客服运维智能人机对话系统设计

王 玮,严文涛,苏 琦,刘 荫,于展鹏,殷齐林

(山东电力集团公司 信息通信公司,山东 济南 250000)

摘要: 在信息系统运维服务过程中,大量客户的咨询问题都是重复的,这为运维人员带来了很多重复性工作。基于文本分词的智能人机对话系统可以很好解决这个问题。提出一种利用运维语料库和基于中文文本分词的智能人机对话模型,该模型可以针对用户提出的问题,自动给出最优的解答,解决了人工运维过程中存在的速度慢、效率低的问题,实现了借助人工智能促进企业运维服务水平不断提升的目的。

关键词: 文本分词; 智能人机对话; 朴素贝叶斯分类; 客服运维

中图分类号: TP312.8 **文献标识码:** B **文章编号:** 1673 -5382(2017)02 -0083 -03

1 概述

在信息系统运维服务过程中,解决用户问题的及时性已经是企业提升服务质量的关键所在。当用户发起咨询时,解决问题是否及时会对用户满意度产生巨大影响。但是,由于用户咨询问题涉及面广且重复性问题较多,要达到用户满意则需配置数量较多的运维人员,不利于企业降低运行成本。大部分运维服务信息都是以文本信息的形式存在,海量运维客服数据中包含很多用户重复性咨询的问题,基于文本分词技术,可以对关键咨询的问题进行内容提取,识别出关键信息,通过编写专门的运维语料库,利用人工智能方法自动实现对问题的最优答案寻找。

从语法角度来看,汉语是一种以词根为特征的语言,这就决定了对其他语言有效的分词处理方法,不能直接在汉语分词处理中应用。汉语分词处理方法经过多年的发展,目前较为成熟的中文分词处理方法主要有 3 大类别: 基于词典分词法、基于统计分词法和基于规则分词法[1]。

基于词典分词法的核心技术是构建分词词典和定义词语匹配算法[2],词典中的词汇要保证它的代表性和完整性,词语匹配算法要考虑匹配的效率和速度以便于实际中应用[3]。长词优先法和短词优先法[4]是目前较为常见的词语匹配算法,长词优先法可以最大限度地解决分词后的歧义性,便于筛选出更多的专用词汇,其突出的问题是分词效率低下和分词后结构多种多样。

基于统计分词法是以词与词之间的相互关系作为依据进行分词,目前较常用的有 3 种模型算法: 词语互信息模型算法、N - gram 模型算法[5]和期望最大值 EM 模型算法[6]。

在词语互信息模型算法中,通过定义不同字符串之间的相关性来描述它们之间的互信息[7],对于字符串 c 和 d,其互信息可以描述为:

$$MI(c,d) = \log_2 \frac{p(c,d)}{p(c)\,p(d)}$$

其中,$p(c,d)$ 表示字符串 c 和 d 共同出现的概率,$p(c)$ 和 $p(d)$ 表示字符串 c 和 d 各自出现的概率。互信息 $MI(c,d)$ 表示字符串间的相关程度,若 $MI(c,d)$ 大于 0,则说明字符串 c 和 d 之间有可信的关系,两个字符串之间的结合度越高; 若 $MI(c,d)$ 等于 0,则说明字符串 c 和 d 之间的结合度不明确; 若 $MI(c,d)$ 小于 0,则说明字符串 c 和 d 之间没有可信的关系。

在 N - gram 模型算法中,文本的分词功能是通

收稿日期: 2017 -04 -03

作者简介: 王玮(1970 -),女,山东济南人,山东电力集团公司信息通信公司工程师,硕士。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

过利用 Viterbi 搜索算法来实现的，算法可以根据上下文自己组成词组，避免了人工预先设定词典带来的繁琐和局限性。

期望最大值 EM 模型算法是一种启发式迭代算法，根据词语在文本中的出现频率，按照极大似然估计原则，构建马尔科夫链，利用期望最大值 EM 算法进行迭代训练。

基于规则分词法[8]是让计算机通过模拟人类对语句的理解来对词语进行自动识别。

2 基于朴素贝叶斯的文本分类算法

朴素贝叶斯算法(Naive Bayesian)是贝叶斯数据挖掘分类方法中应用最为广泛的算法之一。文本的分类过程是将数据中预先分类过的文档作为训练集，在该训练集上运用贝叶斯数据挖掘分类技术建立分类算法模型，将模型应用于尚未进行分类的数据中进行分类。

