基于自适应隐马尔可夫模型的石油领域文档分词

宫法明 朱朋海
(中国石油大学(华东)计算机与通信工程学院 山东 青岛 266580)

摘 要 中文分词技术是把没有分割标志的汉字串转换为符合语言应用特点的词串的过程,是构建石油领域本体的第一步。石油领域的文档有其独有的特点,分词更加困难,目前仍然没有有效的分词算法。通过引入术语集,在隐马尔可夫分词模型的基础上,提出了一种基于自适应隐马尔可夫模型的分词算法。该算法以自适应隐马尔可夫模型为基础,结合领域词典和互信息,以语义约束和词义约束校准分词,实现对石油领域专业术语和组合词的精确识别。通过与中科院的 NLPIR 汉语分词系统进行对比,证明了所提算法进行分词时的准确率和召回率有显著提高。

关键词 中文分词,隐马尔可夫模型,组合词,石油
中图法分类号 TP391 **文献标识码** A

Word Segmentation Based on Adaptive Hidden Markov Model in Oilfield

GONG Fa-ming ZHU Peng-hai
(College of Computer & Communication Engineering, China University of Petroleum, Qingdao, Shandong 266580, China)

Abstract The Chinese word segmentation is the first step in constructing the petroleum field ontology. Documents in petroleum field have their own unique characteristics which make word segmentation more complex. Until now, there is no effective word segmentation algorithm, especially for Chinese characters. Based on the hidden Markovian model, an adaptive hidden Markovian word segmentation model was proposed in this paper, which combines the domain-knowledge dictionary and user-defined information, by introducing the terminology set. The proposed algorithm calibrates word segmentation under semantic constraints and word meaning constraints, and can identify professional terms and character combinations in the field of petroleum accurately. It is also proved that the proposed algorithm achieves remarkable improvements in both accuracy and recall rate in word segmentation, compared to the NLPIR Chinese word segmentation system invented by Chinese Academy of Science.

Keywords Chinese word segmentation, Hidden Markov model, Combined character, Petroleum

1 引言

词是“最小的能独立运行的语言单位”[1]。由于中文具有连续书写的特点,如果不进行分词,计算机很难理解文本包含的信息,无法进行本体构建。石油文档具有独特的领域特点,比如专业术语和组合词较多、专业术语更新快等,针对石油领域文档的特点进行准确分词是进行石油领域本体构建的关键一步。

目前,常用的中文分词算法有:基于统计的分词算法[2-3]、基于规则的分词算法[4-5]、基于两者相结合的分词算法和基于语义的分词算法。

Aken[6]提出了一种分词方法,该方法利用字母之间的统计关系推断字边界的位置,然后通过计算相邻字母之间的贡献频率进行分词。Tohit 等[7]提出了一种基于关联规则的分词方法,该方法首先从语料库中提取上下文信息,然后利用关联规则组合进行分词。Hongbo[8]提出了一种字典和统计分析相结合的分词方法,该算法解决了未登录词和歧义词识别两个问题。基于统计的分词算法没有考虑语义信息,不能进行准确分词。基于语义的分词算法虽然考虑了语义信息,能够进行语法分析,但并没有领域适应性,不能针对领域进行分词。基于统计和规则相结合的分词算法虽然克服了以上两种方法的缺点,但仍存在分词准确性不高以及不能对一些新兴的领域专业词语进行准确分词的缺点。

为了克服以上算法的缺点,本文提出了一种基于自适应隐马尔可夫的分词模型。该模型克服了隐马尔可夫分词模型假设适用范围小的缺点,考虑了具体词汇信息对句子结构的重要作用,结合一阶和二阶隐马尔可夫模型,自适应地选取合适的隐马尔可夫模型进行分词。具体地,首先以自适应隐马尔可夫模型为基础,进行初步分词;然后利用石油领域词典进行词语匹配,初步识别石油领域专业词;接着计算相邻词之间的互信息,通过与阈值比较,初步识别出石油领域组合词;最后利用语义约束矩阵和语法约束矩阵进行约束,通过约束结果反馈校准分词结果,实现对石油领域专业术语和组合词的精确识别。与 NLPIR 系统的比较结果表明,文中所提模型的分词效果更好。本文的贡献如下所示:

1)本文提出的方法在石油领域文档分词方面属首次提出;

