DOI:10.3979/j.issn.1673-825X.2018.02.017

基于 Word2vector 的文本特征化表示方法

周顺先[1],蒋　励[1,2],林霜巧[2],龚德良[1],王鲁达[1]
(1.湘南学院 软件与通信工程学院,湖南 郴州 423000,2.中南大学 信息科学与工程学院,长沙 410075)

摘　要:针对基于词语统计的特征化表示无法有效提取文本的词义特征的问题,提出一种基于上下文关系的文本特征化表示方法。该方法利用 Word2vector 提取词义特征,获得词向量;再对词向量进行“最优适应度划分”的聚类,并根据聚类结果将词语替代表示为聚类质心;根据质心及其所代表的词语的词频,构成词向量聚类质心频率模型(semantic frequency-inverse document frequency,SF-IDF),用于特征化表示文本。在不依赖语义规则的情况下,分别以路透社文本集 Reuter-21578、维基百科(extensible markup language,XML)数据为文本数据集,采用神经网络语言模型(neural network language model ,NNLM)算法进行文本分类实验,并采用 F1-measure 标准进行样本分类的效果评估,词向量聚类质心频率模型 SF-IDF(semantic frequency-inverse document frequency,SF-IDF)向量与现有技术中词频-逆向文件频率(term frequency-inverse document frequency,TF-IDF)向量的分类效果对比,与 TF-IDF 模型进行对比实验;在 Reuter-21578 数据集上平均准确率由原有的 57.1%提高到 63.3%,在 Wikipedia XML 数据集上平均准确率由原有的 48.7%提高到 59.2%。SF-IDF 模型可适用于现行的基于特征向量的信息检索算法,且较 TF-IDF 模型有更高的文本相似性分析效率,可提升文本分类准确率。

关键词:Word2vector;上下文关系;特征化表示;文本分类
中图分类号:TP391.9　　**文献标志码**:A　　**文章编号**:1673-825X(2018)02-0272-08

Characteristic representation method of document based on Word2vector

ZHOU Shunxian[1],JIANG Li[1,2,*],LIN Shuangqiao[1],GONG Deliang[1],WANG Luda[1]
(1. School of Software and Communications Engineering, Xiangnan University, Chenzhou 423000, P. R. China;
2. School of Information Science and Engineering, Central South University, Changsha 410075, P. R. China)

Abstract: Document representations based on statistical term measure can not extract lexical semantics effectively. Therefore, this work proposed a document representation method based on context. Using Word2vector, the method is able to extract lexical semantics in the form of word vector. And it can carry out clustering on word vector with ‘optimized fitness value partition’,then make cluster centroids represent words in each word vector cluster. On the basis of cluster centroids representing and word frequency, to characterize document, the method constructed cluster centroids frequency model, semantic frequency-inverse document frequency(SF-IDF). Without semantic database, respectively by Reuter 21578 and Wikipedia extensible markup languag(XML) as text data sets, using neural network language model (NNLM) algorithm for text classification experiment, and the F1-measure standard to evaluate the effect of sample classification, SF-IDF vector with the existing technology of term frequency-inverse document frequency(TF-IDF) vector classification result contrast, comparative experiment with the TF-IDF model was carried out. The average accuracy on Reuter 21578 data sets increases from the original 57.1% to 57.1%; average Wikipedia XML data set improves from the original 48.7% to 48.7% accuracy. SF-IDF could apply to VSM-based algorithms for information retrieval. And it shall perform better in text similarity analyzing, leading to higher precision in text classification work.

Keywords: Word2vector; context; characteristic presentation; text classification

收稿日期:2016-09-09　**修订日期**:2017-04-13　**通讯作者**:蒋　励　623887366@qq.com
基金项目:湖南省教育厅科研项目(15C1288);国家自然科学基金(61379109,61402165);郴州市科技计划项目(cz2015036)
Foundation Items:The Scientific Research Fund of Hunan Education Department(15C1288);The National Nature Science Foundation of China (61402165); The Chenzhou City Science and Technology Plan Projects (cz2015036)

0 引　言

目前,针对信息检索任务中的文本,在无法直接识别其词语语义的情况下,多采用基于词语统计的样本特征化表示方法,例如词频-逆向文件频率模型[1-2](term frequency-inverse document frequency, TF-IDF)模型与词袋[3](bag of words, BOW)模型。现行的基于词语统计的特征化表示方法可在无语义规则支持的情况下实现文本特征化表示,但将忽略词语语义,无法有效地提取其词语语义特征。

Word2vector 最先由 Mikolov 在 2013 年提出,该方法能够简单、高效获取词义的向量化特征,引起业界的极大关注。当文本作为信息检索的样本时,针对在不同文本中的每个词语,Word2vector 可依据其上下文关系有效地提取其语义(即词语语义特征),并以词向量提供形式化表达[4]。因此,Word2vector 的词义特征提取无需语义规则库。由于 Word2vector 的词义特征提取机制的复杂性,使得不同文本中相同的词所对应的词向量并不相同。难以根据 Word2vector 的词向量形成文本的特征化表示,尤其难以形成向量空间模型(vector space model, VSM)[5]形式的样本特征化表示。

