基于深度学习的科技演化趋势算法研究

物联网工程专业学生　　陈　熙

指导教师　　张士庚

摘要

随着科学研究的高速发展，预测科技趋势引起人们的密切关注。本文提出了科技演化趋势预测算法，首先科技热度根据多项式岭回归模型被预测，之后科技演化的过程被看作是迭代式的重组与变异，重组包括领域分支的分叉和汇合两步。按照距离概率和近邻相似的思想确定分叉数，而汇合步骤采用了自适应数据规模的 K-mean++算法,分叉和汇合之间通过距离变异和分支变异提升泛化性能。输出结果最后在 Sankey 图中进行了展示。评估实验基于权威学术平台 Arnetminer 科技大数据集，验证了科技演化趋势预测算法对深度学习领域在 1973-2013 年间的演化预测能力。

关键词：科技演化　预测算法　Arnetminer　重组和变异　K-mean++

Study on Deep Learning-based Prediction Algorithm of Technology Evolution Trend

ABSTRACT

Resulting from rapid growth of research, trend prediction of science and technology has attracted public concern. In this paper, a prediction algorithm for the evolution trend of science and technology is proposed. First, heat of each term is predicted by the polynomial ridge regression model. After that, evolution process is regarded as iterative reorganization and variation. Reorganization is further divided into two steps -- split and confluence. According to the distance probability and neighbor similarity, the number of branches is determined. Adaptive data scale K-mean++ algorithm is then applied during confluence. To optimize generalization performance, distance variation and branch variation occurs between the split step and confluence. Output is present in Sankey diagram. Based on big data of well-known academic platform Arnetminer, experiment evaluates the performance of the technology evolution trend prediction algorithm for deep learning field from 1973 to 2013.

Key Word: The evolution of Science and Technology Prediction Algorithm Arnetminer Recombination and Mutation K-mean++

目录

引言

随着时代的发展，科技在现代社会中的地位日趋重要。在学者们的研究中，总是通过不断地借鉴、改进、创新、否定……前人的研究成果实现将科学技术水平不断推向前进。从宏观角度上来说，技术发展和生物进化有着相似之处，它们都是通过量变和质变的相互交替，实现由低级到高级、由简单到复杂的发展、继承和变化，而且这些过程往往是有规律可循的。

目前大型的科技智库[11]日渐成熟，通过对科研数据的深度挖掘，我们可以方便地梳理技术领域的演化历程。因此设计合适的预测算法来挖掘科学技术演化的历史大数据成为可能，实现对技术趋势的可靠预测[12]。

从科学研究的意义上来说，该研究将机器学习的方法和科学技术发展的大数据紧密结合，开辟了机器学习预测研究的新领域；从国家层面上来说，该研究将对国家科学技术研究的战略布局提供积极的参考，有利于优化科研资源配置、避免无效科研投入、实现关键领域的突破；从人类社会的角度来说，该研究将为人们认识科学技术的发展规律提供更全面的视角。

第 1 章　绪论

1.1 机器学习

21 世纪是信息化的时代，而信息时代注定离不开一个主题——“人工智能”。而机器学习，又是人工智能中最核心的部分。它使得程序能自动化且掌握数据内部的一些特征和规律，从而高效地完成许多日常生活息息相关的工作——如物品分类，用户画像等等[2]。大体来说，目前利用机器学习解决的常见问题类型包括：分类、聚类、预测问题等。

1.2 科技演化

随着时代的发展，科技在现代社会中的地位日趋重要。同生物的进化过程一样，技术领域内部的技术活动、子技术或技术主题也是随着时间推移存在发展、继承和变化的，而且这些过程往往是有规律可循的。斯坦福教授 Vinodkumar Prabhakaran 研发了工具 ArgZoneTagger[9]，可将摘要的片段划分为理论、方法、实验、结果等 7 个类型，Tom Hope 也将论文核心语句按原理和方法两类给定不同的创新因子，从而激发工程师们产生更多的设计灵感[10]。这些前沿研究的成果都说明科学技术本身是存在结构特征且可以被学习和利用的。

目前科技智库[11]日渐成熟，通过对科研数据的深度挖掘，我们可以让机器不断地去学习技术发展的规律，并在此基础上实现对未来技术发展的可靠预测。

1.3 学术搜索引擎 Arnetminer

Arnetminer 是 2006 年由清华大学知识工程研究室的唐杰等人开发的国际知名学术搜索引擎。ArnetMiner 利用数据挖掘和社会网络分析与挖掘技术，提供研究者语义信息抽取、面向话题的专家搜索、权威机构搜索、话题发现和趋势分析、基于话题的社会影响力分析、研究者社会网络关系识别、即时社会关系图搜索、研究者能力图谱、审稿人推荐在内的众多功能，为研究者提供更全面的领域知识，和更具针对性的研究话题和合作者信息，为科研的更好发展提供服务。

Arnetminer 系统多年来为世界顶级出版刊物提供审稿推荐服务，目前已收集了 7900 多万论文信息、3900 多万研究者信息，1.3 亿论文引用关系、780 万知识实体以及 3 万多学术

会议/期刊。吸引了全球 220 多个国家的 600 多万用户访问[15]。

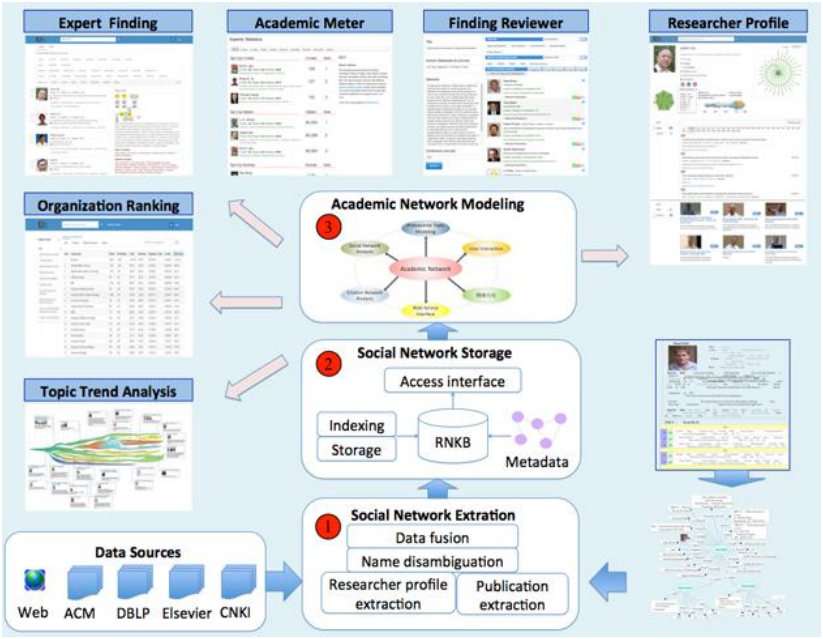

图 1-1 Arnetminer 学术搜索引擎的核心架构[13]

