科研信息化技术与应用
2015, 6(4): 54–59

· 技术 / TECHNOLOGY ·

基于 word2vec 的关键词提取算法

李跃鹏[1,3]，金翠[2]，及俊川[1]

1. 中国科学院计算机网络信息中心, 北京 100190
2. 北京科技大学, 北京 100083
3. 中国科学院大学，北京 100049

摘　要： 随着近些年深度学习的兴起，词语在计算机中的表示有了重大突破；而长期以来关键词提取算法均以词语作为特征进行计算，效果并不理想。因此，本文提出了一种基于深度学习工具 word2vec 的关键词提取算法。该算法首先使用 word2vec 将所有词语映射到一个更抽象的词向量空间中；然后基于词向量计算词语之间的相似度，最终通过词语聚类得到文章关键词。实验表明该算法对于篇幅长文章的关键词提取的准确率要明显高于其他算法。

关键词： word2vec；关键词提取；词向量

doi:10.11871/j.issn.1674-9480.2015.04.007

A Keyword Extraction Algorithm Based on Word2vec

Li Yuepeng[1,3], Jin Cui[2], Ji Junchuan[1]

1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China
2. University of Science and Technology Beijing，Beijing 100083, China
3. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: With the rapid development of deep learning, a major breakthrough has been made to the word representation of computers, while for a long time the keyword extraction algorithms is based on the feature of words, and it is not very ideal. In this paper, we present a keyword extraction algorithm based on word2vec, which is a well known tool for deep learning. Firstly, this algorithm projects all the words into a more abstract word vector space, then based on the word vectors, it calculates the similarity between words to cluster all the words in the target article, and the center of the cluster can be selected as the keyword. According the result of the experiment, this algorithm is better than other algorithms for long articles.

Keywords: word2vec; keyword extraction; word vector

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

引言

随着计算机与互联网的飞速发展，人们积累的文档数据越来越多。然而面对如此庞大的数据，如何从中挖掘出有用信息，如何快速检索数据一直是人们面临的一个重要问题。其中关键词提取在文档数据的利用方面发挥着重大作用，比如根据关键词可以进行文章分类聚类、建立索引、主题搜索、主题爬虫以及推荐系统等等。除了以上应用之外，关键词提取的另一个常见的用处是新闻或博客。通过对新闻或博客进行关键词提取，读者可以在很短时间内了解文章的内容，从而决定是否深入阅读。比如国内新浪、网易等各大网站中的新闻都给出了新闻的关键词；如 CSDN 这样的技术论坛不仅给出了新闻的关键词，所有博客与随笔都进行了关键词提取。

1 背景

关键词提取属于信息抽取的一个子领域，其目的是从文章中挑选出代表性的词语。该问题研究最早始于 20 世纪 70 年代前后，在近 30 多年的发展过程中，其应用从科技论文的数字化推广到了文本处理的各个方面。然而其算法的发展过程却主要包含两个方面：(1) 如何在计算机上来表示一个词语（也即是特征提取）；(2) 如何根据这些词的特征选择出关键词。

早期的关键词提取算法使用词频、词性、词在文章中位置等属性来表示词语，然后根据某个规则计算出每个词的得分，选择得分高的词作为关键词。比如文献 [1] 中提出了基于 TF-DF 特征计算得分的关键词提取算法；针对中文关键词的长度问题，文献 [2] 提出了一种基于 PAT-tree 关键词提取算法，该算法可以计算不同长度关键词的得分，从而可以将一些基本的词语组合成短语；此外以文献 [3] 为代表的基于共现网络的关键词提取算法 TextRank，通过词语共现窗口构建共现网络，并根据共现网络计算词语的得分。受共现网络的启发，以文献 [4] 为代表的方法将文章中的词语组成语义网络结构并计算每个词的得分；而以文献 [5] 为代表的算法通过词典将文章中的词组织成多条词汇链，根据词汇链与词语的得分来挑选关键词。