2)本文提出的方法能够精确识别石油领域的专业词和组

本文受科技部创新方法工作:大数据环境下的油气开采创新方法研究与应用示范(2015IM010300)资助。
宫法明(1970—),男,硕士,副教授,主要研究方向为计算机图像图像处理、大数据智能处理与云计算,E-mail:z15070507@s.upc.edu.cn;朱朋海(1992—),男,硕士生,主要研究方向为计算机图像处理、自然语言处理。

合词，在石油领域取得的分词结果是目前最好的；

3)本文为专业领域文档分词提供了一个分词框架。

本文第 2 节介绍相关理论与方法；第 3 节介绍基于自适应隐马尔可夫模型的分词算法的实现；第 4 节为实验结果及分析；最后对本文工作进行总结，并指出未来的工作方向。

2 相关工作

基于语料库的统计语言学分词方法，通过对语料库中的文本进行统计分析，获取该类文本的某些整体特征或规律，并利用这些统计现象、规律对目标文本进行分词。其中最重要的方法是隐马尔可夫模型(HMM)，该算法有研究透彻、算法成熟、分词效果好、易于训练等优点，目前已被成功应用于语音识别、行为识别、文字识别以及故障诊断等领域；但是，该算法也存在独立性假设适用范围小、考虑不全面等缺点。PB-heganan 等[9]提出了一种新颖的隐马尔可夫模型，其结合了非字典和字典的方法。该方法先将段落分成单个音节，利用隐马尔可夫模型进行合并音节，然后利用字典确定和验证分词结果。该方法的缺点是仅利用领域字典来验证分词结果，不能识别词典中没有的新兴专业术语，不能识别未登录词。Li 等[10]将 M3N(Max Margin Markov Networks)网络模型和结构模型应用于中文分词。该方法基于一定的训练和测试语料库，取得了不错的精度。但该方法需要大量的石油资料，由于石油是国家能源行业，对石油资料具有一定的保密机制，获取语料库困难，其并不适用于石油领域的文档分词。Pang 等[11]提出了一种改进的分词方法，该方法通过使用双向马尔可夫链，应用一种改进的正向最大匹配算法和词频统计算法，对给定文本进行迅速且准确的分词。但是，该方法没有考虑上下文信息，不包含语义信息，分词效果也不理想。Kwon 等[12]利用马尔可夫链生成一个句子，然后使用最大似然估计进行句法分析进而分词。但是，此方法没有校准验证的处理，缺乏足够的容错性。

尽管上述方法能够取得比较好的分词结果，但它们都没有领域针对性，对领域专业词语的识别效果并不好。本文提出一种基于自适应隐马尔可夫模型的算法，该算法能够很好地识别领域专业词语，特别适用于石油领域这种组合词多的领域。

3 基于自适应隐马尔可夫的分词模型

3.1 隐马尔可夫模型的简介

3.1.1 马尔可夫(Markov)过程的定义

一般地，考虑随机过程$\{X_n \mid n=1,2,\cdots\}$，若 $X_n=i$，就说过程在 n 时刻处于状态 i，假设每当过程处于状态 i，则过程在下一时刻处于状态 j 的概率 P_{ij} 就是一个定值，即对于 $\forall n \geqslant 1$ 有：

$$\begin{aligned} P_{ij} &= P(X_{n+1}=j \mid X_n=i, X_{n-1}=i_{n-1}, \cdots, X_1=i_1) \\ &= P(X_{n+1}=j \mid X_n=i) \end{aligned} \tag{1}$$

这样的随机过程被称为 Markov 链(给定过去状态 $X_1,\cdots,X_{n-1}$和现在状态 X_n，将来状态 X_{n+1}的条件独立分布于过去状态，只依赖于现在的状态——这就是 Markov 性)。

3.1.2 隐马尔可夫模型(HMM)

