基于 Word2fea 模型的文本建模方法

卫 华,韩立新,夏建华
(河海大学 计算机与信息学院,江苏 南京 211100)

摘 要:文本聚类在数据挖掘和机器学习中发挥着重要作用,该技术经过多年的发展,已产生了一系列的理论成果。传统向量空间模型的文本建模方法存在维度高、数据稀疏和缺乏语义信息等问题,然而仅仅引入词典的文本建模部分解决了语义问题却又受限于人工词典词量少、人工耗力大等多种问题。文中借鉴主题模型的思想,提出一种以 word2vec 算法得到词向量为基础,词聚类的类别为主题,结合文本中主题的频率、分布范围、位置因子等特征以获得文本在类别空间上的特征向量,完成文本建模的方法 word2fea。将其与两种文本建模方法 VSM 和 word2vec_base 进行比较,实验结果表明该方法能够明显提高文本分类准确率。

关键词:word2vec;文本建模;文本分类;word2fea
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2016)02-0165-03
doi:10.3969/j.issn.1673-629X.2016.02.037

Text Modeling Method Based on Word2fea Model

WEI Hua, HAN Li-xin, XIA Jian-hua
(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: Text classification plays an important role in data mining and machine learning, which has produced a series of theory after years of development. The traditional text modeling method of vector space model has the problems of high dimension, sparse data, and the lack of semantic. However, the text modeling introduced the artificial dictionary is constrained by quantity of words, artificial power consumption and other problems. By referencing the idea of topic model, a text modeling method word2fea was presented which based on the model of word2vec for the topic clusters with the word vectors, meanwhile combined with the frequency, distribution and location of the topic on documents to obtain the feature of the text. Compared with two text modeling methods, VSM and word2vec_base, the experimental results show that this method can significantly improve the accuracy of text classification.

Key words: word2vec; text modeling; text classification; word2fea

0 引 言

随着互联网信息的飞速增长,计算机信息处理已然进入大数据时代。文本形式是互联网信息呈现的主要方式,而对互联网信息的挖掘主要涉及两方面的问题:一是文本信息的挖掘,二是文本信息的组织。可见,文本挖掘是进行文本信息融合的前提与基础。

文本建模是文本挖掘的基石,在文本聚类,分类,信息检索,自动问答系统,自动摘要等场景中均有着重要的地位。其中最流行的是基于向量空间模型(VSM)[1],但是存在中文词维数大,稀疏度高,同义词、多义词等语义问题。基于词项语义来考察文本相似度的方法利用外部词典,如知网、同义词词林等[2-3],虽然解决了部分语义问题,但又存在词典词数小、词典构建困难等问题。在主题模型 LSI、PLSI 和 LDA 等[4-6]提出以后,以其可以发现潜在主题等优势,被广泛地用于文本主题挖掘[7-9],弥补了前两种问题的不足。然而这三种模型均需要大量训练样本学习,训练难度大并且非常耗时,学习到的隐含主题有噪声。基于 word2vec 模型和 tf-idf 进行文本建模[10],在文本分类中,对效率和准确率都有所提升,但是未考虑文本结构特性。

文中通过主题模型对文本进行建模,首先通过 word2vec 对词向量进行聚类的主题分布,利用文本的上下文统计信息,有效降低文本向量维度,同时解决同

收稿日期:2015-04-24 **修回日期**:2015-07-28 **网络出版时间**:2016-01-04
基金项目:中央高校基本科研业务费专项资金(2014B33014)
作者简介:卫 华(1991-),男,硕士研究生,研究方向为信息检索、数据挖掘;韩立新,教授,博士生导师,研究方向为信息检索、模式识别、数据挖掘。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160104.1505.040.html

义词、多义词以及错别字问题。其次,针对文本结构特性,以主题的频率、分布范围、位置等因素对主题进行特征提取并进行建模,命名为 word2fea 算法。在复旦中文语料库进行测试,结果表明在文本分类效果上有所提高。

1 word2fea 算法对文本建模

1.1 神经网络语言模型

神经网络语言模型(Neural Network Language Model)由 Bengio 于 2003 年提出[11],利用神经网络训练语言模型的思想最早由徐伟提出[12],使用一个三层神经网络来构建语言模型,并且假设这种语言遵循 n -gram 语言模型。该模型采用的是词向量(Distributed Representation),即将每个英文单词表示成一个浮点向量,模型见图 1。

