基于改进 BP 网络的网络论坛热点主题挖掘①

李法运, 陈 亮

(福州大学 经济与管理学院, 福州 350108)

摘 要: 针对BP网络收敛速度慢的问题, 提出将改善学习速率和增加动量项结合起来的方法对BP网络进行改进, 通过 Java 编程开发一个以改进 BP 网络为核心的热点主题挖掘原型系统. 实验结果表明, 该系统可以快速并准确地获取符合现实舆情的热点主题分类结果. 研究成果可以帮助民众及时了解网络热点, 帮助政府和企事业单位及时跟踪民意并做出相应决策.

关键词: 收敛速度; Java 编程; 改进 BP 网络; 热点主题; 挖掘; 舆情

Mining for Hot Topics in Online Forum Based on Improved BP Network

LI Fa-Yun, CHEN Liang

(School of Economics and Management, Fuzhou University, Fuzhou 350108, China)

Abstract: To solve the problem of slow convergence speed of BP network, a method is proposed to improve BP network by combining ameliorating learning rate and adding momentum factor. This paper developed a prototype system with improved BP network as the core of hot topics mining by Java programming language. Experimental results show that the system could acquire the classification results of hot topics according with the real public opinions quickly and exactly. The research result can help inform residents with instant hot spots as well as track public opinions so as to respond accordingly for governments and public enterprises.

Key words: convergence speed; Java programming; improved BP network; hot topics; mining; public opinions

近年来, 随着网络技术的推陈出新, 各种网络论坛井喷式出现, 它们构成了新形态的网络信息交互模式. 网民在网络论坛上各抒己见, 形成了一个庞大的社会舆情交流场所, 在这种背景下网络热点主题不断产生, 有些甚至形成了社会热点事件. 网络论坛蕴含着海量的信息, 这些信息以几何级数增长, 如何快速并准确地从这些信息中挖掘热点主题, 成为众多学者和专家感兴趣的研究课题[1].

1996 年, 美国国防先进研究计划局开始了 TDT (主题检测与跟踪)系统的研究[2]. 由于 TDT 与信息挖掘、信息提取等信息处理技术存在许多相似性, 并且主要检测与跟踪具有突发性和延续性特征的新闻语料, 因此逐渐成为当前信息处理领域的研究热点. 近年来中国科学院计算技术研究所(ICT)和北京大学等相关研究单位已经进行了该领域的研究[3]. 互联网的发展带动了国内关于热点主题检测的研究[4], 目前国内已经出现了很多网络舆情科研单位和校企合作的相关商业技术服务单位.

人工神经网络[5]通过模拟生物神经网络结构和功能进行并行信息处理继而进行人类智能活动机理的探讨和研究, 它与一般的逻辑学推理演绎相比更具有优势. BP(Back Propagation)网络是一种根据误差反传播算法(即BP学习算法)进行样本训练和测试的多层前向网络. BP 网络因其较强的收敛性、容错性、鲁棒性以及信息综合能力, 在热点主题挖掘领域有着良好的应用[6]. 然而它的缺点也很明显, 主要表现在网络训练的收敛速度较慢, 因此本文通过改善学习速率和增加动量项对 BP 网络进行改进从而改善其性能, 并提出

① 基金项目:“十三五”数字福建专项规划前期重点课题(822924)
收稿时间:2015-06-15;收到修改稿时间:2015-09-06

© 中国科学院软件研究所 http://www.c-s-a.org.cn

一种基于改进 BP 网络的网络论坛热点主题挖掘方法, 采用 Java 语言开发对应的原型系统.

本文通过开发原型系统下载央视网复兴论坛的主题帖子“2014 我对总书记说”的 1 万余条网友评论[7]并从中挖掘热点主题, 研究意义在于: 一、使民众快速地获取当前社会热点问题; 二、使企事业单位及时获取各自领域的最新动态和前沿信息, 并做出应对之策; 三、使政府了解当前民众最为关心的问题以及对这些问题的看法和态度, 帮助政府对公共事务做出科学决策.