Mikolov 在相关论文[6]中说明了 Word2vector 的机制。部分技术开发人员已对 word2vec 中的数学原理进行了深入的剖析[7]。在此基础之上,西南大学唐明等[8]提出一种文档向量表示的方法,应用与中文文档的分类;华东师范大学计算机应用研究所杨河彬等[9]提出 CT-Word2Vec 神经网络语言模型,利用词汇的上下文信息将词转化成向量,在词向量的学习过程当中融入了用户的搜索点击行为。上述方法在执行过程中,前者是对词语语义明细的文本进行特征化表示,存在对词语语义规则的利用(如中文分词的划分粒度),后者也可用于支持文本进行特征化表示,但存在人为语义识别的因素(如用户搜索行为的干预)。

单纯依靠 word2vec 工具,可根据上下文关系提取词语的语义特征,并可不依赖语义规则,且针对词语语义不明的文本同样有效。而在近期文献中,未有此类基于 Word2vector 的文本特征表示方法被提出。

文本提出的特征化表示可采用 Word2vector 作为基于上下文的词义特征提取方法,并适用于现行基于向量空间模型(vector space model, VSM)的信息检索算法。该文本特征化表示方法,能够根据 Word2vector 词义特征,在无语义规则支持、词语语义不明的情况下,形成 VSM 形式的文本特征化表示,适用于以 VSM 形式为特征化表示的文本分类算法。

1 研究动机及相关技术

本文的研究动机是构建一种基于上下文关系的文本特征化表示方法,采用 Word2vector 提取文本中的词语语义(词义)特征,并最终形成 VSM 形式的文本特征化表示。实现 Word2vector 词义特征提取方法的 Word2vec 工具是其相关的技术基础。

1.1 基于 Word2vector 的文本特征化

基于 Word2vector 信息检索文本特征化可以解决 2 个方面的问题:①根据 Word2vector 词向量难以形成文本特征化表示的问题;②在缺少语义规则库的情况下,文本特征化过程中词义特征提取的问题。本文研究动机的具体思路如下。

1)根据文本中的空格或统一粒度的分词规则划分每个词语。

2)针对由分词得到的词语,采用 Word2vector 方法提取其词义特征,并以词向量形式表示。

3)采用适当的聚类划分数量,对得到的词向量进行聚类,即实现对词语词向量的“最优适应度划分”的聚类。根据最终的聚类结果将词语替代表示为其词向量所属聚类划分的质心 $\boldsymbol{S}$,即用质心 $\boldsymbol{S}$ 代表其聚类划分内的词语,将词语语义特征近似认同为所属聚类划分的质心。

4)将“质心 $\boldsymbol{S}$ 所代表的词语”在该文本中出现频率计为质心 $\boldsymbol{S}$ 的频率,并统计词向量聚类质心 $\boldsymbol{S}$ 的逆向文件频率;参照 TF-IDF 模型构成词向量聚类质心频率模型,并生成 VSM 形式的特征化表示。

5)根据基于 Word2vector 的 VSM 特征化表示进行文本相似性分析。

根据研究动机,将利用 Word2vector 提取词义特征,获得文本中所有词语的词向量;而后,根据最优聚类效果适应度的划分对词语的词向量进行聚类,并根据聚类结果将词语替代表示为其词向量所属聚类划分的质心($\boldsymbol{S}$);最后,将质心所代表的词语在文本中的出现频率计为质心 $\boldsymbol{S}$ 的频率,并构成词向量聚类质心频率模型,用于特征化表示文本。词向量聚类质心频率模型蕴含词义特征,且属于 VSM 形式,可适用于现行的基于特征向量的信息检索算法

(如分类、回归、聚类)。

基于 Word2vector 的文本特征化与传统的词语统计机制不同,可通过分析 Example 1 表明。Example 1 中,2 个简单的句子可视为 2 个文本样本,并且构成一个极小的语料库。

Example 1.

Sample A. Men love holiday.

Sample B. Human enjoys vacation.

Sample A 和 Sample B 的含义极为相似,2 个文件之间的相关性和语义相似性是相当大的。

词语统计机制的文本特征化对 Example 1 中文本的向量化表示如表 1 所示。其中,在 $\boldsymbol{A}$,$\boldsymbol{B}$ 2 个向量中,不为零的词频值没有同时出现在 2 个文本样本中的相同词语上。这 2 个词频的正交向量表明,用于文本特征化表示的词语统计机制,不能有效表示 Example 1 中的语义相似性。而基于 Word2vector 的文本特征化表示则可依靠词向量所属聚类划分质心对词语的替代表示,实现词义特征统计,从而有效表示 Example 1 中的语义相似性。

表 1 词语统计机制的 Example 1 文本向量化表示

Tab.1 Statistical term measures on example 1

词汇	Men	love	holiday	Human	enjoys	vacation
词汇词频向量 $\boldsymbol{A}$	1	1	1	0	0	0
词汇词频向量 $\boldsymbol{B}$	0	0	0	1	1	1