对于马尔可夫模型而言，每个状态都是决定性地对应于一个可观察的物理事件，因此其状态的输出是有规律的。然而，这种模型的限制条件过于严格，在许多实际问题中无法应用。于是，人们将这种模型加以推广，提出了隐马尔可夫模型(HMM)。隐马尔可夫过程是一种双重随机过程，即观察事件是依存于状态的概率数，这是在 HMM 中的一个基本随机过程；另一个随机过程为状态转移随机过程，这一过程是隐藏的，不能直接观察到，只有通过生成观察序列的另外一个概率过程才能间接观察到。隐马尔可夫模型的示意图如图 1 所示，该图分为上、下两行，上面一行是一个马尔可夫转移过程，下面一行是输出，即我们可以观察到的值。在分词模型中，输出序列 $O_1O_2\cdots O_T$ 就是待分的字符串，隐藏状态 $X_1\ X_2\cdots X_T$ 就是分好的单词，通过训练数据调整参数，找到最佳序列 $X_1X_2\cdots X_T$ 使输出序列 $O_1O_2\cdots O_T$ 出现的概率最大，则最佳序列 $X_1X_2\cdots X_T$ 就是分好的单词序列。

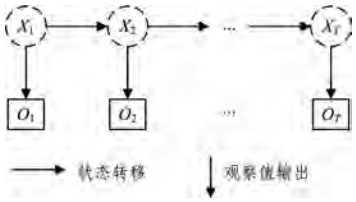

图 1 隐马尔可夫模型示意图

3.1.3 自适应隐马尔可夫模型

首先定义一个术语集，术语集是一系列具有代表性的专业术语集合。对石油领域中的各个子学科从 1 到 i 进行编号，统计每个子学科最常用的 n 个专业术语，如石油勘探开发领域最常用的术语有勘探、测井、储量、井、压差等。然后将各个子学科的代表性专业术语构成一个集合，编号为 i 的子学科的术语集为 Q_i，则总的术语集为 $Q=\bigcup_{i=1}^{N} Q_i$。自适应隐马尔可夫模型的主要思想为：首先根据术语集预先判断待分文档包含专业术语的多少，然后与阈值进行比较，若所包含专业术语的数量大于阈值，则说明该段落包含的专业术语较多，应该进行准确分词，此时调用二阶隐马尔可夫模型进行分词；否则进行快速分词，调用一阶隐马尔可夫模型进行分词。

首先根据输入的文档判断本文档属于哪一个子学科领域，假设输入文档为 D，所属的子学科编号为 m，提取术语集 $Q=Q_m$。假设 Q_m 中含有的代表性术语个数为 n，遍历 Q_m 搜索文档 D 中含有的代表性专业术语数量，$X=[x_1,x_2,\cdots,x_n]$，x_i 表示文档 D 中含有的编号为 i 的代表性术语的数量，则文档 D 含有的代表性专业术语的数量为：

$$num_D=\sum_{i=1}^{n} x_i \tag{2}$$

根据文档字数 num 以及比例系数 α 确定专业术语数量阈值 $s=num\cdot\alpha$。

$$model=\begin{cases} \text{一阶隐马尔可夫模型}, & num_D<s \\ \text{二阶隐马尔可夫模型}, & num_D>s \end{cases} \tag{3}$$

3.2 领域词典和互信息算法的描述

基于词典的方法，按照一定的策略将待分析的汉字串与词典中的词条进行匹配，若在词典中找到某字符串，则匹配成功。按照扫描方向的不同，可以分为正向匹配、逆向匹配；按照长度的不同，可以分为最大匹配和最小匹配。

互信息算法[13]的主要思想是：对于汉字 x 和汉字 y，可以用公式计算出它们的互信息值 $P(x,y)$，然后用 $P(x,y)$的大小来判断汉字 x 和汉字 y 之间的紧密程度，以此来判断 x 与 y 是否能组合成词。互信息的计算公式如下：

$$P(x,y)=\log_2 \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

其中，$p(x,y)$为汉字 x 和 y 共现的频率，$p(x)$和 $p(y)$分别为

汉字 x 和 y 出现的频率。互信息值越大，说明两个字结合的程度越高；互信息值越小，则代表结合的程度越低。

3.3　语法语义约束矩阵算法的描述

矩阵约束法的基本思想是：先建立一个语法约束矩阵和语义约束矩阵，其中元素分别标明具有词性的词和具有另一词性的词相邻是否符合语法规则，属于某语义类的词和属于另一词义类的词相邻是否符合逻辑，计算机在分词时以之约束分词结果。