1　BP网络

美国的 Rumelhart、McClelland 等人于上个世纪 80 年代提出了 BP 神经网络[8], 它可以学习和储存大批量的输入—输出映射关系, 并采用梯度下降法利用反向传播来不断修改神经元节点的连接权值及阈值, 使误差控制在一定范围内.

1.1 BP 神经元模型

图 1 是一个简单的 BP 神经元模型, 输入向量为 $X=(x_1,x_2,...,x_n)^T$, 权值向量 $W=(w_1,w_2,...,w_n)^T$, 输出函数 $y=f(\alpha,\theta)$, α 为全部输入的加权求和, 即 $\alpha=\sum_{i=1}^{n}w_ix_i$, α 也是激励函数 f 的输入, 阈值 θ 是 f 的另一输入.

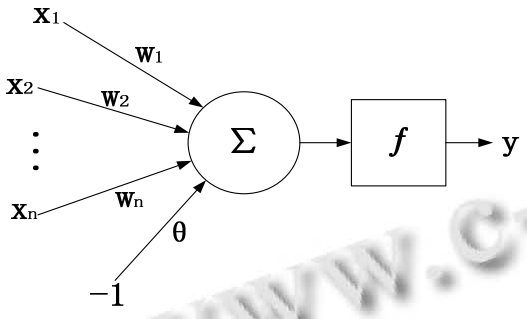

图 1　BP 神经元模型

1.2 BP 学习算法

BP 网络的学习过程由信息的正向传播和误差的反向传播构成. 正向传播时, 信息经输入层输入后, 先由隐含层处理, 然后将处理结果传给输出层. 若输出层结果没有达到理想值, 就进行反向传播, 误差信号沿原神经元连接线路返回. 在反向传播过程中, 不断调节每层神经元的连接权值和阈值, 直至误差控制在事先设定范围之内.

在学习过程中每输入一个训练样本即为一个学习周期, 一般要迭代很多周期并不断修改权值, 直至误差达到允许范围内.

1.3 BP 网络的改进

BP 网络因其优秀的非线性映射能力在热点主题挖掘领域有着良好的应用, 但它并不完美, 主要体现在 BP 网络训练时收敛速度较慢. 本文通过改善学习速率和增加动量项对 BP 网络进行改进, 从而提高了 BP 网络的收敛速度和稳定性.

自适应改善学习率可以缩短学习时间. 在初始化 BP 网络的学习率时, 若修改网络的权值后总误差变大, 则可以减小学习率; 否则可以在不引起网络振荡的前提下尽量使学习率取更大的值, 从而使权值误差达到期望值. 而增加动量项减小了网络的振荡性, 更易于找到最优解. 在修正 BP 网络的权值时, 给每个加权调整量增加一个正比于前一次变化量的值, 即动量因子, 每次调整完成后将该因子用于后续的加权调节过程, 它的思想是利用动量因子来传递权值变化的影响.

对 BP 网络的改进公式如下

$$\Delta w_{ij}^{(l+1)}=\alpha\cdot\Delta w_{ij}^{(l)}+\eta(l)\cdot\alpha\cdot f'(E)$$

$$\eta(l+1)=a\cdot b^x\cdot\eta(l)$$

其中 $\eta(l)$ 为学习率, α 是动量因子, 且 $0<\eta(l)<1$, $0<\alpha<1$, $f'(E)$ 为误差函数曲面负梯度方向, $\Delta w_{ij}^{(l)}$ 为权值调整量, α 为权值误差系数, b^x 为自适应学习因子.

本文将自适应改善学习率和增加动量项结合起来, 在改进算法中加入与权值调整有关的自适应学习因子和动量因子, 这样使得网络权值的加权调节过程、误差函数改变方向及网络自适应学习规律有机结合并共同作用于提高网络的收敛速度, 有效地提高了网络的鲁棒性和自适应性.

