问答系统中问题模式分类与相似度计算方法

周建政[1],谌志群[2],李 治[1],王荣波[2],冯 凯[2]
ZHOU Jianzheng[1], CHEN Zhiqun[2], LI Zhi[1], WANG Rongbo[2], FENG Kai[2]

1.天格科技(杭州)有限公司,杭州 310005
2.杭州电子科技大学 认知与智能计算研究所,杭州 310018
1.Tiange Technology(Hangzhou) Limited Company, Hangzhou 310005, China
2.Institute of Cognitive and Intelligent Computing, Hangzhou Dianzi University, Hangzhou 310018, China

ZHOU Jianzheng, CHEN Zhiqun, LI Zhi, et al. Methods of questions pattern classification and similarity measure for question answering system. Computer Engineering and Applications, 2014, 50(1):116-120.

Abstract: At present, question answering system based on Frequently Asked Questions(FAQ) for restricted domains is a research focus in the field of natural language processing due to its practicality. The similarity measure between questions plays a very important role in one question answering system. The traditional questions similarity measure technologies have unsatisfactory effects for those questions with context information. A rule-based question pattern classification algorithm is proposed for dividing all questions into two categories: Simple Mode Questions(SMQs) and Context Mode Questions(CMQs). Then, a similarity measure method for CMQs is presented in which the similarities between context information and that between questions are combined together. The experimental results show that both precision and recall rate of the proposed question pattern classification method exceed 90%, and the accuracy of similarity measure for context mode questions reaches 74.3% with lower time complexity.
Key words: similarity measure; pattern classification; context information; question answering system

摘 要:基于FAQ库的限定域自动问答系统由于更具实用性而成为自然语言处理领域的研究热点,而问题之间的相似度计算是其中最关键的技术。现有的问句相似度计算技术在处理带有上下文情景描述的问题时效果较差。针对现有技术存在的问题,提出将用户问题分为简洁模式问题(SMQs)和情景模式问题(CMQs),并提出了基于规则的问题模式分类算法。在此基础上,进一步提出了综合考察情景相似度和问句相似度的情景模式问题(CMQs)相似度计算方法。实验结果表明,问题模式分类算法取得了90%以上的准确率和召回率,情景模式问题相似度计算方法在时间复杂度较低的情况下也取得了74.3%的正确率。
关键词:相似度计算;模式分类;上下文信息;问答系统
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1305-0280

1 引言

自动问答系统(Question Answering System,QA系统)是一种新的信息检索系统,也是一种新的人机交互方式[1]。用户可以用自然语言向QA系统提问,QA系统能自动获取知识并返回所需答案。QA系统可以定义为一个三元组$<Q,A,F>$,其中Q为问题集,A为答案集,F为处理方法[2]。对于任一问题$q_i(q_i \in Q)$,有$f:q_i \to a_j$,其中$a_j \in A$,$f \in F$。即通过某种处理或计算,寻找与问题q_i对应的答案a_j,也就是寻找q_i在F上的映射值a_j。

可以根据不同的标准对QA系统进行分类,目前比较常见的分类方式是根据系统结构将它分为四类[3]:聊天机器人、基于FAQ(Frequently Asked Questions,常问问

基金项目:杭州市科技发展计划重大科技创新专项(No.20122511A18);国家自然科学基金青年项目(No.61202281)。
作者简介:周建政(1963—),男,工程师,研究领域为智能信息处理;谌志群(1973—),男,副教授,研究领域为中文信息处理;李治(1977—),男,工程师,研究领域为智能信息处理;王荣波(1978—),男,博士,副教授,研究领域为中文信息处理;冯凯(1987—),男,硕士生,研究领域为中文信息处理。E-mail:chenzq@hdu.edu.cn
收稿日期:2013-05-23 **修回日期:**2013-09-11 **文章编号:**1002-8331(2014)01-0116-05

题)库的QA系统、问答式检索系统以及基于自由文本的QA系统。根据面向的领域不同,QA系统又可以分为面向开放域的系统和面向限定域的系统[4]。本文研究基于FAQ库的限定域QA系统的关键技术,这类系统由于其实用性目前已成为自动问答研究领域的研究热点。

