面向游戏客服场景的自动问答系统研究与实现

王丽月,叶东毅
WANG Liyue, YE Dongyi

福州大学 数学与计算机科学学院,福州 350100
College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350100, China

WANG Liyue, YE Dongyi. Research and implementation of automatic question-answer system in game customer service scenarios. Computer Engineering and Applications, 2016, 52(17):152-159.

Abstract: In view of players' professional and colloquial way of querying in game customer service scenarios, this paper presents a sentence similarity model that takes into account synonymous substitutions, weights, sentence length, word order and other factors with semantic word vectors being established using the deep learning tool word2vec. Based on this model, the drawbacks of both dominance of majority classes and high computational cost associated with KNN classification algorithm are improved by pre-classification and re-defining classification rules. Furthermore, this paper implements an automatic question-answer system based on text classification for the game customer service scenarios. Experimental results show that this system has higher accuracy and efficiency of queries classification.

Key words: word2vec; sentence similarity; text classification; automatic question-answer; natural language processing

摘 要: 针对游戏客服场景中玩家领域化、口语化的提问方式,应用深度学习工具word2vec建立带有语义的词的向量表示,设计了一种利用词向量距离,结合同义词替换、权重、句子长度、词序等因素的句子相似度计算模型。在该模型基础上,通过预分类、重定义分类规则,对KNN分类算法的大类占优、全局匹配计算代价高等问题进行改进,实现了一种基于文本分类的面向游戏客服场景的自动问答系统。实验结果表明,该系统具有较高的问题分类准确率和分类效率。

关键词: word2vec; 句子相似度; 文本分类; 自动问答; 自然语言处理
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1512-0341

1 引言

随着网络数据急剧增长,信息检索技术已经成为互联网时代的研究热点。搜索引擎是一种常用的信息检索手段,它通过关键词匹配的方式,搜索相关文档,并对结果进行排序、返回。然而,传统搜索引擎无法深入理解问题的语义,且其返回结果为链接列表,无法直接获取答案。伴随着人工智能的兴起,研究学者们纷纷投入自动问答系统的研究,期待利用自然语言处理技术克服搜索引擎的局限性。

现有的自动问答系统实现方式主要有两种:一是,基于常见问题集的自动问答系统[1]。常见问题系统中的问句一般为标准表达方法,问题数量有限,难以模拟和覆盖用户真实的提问方式。二是,基于文档库的答案抽取方法[2]。该方法需要构建大量答案模式,代价高。目前,学术界采取答案抽取方法的研究大多是局限在答案为人名、机构名、时间、地点等实体对象的问题,可结合命名实体等技术抽取答案[3-4]。且在具体应用场景中,基于模式构建的答案未必适合直接作为官方的回复。因此,本文结合游戏客服场景,提出了一种通过计算新问题与历史问题的相似度,根据历史问题的节点对新问题进行分类,给出节点答案的自动问答系统解决方案。

2 综合多因素的句子相似度模型

常见的中文句子相似度计算方法主要有基于统计

基金项目: 国家自然科学基金(No.61473089)。
作者简介: 王丽月(1989—),女,硕士,研究领域为自然语言处理,E-mail:wangly2009@163.com;叶东毅(1964—),男,博士,教授,研究领域为计算智能和数据挖掘。
收稿日期: 2015-12-28 **修回日期:** 2016-04-21 **文章编号:** 1002-8331(2016)17-0152-08
CNKI **网络优先出版:** 2016-05-27, http://www.cnki.net/kcms/detail/11.2127.TP.20160527.1527.032.html

信息、基于语义信息以及基于句型、句法信息等的计算模型[5-6]。基于统计信息的方法以TF-IDF算法及其改进算法为主,主要实现方法是将词频、共现关系等统计信息融入向量空间模型,利用向量距离表示句子相似度[7]。该方法实现简单、计算效率高,但缺乏语义、句法信息,无法深入理解句子含义[8]。基于语义信息的方法,利用事先建立的语义知识网络,根据词在语义树的距离,度量词汇的语义相似度,并结合其他信息,构成句子相似度算法,主要有基于《知网》或领域本体知识库的词语相似度计算方法[9-10]。前者受限于知识源,不适用于特定领域。而后者需构建领域本体知识库,对人力成本和语言学知识要求高,且不易自动扩展到其他领域。基于依存树的句子相似度计算方法,对词语在句子中的成分进行分析,按依存关系得到有效配对,进而计算句子相似度[11]。该方法引入语法特征,加强了句子相似度模型的句法相似性[12],但无法表达语义层面的相似性,且该方法依赖于句法分析的准确率,而现有的依存句法分析准确率不高,需要人工加以修正[13]。因此,本文提出一种基于深度学习工具word2vec提供的词向量,结合词频统计信息构建的权重以及词序、句子长度等因素的句子相似度计算模型。

2.1 词向量

文本表示是自然语言处理的基础性技术,本文应用Google开源的深度学习工具word2vec[14],对基于隐马尔可夫分词器[15]分词的结果进行训练,得到固定维数的词向量。词向量各维度上的数值无显示含义,但向量之间的差异代表了词语的语义间隔[16]。一方面,借助word2vec工具可实现词语的知识表示,另一方面,利用词向量的余弦距离可表示词的语义相似度。