2　基于改进BP网络的网络论坛热点主题挖掘流程设计

如图 2 所示, 基于改进 BP 网络的网络论坛热点主题挖掘流程主要包括 Web 信息获取、信息预处理[9]、热点主题发现、热点主题结果评估等过程.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

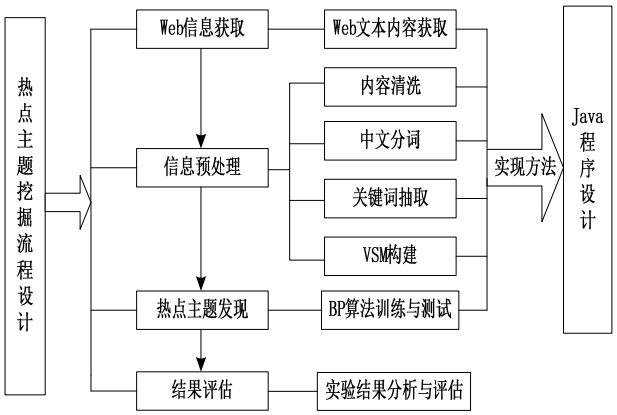

图 2 网络论坛热点主题挖掘流程设计

3 基于改进BP网络的网络论坛热点主题挖掘原型系统的开发

3.1 Web 文本内容获取

Web 文本内容获取过程分为获取网址和抓取网页内容两个步骤.

本文通过参考现有的网页设计原理和机制, 编写 Java 程序下载央视网复兴论坛的主题帖子“2014 我对总书记说”的 1 万余条网友评论并将其转换成数据库格式, 存入相应的 Access 数据表中, 从而实现 Web 文本内容获取.

3.2 内容清洗

运用一定的方法消除与主题无关的内容的过程称为内容清洗.

由于网络论坛中网友的跟帖评论存在主观随意性、恶意刷屏、语言表达不规范等现象, 所以需要对下载的文本内容进行清洗. 清洗的内容主要包括: 完全重复的记录、空记录、全为标点符号的记录、没有汉字的记录、全是数字的记录、全是英语字母的记录、含有“好好”“很好”“好的”等无关紧要的副词的记录、含有“习大大”“总书记”“谢谢”“你好”等问候语的记录. 此外, 如果删除后该字段为空, 则将该记录删除.

3.3 中文分词

按照一定的规则将连续的汉字字符串切分成若干词语的过程称为中文分词, 它是文本信息挖掘的基础.

现有的中文分词算法主要包括统计法、字符匹配法、理解法等. 本文提出基于 PAT 树的统计分词方法, 具体思路是:

①从数据表中提取文本, 然后进行文本切分, 形成短语集合.

②抽取正序数组和逆序数组. 将切分后的短语转换成半无穷大字串数组并去重、合并, 统计各字串的频次.

③构建中文 PAT 树. 将正序数组和逆序数组分别插入 PAT 树中, 然后用布尔型变量取代字符串型变量对 PAT 树进行改进.

④中文 PAT 树的检索和遍历. 先统计某个树的所有叶节点的词频, 执行最大词频查找, 返回两个布尔型数组是否相等, 当遍历到叶节点时取该叶节点的计数作为返回结果, 完成分词过程.

实际分词时可能会切分出一些共现频率高而不是词的字组, 需结合停用词表删除这些“噪音”字组.

3.4 关键词抽取

传统的关键词抽取方法如统计法和词语网络法等通常存在缺乏语义理解或者局限于字面匹配的缺点, 因此本文提出从总正序半无穷大字符串数组中逐个提取字符串, 再根据以下步骤, 完成关键词抽取.

①抽取正序字串. 将中文分词产生的子串转换成二进制字符串 x, 计算该子串的后继字串词频(不包括本身) S_xCount, 如果 $S\_xCount < t_1$ (t_1 是一个阈值), 则返回, 继续统计其他子串的后继字串词频, 以此类推, 求出后继字串的最高词频 $S\_xCount_{\max}$.

②计算包含 x 的所有子串的数量 $f(x)$, 若 $\frac{S\_xCount_{\max}}{f(x)} > t_2$ (t_2 是一个阈值), 返回.

