基于 BP 神经网络的互相关注用户兴趣的演变与同化预测

Evolution and Assimilation of User Interest based on BP Neural Networks in a Cluster of Mutually Following Users

周天祎 *
ZHOU Tian-yi

摘 要 BP 神经网络用于建立一个互相关注的用户兴趣变化的简单模型。为了便于分析，国外知名在线知识市场 Quora 的用户数据被作为分析的在线社区的数据来源。通过开源的爬虫软件对 Quora 上的用户数据进行下载并在本地进行分析；通过对话题类型的抽象和归类来量化用户的兴趣指标；通过建立 BP 神经网络模型，对现有用户兴趣的变化规律进行学习，从而由用户当下的兴趣变化规律来预测用户未来的兴趣；该模型取得了 88% 的预测精度。最后，由生成的足够多的未来预测数据和下载的实际用户数据的校验比对，得出了用户兴趣的演变规律：在线社区中互相关注的用户之间的用户兴趣会有同化趋势。

关键词 BP 神经网络；用户兴趣；数据挖掘；数据分析；数据预测

Abstract The technology of BP neural networks provided an elementary model for the evolution of interest of mutually following users in the online community. For the convenience of analysis, in this thesis, Quaro data was used as the input data source of the online community. First of all, Quora data was downloaded by using an open-source data crawler; the topics assigned to each question was used as an indicator of users' interest; BP neural networks model was applied to the data to predict future user interest based on previous inputs. The model can predict future data with average accuracy 88%. Finally, enough future data was generated and compared with the downloaded data to arrive at the conclusion of user interest evolution: user interest between mutually following users on the online community will assimilate.

Key words BP neural networks; user interest; data mining; data analysis; data prediction

doi：10.3969/j.issn.1672-9528.2016.09.031

1 引言

随着机器学习技术与自然语言处理技术的发展，对于用户在社交网站上的行为和互动已经有了很多深入的研究。例 如，Investigation and Analysis of Research Gate User's Activities using Neural Networks 专门研究了在线学术平台 Research Gate 的用户活动[1]。但是，这些的研究基本都是对当前用户的行为和互动进行分析刻画，而对于用户未来行为与兴趣的研究则较为匮乏。这意味着商家只能了解用户对于现有产品的喜好，而不能预测用户对于未来产品的兴趣。因此，一个能够用于精准预测用户未来兴趣的模型对于商家显得尤为重要。本文提出了一种较为初等的方法，通过建立一个 BP 神经网络模型，来分析在线社区 Quora 上结构化的用户数据，并以平均 88% 的精度预测用户的未来兴趣。

2 研究方法

本研究有一个基本的前提假设：在一个彼此相关的群体中，人们会互相影响并逐渐改变他人的兴趣。所以，研究首先要找到一个强相关的群体。然后，该群体中的人在 Quora 上的行为动态会被抓取并下载到本地。相邻两天的用户动态关系会通过人工 BP 神经网络进行分析。当 BP 神经网络模型被训练而有一个较高的精度时，该模型就被用于生成对未来用户兴趣的预测数据，从而找到一个用户兴趣的演变趋势。

* 上海市世界外国语中学 上海 200233

经过研究发现，在线问答网站 Quora 上的数据具有高度结构化的特点，因此 Quora 被选为本研究的在线社区的数据源。这样就不必先对数据源进行聚类分析，同时最终模型也有更高的可信度。Quora 上高度结构化的数据可以如下图所示：

• 用户动态（关注一个用户，关注一个话题，提出问题，回答问题，支持一个答案）

• 问题（问题名称，问题的回答，问题所属的话题）

本课题只研究支持一个答案这一种用户动态，因为这是 Quora 上最常见的，而且这种动态最能够代表用户的兴趣。

为了简化模型，必须同时满足另一个条件：用户群体在给定的时间段内互相关注。

3 数据预处理

3.1 数据的下载及分类

为了进行本地的分析和建模，需要下载 Quora 的用户数据；而 Quora 的数据都是基于网页的。因此，本研究使用了一款开源的基于 Python 3 框架的软件来自动下载所需的 Quora 用户数据。需要的 Quora 用户数据包括：

