基于BP神经网络的文档聚类研究

Research on Document Clustering Based on BP Neural Net

田 萱 刘希玉 孟 强

(山东师范大学信息管理学院 济南250014)

Abstract Document clustering has been used in a number of different areas of text mining and information retrieval. This paper first introduces the presentation of document clustering and it's ground, VSM(Vector Space Mode). On the other hand, comparing with the VSM, we present a new model to calculate the word weight in a document based on BP neural net. On the ground of it, two document clustering algorithms are described aiming at scientific literature on the Web. One is to get document sets relevant to user's query, and the other is to extract more personalized interesting items.

Keywords Document clustering, Vector Space Model, BP neural net, Word weight, Scientific literature

1. 引言

近年来,随着互联网的迅速发展,基于Web的数据挖掘技术受到越来越多的关注,经常用在文本挖掘和信息检索等多个领域的聚类(Clustering)技术也成为人们研究的热点。对一组实际或抽象的元素进行处理,把相似的元素归为同类的过程称之为聚类[1]。对文本信息,如科技文献、Web文档等的聚类,称之为文档聚类(Document Clustering)。最初,文档聚类常用于提高信息检索系统的查准率和查全率(recall),或用来寻找与一篇文档最为相似的文档[2]。现在,人们利用文档聚类来获得一组满足用户要求的文档集合并按用户需求对其进行排序。另外在Internet上,文本聚类也可用来自动产生文档的层次聚类,从而实现对Web文档的分类。

聚类算法多种多样,常分为以下几类:构建和评估各种分区的分割算法(Partitioning Methods);创建层次分解的层次算法(Hierarchical Methods);基于连通和密度函数的基于密度算法(Density-Based Methods);基于多层粒度的网格算法(Grid-Based Methods);构建聚类假设模型的模型算法(Model-Based Clustering Methods)。其中,文档聚类中最为常用的算法为分割算法中的K-means算法和层次算法中的凝聚层次算法(Agglomerative hierarchical clustering)。相比较而言,凝聚层次算法的聚类效果比K-means算法效果要好,但其时间复杂度为$O(n^2)$(n为聚类元素的个数),效率不如K-means算法。然而,不论哪种算法,常使用向量空间模型法[7](Vector Space Model)来表示文档,而向量空间模型法具有庞大的计算复杂度。

本文针对网上科技文献文档所具有的规范格式的特点,提出一种基于人工神经网络的文档表示模型,并在此模型的基础上给出一种对中文科技文献进行聚类的方法。由于充分利用科技文献文档所具有的规范格式特点,该方法具有以下优点:

* 训练BP神经网络获得词语在文档中出现的权重,实现容易,计算复杂度低。

* 针对网上中文科技文档的规范格式特点,以简单易行的聚类方法统计满足用户需求的文档集合和感兴趣的主题。

本文首先介绍向量空间模型以及基于此模型的聚类技术,然后对本文的聚类技术所基于的BP神经网络的文档表示模型进行详细阐述,最后给出对中文科技文献文档进行聚类的方法。

2. 基于向量空间模型的聚类算法

2.1 向量空间模型介绍

向量空间模型(Vector Space Model, VSM)是Salton教授在1968年提出的,一直以来都是信息检索领域最为经典的计算模型。其思想是:把文档集合中的每篇文档都形式化为高维空间中的一个向量;同样把每个查询也转化为高维空间中的一个向量;然后按照TF·IDF(Term Frequency·Inverse Document Frequency)方法计算每一个文档词语权重;最后通过计算查询向量和文档向量夹角的余弦值来得到它们之间的相似度(距离)。

向量空间模型最大的优点在于[10]:①将文档映射到连续域的向量空间中,从而为使用多元统计方法提供了进一步分析处理的基础;②通过距离计算的远近可以给出查询和答案文档相关程度的级别表。

2.2 基于向量空间模型的常见聚类算法

基于向量空间模型的聚类算法在计算文档之间距离和文档与聚类中心(cluster centroid)距离时往往采用余弦值来测量。如[3]:

$$Cos(d,c)=(d\cdot c)/\|d\|\|c\|=(d\cdot c)/\|c\| \quad (1)$$

$$Cos(c_1,c_2)=(c1\cdot c2)/\|c_1\|\|c_2\| \quad (2)$$

式1用向量之间的余弦值表示出文档d和聚类中心c之间的相似度,而式2则计算出两个聚类中心之间的相似度。基于以上计算相似度的向量空间模型,常见的文档聚类算法如下:

I. 产生k个聚类的基本K-means算法 a)在n个元素中选择k个元素作为初始的聚类中心;b)把其余n-k个元素归到距离最近的聚类中;c)重新计算每一个聚类的中心;d)重复b),c),直到每一聚类的中心不再改变。

田 萱 研究生,研究方向为知识发现、数据挖掘.刘希玉 博士,教授,硕士生导师,研究方向为神经网络、微分方程.孟 强 博士,教授,硕士生导师,研究方向为交通规划理论、网络路由.