1.4 数据可视化

人类总是从视觉起始观察和理解事物，这种奇妙的能力使得我们能迅速抓住日常场景的特征。当数据的规模相当庞大，且元素间充满复杂多变的关系时，展示数据变成了一件极具挑战性的任务。至今市场上已经出现了成百上千种各具特色的可视化技术和应用。

展现数据的机制有上百种，但其实它们都可以抽象成一个统一的模型。系统地说，可视化数据一般要经过获取数据、数据转型、可视化映射、浏览视图四个步骤。这个过程其实可以用一个 “数据流模型”来描述。图 1-2 是一个典例。

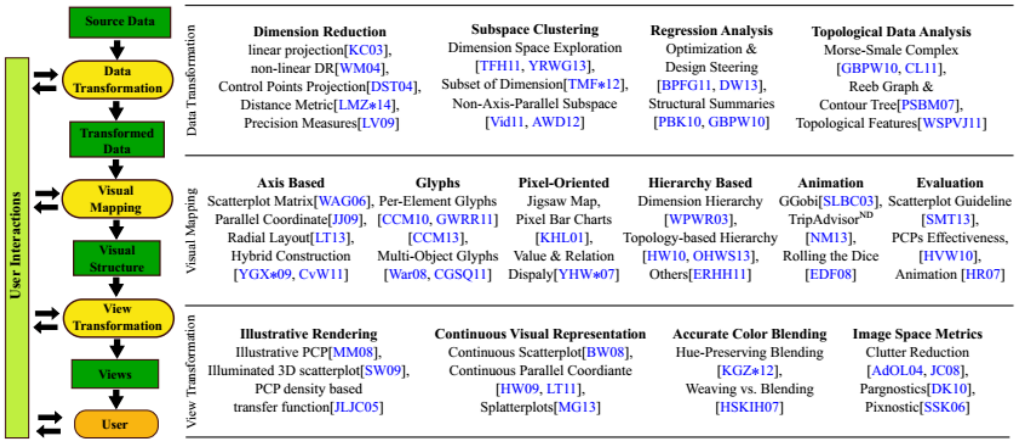

图 1-1 信息可视化流水线中的转换步骤[14]

第 2 章　模型分析与设计

本章首先会对算法进行需求分析，之后再详细分析各个模块的设计思路。

2.1 需求分析

实现有效的科技趋势预测，我们首先获取质量高的科技大数据，按照一定的规范格式进行组织，然后根据时间和空间特征的挖掘和学习方法建立数学模型，用代码实现对应的算法，最后需要得到预测的结果并分析其可靠性。

2.2 设计思路

对于一个算法的研究而言，首先要解决的问题就是——明确算法的输入内容和输出内容。由于我们需要对不同科技领域的发展趋势进行预测，实质上输出的内容应当是一个类似图的数据结构，输出图中的点表示各个技术分支，点的大小反映当前分支的热度，而不同点之间连成的边则是反映这些技术分支之间是否关联以及关联的程度。

而要实现数据的预测，显然输入和输出需要存在时间依赖关系。循环神经网络（RNN）是目前运用于时间序列分析的主流模型[17]，然而科技趋势演化的预测需要考虑到不同技术分支之间的相互影响，更普适的说法可以称作是空间特征。卷积神经网络（CNN）正是一种引入了强空间特征的流行模型[18]，然而实验条件受限，所以并未采用这种模型。

综合考虑时序和空间特征关系后，科技趋势的演化过程被发现与物种繁衍相似，借鉴遗传算法来实现演化预测的功能。即科技特征元组作为演化的载体——基因。初代种群产生后，按照适者生存和优胜劣汰原理，逐代演化，在每一代，根据问题域中个体的适应度大小选择个体，借助遗传算子进行重组和变异，产生出代表新一代科技演化关系的解集。

基于以上分析，我们最终提出使用这样一种策略来预测科技趋势：首先在 Arnetminer 引擎上使用 API 提取期刊论文和会议论文中各领域的词频、学者数等组合的综合评价指标作为发展热度；然后根据人工标注的方式对领域关联关系进行初始化，再将发展热度和关联关系作为演化预测模型的输入，最后得到下一年代的发展热度和领域间关联关系。

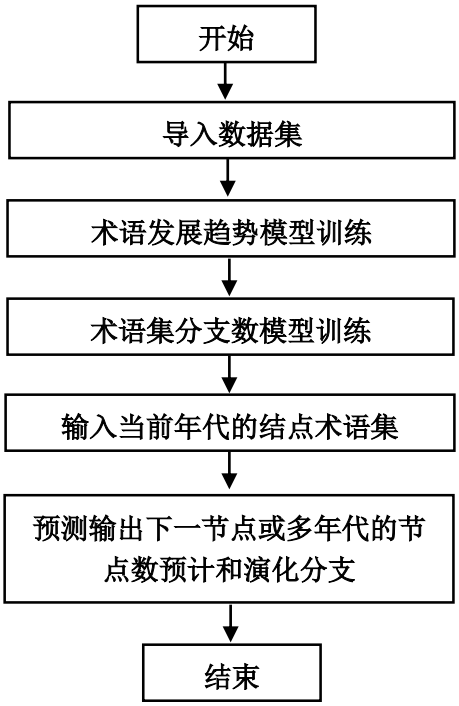

图 2-1 科技演化趋势预测算法流程图

2.2.1 数据获取

科研数据为基于作者-论文、论文-术语集、术语-聚类-权重关系的 json 数据，获取过程中涉及到知识图谱、自然语言处理、文本聚类、矩阵分析等一系列的知识，详细处理过程如图 2-2 所示：