在限定域自动问答中,用户的提问可以根据其特点分为两类:一类是以询问领域概念、实体定义等为目的,其特点为结构简单且大多包含领域专有词语;另一类是先描述提问背景,然后在此基础上提出问题,其特点为背景描述一般比较长,问题与背景描述密切相关。现有的基于FAQ库的自动问答系统,均是以问句(或句子)的相似度计算为基础[5],对于处理第一类问题比较有效,而在处理第二类问题时效果较差。本文针对现有技术的不足,提出一种根据问题特点对问题进行模式分类的方法;并在问题模式分类基础上,提出针对不同模式问题的相似度计算方法,以达到大幅改善限定域QA系统性能的目的。

2　问题模式分类方法

2.1　问题模式及其特点

模式一词指代事物的标准样式,符合某种模式的事物都具有一些共同的特点。在对限定域自动问答系统进行研究过程中,通过收集整理大量用户的常问问题并对其进行分析,发现主要有两种模式。

第一种问题简明、直接,通常询问定义、方法、意见或者事实,同时包含大量领域内的专有词语,如下面的税务领域相关问题。

例1 “什么是摊销?”(询问定义)

例2 “国税都包含哪些税种?”(询问实体)

例3 “挂账工资是否缴纳个人所得税?”(询问意见)

例4 “小规模纳税人要升为一般纳税人要有几个条件要求?”(询问方法)

本文将这种模式的问题称为:简洁模式问题(Simple Mode Questions,SMQs)。

另外一类问题的特点是结构比较复杂,一般提问者在提出问题前会先给出一段比较长的背景描述,为后面的提问设定一个情景,然后再提出问题。见如下例子。

例5 “我公司新买了一辆轻型客车(运输用),开过来的发票是机动车销售统一发票(有发票联与抵扣联),请问这种发票是否可以抵扣?”

该例子中前两个分句是背景描述,后一个分句才是真正的问题。对于这类问题如果仅仅采用问句(或句子)的相似度计算来从FAQ库中检索答案,效果会比较差。本文将这种模式的问题称为:情景模式问题(Context Mode Questions,CMQs)。

2.2　问题模式分类算法

对问题进行模式分类,首先需要对用户问题进行分词和词性标注,本文使用的工具为中科院的ICTCLAS5.0汉语分词系统[6],其主要功能包括中文分词、词性标注、命名实体识别等,标注集也采用ICTCLAS标准。该系统经过多年的完善和改进,能够提供较为精确的分词结果,提高了后续算法的正确率。为有效处理特定领域问题,给该系统配置了领域词典。

算法思想是基于规则的方法,其中规则是对人工收集并整理的大量问题进行分析总结后获得的。不难发现简洁模式问题通常只由一句话组成,而情景模式问题由于需要描述背景,一般需要由多个分句构成,这里的分句是以逗号、分号、句号、问号等来分隔的。但是仅仅以分句数量作为特征进行判别会出现一些误判,因为有的用户提问方式比较灵活,可能会一次提若干个简洁模式问题,见例6。

例6 “新办企业固定资产可以全部用来抵扣税额吗?抵扣时间有没有限制?”

上述问题实际上是由两个简洁模式问题组成的,如果仅仅通过分句数量判别就会产生误判。因此还要考虑疑问句是第几个分句。

算法步骤如下:

步骤1　问题预处理。

将用户问题记为如下字符向量 $\boldsymbol{Q}_u$:

$$\boldsymbol{Q}_u=(c_1,c_2,\cdots,c_n) \tag{1}$$

分词和词性标注后,将 $\boldsymbol{Q}_u$ 转化为词语和词性构成的二元组序列 $\boldsymbol{V}_u$:

$$\boldsymbol{V}_u=\{(w_1,p_1),(w_2,p_2),\cdots,(w_m,p_m)\} \tag{2}$$

步骤2　问题特征向量获取。

对 $\boldsymbol{V}_u$ 进行一趟扫描获取 $\boldsymbol{V}_r$,作为问题特征向量:

$$\boldsymbol{V}_r=(S_n,N_1,N_2,\cdots,N_n) \tag{3}$$

其中,S_n 用来存储问题中的分句数量,后面的各维对应各个分句中包含的疑问词和疑问句特征词的数目。分句标志可以是(wj,ww,wt,wd,wf),即句号、问号、叹号、逗号、分号。疑问词的标注标志是疑问代词标记ry;疑问句特征词包括“能否”、“是否”、“是不是”等词语。