基于词向量的文档(句子)表示有多种组织方式,可直接叠加词向量,合成文档(句子)向量。下文将其简称为W2V文本表示法,仅用以验证词向量的有效性。由于该方法将文档简化为词语的堆砌,未考虑句子的中心词、语法等信息。因此,本文将句子视为由词向量构成的矩阵,如式(1),以便结合词权重、句法等信息。

$$\boldsymbol{S}=\begin{pmatrix}V_1\\ \vdots\\ V_m\end{pmatrix}=\begin{pmatrix}v_{11} & \cdots & v_{1n}\\ \vdots & v_{ij} & \vdots\\ v_{m1} & \cdots & v_{mn}\end{pmatrix} \tag{1}$$

2.2 构造同义词集

本文研究的游戏客服场景中玩家的提问方式随意,存在大量的错别字、词,例如“被盗”、“充值”常被误打成“被到”、“冲值”等。另外,游戏中如“魔石”、“宝宝”等专有词汇存在“MS”、“BB”等别称。在该问题领域中,无论是错别字还是别称,只要是表达的语义上等价、可替代的词,都将其视为同义词。由于相同语义存在多种表达方式,直接利用词向量,进行句子相似度计算,将无法得到准确的相似度信息。例如,“魔石被盗”和“MS被到”,它们表达的语义相同,而在词相似度层面,“魔石”与“MS”的相似度为0.58,“被盗”与“被到”相似度为0.57,直接用于计算句子相似度,无法得到句子客观上等价的结果。因此,本文通过在计算句子相似度之前,进行同义词替换,以有效地解决此类问题。由此,引入了新的问题,如何获得领域同义词?

本文基于word2vec训练得到的词向量库,结合人工标记,实现了一个半自动化同义词查找工具。其基本原理如图1,对输入词 A,根据相似度阈值,查找出候选同义词,进行人工筛选,当用户确定候选集中的 B 词确实为 A 的同义词,则将 B 加入同义词集,并且开始级联查找 B 词的候选同义词,以此类推,在词向量库中进行横向和纵向的搜索,直至所有候选词均通过筛选,或满足用户需求为止。

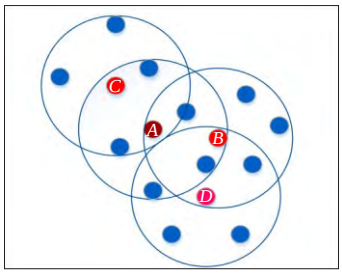

图1 同义词搜索示意图

利用该工具,能够有效地发现领域同义词,例如:

账号:游戏号/帐号/账户/张号/帐好/号

充值:冲值/直冲/直充/充直/冲直

2.3 词权重算法

由于句子中的词语存在实词、虚词之分,实词是具有实际意义的词,而虚词通常在句子中起语法作用,没有具体含义。另外,不同词语对句子主题的贡献度也有所差异。而采用直接叠加词相似度,得到句子相似度的方式,将丢失词语在语料范围中的其他有效信息。因此,本文在去停用词基础上,利用词频统计信息,构建词汇权重表,作为词相似度的调节因子。

经典的TF-IDF算法,考虑了词条在多文档的分布情况,却没有考虑词条在类别之间的分布情况。当词条分布相对集中在某个或某些类别,认为它具备较高的区分度,应当相应增加其权重。本文借鉴徐凤亚[17]等人的权重改进方法,引入词频分布标准差,当词条在类间分布越集中,标准差项越大,词权重越大。改进后的词权重计算公式如下:

$$w(t)=tf(t)\times idf(t)\times De(t) \tag{2}$$

$$idf(t)=\ln\left(\frac{N}{n}\right) \tag{3}$$

$$D_e(t)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(tf_i(t)-\overline{tf(t)})^2} \tag{4}$$

其中,$w(t)$ 表示词条 t 的权重,$tf(t)$ 为词条 t 在所有语料中出现的总词频,$idf(t)$ 表示词条的逆向文件频率,

$D_e(t)$ 表示词条 t 的词频在所有类别的分布标准差。$tf_i(t)$ 为词条 t 在第 i 类中的出现词频。

经以上公式计算得到的词权重,按权值高低排序,结果列表与词语的类别区分度基本相符,表明该方法能够较客观地评价词语的重要性。然而,由公式(2)计算得到的权值跨度大,直接用于计算句子相似度,将因权重影响过大,而降低词相似度的作用,效果反而不好。因此,本文采用对数标准化方法,将权值范围映射到[0,1]区间。标准化公式如下:

$$w(t)=\lg(w(t)) \tag{5}$$

$$w(t)=\frac{w(t)-\min}{\max-\min} \tag{6}$$

在实际应用过程中,权重也可根据经验,人为加以调整,以突出类别特征明显的词。

2.4 新句子相似度模型

杨思春[18]、吕学强[7]等人研究的句子相似度是基于相同词数的计算方法,句子 A 对 B 的相同词数与 B 对 A 的相同词数一致,因此,A 与 B 的相似度是对称的。而本文提出的基于词向量的相似度计算方法,句子 A 对 B 的相似度并不等价于 B 对 A 的相似度,孔胜[19]、李茹[20]等人的研究采用求二者均值的方式,得到 A、B 的最终相似度。结合本文的词权重,定义相关概念如下:

定义 1 将句子分词、去停用词,并进行同义词替换,将句子转为词汇向量的表示方式:

$$\boldsymbol{S}=(t_1,t_2,\cdots,t_n)^{\mathrm{T}} \tag{7}$$

定义 2 计算词语相似度,优先通过字符串匹配方式,相同则词语相似度为1;否则,计算词向量的余弦距离。由于余弦距离的值范围为[−1,1],大量实验表明,当词向量间余弦距离小于0时,词语之间几乎无语义相关性,也不存在反面语义现象,因此,本文对词语相似度做非负处理,如下:

$$sim(t_1,t_2)=\max\{\cos(v_1,v_2),0\} \tag{8}$$

其中,v_1、v_2 分别为词条 t_1、t_2 的词向量。

定义 3 词条 t 对句子 S 的相似度为 t 与 S 中最相似词的相似度:

$$sim(t,S)=\max(sim(t,t_i)),i\in(1,n) \tag{9}$$

定义 4 句子 A 对 B 的相似度,定义为 A 中词条对句子 B 相似度的加权和:

$$SimA2B=\frac{\sum_{i=1}^{m}w_iS(t_i,B)}{\sum_{i=1}^{m}w_i} \tag{10}$$

由此,得到句子相似度公式:

$$Sim(A,B)=\frac{(SimA2B+SimB2A)}{2} \tag{11}$$

然而,经大量实验发现,当句子 A、B 的长度有明显差异时,公式(11)相似度计算结果将明显受短句影响。

例 1 句子A:“还在吗?”。

分词结果:“还在/吗”。

句子B:“我在黑网吧登了一下游戏,可能是中了木马了,我的3洞攻击装备还有90星的宝宝,不知道还在吗?”。

分词结果:“在/黑/网吧/登录/游戏/中/木马/ 3/洞/攻击/装备/还有/90/星/宝宝/不知道/还在/吗”。

句子 A 中仅包含“还在”、“吗”两个词,句子 B 除“还在”、“吗”以外,还包含其他16个词语。计算句子相对相似度,$SimA2B=1.0$,$SimB2A=0.11$,得句子 A、B 的最终相似度为0.56。然而,事实上,长句 B 比短句 A 包含更多更重要的语义信息,尤其是当短句中的词语语义信息不强时,以上公式不能准确地表达二者的相似性。因此,本文对以上公式加以改造,引入句子长度调整因子,修正句子相似度公式。

说明 由于本文计算相似度的基本单位为词,而非字,因此,引入句子长短影响因子时,考虑的是词数而非字数。

引入句子长度因子的相似度计算公式如下:

$$Sim(A,B)=\left(\frac{\sum_{i=1}^{m}w_ia_i}{\sum_{i=1}^{m}w_i}\times\alpha+\frac{\sum_{j=1}^{n}w_jb_j}{\sum_{j=1}^{n}w_j}\times\beta\right) \tag{12}$$

$$a_i=\max(sim(t_i,t_j)),j\in(1,n) \tag{13}$$

$$b_j=\max(sim(t_j,t_i)),i\in(1,m) \tag{14}$$

$$\alpha=\frac{lenA}{lenA+lenB} \tag{15}$$

$$\beta=\frac{lenB}{lenA+lenB} \tag{16}$$

其中,m、n 分别表示句子 A、B 所包含的有效词语数量,w_i、w_j 表示相应词语的权重,a_i 表示句子A的第 i 个词与句子 B 最相似词的相似度值。b_j 同理。$lenA$、$lenB$ 分别表示句子 A、B 的词数,α、β 分别表示句子 A、B 各自的词数占总词数的比例,它们作为权重,调节句子间相似度。回到例1,使用改造后的句子相似度公式,$Sim(A,B)=\left(1.0\times\frac{2}{2+18}+0.11\times\frac{18}{2+18}\right)=0.20$,表明加入句子长度影响因子能够更合理地表达句子相似度关系。

此外,由于相同词语作为句子的不同成分,词的语义和作用会有所差异。不少研究句子相似度的学者认为,通过词序的差异,能够在一定程度上反应句法相似信息[21]。因此,本文尝试引入词序信息,使句子相似度计算模型更具合理性。

定义 5 句子 A、B 分词后,分别存于数组 $arrA$、$arrB$ 中,对 $arrA$ 中的每个词,从 $arrB$ 中寻找最相似词,记录其下标,构成 A 对 B 的词序向量 ($orderVec$),同理可得,B 对 A 的词序向量。