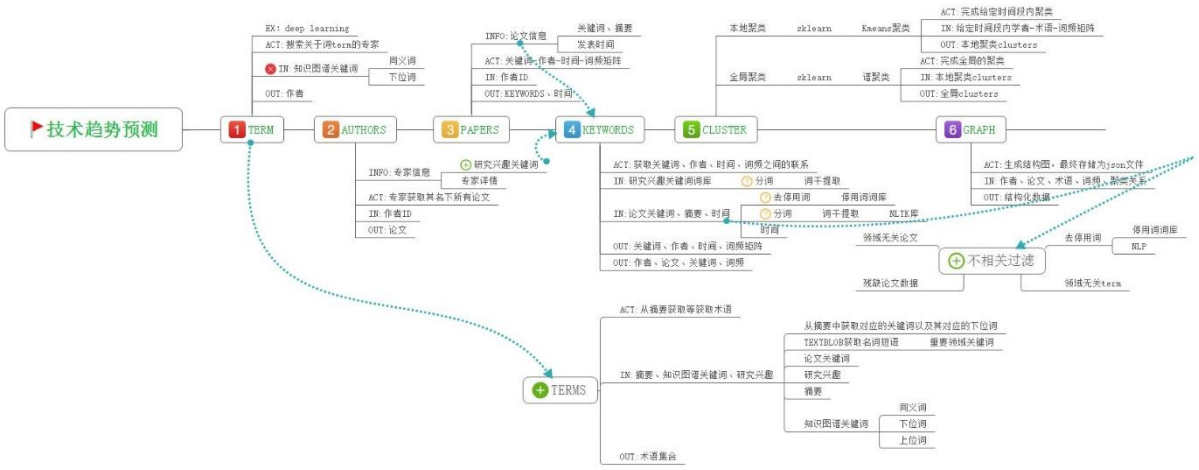

图 2-2 技术趋势预测基础处理详图

后面数据分析的实验中，我们截取了“深度学习”相关领域的数据作为我们实验的基础数据。其主要内容如下：

(1) 来自 1973-2013 年间“深度学习”领域热度最高的 100 个术语的发展趋势年谱和专家年谱、以及相关论文；

(2) 1973-2013 年间这些术语之间的演化关系。

其数据可视化结果如图 2-3 所示：

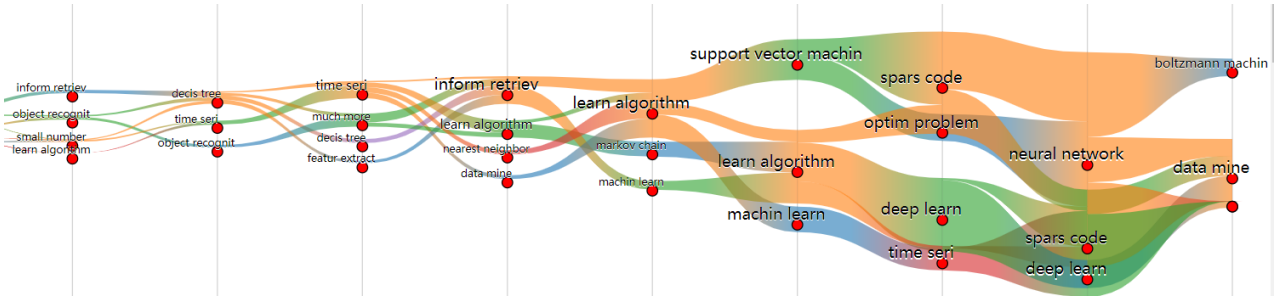

图 2-3 Deep Learning 领域科技演化图

2.2.2 术语发展趋势预测模型

根据之前的假设，可直接利用已有的术语发展趋势数据（即热度大小）建立回归模型进行多项式拟合，数学表达式为

$$y = W \cdot X + b \tag{1}$$

其中W为$\{x_1, x_2, x_3 \dots\dots\}$，$x_i$对应热度时间序列的第$i$个数据；$W$为$\{w_1, w_2, w_3 \dots\dots\}$，$w_i$对应$x_i$在拟合方程中的权重系数。而拟合的目标（代价函数）则是最小二乘的无偏估计：

$$\min\left(\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} w_i x_{ij}\right)\right) \tag{2}$$

考虑到数据集规模有限，如果多项式最高次项比较大，模型就容易出现过拟合。为了同时保证拟合精度和减少过拟合的影响，在公式 1 和公式 2 的基础上又采用引入正则化的岭回归代价函数。正则化是一种常见的防止过拟合的方法，一般原理是在代价函数后面加上一个对参数的约束项，这个约束项被叫做正则化项。引入正则向后的代价函数形式如下：

$$\min\left(\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} w_i x_{ij}\right)\right) + \gamma \sum_{j=0}^{p} w_j^2 \tag{3}$$

$\gamma \sum_{j=0}^{p} w_j^2$是用于参数正则化带来的罚函数。图 2-4 更直观地展示了正则化项对代价函数的防止过拟合的作用。

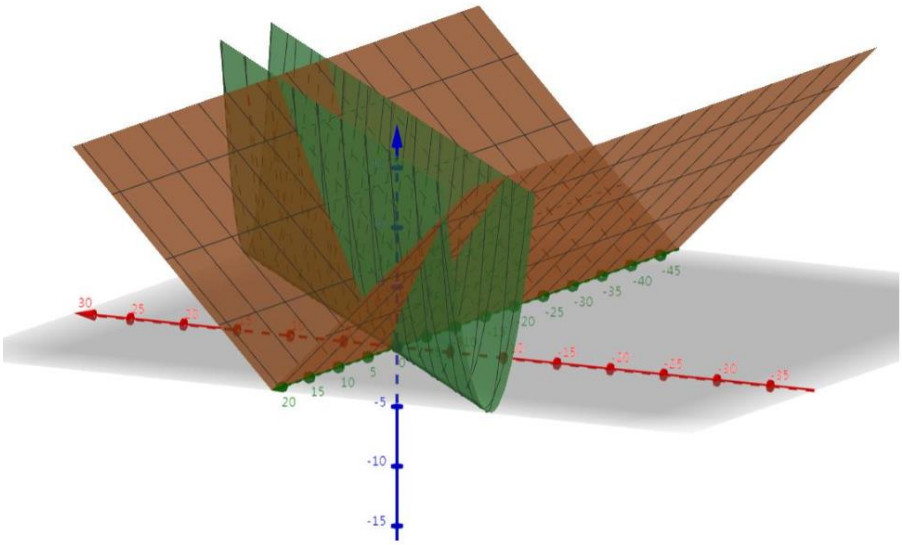

图 2-4 代价函数与正则化项图像的叠加的示意图[19]

2.2.3 术语集演化分支数预测模型

根据假设，术语集的分支数与术语集大小间存在联系。因此，可以统计各分支数对应的数据，建立分类模型，而后进行分支数预测。对于较小的数据集也可以遍历比较进行预测。我们比较了术语集大小按分支数取平均值的方法和直接取大小最相近的术语集分支数进行预测两种方法。我们以预测与实际的分支数相等的结点数和总结点数的比值作为准确度来评估不同方法的性能。实验中，我们发现取大小最近的分支数这种方法准确度较高。后面的实验中均采取此方法进行分支数预测。

表 2-1 中，显示了使用前 30 个节点数据做训练集，两种方法下对所有的 40 个节点数据的验证效果。

表 2-1 两种分支数预测方法下的验证效果

指标	按分支数取平均法	最近取相同法
准确度	19/40	37/40

2.2.4 术语集演化关系预测模型

演化关系预测包括分支的产生预测与合并预测。考虑到演化的上下位节点之间有明显的继承关系，这里将每个年代的演化分支产生看作是对每个术语集内部重新聚类，聚类后的每个类即一个分支；而演化分支的合并则看作是这些类的进一步聚合，如图 2-5 所示。