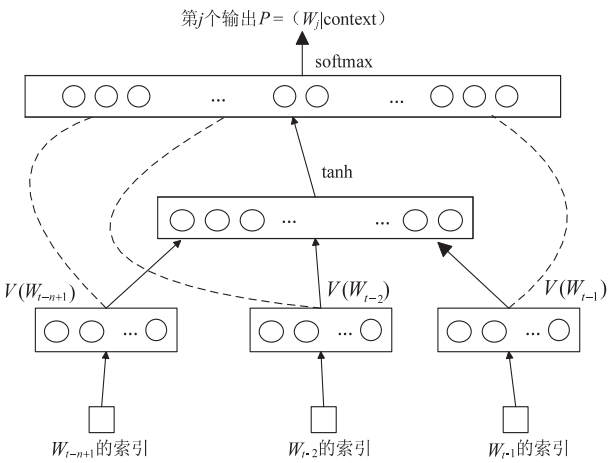

图 1 神经网络语言模型结构示意图

目标是要学到的 n -gram 模型如式(1):

$$f(W_t, W_{t-1}, \cdots, W_{t-n+1}) = p(W_t \mid W_1^{t-1}) \tag{1}$$

需要满足的约束如公式(2)、(3):

$$f(W_t, W_{t-1}, \cdots, W_{t-n+1}) > 0 \tag{2}$$

$$\sum_{i-1}^{|V|} f(W_t, W_{t-1}, \cdots, W_{t-n+1}) = 1 \tag{3}$$

首先将输入的词都映射为一个向量,该映射用 $V(W)$ 表示,$V(W_{t-1})$ 即为 W_{t-1} 的词向量。网络的输入层将 $V(W_{t-n+1}), V(W_{t-n+2}), \cdots, V(W_{t-1})$ 这 $n-1$ 个向量首尾相连接,组成一个 $(n-1)\times m$ 维的向量,记为 x。网络用 $d + H_x$ 计算得到,其中 d 为偏置项,随机初始化值。使用 tanh 函数作为激活函数。网络的第三层用节点 y_i 表示,一共有 $|N|$ 个节点,其中 $|N|$ 表示词表的大小。最后使用 softmax 激活函数,将输出值 y 归一化成概率,y 的计算如式(4):

$$y = b + W_x + U \times \tanh(d + H_x) \tag{4}$$

最后使用随机梯度下降法将模型优化。优化结束之后,训练得到词向量,进而得到语言模型。Softmax 模型使得概率取值为(0,1),因此不会出现概率为 0 的情况,也就是自带平滑,无需传统 n -gram 模型中那些复杂的平滑算法。实验也表明神经网络语言模型比带有平滑算法的 n -gram 模型的算法效果要好。

word2vec 是 Google 开源的用于计算词向量的工具,主要有模型 CBOW(Continuous Bag-Of-Words model)和 Skip-gram(continuous Skip-gram model)两种[13],基本思想来自于神经网络语言模型。word2vec 通过对大批文本进行训练,将文本中的词转化为 N 维向量空间中的词向量,而向量空间上的相似度可以用来计算词或文本等语义上的相似度。因此,word2vec 输出的词向量可以被用来做很多与自然语言处理相关的工作,比如聚类、找同义词、自动翻译等等。

1.2 Skip-gram 模型

Skip-gram 模型的网络结构见图 2,包括三部分:输入层、投影层、输出层。

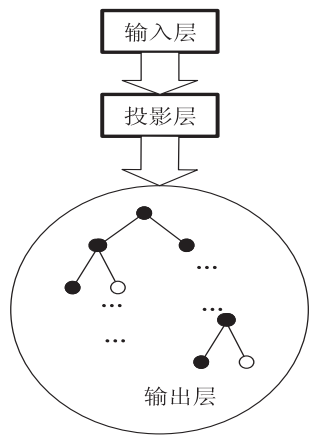

图 2 Skip-gram 模型结构示意图

输入层:只含当前样本的中心词 w 的词向量 $V(w)$。

投影层:恒等投影,把 $V(w)$ 投影到 $V(w)$。

输出层:对应一棵哈夫曼树,以语料中的词作为叶子节点,每个词在语料中出现的次数作为权值构造的哈夫曼树,在这个哈夫曼树中,叶子节点数对应这词典中的词数。

1.3 word2fea 文本建模方法

word2fea 的文本建模方法主要包含 4 部分:预处理、主题聚类、文本主题特征计算、文本向量化。其流程如图 3 所示。

图 3 word2fea 文本建模算法流程图

首先对文本库进行预处理,主要包括中文分词、去除停用词等,分词系统使用中科院的 ICTCLAS[14],并将另存处理后的文本库以一篇文档的形式用于 word2vec 工具训练词向量。

在主题聚类中,采用 word2vec 中的 Skip-gram 对本文进行词向量训练,将训练后的词向量使用 K-means 进行聚类,聚类数 K 即为主题数, K 的取值范围为 50~400,间隔为 50。经过聚类后的词袋即代表不同的主题。

