(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 109815415 A
(43)申请公布日 2019.05.28

(21)申请号 201910061663.7

(22)申请日 2019.01.23

(71)申请人 四川易诚智讯科技有限公司
地址 610041 四川省成都市武侯区武侯新城管委会科技园武青南路51号1幢2层1号
申请人 电子科技大学

(72)发明人 占梦来 王旭 张棚 罗爽 徐晓龙

(74)专利代理机构 成都虹盛汇泉专利代理有限公司 51268
代理人 王伟

(51)Int.Cl.
G06F 16/9536(2019.01)

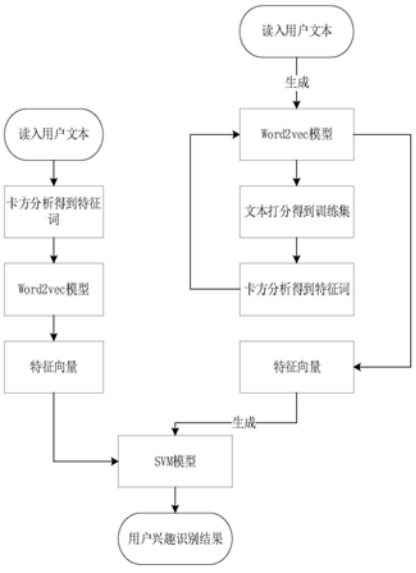
权利要求书2页 说明书8页 附图2页

(54)发明名称

基于卡方词频分析的社交媒体用户兴趣识别方法

(57)摘要

本发明公开一种基于卡方词频分析的社交媒体用户兴趣点识别方法，包括：S1、构建Word2vec模型；S2、基于Wrod2vec模型对文本进行打分，获取正例训练集、负例训练集；S3、采用卡方检验的原理计算正例训练集、负例训练集中的词汇卡方值，根据卡方值得到特征词汇，从而获取代表文本的特征向量；S4、采用步骤S3得到的代表文本的特征向量对SVM模型进行训练；S5、采用步骤S4训练好的SVM模型进行文本内容的人物兴趣识别；本发明采用了打分的方式来筛选训练文本，采用卡方统计的方法来提取关键词，并结合Word2vec模型向量方法，能显著提高了兴趣识别的准确率。



CN 109815415 A

1.基于卡方词频分析的社交媒体用户兴趣点识别方法,其特征在于,包括:

S1、构建Word2vec模型;

S2、基于Wrod2vec模型对文本进行打分,获取正例训练集、负例训练集;

S3、采用卡方检验的原理计算正例训练集、负例训练集中的词汇卡方值,根据卡方值得到特征词汇,从而获取代表文本的特征向量;

S4、采用步骤S3得到的代表文本的特征向量对SVM模型进行训练;

S5、采用步骤S4训练好的SVM模型进行文本内容的人物兴趣识别。

2.根据权利要求1所述的基于卡方词频分析的社交媒体用户兴趣点识别方法,其特征在于,步骤S1具体为:获取到的兴趣相关用户的所有推文,将其正则化训练成Word2vec模型。

3.根据权利要求2所述的基于卡方词频分析的社交媒体用户兴趣点识别方法,其特征在于,所述正则化处理过程为:

A1、去掉...与…结尾的单词;

A2、去掉所有格与助词缩写;

A3、去掉文本中的特有无意义词汇或内容;

A4、把用.连接的词汇中的.去除;

A5、只保留文中的基本和扩展拉丁字母,数字以及基本的ASCII码;

A6、将首字母大写连词分开;

A7、对词进行词干化以及去除停等词。

4.根据权利要求3所述的基于卡方词频分析的社交媒体用户兴趣点识别方法,其特征在于,步骤S2具体包括以下分步骤:

S21、根据Word2vec模型获取与兴趣关键词相似度最高的若干个单词;

S22、通过Word2vec模型计算索要筛选的文本内容中的语句词语与步骤S21中单词的相似度;

S23、取语句中每个词语与这些词计算的相似度值的最大值,累加作为语句的分值;

S24、将语句分值最高的N条文本内容作为正例训练集,将语句的分值最低的N条文本内容作为负例训练集。

5.根据权利要求4所述的基于卡方词频分析的社交媒体用户兴趣点识别方法,其特征在于,步骤S3具体为:

S31、将正例训练集、负例训练集文本内容中的词汇进行词频统计;

S32、根据步骤S31的词频统计采用卡方检验的原理计算卡方值;

S33、选取卡方值能代表与兴趣相关与不相关的词汇作为特征词汇;