除了基于得分的关键词提取方法之外，还有一类是基于机器学习的关键词提取方法。其中包括如 SVM[6]、朴素贝叶斯 [7] 等有监督学习方法；以及 K-means、层次聚类 [8] 等无监督学习方法。相对于基于得分的方法而言，这些方法虽然利用了数据集中的信息，但是并没有改变词的表示方式。其中词的特征仍然是词的词性、词频等，这种表示方式忽略了词语之间的语义联系，比如同义词、反义词等。因此不论是聚类还是分类过程中，词语的特征并不能给出关于词语语义充分信息，所以这些关键词提取算法的准确率并不理想。

不仅对于关键词提取算法，几乎所有机器学习算法都在特征提取方面遇到了瓶颈。其主要原因在于人工选择的原始特征含有的用于如分类聚类这些机器学习任务的信息量不够抽象，也不够充分。为此人们提出了各种用于浓缩原始特征的降维方法 [9]，比如早期提出的通用的 PCA 降维方法、Fisher 判别方法等；以及近期在图像处理领域提出的 DSNPE[10] (Discriminant Sparse Neighborhood Preserving Embedding)、GSM-PAF[11] (Group Sparse Multiview Patch Alignment Framework) 等降维方法。这些方法都是旨在将原始特征通过某种方式投影到一个既能够保持用于机器学习任务的信息，同时维数又低的空间中，从而减小算法复杂度，减少原始特征的噪声。

近几年兴起的深度学习正是特征提取与数据降维的一种方式，在自然语言处理领域，其中一个重要的突破就是词语的表示方式。它在训练语言模型的过程中将词映射到一个更抽象的向量空间中。在大数据的环境下，可以认为该向量空间中两点之间的距离就是对应两个词语的相似程度。这便在一定程度上解决了同义词问题，以及如何用计算机表示词语的问题。更重要的是，当词语被映射到某个向量空间中之后，我们就可以在该空间中应用各种机器学习方法。其中在自然语言处理领域具有代表性的深度学习工具就是 google 的 word2vec。

正是基于这个背景，本文提出了一种应用 word2vec 进行关键词提取的算法。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2 研究内容

2.1 word2vec 简介

自 2006 年 Hinton 等人提出深度学习的概念之后，该方法就在机器学习领域得到了广泛的关注。该方法基于人工神经网络，通过多层感知机将初始的底层特征组合为更抽象的高层特征，并将高层特征用于普通的机器学习方法以得到更好的效果。由于人工神经网络的通用性，可以很自然的对各种特征进行整合，因此深度学习在各个领域都得到了广泛应用；比如在图像处理领域提出的卷积神经网络（Convolutional Neural Network, CNN）、在语音识别领域提出的深度神经网络（Deep Neural Network, DNN）和受限波兹曼机 (Restricted Boltzmann Machine, RBM) 以及在自然语音领域提出的词向量模型与文档向量模型等。

正是在这个背景下，谷歌于 2012 年实现了开源语言建模工具 word2vec，并在自然语言处理领域得到了广泛关注。该工具实现了连续 bag-of-word 模型 [12]，以及计算词向量的 skip-gram 结构 [13]。它以文本集为输入，通过训练生成每个词对应的词向量。这些词向量可以作为词的特征应用到其他自然语言处理问题中去。比如可以根据词向量计算两个词的相似程度；使用词向量可以避免词语表示的“维灾难”现象。

word2vec 是对神经网络概率语言模型的实现，它将概率模型与人工神经网络相结合，对 n-gram 模型进行了改进。n-gram 模型的目的是要得到给定前 n-1 个词 $w_1,.....,w_{n-1}$ 取值的条件下，第 n 个词 w_n 取值的条件概率分布。也即是说，如果词典中共有 D 个词的话，那么 n-gram 模型的学习内容就是要从数据中学习出 A_D^{n-1} 个上下文 $w_1,.....,w_{n-1}$ 下的条件概率分布 $P(w_n \mid w_1,.....,w_{n-1})$。word2vec 的目标也是如此，不过 word2vec 假设 $P(w_n \mid w_1,.....,w_{n-1}) = f(w_1,.....,w_{n-1},w_n)$，其中 $f(w_1,.....,w_{n-1},w_n)$ 是用一个神经网络表示的函数。