③抽取逆序字串. 原理同正序字串的抽取.

④计算有效估计函数 SE_x. 若 $SE_x = \frac{f(x)}{f(y)+f(z)-f(x)} \le t_2$, 则返回. 其中 $f(y)$ 和 $f(z)$ 分别为 x 的正序、逆序子串.

⑤对筛选出来的关键词进一步处理, 去掉来自同一个字串的子串, 只保留最大字串, 至此关键词抽取结束.

3.5 向量空间模型构建

构建向量空间模型(VSM)的过程就是将文本处理简化为向量运算, 并以向量相似度表示语义相似度, 直观易懂. 本文采用 TF-IDF 统计方法计算各关键词在文本记录中的权值, 以此构建向量空间模型. 该方法的基本思想是: 如果某一词或短语在某一类文本中出现的频度很高, 而在其他文本中很少出现, 则认为此词或短语具有良好的类别区分能力, 适合用来分类.

构建向量空间模型的过程如下.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

① 关键词抽取结束之后统计各关键词在所有文档中出现的总数及其在各文档中的词频, 以此构建一个关键词词表数组.

② 计算各关键词在各文档中的权值, 然后将计算结果构建一个二维数组, 结构是以关键词为序, 保存各关键词在各文档中的权值. 这样就完成了向量空间模型的构建.

3.6 BP 网络训练与测试

3.6.1 手工分类

BP 网络样本训练与测试之前需要有一个网络模型, 本文首先将数据库中经过处理的前 200 条跟帖评论进行手工分类, 分类以 3.4 节抽取出来的关键词表为准, 将每条记录的分类结果(每条记录所属类别)加以保存, 而对于无类可归的记录则以-1 表示(如图 3 所示).

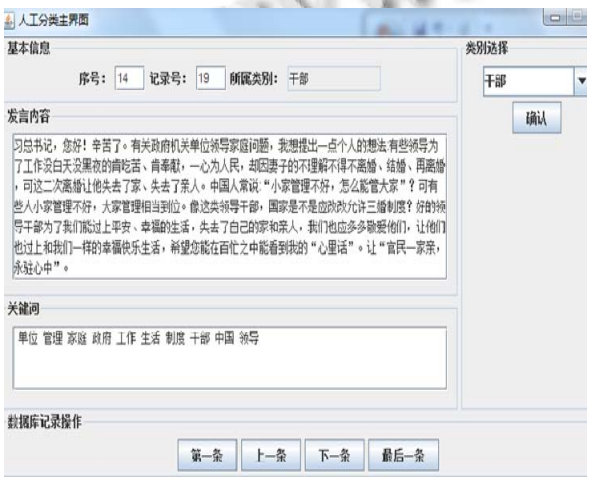

图 3　人工分类主界面

由自然语言组成的文本数据在抽取关键词后呈现出明显的数据稀疏性, 因此在手工分类中, 若不参考关键词进行分类, 就有可能造成手工分类的结果与样本学习后的结果存在较大的出入, 从而影响训练和测试的效果. 此外, 本研究还将手工分类的结果(类别)与原有关键词表对比去重, 然后添加到各自相对应的关键词中, 这样可以有效提高分类的准确率.

3.6.2 新样本训练

构造 BP 网络模型时, 需要设定相关参数, 比如动量项、学习的极限次数、允许误差值等. 本模型采用 Delta 学习规则, 网络的主要结构如表 1.

表 1　改进 BP 网络的主要结构

名称	描述
样本个数	800
网络层数及神经元个数	输入层(3), 隐含层(4), 输出层(1)
学习速率和动量项	初始学习率 0.2, 动量项 0.3
训练极限次数	100000
允许误差	0.001
激励函数	S 型函数

本文选取数据集中前 100 条记录对 BP 网络进行样本训练, 步骤如下:

①为网络提供 100 个训练样本, 初始输入样本包含记录号、关键词及所属类别 3 个神经元节点, 输出节点为主题分类预测结果.