S34、通过Word2vec模型获取步骤S33所述特征词汇的向量,通过权重分配累加作为文本的特征向量。

6.根据权利要求5所述的基于卡方词频分析的社交媒体用户兴趣点识别方法,其特征在于,卡方值的计算公式为:

$$X^2=\sum\frac{(A-T)^2}{T}$$

其中,X^2为卡方值,A为实际值,T为理论值。

7.根据权利要求6所述的基于卡方词频分析的社交媒体用户兴趣点识别方法，其特征在于，X^2用于衡量实际值与理论值的差异程度，包含了以下两个信息：

实际值与理论值偏差的绝对大小，差异程度与理论值的相对大小。

8.根据权利要求7所述的基于卡方词频分析的社交媒体用户兴趣点识别方法，其特征在于，步骤S33所选取的特征词汇为卡方值最大的词汇。

基于卡方词频分析的社交媒体用户兴趣识别方法

技术领域

[0001] 本发明属于数据挖掘领域,特别涉及一种社交媒体用户个性化推荐技术。

背景技术

[0002] 在网络信息大爆炸的时代的各种社交网络亦或是软件上,人们每天会接收或分享成百上千条信息,内容涉及方方面面,并且这些信息还有碎片化,更新速度快等特点,这样的情况加大了人们快速,准确获取信息以及知识的难度,从而有信息过载的问题。最典型的微博平台为例,商家,用户与平台,若能解决定向用户广告投放问题,则能使三方都受益,而这其中的重点问题就用户兴趣模型的建立以及用户兴趣的识别。从技术层面来讲,目前在用户兴趣模型建立以及用户兴趣识别方法上,有许多难题值得研究。不论是从商业角度还是技术角度来说,实现用户兴趣的识别是非常具有意义的研究课题。

[0003] 要准确描述用户兴趣并非易事,要想清楚描述用户兴趣,要对用户信息内容进行深层次的挖掘和分析,并建立相应的用户模型,目前的三种用户兴趣模型分别为典型用户模型与个体集合模型,显式模型与隐式模型,长期模型与短期模型。而三种兴趣表示方法为空间布尔模型,潜在语义索引模型和向量空间模型。近年来,研究人员在三种用户兴趣模型以及三种用户兴趣表示方法的基础上,发展出两种用户兴趣识别方法:即传统分类方法及主题模型方法。

[0004] (1)传统分类方法

[0005] 传统分类方法包含的内容很多,从最初的朴素贝叶斯,Logistic回归,到如今的机器学习中的svm,rnn,cnn等分类器,到最新的神经网络等。其流程都是训练模型再用于预测,主要的区别在于用于分类的依据即特征的选择上,目前的主流包括在用户文本内容信息,用户行为以及用户的社交关系上进行选择兴趣。具体体现在特征向量的表征上,包括文本内容的关键词选取以及权重分配,好友关系等等都可以加入到向量中。

[0006] (2)主题模型方法

[0007] 主题模型方法是早期的人物兴趣识别方法的主流,主要使用的就是LDA模型,即隐藏狄利克雷分布,是比较典型的词袋模型,模型中词语之间没有顺序关系,文档表示为词语的简单组合,不体现词语的先后顺序。文档中的单个词语在一般情况下,都由某个主题产生的,每个文档所对应的主题数量不是确定的,可以是一个,也可以是多个。目前的研究中有着不计其数的改进过的LDA主题模型,用于挖掘用户兴趣爱好,包括LSA,PLSA,Labled LDA等等。其中研究者常用的三种为Twitter Rank模型,Author-Topic模型和TwitterLDA模型。

发明内容

[0008] 为解决上述技术问题,本发明提出一种基于卡方词频分析的社交媒体用户兴趣识别方法,基于Word2vec模型以及SVM方法,并根据自创的寻找训练集的打分方法以及卡方词频分析的选取特征向量方法来进行社交媒体用户兴趣识别。