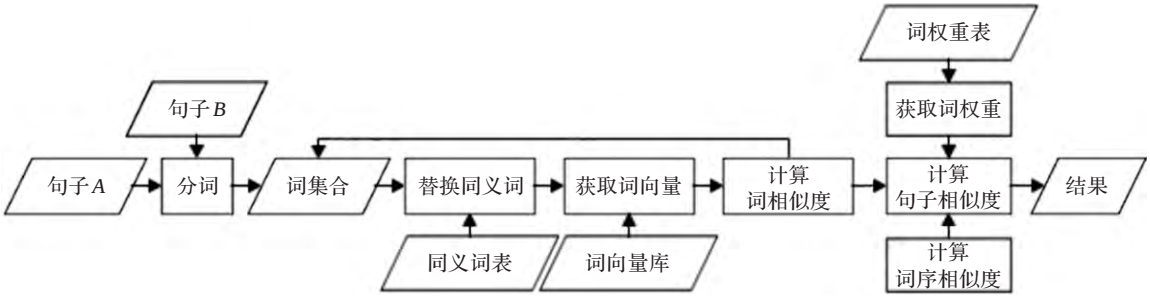

图2　句子相似度计算流程

通过计算词序向量中的逆序对,得到词序相似度,计算公式如下:

$$orderSim = 1 - \frac{rev}{\max Rev} \tag{17}$$

其中,rev 为逆序对数,$\max Rev$ 为最大逆序对数。

最后,本文综合多因素的句子相似度计算公式如下:

$$Sim(A,B) = \lambda_1(SimA2B \times \alpha + SimB2A \times \beta) + \lambda_2 orderSim \tag{18}$$

其中,$\lambda_1 + \lambda_2 = 1$,$\lambda_1$、$\lambda_2$ 分别为语义相似度和词序相似度的占比。

本文计算句子相似度的基本流程如图2。

3　改进KNN的文本分类方法

本文将自动问答系统简化为分类问题,省去了构造问题答案的难题,却提高了对文本分类的要求。由于要将问题定位到有唯一答案的节点上,势必要求问题节点足够细,而节点越细,类别越多,分类的难度越大。而基于特征抽取的文本分类方法,不可避免地损失了原始数据中的细节信息[22],分类精度难以达到要求。因此,本文采用基于实例的文本分类方法。

KNN(K-Nearest Neighbor)算法是经典的基于实例的分类算法,但它存在以下缺陷:(1)基于全局的实例相似度计算,当知识库规模庞大,KNN算法的计算代价不容忽视[23];(2)当各个类别的语料数量不均时,KNN算法偏向于语料数量多的类别[24]。

本文针对这些问题,提出了相应的解决方案。首先,对全局匹配导致的效率问题,采用预分类,限定候选集,缩小匹配范围的方式来解决。由于KNN算法以最近邻的 k 个实例的类别标注决定分类结果,算法效果依赖于参数 k 的恰当取值。当 k 的取值过小,结果易受噪声数据干扰,如图3;当 k 的取值过大,样本数多的大类将明显占优势,如图4。

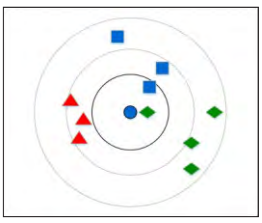

图3　k 的不同取值

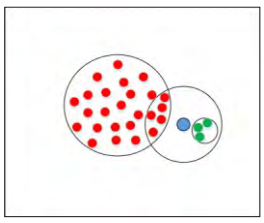

图4　大类占优

因此,本文提出通过对每个邻近类取相同数量的最近实例,以其中平均距离最小的类别作为分类结果,以克服大类占优问题,如图5。从相邻类别中,各取top3近邻的样本,以平均距离最小的 C 类作为分类结果,能够克服 A 类在 $K(K=10)$ 近邻中占优的问题,并且,该方法能有效避免噪声数据的干扰,如 B。

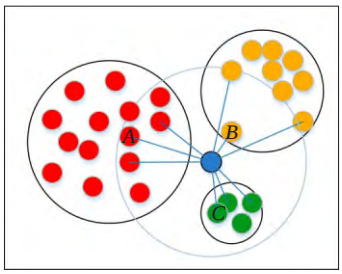

图5　对邻近类别取均等的最近实例

本文利用基于特征抽取的文本分类方法,预选出若干候选类别,并从这些类别中随机抽取一定数量的语料构成候选集,进行下一步基于实例匹配的文本分类。

3.1　特征词抽取算法

吕佳等对特征提取算法进行了改进,它的基本思想是认为特征词在类别间分布越集中,在类内分布越均匀,越具有类别代表性[25]。而方差作为样本离散情况的衡量指标,能够体现样本在类别间的分布情况,因此,引入方差,以克服TFIDF算法不能体现类内、类别间分布情况的缺陷。据此,提出的词权重计算公式如下:

$$w_i(t) = tf_i(t) \times idf(t) \times D_e \times (1 - D_{ii}) \tag{19}$$

$$D_e(t) = \frac{1}{n}\sum_{i=1}^{n}(tf_i(t) - \overline{tf(t)})^2 \tag{20}$$

$$D'_{ii} = \frac{\frac{1}{m}\sum_{j=1}^{m}(tf_{ij}(t) - \overline{tf_i(t)})^2}{\frac{1}{m}\sum_{j=1}^{m}(tf_{ij}(t))^2} \tag{21}$$

其中,$w_i(t)$ 表示词条 t 在第 i 类的权重,$tf_i(t)$ 为词条 t 在第 i 类的出现频率,$idf(t)$ 表示词条 t 的逆向文档频率,$D_e(t)$ 表示词条 t 的词频在所有类别的分布方差。当 $D_e(t)$ 越大,说明词条在类别间的分布越集中,越具有类别代表性。D'_{ii} 表示词条 t 在第 i 类内的平均方差,其中 m 为类内样本数。当类内平均方差 D'_{ii} 越小,说明