从图 1 与图 2 可以看出，对于 n-gram 模型，在已知上下文 $w_1,.....,w_{n-1}$ 取值情况求 w_n 取值概率的方法是从一个巨大的概率分布表中进行查询；而对于 continue bag of word 模型，这个概率的值只需要经过函数 $f(w_1,.....,w_{n-1},w_n)$ 计算即可。该函数首先将词映射为一个词向量，然后将词向量相加，最后计算出 $f(w_1,.....,w_{n-1},w_n)$ 的值。

我们知道，使用 logistic 回归进行分类要比使用朴素贝叶斯模型分类对数据更不敏感。同样，相对于 n-gram 模型，word2vec 在表示模型的过程中使用了更少的参数，这就在一定程度上避免了过拟合现象的发生。最关键的是经过训练之后每个词语得到了相应的词向量，可以认为，该向量是词在某个语义空间中的投影。

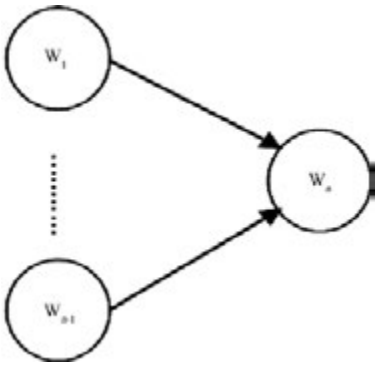

图 1 n-gram 模型
Fig.1 n-gram model

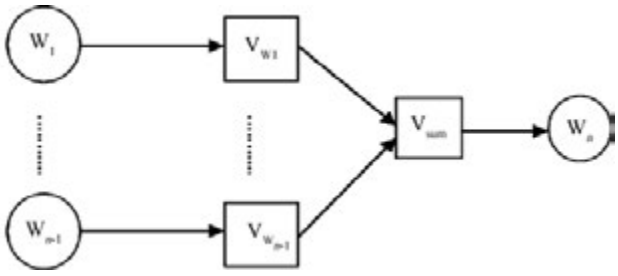

图 2 Continue bag of word 模型
Fig.2 Continue bag of word model

2.2 关键词提取过程

本文关键词提取算法基于如下假设：（1）文章中的词语是围绕关键词展开的，关键词即是文章的中心思想；（2）文章中大部分词与关键词在语义上相似，少部分词语与文章中心思想不相关。

因此本文关键词提取的主要思路是通过 k-means 对文章中的词进行聚类，选择聚类中心作为文章的一个关键词。其中在 k-means 聚类过程中的词间相似度用 word2vec 生成的向量计算。词向量的训练使用了

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

搜狗语料库中的新闻语料，该语料共有九个文件夹，每个文件夹代表一类新闻，整个关键词提取过程共分为 5 步。

第一步：通过 ICAS 分词工具将新闻语料进行分词，并去除停用词列表中的停用词。

第二步：去除在各类中都出现，且出现次数都大于 100 的词。这是因为如果一个词不能为分类提供信息，实际上它也不会为聚类提供信息。

第三步：将去噪之后的所有词串联成一个文件 F，词语之间用空格分割；运行 word2vec 对文件 F 进行训练、得到词向量。

第四步：对于新的文本，按照第一步到第三步对文本进行预处理。然后对处理好的词进行 kmeans 聚类，在聚类过程中使用词向量计算两个词之间的距离，最后选择每个类别中距离聚类中心最近的词作为关键词。虽然经过停用词过滤，但考虑到仍可能出现少数几个与文章中心思想不相关词语聚成一类的情况。因此本文舍去少于 10 个词的簇，并从词个数最多的簇中选择其他词作为补充。

3 实验结果与分析

本文使用的语料库共有 90MB，在 4G 内存电脑上历时 40 分钟训练产生了一个 180MB 的词向量模型文件。

此外基于同样的语料库，本文实现了目前主流的基于 TF-IDF 的关键词提取算法，以及基于 textrank 的关键词提取算法，并对这 3 种算法的关键词提取结果进行了分析比较。