1.2 Word2vector 及 Word2vec 工具

当文档作为信息检索的样本时,针对在不同文档中的每个词语,Word2vector 可依据其上下文关系有效地提取其语义(即词义特征),并以词向量的形式给出[10]。Word2vec 是 Word2vector 方法的模型实现软件工具包,能够基于词语的上下文关系,快速有效地训练并生成词向量。Word2vec 工具包含了 2 种训练模型,CBOW(continuous bag of word)与 Skip_gram。Word2vec 中训练模型的基础是神经网络语言模型(neural network language model ,NNLM)[11],其基本原理如图 1 所示。必须注意的是,Word2vector 的词义特征提取机制使得不同文档中相同的词所对应的词向量并不相同。所以,导致难以根据 Word2vector 的词向量形成信息检索样本的特征化表示,特别是难以形成 VSM 形式的样本特征化表示。

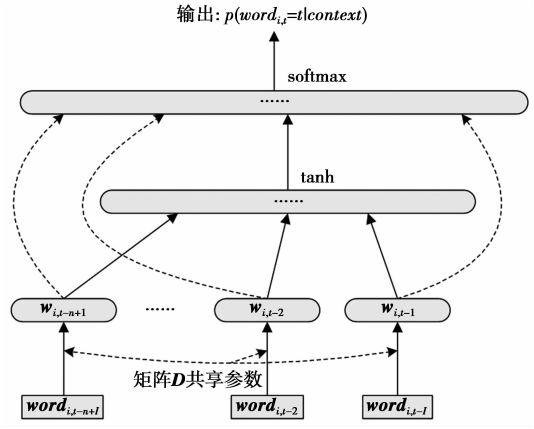

图 1 Word2vector 方法原理

Fig.1 Word2vector mechanism

Word2vec 作为一款将词表征为实数值向量的高效工具包[6]。其利用深度学习的思想,可通过训练把对文本内容的处理简化为多维向量空间中的向量运算,而向量空间上的相似度可以用来表示文本语义上的相似度。Word2vec 输出的词向量可以被用于自然语言处理相关的工作,如聚类、同义词查找、词性分析等。若将词语当做特征,则 Word2vec 就可将特征映射到多维向量空间,可为文本数据寻求更加深层次的特征表示。

Word2vec 使用的是 Distributed representation 的词向量表示方式[6]。Distributed representation 的基本思想是通过训练将每个词映射成 N 维实数向量,通过词之间的距离(如余弦相似度、欧氏距离等)判断它们之间的语义相似度[12]。Word2vec 采用一个三层的神经网络(含输入层-隐含层-输出层),Word2vec 的三层神经网络本身是对语言模型进行建模,同时获得一种词语在向量空间上的表示是 Word2vec 的真正目标(见图 1)。Word2vec 三层神经网络可根据词频用 Huffman 编码使得所有词频相似的词隐藏层激活的内容基本一致,出现频率越高的词语,所激活的隐藏层数目越少,可有效地降低计算的复杂度。因此,Word2vec 具备高效性。

2 基于 Word2vector 的文本特征化表示方法

基于 Word2vector 的文本特征化表示利用 Word2vector 提取词义特征,获得文本中所有词语的词向量。之后,根据最优聚类效果适应度的划分对

词语的词向量进行聚类,并根据聚类结果将词语替代表示为其词向量所属聚类划分的质心 $\boldsymbol{S}$。质心所代表的词语在文本中的出现频率计为 $\boldsymbol{S}$ 的频率,构成用于特征化表示文本的词向量聚类质心频率模型(semantic frequency-inverse document frequency, SF-IDF)。

2.1　方法说明

基于 Word2vector 的文本特征化表示方法,主要由以下步骤构成。

2.1.1　对样本进行词语分词

将文本中的词语视为 ASCII 字符串,根据空格或划分每个词语。将词语记为 $\boldsymbol{word}_{i,t}$,表示第 i 个样本中的第 t 种词语的分词,有 $i=\{1,2,\cdots,|D|\}$,$|D|$ 为数据集中 D 的样本数,$t=\{1,2,\cdots,n\}$,n 为词语种类数,所有文本中词语 $\boldsymbol{word}_{i,t}$ 的数量为 N,不同文本中的相同 ASCII 字符串识别为同一词语。

2.1.2　采用 Word2vec 工具提取词义特征

词向量初始化赋值时,不同文本中的相同词语具有一致的词向量,有 $\boldsymbol{w}_{i,t}=\boldsymbol{w}_{j,t}$。

针对由 2.1.1 节得到的词语,采用 Word2vector 方法,基于词语的上下文关系提取其词义特征,并以词向量形式表示。本步骤运用 Word2vec 工具包中的训练模型,可获得词语的词向量。训练模型以神经网络语言模型 NNLM 为基础,其原理如图 1 所示。

