基于词向量的Jaccard相似度算法

田 星 郑 瑾 张祖平
(中南大学信息科学与工程学院 长沙 410083)

摘 要 通过对传统Jaccard算法的研究和改进,提出了一种基于词向量的Jaccard句子相似度算法。传统的Jaccard算法以句子的字面量为特征,因而在语义层面的相似度计算方面受到了一定的限制。而随着深度学习的兴起,尤其是词向量的提出,词语在计算机中的表示有了突破性的进展。该算法首先通过训练将每个词语映射为语义层面的高维向量,然后计算各个词向量之间的相似度,高于阈值α的作为共现部分,最终计算句子的相似度。实验表明,相较于传统的Jaccard算法,该算法在短文本相似度计算的准确率上有较明显的提升。

关键词 词向量,Jaccard算法,句子相似度
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.07.032

Jaccard Text Similarity Algorithm Based on Word Embedding

TIAN Xing ZHENG Jin ZHANG Zu-ping
(School of Information Science and Engineering,Central South University,Changsha 410083,China)

Abstract Based on the research and improvement of the traditional Jaccard algorithm,this paper proposed a Jaccard sentence similarity algorithm based on word embedding. Traditional Jaccard algorithm is characterized by literals of the sentence,so it is restricted in the respect of semantic similarity calculation. While with the rapid development of deep learning,especially the proposal of word embedding,there is a breakthrough on the expression of words in computer. This algorithm firstly maps each word into a high-dimensional vector on semantic level by training,and then calculates the similarity between the respective word vector. The results which are higher than the threshold α are regarded as the intersection,and finally the sentence similarity is calculated. Experiments show that the algorithm significantly improves the accuracy of short text similarity calculation comparing with traditional Jaccard algorithm.

Keywords Word embedding,Jaccard algorithm,Text similarity

1 引言

在信息检索领域,句子相似度的计算是一个非常基础和重要的问题,在文本分类、文本摘要、机器翻译、自动问答等方面均有广泛的应用。因此,如何有效地衡量句子间的相似度,一直是相关学者研究的重点和难点,针对该问题也涌现出了大量的研究成果。文献[1]总结了3类计算句子相似度的算法,包括基于共现词的方法,基于VSM和TF-IDF的方法,基于句法、词序等语言学特征的方法,并对它们进行了对比。文献[2]提出了一种基于词汇概率分布的混合模型,并将其应用于关键字检索领域,取得了不错的实验效果。文献[3]提出了一种同时考虑语义和词序的句子相似度计算方法,该方法通过计算两个句子的语义相似度和词序相似度,然后加权得到两个句子的最终相似度。文献[4]借助语义词典WordNet进行相似度计算,其作者在计算相似度时运用概念层次树的思路,将在语义词典中两个词语的节点深度和路径长度作为衡量相似度的指标。在中文文本相似度的研究方面,近年来也取得了较大的进展。文献[5]提出了一种基于汉明距离的文本相似度计算方法,即通过把文本表示为0/1向量,并计算编辑距离,来得到文本的相似度。文献[6]提出了一种基于VSM的文本相似度计算方法,并以加权的方式对传统TF-IDF算法的权重计算部分进行了改进。文献[7]针对基于词频统计的传统相似度计算算法未考虑语义的缺陷,提出了一种基于加权语义网的文本相似度计算方法,通过构建特征词的加权语义网来计算文本相似度,取得了不错的效果。文献[8-9]提出了一种基于语法语义的短文本相似度计算方法,通过分析句子的组成成分和句法结构,使得计算结果更符合人们的主观判断。

但是,上述方法仍然有其各自的缺陷。基于共现词的相似度算法只以句子的字面量作为特征,未考虑词语蕴含的语义信息。比如,考虑这两个短句:1)这台电脑价格昂贵;2)这个计算机不便宜。如果使用基于共现词的相似度算法,对句子分词之后,这两个短句之间几乎无交集,那么按照定义,它们之间的相似度为0,然而事实上这两句话显然是等同的。