II. 简单的凝聚层次聚类算法 a)把 n 个元素分别作为 n 个聚类，计算所有聚类两两之间的相似度；b)合并最为相似(距离最近)的聚类；c)重新计算更新以后所有聚类两两之间的相似度；d)重复 b)，c)，直到只余下一个聚类。

除了以上基于向量空间模型的基本的文档聚类方法以外，还有许多改进的算法，这些算法大多以向量空间模型为基础，如文[9]给出了一种产生 k 个聚类的二分 K-means 算法，并进行了结果比较。

虽然向量空间模型表示起来容易理解，但向量空间模型的量化基础是词语的出现频率和出现文档的频率。对中文文档而言，这就需要进行繁琐的预处理。一种方法是：手工建立停用词表(stoplist)，并对中文文档按照词典进行分词，然后把文档投影到一个高维的词向量空间中；另一种方法是采用主题识别中的特征(词语)提取技术。无论哪一种方法，文档中存在的词语的巨大数量都让人不得不考虑到庞大的计算复杂度。

3. 基于 BP 神经网络的文档词语权重统计

针对向量空间模型所存在的固有缺点，本文提出采用 BP 神经网络的高度非线性拟合特点来对词语在文档中的权重进行非线性回归。

BP 网络作为目前应用最广泛的人工神经网络之一，是一种前馈型多层映射网络。它把一组样本的输入输出问题转化为一个非线性优化问题，使用了最优化方法中最普遍的梯度下降算法，运用迭代运算求解权相应于学习记忆问题，加入隐节点使优化问题的可调参数增加，从而得到更精确的解。

3.1 BP 网络输入输出的设计

在训练 BP 网络求解文献词语权重统计过程中，把一篇文档中给定词语出现的次数和文档的正文字数作为 BP 网络的输入，BP 网络的输出就是各个词语在该篇文档中的权重。为了训练网络，输入矩阵 $P_{m\times n}$，输出矩阵 $O_{m\times(n-1)}$，样本输入的理想输出就是系统的训练者根据每个词语在文档中的地位所确定的权重。其中：

(1) $$P=\begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1n} \\ x_{20} & x_{21} & \cdots & x_{2n} \\ x_{m0} & x_{m1} & \cdots & x_{mn} \end{pmatrix}$$

$$O=\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{pmatrix}$$

(2)m 表示训练样本集中有 m 篇文档，即 m 个样本；

(3)$x_{i0}(i=1,2,\cdots,m)$是每篇文档正文中所包含的字数，之所以作为一个输入参数是因为一个词语在文档中的权重与此篇文档所包含的字数也有关系；

(4)$x_{ij}(i=1,2,\cdots,m,j=1,2,\cdots,n)$是给出的第 i 篇文档中第 j 个词语出现的次数；$y_{ij}(i=1,2,\cdots,m,j=1,2,\cdots,n)$是给出的第 i 篇文档中第 j 个词语经过 BP 网络的非线性回归获得的实际的权重输出。

3.2 BP 网络的训练

这里在设计 BP 网络时，神经元的激发函数取 Sigmoid 型函数，即 $S(x)=(e^{x}-e^{-x})/(e^{x}+e^{-x})$。另外，BP 网络由输入层、隐层和输出层构成。其中隐层神经元节点的多少取决于网络的容量以及精度的要求。在人工神经网络理论上对 BP 网络神经元数目的选取尚无理论上的指导，一般通过经验及多次计算的调整结果来选定的。至于网络的初始权值，原则上在[-1,1]范围内任取。同时采用网络的实际输出与期望输出均方差最小来控制网络训练停止。该 BP 网络训练结果示意图如图1所示。

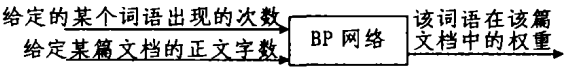

图1 BP 网络训练结果示意图

4. 聚类算法

一般中文科技文献文档都具有规范的格式，即从开头到结尾分别是标题、作者、摘要、关键字、正文和参考文献。基于此特点，下面提出了两种新的聚类算法。新的聚类算法在计算词语的权重时均采用上述训练的 BP 网络来计算。