• 用户关系：

程序以一个用户为起始点，通过多层搜索，最终找到了 5 个互相关联的用户。

• 用户动态：

程序以一个用户的“支持一个答案”这种动态为起始点，获取答案所在问题及其所属的话题。该话题数据需要被进一步处理。

最终，获取到的数据被存入数据库准备进一步分析。

因为 Quora 上有上千个话题，为了简化模型，需要对原话题进行抽象处理，从而可以用更上层的抽象话题来代替原话题。

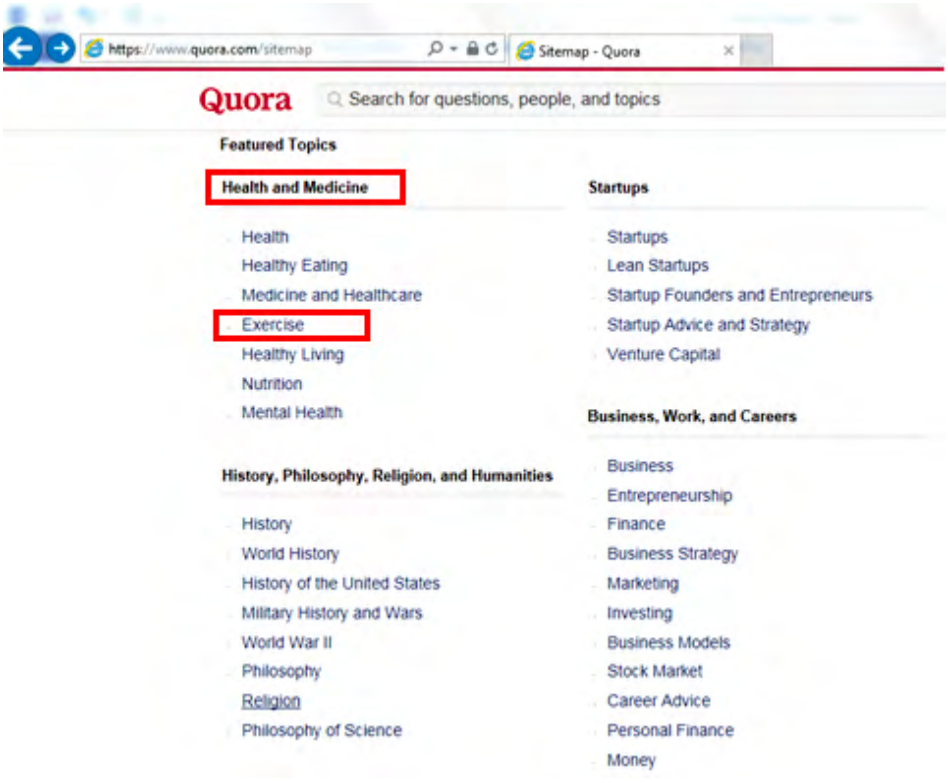


图 1：Quaro 上典型的话题结构

上图展示了一个典型的话题树形结构。在截图中，健身锻炼是一个原话题，而健康及医疗是一个抽象话题。本研究使用一个 HTTP 爬虫和一个 HTTP 分析器来自动提取出诸如“健康及医疗”这种抽象话题。

在通过程序的自动化分类处理后，本课题从其中提取了三个主话题：“历史、哲学及宗教”，“生活、关系及个人”，“科学、技术及工程”。

3.2 用户每日动态流矩阵与数列表示

按照选取的话题内容和用户的行为动态，用户每天的动

态被转换为一个 1×3 的行矩阵 Lj，其中 j 表示用户的索引。每个在 Lj 中的元素是一天中该话题出现的次数。这个矩阵可以理解为一个用户在这一天的兴趣向量。对于这 5 个用户，L1、L2、L3、L4、L5 可以合并成一个 5×3 的矩阵 Mi，其中 i 表示日期的索引。每个代表每日所有用户动态流的矩阵 Mi 被放入一个有序数列 S 中。

如上述研究方法中的介绍，用于研究的用户必须在给定的时间段内相互关注。因此，仅仅在 2014 年 9 月 14 日到 2015 年 6 月 16 日得到的用户每日动态流矩阵 Mi 才被保留在 S 中，其余的都被删除。