步骤3　问题模式类别判定。

按照公式(4)对用户问题模式进行判定,如果输出结果 $R=1$,则说明问题是简洁模式问题(SMQs),$R=0$,则为情景模式问题(CMQs)。

$$R=\begin{cases}1, & V_r[0]=1 \text{ or } (V_r[0]\geqslant 1 \text{ and } V_r[1]\neq 0)\\ 0, & \text{other}\end{cases} \tag{4}$$

3　问题相似度计算方法

对于简洁模式问题的匹配,可采用问句(或句子)相

似度计算方法来解决。句子之间的相似度计算目前已有较多研究,也取得不少成果[7]。针对自动问答系统中简洁模式问题的匹配也提出了一种基于动态规划的汉语问句相似度计算算法,无需分词而是通过获取两个问句中的最长公共子串集合来进行问句之间的相似度计算。该算法在实践中取得了令人满意的效果[8]。

而对于情景模式问题之间的相似度计算,目前还未见文献报道。情景模式问题之间的相似度计算情况比较复杂,在该类问题中,情景描述占有很大的比重,也是后面所提问题的语境信息,因此针对这类问题的相似度计算需要综合考虑情景相似度以及在此情景下所提问题的相似度。为此,本文构建的情景模式FAQ库由情景库和情景问题库组成。对于用户给出的问题,首先切分出情景部分和问句部分,并通过情景相似度计算从情景库中检索出类似情景,然后获取这些情景对应的问句集合,最后计算用户问题的问句部分与该问句集合中问句的相似度以获取答案。

3.1 情景模式FAQ库的组织

情景模式FAQ库包括情景库和情景问题库,结构如图1所示。

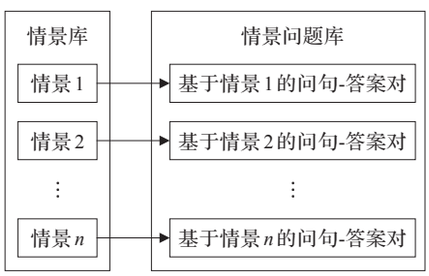

图1 情景模式问题FAQ库

情景从问题-答案语料库中提取,采用情景特征词集合的形式来表达。对于语料库中每一个情景模式问题,提取如式(3)所示的特征向量 $\boldsymbol{V}_r$,其中记录了问题中的分句数量以及各分句情况,可以将不包含疑问词或者疑问句特征词的分句作为情景描述,从而实现情景描述部分与问句部分的分离。

对于情景描述部分,首先通过扫描停用词表去除停用词,然后根据词性去掉标点。在一些情景描述中可能包含日期和时间等信息,而在情景相似度计算中一般只需对用户情景中的行为进行匹配,只有少量情景的匹配需要考虑时间因素,同时考虑到时间匹配计算十分复杂,为了兼顾实用系统中的精度和效率,因此将表示时间的词去掉。对于公司名之类的命名实体,用一个通配符号替代。这样经过处理后,情景描述部分转化为一个情景特征词集合。对于相似度大于某一阈值的多个情景描述(计算方法见3.2节),可以当做一个情景来处理,因此在FAQ库中每一情景可能对应多个问句-答案对。

3.2 情景相似度计算

从原理上来说,要度量两个情景之间的相似度,需要对情景描述进行语法、语义分析,在获取情景描述深层语义的基础上来计算情景之间的相似度。目前汉语语义分析研究主要集中在词义消歧和语义角色标注两个方面[9-10],面向的主要是规范的汉语句子,在面对实际应用系统中的开放语料时,无论是在分析结果还是在分析效率方面都难以满足要求。因此本文将情景特征词集合作为情景描述的形式化表达,并在此基础上计算情景之间的相似度。

设有两个情景,其特征词集合分别可以表示为两个向量:$\boldsymbol{V}_{d1}$ 和 $\boldsymbol{V}_{d2}$,两个情景之间的相似度即为两个向量的语义相似度。设:

$$\boldsymbol{V}_{d1}=(W_{d11}, W_{d12}, \cdots, W_{d1m}) \tag{5}$$

$$\boldsymbol{V}_{d2}=(W_{d21}, W_{d22}, \cdots, W_{d2n}) \tag{6}$$

则其相似度可由公式(7)计算:

$$\mathrm{sim}(\boldsymbol{V}_{d1}, \boldsymbol{V}_{d2})=\frac{\dfrac{\sum_{i=1}^{m} x_i}{m}+\dfrac{\sum_{j=1}^{n} y_j}{n}}{2} \tag{7}$$