词条在某个类别分布越均匀,越能代表该类的特征。

由于自动问答应用场景中分析的文本以短文本为主,一般情况下,在一条记录(一个提问句)中,关键词不会频繁地重复出现,即不会因 $tf_i(t)$ 可能主要来源于某条记录的贡献,而需要引入 $(1-D'_{ii})$ 调整项。因此,本文在文献[25]基础上,针对短文本的特点,对其进行简化。改进后类特征词权重计算公式如下:

$$w(t_i)=tf_i(t)\times idf(t)\times D_e(t)'\times df_i(t) \tag{22}$$

$$D'_e(t)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(tf_i(t)-\overline{tf(t)})^2} \tag{23}$$

$$df_i(t)=\lg(\frac{wsf_i}{csf_i}+1) \tag{24}$$

文献[25]引入的 $D_e(t)$ 是词频分布方差,实验表明,该项影响偏大,本文将其改为词频分布标准差 $D'_e(t)$。$df_i(t)$ 表示词条 t 在第 i 类的文档频率,wsf_i 表示第 i 类中包含词条 t 的样本频数,csf_i 表示第 i 类的样本总数。基于与IDF相反的思想,认为词语在类内分布在越多条记录中,即在类内分布越分散,说明其越能代表该类的特征,权重越高。由于 $wsf_i/csf_i\in[0,1]$,对其取对数,结果非正值,此处通过加1,平移对数曲线,加以修正。

基于公式(23),从各个类别中抽取出权重最高的前 m 个词,作为类别特征词。

3.2 文本预分类

在抽取出类别特征词的基础上,对新问题进行预分类,即计算新问题与各个类别的特征词的相似度,取最相似的前 n 个类别的语料作为基于句子相似度分类的候选集。候选类别的数量需兼顾效率和准确率,结合具体语料分布情况而定。问题与类别的相似度计算公式如下:

$$Sim(Q,C_k)=\frac{\sum_{i=1}^{m}w_i a_i}{\sum_{i=1}^{m}w_i} \tag{25}$$

$$a_i=\max(sim(t_i,t_j)),j\in(1,n) \tag{26}$$

其中,m 为新问题 Q 的关键词数,n 为类别 C_k 的特征词个数,t_i、t_j 分别表示新问题 Q 中的关键词和类别 C_k 的特征词。w_i 为关键词权重,由2.3节中词权重算法构建的权重表提供。

通过预分类,排除了大量几乎不可能的类别,为实例级别的相似度计算缩小了范围,大大提高计算效率。然而,该方法仍无法解决由于某一类别,语料数量巨大而带来的计算性能问题。据观察发现,同一问题节点的语料会有多种形式,但并非无限可能。因此,本文借鉴苏金树等人提出的随机采样的思想[26],从各个候选类别中,抽取出一定数量(MAX_LIMIT)的语料作为训练样本,待后续的句子相似度计算。若该类语料不足上限,则取全部。

经过上述调整后,基本能够保证分类模型的计算规模在可控范围内,不会随着语料的扩充而无限增长。

3.3 基于句子相似度的文本分类

在预分类基础上,计算新问题与候选类别语料的句子相似度。新问题与类别的相似度定义为类别语料中与新问题最相似的Top$M$个问题的相似度均值。算法如下:

```
input:newQues
output:similarity
quesList=getCorpus(Ck)
for (String q:quesList)
    ques_sim=comSenSim(newQues,q)
    if (ts.size()<topM)
      ts.add(q);
    else if (ques_sim>ts.last().getSim())
      ts.pollLast();
      ts.add(q);
size=ts.size();
while(!ts.isEmpty())
  similarity+=ts.pollFirst().getSim();
  similarity=similarity/size;
```

以上分类方法的实现,通过维护一个大小为 n 的有序集合(TreeSet),避免了对整个类别的所有实例进行相似度排序,提高的分类效率。类似的,也可采用优先队列(PriorityQueue)优化算法。

综上,本文提出的文本分类策略如下:

(1)基于特征抽取算法抽取出的类别特征词,对问题进行预分类,筛选出若干候选类别。

(2)从候选类别中,随机抽取出一定数量的语料,与新问题计算句子相似度,并以每个类别最相似的 n 个问题的平均相似度,作为新问题与该类别的相似度。

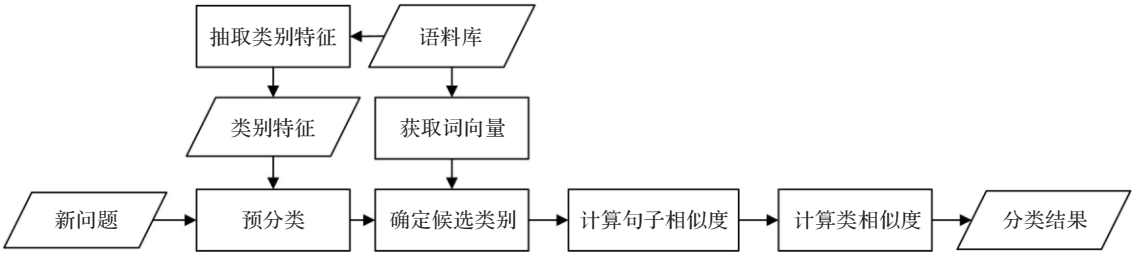

图6 文本分类的基本流程图

(3)对候选类别按相似度进行排序,取相似度最大者作为分类结果。

(4)结合应用场景需要,可设置相似度阈值。当分类相似度低于该值时,可放弃本次分类,以降低误分类的代价。

4 实验

4.1 数据说明

数据集1:搜狐的新闻语料(http://www.sogou.com/labs/dl/ca.html),包含9个类别,共17 906篇文章,各个类别语料分布均匀。

数据集2:复旦中文语料(http://www.datatang.com/data/44139/),包含训练集9 804篇,测试集9 833篇,语料在20个类别分布不均。