目前关键词提取算法的召回率计算大致有两方面不同：（1）文章关键词来源，其中主要分为实验人员标注与寻找现存的已标注语料；（2）召回率是否考虑语义，其中一类方法认为提取到的关键词如果与标注关键词语义相似的话则准确率会增加，而另一类方法则认为只有提取到的关键词与标注的关键词完全匹配才是正确的。

本文认为，现存标注语料陈旧，未出自权威机构；因此，通过实验人员进行关键词标注可以保证实验的正确性。此外，关键词的语义只能作为不同算法分析比较过程的一个方面，不能引入到召回率的计算公式中。因为词语的语义本身就尚未定论，所以如果计算召回率过程中考虑语义的话会引入更多的误差。

因此，本文随机选择了 20 篇搜狐新闻，共涉及 10 个不同的版块。然后分别为每篇新闻人工标注了 3、5、7、10 个关键词。最后分别使用 3 种算法进行关键词提取，按照公式：

$$召回率 = \frac{与人工标注相同的关键词个数}{总的关键词个数}$$

得到 3 种算法在提取不同个数关键词时的召回率如表 1 所示。

表 1　三种算法召回率的比较

Table 1　Comparison of the recall rate of 3 algorithms

个数 / 算法	3	5	7	10
TF-IDF	31.2%	26.5%	25.7%	23.4%
Textrank	34.1%	33.4%	32.9%	32.4%
Word2vec	27.4%	32.0%	35.1%	37.8%

根据上表可以看出，在关键词个数较少的情况下，基于得分的关键词提取算法的准确率要高于基于词向量聚类的关键词提取方法。而随着关键词个数的增加，基于得分的方法的准确率开始逐渐下降，而基于词向量聚类方法的准确率却逐渐上升。这种现象说明，随着关键词个数的增加，得分已经不能作为判断一个词对于文章的重要性的依据。针对这种现象，本文对关键词提取过程进行了深入分析，并得出如下结论：（1）对于篇幅较短，内容简单的介绍类文章，基于得分的方法准确率要比基于词向量聚类的方法高。（2）而对于篇幅较长，内容丰富的文章，基于词向量的方法要比基于得分的方法准确率高。（3）在一定范围之内，关键词个数越少，基于词向量方法的准确率就越低；相反，关键词个数越少，基于得分方法的准确率就越高。（4）虽然基于得分与基于词向量的方法的召回率都很低，但是基于得分方法中除了完全正确的关键词外，其他的关键词与人工标注的关键词几乎不相关；而基于词向量的方法中，除了正

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

确的关键词外，其他关键词与人工标注的关键词在语义上大部分都是相似的。

出现上述现象的原因来自三个方面。首先，对于篇幅短、内容简单的介绍类文章（比如对某个上市公司的介绍），其关键词会频繁的出现在所有段落中，并且文章中的句子大都是同样的语法结构。这就意味着以词频与词在文章中的位置作为特征能够给出一个词是否为关键词的信息。然而，对于篇幅长且内容丰富的文章（比如对某个上市公司在某一天股票行情描述），其关键词主要是在语义上能够关联文章各个段落的词语；基于词向量聚类的方法能够将所有词语聚在不同的簇上，因此关键词提取的结果要更好。最后，当关键词个数少时，两个或多个不同的簇揉合在一起，会改变簇的中心，从而使得距离聚类中心最近的点并不是该簇中作为关键词的最佳点；因此，当关键词个数较少时，基于得分的方法要比基于词向量聚类的方法准确率高。

4 结语与展望

本文分析了 word2vec 的工作原理，以及词向量的优点。并基于文章的词向量进行了关键词提取。实验表明，基于词向量的关键词提取方法可以充分利用语料的信息，通过语义解决当关键词个数增加时词频特征无法提供词语重要程度信息的问题，对篇幅长、内容丰富文章进行关键词提取的准确率与实用性要明显优于其他方法。