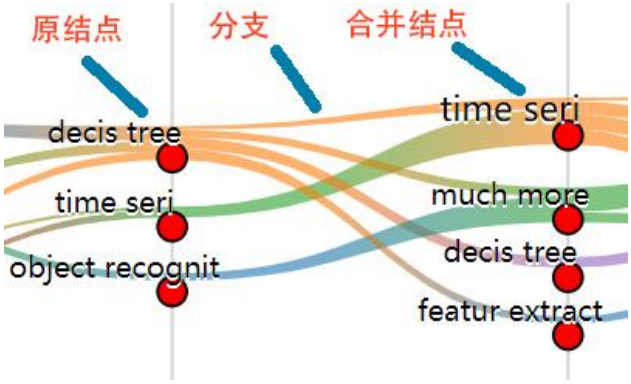

图 2-5 演化关系示意图

根据假设术语集产生的分支数可以预计，因此分支产生步骤采取 KMEAN++方法[20]来重新聚类，聚类依据为术语间的关联距离，类间距离看作所有类间术语距离的平均值。汇合步骤无法确定最终的术语集个数，但每个节点至多产生 1 个分支到单个演化后的新术语集中，即对每个节点取 0~1 个分支轮询进行合并，直至所有分支处理完毕，聚类依据仍然为术语关联程度，具体是为合并前后的距离比值设置一个阈值。

此外，数据中还存在由一组术语集部分演化到另一组术语集的情况，这可以解释为演化过程中存在突变环节。突变环节也分为两个因素，一个是是否突变，由突变概率决定；另一个是突变为哪个术语，由术语突变距离决定。因为突变为小概率事件，学习梯度应该随变异发生的次数显著累加，用指数函数对突变距离转换取值空间来反映这一点。

相比 K-mean 算法，K-mean++可以大大减小初始化造成的精度缺陷；相比其他聚类数目确定的算法，K-mean++则具有无监督学习的特性，能够根据数据规模和分布自适应地给出聚类结果，这两个特点使得科技趋势的预测效果变得更加精准而自然[24]。

Kmean ++ 算法的具体步骤如算法 1。

算法 1：K-mean++ 聚类算法

1： 随机选取一个样本作为第一个聚类中心c_1；
2： While 没有选出k个聚类中心时：
3： 计算每个样本与当前已有类聚中心最短距离（即与最近一个聚类中心的距离），用$D(x)$表示；这个值越大，表示被选取作为聚类中心的概率较大；最后，用轮盘法选出下一个聚类中心；
4： End
5： While 一轮循环中仍有聚类中心的位置发生改变时：
6： 针对数据集中每个样本x_i计算它到k个聚类中心的距离并将其分到距离最小的聚类中心所对应的类中；
7： 针对每个类别c_i，重新计算它的聚类中心$c_i = \frac{1}{|c_i|}\sum_{x \in c_i} x$（即属于该类的所有样本的质心）；
8： End

该预测模型的训练，也按年代先后对节点依次进行分支训练、变异训练和合并训练：

(1) 分支训练： 选择当前年代的一个节点，找出它与分支到的所有下一年代节点之间的交集，看做是分支的术语内容，对每个分支的所有术语间距离乘以学习因子α_1（小于 1），即加强它们的关联度。

(2) 变异训练： 对当前选择节点在第一步得到的所有分支取合集，然后用该节点的术语集与该合集取差集，看作是变异部分。然后对变异部分每个术语的变异概率乘以学习因子β（大于 1），并找出它在下一年代哪个节点中，对它与找出节点内每个术语的变异距离乘以学习因子γ（小于 1），即提高变异概率。

(3) 合并训练：对下一年代的每个节点即分支合并的每个节点中的所有术语，它们之间的距离乘以学习因子α_2（小于 1）。

2.2.5 Sankey 力学图

在科技趋势预测系统的可视化界面中，演化关系是通过 Sankey 图的形式来呈现的。Sankey 图最明显的特征就是，始末端的分支宽度总和相等，即所有主支宽度的总和应与所有分出去的分支宽度的总和相等，保持能量的平衡，从而使展现效果具有均衡的美感。图 2-6 给出了 Sankey 图的一个典型可视化效果。

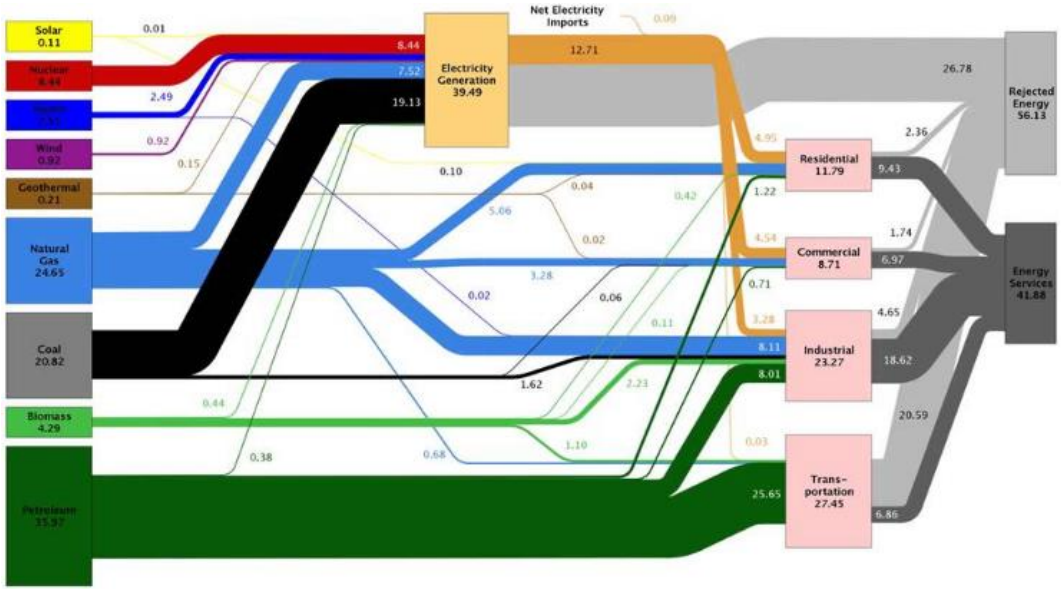

图 2-6 Sankey 力学图示例[22]

第 3 章　实验结果

在这一章，用于验证和评估科技趋势预测算法的实验被介绍和分析。首先，给出的是实验的运行环境和参数，然后关于算法性能评估的实验结果将被分析。

3.1 实验介绍

科技演化趋势算法的评估实验在 Pycharm[25] 2017.2.4 编程环境上运行，主要参数设置是：初始变异率为 0.01，初始术语距离为 1，演化率为 0.8，变异率为 1.1，类内演化度为 0.95，变异演化度为 0.9。参数设置如表 4-1 所示。训练数据集和验证数据集的来源是清华大学知识工程研究室研发的学术搜索引擎 Arnetminer，数据内容为 1973-2013 年间“deep learning”领域热度最高的 100 个术语的发展趋势年谱和专家年谱、以及相关论文中提取的热度值以及这些术语之间的关联情况。

表3-1 参数设置

符号	参数名	数值
ch_rate	初始变异率	0.01
ch_dist	术语距离	1
DIST_TRAIN_RATE	演化率	0.8
vari_rate	变异率	1.1
inner_ch_rate	类内演化度	0.95
vari_ch_rate	变异演化度	0.9

