(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 108986907 A
(43)申请公布日 2018.12.11

(21)申请号 201810818355.X

(22)申请日 2018.07.24

(71)申请人 郑州大学第一附属医院
地址 450001 河南省郑州市二七区建设东路50号

(72)发明人 翟运开 赵杰 石金铭 陈昊天 孙东旭 卢耀恩 陈保站 王振博

(74)专利代理机构 常州佰业腾飞专利代理事务所(普通合伙) 32231
代理人 张宇

(51)Int.Cl.
G16H 40/67(2018.01)
G16H 10/60(2018.01)

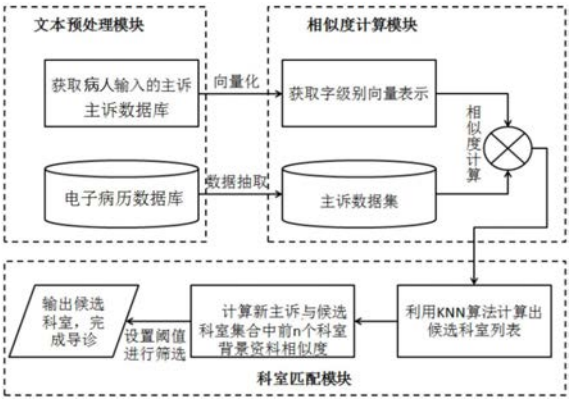
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种基于KNN算法的远程医疗自动分诊方法

(57)摘要

本发明公开了一种基于KNN算法的远程医疗自动分诊方法，属于大数据技术领域，包括建立文本预处理模块、相似度计算模块和科室匹配模块，采用Jaccard相似系数作为KNN算法中相似性度量算法，并采用KNN算法进行分诊，解决了远程医疗系统中多病症快速准确分诊的技术问题，本发明利用数据挖掘技术中的K最近邻(kNN，k-Nearest Neighbor)算法，根据提交的患者主诉，计算患者主诉与数据库中其他主诉的语义相似度，实现自动分诊，针对远程医疗服务中的分诊需求，获取患者主诉与就诊科室数据，通过计算不同患者主诉之间的相似度，基于KNN分类算法实现了智能分诊，在用户提交远程医疗病历资料后，引导用户进行正确就诊，提高了分诊的速度和准确度，降低了维护难度。



CN 108986907 A

1.一种基于KNN算法的远程医疗自动分诊方法,其特征在于:包括如下步骤:

步骤1:建立分布式云服务器平台,建立若干远程医疗客户端,所有远程医疗客户端均通过互联网与分布式云服务器平台通信,在分布式云服务器平台中建立文本预处理模块、相似度计算模块和科室匹配模块;

步骤2:远程医疗客户端用于收集病人的主诉信息和电子病历,并将主诉信息和电子病历上传到云服务器平台;

步骤3:云服务器平台的文本预处理模块对病人的主诉信息和电子病历进行文本预处理,其步骤如下:

步骤A1:文本预处理模块设立主诉数据库和电子病历数据库,主诉数据库用于存储病人的主诉信息,电子病历数据库用于存储病人的电子病历;

步骤A2:文本预处理模块建立停用词表,在停用词表中预存停用词;

步骤A3:文本预处理模块将主诉信息和电子病历中的文本转化成为主诉文本向量:文本预处理模块根据停用词表将主诉信息和电子病历中的停用词删除,并利用正则表达式删除时间词,合并主诉信息和电子病历中的剩余文本,以字为单位构件文本向量,即,生成主诉文本向量;

步骤4:文本预处理模块将主诉文本向量上传给相似度计算模块,相似度计算模块对主诉文本向量进行相似度分析,其步骤如下:

步骤B1:在相似度计算模块中建立主诉数据集,主诉数据集中包含数个表达病症的字向量,以病症对应的科室为归类规则,对字向量进行归类,生成数个字向量集;