采用 NNLM 计算某一个上下文中一个词语 $\boldsymbol{word}_{i,t}$ 的概率,即 $p(\boldsymbol{word}_{i,t}=t\,|\,context)$,词向量是其训练的副产物。NNLM 根据数据集 D 生成一个对应的词汇表 V,其中的每一个词语都对应着一个标记 $\boldsymbol{word}_{i,t}$。通过数据集来构建训练样本并作为神经网络的输入,以确定神经网络的参数。NNLM 词语上下文样本的构建过程为:对于 D 中的任意一个词 $\boldsymbol{word}_{i,t}$,获取其上下文 $context(\boldsymbol{word}_{i,t})$(例如前 $n-1$ 个词),从而得到一个元组($context(\boldsymbol{word}_{i,t})$, $\boldsymbol{word}_{i,t}$)。以该元组作为神经网络的输入进行训练。NNLM 的输入层和传统的神经网络模型有所不同,输入的每一个节点单元是一个向量,向量的每一个分量为变量,在训练过程中对其进行变更,该向量即为词向量。由图 1 可知,对于每一个词 $\boldsymbol{word}_{i,t}$,NNLM 都将其映射成一个向量 $\boldsymbol{w}_{i,t}$,即为词向量。

Word2vec 生成的词向量 $\boldsymbol{w}_{i,t}$ 具体表示第 i 个文本中的第 t 种词语的词义特征,有 $i=\{1,2,\cdots,|D|\}$,$|D|$ 为样本数,全体样本中词语的词向量 $\boldsymbol{w}_{i,t}$ 的数量为 N。

2.1.3　词语语义特征替代表示

首先,采用最优聚类效果适应度下的聚类划分数量,对词向量进行 K-means 算法聚类[13],即实现对词语词向量的“最优适应度划分”的聚类。词向量的 K-means 聚类中,采用两词向量夹角的余弦值计算二者间的距离。

根据 2.1.2 节,所有样本中词语的词向量 $\boldsymbol{w}_{i,t}$ 的数量为 N,词向量 $\boldsymbol{w}_{i,t}$ 具体表示第 i 个样本中的第 t 种词语的词义特征。已知的样本分类数量为 C,而样本数量为 M。本步骤中,将词向量聚类划分的质心称为 $\boldsymbol{S}$(表示为词向量空间中的向量),$\boldsymbol{S}$ 的数量 k 即是聚类划分个数。

为度量词向量空间中的 K-means 聚类效果,本文给出聚类划分数量适应性的计算。为表示聚类划分数量适应性,令 $f(k)$ 为体现聚类效果适应度的函数,表示为

$$f(k)=\frac{\alpha}{\beta},\quad N\leqslant k\leqslant N\times C \tag{1}$$

(1)式中:α 为 k 个 $\boldsymbol{S}$ 向量间的平均余弦距离;β 为 k 个聚类划分内的词向量间平均余弦距离的均值,具体地有

$$\alpha=\frac{1}{k}\sum\cos(\boldsymbol{S},\boldsymbol{S}') \tag{2}$$

$$\beta=\frac{1}{k}\sum_{b=1}^{k}\overline{\cos(\boldsymbol{w}_{i,t},\boldsymbol{w}'_{i,t})} \tag{3}$$

(2)—(3)式中:$\boldsymbol{S}$ 与 $\boldsymbol{S}'$ 为不同聚类划分的质心向量;$\boldsymbol{w}_{i,t}$ 与 $\boldsymbol{w}'_{i,t}$ 是类属于第 b 个聚类划分中的不同词语的词向量。

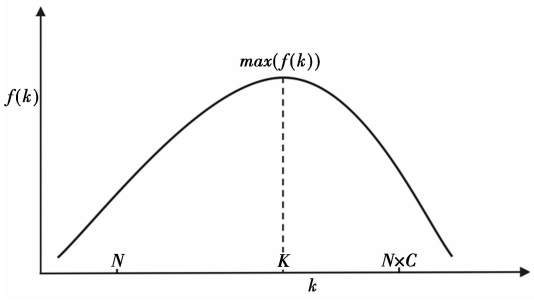

图 2　聚类效果适应度函数

Fig.2　Clustering effect of fitness function

设聚类划分个数 $k\in[N,N\times C]$,且为正整数,当 $f(k)=\max(f(k))$ 时,令最优聚类效果适应度下的聚类划分数量 $K=k$,$f(K)$ 是聚类效果适应度的最大值。经计算可知,函数 $f(k)$ 在 N 到 K 的区间是单调递增的,在 K 到 $N\times C$ 的区间是单调递减的,函数

$f(k)$的分布如图 3 所示。

当$f(k)=\max(f(k))$时,$K=k$,$f(K)$是聚类效果适应度函数的极值,即最优聚类效果适应度,K-means 聚类质心 $\boldsymbol{S}$ 的数量最终确定为 K。

根据最终的聚类结果将词语替代表示为其词向量所属聚类划分的质心 $\boldsymbol{S}$。具体地,当$f(k)=\max(f(k))$时,最优聚类效果适应度下的聚类划分数量 $K=k$,将任意词语 $\boldsymbol{w}_{i,t}$ 替代表示为其词向量所属聚类划分的质心 $\boldsymbol{S}$,即将词语的特征近似认同为所属聚类划分的质心。在任意局部词向量空间中,用质心 $\boldsymbol{S}$ 代表其聚类划分内的词语,其对应关系如图 3 所示。

图 3 中的具体替代表示关系为

$$\boldsymbol{S}_b \leftrightarrow \{\boldsymbol{word}_{i,t} \mid \boldsymbol{w}_{i,t} \in \boldsymbol{W}_b\} \tag{4}$$