[0009] 本发明采用的技术方案为:基于卡方词频分析的社交媒体用户兴趣点识别方法,

包括：

[0010] S1、构建Word2vec模型；具体为：获取到的兴趣相关用户的所有推文，将其正则化训练成Word2vec模型。

[0011] S2、基于Wrod2vec模型对文本进行打分，获取正例训练集、负例训练集；

[0012] S3、采用卡方检验的原理计算正例训练集、负例训练集中的词汇卡方值，根据卡方值得到特征词汇，从而获取代表文本的特征向量；

[0013] S4、采用步骤S3得到的代表文本的特征向量对SVM模型进行训练；

[0014] S5、采用步骤S4训练好的SVM模型进行文本内容的人物兴趣识别。

[0015] 进一步地，步骤S1所述正则化处理过程为：

[0016] A1、去掉...与…结尾的单词；

[0017] A2、去掉所有格与助词缩写；

[0018] A3、去掉文本中的特有无意义词汇或内容；

[0019] A4、把用.连接的词汇中的.去除；

[0020] A5、只保留文中的基本和扩展拉丁字母，数字以及基本的ASCII码；

[0021] A6、将首字母大写连词分开；

[0022] A7、对词进行词干化以及去除停等词。

[0023] 进一步地，步骤S2具体包括以下分步骤：

[0024] S21、根据Word2vec模型获取与兴趣关键词相似度最高的若干个单词；

[0025] S22、通过Word2vec模型计算索要筛选的文本内容中的语句词语与步骤S21中单词的相似度；

[0026] S23、取语句中每个词语与这些词计算的相似度值的最大值，累加作为语句的分值；

[0027] S24、将语句分值最高的N条文本内容作为正例训练集，将语句的分值最低的N条文本内容作为负例训练集。

[0028] 进一步地，步骤S3具体为：

[0029] S31、将正例训练集、负例训练集文本内容中的词汇进行词频统计；

[0030] S32、根据步骤S31的词频统计采用卡方检验的原理计算卡方值；卡方值的计算公式为：

[0031] $$X^2 = \sum \frac{(A-T)^2}{T}$$

[0032] 其中，X^2为卡方值，A为实际值，T为理论值。

[0033] X^2用于衡量实际值与理论值的差异程度，包含了以下两个信息：

[0034] 实际值与理论值偏差的绝对大小，差异程度与理论值的相对大小。

[0035] S33、选取卡方值能代表与兴趣相关与不相关的词汇作为特征词汇；所选取的特征词汇为卡方值最大的词汇；

[0036] S34、通过Word2vec模型获取步骤S33所述特征词汇的向量，通过权重分配累加作为文本的特征向量。

[0037] 本发明的有益效果：本发明采用了打分的方式来筛选训练文本，从而准确定位识别的兴趣的范围内容；引入了卡方统计的方法来提取关键词，并结合Word2vec模型向量方

法,从而提高了兴趣识别的准确率;具体为:采用Word2vec模型:通过将大量的兴趣相关用户的推文用于训练Word2vec模型,尽可能将得到的模型中尽可能包含足够多社交媒体场景下该兴趣的相关词汇;从而用于后续的计算词汇相似度,词向量等等;采用文本内容打分方法,提取用于SVM的训练文本,得到在社交文本的内容级别上与兴趣在Word2vec的向量空间中与兴趣最相关的文本;从而用于SVM训练提高对兴趣识别的准确率;采用卡方检验的方法得到文本内容的关键词汇,从而用于得到训练文本的特征向量,这样提取的关键词可用于代表文本内容进行SVM训练以及后续的文本兴趣识别。

附图说明

[0038] 图1为本发明实施例提供的方案流程图;

[0039] 图2为本发明实施例提供的Word2vec模型的生成过程示意图;

[0040] 图3为本发明实施例提供的CBOW与Skip-gram两种模型。

具体实施方式

[0041] 首先对本发明涉及到的现有技术进行简要说明:

[0042] 1、基于用户用户文本内容信息的兴趣识别研究

[0043] 基于用户文本信息的用户兴趣挖掘识别是最显而易见的一种方法,用户文本包含的信息很多,例如微博文本,包含的用户标签,话题标签,以及其中的关键词等等,都是很重要的信息。Shepitsen(参考:J.Stoyanovich,S.Amer-Yahia,C.Marlow,C.Yu:Leveraging Tagging to Model User Interests in del.icio.us:AAAI Spring Symposium on Social Information Processing(2008))就通过对标签使用继承聚类的方法,来挖掘用户的兴趣。但是用户的标签有很多并非兴趣的表征,且噪声很大。所以更多的学者从文本内容挖掘关键词来代替标签。Thuy(参考:T.Vu,V.Perez:Interest Mining from User Tweets.CIKM,Oct.27-Nov.1,2013,SanFrancisco,CA,USA)等人用语言规则及关键字按掘算法来挖掘用户兴趣,通过制定语言模式匹配微博文本中兴趣相关的语句,然后结合TFIDF和TextRank算法提取topN个重要度高的关键词作为对用户兴趣的表征,实验表明当用户兴趣关键词为5个时,TextRank算法优于TFIDF算法,当用户兴趣关键词为10个时,TFIDF比TextRank算法效果好。宋巍等人(参考:宋巍,张宇,谢毓彬,等.基于微博分类的用户兴趣识别[J].智能计算机与应用,2013,3(4):80-83)利用卡方统计得到微博文本关键词作为局部信息,结合微博平台的全局信息,用svm分类器对每条微博进行兴趣识别。杜雨萌等人(参考:杜雨萌,张伟男,刘挺.基于主题增强卷积神经网络的用户兴趣识别[J].计算机研究与发展,2018,55(01):188-197)用Labeled LDA得到用户文档的兴趣主题关键词,再用cnn分类器对每条微博进行兴趣识别。朱凯歌(参考:朱凯歌.面向个性化服务的用户兴趣挖掘方法研究与实现[D].北京交通大学,2018)用Word2vec模型的方法提取文本关键词作为潜在兴趣,再将其映射到构建好的开放式标签空间,从而对用户兴趣进行描述。