另外，有时在一天中存在用户没有任何动态的情况，因此数列 S 中会有部分为 0 的数据。这样的训练数据会使得 BP 神经网络将 NaN 作为预测结果。为了解决这个问题，可以通过合并相邻的含有大量 0 的数据来实现。合并的结果同样代表了在一段时间内用户受到影响的总和，具有一致的参考意义。而且在合并的过程中，数列 S 的长度会缩短，这更进一步加快 BP 神经网络训练的过程。

假设数列 S 的长度为 n，那么将数列中的第一个到第 (n-1) 个 Mi 矩阵作为训练数据中的输入变量，数列中的第二个到第 n 个 Mi 矩阵作为训练数据中的输出变量。这些训练数据会被导入人工 BP 神经网络进行学习。

4 BP 神经网络的构建及模型训练

4.1 BP 神经网络的构建

BP 神经网络的基本结构包括为：输入输出模型、作用函数模型、误差计算模型和自学习模型 [3][6]。

4.1.1 节点输出模型

隐节点输出模型：$Oj = f(\sum W\_ij \times X\_(i - qi))$ (1)

输出节点输出模型：$Yk = f(\sum Tjk \times Oj - qk)$ (2)

f- 非线形作用函数；

q - 神经单元阈值。

4.1.2 作用函数模型

作用函数是反映下层输入对上层节点刺激脉冲强度的函数又称刺激函数。一般取为 (0,1) 内连续取值 Sigmoid 函数：

$$f(x) = 1/(1 + e^{-x}) \quad (3)$$

4.1.3 误差计算模型

误差计算模型是反映神经网络期望输出与计算输出之间误差大小的函数：

$$Ep = \frac{1}{2}\sum(tpi - Opi)^2 \quad (4)$$

tpi-i 节点的期望输出值；

Opi-i 节点计算输出值。

4.1.4 自学习模型

BP 神经网络的学习过程，即连接下层节点和上层节点之间的权重矩阵 Wij 的设定和误差修正过程。BP 网络有师学习方式 - 需要设定期望值和无师学习方式 - 只需输入模式之分。自学习模型为：

$$\Delta Wij(n+1) = h \times \Phi i \times Oj + a \times \Delta Wij(n) \quad (5)$$

h- 学习因子；

Φi- 输出节点 i 的计算误差；

Oj- 输出节点 j 的计算输出；

a- 动量因子。

如下图 2 所示，BP 神经网络一般被称为：三层前馈网或三层感知器，即：输入层、中间层（也称隐层）和输出层 [4]。

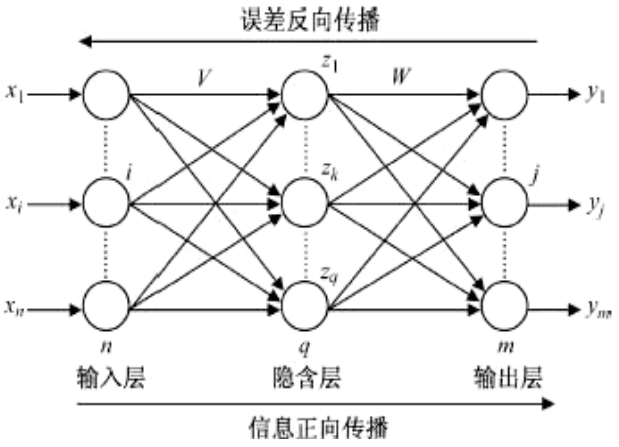


图 2 BP 神经网络基本结构

建立 BP 神经网络前首先要确认整个网络结构：确定网络的节点层数和每层的节点数量。该结构确认，仍无成熟的理论，故需依据经验值来确认其网络结构参数。

本文中选择了 4 种网络结构，最终选择一种最优的方案作为解决问题的神经网络模型 [5]。

网络结构 1：

- ✓ 输入层：6 个节点；
- ✓ 第一隐层：7 个节点；
- ✓ 输出层：3 个节点；

网络结构 2：

- ✓ 输入层：6 个节点；
- ✓ 第一隐层：4 个节点；
- ✓ 输出层：3 个节点；

网络结构 3：

- ✓ 输入层：6 个节点；
- ✓ 第一隐层：4 个节点；
- ✓ 第二隐层：4 个节点；
- ✓ 输出层：3 个节点；

网络结构 4：

- ✓ 输入层：6 个节点；

- ✓ 第一隐层：5 个节点；
- ✓ 第二隐层：4 个节点；
- ✓ 输出层：3 个节点；

4.2 BP 神经网络的模型

BP 神经网络的搭建环境为 MATLAB，共选用 36 组训练样本用于训练集，前 3 个网络结构的训练周期为：100，第 4 个网络结构的训练周期为：1000。通过调整学习步长以及反复训练，4 种网络结构均达到了收敛的预设训练要求。

