提取文本流主题的神经网络新算法*)

蒲晓蓉　叶　茂　游明英

(电子科技大学计算机科学与工程学院　成都 610054)

摘　要　目前,关于动态文本数据处理已逐渐成为数据挖掘的研究热点,例如,在聊天室中提取热门主题以及所有的讨论主题。目前已有的神经网络方法能较好地提取所讨论的主题,但不能决定哪个主题是热门主题,而且,提取到的主题之间相互干扰。利用主题之间相互独立和主题自相关的特性,基于自相关矩阵以及独立主元分析数学模型,本文提出一种新的神经网络方法,该算法能成功解决这些问题。在 Yahoo 聊天室上的实验结果表明,本文算法能准确提取主题以及热门主题,并且主题之间相互干扰大大减小。

关键词　独立主元分析,神经网络,自相关矩阵,时间序列

A New Neural Network to Extract Topics in Dynamical Text

PU Xiao-Rong　YE Mao　YOU Ming-Ying

(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054)

Abstract　Recently, the analysis of dynamically evolving textual data has become to an active research field in Data Mining. For example, extracting topics in the Internet chat lines. The existing neural network methods are based on linear time-series model, which could extract topics very well. But it cannot decide which topic is the hot topic and the topics disturb each other. Since the topics is independent each other and the topics are self- correlation, a new neural network is derived. It can solve the mentioned problems. Simulation results on Yahoo chat room illustrate that our neural network indeed extract meaningful and hot topics. And the disturbance between topics is very small.

Keywords　Independent component analysis, Neural network, Self-correlation, Time-series

在信息时代,Web 数据流量成指数级增长,海量的数据要求自动的数据分析工具。目前已有的算法和数学模型包括 Web 文档聚类算法,基于日志的 Web 挖掘算法等等。其中,值得一提的是 LSI(Latent Semantic Indexing)算法[1]。该算法的基本思想是通过奇异值分解,求解数据矩阵的奇异向量,并由较重要的奇异向量组成一个低维空间(该空间有时也称为语义空间),通过将高维数据投影到低维空间,实现数据的降维。在 Web 挖掘中,Web 文档通常被看成是由关键词组成的向量,这些向量是高维的而且稀疏。通过 LSI 算法对 Web 文档向量进行数据降维,能有效地避免维数灾难。尽管 LSI 算法能找到代表最大数据分布的奇异向量,但文本流的主题并不在这些奇异向量上。要找到文本流的主题,还需要进一步处理语义空间的数据。

动态文本流一般由许多主题组成。例如,一个聊天室内可能有许多人同时聊天,他们同时讨论许多主题,聊天室的数据形成一个动态的文本流。目前,已经出现一些方法用于提取聊天室主题。例如,假设各个主题之间相互统计独立,文[2]利用了 ICA(独立主元分析)神经网络方法;假设主题是线性时间序列,文[3]利用了非高斯方法。其中 Bingham 利用文[4]中的方法提取聊天室主题取得了较好的实验结果[5]。但这些算法都不能决定热门主题,并且各个主题之间也存在一定的干扰。

如果各个主题相互独立且主题自相关,而各个主题拥有不同的自相关函数,那么,本文提出的神经网络新算法将能有效地提取主题,且使得主题之间的干扰最小。该算法在空间上利用独立主题的非高斯性,在时间上利用主题的自相关特性,提取相对独立的主题并使得主题之间的干扰最小。同时通过比较自相关函数值的大小,也可确定在某段时间内的热门主题,即讨论得最多的主题。由此解决了原有神经网络方法的一些缺点。

本文首先通过 PCA 神经网络计算奇异向量以及语义空间[6],然后在语义空间中提取相对独立的主题。为了提取相互独立的主题,本文建立了结合时间和空间信息的目标函数以及优化该目标函数的迭代算法。实验结果表明,本文算法提取的主题相互干扰较小,并且所提取的热门主题与实际情况非常相符。