到稿日期:2017-03-14 返修日期:2017-05-01 本文受国家自然科学基金(61379109)资助。

田 星(1993-),男,硕士生,主要研究方向为机器学习、自然语言处理,E-mail:grubbyskyer@qq.com;郑 瑾(1970-),女,副教授,主要研究方向为软件工程、机器学习,E-mail:zhengjin@csu.edu.cn;张祖平(1966-),男,教授,主要研究方向为软件工程、数据挖掘、信息检索等。

基于 WordNet 等语义词典的相似度算法为语义层面的相似度计算提供了可能，在实际中也取得了不错的效果。但这类算法存在一个严重缺陷：过于依赖所使用的语义词典，词典的质量参差不齐，而最终的结果很大程度上由词典中词语的相似度决定，从而使得该类算法对一些未登录词的相似度计算产生了较大的误差。基于语言学特征的相似度算法对一些复杂句的相似度计算有较好的效果，但这类算法需要对句法、词序等特征进行实验分析，计算效率较低，而且对常见的简单短句的相似度比对并无明显的效果提升。

近年来，随着深度学习技术的逐渐兴起，尤其是词向量的提出，使得其在 NLP 领域大放异彩。词向量指通过训练将词语映射到一个高维的空间向量，通过计算这些向量之间的相似度来得到词语在语义层面的相似度；而且词向量的训练过程是无监督的，这就意味着它不会有语义词典之类的限制，实际效果更好。文献[10]提出了一种基于 n-gram 模型的三层神经网络模型来训练词向量，实验结果表明，相比深度调优的传统 n-gram 模型，该模型将效果提升了 10%～20%。文献[11]利用窗口对神经网络的输入层进行简化，从而降低了传统三层模型的计算复杂度。文献[12]提出了 CBOW 和 Skip-gram 两种模型训练词向量，减少了训练过程中所需要的参数，从而避免了过拟合，在保证词向量质量的同时，提升了训练效率。文献[13]提出了一种用全局信息辅助局部信息进行训练的思路，同时通过在原有模型上加入 k-means 算法，对一词多义的表示效果进行了有效改进。

考虑到词向量是每个词语在语义层面的高维向量表示，利用这些向量进行相似度计算，将结果高于阈值 α 的部分作为两个短句的交集，可以将传统 Jaccard 算法在语义层面进行扩展。文献[14]使用词向量带来的语义信息对 ROUGE 进行改进，在文本摘要的评测方面取得了不错的成果。文献[15]利用词向量提出了一种 WMD(Word Movers's Distance)模型，该模型在文本分类方面取得了不错的效果。在文本相似度计算方面，使用语义词典可以改进传统的词频统计方法取得的效果，但又受词典规模和语义的主客观限制；而使用词向量既可以保留和扩展语义信息，又能突破词典的限制。正是基于这样的思路，本文提出了一种基于词向量的 Jaccard 相似度算法。

2　研究内容

2.1　基于共现词的 Jaccard 算法

直观上来说，两个句子的相同部分越大，共现的词汇数目越多，它们的相似度应该越高，传统的 Jaccard 算法正是基于这样的思想。而共现词相对于所有词汇所占的比例可以从数值上反映这两个句子的相似度，用公式表示为：

$$Sim(S,T)=\frac{\| Inter(S,T) \|}{\| Union(S,T) \|} \tag{1}$$

其中，$Inter(S,T)$ 表示句子 S 和句子 T 的词汇交集，$Union(S,T)$ 表示句子 S 和句子 T 的并集。

但是，这样得到的 Jaccard 算法未考虑交集中各个词汇的重要程度，因此有学者利用 TF-IDF 对其进行了改进：

$$Sim(S,T)=\frac{\sum_{w\in Inter(S,T)} IDF(w)}{\| Union(S,T) \|} \tag{2}$$

在共现部分的计算方面，为交集中的每一个词汇 w 加入了权重(如逆文档频率 IDF)，用于体现词汇的重要程度及对整句相似度计算的影响。显然，越重要的词汇，其 IDF 值越大，对相似度的影响也越大。改进的算法在特定的数据集上有一定的效果提升，但是仍然未突破 Jaccard 算法在语义层面的限制。