[0044] 2、基于用户用户行为的兴趣挖掘识别研究

[0045] 基于用户行为的兴趣识别研究主要是集中在用户的搜索以及浏览行为上,用户的兴趣都是通过分析点击数据以及浏览网页内容所获得。Kim(参考:Marco Pennacchiotti,Fabrizio Silvestri,Hossein Vahabi,Rossano Venturini:Making Your Interests

Follow You on Twitter.CIKM,Maui,HI,USA October 29-November 2,2012.)提出一种分层聚类算法,然后从一系列与用户相关的网页中学习用户的兴趣层次结构。这种方法适用于网页浏览的用户,不适用于社交网络用户。

[0046] 3、基于用户社交关系的兴趣挖掘识别研究

[0047] 基于用户社交关系的兴趣识别研究,主要适用于那些将网络作为浏览内容或者获取订阅内容的平台的用户,例如很多微博用户只是浏览别人的分享内容,自己并不转发或者原创微博。Macro等人(参考:Kim,H.R,Chan,P.K.:Learning implicit user interest hierarchy for context in personalization.In:Proceedings of the 8th International Conference on Intelligent User Interfaces,IUI 2003,pp.101-108.ACM,New York(2003).)提出了通过融合用户及用户朋友的所有微博来挖掘用户感兴趣的微博类型,然后向用户推荐相关兴趣微博。Globerg等人(参考:Glodberg,D,Nichols,D,Oki,B.M.,Terry,D.:Using collaborative filtering to weave an information tapestry.Communications of the ACM-Special Issue on Information Filtering 35(12),61-70(1992))提出了从跟用户在某个问题上有相似观点的其他用户身上来挖掘兴趣,其想法基于有相似行为的两个用户也有着同样的兴趣。另外,Wen(参考:Wen,Z,Lin,C.-Y:On the quality of inferring interests from social neighbors.In:Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,KDD 2010,pp.373-382.ACM,New York(2010))提出通过挖掘相邻用户的社交关系来挖掘用户的兴趣。对于一个用户而言,考虑其关注人或者粉丝及关注人所关注的人,由此通过构建一张该目标用户为中心的用户关系网络来挖掘识别用户的兴趣。

[0048] 为便于本领域技术人员理解本发明的技术内容,下面结合附图对本发明内容进一步阐释。

[0049] 如图1所示,本发明的技术方案实现过程包括以下步骤:

[0050] S1、构建Word2vec模型;

[0051] 本发明主要是针对社交网络媒体的用户文本内容中的兴趣识别,那么,针对各种兴趣,先要训练出一个足够大(这里的足够大应理解为:尽可能采用覆盖范围比较全的数据集来训练的模型,较大的兴趣数据集是几十万条数据,较小的是几万条)的Word2vec模型,将足够多的兴趣相关用户的所有推文在正则化处理后,训练成Word2vec模型。

[0052] 正则化处理过程为:

[0053] A1、去掉...与…结尾的单词;

[0054] A2、去掉所有格与助词缩写;

[0055] A3、去掉文本中的特有无意义词汇或内容;

[0056] A4、把用.连接的词汇中的.去除;

[0057] A5、只保留文中的基本和扩展拉丁字母,数字以及基本的ASCII码;

[0058] A6、将首字母大写连词分开;

[0059] A7、对词进行词干化以及去除停等词。

[0060] 将足够多的用户文本训练成Word2vec模型后,用于后面的训练文本的提取以及特征向量的获取。而实际上Word2vec的作用是将所有的词语都投影到K维的向量空间,每个词语都可以用一个K维向量表示,K的取值取决于实验的效果例如本实施例中取值为100。

[0061] Word2vec模型其实是简单化的神经网络，输入是One-Hot Vector，Hidden Layer没有激活函数，也就是线性的单元。Output Layer维度跟Input Layer的维度一样，用的是Softmax回归。要获取的dense vector其实就是Hidden Layer的输出单元。其过程如图2所示。