(4)式中:第 b 个聚类质心 $\boldsymbol{S}_b$ 所代表的词语 $\boldsymbol{word}_{i,t}$ 构成一个词语集合;$\boldsymbol{w}_{i,t}$ 是词语 $\boldsymbol{word}_{i,t}$ 的词向量;$\boldsymbol{W}_b$ 是类属于质心 $\boldsymbol{S}_b$ 所在聚类划分的词向量所对应的词语的集合。

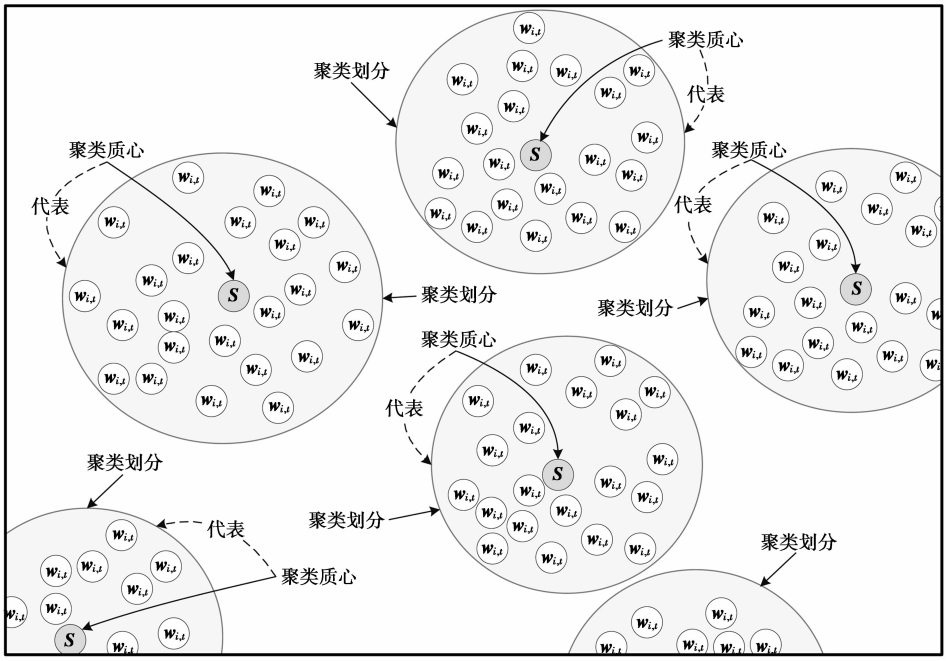

图 3 词向量空间中根据聚类的替代表示关系

Fig.3 Substitution in word vector space

词语语义特征替代表示是采用聚类质心代表该聚类划分中所有词语及其语义,有可能造成信息损失。将由结果表明其所造成的信息损失可视为在文本特征表示过程中可承受的影响。

2.1.4 基于词义特征替代表示构建文本特征化表示模型

首先,统计每个词语在一个样本中出现的频率,根据 2.1.3 节给出的质心 $\boldsymbol{S}$ 与词语的替代表示关系,将第 b 个质心 $\boldsymbol{S}_b$ 所代表的词语在该样本中的出现频率计为质心 $\boldsymbol{S}_b$ 的频率;并统计词向量聚类质心 $\boldsymbol{S}_b$ 的逆向文件频率,有 $b=\{1,2,\cdots,K\}$。而后,参照 TF-IDF 模型构成词向量聚类质心频率模型——SF-IDF。

TF-IDF 模型[1]中,样本 $\boldsymbol{doc}_i$ 的特征化表示由特征向量 $\boldsymbol{d}_i$ 实现,有

$$\boldsymbol{d}_i = (\boldsymbol{d}_{i(1)}, \boldsymbol{d}_{i(2)}, \cdots, \boldsymbol{d}_{i(n)}) \tag{5}$$

向量 $\boldsymbol{d}_i$ 中第 t 维元素 $\boldsymbol{d}_{i(t)}$ 计算方式为

$$\boldsymbol{d}_{i(t)} = TF(\boldsymbol{word}_t, \boldsymbol{doc}_i) \cdot IDF(\boldsymbol{word}_t) \tag{6}$$

(6)式中,$TF(\boldsymbol{word}_t, \boldsymbol{doc}_i)$是词语 $\boldsymbol{word}_t$ 在样本 $\boldsymbol{doc}_i$ 中的频率,其计算方式为

$$TF(\boldsymbol{word}_t, \boldsymbol{doc}_i) = \frac{count(\boldsymbol{word}_t)}{\sum_{j=1}^{n} count(\boldsymbol{word}_j)} \tag{7}$$

(7)式中的分子是该词语在样本中的出现次数,而分母则是在文件中所有词语的出现次数之和;$IDF(\boldsymbol{word}_t)$为词语 $\boldsymbol{word}_t$ 的逆向文件频率,其计算方式为

$$IDF(\boldsymbol{word}_t) = \frac{|D|}{|\{\boldsymbol{doc}_i \mid \boldsymbol{word}_t \in \boldsymbol{doc}_i\}|} \tag{8}$$