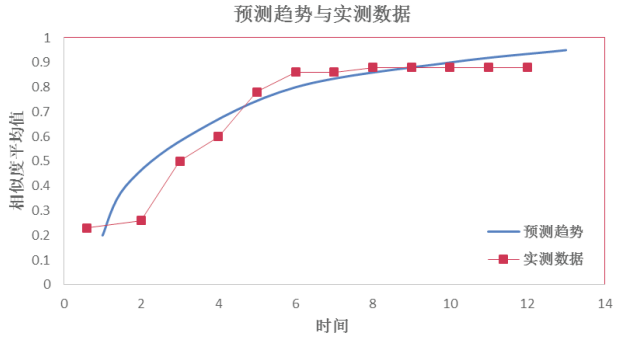
在满足预设目标的前提下，选择了结构最为精简的网络结构 2 作为本文的最终选定的网络模型结构。

4.3 BP 网络模型的数据预测

训练好 BP 网络模型后，模型被用于预测之后 500 个用户每日信息流矩阵 Mi。每一个用户每一天的动态用一个向量 V 代表。两个向量间的角度可以用来代表用于兴趣的相似程度。因为这个角度可以用公式：$\mathrm{Cos}(\theta)=\frac{V1*V2}{|V1||V2|}$，这个 cosine 值可以直接用于代表兴趣的相似度。一个新变量 S 被用于表示相似度。$S=\mathrm{Cos}(\theta)=\frac{V1*V2}{|V1||V2|}$。当 S=0 时，两个向量相互垂直，因此相似度最小；当 S=1 时，两个向量重合，相似度最大。

平均相似度 $\bar{S}$ 从所有 24 种不同的用户兴趣的两两组合的平均值获得，并形成所有 $\bar{S}$ 的一个数列。为了减少数列中的数据的短期波动，每 50 个 $\bar{S}$ 再求一次平均，最终减少每日信息流的随机性的影响。

从在线社区上获得的实际数据和 BP 神经网络预测的结果，如图 3 所示：



图表 3：用户兴趣相似度演化趋势

从长期趋势角度，用户的兴趣吧被逐步同且其相似度达到最终饱和值。该结果与初始假设相符：在线社区中，相互关注的用户兴趣有被同化的趋势。

5 结论

本文提出利用 BP 神经网络对用户在在线社区中的兴趣变化进行建模预测，并通过与社区的实际数据进行比对，最终验证了初始假设，在线社区中，相互关注的用户兴趣有被同化的趋势。

后续仍需持续增加模型的训练样本数量，进而提升模型的可靠性。同时需要考虑增加其他推荐性的算法，加快 BP 神经网络训练收敛速度慢的问题，并结合模糊技术，进一步优化 BP 神经网络的输出。

参考文献

1. Omar Alheyasat. Investigation and analysis of research gate user's activities using neural networks.

2. M.T. Hagan, H.B. Demuth, M.H. Beale. Neural Network Design. Number v. 10 in Neural network design. Campus Pub. Service, University of Colorado Bookstore, 2002.

3. 王正勤，基于优化 BP 神经网络的车型分类识别技术的研究．化学工业出版社，第 1 版（2014 年 1 月 1 日），2014

4. 蒋天一，胡德金，许开州，等．改进型 BP 神经网络对球面磨削最高温度的模拟及预测 [J]．上海交通大学学报，2011,45(6):901-906

5. 王宁．一种基于 BP 神经网络的即时在线推荐系统．计算机技术与发展，2009,19(7)

6. 于婷婷．基于 BP 神经网络的滚动轴承故障诊断方法 [D]．大连：大连理工大学电子与信息工程学院，2008.

（收稿日期：2016-08-26）