设 $TCN=(T,C)$ 是字符串函数，其中，$T=\{g_0,g_1,\cdots,g_{n-1}\}$，则语法约束矩阵 $P=(p_{ij})$ 是一个 n 阶方阵。

$$P_{ij}=\begin{cases}1, & (g_i,g_j)\in C\\ 0, & \text{其他}\end{cases} \tag{5}$$

其中，C 为《现代汉语语法信息词典》。由定义可知，约束矩阵是一个布尔矩阵，表达了约束系统中的约束情况。对于矩阵中的元素 p_{ij}，如果第 i 行对应的单词词性有约束指向第 j 列单词的词性，那么该元素值为 1，否则为 0。同理，语义约束矩阵采用同样的方法，语义规则采用《现代汉语语义词典》。

3.4　基于自适应隐马尔可夫模型分词算法的实现

本算法基于自适应隐马尔可夫模型，结合领域词典、互信息和语义语法约束矩阵进行校准，主要分为两部分，即预处理阶段和分词阶段。算法的流程图如图 2 所示。

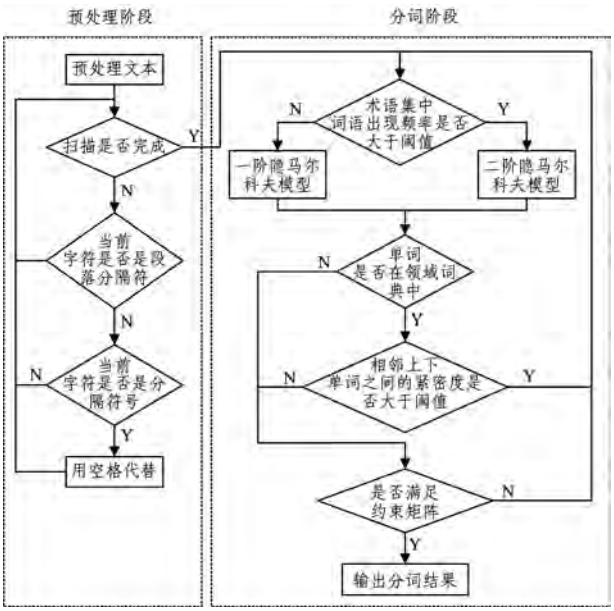

图 2　算法流程图

3.4.1　数据预处理阶段

在分词计算前，需要对待分词文档进行预处理，即利用一些具有分隔作用的符号将文档切分成较短的句子或字符串，以减少匹配次数，提高分词效率，降低分词难度。先按照段落分隔符（回车符）进行段落切分，将文档分成多个段落，然后利用标点符号、数字、英文以及构词能力差的单字等分隔符再将段落细分成较短的句子或字符串。

3.4.2　分词阶段

假设待分文档为 D，经过预处理后有 r 个段落和 s 个最小字符串。字串 $Y_i=y_{i1},y_{i2},\cdots,y_{im}$，$y_{ij}$ 表示一个单字。调用自适应隐马尔可夫模型进行分词，然后判断分好的单词是否在领域词典中，若在领域词典中，则判断与该单词相邻的上下单词之间的紧密度，以及其是否需要进行重新分词，否则代入约束矩阵再进行语法和语义约束校准，最后输出分词结果。具体分词步骤如下：

1）输入字串 Y_i，调用自适应隐马尔可夫模型进行分词，计算该句所在段落包含术语集中专业术语的数量，如果大于某一阈值，则转入步骤 2），否则转入步骤 3）。

2）调用二阶隐马尔可夫模型进行分词，将字串 Y_i 分成单词序列 $X_i=x_{i1},x_{i2},\cdots,x_{in}$，转入步骤 4）。

3）调用一阶隐马尔可夫模型进行分词，将字串 Y_i 分成单词序列 $X_i=x_{i1},x_{i2},\cdots,x_{in}$，转入步骤 4）。

4）遍历 X_i，判断 x_{ij}（$j=1,\cdots,n$）是否在领域词典中，若在领域词典中，则转入步骤 5），否则转入步骤 6）。

5）查找单词 x_{ij} 的相邻上下单词，并记录单词 x_{ij}，$x_{i,j-1}$，$x_{i,j+1}$ 和句子编号 i 到数组 S，转入步骤 8）。