最后对每篇文档进行主题特征计算,将文本的主题特征转化为文本向量。使用 LibSVM[15] 作为分类器,对语料库进行训练,并预测分类准确率。

1.4　文本主题特征计算

对于主题权重的定义,唐晓丽等[10]统计每个词所属的类别,对同一类别下所有特征词的 tf-idf 值求和并进行归一化。文中在 tf-idf 之外,综合考虑文本中不同主题出现的频率、范围和位置等特征,主要从 3 个方面对主题权重进行定义:

$$D_{t,d} = \mathrm{loc}_{t,d} + \mathrm{fre}_{t,d} + \mathrm{sca}_{t,d} \tag{5}$$

(1)主题词语在文本出现的频率。频率越大表明该主题对该文本贡献越大。定义式(6):

$$\mathrm{fre}_{t,d} = \frac{N_t}{N_d} \tag{6}$$

其中, N_t 为主题 t 的频次; N_d 为文档 d 的频次。

(2)主题词语出现的范围。若该主题词语在某一类中频繁出现,则认为它在此类文本中价值较大,即该主题词语在此类中出现频率不仅高且范围较小。定义式(7):

$$\mathrm{sca}_{t,d} = \frac{S_t}{S_d} \tag{7}$$

其中, S_t 为主题 t 在语料库中所有出现的类别数; S_d 为语料库中总的类别数。

(3)主题词语位置因子。主题词语在文本出现的位置不同贡献也有所不同,出现在段首和段尾中的主题词语要比在内容中的贡献大。定义式(8):

$$\mathrm{loc}_{t,d} = \{c \mid 0.5, 0.2, 0.3\} \tag{8}$$

其中,段首的权重最高为 0.5,段尾为 0.3,段中为 0.2。

1.5　文本向量化

将词向量聚类为主题后,并通过 1.4 为每篇文档进行主题特征计算,将每篇文档主题分布的特征转化为向量的形式如式(9):

$$\mathrm{Doc}_i = (D_{t_1,i}, D_{t_2,i}, \cdots, D_{t_n,i}) \tag{9}$$

其中, Doc_i 为第 i 篇文档的向量表示形式; $D_{t_i,i}$ 为第 i 篇文档中主题 i 的权重,其中共有 n 个主题。

2　实验设计与分析

2.1　数据集与度量标准

文中在中文语料上进行了实验,采用复旦中文语料库,挑选其中 10 个类别,分别是“环境”“交通”“计算机”“教育”“经济”“军事”“体育”“医药”“艺术”“政治”,每个类别挑选 200 篇文本作为语料集,每个类均按照 4:1 的比例划分,80% 作为训练集,20% 作为测试集。实验采用 SVM 分类器对训练集进行训练,用测试集验证最终分类结果,实验采用分类准确率 P 作为最终的评测指标。

2.2　实验结果分析

从图 4 中可知,当主题数选择 300 时,准确率达到最高值。选择最优主题数之后就得到基于 word2fea 模型进行文本建模的分类结果。从图 5 中可以看出,文中方法比基于 VSM 和 word2vec_base 的分类准确率有明显提升。

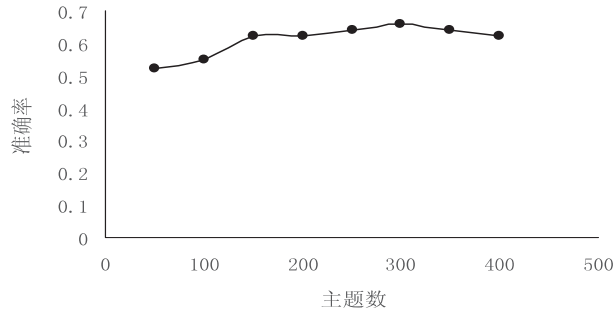

图 4　不同主题数 K 下的分类结果

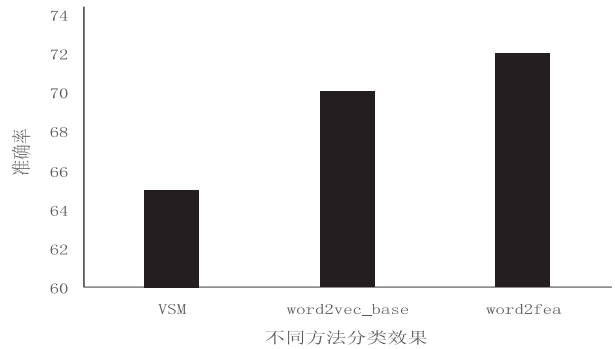

图 5　VSM,word2vec_base 与 word2fea 对比结果图

3　结束语

文中将 word2vec 模型应用到文本建模中。利用了 word2vec 模型的词向量高效性,加入了文本的深层语义知识,从而使分类更加精准。利用隐主题映射文本主题空间,在文本主题特征计算中,综合考虑文本主题频次、范围以及位置因子,提高了分类效果。实验结果表明,文中所采用的方法是一种能够有效提高文本分类准确率的方法。

由于 word2vec 非常容易扩展,后续研究将在 word2vec 模型的基础上继续探讨文本建模方法以及基于其上的文本挖掘,如文本分类、相似项挖掘等。

参考文献:

[1]　Salton G, Others A. A vector space model for automatic inde-

(下转第 173 页)

energy consumption via sleeping and rate-adaptation[C]// Proc of 5th USENIX symposium on networked systems design and implementation. [s. l.]:USENIX,2008:323-336.