数据集3:游戏客服语料,约12 000条数据,包含8个问题节点,各个节点的语料数量不均等。该语料存在较多复合节点问题,更贴近系统使用过程中可能面临的实际情况。

数据集4:游戏客服语料,包含12个问题节点,共5 002条数据,语料在各个节点分布不均。该语料经严格筛选的,基本排除了错误分类等问题,语料质量较高。

4.2 评估指标

分类器性能评估指标主要包括准确率(accuracy),指正确分类数与样本总数的比值。分类器对类别的评估指标包括精确率(precision,P)、召回率(recall,R)和综合测度F_1值(F_1-measure),它们的定义见表1及如下[27]:

$$P=\frac{a}{a+b} \tag{27}$$

$$R=\frac{a}{a+c} \tag{28}$$

$$F_1=\frac{2PR}{P+R}=\frac{2a}{2a+b+c} \tag{29}$$

表1 混淆矩阵

	相关	不相关
被检索到	a	b
未被检索到	c	d

这些指标根据关注点不同,分为微平均和宏平均指标。例如,微平均精确率表示各个类别精确率按样本比例的加权均值,而宏平均精确率指各个类别精确率的算术平均数。微平均指标侧重表现大类的分类效果,而宏平均指标关注所有类别的分类效果,易受小类影响。

4.3 实验结果及分析

本文较已有的研究工作,主要区别包括:(1)应用word2vec工具,实现词的文本表示;(2)结合应用场景特点,设计了一种综合多因素的句子相似度计算模型;(3)针对语料分布不均的场景,提出了一种改进的KNN分类方法。

本节将分别对这些新特性的有效性加以验证。

4.3.1 文本表示法

选用数据集1、2,通过与传统向量空间模型对比,验证本文引入词向量的W2V表示法的有效性。实验条件如训练集与测试集的比例、SVM参数等,与王进[28]等人的实验保持一致。实验结果如表2、3。

表2 数据集1实验结果 %

分类器	文本表示法	宏准确率	宏召回率	宏$F1$值
KNN	VSM	68.7	67.6	68.2
	W2V	86.1	85.6	85.6
SVM	VSM	75.0	64.6	66.0
	W2V	78.8	77.0	77.3

表3 数据集2实验结果 %

分类器	文本表示法	宏准确率	宏召回率	宏$F1$值
KNN	VSM	85.4	83.7	84.5
	W2V	87.9	88.3	87.7
SVM	VSM	84.4	83.5	83.9
	W2V	88.4	89.7	88.8

实验结果表明,基于词向量的W2V文本表示法,结合KNN和SVM分类器,较传统向量空间模型,在各项分类指标中均有提升,说明基于word2vec训练得到词向量是一种有效的文本表示法,为后续的文本分析提供了更多语义方面的信息。

4.3.2 句子相似度模型

选用数据集3,对比引入同义词替换、权重前后的句子相似度模型,在传统KNN(K取值为20)文本分类器上的效果。

表4 在不同句子相似度模型下的分类效果 %

句子相似度	宏精确率	宏召回率	宏$F1$值	准确率
词叠加	77.4	71.5	72.7	74.8
有权重	77.9	72.0	72.9	75.4
有同义词	78.1	73.1	74.3	75.6
有权重有同义词	80.0	73.9	75.0	76.7

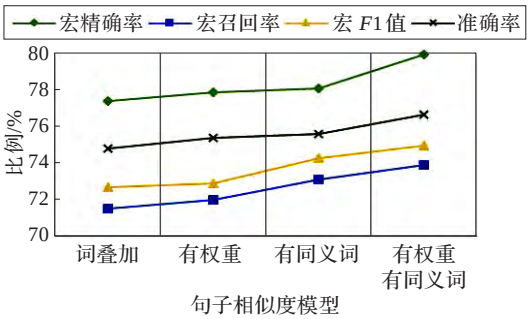

图7 不同句子相似度模型在KNN下的分类效果

实验结果表明,本文引入同义词替换及词权重后的句子相似度模型具有更好的分类效果,说明该句子相似度模型是有效的,较直接叠加词相似度的方法更具合理性。

4.3.3 特征抽取算法

选用数据集1,根据类别特征词进行文本分类,验证

本文采用的特征抽取算法的有效性。

步骤1 基于3.1节特征抽取算法,从语料中抽取出各个类别的关键词,每个类取前100个作为特征词。

步骤2 基于特征抽取的结果,计算测试样本与类别的相似度,记录取Top1和Top5相似类别的命中率。

表5 不同特征抽取算法的预分类效果 %

方法	Top1	Top5
TFIDF	75.2	95.7
本文方法	76.6	96.4

实验数据表明,本文的特征抽取算法较传统TFIDF具有更好的分类效果。并且,由于该方法考虑了应用领域的语料特征,下一节的实验数据将进一步说明基于该特征抽取算法的文本预分类方法能够保证分类准确性。