2.2　word2vec 简介

在深度学习研究兴起的大背景下，Google 公司实现了开源的词向量训练工具 word2vec。该工具主要实现了 Mikolov 论文中的 CBOW 和 Skip-gram 两种模型，同时加入了k-means 算法来解决一词多义的问题。

以 Skip-gram 模型为例(见图 1)，该模型是一个包含输入层、隐藏层和输出层的三层神经网络模型。在给定文本集的训练过程中，每个词向量都需要最大化其邻近词汇的 log 概率，对于给定的词汇序列：$w_1, w_2, \cdots, w_T$，Skip-gram 模型的训练目标就是使式(2)的值最大：

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j\in nb(t)}\log p(w_j \mid w_t) \tag{3}$$

其中，$nb(t)$是文本集中词汇 w_t 的邻集，而 $p(w_j \mid w_t)$是词向量 V_{w_j} 和 V_{w_t} 的层次 softmax 回归值，更详细的解释请参考文献[12]。

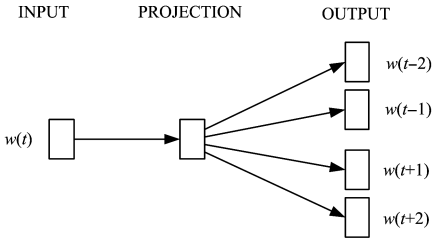

图 1　Skip-gram 模型

Fig. 1　Skip-gram model

由于模型简单且训练参数较少，因此其训练效率很高，可以在单机上每小时完成数亿词向量的训练。word2vec 以任意大小的文本集作为输入，通过无监督的训练得到一份包含所有词向量的二进制文件。利用该高维向量作为输入，可以把深度学习应用到 NLP 的诸多领域中，而词向量本身在语义层面蕴含的信息也可以直接作为词汇相似度计算的依据。比如，Mikolov 在其 2013 年的论文中发现，$vec(Berlin)-vec(Germany)+vec(France)\approx vec(Paris)$，这表明词向量之间的相似度可以表示词汇在语义层面的相似度。

2.3　基于词向量的 Jaccard 算法

无论是传统的 Jaccard 算法还是利用 TF-IDF 改进之后的算法，显然都未涉及语义层面的相似度对比，遇到同义词、近义词时将显得无能为力。而词向量在语义层面有强大的表示能力，因此利用词向量改进原有的 Jaccard 算法，使其能够在语义层面进行相似度计算。以句子 S 和句子 T 为例，该算法的思路如下：

Step1　利用 word2vec 进行词向量的训练，得到每个词汇的词向量。

Step2 使用分词工具将句子 S 和 T 分词(英文文本不需要),并使用 2-gram 构成词组或短语,此时句子 S 和 T 分别表示为:$S(s_1 s_2 \cdots s_S)$ 和 $T(t_1 t_2 \cdots t_T)$。

Step3 计算 S 中每一个词 s 对应的词向量 V_s 在句子 T 中所能得到的最大匹配值,将其作为该词汇在句子 S 和 T 中的共现度,公式为:

$$Sim(s,T)=\max_{t_j\in T}\cos(V_s,V_{t_j}) \tag{4}$$

Step4 将得到的每个词向量的最大匹配值与阈值 α 进行比较,高于阈值 α 的判定为两句的交集,低于 α 的记为不相似部分,以 $1-Sim(s,T)$ 计算其不相似度,然后将其得到的最大匹配 $Sim(s,T)$ 置为 0:

$$Sim'(s,T)=\begin{cases}Sim(s,T), & Sim(s,T)>\alpha\\ 1-Sim(s,T), & Sim(s,T)\leqslant\alpha\end{cases} \tag{5}$$

$$Sim(s,T)=0(Sim(s,T)\leqslant\alpha) \tag{6}$$

Step5 交换 S 和 T,重复 Step3 和 Step4。

Step6 最终的相似度计算公式为:

$$Sim(S,T)=\frac{\sum_{i=1}^{S}Sim(s_i,T)+\sum_{j=1}^{T}Sim(t_j,S)}{\sum_{i=1}^{S}Sim'(s_i,T)+\sum_{j=1}^{T}Sim'(t_j,S)} \tag{7}$$