3.2 输出展示

算法的研究目的是优化华为知识洞察系统中的话题趋势预测模块，图 3-1 展示了术语演化的实际情况，图 4-2 展示了根据我们所设计的算法预测出来的结果，图 3-3 是当前华为知识洞察系统的界面效果。

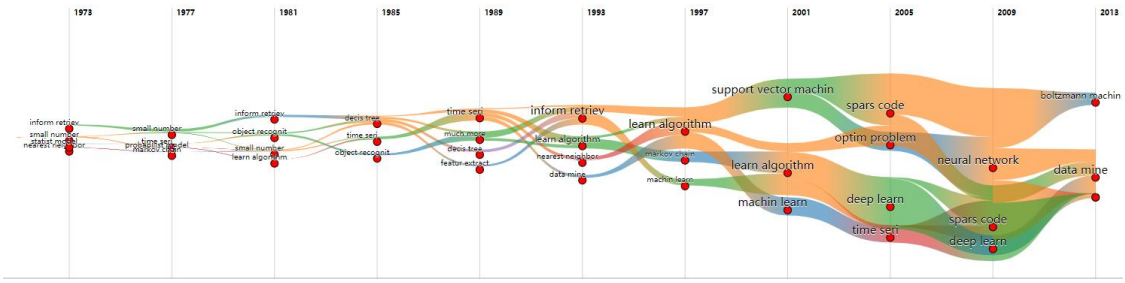

图 3-1 实际术语演化情况

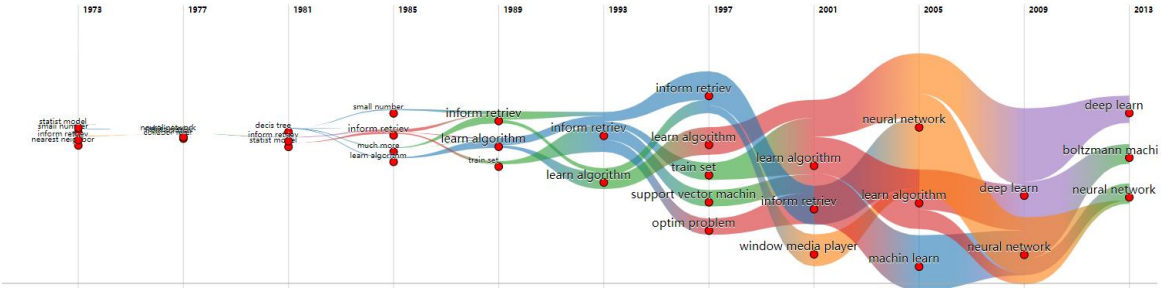

图 3-2 算法预测的术语演化情况

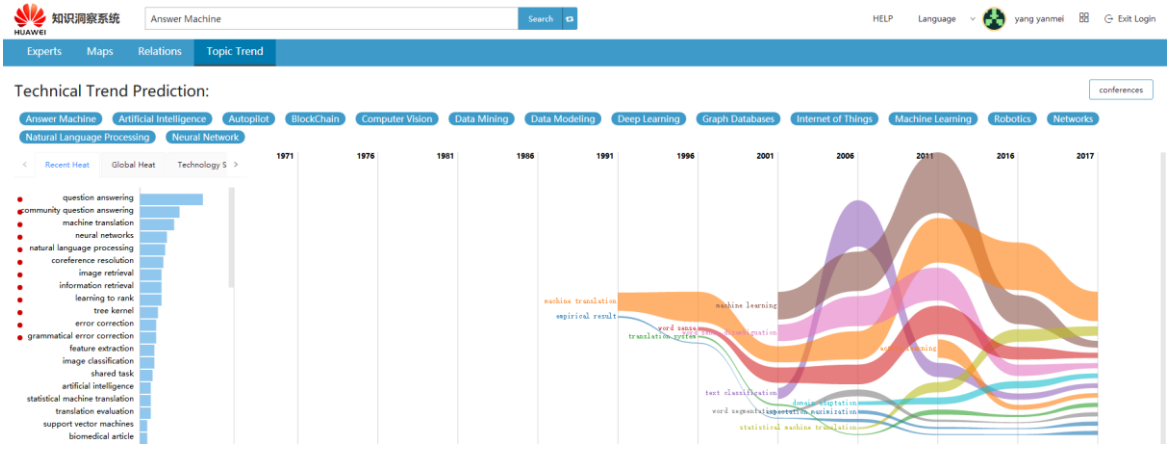

图 3-3 华为知识洞察系统技术趋势预测模块

其中横轴为时间轴，代表不同年代，而 Sankey 图中的每个结点表示一个技术分支，结点之间的关联即代表领域技术随时间的演化关系，关联的粗细反映了对应技术的热度大小。可以发现实际演化过程中的很多重点术语在算法应用中成功预测。相比华为知识洞察系统的可视化效果，科技演化趋势算法不但给出了各个分支的热度变化，还反映了分支间的关联性，以及各个时期的主流技术流派，科技结构的可视化效果有了显著提升。

图 3-4 显示了以 neural network 为中心的一组术语集热度，包括 neural network, time seri, deep learning, gene model, unsupervised learn, supervised learn, belief network, machine learning, feature learn. 这些与当前领域研究的发展情况相符。

neural network: 0.0006998958911283162
time seri: 0.0015243849184809967 deep learn: 0 gener model: 0 unsupervis learn: 0 belief network: 0 machin learn: 0 supervis learn: 0 spars code: 0 featur learn: 0 time seri： 0 Times 160.38989978748097

图 3-4 以神经网络为中心的术语集热度

表 3-2 通过比较数据进一步分析算法的预测效果，可以看到预测结果的结点数和分支数大致符合实际情况，分支精度较低是因为分支的确定比较主观，不同粒度的展示效果可能差别很大[26]。

表 3-2 实际演化情况和算法预测结果的比较

指标	实际值	预测值
节点数	39	37
分支数	63	59
节点精度	100%	67.5%
分支精度	100%	43.2%

3.3 精度与时间关系分析

图 3-5 显示了算法的预测节点在不同年份的精度，可以看出，精度随时间的推移呈现上升趋势。这是因为在算法的训练部分，我们对模型使用了所有年份的数据来学习。预测早期演化时，模型已经掌握了未来演化的一些特征，由此产生了“与阅历不符”的预测判断[29]。另外，早期演化中许多术语刚刚兴起，短期内发展差距并不大，这也是促成预测失准的重要因素。