其中 x_i 定义见式(8):

$$x_i=\max(\mathrm{sim}(W_{d1i}, W_{d21}), \mathrm{sim}(W_{d1i}, W_{d22}), \cdots, \mathrm{sim}(W_{d1i}, W_{d2n})) \tag{8}$$

即 $\boldsymbol{V}_{d1}$ 中的第 i 个词项与 $\boldsymbol{V}_{d2}$ 中词项相似度的最大值。y_j 的定义类似,即定义为 $\boldsymbol{V}_{d2}$ 中的第 j 个词项与 $\boldsymbol{V}_{d1}$ 中词项相似度的最大值,见式(9)。

$$y_j=\max(\mathrm{sim}(W_{d2j}, W_{d11}), \mathrm{sim}(W_{d2j}, W_{d12}), \cdots, \mathrm{sim}(W_{d2j}, W_{d1m})) \tag{9}$$

将词项分为名词词项(包括命名实体)和其他词项。名词词项之间的相似度本文采用基于Wikipedia的计算方法,其他词项采用完全匹配的方法。Wikipedia(维基百科)是目前世界上最大的、多语种的、开放式的在线百科全书,已在自然语言处理的多个领域得到应用[11]。Wikipedia覆盖面极其广泛,覆盖几乎所有专门领域的专有词汇,如税务领域的“个人所得税”、“小规模纳税人”、“摊销”等等。分析整理了最新版本的中文Wikipedia数据,并提出了有效的词语相关度计算方法[12]。本文在已有工作基础上,借鉴文献[13]提出的利用Wikipedia分类信息和页面链接信息的词语相似度方法,实现了名词词项之间的相似度计算。

3.3 情景模式问题相似度计算

情景模式问题的相似度计算需要综合考虑情景部分的相似度和问句部分的相似度。设有两个情景模式问题 Q_1 和 Q_2,其情景特征词向量分别为 $\boldsymbol{V}_{d1}$ 和 $\boldsymbol{V}_{d2}$,问句部分分别为 q_1 和 q_2,则 Q_1 和 Q_2 的相似度可由公式(10)计算:

$$\mathrm{sim}(Q_1, Q_2)=\alpha\mathrm{sim}(\boldsymbol{V}_{d1}, \boldsymbol{V}_{d2})+\beta\mathrm{sim}(q_1, q_2) \quad (10)$$

其中 α 和 β 分别为情景相似度和问句相似度的权重，$\alpha+\beta=1$，本文 α 和 β 均取值为0.5。

$\mathrm{sim}(\boldsymbol{V}_{d1}, \boldsymbol{V}_{d2})$ 由公式(7)计算，$\mathrm{sim}(q_1, q_2)$ 即问句的相似度计算采用已有的基于动态规划的汉语句子相似度计算方法[8]。

4　实验结果与分析

4.1　原型系统结构

针对本文提出的问题模式分类算法和情景模式问题相似度计算算法，分别进行了实验，并实现了完整的自动问答原型系统。原型系统结构如图2所示。

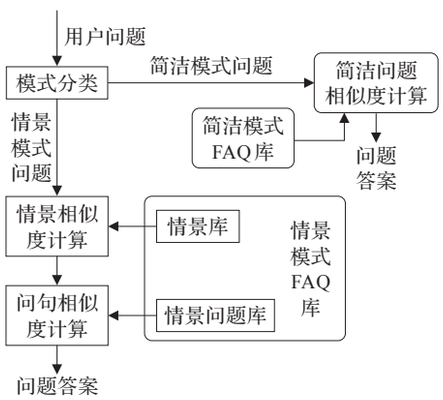

图2　原型系统结构图

4.2　实验语料

实验语料来自南方某市地方税务局的在线咨询问答系统，收集了该系统2011年3月到2012年3月期间的所有问题，每条问题均包含用户提问和人工回答。通过分析，发现简洁模式问题约占整个问题总量的30%，而带有情景描述的情景模式问题达到了70%左右。