[0062] Word2vec根据上下文之间的出现关系去训练词向量，主要分为CBOW(Continuous Bag of Words)和Skip-Gram两种模式。CBOW是从原始语句推测目标字词；而Skip-Gram正好相反，是从目标字词推测出原始语句。CBOW对小型数据库比较合适，而Skip-Gram在大型语料中表现更好。模式示意图如图3所示，其中，input为输入层，projection为投影层，output为输出层。

[0063] S2、文本打分获取训练集方法；

[0064] 从海量的数据集中提取足以用于训练兴趣的SVM模型的少量具有代表性的内容文本，不同于常用的人工筛选，从数据源到选取都采用程序自动化的方法。

[0065] 首先是采用从足够多的兴趣相关用户中爬取海量的推文，并将其正则化后训练成Word2vec模型，随后将衡量这些兴趣相关的用户文本内容与兴趣关键词的相关度，将相关度排序。取TopN的文本内容作为训练集，N的取值取决于所采用的分类器以及机器性能，跟数据集大小有一定关系，本实施例取值为5000。而衡量相关度的打分方法具体如下：

[0066] 通过训练好的Word2vec模型获取与兴趣关键词相似度最高的几千个单词。通过简单神经网络训练得到的Word2vec模型获取的与兴趣关键词相似度最高的这些单词，是在文本内容以及文本语义以及文本位置上与兴趣关键词最相似的一些词，并且有一个相似度值作为衡量标准。随后，将要筛选的所有文本内容，其中的语句词语与这些词通过Wrod2vec模型来计算相似度值，取语句中每个词语与这些词计算的相似度值的最大值，累加作为语句的分值。最后排序获取TopN的文本内容作为正例训练集，而分值最低的N条文本内容作为负例训练集。

[0067] 这样获取的训练集可以很明显的发现其内容与兴趣关键词非常相关，并且在修改其中的关键词的相关词汇的数量，可以对获取到的与兴趣相关的训练集的文本包含的内容的范围有所控制。举例来说，获取到的体育兴趣的训练集中的包含的相关内容以及词汇，扩展更多的相关词汇数量可以使得最后获得的训练集尽可能得包含足够多的体育运动词汇以及赛事词汇，包括各种足球篮球乒乓球等运动，各种体育名人，各大赛事等等。

[0068] S3、运用卡方词频统计获取特征向量方法；

[0069] 获取到海量的原始文本以及足够相关的训练的文本后，如何将其转化为特征向量是方法流程中的重点。针对文本内容的兴趣识别挖掘，前面也是通过词汇的相似度计算等来对文本打分获取训练集。那么接下来的卡方词频统计也是基于此前工作的基础之上的。

[0070] 将正负例的训练集文本内容中的词汇进行词频统计，随后再用卡方检验的原理计算卡方值。选取卡方值能代表与兴趣相关与不相关的词汇作为特征词汇，再通过Word2vec模型获取词汇的向量，通过权重分配累加作为文本的特征向量，再用于SVM模型的训练。

[0071] 卡方检验是一种用途很广的计数资料的假设检验方法。属于非参数检验，主要是比较两个以及两个以上样本率(构成比)以及两个分类变量的关联性分析。根本思想在于比较理论频数和实际频数的吻合程度或者拟合优度问题。

[0072] 所用的为四格卡方检验，统计表格样式如表1所示：

[0073] 表1四格卡方检验统计表格

[0074]

某个词	含有该词的文本数量	不含有该词的文本数量	合计
兴趣相关训练集			
兴趣不相关训练集			

[0075] 卡方值的计算公式为：

[0076] $$X^2 = \sum \frac{(A-T)^2}{T}$$

[0077] 其中,A为实际值,T为理论值。

[0078] X^2用于衡量实际值与理论值的差异程度(也就是卡方检验的核心思想),包含了以下两个信息：

[0079] 1.实际值与理论值偏差的绝对大小(由于平方的存在,差异是被放大的)

[0080] 2.差异程度与理论值的相对大小

[0081] 在通过卡方值以及查表来判定词汇是否与兴趣相关的假设是否可靠。这里需要用到一个自由度的概念,自由度等于V＝(行数-1)*(列数-1),对四格表,自由度V＝1。例如对V＝1,词汇与兴趣95％概率相关的卡方分布的临界概率是:3.84。即如果卡方大于3.84,则认为词汇和兴趣有95％的概率相关。在判断文本与兴趣是否相关,在训练集中只要挑选与兴趣相关或不相关的词汇即可,那么计算卡方值并进行比较与临界值即可。而在预测中,使用卡方词频统计的检验时,要用的就只能是最具有代表可能相关词,那么就应该挑选卡方值最大的词而不用比较与临界值的大小。