在经典的分类模型中，决策树模型和朴素贝叶斯模型是应用最为广泛的两种分类算法。决策树模型是根据已知条件下各种事件发生的概率，通过构成决策树来获取净现值的期望值不小于零的概率。朴素贝叶斯分类器起源于古典数学理论，其所需的参数较少，与其他分类方法相比，误差率较小。但在文本分词的实际应用中并非如此，在实际计算先验概率时，朴素贝叶斯模型以概率的形式被加入到计算中去，而不是作为自然语言被人理解，所以，获得的结果是相同的。

基于朴素贝叶斯的文本分类是目前应用较为广泛的一种分词方法[9,10]。在文本分类过程中，对于任意给定的文本字符串 C，假定文本字符串中每个单词 w_i 和 w_j 相互之间始终是独立的，单词之间的关系如图 1 所示。

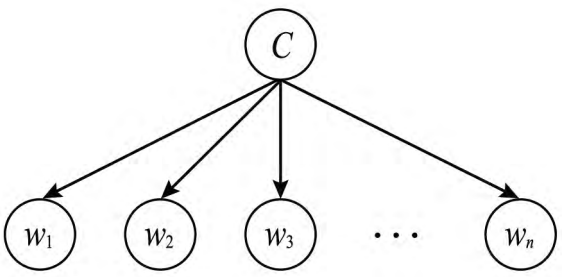

图 1 朴素贝叶斯文本分类模型

假定样本空间为 n 维特征空间向量，空间向量可表示为 $X=(x_1,x_2,\cdots,x_n)$，x_i 分别表示 n 个属性 $A_1,A_2,\cdots,A_n$ 的样本的 n 个度量，将样本划分为 m 类，可表示为 $C=(c_1,c_2,\cdots,c_m)$，假设每个类别 c_i 的先验概率为 $P(c_i)$，$i=1,2,\cdots,m$，这样，对于新样本 d，它属于 c_i 的条件概率为 $P(d|c_i)$，由此可以计算出 c_i 类的后验概率为 $P(c_i|d)$，由贝叶斯定理可知:

$$P(c_i|d)=\frac{P(d|c_i)P(c_i)}{p(d)}$$

根据上述公式，样本 d 可由其包含的特征表示，$d=(d_1,d_2,\cdots,d_j)$，j 表示 d 的属性个数，由此可得:

$$P(d|c_i)=P((d_1,d_2,\cdots,d_n)|c_i)=\prod_{k=1}^{n}P(d_k|c_i)$$

其中，求解 $P(d_k|c_i)$:

$$P(d_k|c_i)=\frac{TF(d_k,c_i)}{TF(c_i)}$$

其中，$TF(d_k,c_i)$ 表示属性 d_k 在类别 c_i 中出现的频次之和，$TF(c_i)$ 表示类别 c_i 的训练样本数。

3 客服运维智能人机对话系统

在客服运维智能人机对话系统中，首先需要进行人工训练，获取客户问题语句以及客服对应该问题语句的应答语句，并按对应关系存储在数据库中。训练可以随时进行，并不断完善。系统在收到用户问题语句后，将该问题语句与数据库中保存的用户问题语句进行查询匹配，若匹配成功，则直接输出对应的客服应答语句; 若匹配失败，则提示用户无答案。系统流程如图 2 所示。

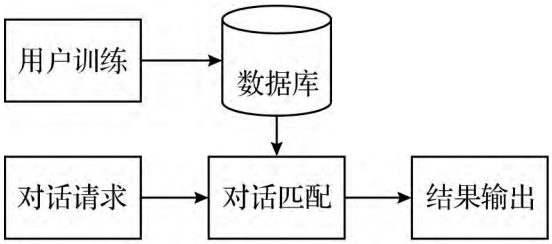

图 2 用户训练流程

根据以上原理，设计了客服运维智能人机对话系统模型，该模型利用多年来运维过程中积累的大量数据作为训练集样本，通过中文文本分类、分词技术来拆分确认客户问题对话主题，以确认的主题作为关键字在历史运维数据中搜寻适当的语句作为客服的应答。模型的系统流程如图 3 所示。