步骤B2:相似度计算模块获取文本预处理模块生成的主诉文本向量后,以字为最小单位,将主诉文本向量与主诉数据集中的字向量进行相似度对比;找出所有与主诉文本向量相似的字向量;

步骤5:根据KNN算法计算出候选科室列表,其步骤如下:

步骤C1:根据KNN算法,找出与主诉文本向量具有最多的相似度字向量的字向量集;

步骤C2:根据步骤C1得到的字向量集查找对应的科室,将该科室作为伪候选科室;

步骤C3:根据步骤C1和步骤C2的方法,选择出数个伪候选科室,并将所有伪候选科室按相似度字向量的数量进行顺序排序;

步骤C4:选择相似度字向量的数量最高的3个伪候选科室作为候选科室,并将这3个候选科室作为分诊结果输出;

步骤6:结束分诊。

2.如权利要求1所述的一种基于KNN算法的远程医疗自动分诊方法,其特征在于:在执行步骤B1时,每一个科室均提供一个科室背景资料,科室背景资料通过统计每个科室对应的病症主诉,由对应病症主诉集合关键词构建科室背景资料,在对主诉数据集中的表达病症的字向量进行归类时,采用相似度对比的方式,对所述表达病症的字向量与科室背景资料中的病症主诉集合关键词继续对比,将对比结果作为主诉数据集中的表达病症的字向量进行归类的依据。

3.如权利要求1所述的一种基于KNN算法的远程医疗自动分诊方法,其特征在于:在执行步骤A3时,所述时间词为主诉信息中表达时间的词汇。

4.如权利要求1所述的一种基于KNN算法的远程医疗自动分诊方法,其特征在于:所述

远程医疗客户端为电脑、远程医疗终端或挂号终端。

一种基于KNN算法的远程医疗自动分诊方法

技术领域

[0001] 本发明属于大数据技术领域,特别涉及一种基于KNN算法的远程医疗自动分诊方法。

背景技术

[0002] 随着互联网医疗的快速发展,远程医疗作为其中的重要应用,得到了快速发展,在远程医疗的诸多应用中,诸如远程会诊、远程门诊等应用,需要人工选择相应科室,对于远程医疗服务的申请方而言,由于不同医院间的科室设置不同,在提交远程电子病历后,通常需要根据经验手动选择相关科室,经常出现错选的状况。

[0003] 自动分诊旨在根据患者情况为患者指引正确科室,国外发达国家对于该应用需求不大,因此相关研究较少,主要相关工作是针对自动导医进行研究,导医是指根据患者症状判断其病症并引导至相关科室。

[0004] 目前,自动导医系统分为两种:

[0005] 一种是基于专家系统的自动导医系统,专家系统(ExpertSystem,ES)是一种模拟一个领域内专家的思维进行推理判断以解决某些问题的计算机系统。INTERNIST1是由匹兹堡大学Miller等人在1982年开发的计算机辅助诊断工具,根据领域专业人员预先输入的规则与数据库,INTERNIST1可以根据患者症状判断患者疾病,从而达到对患者的诊断过程。斯坦福大学的Shortliffe等人开发了MYCIN系统,用于鉴别细菌感染及治疗。专家系统能够有效的解决大多数领域内人们所不能有效解决的问题,但将其应用在自动导诊领域内也有明显的缺点:由于推理规则的复杂性,推理时会有组合爆炸的问题;专家知识库如果过于庞大,会明显的降低时效性;专家知识库需要有专业人员定期进行维护,维护成本较高。

[0006] 另外一种是基于相似度计算的导医系统,通过计算患者症状与疾病症状的相似度来计算可能患有某种疾病的概率。现有技术中提出了改进的TF-IDF算法,根据症状的用户关注度来计算症状的权重,使其更适用于医疗诊断,但是基于相似度计算的方法计算速度虽快,但是未考虑多病症同时出现的情况。

发明内容

[0007] 本发明的目的是提供一种基于KNN算法的远程医疗自动分诊方法,解决了多病症快速准确分诊的技术问题。

[0008] 为实现上述目的,本发明采用以下技术方案:

[0009] 一种基于KNN算法的远程医疗自动分诊方法,包括如下步骤:

[0010] 步骤1:建立分布式云服务器平台,建立若干远程医疗客户端,所有远程医疗客户端均通过互联网与分布式云服务器平台通信,在分布式云服务器平台中建立文本预处理模块、相似度计算模块和科室匹配模块;

[0011] 步骤2:远程医疗客户端用于收集病人的主诉信息和电子病历,并将主诉信息和电子病历上传到云服务器平台;

[0012] 步骤3:云服务器平台的文本预处理模块对病人的主诉信息和电子病历进行文本预处理,其步骤如下:

[0013] 步骤A1:文本预处理模块设立主诉数据库和电子病历数据库,主诉数据库用于存储病人的主诉信息,电子病历数据库用于存储病人的电子病历;

[0014] 步骤A2:文本预处理模块建立停用词表,在停用词表中预存停用词;

[0015] 步骤A3;文本预处理模块将主诉信息和电子病历中的文本转化成为主诉文本向量:文本预处理模块根据停用词表将主诉信息和电子病历中的停用词删除,并利用正则表达式删除时间词,合并主诉信息和电子病历中的剩余文本,以字为单位构件文本向量,即,生成主诉文本向量;

[0016] 步骤4:文本预处理模块将主诉文本向量上传给相似度计算模块,相似度计算模块对主诉文本向量进行相似度分析,其步骤如下:

[0017] 步骤B1:在相似度计算模块中建立主诉数据集,主诉数据集中包含数个表达病症的字向量,以病症对应的科室为归类规则,对字向量进行归类,生成数个字向量集;

[0018] 步骤B2:相似度计算模块获取文本预处理模块生成的主诉文本向量后,以字为最小单位,将主诉文本向量与主诉数据集中的字向量进行相似度对比;找出所有与主诉文本向量相似的字向量;

[0019] 步骤5:根据KNN算法计算出候选科室列表,其步骤如下:

[0020] 步骤C1:根据KNN算法,找出与主诉文本向量具有最多的相似度字向量的字向量集;

[0021] 步骤C2:根据步骤C1得到的字向量集查找对应的科室,将该科室作为伪候选科室;

[0022] 步骤C3:根据步骤C1和步骤C2的方法,选择出数个伪候选科室,并将所有伪候选科室按相似度字向量的数量进行顺序排序;

[0023] 步骤C4:选择相似度字向量的数量最高的3个伪候选科室作为候选科室,并将这3个候选科室作为分诊结果输出;

[0024] 步骤6:结束分诊。

[0025] 在执行步骤B1时,每一个科室均提供一个科室背景资料,科室背景资料通过统计每个科室对应的病症主诉,由对应病症主诉集合关键词构建科室背景资料,在对主诉数据集中的表达病症的字向量进行归类时,采用相似度对比的方式,对所述表达病症的字向量与科室背景资料中的病症主诉集合关键词继续对比,将对比结果作为主诉数据集中的表达病症的字向量进行归类的依据。

[0026] 在执行步骤A3是,所述时间词为主诉信息中表达时间的词汇。

[0027] 所述远程医疗客户端为电脑、远程医疗终端或挂号终端。

[0028] 本发明所述的一种基于KNN算法的远程医疗自动分诊方法,解决了多病症快速准确分诊的技术问题,本发明利用数据挖掘技术中的K最近邻(kNN,k-Nearest Neighbor)算法,根据患者主诉,计算患者主诉与数据库中其他主诉的语义相似度,实现自动分诊,针对智能导诊这一需求,从电子病历中抽取患者主诉与就诊科室数据,通过计算不同患者主诉之间的相似度,基于KNN分类算法实现了智能导诊,在用户提交远程医疗病历资料后,引导用户进行正确就诊,提高了分诊的速度和准确度,降低了维护难度。