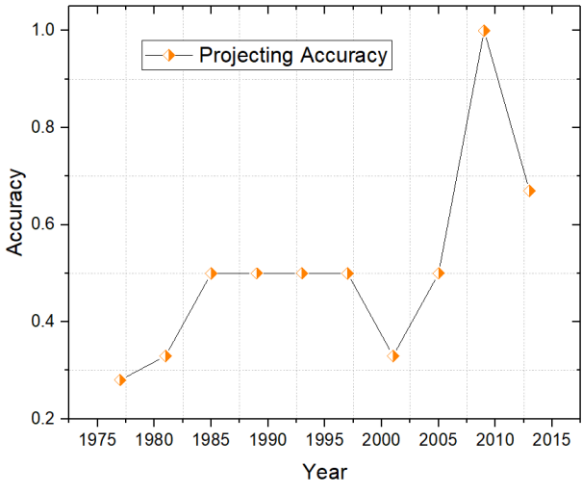

图 3-5 节点预测精度与时间的关系

3.4 多项式岭回归分析

在预测节点的发展趋势时用到了岭回归的方法，其中罚函数参数 alpha 的取值对于结果有着显著影响： alpha 过小则会淡化罚函数的作用，退化为最小二乘法的无偏估计；alpha 过大则会削弱最小二乘因子的作用，降低曲线模型对数据的拟合程度。图 3-6 的(a)-(d)分别显示了 $\alpha = \lg 5 \times [10^5, 10^8, 10^{15}, 10^{16}]$ 时对术语 learn algorithm 的预测情况(1973-2003 年为训练集，2003-2013 年为验证集)。可以看到(a)图中曲线很好地拟合了训练集数据，但在验证集的表现却非常糟糕；(b)和(c)相较(a)则显然在验证集上误差更小；(d)图则反映出 alpha 设置过大，这时曲线甚至都无法很好地拟合训练集数据。这些情况符合分析（图中横轴为年代，纵轴为术语热度）。

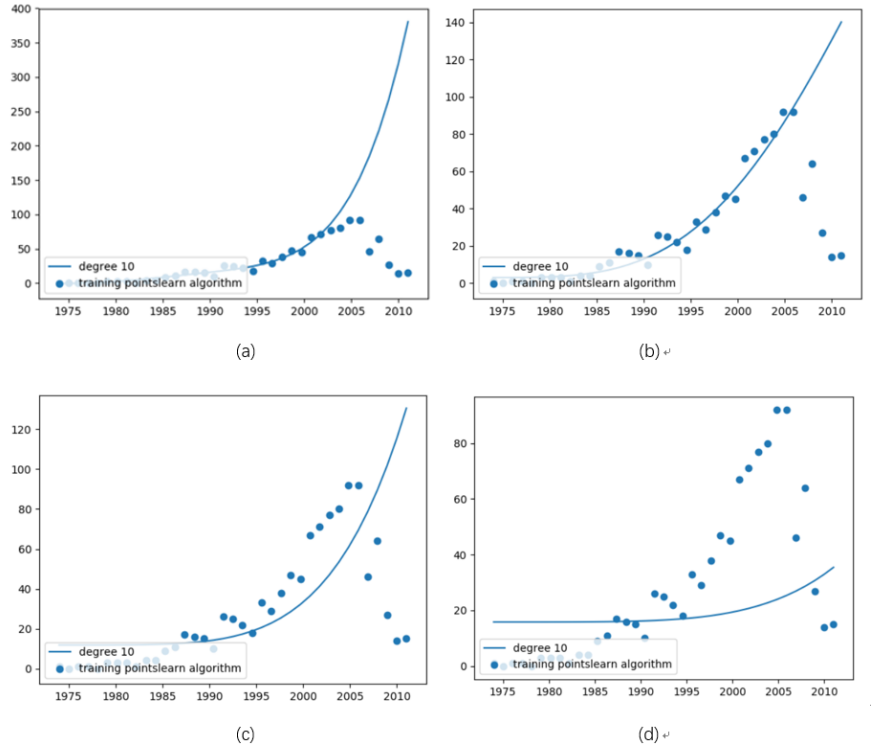

图 3-6 alpha 取值对发展趋势预测的影响

3.5 距离训练速度分析

算法的模型训练部分对于预测结果影响至关重要，在训练集较小时，距离训练速度（即学习因子α_1）决定了模型间关联度是否学习充分。图 3-7、图 3-8、图 3-9 反映了α_1分别为 0.9、0.8、0.7 时的预测结果。对比图 3-7 和图 3-9 可以明显发现图 12 中更容易有多分支聚合为一，这是因为图 3-9 对应的学习速度较快，模型训练后术语关联程度更强。但三个图