其中,分子部分是句子 S 和句子 T 词向量的共现值,即句子 S 和句子 T 在语义上的交集。分母部分包含了分子,而且包括了所有匹配值低于阈值 α 的不相似度,因此:

$$0\leqslant Sim(S,T)\leqslant 1 \tag{8}$$

Step7 将得到的句子 S 和句子 T 的相似度 $Sim(S,T)$ 与阈值 β 进行比较,若大于 β 则判定为相似,否则视为不相似。

3 实验分析

3.1 数据集及评测指标

在 3 个数据集上进行了传统 Jaccard 算法、基于 TF-IDF 改进的 Jaccard 算法以及基于词向量的 Jaccard 算法的对比实验。其中数据集 I 是取自 MSR 的语料,共包含 5801 对英文句子对,其中有 3900 对句子被人工标注为相似句,记为 DATASET-I;数据集 II 是通过不同的翻译工具生成的 5000 组相似句子对,再随机生成 5000 组负类,共计 10000 组句子对,记为 DATASET-II;数据集Ⅲ是人工组合的 100 组中文相似句子对以及 100 组负类,共计 200 组句子对,记为 DATASET-III。其中,DATASET-I 和 DATASET-II 为英文数据集,DATASET-III 为中文数据集。

在词向量的训练方面,英文词向量的训练采取了 3 种不同规模的训练语料,分别是:取自 word2vec 的自带语料,共计 97M,得到的词向量记为 WE-S;取自 Twitter 的评论语料,共计 470M,得到的词向量记为 WE-M;取自英文维基百科的语料,共计 1.26G,得到的词向量记为 WE-B。中文词向量的训练语料取自搜狗新闻语料,共计 320M,分词工具采用的是 Jieba 分词,得到的词向量记为 WE-C。

在实验结果的评测方面,采取信息检索领域常用的准确率(Precision)、召回率(Recall)以及 F 值(F-measure)作为算法的评价指标。其中,F 值的计算公式为:

$$F\text{-}measure=\frac{2\cdot P\cdot R}{P+R} \tag{9}$$

3.2 实验结果及分析

表 1 和表 2 分别列出了 3 种算法在 DATASET-I 和 DATASET-II 上的评测结果,其中在基于词向量的 Jaccard 算法中,判定词向量相似的阈值 α 分别设为 0.4,0.4,0.5,判定句子相似的阈值 β 设为 0.8。

表 1 DATASET-I 的评测结果

Table 1 Evaluation results of DATASET-I

算法	准确率	召回率	F 值
Jaccard	0.702	0.583	0.637
Jac-IDF	0.679	0.598	0.636
Jac-WE-S	0.610	0.783	0.686
Jac-WE-M	0.691	**0.880**	0.774
Jac-WE-B	**0.779**	0.862	**0.818**

表 2 DATASET-II 的评测结果

Table 2 Evaluation results of DATASET-II

算法	准确率	召回率	F 值
Jaccard	0.934	0.912	0.924
Jac-IDF	0.931	0.924	0.927
Jac-WE-S	0.837	0.858	0.847
Jac-WE-M	0.892	0.938	0.914
Jac-WE-B	**0.945**	**0.943**	**0.944**