(8)式中:D 为样本 $\boldsymbol{doc}_i$ 的构成数据集;$|D|$为数据集 D 中样本的总数;$|\{\boldsymbol{doc}_i \mid \boldsymbol{word}_t \in \boldsymbol{doc}_i\}|$为包含词语 $\boldsymbol{word}_t$ 的样本数量。

参照 TF-IDF 模型,SF-IDF 模型具体构成如下。

$SF(\boldsymbol{S}_b,\boldsymbol{doc}_i)$是词向量聚类质心 $\boldsymbol{S}_b$ 在文本 $\boldsymbol{doc}_i$ 中的频率,其计算方式为

$$SF(\boldsymbol{S}_b,\boldsymbol{doc}_i)=\sum_{\boldsymbol{w}_{i,t}\in\boldsymbol{W}_b}TF(\boldsymbol{w}_{i,t}) \tag{9}$$

(9)式中:$TF(\boldsymbol{w}_{i,t})$表示词语 $\boldsymbol{w}_{i,t}$在文本 $\boldsymbol{doc}_i$ 中出现的频率;$SF(\boldsymbol{S}_b,\boldsymbol{doc}_i)$仅累计文本 $\boldsymbol{doc}_i$ 中由质心 $\boldsymbol{S}_b$ 所代表的词语的频率。

$IDF(\boldsymbol{S}_b)$为词向量聚类质心 $\boldsymbol{S}_b$ 的逆向文件频率,其计算方式为

$$IDF(\boldsymbol{S}_b)=\frac{|D|}{|\{\boldsymbol{doc}_i\mid \boldsymbol{w}_{i,t}\boldsymbol{w}_{i,t}\in\boldsymbol{W}_b\in\boldsymbol{doc}_i\}|} \tag{10}$$

(10)式中:D 为文本 $\boldsymbol{doc}_i$ 的构成数据集;$|D|$为数据集 D 中样本的总数;$|\{\boldsymbol{doc}_i\mid\boldsymbol{w}_{i,t}\boldsymbol{w}_{i,t}\in\boldsymbol{W}_b\in\boldsymbol{doc}_i\}|$为包含由质心 $\boldsymbol{S}_b$ 所代表的词语的样本的数量。

SF-IDF 模型中,文本 $\boldsymbol{doc}_i$ 的特征化表示由特征向量 $\dot{\boldsymbol{d}}_i$ 实现

$$\dot{\boldsymbol{d}}_i=(\dot{\boldsymbol{d}}_{i(1)},\dot{\boldsymbol{d}}_{i(2)},\cdots,\dot{\boldsymbol{d}}_{i(K)}) \tag{11}$$

向量 $\dot{\boldsymbol{d}}_i$ 中第 b 维元素 $\dot{\boldsymbol{d}}i(b)$计算方式为

$$\dot{\boldsymbol{d}}_{i(b)}=SF(\boldsymbol{S}_b,\boldsymbol{doc}_i)\cdot IDF(\boldsymbol{S}_b)。 \tag{12}$$

SF-IDF 模型属于 VSM(向量空间模型)形式,用于特征化表示一个文本。

2.1.5　文本相似性分析

根据 SF-IDF 模型特征化表示,计算 2 个文本间的相似度;并据此进行信息检索领域中样本分类算法的执行。

采用 SF-IDF 模型特征化表示文本,任意 2 文本 $\boldsymbol{doc}_i$ 与 $\boldsymbol{doc}'_i$ 间相似性由相似度函数 $Sim(\boldsymbol{doc}_i,\boldsymbol{doc}'_i)$表示,其具体计算方式为

$$Sim(\boldsymbol{doc}_i,\boldsymbol{doc}'_i)=\cos(\dot{\boldsymbol{d}}_i,\dot{\boldsymbol{d}}'_i), \tag{13}$$

(13)式中,$\cos(\dot{\boldsymbol{d}}_i,\dot{\boldsymbol{d}}'_i)$ 为 SF-IDF 向量空间中特征向量$\dot{\boldsymbol{d}}_i$ 与$\dot{\boldsymbol{d}}'_i$ 间夹角的余弦值。

2.2　方法分析

据互信息理论,可给出基于 Word2vector 的文本特征化表示方法的有益性分析。

假定 $\mathscr{X}$与 $\mathscr{Y}$为表示样本 X 与 Y 词语语义内容的随机变量。若样本已知,$\mathscr{X}$和 $\mathscr{Y}$的互信息(mutual information)表示两者间的不确定性归纳。样本 X 与 Y 之间的互信息 $I(\mathscr{X};\mathscr{Y})$定义为

$$I(\mathscr{X};\mathscr{Y})=\sum_{x_i\in X}\sum_{y_j\in Y}P(x_i,y_j)\log\frac{P(x_i,y_j)}{P(x_i)P(y_j)} \tag{14}$$

在词语统计机制下,概率 $P(x_i)$或 $P(y_j)$由统计样本 X 或 Y 中 x_i 或 y_i 的出现次数(词频)计算,并通过文本语料库规模(N)进行归一化。联合概率 $P(x_i,y_j)$由 x_i 与 y_i 之间的存在关系的次数(相关频率)统计,并根据 N 进行归一化。x_i 与 y_i 之间存在的关系为“相同”或特定关系[14]。