中效果较好的为图 3-8，所以相较起来可能α_1取值 0.8 更为合适。

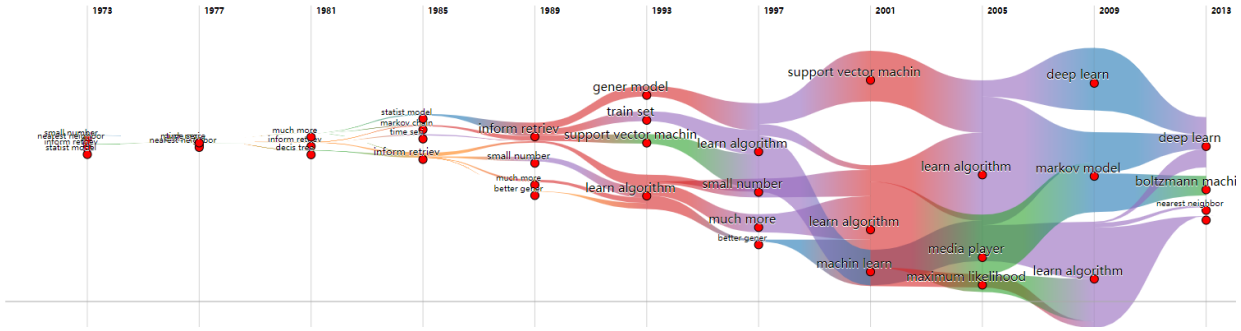

图 3-7 alpha 取 0.9 时的科技演化趋势预测结果

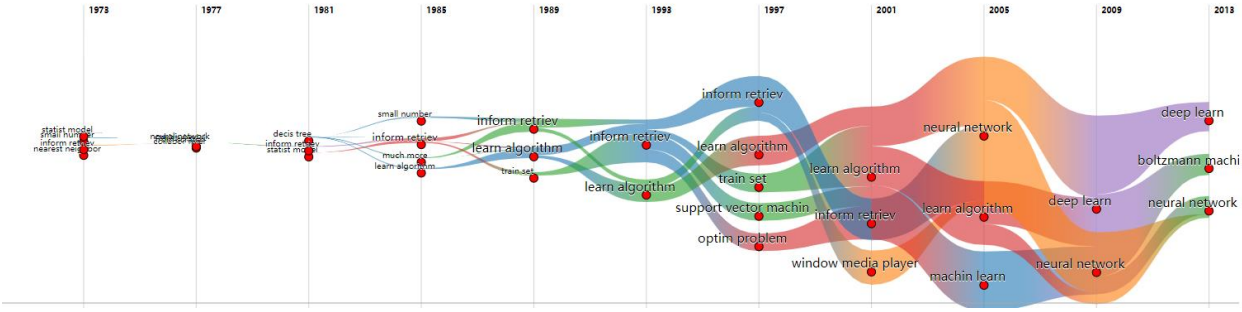

图 3-8 alpha 取 0.8 时的科技演化趋势预测结果

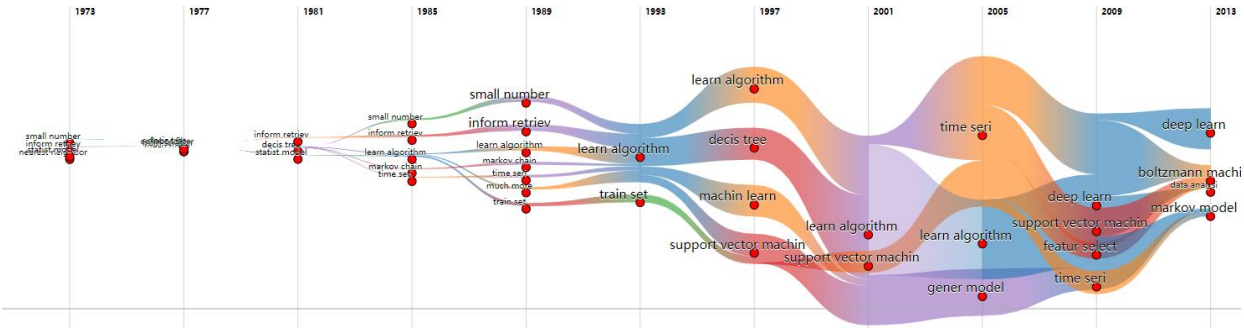

图 3-9 alpha 取 0.7 时的科技演化趋势预测结果

3.6 分支合并阈值分析

在术语分支重新合并为新节点时，决策依据为合并前后的距离比值是否超过阈值，不超过则合并，反之不合并。阈值较大，则合并条件宽松，更容易产生分支汇合；阈值较小，则合并条件严格，分支汇合较困难。图 3-10 和图 3-11 分别是阈值为 10 和 20 时的预测结果，对比两图 2001-2013 年的演化情况，显然阈值为 10 时产生了更多新节点，即说明分支的汇合较少。

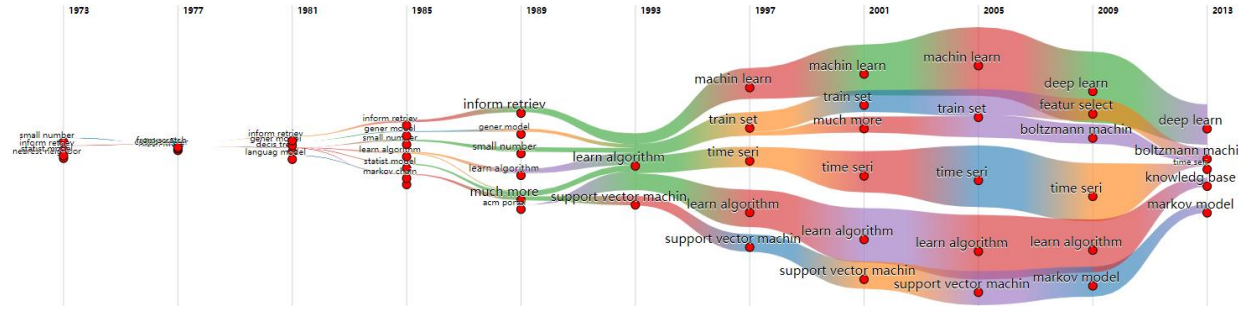

图 3-10 阈值为 10 时的科技演化趋势预测结果

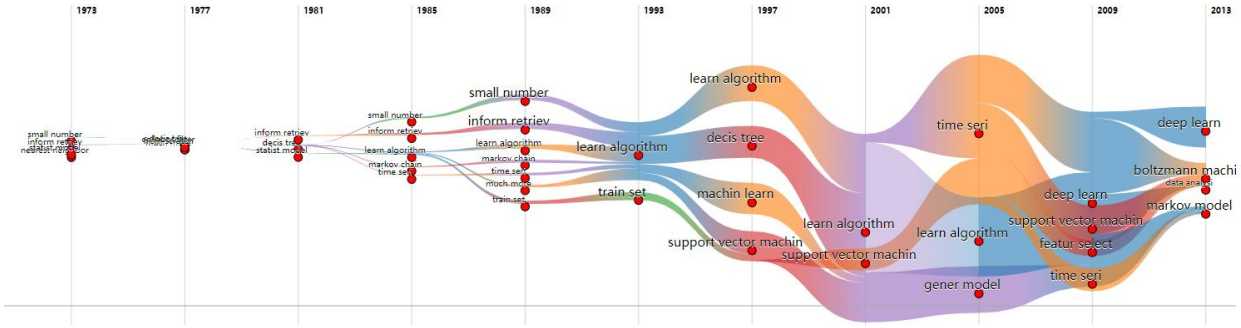

图 3-11 阈值为 20 时的科技演化趋势预测结果

第 4 章　总结及改进

本章首先对算法的研究工作做一下总结，然后指出存在的缺陷，最后对未来的优化思路进行展望。

4.1 研究总结

本次研究采用了 K-mean++、多项式岭回归、遗传算法思想等数据挖掘和机器学习的方法建立了科技演化趋势预测模型进行训练和预测，并使用 D3 可视化库的 Sankey 力学图对结果进行了展示，通过对结果和参数的分析验证了模型的有效性。

用于研究的科技大数据来自清华知识工程研究室研发的学术搜索引擎 Arnetminer，数据经过采集和预处理后在一定程度上反映了科技领域的热度和关联情况。

算法使用多项式岭回归方程实现了对热度数值的预测，正则化处理抑制了学习模型的过拟合现象；每次演化过程主要分为术语分支的分叉与汇合两个步骤。距离概率和近邻相似的思想实现了分叉数预测模块，简便而高效；使用 K-mean ++算法实现了术语的汇合，可自适应数据规模。分叉和汇合两个步骤在借鉴遗传算法的框架下通过术语重组与变异过程完成迭代。

最终算法的预测结果在 Sankey 图中得以显示，能量均衡的特性使得预测结果具有美感，便于使用者直观洞察科技发展的趋势。

算法的性能在实验中得到了验证和评估。以深度学习领域为例，算法较好地对 1973-2013 年的科技演化趋势进行了预测，受限于模型的训练机制，算法在对早期科技的预测中表现不如对近期科技的预测；调整正则化项 alpha 的数值会同时影响模型在训练集和验证集上的准确度；训练速度的改变也对模型精度有影响，但不是单调变化；分支阈值对可视化效果有显著影响，反映了结果展示的精度。

4.2 算法缺陷

目前算法存在如下已知缺陷：

(1) 我们的预测算法将术语演化看作是重新分类再聚类并伴有变异的过程，但这不一定完全符合实际的演化规律，更换数据集后性能难以保证;