在 DATASET-I 上的实验结果表明,基于词向量的 Jaccard 算法能有效改进传统的 Jaccard 算法,特别是在召回率和 F 值方面有较明显的提升。DATASET-II 上的各项评测指标均高于 DATASET-I,这是由 DATASET-II 的数据特点决定的。由于 DATASET-II 是由不同的翻译工具对同一文本的翻译结果组成的,使得出现共现词的概率更高,因此原生的 Jaccard 算法已经取得了不错的效果。但是从实验结果来看,在采用了质量较高的词向量之后,基于词向量的 Jaccard 算法仍然有着不错的效果。

利用 3 个不同大小的训练语料得到的词向量分别进行实验,通过对其评测结果的纵向对比可以发现,WE-B和WE-M 的实验结果明显优于 WE-S,这表明在训练语料更全面、更充足的情况下,词向量对词汇在语义层面的表示更为准确,这与 Mikolov 在其论文中的论述一致。

表 3 列出了 3 种算法在中文数据集 DATASET-III 上的评测结果。由于中文文本相似度计算受到分词质量的影响,因此在基于词向量的 Jaccard 算法中,将判定词向量相似的阈值 α 设为 0.3,将判定句子相似的阈值 β 设为 0.6。

表 3 DATASET-III 的评测结果

Table 3 Evaluation results of DATASET-III

算法	准确率	召回率	F 值
Jaccard	0.55	0.52	0.53
Jac-IDF	0.54	0.55	0.54
Jac-WE-C	**0.61**	**0.68**	**0.64**

可以看出,在中文文本相似度计算方面,基于词向量的 Jaccard 算法虽然不如在英文文本对比中的效果好,但在这 3 种算法中仍然是最优的。这是因为中文的句子结构更加复杂,句法、词序变化更大,仅仅依靠语义来进行相似度计算而得到的效果可能并不理想。

结束语 本文提出了一种基于词向量的 Jaccard 相似度算法,该算法首先通过对外部语料的训练得到词向量,然后对

两个句子中的词向量进行最大匹配，相似度高于共现阈值 α 的作为这两个短句语义层面的交集，最终计算出句子的相似度。本文在中英文数据集上分别进行了实验，并将该算法与传统的 Jaccard 算法进行了对比，从而证明了该算法的有效性。但是，该算法在中文文本的相似度计算方面的效果仍不尽如人意。如何得到一份高质量的中文词向量，以及如何将词向量与句法、词序等语言学特征结合起来提升中文文本相似度计算的效果，将是我们下一步工作的重点。

参 考 文 献

[1] ACHANANUPARP P, HU X, SHEN X. The Evaluation of Sentence Similarity Measures[C]// International Conference on Data Warehousing and Knowledge Discovery. 2008: 305-316.

[2] METZLER D, DUMAIS S, MEEK C. Similarity Measures for Short Segments of Text[C]// Advances in Information Retrieval, European Conference on Ir Research (ECIR 2007). Rome, Italy, 2007: 16-27.

[3] LI Y, MCLEAN D, BANDAR Z A, et al. Sentence Similarity Based on Semantic Nets and Corpus Statistics[J]. IEEE Transactions on Knowledge & Data Engineering, 2006, 18(8): 1138-1150.

[4] AGIRRE E, ALFONSECA E, LACALLE O L D. Approximating hierarchy-based similarity for WordNet nominal synsets using topic signatures[C]// Proceedings of Gwc. 2004.

[5] ZHANG H J, WANG G S, ZHONG Y X. Text Similarity Computing Based on Hamming Distance[J]. Computer Engineering and Applications, 2001, 37(19): 21-22. (in Chinese)
张焕炯，王国胜，钟义信. 基于汉明距离的文本相似度计算[J]. 计算机工程与应用，2001，37(19)：21-22.

[6] GUO Q L, LI Y M, TANG Q. Similarity computing of documents based on VSM[J]. Application Research of Computers, 2008, 25(11): 3256-3258. (in Chinese)
郭庆琳，李艳梅，唐琦. 基于 VSM 的文本相似度计算的研究[J]. 计算机应用研究，2008，25(11)：3256-3258.