以 Example 1 为例,在任意 Sample A 中的词语与 Sample B 中的词语之间,并没有可统计的存在关系的次数,它们并不“相同”,也没有表现出特定关系。故词语统计的特征提取显示 $P(x_i,y_j)=0$,且样本间的互信息 $I(\mathscr{X};\mathscr{Y})=0$。可以证明词语统计机制的特征提取丢失词语语义内容所产生的互信息。

基于 Word2vector 的文本特征化表示的语义特征提取方式,是进行词语语义特征替代表示。因而在不同的样本中,词语可由词语语义特征替代表示产生关系。在 Example 1 的 Sample A 与 Sample B 中的词语间,存在可统计的存在关系的次数,尽管它们并不“相同”,却可表现出特定关系。例如词语“Men”与“Human”的语义根据上下文关系提取,并通过词语语义特征替代表示进行近似认同。上述分析表明,基于 Word2vector 的文本特征化表示可提供文本词语语义层面的信息概率加权量(probability weighting information ,PWI)[15]。

3　实验及结果分析

根据词向量聚类质心频率模型,采用信息检索领域中的经典样本分类算法——权重邻居不均衡分类样本集分类算法(neighbor-weighted k-nearest neighbor for unbalanced text corpus ,NWKNN)执行文本分类。NWKNN 是权重邻居(k-nearest neighbor,KNN)算法,用于不均衡分类样本集的样本分类判别。该算法在信息检索领域中被视为一种高效的分类算法,其公式为[16]

$$score(\boldsymbol{doc},\boldsymbol{c}_i)=Weight_i\Big(\sum_{\boldsymbol{doc}_j\in KNN(d)}Sim(\boldsymbol{doc},\boldsymbol{doc}_j)\delta(\boldsymbol{doc}_j,\boldsymbol{c}_i)\Big), \tag{15}$$

函数 $score(\boldsymbol{doc},\boldsymbol{c}_i)$求得将文本 $\boldsymbol{doc}$ 归于分类 $\boldsymbol{c}_i$ 的评估值,用于判定文本 $\boldsymbol{doc}$ 归属于拥有最高评估值的分类;函数 $score(\boldsymbol{doc},\boldsymbol{doc}_i)$表示样本 $\boldsymbol{doc}$ 与已知类别样本 $\boldsymbol{doc}_i$ 的相似度,采用向量余弦距离计算;$Weight_i$为分类权重设定值,根据 NWKNN 算法经验化赋值为 3.5[16];函数 $\delta(doc_j,c_i)$表示样本 $\boldsymbol{doc}_j$

是否属于类别 $\boldsymbol{c}_i$,若样本 $\boldsymbol{doc}_j$ 属于类别 $\boldsymbol{c}_i$,则该函数取值为1,否则,该函数取值为0。

样本分类的性能评估采用 F1-measure 标准。该标准结合召回率 *Recall* 和准确率 *Precision* 的评估度量 *F*1 如下

$$F1 = \frac{2 \times Recall \times Precision}{(Recall + Precision)} \tag{16}$$

运用 F1-measure 标准,可观察分类系统针对数据集的分类效果。为便于比较,将总结文本分类结果的宏观 F1 度量值 *Macro-F*1。同时,可以得到文本分类结果的平均准确率。

由于实验步骤中将文本中的词语视为 ASCII 字符串,根据空格或划分每个词语,所选用的文本数据集均可视为无语义规则支持、词语语义不明的文本集合。

分别以路透社文本集 Reuter-21578、维基百科 XML 数据 Wikipedia XML 为文本数据集,采用 NWKNN 算法进行文本分类实验,并采用 F1-measure 标准进行样本分类的效果评估,SF-IDF 向量与现有技术中 TF-IDF 向量的分类效果对比如表2,表3所示。

表2 Reuter-21578 数据集上 TF-IDF 向量与 SF-IDF 向量的分类效果比较

Tab.2 Classifications results of TF-IDF and SF-IDF on Reuter-21578 %

特征化表示	*Macro-F*1	平均召回率	平均准确率
TF-IDF	54.2	54.7	57.1
SF-IDF	62.1	62.7	63.3

表3 Wikipedia XML 数据集上 TF-IDF 向量与 SF-IDF 向量的分类效果比较

Tab.3 Classifications results of TF-IDF and SF-IDF on Wikipedia XML %

特征化表示	*Macro-F*1	平均召回率	平均准确率
TF-IDF	45.6	43.1	48.7
SF-IDF	45.3	51.9	59.2

据表2,表3所述,可见 SF-IDF 向量的分类效果明显优于现有技术中 TF-IDF 向量。在没有语义规则支持且词义不明的情况下,在 Reuter-21578 数据集上平均准确率由原有的57.1%提高到63.3%,在 Wikipedia XML 数据集上平均准确率由原有的48.7%提高到59.2%。

实验结果显示,在没有语义规则支持且词义不明的情况下,针对文本相似性分类任务,SF-IDF 模型相较 TF-IDF 模型拥有更优良的 F1-measure 评估结果,说明本文所提出的特征化表示方法具备文本词义特征提取方面的优势。