基于词向量的关键词提取算法对文本搜索有重要意义。传统的搜索引擎首先对文本建立索引，然后根据查询词与索引词的匹配程度返回搜索结果。这种方式有两个弊端：首先，对全文索引是一个工作量非常大的问题，不仅在存储上会有压力，查找也会很困难；其次，基于词匹配的方法无法解决同义词的问题。如果能提取出文章的关键词，使用关键词对文章建立索引，并根据查询词的词向量与关键词的词向量进行比较来返回结果的话，会大大减少索引的存储量以及查询时间，并且可以解决搜索过程中的同义词问题。

此外，基于词向量的关键词提取方法的另一个优势在于不需要人工确定词的特征，它可以利用不断增长的数据提升模型的准确性。因此，本文提出的算法还可以在如下方面进行改进：（1）使用更大的数据集训练更精确的词向量；（2）使用更全面的停用词列表去除无关词语对模型的干扰；（3）使用层次聚类自动确定关键词的个数；（4）对 word2vec 的语言模型（也即是图 2 中的图结构）进行改进，增加神经网络的层次，提高词向量的语义抽象层次；（5）使用 hadoop 分布式存储与 mapreduce 分布式计算技术提高算法的运行速度。

综上所述，应用词向量进行关键词提取的方法的主要优势在于通过深度学习将词语特征投影在一个更抽象的空间中，并在该空间中进行关键词的提取。这种方法能够克服选择关键词过程中原始特征噪声高的问题。随着大数据环境的不断完善，如基于词向量关键词提取算法一样，通过深度学习提取特征的方式会成为人工特征提取的另一种选择。

参考文献

[1] Yih W, Goodman J, Carvalho V R. Finding advertising keywords on web pages[C]//Proceedings of the 15th international conference on World Wide Web. ACM, 2006: 213-222.

[2] Chien L F. PAT-tree-based keyword extraction for Chinese information retrieval[C]//ACM SIGIR Forum. ACM, 1997, 31(SI): 50-58.

[3] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]//Proceedings of EMNLP. 2004, 4(4): 275

[4] 王立霞, 淮晓永. 基于语义的中文文本关键词提取算法[J]. 计算机工程, 2012, 38(01): 1-4.

[5] 张颖颖,谢强,丁秋林. 基于同义词链的中文关键词提取算法[J]. 计算机工程,2010,36(19):93-95.

[6] Zhang K, Xu H, Tang J, et al. Keyword extraction using support vector machine[M]//Advances in Web-Age Information Management. Springer Berlin Heidelberg, 2006: 85-96.

[7] Uzun Y. Keyword Extraction Using Naïve Bayes[C]// Bilkent University, Department of Computer Science,

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Turkey www. cs. bilkent. edu. tr/~ guvenir/courses/CS550/Workshop/Yasin_Uzun. pdf. 2005.

[8] 高学东,吴玲玉. 基于高维聚类技术的中文关键词提取算法[J]. 中国管理信息化,2011,09:23-27.

[9] Gui, Jie, et al. “How to estimate the regularization parameter for spectral regression discriminant analysis and its kernel version?.” Circuits and Systems for Video Technology, IEEE Transactions on 24.2 (2014): 211-223.

[10] Gui, Jie, et al. “Discriminant sparse neighborhood preserving embedding for face recognition.” Pattern Recognition 45.8 (2012): 2884-2893.

[11] Gui, Jie, et al. “Group sparse multiview patch alignment framework with view consistency for image classification.” Image Processing, IEEE Transactions on 23.7 (2014): 3126-3137.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.

收稿日期：2015 年 6 月 6 日

李跃鹏：中国科学院计算机网络信息中心，中国科学院大学，在读硕士研究生，主要研究方向为机器学习。
E-mail: *908065729@qq.com*

金　翠：北京科技大学，在读硕士研究生，主要研究方向为机器学习。
E-mail: *cuicuijin@163.com*

及俊川：中国科学院计算机网络信息中心，正研级高级工程师，主要研究方向为科研信息化。
E-mail: *jcji@cashq.ac.cn*

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net