1　文本流模型以及预处理

处理动态文本流的一般方法是将动态文本分解为一串可以重叠的窗口,并将每一个窗口看作一篇文档,每一篇文档形成一个 n 维向量,n 代表不同关键词的总数。本文将这些关键词组成的高维空间命名为词空间。n 维向量中第 i 项的数值表示第 i 个关键词在文档中出现的频率。用 $x(t)$,$t=1$,…,N 表示这些文档向量,N 为文档的数目。由于 $x(t)$ 向量维数很高且非常稀疏,这些向量需要被投影到一个低维空间以避免维数灾难,且该低维空间能最大化表示这些数据,该空间也称为语义空间。令 $A=E\{x(t)x^{T}(t)\}$,其中 $E\{\cdot\}$ 是数学期望。由主元分析(也称 K-L 变换,PCA)理论,A 的主要特征向量组成该语义空间,主要特征向量指对应于矩阵 A 较大特征值的那些特征向量。这里需要特别指出的是,这些特征向量与 LSI 算法的奇异向量相同。由于矩阵 A 的维数非常高,不可能用传统方法直接计算矩阵 A 的特征向量,所以本

*)本研究获电子科技大学青年基金资助(编号:L08010601JX04030)。**蒲晓蓉**　博士生,讲师,主要研究领域包括神经网络、基于生物特征的模式识别等。**叶　茂**　博士,讲师,主要研究领域为神经网络,模式识别,智能计算。

文采用 PCA 神经网络在线计算 A 的特征向量。

PCA 神经元的输入为 $x(t)$，权向量为 $w(t)\in R^n$，输出为 $y(t)=w^T(t)x(t)$。通过权向量学习算法，$w(t)$将会收敛到 A 的最大特征向量。最大特征向量指，对应矩阵 A 最大特征值的特征向量。因为 A 为正定的对称矩阵，所以 A 存在对应特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n>0$ 的正交特征向量 $v_1,v_2,\cdots,v_n$。由文[6]，计算矩阵 A 最大特征向量的算法如下

$$w(k+1)=w(k)+\eta(t)y(t)(x(t)-y(t)w(t)) \tag{1}$$

其中 $0<\eta(t)<1$ 为学习速率。为确保算法(1)的全局收敛性，算法在每一步迭代中都需要规格化权向量

$$w(k+1)=\frac{w(k+1)}{\|w(k+1)\|} \tag{2}$$

那么，$w(k)$将全局收敛到矩阵 A 的最大特征向量 v_1[7]。算法(1)，(2)只能计算一个特征向量，为了计算更多的特征向量，我们将采用一种称缩小(deflation)过程的方法[8]。为计算第 j 个特征向量，迭代权向量 $w_j(k)$：

$$e_1(k)=x(k) \tag{3}$$

$$w_j(k)=\frac{w_j(k)}{\|w_j(k)\|} \tag{4}$$

$$y_j(k)=w_j^Te_j(k) \tag{5}$$

$$w_j(k+1)=w_j(k)+\eta(k)y_j(k)(e_j(k)-y_j(k)w_j(k)) \tag{6}$$

计算第 j 个特征向量时，神经元的输入为

$$e_j(k)=e_{j-1}(k)-y_{j-1}(k)w_{j-1}(k) \tag{7}$$

通过提取 K 个特征向量可得到语义空间为 $V_K=\{v_1,v_2,\cdots,v_K\}$。计算语义空间需要确定空间的维数，即 K 的大小，模式识别理论的一般规则是如果已提取的 K 个特征向量所对应特征值之和占矩阵特征值总和 90% 以上，一般认为这个 K 维空间已经最大化表示了 n 维空间的数据。将文档向量 $x(t)$ 投影到语义空间有

$$z(t)=V_k^Tx(t) \tag{8}$$

在进行主题提取时，还需将输入向量的模归一化处理，即 $E\{z(t)z^T(t)\}=I$，I 为单位矩阵。定义 $D=\mathrm{diag}\{\frac{1}{\sqrt{\lambda_1}},\frac{1}{\sqrt{\lambda_2}},\cdots,\frac{1}{\sqrt{\lambda_n}}\}$，令

$$z(t)=DV_k^Tx(t) \tag{9}$$

此时的 $z(t)$满足 $E\{z(t)z^T(t)\}=I$，下节将利用 $z(t)$提取主题。

2 算法描述

在动态文本流中，主题是指在一些关键词全体上的概率分布。一般假设在语义空间中文档向量 $z(t)$是一些相对独立主题的线性组合，记这些主题为 s_j，$1\leqslant j\leqslant m$，其数学模型为