②建立 BP 网络结构模型. 主要包括确定网络层数和各层神经元个数, 定义各层间连接权值矩阵并初始化权值矩阵值和阈值(系统随机取值).

③ 初始化学习率 η、动量因子 α 和允许误差值 ε 等参数(见表 1).

④取第 k 个样本 X_k 进行正向传播, 统计每层连接单元的实际输入和输出.

⑤计算输出层(设为第 L 层)各神经元节点误差 E_{jk}, 若对所有训练样本的任一样本 k 均存在 $E_{jk} \le \varepsilon$, 则学习过程结束; 否则, 进行误差反向传播.

⑥误差反向传播计算. 改善学习率, 增加动量项, 并更新权值矩阵, 修改第 L-1 隐含层至第 L 输出层连接权值, 依此类推, 逐层反向修改各节点连接权值. $k = k + 1(\mathrm{mod}\, N)$, 转步骤④.

样本训练结果如图 4.

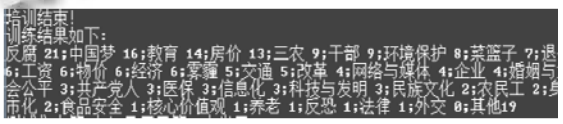

图 4　新样本训练结果

3.6.3 新样本测试

完成新样本的训练后, 需要对测试集所属的关键词类别进行测试, 以此来测试分类结果的正确率.

本次测试另外选取数据表中的 100 条记录, 经统计得知测试的正确率约为 85%, 考虑到计算机对自然语言的识别能力较低, 测试效果尚可. 在测试期间之所以会出现错误, 与样本学习过程中关键词分类时产生的误差有关系. 比如, 我们想划分出属于“中国梦”

© 中国科学院软件研究所 http://www.c-s-a.org.cn

这一类别的记录, 而有些记录中同时出现“中国”和“改革开放”两个词条的概率特别高, 因此 BP 网络对这些记录进行学习时可能会将“中国”归入到“改革”这一类别, 从而出现语言识别错误. 多次实验证明, 本程序的测试效果比较理想.

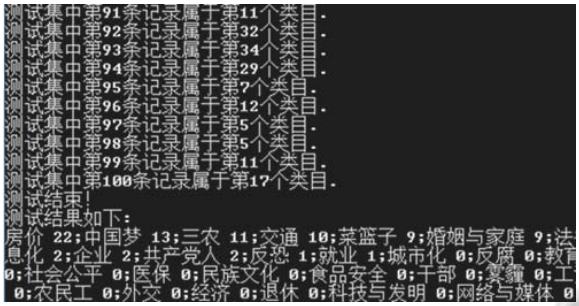

图 5 新样本测试结果

3.7 网络论坛热点主题发现与结果分析

针对 BP 网络收敛速度慢的不足将自适应改善学习率和增加动量项两种方法结合起来, 这种改进方法不易造成网络冗余, 减少了网络的学习迭代次数, 提高了网络的自适应学习能力, 使网络的稳定性和收敛速度达到一个理想的效果. 如图 6 所示, 改进后的 BP 网络大幅度减少了迭代次数, 更容易收敛, 而且网络更加稳定, 鲁棒性更强.

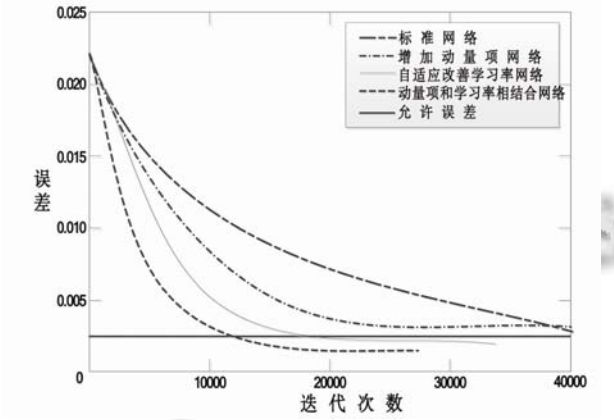

图 6 改进 BP 网络和标准 BP 网络迭代曲线对比图

通过对样本的不断训练和测试, 不断调整各参数的值, 优化网络结构, 然后对数据表中剩余的记录继续分类, 统计并分析分类结果. 分类结果即为网络论坛热点主题的预测集合, 该集合是一个降序排序的类标号与类中所含记录数的表, 从中可以发现网络论坛热点主题.

