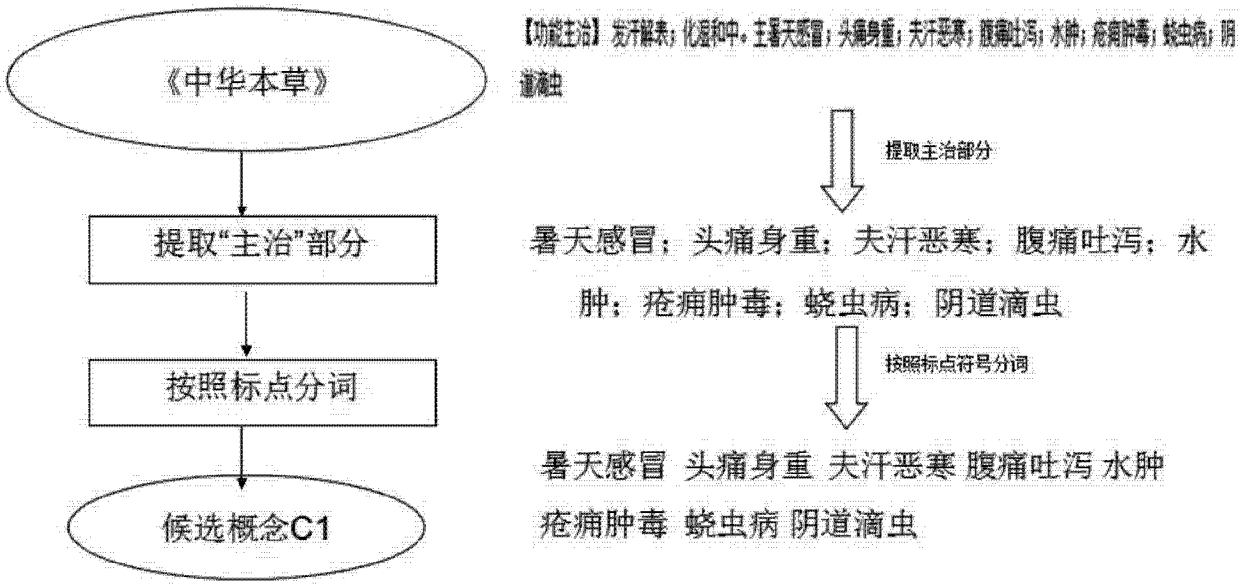
《中华本草》
提取“主治”部分
按照标点分词
候选概念C1
【功能主治】发汗解表；化湿和中。主暑天感冒；头痛身重；夫汗恶寒；腹痛吐泻；水肿；疮痈肿毒；蛲虫病；阴道滴虫
提取主治部分
暑天感冒；头痛身重；夫汗恶寒；腹痛吐泻；水肿；疮痈肿毒；蛲虫病；阴道滴虫
按照标点符号分词
暑天感冒 头痛身重 夫汗恶寒 腹痛吐泻 水肿 疮痈肿毒 蛲虫病 阴道滴虫

图 2

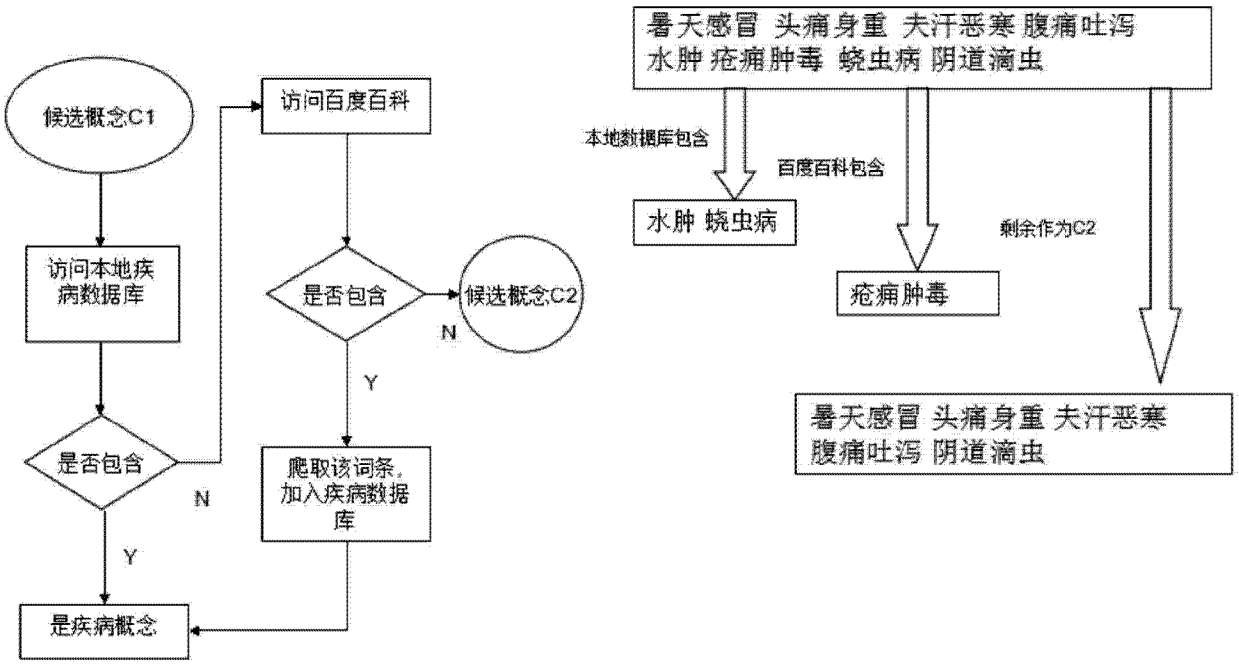
候选概念C1
访问本地疾病数据库
是否包含
N
Y
访问百度百科
是否包含
候选概念C2
N
Y
爬取该词条，加入疾病数据库
是疾病概念
暑天感冒 头痛身重 夫汗恶寒 腹痛吐泻 水肿 疮痈肿毒 蛲虫病 阴道滴虫
本地数据库包含
百度百科包含
剩余作为C2
水肿 蛲虫病
疮痈肿毒
暑天感冒 头痛身重 夫汗恶寒 腹痛吐泻 阴道滴虫