$$z(t)=Bs(t) \tag{10}$$

其中 $s(t)\in R^m$，$B\in R^{K\times m}$，矩阵 B 未知。另外假设 s_j，$1\leqslant j\leqslant m$ 是任意的概率分布，但它们有不同的自相关函数，即对时间延迟 p，有 $E\{s_i(t-p)s_i(t)\}\neq E\{s_j(t-p)s_j(t)\}$。需要特别指出的是，在一般情况下，这两个假设对于动态文本流的主题都是满足的。

考虑一线性神经元，输入为 $z(t)$，权向量为 $w(t)$，输出为 $y(t)=w^T(t)z(t)$。根据独立主元分析神经网络的基本原理，定义非高斯目标函数如下[4]

$$J_1(w)=\frac{1}{2}\{E[G(z^T(t)w)]-E[G(v)]\}^2 \tag{11}$$

其中 v 是标准高斯随机向量，$G(\cdot)$是非线性函数，约束条件 $\|w\|=1$。如果选取的权向量 w 使得 $J_1(w)$最大，那么输出的 y 将是其中的一个主题。虽然通过 $J_1(w)$提取的主题相互独立干扰最小，但由于主题之间只是相对独立，而并非完全统计意义下的独立，所以提取的主题有可能并不准确。另外，通过 $J_1(w)$也无法确定哪个主题是热门主题。

假设主题 s_j，$1\leqslant j\leqslant m$ 满足一阶自回归过程，即 $AR(1)$模型，对某一时间延迟 p，有

$$s_j(t)=\tilde{s}_j(t)+b_js_j(t-p) \tag{12}$$

其中 b_j 为常数，$\tilde{s}_j(t)$为高斯白噪声。我们还可以用 n 阶 $AR(n)$模型来表示主题时间序列模型，为简单起见，本文仅采用一阶 AR 模型。定义 $\varepsilon(t)=y(t)-by(t-p)$，$b$ 为 b_j，$1\leqslant j\leqslant m$ 中的某一个常数。如果有单位权向量 w 使得 $E\{\varepsilon^2(t)\}$最小，从时间序列模型的意义上讲，此时的 $y(t)$最接近某个主题。而

$$E\{\varepsilon^2(t)\}=w^TE\{z(t)z^T(t)\}w-2bE\{y(t-p)y^T(t)\}+b^2E\{y(t-p)^2\} \tag{13}$$

其中 $E\{z(t)z^T(t)\}=I$，$E\{y^2(t-p)\}=w^TE\{z(t)z^T(t)\}w=I$。于是求单位向量 w 使 $E\{\varepsilon^2(t)\}$最小与求 w 使得 $E\{y(t-p)y^T(t)\}$最大等价。定义如下目标函数

$$J_2(w)=\frac{1}{2}w^TBw \tag{14}$$

其中 $B=\frac{1}{2}(E\{z(t)z^T(t-p)\}+E\{z(t-p)z^T(t)\})$为自相关矩阵，约束条件 $\|w\|=1$。将 $E\{y(t-p)y^T(t)\}$写为两个期望之和的目的是使 B 为对称阵。至此，极大化 $J_2(w)$的最优解实质上是求 B 的最大特征向量。优化 $J_2(w)$所求得主题的优点是自相关较强。同时，自相关矩阵最大特征值对应的那个特征向量实际上就是热门主题。因为自相关程度最高，也说明该主题讨论的频率最高。它的缺点是主题之间存在相互干扰，有时可能无法通过关键词集归纳出主题的具体意义。

结合 $J_1(w)$与 $J_2(w)$的优缺点，在约束条件 $\|w\|=1$ 条件下，定义目标函数

$$J_3(w)=aJ_1(w)+J_2(w) \tag{15}$$

其中 a 是由实际计算环境决定的经验值。通过极大化目标函数 $J_3(w)$求得的主题相对独立而且自相关程度很高。目标函数 $J_3(w)$只能提取一个主题，如果要提取更多主题，需要利用前述的缩小过程，第一个提取出来的主题就是热门主题。由数学分析基本知识，使 $J_3(w)$值最大的 w 使得其梯度为零，对 $J_3(w)$求梯度得

$$\begin{aligned}\nabla_wJ_3(w)=&a(E[G(z^T(t)w]-E[G(v)])\cdot E[G'(z^T(t)w)\cdot z(t)]\\&+(Bw\cdot w^Tw-w^TBw\cdot w)\end{aligned}$$