附图说明

[0029] 图1是本发明的流程图；

[0030] 图2是本发明的KNN算法示意图。

具体实施方式

[0031] 如图1和图2所示的一种基于KNN算法的远程医疗自动分诊方法，包括如下步骤：

[0032] 步骤1：建立分布式云服务器平台，建立若干远程医疗客户端，所有远程医疗客户端均通过互联网与分布式云服务器平台通信，在分布式云服务器平台中建立文本预处理模块、相似度计算模块和科室匹配模块；

[0033] 步骤2：远程医疗客户端用于收集病人的主诉信息和电子病历，并将主诉信息和电子病历上传到云服务器平台；

[0034] 步骤3：云服务器平台的文本预处理模块对病人的主诉信息和电子病历进行文本预处理，其步骤如下：

[0035] 步骤A1：文本预处理模块设立主诉数据库和电子病历数据库，主诉数据库用于存储病人的主诉信息，电子病历数据库用于存储病人的电子病历；

[0036] 步骤A2：文本预处理模块建立停用词表，在停用词表中预存停用词；

[0037] 步骤A3；文本预处理模块将主诉信息和电子病历中的文本转化成为主诉文本向量：文本预处理模块根据停用词表将主诉信息和电子病历中的停用词删除，并利用正则表达式删除时间词，提高文本向量的表现力，合并主诉信息和电子病历中的剩余文本，以字为单位构件文本向量，即，生成主诉文本向量；

[0038] 所述时间词如主诉信息中表达时间的词汇：“头痛2天”中的“2天”，文本预处理模块首先建立时间词词库，预先收录相关时间词。

[0039] 通常自然语言处理的第一步是分词，而针对主诉信息的分词效果较差，所以本发明以字代词，以字为单位构建文本向量；

[0040] 步骤4：文本预处理模块将主诉文本向量上传给相似度计算模块，相似度计算模块对主诉文本向量进行相似度分析，其步骤如下：

[0041] 步骤B1：在相似度计算模块中建立主诉数据集，主诉数据集中包含数个表达病症的字向量，以病症对应的科室为归类规则，对字向量进行归类，生成数个字向量集；

[0042] 步骤B2：相似度计算模块获取文本预处理模块生成的主诉文本向量后，以字为最小单位，将主诉文本向量与主诉数据集中的字向量进行相似度对比；找出所有与主诉文本向量相似的字向量；

[0043] 步骤5：根据KNN算法计算出候选科室列表，其步骤如下：

[0044] 步骤C1：根据KNN算法，找出与主诉文本向量具有最多的相似度字向量的字向量集；

[0045] 步骤C2：根据步骤C1得到的字向量集查找对应的科室，将该科室作为伪候选科室；

[0046] 步骤C3：根据步骤C1和步骤C2的方法，选择出数个伪候选科室，并将所有伪候选科室按相似度字向量的数量进行顺序排序；

[0047] 步骤C4：选择相似度字向量的数量最高的3个伪候选科室作为候选科室，并将这3

个候选科室作为分诊结果输出；

[0048] 步骤6:结束分诊。

[0049] 本发明利用KNN算法计算出候选科室列表,对于候选科室列表中的候选科室,计算用户所输入主诉与候选科室集合中前n个科室背景资料的相似度,根据相似度进行排序,其中科室背景资料通过统计每个科室对应主诉,由对应主诉集合关键词构建科室背景资料。鉴于用户主诉通常较短,且同一病症可能对应不同的科室,只输出一个科室会导致准确率较低,故输出相似度最高的3个候选科室,完成导诊流程。