表 2 热点主题分类结果

次数 / 结果 / 关键词	标准BP网络分类结果	改进BP网络分类结果(4 次实验)			
		第一次	第二次	第三次	第四次
反腐	64	78	77	81	84
中国梦	70	63	53	59	64
教育	58	55	58	61	54
房价	28	39	40	34	33
改革	16	28	23	24	27
菜篮子	0	21	18	17	17
就业	9	16	18	19	20
交通	45	13	15	16	13
共产党人	21	10	7	6	11
农民工	5	0	3	0	5
医保	2	1	4	0	2
其他	38	24	24	22	25

如表 2 所示, 改进 BP 网络后热点主题分类结果更加准确. 从实验结果可以看出“反腐”、“中国梦”、“教育”、“房价”等主题的记录相对其他主题明显较多, 反映了民众对这些热点主题的特别关注, 符合近几年网络热点舆情现状, 体现了改进的 BP 网络在热点主题预测方面的科学性和实用性. 然而实验结果还有不尽人意的地方, 比如对“农民工”和“医保”等主题的预测结果显然不符合预期, 经过认真分析, 发现这是由原始数据的稀疏性和不规范性以及计算机对自然语言的识别能力与人脑存在偏差造成的.

4 结语

本文采用 Java 编程实现了 Web 文本内容下载、内容清洗、中文分词和关键词抽取以及向量空间模型构建等一系列数据处理技术, 提出了基于 PAT 树的统计分词方法、从总正序半无穷长字符串数组中逐个提取字符串的关键词抽取方法等, 并通过对 BP 网络进行改进, 提高了其收敛速度和稳定性, 最后采用 Java 语言开发了一个基于改进 BP 网络的网络论坛热点主题挖掘原型系统. 实验结果表明: 该系统可以从网络论坛即时下载并处理大批量的文档数据, 从而快速地获取符合现实情况的热点主题, 可实践性强, 具有一定的现实意义和创新性. 不可否认的是, 本文对部分热点主题的分类结果不是很理想, 在下一步的研究中将继续对 BP 网络进行研究与改进, 并从语义的层面进

© 中国科学院软件研究所 http://www.c-s-a.org.cn

行文档数据处理, 提高数据处理的精度.

参考文献

1 Hu CJ, Weng Y, Zhang XC, et al. Hot Topic Detection Based on Opinion Analysis for Web Forums in Distributed Environment. Studies in Computational Intelligence, 2009, 237:101–110.

2 智库百.TDT.http://wiki.mbalib.com/wiki/TDT. [2014-4-20].

3 侯晓冲.话题检测与跟踪算法改进研究[学位论文].武汉:华中科技大学,2013.

4 杨梅.网络舆情热点发现的研究[学位论文].北京:北京交通大学,2008.

5 邢红杰,哈明虎.前馈神经网络及其应用[学位论文].北京:科学出版社,2013.

6 Xu T, Xu M, Ding H. BBS topic's hotness forecast based on back-propagation neural network. 2010 International Conference on Web Information Systems and Mining. Washington. IEEE Computer Society, 2010:57–61.

7 复兴论坛.2014 我对总书记说. http://bbs.cntv.cn/thread-26816971-1-1.html.[2014-2-28].

8 曾喆昭.神经计算原理及其应用技术.北京:科学出版社,2012.

9 胡静,蒋外文,朱华.Web 文本挖掘中数据预处理技术研究.现代计算机(专业版),2009,(3):48–51.

www.c-s-a.org.cn

© 中国科学院软件研究所 http://www.c-s-a.org.cn