由梯度上升算法，使得 $J_3(w)$最大的权向量学习算法为

$$\begin{aligned}&w(k+1)=w(k)+\eta\nabla_wJ_3(w(k))\\&w(k+1)=w(k+1)/\|w(k+1)\|\end{aligned} \tag{16}$$

由文[4]，取非线性函数 $G(\cdot)$如下，

$G(u)=u^4$，$G'(u)=4u^3$ 或

$$G(u)=\frac{1}{a}\log\cosh(au),G'(u)=\tanh(au)$$

其中 $1\leqslant a\leqslant 2$。假设 $w(k)$收敛到权向量 w^*，则 $y(k)=|z^T(t)w^*|$表示了主题在不同时间的讨论强度。

算法(16)仅能提取一个主题，如果运行算法(16) p 次，可以提取 p 个主题，$w_1^Tz(t),w_2^Tz(t),\cdots,w_p^Tz(t)$，这些主题可能相关。所以，为了提取多个主题，需要在算法中使被提取的主题与已提取出来的主题不相关。如果已计算出 p 个主题，或向量 $w_1,w_2,\cdots,w_p$，在计算 w_{p+1}时，首先，结合前述中的缩小过程，从输入数据中减去已提取出来的 p 个主题分量；然后，利用 Gram-Schmidt 正交思想，在每一步迭代中减去 w_{p+1}在

已提取出来的 p 个主题向量上的投影 $w_{p+1}^T w_j w_j$，并规格化向量 w_{p+1}：

$$w_{p+1}=w_{p+1}-\sum_{j=1}^{p} w_{p+1}^T w_j w_j$$

$$w_{p+1}=w_{p+1}/\sqrt{w_{p+1}^T w_{p+1}}$$

3 实验结果

本文实验基于 Yahoo 聊天室，在线聊天数据从 Yahoo 电脑与互联网聊天室收集。日期为 2004-6-11，时间为下午 1:00 至下午 3:00。聊天室内大约有 30 个聊天者，在这 2 个小时内有 2850 行聊天数据。将动态文本流分为宽度为 12 行的 N 个窗口，并且窗口之间允许有 2/3 左右的重叠。移去一些没有意义的单词和标点符号，例如，I，Hc，is，was 等等，文本流将剩余有大约 1000 个不同的单词。将这些单词按照序列排列形成集合。为便于计算机处理，将每一个窗口写成一个文本向量 $x(k)$。如果单词集中的第 i 个单词在这个窗口中，则该文本向量的第 i 分量为 1，反之，为 0。

首先，利用算法(3)计算 K 阶 LSI 语义空间。计算 K 值的方法如下，如果已经提取的 K 个特征向量对应的特征值之和大于关联矩阵 A 特征值之和 90% 以上，那么一般认为此时的 K 个特征向量已经足以表示原有的数据。虽然算法(3)事先并不知道矩阵 A，由矩阵理论，矩阵 A 所有特征值之和等于它的迹。而矩阵 A 的迹恰好等于所有输入向量 $x(k)$ 平方和的平均数。所以无须计算矩阵 A，仍然可以得到矩阵 A 所有特征值之和。

在获得 K 维特征向量后，将输入向量 $x(k)$，$k=1,2,\cdots,N$ 投影到由其组成的 LSI 空间得到新的向量 $z(k)$，$k=1,2,\cdots,N$。对向量 $z(k)$，利用算法(16)能提取相对独立的主题。为简单起见，令 $G(u)=u^4$。在实验中，取时间延迟 $p=5$，$\alpha=\frac{1}{4}$，能取得较好的结果。算法需要提取主题的数目与前述决定 K 值的方法相同。本文实验共提取了 3 个主题。图 1 显示了不同的主题时间序列 $|w^T z(k)|$，$k=1,2,\cdots,N$，在不同时刻的活动情况。其中 w 表示某个主题向量。