[0082] S4、采用步骤S3得到的代表文本的特征向量对SVM模型进行训练；

[0083] SVM(Support Vector Machine,支持向量机)是一种线性分类器,于1995年由Cortes和Vapnik提出,目前已经应用在手写体识别以及文本分类等领域。

[0084] SVM建立在统计学习理论中的VC维理论和结构风险最小化的基础之上,在模型的复杂度和学习能力之间寻求最佳折衷,以期获得最好的推广能力,所得到的分类器一般是全局最优的。

[0085] 支持向量的目标就是找到一个分割面能够将两类给区分开来,同时和两类中离分割面最近的那些样本点保持最大的距离。运用SVM就是用已经分好的正负文本转化的向量来训练SVM模型,再用于预测其他文本属于与兴趣相关或者与兴趣不相关,从而识别出文本内容中的人物兴趣。

[0086] 用训练集中的文本的卡方词频分析得到的词作为兴趣的相关与不相关的两类的特征词,再通过Word2vec模型得到特征词的向量,随后根据词汇的词性进行权重分配后累加,得到代表文本的特征向量。输入SVM模型训练参数得到该兴趣的文本的SVM模型。随后,通过十折检验验证模型的训练效果,通过输入其他文本来预测文本中的兴趣得到模型对于人物兴趣识别挖掘的实际能力。

[0087] S5、采用步骤S4训练好的SVM模型进行文本内容的人物兴趣识别。

[0088] 本发明能够有效提高SVM模型在人物兴趣预测上的准确率,并对人物兴趣的大致兴趣预测有较好效果,具体数据如表2所示:

[0089] 表2预测准确率

[0090]

测试类别	十折检验正确率	测试集正确率
军事(military)	97.68%	87%
政治(political)	95.34%	81.5%
体育(sport)	88.25%	80.1%

[0091] 以文本内容预测以及人物文本兴趣相关预测比例来识别人物兴趣结果来看,本发明能够有效准确的识别人物兴趣。具体例子如下:

[0092] 1、"text":"It's undeniable:#Helicopters are awesome.And so is the Army#NationalGuard hangar life.Let's get you there:…https://t.co/bnl4lh8u4D"

[0093] 翻译:不可否认的是:直升机是令人敬畏的。陆军和国家警卫队的机库生活也是如此。

[0094] 让带你去……

[0095] 2、"text":"As a professional#NationalGuard Soldier,we'll train you to#protect and#defend with the M249Squad Automatic Wea…https://t.co/6IFzMrfCuP"

[0096] 翻译:作为一名专业的"国民警卫队"士兵,将训练你用M249小队自动武器保护和防

[0097] 御……

[0098] 3、"RT@SimonBanksHB:Because being tricky with the electoral system in order to call a long election campaign has a history of working out jus…"

[0099] 翻译:因为要发起一场长期的选举活动,对选举制度很棘手,所以有制定强制措施的

[0100] 历史……

[0101] 4、"RT@Bowenchris:BREAKING:Malcolm Turnbull's Liberal Government is voting RIGHT NOW against bigger,better and fairer personal income tax c…"

[0102] 翻译:突破:马尔科姆·特恩布尔的自由政府现在投票反对更大、更好、更公平的个人

[0103] 所得税...

[0104] 5、"Seeing our favorite#Volleyball and#BeachVolleyball players off-court is always special!And we are obviously happ…https://t.co/KUhmG430Cf"

[0105] 翻译:看到最喜欢的排球和沙滩排球的场外选手总是特别的!很明显…

[0106] 6、"RT@CEVEuroVolley:We are pleased to announce organisers of upcoming age-group European Championships!\n\n2019#EuroVolleyU16W

[0107] 翻译:很高兴地宣布即将到来的年龄组欧洲锦标赛的组织者

[0108] 上面的1,2都是预测判为与军事相关的文本,而3,4为预测判为与政治相关的文本,5,6为预测中判定为与体育相关的文本。下面的7,8则为人物的文本中预测判为分别与军事无关的文本,9,10为预测判断中与政治无关的文本,11,12为预测判断中与体育无关的

文本。

[0109] 7、"text":"If you want something#new,you have to stop doing something old.Tired of the same routine?We can fix that.Send…https://t.co/jyR2ixgPyd"

[0110] 翻译:如果你想要一些新的东西,你就必须停止做一些旧的事情。厌倦了同样的程序吗?