6）将单词代入约束矩阵进行验证，若满足约束矩阵，则转入步骤 8），否则记录并剔除该分词方式，转入步骤 1）。

7）判断组合词频率是否大于阈值，如大于阈值，则将编号为 i 的句子作为字符串输入，转入步骤 1），否则转入步骤 6）。

8）判断数组 S 是否完全遍历，若遍历结束，则结束分词，并输出分词结果，否则遍历整个文档统计组合词 $x_{i,j-1}x_{ij}$，$x_{i,j}x_{i,j+1}$ 的频率，转入步骤 7）。

3.5　评价方法

本文使用准确率、召回率和 F 值[14]对本文提出的方法进行评价。其中，准确率是分词正确的单词数与分词结果的总数的比率，衡量的是分词结果的查准率；召回率是分词正确的单词数与实际单词总数的比率，衡量的是分词结果的查全率。两者取值在 0 和 1 之间，数值越接近 1，准确率或召回率就越高，具体定义如下：

$$\text{准确率}(P)=\frac{\text{分词结果中切分正确的总词数}}{\text{分词结果中的总词数}}\times 100\% \tag{6}$$

$$\text{召回率}(R)=\frac{\text{分词结果中切分正确的总词数}}{\text{标准结果中的总词数}}\times 100\% \tag{7}$$

$$F=\frac{2RP}{R+P} \tag{8}$$

4　实验结果及分析

本文利用中国知网中的石油领域的 100 篇文档进行实验，这些文档涵盖了油气地质勘探、油气田开发与开采等石油领域的 10 个子学科，每个子学科 10 篇文档。与 NLPIR 汉语分词系统的分词结果进行数据对比，对本文提出的算法进行测试。NLPIR 汉语分词系统是我国现有的已经投入市场应用的较为成熟的自然语言处理平台，该系统涵盖了中文分词、词频统计、智能分析、摘要提取、关键词获取及新词获取等多方面的功能。本文以 NLPIR 汉语分词系统的分词结果为标准，对所提出的算法进行对比验证，以更直观地体现本文算法的有效程度。这里以测试文档中的一段为例，原始数据如图 3 所示。经过预处理之后，将标点符号、语气词去掉，结果如图 4 所示。NLPIR 系统进行分词之后的数据如图 5 所示。本文算法分词之后的数据如图 6 所示。

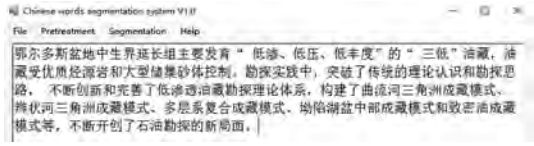

图 3　原始文档

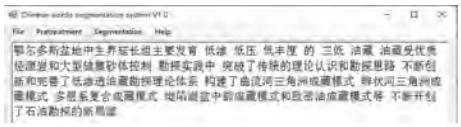

图 4 预处理之后的结果

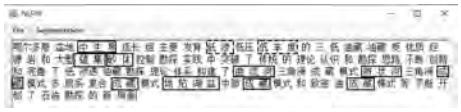

图 5 NLPIR 系统的分词结果

图 6 自适应隐马尔可夫分词模型的分词结果

预先统计这段中包含的专业术语。由于该段文字包含的专业术语较多，因此使用二阶隐马尔可夫模型进行分词，其中相邻词语由空格分开。从上面的分词结果来看，NLPIR 汉语分词系统对组合词和石油领域专业术语的处理效果并不理想。其中，“储集”“砂体”“成藏”等词语是石油领域的专业词语，应该分成一个词语，但 NLPIR 系统并没有识别出此类专业术语，也没能对其进行正确分词，并且“中生界”“曲流河”“辫状河”“坳陷盆地”等石油领域组合词也没有被正确分词；而本文的算法能准确无误地将其提取出来。但是，也有 NLPIR 系统和本文算法都没有正确识别出来的词语，如“低渗”“低丰度”等。实线标记的是 NLPIR 系统没有正确分词而本文算法能正确分词的部分，虚线标记的是 NLPIR 系统和本文算法都没有正确分词的部分。两种算法的准确率、召回率和 F 值比较如表 4 所列。