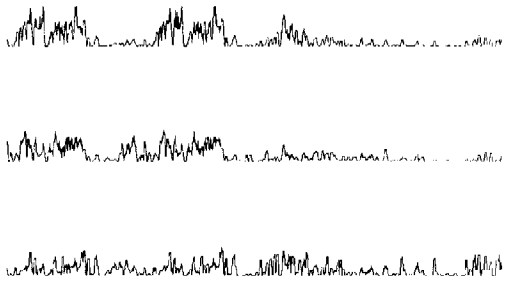

图 1 主题随时间变化图，最上面的时间序列表示第一个主题，依次类推

通过在词空间分析主题向量 w，可以理解各个主题的具体含义。如果主题向量 w 的第 i 个分量的绝对值最大，那么认为单词集中的第 i 个词是最能反映该主题的关键词。表 1 列出了所有主题对应的按大小排列的 6 个最重要的关键词。从表 1 中可以很容易地看出，第 1 个主题在讨论读大学所面临的经济问题，讨论内容包括经济资助，工作等，属于热门话题；第 2 个主题在讨论各种编程语言，讨论内容包括 qb，Java，asp 等。而第 3 个主题讨论大学里教授的工作情况，该主题是由第 1 个主题派生出来的，与第 1 个主题相对独立。实验结果与实际情况非常吻合。同时，也应该注意到提取出来的主题有一定的干扰，例如，第 1 个主题有单词 profs，第 2 个主题有单词 loans。这些微小的干扰并不影响对主题含义的正确理解。

表 1 与图 1 相关的主题

主题 1	主题 2	主题 3
ut	qb	profs
loans	java	job
school	vb	courses
aid	asp	work
work	ms	money
profs	loans	school

利用 Complex pursuit 算法提取的主题如表 2 所示。由于算法也无法确定热门主题，因此表 2 中的主题是无序的。由表 2 可以看出，Complex pursuit 算法在提取主题准确度上不及本文算法，特别当主题之间存在某种关联性但又相对独立的情况下。例如，主题 1 与主题 2 混淆在一起，而主题 3 有主题 1 和主题 2 的单词 ut，money 等。产生误差的原因是 Complexpursuit 算法仅考虑了主题的时间的自相关特性，而并没有考虑主题在空间上的相对独立性。

表 2 Complex pursuit 算法提取的主题

主题 1	主题 2	主题 3
work	aid	qb
qb	course	java
school	financial	vb
ut	job	ut
money	profs	ms
debt	javascript	money

结束语 结合动态文本流主题时间自相关与空间相对对立的特性，在文本流主题有不同的自相关函数条件下，本文建立了提取文本流主题的目标函数，并且，导出了求解目标函数最优解的神经网络新算法。在聊天室的实验结果表明，本文算法不但能较准确地提取主题，而且能决定哪一个是热门主题。与其他算法比较，本文算法准确度得到了较大的提升。将来的研究方向是如何动态地提取文本流主题。

参 考 文 献

1 Deerwester D, Dumais S, Furnas G, Landauer TK. Indexing by latent semantic analysis. Journal of the Am. Soc. for Information Science, 1990, 41: 391～407

2 Isbell CL, Viola P. Restructuring sparse high dimensional data for effective retrieval. Advances in Neural Information Processing Systems, 1998, 11: 480～486

3 Molgedey L, Schuster H. Separation of a mixture of independent signals using time delayed correlations. Physical Review Letters, 1994, 72: 3634～3637

4 Hyvarinen A. Complexity pursuit: separating interesting components from time-series. Neural Computation, 2001, 13: 883～898

5 Bingham E, Kaban A, Girolami JK. Topic identification in dynamical text by complexity pursuit. Neural Processing Letters, 2003, 17: 68～83

6 Oja E. Principal components, Minor components, and linear neural networks. Neural Networks, 1992, 5: 927～935

7 黄克军，叶茂，王雁东，李毅超．一种全局收敛的 PCA 神经网络学习算法．计算机科学，2004，31(5)：153～156

8 Cichocki A, Amari S. Adaptive blind signal and image processing. John Wiley & Sons, 2002. 193～199