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

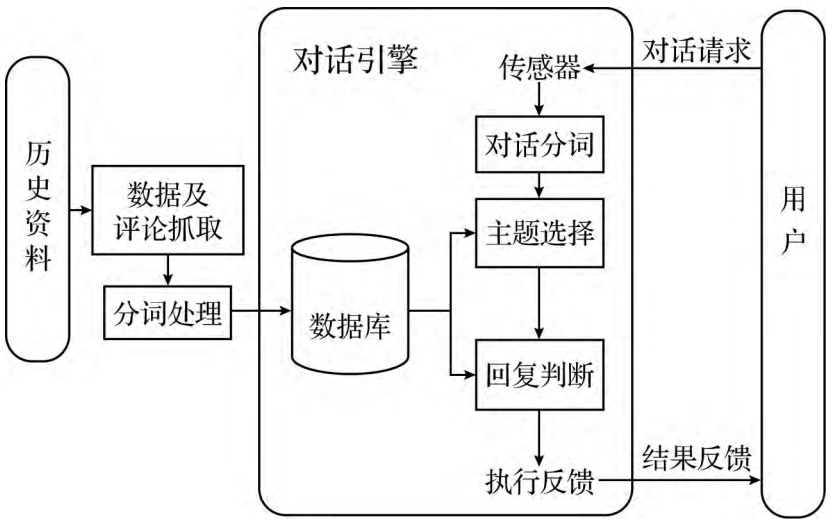

图 3　基于运维历史资料信息库的对话流程

当新用户的问题语句请求达到时,首先利用中文分词技术对用户的问题语句进行分词处理,拆分得到的分词依次作为主主题关键字,在主题与历史运维方案对应数据库中进行匹配检索,得到的包含主关键字的历史运维方案作为备选答案,然后再利用其它分词作为次主题关键字对备选答案集再进行筛选,筛选过程中对分词与主题匹配的历史运维方案进行统计计数,最终选择计数最大的一条历史运维方案作为答案语句,并将该语句回复给提出问题的用户。

4　结束语

人工智能以及数据挖掘的一个重要应用领域是人机对话系统,在分析研究了中文文本分词分类技术的基础上,提出了一种基于历史运维数据主题分析,结合问题文本分词与主题匹配频次统计的改进模型。经过一段时间的实践检验,该模型很好地解决了系统运维过程中大量的重复性工作,大大地提高了运维工作效率。在下一步的研究中,将对模型的预测和准确性等方面进行深入的探讨研究。

参考文献:

[1] 孙铁利,刘延吉. 中文分词技术的研究现状与困难 [J]. 信息技术,2009,9(7):187 -189.
[2] 傅立云,刘新. 基于词典的汉语自动分词算法的改进 [J]. 情报杂志,2006,4(1):40 -41.
[3] 黄河燕,李渝生. 上下文相关汉语自动分词及词法预处理算法 [J]. 应用科学学报,1999,17(2):148 -155.
[4] 文庭孝. 情报检索中汉语语词自动切分研究 [J]. 图书与情报,2001,11(2):57 -58.
[5] 吴应良,韦岗,李海洲. 一种基于 N -gram 模型和机器学习的汉语分词算法 [J]. 电子与信息学报,2001,23(11):1148 -1153.
[6] 李家福,张亚非. 基于 EM 算法的汉语自动分词方法 [J]. 情报学报,2002,21(3):269 -272.
[7] 费洪晓,康松林,朱小娟. 基于词频统计的中文分词的研究 [J]. 计算机工程与应用,2005,26(7):67 -68.
[8] 张江. 基于规则的分词方法 [J]. 计算机与现代化,2005,19(4):18 -20.
[9] 毛伟,徐蔚然,郭军. 基于 n -gram 语言模型和链状朴素贝叶斯分类器的中文文本分类系统 [J]. 中文信息学报,2006,20(3):29 -35.
[10] 李静梅,孙丽华,张巧荣. 一种文本处理中的朴素贝叶斯分类器 [J]. 哈尔滨工程大学学报,2003,24(1):71 -74.

(责任编辑　王国新)

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net