[0111] 可以解决这个问题。发送…

[0112] 8、"text":"An unsupervised model telling a supervised model where to find training data https://t.co/mVCehK5q3m";有词:supervised,training

[0113] 翻译:一种无监督的模型,指导监督模型找到训练数据。

[0114] 9、"RT@billshortenmp:We are ranked 25th in the world for maths.If we finished 25th on the Olympic medal tally,there would be a national out…"

[0115] 翻译:在数学上排名世界第二十五。如果在奥运奖牌榜上名列第二十五,就会有全国

[0116] 性的淘汰赛……

[0117] 10、"RT@RichardMarlesMP:With@ShayneNeumannMP,Group Captain Robert Denney and a Super Hornet at RAAF Base Amberley.@Aus_AirForce https://t.c…"

[0118] 翻译:与@ShayneNeumannMP一起,组长Robert Denney和皇家空军Amberley基地的

[0119] 超级大黄蜂。奥斯空军

[0120] 11、"RT@POINTS_EU:Thank you to all participants for the fruitful exchanges during the last@POINTS_EU meeting.\nWe are right on track to delive…"

[0121] 翻译:感谢所有与会者在上一次@POINTS_European会议期间进行的富有成果的交流。

[0122] 12、"Currently we have openings for two full-time and permanent positions in our office in#Luxembourg

[0123] 翻译:目前在卢森堡的办公室有两个全职和永久性职位的空缺!雇佣新人…

[0124] 除了单条用户文本与兴趣的相关判断上,本发明能基于用户文本内容对人物的兴趣进行有效的判别。对于名为NationalGuard的国防官方账号,对其军事的判断比例为60.5%;名为ShayneNeumannMP的议员用户的账号,对其政治的兴趣判断比例为63.5%;名为CEVolleyball的欧洲的一个排球官方用户账号的体育兴趣判断比例为67.4%;而六个教育领域的大学教授用户的账号对于这三种兴趣的判断比例都低于10%。

[0125] 本领域的普通技术人员将会意识到,这里所述的实施例是为了帮助读者理解本发明的原理,应被理解为本发明的保护范围并不局限于这样的特别陈述和实施例。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权利要求范围之内。