[7] LIAO K J, YANG B B. Similarity Computing of Documents Based on Weighted Semantic Network[J]. Journal of Intelligence, 2012, 31(7): 182-186. (in Chinese)
廖开际，杨彬彬. 基于加权语义网的文本相似度计算的研究[J]. 情报杂志，2012，31(7)：182-186.

[8] LIAO Z F, QIU L X, XIE Y S, et al. A Frequency Enhanced Algorithm of Sentence Semantic Similarity[J]. Journal of Hunan University (Natural Sciences), 2013, 40(2): 82-88. (in Chinese)
廖志芳，邱丽霞，谢岳山，等. 一种频率增强的语句语义相似度计算[J]. 湖南大学学报(自然科学版)，2013，40(2)：82-88.

[9] LIAO Z F, ZHOU G E, LI J F, et al. A Chinese Short Text Similarity Algorithm Based on Semantic and Syntax[J]. Journal of Hunan University (Natural Sciences), 2016, 43(2): 135-140. (in Chinese)
廖志芳，周国恩，李俊锋，等. 中文短文本语法语义相似度算法[J]. 湖南大学学报(自然科学版)，2016，43(2)：135-140.

[10] BENGIO Y, SCHWENK H, SENÈCAL J S, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[11] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural Language Processing (almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.

[12] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[13] HUANG E H, SOCHER R, MANNING C D, et al. Improving word representations via global context and multiple word prototypes[C]// Meeting of the Association for Computational Linguistics: Long Papers. 2012: 873-882.

[14] NG J P, ABRECHT V. Better Summarization Evaluation with Word Embeddings for ROUGE[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.

[15] KUSNER M J, SUN Y, KOLKIN N I, et al. From Word Embeddings to Document Distances[C]// International Conference on Mechine Learning. 2015: 957-966.

(上接第 157 页)

[7] LI F, WU C M. Research on Prevention Fluctuation Control method of Network Intrusion Based on Energy Management [J]. Computer Simulation, 2013, 30(12): 45-48. (in Chinese)
黎峰，吴春明. 基于能量管理的网络入侵防波动控制方法研究[J]. 计算机仿真，2013，30(12)：45-48.

[8] DENG Z H, CAO L B, JIANG Y Z, et al. Minimax probability TSK fuzzy system classifier: A more transparent and highly interpretable classification model[J]. IEEE Transactions on Fuzzy Systems, 2015, 23(4): 813-826.

[9] HESS R A. Aircraft and rotorcraft system identification-engineering methods with flight test examples[J]. Journal of Guidance, Control, and Dynamics, 2013, 36(4): 1249-1250.

[10] ZHANG H B, HE Q B, KONG F R. Stochastic resonance in an underdamped system with pinning potential for weak signal detection[J]. Sensors, 2015, 15(9): 21169-21195.

[11] WANG H X, WANG S Y, WANG X, et al. Analysis of LFM signals and improvement of IFM system[J]. Acta Armamentarii, 2014, 35(8): 1193-1199. (in Chinese)
王洪迅，王士岩，王星，等. 瞬时测频系统的线性调频信号分析及改进[J]. 兵工学报，2014，35(8)：1193-1199.

[12] MAHBOUBI H, MOEZZI K, AGHDAM A G, et al. Distributed deployment algorithms for improved coverage in a network of wireless mobile sensors[J]. IEEE Transactions on Industrial Informatics, 2014, 10(1): 163-174.

[13] MAHBOUBI H. Distributed deployment algorithms for efficient coverage in a network of mobile sensors with nonidentical sensing Capabilities[J]. IEEE Transactions on Vehicular Technology, 2014, 63(8): 3998-4016.

[14] DAI W. Application of Intrusion Detection Technology in Network Security[J]. Journal of Chongqing University of Technology (Natural Science), 2018, 32(4): 156-160, 185. (in Chinese)
代威. 入侵检测技术在网络安全中的应用[J]. 重庆理工大学学报(自然科学)，2018，32(4)：156-160，185.