4.1 寻找满足用户查询需求的聚类算法

假设条件：a)用户要求查询关键字为 W(可以是一个或多个词语，为便于描述算法，这儿假设就一个词语)；b)文献文档数据库中包含的文档记录个数为 N。

算法步骤：

a)检索文档数据库中的每一条记录，获得包含关键字 W 的文档集合 D_w，D_w 中包含的论文篇数为 N_w；

b)对文档集合 D_w 中的每一篇文档，计算它和用户查询关键字之间的相似度(该文档的权重)：

(1)If 文档标题中包含 W，此条文档记录的权重 $Q_i=Q_{i0}+1$，其中 Q_{i0} 表示 D_w 中第 i 篇文档的初始权重，一般设 $Q_{i0}=0$；

(2)If 文档关键字中包含 W，此条文档记录的权重 $Q_i=Q_i+1$；

(3)If 文档摘要中包含 W，此条文档记录的权重 $Q_i=Q_i+1$；

(4)If 文档正文中包含 W，此条文档记录的权重 $Q_i=Q_i+Q$，其中 Q 的计算如下：$Q=f_{BPnet}(W_i,Sum_i)$。W_i 表示关键字 W 在第 i 篇文档正文中出现的次数，Sum_i 表示第 i 篇文档正文的总字数。$f_{BPnet}(W_i,Sum_i)$表示经过上述 BP 网络的计算得出的关键字 W 在正文中的权重。这个值在[0,1]之间。

c)最后，对文档集合 D_w 中所有文档的权重求平均，把平均值作为阈值，即 $\theta=\sum_{i=1}^{N_w}Q_i/N_w$，如果集合 D_w 中文档的权重大于或等于阈值，则把它们作为聚类结果提交给用户。

以上便是根据用户输入的关键字来对科技文档聚类，以产生满足用户需求的文档。在这儿主要充分利用了科技文档固有的特定格式所体现的词语重要性。另外，这个算法里没有考虑科技文档所带有的参考文献和作者信息，其实这也是一种值得文档聚类参考的信息，因为一篇科技文献附带的参考文献内容常常与这篇文献的内容相关，而一篇文档的作者研究的内容大体也是相关的。文[5]就根据科技文档附带的参考文献信息来得到满足检索条件的文档集合进行了讨论。

4.2 产生与用户检索关键字相关主题的聚类算法

一个智能的文档检索系统除了能完成用户提交的检索任务之外，还应该能主动了解用户，学习用户兴趣，从而使返回的结果更符合用户的意图和要求。下面这个聚类算法通过提取满足条件的结果文档中重叠的关键字词语，来实现对用户

检索主题的扩展查询和兴趣学习。

假设条件：a)用户要求查询关键字为 W(可以是一个或多个词语，为便于描述算法，这儿假设就一个词语)；b)经过4.1节聚类算法得到的满足条件的文献文档集合 S，初始值为空；c)提供聚类中心词语的文档集合 S'，初始值为空；d)作为聚类中心的词语集合 C'，初始值为空。

算法步骤：

a)检索文档集合 S 中每一篇文档
If 文档标题中含有检索关键字 W，把该文档加入到集合 S'中
Then
{对该文档中标为关键字的每个词语(注意这儿要排除 W)进行以下处理：
If 该词语不在聚类中心集合 C'中
Then 把它加入到集合 C'中，并把 C'中该词语的初始权值记为0
Else 把 C'中该词语的权值自动加1 //即该词语已存在于聚类中心集合 C'中
}
Else //文档标题中不含有检索关键字 W
{If 文档关键字词语中含有 W，把该文档加入到集合 S'中
Then
{对该文档中标为关键字的每个词语(注意这儿要排除 W)进行以下处理：
If 该词语不在聚类中心集合 C'中
Then 把它加入到集合 C'中，并把 C'中该词语的初始权值记为0
Else 把 C'中该词语的权值自动加1 //即该词语已存在于聚类中心集合 C'中
}
Else // 该文档的标题和关键字词语中均不包含 W
{对该文档中标为关键字的每个词语(注意这儿要排除 W)进行以下处理：
If 该词语在聚类中心集合 C'中
Then 则把 C'中该词语的权值自动加1
}
}

b)最后把聚类中心集合 C'中词语的权值求平均作为阈值，把 C'中那些权值大于阈值的词语作为用户扩展的兴趣主题。

以上就是根据科技文献文档所固有的规范格式特点，通过聚类统计用户感兴趣的主题。这里为了减少计算复杂度，没有利用同义词典和近似词典。虽然不够精确，但因为只要文档数据库库容较大，就不妨碍产生用户的扩展兴趣主题。