4.3　问题模式分类实验及分析

随机抽取语料库中的1 000个问题，其中简洁模式问题300个，情景模式问题700个。采用本文提出的问题模式分类算法对这1 000个问题进行分类，测试结果319个被判定为简洁模式问题，其中将情景模式问题判定为简洁模式问题的有30条(误判)。681个被判定为情景模式问题，其中将简洁模式问题判定为情景模式问题的有11条(误判)。

两种模式问题分类的准确率和召回率(Precision & Recall)[14]，如表1所示。

表1　问题模式分类结果　(%)

分类	准确率	召回率
简洁模式问题	90.6	96.3
情景模式问题	98.4	95.7

由于问题模式分类未见相关研究，无法进行实验比较，但从实验数据来看本文方法是令人满意的。分析误判的实例，发现将简洁模式问题误判为了情景模式问题，主要是因为用户提问不规范，如“办公桌，空调等等这些固定资产进项税可以抵扣吗?”被误判为情景模式，是因为“办公桌”和“空调”两个词语之间没有用顿号“、”，而是误用了逗号“，”，造成问题被分成了两个分句。将情景模式问题误判为简洁模式问题主要是因为情景描述和问句之间的分割标志不规范或者情景和问句之间是连写的。总的来说该算法在进行问题模式判定中同时兼顾到效率和性能，是实用且有效的。

4.4　情景模式问题相似度计算实验及分析

分析处理实验1中的700个情景模式问题，构成情景模式FAQ库，另外从语料库中抽取300个情景模式问题作为测试集，这300个情景模式问题经过原型系统的测试均可返回答案，正确结果由人工标注。

为进行实验比较，实现了三种问题相似度计算算法。第一种即本文提出的将情景描述和问句分开并分别计算相似度，然后计算综合相似度的方法；第二种是基于动态规划的句子相似度计算方法[8]，该方法将问题作为一个整体，不区分情景描述和问句；第三种是基于《知网》(HowNet)的句子相似度计算算法[15]，也不区分情景描述和问句，但需要分词。实验结果如表2所示。

表2　问题相似度计算实验结果与比较

算法	正确数	正确率/(%)	时间/s
本文算法	223	74.3	175.5
动态规划算法	166	55.3	58.1
《知网》算法	233	77.7	246.3

从实验结果来看，本文算法综合性能最优。基于动态规划的方法，由于不分词，也无需其他外部知识库的支持，因此时间效率最优，但正确率很低，没有实用价值。基于《知网》的方法，虽然正确率比本文方法略高，但由于需要查询庞大的外部知识库《知网》，其时间效率很低，同时由于《知网》的商业化使用需要授权，会提高产品开发的成本，因此也不是最优选择。

本文方法的正确率也不是很高，主要原因有以下几点：

(1)用户的输入比较随意，中间可能夹杂错别字，或者网络用语，如：“99我”(救救我)，或者“求助大虾”等。对这种情况，既不能将这些词简单地加入停用词表，同时也没有在现有的语料中实现拼写检查功能，因此还没能很好地解决这个问题。

(2)在情景描述中，有的用户问题会加入时间，并且所提问的问题是关于时间期限的。本文算法为保证效率忽略了时间的匹配计算，从而影响了匹配结果的正确性。不过这种问题在所有情景中占有的比例很小。

(3)有的词语存在一词多义的情况。在这种情况下，即使使用语义词典也没有涉及到对上下文的分析，而如果对句子进行更深层的分析，作为实用问答系统，

效率达不到要求,因此需要根据实际情况进行取舍。

5 结束语

自动问答技术是自然语言处理领域的研究热点,自动问答系统也被称为“第二代搜索引擎”,拥有广阔的应用前景。本文针对特定域自动问答领域现有研究的不足,提出将用户问题分为简洁模式问题和情景模式问题,并提出了问题模式分类算法和情景模式问题的相似度计算方法。本文以税务知识自动咨询为应用背景,实现了一个原型系统并进行了实验。实验结果表明本文算法是可行且有效的,有望推广到其他应用领域。

本文研究也存在不足之处:在进行问题模式分类时,主要采用了基于规则的方法,由于实际语料(特别是网络文本)的复杂性,人工总结的规则无法覆盖所有情况,会造成误判;在计算情景相似度时,简单地采用情景特征词集合作为情景语义的形式化表达,造成总体性能难以进一步提高。解决以上问题,是今后继续研究的方向。

参考文献:

[1] 郑实福,刘挺,秦兵,等.自动问答综述[J].中文信息学报,2002,16(6):46-52.