4 结束语

本文所提出的词向量聚类质心频率(SF-IDF)模型,采用信息检索领域经典样本分类算法 NWKNN,在公用数据集 Reuter-21758 与 Wikipedia XML 之上,与 TF-IDF 模型进行对比实验,展示了明显优势。SF-IDF 模型提高了文本相似度计算的准确性,提升了文本分类准确度,并拓展了信息检索领域中向量空间模型的构建方法。

SF-IDF 模型所实现的方法,解决了根据 Word2vector 词向量难以形成文本特征化表示的问题,可在无语义规则支持的情况下,构成基于上下文的文本特征化表示。因此,SF-IDF 还可应用于分析无法被自然语言直接解读的文本或数据链报文(如 Link-16,Link-22)。

今后基于 Word2vector 的文本特征化表示方法的研究工作将尝试采用密度聚类算法执行词向量最优聚类效果适应度下的聚类,并展开多种文本数据集上的信息检索试验。

参考文献:

[1] ZHANG W,YOSHIDA T,TANG X. A comparative study of TF * IDF, LSI and multi-words for text classification [J]. Expert Systems with Applications, 2011, 38(3): 2758-2765.

[2] TU Shouzhong,HUANG Minlie.Mining microblog user interests based on TextRank with TF-IDF factor[J].The Journal of China Universities of Posts and Telecommunications, 2016, 23(5):40-46.

[3] PURDA L,SKILLICOM D. Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection[J].Contemporary Accounting Research, 2015, 32(3): 1193-1223.

[4] LEQV,MIKOLOV T.Distributed Representations of Sentences and Documents [J]. Computer Science, 2014, 4(32):1188-1196.

[5] JING L,NG M K, HUANG J Z.Knowledge-based vector space model for text clustering[J].Knowledge and Information Systems, 2010, 25(1):35-55.

[6] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compo-

sitionality[J].Advances in Neural Information Processing Systems,2013,10(26): 3111-3119.

[7] Poll 的笔记.文本深度表示模型——word2vec & doc2vec 词向量模型[EB/OL].(2016-04-24)[2017-02-20].http://www.cnblogs.com/maybe2030/p/5427148.html.

[8] 唐明,朱磊,邹显春. 基于 Word2Vec 的一种文档向量表示[J].计算机科学,2016,43(6):214-217.
TANG Ming, ZHU Lei, ZOU Xianchun. Document Vector Representation Based on Word2Vec[J].Computer Science,2016, 43(6):214-217.

[9] 杨河彬,贺樑,杨静.一种融入用户点击模型 Word2Vec 查询词聚类[J].小型微型计算机系统,2016,37(4):676-681.
YANG Hebin, HE Liang, YANG Jing.Query Clustering Using CT-Word2Vec Model [J].Journal of Chinese Mini-Micro Computer Systems , 2016 ,37(4):676-681.

[10] MIKOLOV T, CHEN K, CORRADO G, et al. Computing numeric representations of words in a high-dimensional space:United States,13/841,640[P].2015-05-19.

[11] MARTÍNEZ G E, ESPAÑA B C, TIEDEMANN J, et al. Word's vector representations meet machine translation [C]//Gertjan van Noord.Workshop on Ssst. Doha,Qatar: Association for Computational Linguistics,2014:132-134.

[12] MIKOLOV T, YIH W, ZWEIG G.Linguistic Regularities in Continuous Space Word Representations [C]//Ken Church. HLT-NAACL. Atlanta, Georgia: Association for Computational Linguistics,2013:746-751.

[13] KANUNGO T, MOUNT D M, NETANYAHU N S, et al. An efficient k-means clustering algorithm: Analysis and implementation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,24(7):881-892.

[14] van RIJSBERGEN C J. Information retrieval [M]. London: Butterworths Press,1979:113-252.

[15] CHUM O, PHILBIN J, ZISSERMAN A. Near Duplicate Image Detection: min-Hash and tf-idf Weighting [EB/OL]//(2008-08-01)[2017-04-20]. http://www.cs.jhu.edu/~misha/ReadingSeminar/Papers/Chum08.pdf.

[16] TAN S. Neighbor-weighted k-nearest neighbor for unbalanced text corpus[J].Expert Systems with Applications, 2005,28(4):667-671.

作者简介:

周顺先(1968—),男,湖南新邵人,教授,博士,主要研究方向为机器学习、信息检索。E-mail: zsx_hd@hnu.edu.cn。

蒋　励(1988—),男,湖南郴州人,讲师,工程师,硕士,主要研究方向为网络安全、信息检索。E-mail: 623887366@qq.com。

林霜巧(1991—),女,贵州贵阳人,硕士生,主要研究方向为软件服务质量、信息检索。E-mail: linshq@csu.edu.cn。

龚德良(1964—),男,湖南沅江人,教授,主要研究方向为计算理论、信息检索。E-mail: Gdl2865605@163.com。

王鲁达(1981—),男,山东济南人,讲师,博士,主要研究方向为数据分析、信息检索。E-mail: 9702361@qq.com。

(编辑:王敏琦)