[4] Singh S,Yiu C. Putting the cart before the horse:merging traffic for energy conservation[J]. IEEE Communications Magazine,2011,49(6):78-82.

[5] Christensen K,Reviriego P,Nordman B,et al. IEEE 802.3 az: the road to energy efficient Ethernet[J]. IEEE Communications Magazine,2010,48(11):50-56.

[6] Miao G,Himayat N,Li G Y,et al. Distributed interference-aware energy-efficient power optimization[J]. IEEE Trans on Wireless Communication,2011,10(4):1323-1333.

[7] 郭秉义. 绿色通信网络的节能方法研究[D]. 广州:华南理工大学,2014.

[8] 郭晓达. 下一代接入网节能技术研究[D]. 北京:北京邮电大学,2013.

[9] 苏俊基,杨龙祥,朱乐恒. 未来网络休眠机制的研究[J]. 微型机与应用,2014,33(24):59-61.

[10] Christensen K J,Gunaratne C,Nordman B,et al. The next frontier for communications networks:power management[J]. Computer Communications,2004,27(18):1758-1770.

[11] Bolla R,Bruschi R,Davoli F,et al. Energy efficiency in the future internet:a survey of existing approaches and trends in energy-aware fixed network infrastructures[J]. IEEE Communications Surveys & Tutorials,2011,13(2):223-244.

[12] 张寅翔. 成本与能效优化的虚拟网络映射算法研究[D]. 南京:南京邮电大学,2013.

[13] Silva T,Arsenio A. A survey on energy efficiency for the future internet[J]. International Journal of Computer and Communication Engineering,2013,2(5):589-589.

[14] Botero J F,Hesselbach X,Duelli M,et al. Energy efficient virtual network embedding[J]. IEEE Communications Letters,2012,16(5):756-759.

[15] Fischer A,Botero J F,Beck M T,et al. Virtual network embedding:a survey[J]. IEEE Communications Surveys & Tutorials,2013,15(4):1888-1906.

(上接第 167 页)

xing[J]. Communications of the ACM,1975,18(10):613-620.

[2] 李　峰,李　芳. 中文词语语义相似度计算-基于《知网》2000[J]. 中文信息学报,2007,21(3):99-105.

[3] 梅家驹,竺一鸣,高蕴琦,等. 编纂汉语类义词典的尝试-《同义词词林》简介[J]. 辞书研究,1983(1):133-138.

[4] Deerwester S C,Dumais S T,Landauer T K,et al. Indexing by latent semantic analysis[J]. JASIS,1990,41(6):391-407.

[5] Hofmann T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. [s. l.]:ACM,1999:50-57.

[6] Blei D M,Ng A Y,Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research,2003,3:993-1022.

[7] 张志飞,苗夺谦,高　灿. 基于 LDA 主题模型的短文本分类方法[J]. 计算机应用,2013,33(6):1587-1590.

[8] 王振振,何　明,杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学,2013,40(12):229-232.

[9] 孙昌年. 基于主题模型的文本相似度计算研究与实现[D]. 合肥:安徽大学,2012.

[10] 唐晓丽,白　宇,张桂平,等. 一种面向聚类的文本建模方法[J]. 山西大学学报:自然科学版,2014,37(4):595-600.

[11] Bengio Y,Schwenk H,Senécal Jean-Sébastien,et al. Neural probabilistic language models[J]. Studies in Fuzziness & Soft Computing,2006,16(3):137-186.

[12] Xu W,Rudnicky A. Can artificial neural network learn language models? [C]//Proc of international conference on statistical language processing. Beijing,China:[s. n.],2000.

[13] Mikolov T. Statistical language models based on neural networks[D]. Brno:Brno University of Technology,2012.

[14] 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.

[15] Chang C C,Lin Chih-Jen. LIBSVM:a library for support vector machines[J]. ACM Transactions on Intelligent Systems & Technology,2001,2(3):389-396.

基于Word2 fea模型的文本建模方法

作者： 卫华， 韩立新， 夏建华， WEI Hua， HAN Li-xin， XIA Jian-hua
作者单位： 河海大学 计算机与信息学院,江苏 南京,211100
刊名： 计算机技术与发展 ISTIC
英文刊名：
年，卷(期)： 2016(2)