表 4 准确率、召回率和 F 值对比

（单位：%）

	准确率	召回率	F 值
NLPIR 系统	80.34	88.76	84.34
本文算法	97.50	98.73	98.11

本文选取 10 个石油子学科进行实验，计算它的平均 F 值，实验结果如图 7 所示。从图 7 中可以看出，NLPIR 系统的 F 值分布在 81%～90%之间，而本文方法的 F 值均高于 90%。本文方法的 F 值明显高于 NLPIR 系统的 F 值，其中有的还达到了 99%，从而证明了本文分词算法对石油领域的分词效果很好。

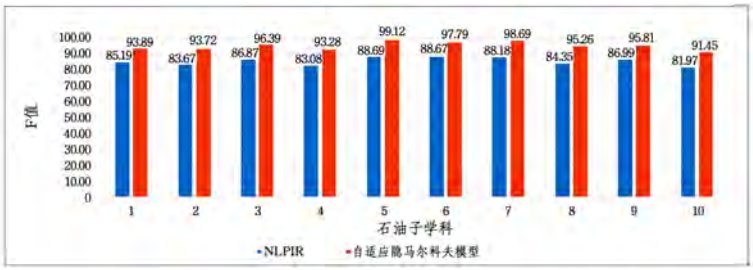

图 7 10 个石油子学科的 F 值对比

结束语 本文在隐马尔可夫分词模型的基础上，提出了一种基于自适应隐马尔可夫模型的分词算法。预先统计代表性专业术语，根据专业术语的数量，采用不同的隐马尔可夫分词模型，利用专业词典提取专业术语，利用互信息提取组合词，引入语义约束矩阵和语法约束矩阵用于辅助分词。本算法能够对专业术语和组合词进行精确分词，更加准确地提取与专业术语有关的组合词，极大地提高了石油领域分词的准确率。但是本算法处理步骤较多，因此运行速度较慢。因此，提高本算法的运行速度是下一步的研究重点。

参 考 文 献

[1] 来斯惟，徐立恒，陈玉博，等. 基于表示学习的中文分词算法探索[J]. 中文信息学报，2013，27(5)：8-14.

[2] JOHNSON E K，TYLER M D. Testing the limits of statistical learning for word segmentation[J]. Developmental Science，2010，13(2)：339-345.

[3] FU G，LUKE K K. A two-stage statistical word segmentation system for Chinese[C]// Sighan Workshop on Chinese Language Processing. Association for Computational Linguistics，2003：156-159.

[4] WANG J. A Rule-based Methodology and Feature-based Methodology for Effect Relation Extraction in Chinese Unstructured Text[D]. Dydney：University of Sydney，2015.

[5] SILVA D C，BRAGA D，RESENDE F G V J. A rule-based method for homograph disambiguation in brazilian portuguese text-to-speech systems[J]. Journal of Communication and Information Systems，2015，27(1).

[6] AKEN J R V. A statistical learning algorithm for word segmentation[J/OL]. Computer Science，https//arixv. org/ftp/arxiv/papers/1105/1105. 6162. pdf.

[7] TOHTI T，MUSAJAN W，HAMDULLA A. Unsupervised Learning and Linguistic Rule Based Algorithm for Uyghur Word Segmentation[J]. Journal of Multimedia，2014，9(5)：627-634.

[8] HONGBO POSTGRADUATE L I. Dictionary and Statistical Analysis Combined Algorithm for Chinese Word Segmentation[J]. Journal of Wuhan University of Technology，2010(12)：907-909.

[9] BHEGANAN P，NAYAK R，XU Y. Thai Word Segmentation with Hidden Markov Model and Decision Tree[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg，2009：74-85.

[10] 李月伦，常宝宝. 基于最大间隔马尔可夫网模模型的汉语分词方法[J]. 中文信息学报，2010，24(1)：8-14.

[11] PANG B，SHI H. Research on Improved Algorithm for Chinese Word Segmentation Based on Markov Chain[C]// International Conference on Information Assurance and Security. IEEE，2009：236-238.

[12] OH-WOOK K. Korean Word Segmentation and Compound-noun Decomposition Using Markov Chain and Syllable N-gram[J]. Journal of the Acoustical Society of Korea，2002，21(3)：274-284.

[13] 刁毓. 基于本体的中文分词算法的研究与实现[D]. 曲阜：曲阜师范大学，2012.

[14] 李良洁. 基于统计和语义信息的中文分词算法研究[D]. 青岛：青岛科技大学，2015.