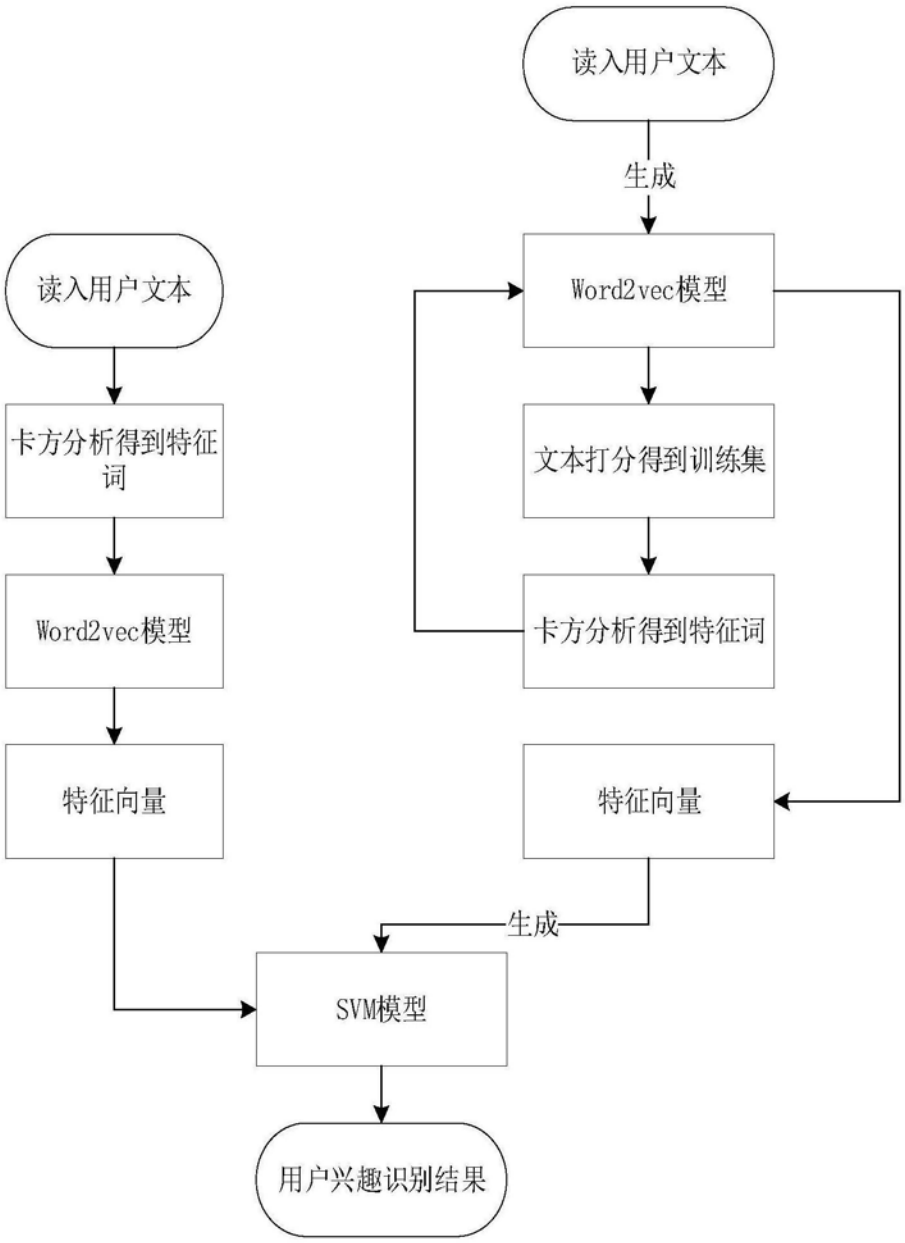

读入用户文本
生成
读入用户文本
Word2vec模型
卡方分析得到特征词
文本打分得到训练集
Word2vec模型
卡方分析得到特征词
特征向量
特征向量
生成
SVM模型
用户兴趣识别结果

图1

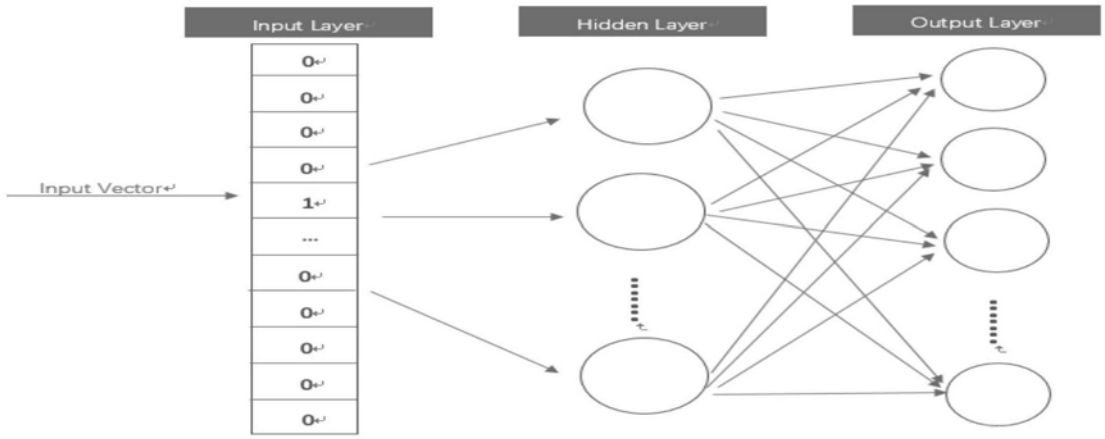

Input Layer
Hidden Layer
Output Layer
Input Vector
0
0
0
0
1
...
0
0
0
0
0

图2

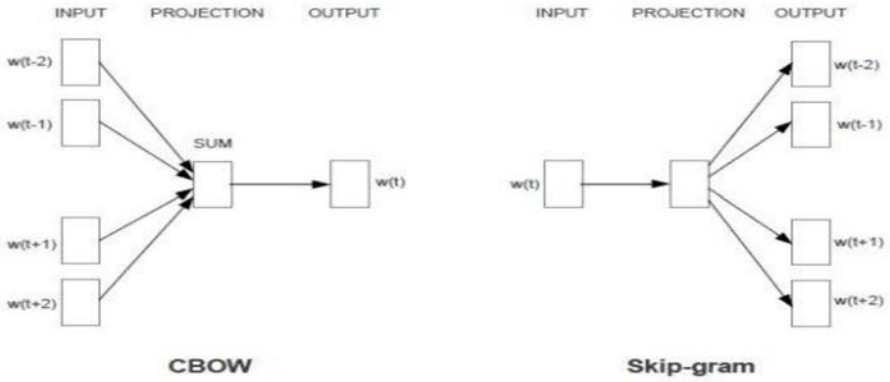

INPUT
PROJECTION
OUTPUT
w(t-2)
w(t-1)
SUM
w(t)
w(t+1)
w(t+2)
CBOW
INPUT
PROJECTION
OUTPUT
w(t)
w(t-2)
w(t-1)
w(t+1)
w(t+2)
Skip-gram

图3