(2) 而且它虽然从时间变迁的角度学习了一定的技术演化规律，但并未考虑技术发展趋势和演化拓扑之间的联系;

(3) 也未考虑领域专家的社交活动和研究兴趣对演化过程产生的影响;

(4) 且我们目前使用的数据集也很有限，数据总量和属性种类都难以训练出有效且强健的预测模型;

(5) 另外训练程度无法根据数据的时序特征自适应调整，可能在预测时因利用了不合时宜的知识与经验而出现异常结果。

4.3 优化思路

针对目前实验中出现的各种问题制定的后续计划如下：

(1) 用 Aminer 中更完整的数据集继续开展后续研究;

(2) 从数据集中统计科技跃迁点(burst)数据，研究跃迁点出现的时序特征;

(3) 研究知识图谱的对科技演化的影响，分析演化时对应知识图谱的上下位结构特征[30,31]。

(4) 研究领域专家的社交网络和研究方向对科技跃进的影响，比如比较专家和他社交圈内其他学者的论文、词频的变化情况，如果是对于某个领域，将每个专家社交圈看做是一个大输入节点，它的输入由该社交圈内所有专家节点决定，然后输出对应到领域知识图谱的发展，再根据相应的时序特征和结构特征进行预测[32]。

(5) 考虑使用深度森林模型[33]优化自适应训练的能力，用 RDA 等在线学习方法动态调整学习梯度，并研究利用其他集成学习与深度学习算法进一步提高预测精度的可行性[34]。

参考文献

[1] 周志华. 机器学习与数据挖掘[M]. 电子工业出版社, 2004.

[2] Xiong R, Donath J. PeopleGarden: creating data portraits for users[C]. ACM Symposium on User Interface Software and Technology, 1999:37-44.

[3] Umanol M, Okamoto H, Hatono I, et al. Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems[C]. IEEE World Congress on Computational Intelligence. Proceedings of the Third IEEE Conference on. IEEE, 2002:2113-2118 vol.3.

[4] Joachims T. Making Large-Scale SVM Learning Practical[J]. Advances in Kernel Methods-Support Vector Learning, 1998, 8(3):499-526.

[5] Wu X, Kumar V, Quinlan J R, et al. Top 10 algorithms in data mining[J]. Knowledge & Information Systems, 2008, 14(1):1-37.

[6] Xu R. Survey of clustering algorithms[M]. IEEE Press, 2005.

[7] Kirubha V, Priya S M. Survey on Data Mining Algorithms in Disease Prediction[J]. International Journal of Emerging Trends & Technology in Computer Science, 2016, 38(3):124-128.

[8] Tsai H H. Global data mining: An empirical study of current trends, future forecasts and technology diffusions[J]. Expert Systems with Applications, 2012, 39(9):8172-8181.

[9] Prabhakaran V, Hamilton W L, Dan M F, et al. Predicting the Rise and Fall of Scientific Topics from Trends in their Rhetorical Framing[C]. Meeting of the Association for Computational Linguistics. 2016:1170-1180.

[10] Hope T, Chan J, Kittur A, et al. Accelerating Innovation Through Analogy Mining[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017:235-243.

[11] Tsai H H. Knowledge management vs. data mining: Research trend, forecast and citation approach[J]. Expert Systems with Applications, 2013, 40(8):3160-3173.

[12] Tang J, Zhang J, Zhang D, et al. ArnetMiner: an expertise oriented search system for web community[C]. International Conference on Semantic Web Challenge, 2007:1-8.

[13] Herman I, Melançon G, Marshall M S. Graph Visualization and Navigation in Information Visualization: A Survey[J]. IEEE Transactions on Visualization & Computer Graphics, 2000, 6(1):24-43.

[14] Liu S, Dan M, Wang B, et al. Visualizing High-Dimensional Data: Advances in the Past Decade[C]. Eurographics Conference on Visualization. 2015.

[15] 唐杰、张静、张宇韬.“AMiner 背后的技术细节与挑战”, https://www.csdn.net/article/2015-06-11/2824931, 2015-06-11/2018-5-22

[16] Seop P J, Hong S G, Jongweon K. A Study on Science Technology Trend and Prediction Using Topic Modeling[J]. Journal of the Korea Industrial Information Systems Research, 2017, 22.

[17] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. Computer Science, 2014.

[18] Yang A Y, Wright J, Ma Y, et al. Unsupervised segmentation of natural images via lossy data compression[J]. Computer Vision & Image Understanding, 2008, 110(2):212-225.

[19] “[机器学习] 正则化的线性回归——岭回归和 Lasso 回归”, https://www.cnblogs.com/Belter/p/8536939.html, 2018-03-16/2018-5-22

[20] Arthur D, Vassilvitskii S. k-means++:the advantages of careful seeding[C]. Society for Industrial and Applied Mathematics, 2007:1027-1035.

[21] Schmidt M. The Sankey Diagram in Energy and Material Flow Management[J]. Journal of Industrial Ecology, 2008, 12(1):82-94.

[22] Olivier Catherin,“Sankey diagram made of dynamically generated polygons”https://community.tableau.com/thread/152115, 2014-11-22/2018-5-22

[23] Mcdonald G C. Ridge regression[J]. Wiley Interdisciplinary Reviews Computational Statistics, 2010, 1(1):93-100.

[24] Xu R. Survey of clustering algorithms[M]. IEEE Press, 2005.

[25] Goldberg D E. Genetic Algorithm in Search Optimization and Machine Learning[J]. Addison Wesley, 1989, xiii(7):2104–2116.

[26] Weiss S M, Kapouleas I. An empirical comparison of pattern recognition, neural nets, and machine learning classification methods[C]. Proc. of International Joint Conference of Artificial Intelligence. 1989:781-787.

[27] Developer H. PyCharm 3.1 kommt mit neuem Interface[J]. Heise Zeitschriften Verlag, 2014.

[28] Miller F P, Vandome A F, Mcbrewster J, et al. Data Flow Diagram[J]. Alphascript Publishing, 2010, 4(1):66--78.

[29] Liu, Yamei. Overfitting and forecasting: linear versus non-linear time series models[J]. Isu General Staff Papers, 2000.

[30] Pujara J, Miao H, Getoor L, et al. Knowledge Graph Identification[C]. International Semantic Web

Conference. Springer-Verlag New York, Inc. 2013:542-557.

[31] Wang Z, Zhang J, Feng J, et al. Knowledge Graph Embedding by Translating on Hyperplanes[J]. AAAI - Association for the Advancement of Artificial Intelligence, 2014.

[32] Plantié M, Crampes M. Survey on Social Community Detection[J]. 2013:65-85.

[33] Zhou Z H, Feng J. Deep Forest: Towards An Alternative to Deep Neural Networks[C]. Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017:3553-3559.

[34] Dietterich T G. Ensemble Methods in Machine Learning[M]. Springer Berlin Heidelberg, 2000:1-15.