总结 本文先就文档聚类的现状及常用的模型——向量空间模型进行了介绍，并提出基于 BP 神经网络的文档词语权重计算模型。基于这种计算模型，针对科技文献文档所固有的规范格式特点，给出了两种新的文档聚类算法，一种用来产生和用户查询相关的文档集合，一种用来产生用户的扩展兴趣主题。这两种聚类算法对设计个性化智能化的文档检索系统是有借鉴意义的。

当然，还有一些问题需要解决，例如：在训练 BP 网络获取文档词语权重时，网络的初值选取问题；在对科技文档进行聚类时，不同格式文档中词语的提取问题等。

参 考 文 献

1 (美)韩(Han, J.)、数据挖掘：概念和技术. 北京：高等教育出版社，2001
2 闻新，周露，王丹力，熊晓英. MATLAB 神经网络应用设计. 北京：科学出版社，2000
3 Steinbach M, Karypis G, Kumar V. A Comparison of Document Clustering Techniques. www.acm.org
4 Huang L. A survey On Web Information Retrieval Technologies. www.acm.org
5 Bollacker K D, lawrence S, lee Giles C. Discovering relevant scientific literature on the web. IEEE Intelligent systems, 2000, 15(2): 42～47
6 Martin J D. Clustering Full Text Documents. www.acm.org
7 Information Retrieval Survey —1997. www.acm.org
8 王实，高文. 数据挖掘中的聚类方法. 计算机科学，2000，27(4)：42～45
9 赵仲孟，张蓓，沈均毅. 对搜索引擎未来发展的探讨. 计算机科学，2001，28(3)：60～61
10 鲁松. 自然语言处理中词相关性知识无导获取和均衡分类器构建：[博士论文]. 中国科学院，2001

(上接第108页)

图像大小的稳定性其次。对于闭眼、戴眼镜等干扰较大的图像有一定的稳定性，而对于大胡子、复杂背景的照片则识别效果稍差。

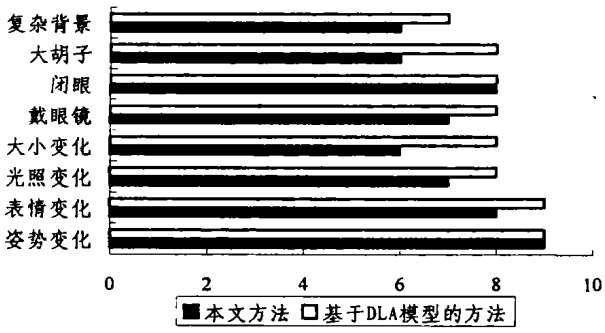

图4 各类情况下两种识别方法的结果比较

结论 本文运用弹性匹配的思想，主要从脸部对称性的利用、脸部模型的选取、人脸图像的各层小波子图的分析和利用、相似度和代价函数的计算等几个方面对现有弹性图识别方法作了改进，提出了基于小波变换域的人脸弹性识别方法。实验证明，本文方法在总体识别率上接近于基于 DLA 模型的方法，但在识别时间上仅为原方法的1/40，可大量节省查询时间。并且从具体实例上看，本文方法对人脸姿势、表情、光照以及大小变化等有较为稳定的识别效果，对于戴眼镜、闭眼、大胡须等易混淆的照片也具有较好的稳定性。

参 考 文 献

1 Valentin D, Abdi H, et al. Connectionsist Models of Face Processing: A Survey, Pattern Recognition, 1994, 27(9): 1209～1229
2 Lades M, Vorbuggen J, Buhman J et al. Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. on computer, 1991, 42(3): 300～311
3 Zhang Jun, et al., Face Recognition: Eigenface, Elastic Matching, and Neural Nets, Proc. of the IEEE, 1997, 85(9): 1422～1435
4 Averbuch, Lazar D, et al. Image Compression Using Wavelet Transform and Multiresolution Decomposition, IEEE Trans. On IP 1996, 5(1): 4～15
5 田金文，等. 变窗 Gabor 变换理论及其在图像处理中的应用. 红外与激光工程，1998，27(4)：1～5
6 Campell FW, Robson J G. Application of Fourier analysis to the visibility of gratings, J. Physiol. 1968, 19(7): 551～556