[0050] KNN算法具体为:K最近邻(k-Nearest Neighbor,KNN)分类算法是数据挖掘中最经典的算法之一,其基本思想是:如果一个样本在特征空间中的k个最相似(即,特征空间中最邻近)的样本中的大多数属于某一个类别,则该样本也属于这个类别。KNN算法中,所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。算法示意图如图2所示,对于用户输入的样本点Xi,计算出与其最相近的k个样本点,统计这k个样本点所属类别,出现最多的类即为样本点的类别。在图2中,选取了6个与样本点Xi最为相近的样本,其中3个属于类别W1,2个属于类别W2,1个属于类别W3,因此判断样本点Xi属于类别W1;在本发明中,所述样本点Xi即为主诉文本向量中的任意一个字向量Xi,类别W1、类别W2和类别W3即为相似度计算模块中的3个字向量集:W1、W2和W3,这3个字向量集分别对应三个科室,字向量A经过相似度对比后,对应在W1中有3个相似的字向量,在W2中有2个相似的字向量,在W3中有1个相似的字向量,那么字向量Xi属于W1所对应的科室。

[0051] 本发明采用KNN算法,无需训练,对于多分类问题表现较好。

[0052] 在执行步骤4对主诉文本向量进行相似度分析时,采用考虑到主诉文本通常偏短且智能导诊对算法效率要求较高,采用Jaccard相似系数作为相似性度量算法:Jaccard相似系数主要用于计算符号度量或布尔值度量的个体间的相似度,其计算公式如下：

[0053] $$\mathrm{Jaccard}(X,Y)=\frac{X\cap Y}{X\cup Y}$$ ；

[0054] 对于经过预处理的主诉文本,构建词表作为分母,用词袋模型生成文本向量X＝{$x_1,x_2,x_3,\ldots,x_n$}与文本向量Y＝{$y_1,y_2,y_3,\ldots,y_m$},统计文本向量X＝{$x_1,x_2,x_3,\ldots,x_n$}与文本向量Y＝{$y_1,y_2,y_3,\ldots,y_m$}的交集作为分子,然后利用上述公式计算Jaccard相似。

[0055] 在执行步骤B1时,每一个科室均提供一个科室背景资料,科室背景资料通过统计每个科室对应的病症主诉,由对应病症主诉集合关键词构建科室背景资料,在对主诉数据集中的表达病症的字向量进行归类时,采用相似度对比的方式,对所述表达病症的字向量与科室背景资料中的病症主诉集合关键词继续对比,将对比结果作为主诉数据集中的表达病症的字向量进行归类的依据。

[0056] 在执行步骤A3是,所述时间词为主诉信息中表达时间的词汇。

[0057] 所述远程医疗客户端为电脑或远程医疗终端。

[0058] 本发明所述的一种基于KNN算法的远程医疗自动分诊方法,解决了多病症快速准确分诊的技术问题,本发明利用数据挖掘技术中的K最近邻(kNN,k-Nearest Neighbor)算法,根据患者主诉,计算患者主诉与数据库中其他主诉的语义相似度,实现自动分诊,针对智能导诊这一需求,从电子病历中抽取患者主诉与就诊科室数据,通过计算不同患者主诉