[2] 张亮.面向开放域的中文问答系统问句处理相关技术研究[D].南京:南京理工大学,2005.

[3] 王树西.问答系统:核心技术、发展趋势[J].计算机工程与应用,2005,41(18):1-3.

[4] Huang Gaitai,Yao Hsiuhsen.Chinese question answering system[J].Journal of Computer Science and Technology,2008,19(4):479-488.

[5] 熊大平,王健,林鸿飞.一种基于LDA的社区问答问句相似度计算方法[J].中文信息学报,2012,26(5):40-45.

[6] 刘克强.2009共享版ICTCLAS的分析与使用[J].科教文汇,2009,8:271-272.

[7] 岳大鹏,饶岚,王挺.一种针对新闻话题的多文档文摘技术[J].中文信息学报,2012,26(6):79-84.

[8] 冯凯,王小华,谌志群.基于动态规划的汉语句子相似度算法[J].计算机工程,2013,39(2):220-224.

[9] 罗森林,韩磊,潘丽敏,等.汉语句义结构模型及其验证[J].北京理工大学学报,2013,33(2):166-171.

[10] 刘挺,车万翔,李正华.语言技术平台[J].中文信息学报,2011,25(6):53-62.

[11] 赵飞.维基百科研究综述[J].电子科技大学学报,2010,39(3):322-335.

[12] 谌志群,高飞,曾智军.基于中文维基百科的词语相关度计算[J].情报学报,2012,31(12):1265-1270.

[13] 盛志超,陶晓鹏.基于维基百科的语义相似度计算方法[J].计算机工程,2011,37(7):193-195.

[14] Euzenat J.Semantic precision and recall for ontology alignment evaluation[C]//Proceedings of the 20th International Joint Conference on Artificial Intelligence(IJCAI),2007:348-353.

[15] 程传鹏,吴志刚.一种基于知网的句子相似度计算方法[J].计算机工程与科学,2012,34(2):172-175.

(上接104页)

[2] Maji P K,Biaswas R,Roy A R.Soft set theory[J].Computers and Mathematics with Applications,2003,45(4/5):555-562.

[3] Aktas H,Cagman N.Soft sets and soft groups[J].Information Sciences,2007,178:2726-2735.

[4] Feng F,Jun Y,Zhao X.Soft semirings[J].Computers and Mathematics with Applications,2008,56:2621-2628.

[5] Maji P K,Roy A R,Biswas R.An application of soft sets in a decision making problem[J].Computers and Mathematics with Applications,2002,44:1077-1083.

[6] Chen D,Tsang E,Yeung D,et al.The parameterization reduction of soft sets and its applications[J].Computers and Mathematics with Applications,2005,49:757-763.

[7] Kong Z,Gao L,Wang L,et al.The normal parameter reduction of soft sets and its algorithm[J].Computers and Mathematics with Application,2008,56:3029-3037.

[8] Feng F,Li C,Davvaz B,et al.Soft sets combined with fuzzy sets and rough sets:a tentative approach[J].Soft Comput,2010,14:899-911.

[9] Feng F,Liu X,Violeta L,et al.Soft sets and soft rough sets[J].Information Sciences,2011,181:1125-1137.

[10] 马振明.Pawlak近似空间中软集的软上(下)近似[J].计算机工程与应用,2011,47(18):60-61.

[11] Muhammad I A.A note on soft sets,rough soft sets and fuzzy soft sets[J].Applied Soft Computing,2011,11:3329-3332.

[12] Meng D,Zhang X,Qin K.Soft rough fuzzy sets and soft fuzzy rough sets[J].Computers and Mathematics with Applications,2011,62:4635-4645.

[13] Zhang W,Leung Y.Theory of including degrees and its applications to uncertainty inferences[C]//Proceedings of Asian Fuzzy Systems Symposium on Soft Computing in Intelligent Systems and Information Processing,New York,1996.[S.l.]:IEEE,1996:496-501.

[14] Ziarko W.Variable precision rough set model[J].Journal of Computer System Science,1993,46(1):39-59.