4.3.4 改进的KNN算法

由于预分类阶段选取的候选类数量Top*M*,对分类器效果和时间性能有直接影响。本节选用数据集4,从各个类别中抽取不超过400条语料,对比不同Top*M*对预分类准确率和平均分类时间的影响,以选定合适的Top*M*值。

由图8可见,当 *m* 值小于5,预分类准确率随着候选类数量的增加而提升显著。当 *m* 值达到5以后,预分类准确率提升有限,而分类时间随候选类数量增加成线性增长趋势。因此,综合预分类准确率和分类的时间代价,对数据集4选定候选类数量为5。

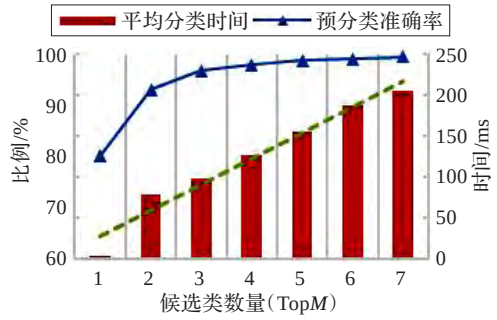

图8 不同Top*M*对预分类准确率及分类时间的影响

确定候选类数量后,对比取全部候选类语料和部分语料作为训练样本的实验效果,验证在牺牲一定程度的分类准确率情况下,能够较大地提升分类效率。同时,就相同实验条件,与KNN分类效果进行对比,验证本文重定义分类规则的方法在准确率方面的优势。

由表6可见,在语料数量相同的情况下,本文方法较传统的KNN方法及距离加权的KNN方法具有更高的分类准确率。由于本文方法经过预分类,因此,分类效率明显高于前两者。此外,基于本文方法在准确率方面的优势,使得可以通过缩减候选类语料,进一步提升分类效率。如表6中,第2、4、5行分类效果相当,而本文方法(第5行)的分类效率提升了5倍以上。

表6 KNN方法及本文方法在不同语料量的分类结果

文本分类法	MAX_LIMIT	准确率/%	分类时间/ms
KNN	400	81.8	163.423
	不限	89.9	896.235
距离加权KNN	400	85.4	397.888
	不限	91.9	927.507
本文方法	400	89.7	169.338
	不限	93.9	431.189

由表7可见,面向游戏客服场景,本文方法的分类性能优于前两者,尤其是在宏召回率指标上有明显提升,说明本文重新定义的文本分类策略,确实能够在一定程度上避免小类因样本数量的弱势而被误分到大类,从而提升了整体召回率。

表7 KNN方法及本文方法的分类效果

文本分类法	MAX_LIMIT	宏精确率/%	宏召回率/%	宏 *F*1值/%
KNN	不限	85.3	60.0	61.0
距离加权KNN	不限	85.8	65.6	70.0
本文方法	400	89.1	79.8	83.4

5 结束语

本文针对游戏客服领域,提出了一个结合词向量及同义词替换、权重、词序、句子长度等因素的句子相似度计算模型。在此基础上,通过预分类及重定义分类规则,对KNN分类算法加以改进,较好地克服了KNN分类算法存在的大类占优、计算代价高等问题。

由于游戏客服场景中玩家的表达方式随意,不尽符合语言规范,因此,本文的句子相似度模型尚未深入考虑句法信息。然而,从通用性角度考虑,句法分析对计算句子相似度具有重要意义。因此,本文下一步研究将结合依存句法,设计更加通用的句子相似度模型。

参考文献:

[1] 钟敏娟,万常选,刘爱红,等.基于词共现模型的常问问题集的自动问答系统研究[J].情报学报,2009,28(2):242-247.
[2] 郑实福,刘挺,秦兵,等.自动问答综述[J].中文信息学报,2002,16(6):46-52.
[3] 刘宁锋,史晓东.中文问答系统中答案抽取的研究[J].电脑知识与技术,2011,7(12):2865-2868.
[4] 吴友政,赵军,段湘煜,等.问答式检索技术及评测研究综述[J].中文信息学报,2005,19(3):1-13.
[5] 周法国,杨炳儒.句子相似度计算新方法及在问答系统中的应用[J].计算机工程与应用,2008,44(1):165-167.
[6] 吕学强,任飞亮,黄志丹,等.句子相似模型和最相似句子查找算法[J].东北大学学报:自然科学版,2003,24(6):531-534.
[7] 庞剑锋,卜东波,白硕.基于向量空间模型的文本自动分类系统的研究与实现[J].计算机应用研究,2001,18(9):23-26.
[8] 殷耀明,张东站.基于关系向量模型的句子相似度计算[J].计算机工程与应用,2014,50(2):198-203.