之间的相似度，基于KNN分类算法实现了智能导诊，在用户提交远程医疗病历资料后，引导用户进行正确就诊，提高了分诊的速度和准确度，降低了维护难度。

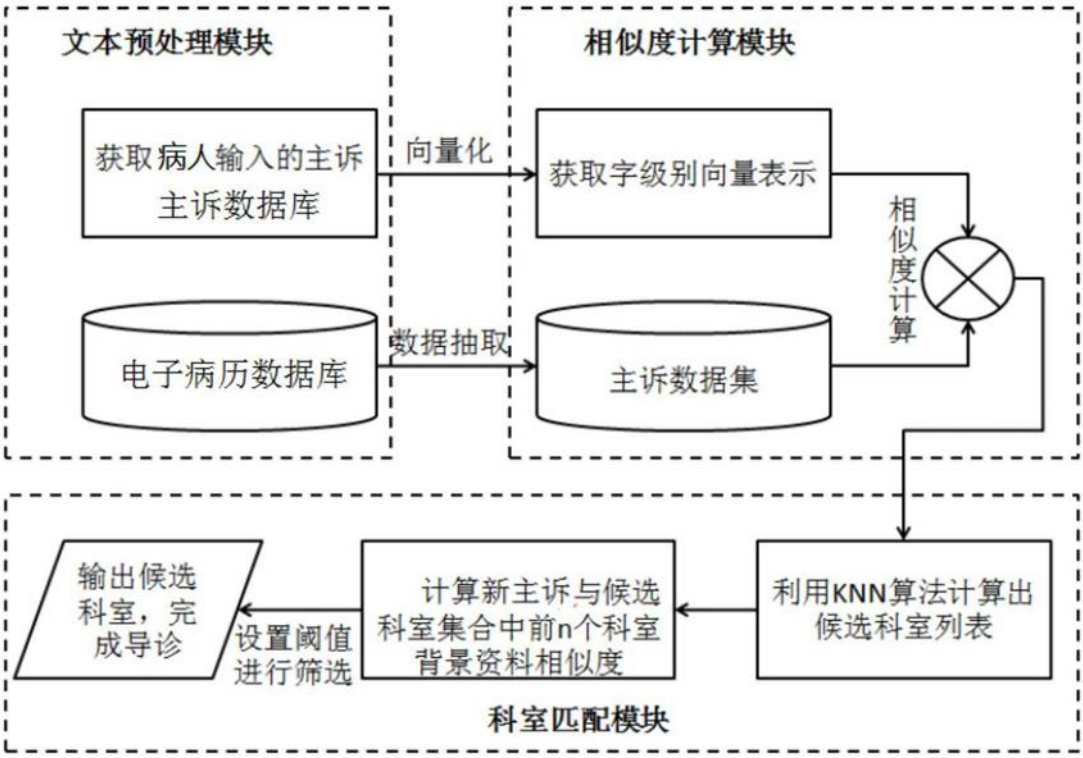

文本预处理模块
相似度计算模块
获取病人输入的主诉
主诉数据库
向量化
获取字级别向量表示
相似度计算
电子病历数据库
数据抽取
主诉数据集
输出候选科室，完成导诊
设置阈值进行筛选
计算新主诉与候选科室集合中前n个科室背景资料相似度
利用KNN算法计算出候选科室列表
科室匹配模块

图1

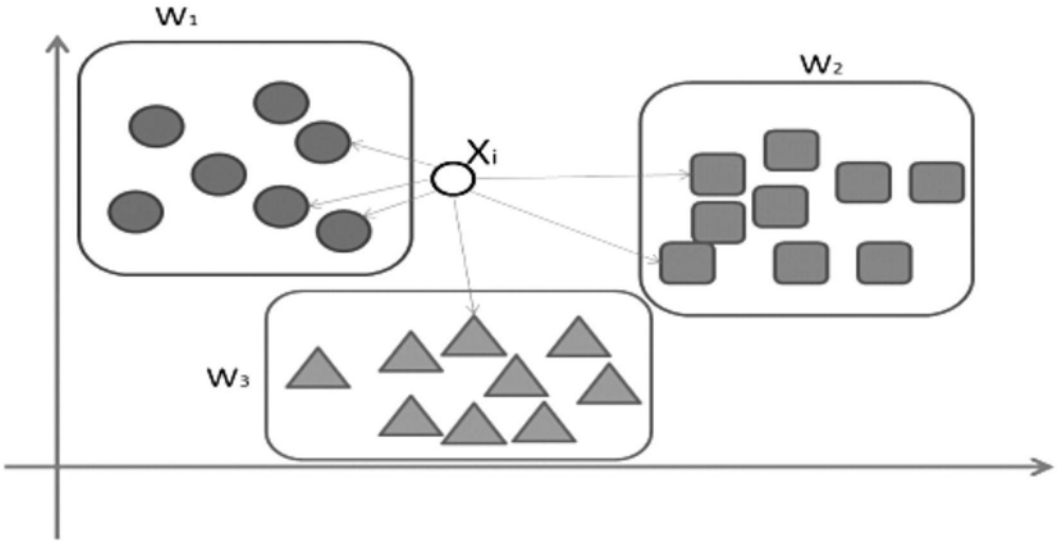

W1
W2
W3
Xi

图2