图 3

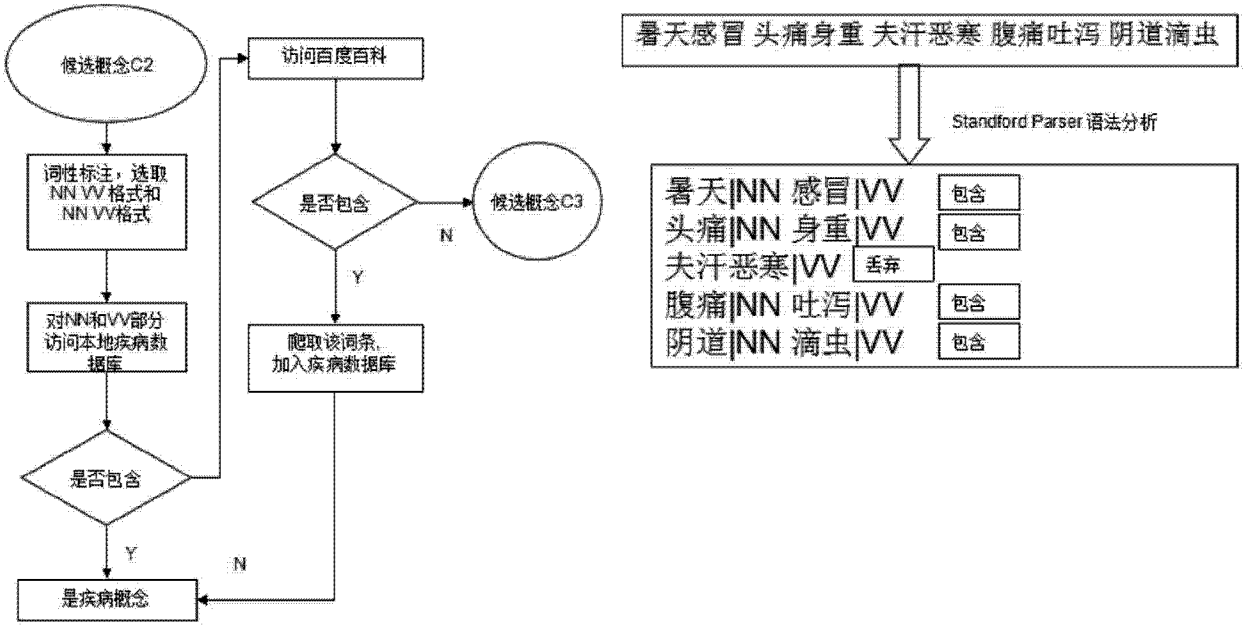
假选概念C2
词性标注，选取NN VV格式和NN VV格式
对NN和VV部分访问本地疾病数据库
是否包含
Y
是疾病概念
N
访问百度百科
是否包含
候选概念C3
N
Y
爬取该词条，加入疾病数据库
暑天感冒 头痛身重 夫汗恶寒 腹痛吐泻 阴道滴虫
Standford Parser 语法分析
暑天|NN 感冒|VV
包含
头痛|NN 身重|VV
包含
夫汗恶寒|VV
舍弃
腹痛|NN 吐泻|VV
包含
阴道|NN 滴虫|VV
包含

图 4