[9] 程传鹏,吴志刚.一种基于知网的句子相似度计算方法[J].计算机工程与科学,2012,34(2):172-175.
[10] 刘宏哲.一种基于本体的句子相似度计算方法[J].计算机科学,2013,40(1):251-256.
[11] 李彬,刘挺,秦兵,等.基于语义依存的汉语句子相似度计算[J].计算机应用研究,2003,20(12):15-17.
[12] 吴佐衍,王宇.基于HNC理论和依存句法的句子相似度计算[J].计算机工程与应用,2014,50(3):97-102.
[13] 贾宗福,王知非.中文句子相似度计算的研究[J].科技信息,2009(11):10-11.
[14] 周练.Word2vec的工作原理及应用探究[J].科技情报开发与经济,2015(2):145-148.
[15] 刘群,张华平,俞鸿魁,等.基于层叠隐马模型的汉语词法分析[J].计算机研究与发展,2004,41(8):1421-1429.
[16] Mikolov T,Chen K,Corrado G,et al.Efficient estimation of word representations in vector space[J].arXiv preprint arXiv,2013.
[17] 徐凤亚,罗振声.文本自动分类中特征权重算法的改进研究[J].计算机工程与应用,2005,41(1):181-184.
[18] 杨思春.一种改进的句子相似度计算模型[J].电子科技大学学报,2006,35(6):956-959.
[19] 孔胜,王宇.基于句子相似度的文本主题句提取算法研究[J].情报学报,2011,30(6).
[20] 李茹,王智强,李双红,等.基于框架语义分析的汉语句子相似度计算[J].计算机研究与发展,2013,50(8):1728-1736.
[21] 杨思春,陈家骏.中文自动问答中句子相似度计算研究[J].情报学报,2008,27(1):35-41.
[22] 刘海峰,刘守生,姚泽清,等.文本分类中基于训练样本空间分布的K近邻改进算法[J].情报学报,2013,32(1):80-85.
[23] 郝秀兰,陶晓鹏,徐和祥,等.kNN文本分类器类偏斜问题的一种处理对策[J].计算机研究与发展,2009(1):52-61.
[24] 李荣陆,胡运发.基于密度的kNN文本分类器训练样本裁剪方法[J].计算机研究与发展,2004,41(4):539-545.
[25] 吕佳.基于改进分类模型的文本分类系统实现[J].重庆师范大学学报:自然科学版,2009,26(2):79-83.
[26] 苏金树,张博锋,徐昕,等.基于机器学习的文本分类技术研究进展[J].软件学报,2006,17(9):1848-1859.
[27] 李航.统计学习方法[M].北京:清华大学出版社,2012.
[28] 王进,金理雄,孙开伟,等.基于演化超网络的中文文本分类方法[J].江苏大学学报:自然科学版,2013,34(2):196-201.

(上接83页)

[3] 覃雄派,王会举,杜小勇,等.大数据分析—RDBMS与MapReduce的竞争与共生[J].软件学报,2012,23(1):32-45.
[4] The Apache Software Foundation.Hadoop[EB/OL].(2015-12-18) [2016-1-12].http://hadoop.apache.org/.
[5] Dean J,Ghemawat S.MapReduce:simplified data processing on large clusters[J].Communications of the ACM,2008,51(1):107-113.
[6] 宛婉,周国祥.Hadoop平台的海量数据并行随机抽样[J].计算机工程与应用,2014,50(20):115-118.
[7] 丁玉成,诸葛晴凤,沙行勉.云计算环境下排序算法的性能分析[J].重庆大学学报,2014,37(4):58-64.
[8] White T.Hadoop权威指南[M].华东师范大学数据科学与工程学院,译.3版.北京:清华大学出版社,2015.
[9] 张海军,丁溪源,朱朝勇.一种改进的中文字符串排序方法[J].计算机工程与应用,2010,46(19):129-131.
[10] 金菁.基于MapReduce模型的排序算法优化研究[J].计算机科学,2014,41(12):155-159.
[11] 梁建武,周杨.一种异构环境下的Hadoop调度算法[J].中国科技论文,2012,7(7):495-497.
[12] 杨勇,王伟.一种基于MapReduce的并行FP-growth算法[J].重庆邮电大学学报:自然科学版,2013,25(5):651-657.

(上接132页)

[9] Morais R,Fernandes M A,Matos S G,et al.A ZigBee multi-powered wireless acquisition device for remote sensing applications in precision viticulture[J].Computers and Electronics in Agriculture,2008,62(2):94-106.
[10] 黄智水.无线传感器网络免测距定位技术研究[D].南昌:华东交通大学,2012.
[11] Yedavalli K,Krishnamachari B.Sequence-based location in wireless sensor networks[J].IEEE Transactions on Mobile Computing,2008,7(1):81-94.
[12] 伍春燕.基于ZigBee技术的井下人员定位算法研究[D].长沙:中南大学,2012.
[13] 罗臻,刘宏立,徐琨.无线传感器网络在协同攻击环境中的安全定位研究[J].传感器与微系统,2014(7):38-41.
[14] 刘雪兰,王宜怀,陆全华,等.无线传感器网络RSSI定位算法改进[J].计算机应用与软件,2013(11):87-89.
[15] 曾超,刘宏立,徐琨,等.室内三维空间定位系统的RSSI特性分析[J].计算机工程与应用,2014,50(11):70-74.
[16] Garg R,Varna A L,Wu M.An efficient gradient descent approach to secure localization in resource constrained wireless sensor networks[J].IEEE Transactions on Information Forensics and Security,2012,7(2):717-730.