问题分类的计算模型研究*)

张 亮[1,2,3] 陈肇雄[2] 黄河燕[2]
(南京理工大学计算机系 南京 210000)[1] (中国科学院计算机语言信息工程研究中心 北京 100083)[2]
(江苏警官学院 南京 210000)[3]

摘 要 问题分类是问答系统技术处理的基础与核心,它决定答案抽取的范围和方法,进而影响整个系统的性能。本文提出了一个基于贝叶斯理论的问题分类计算模型,并给出其详细算法。研究分析了问句内部结构与问题类型之间的关系,将基于疑问词的2-gram组合和问句特征项同义近义扩展应用到具体计算中。实验表明,效果较为理想。

关键词 问答系统,问题分类,贝叶斯模型

Research of Question Classification Computation Modal

ZHANG Liang[1,2,3] CHEN Zhao-Xiong[2] HUANG He-Yan[2]
(Dept. of Computer Science, NJUST, Nanjing 210000)[1]
(Research Center of Computer & Language Information Engineering, CAS, Beijing 100083)[2] (Jiangsu Police Institute, Nanjing 210000)[3]

Abstract Question classification is the basic and core of question answering system process. It rules answer extraction range and method, and effects entire system performance. This paper proposes a new Bayes-based question classification computation modal and its detail algorithm, studies and analyzes relation of question structure and question type, and applies interrogative-based 2-gram and feature vector synonym to extend classification computation. Experiment shows that result is ideal.

Keywords Question answering, Question classification, Bayes model

1 引言

传统的信息检索主要是基于关键字,检索结果是相关的大量文本或网页,需要用户花费大量的时间从中筛选。问答系统是信息检索的高级形式,近年来成为国内外信息技术领域研究的热点。它有别于传统检索系统的特征或优势表现为:一是检索入口是自然语言形式的问句,而非关键字;二是检索得到的结果是一句或一段话,答案简洁,检索效率高[1]。

目前的问答系统虽各有不同的处理技术,但总体流程上都有一些相似或相近的处理模块,即问题分析(Question Parsing)、问题分类(Question Classification)、问题形式化(Query Formulation)、形式化扩展(Query Expansion)、文本检索(Document Retrieval)、候选句检索(Candidate Sentence Retrieval)、答案抽取(Answer Extraction)。图1即为一个典型的问答系统处理流程[2],其中问题分类是问答系统处理的基础与核心。不同的分类标准和方法,决定了不同的问题形式化和形式化扩展,同时决定了不同的答案抽取范围和方法[3]。

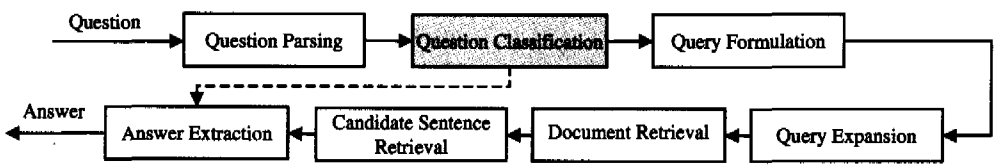

图1 问答系统的一般处理流程

2 问题分类研究现状

与文本分类类似,问题分类也是先根据一定的分类标准,定义一个类型集,对于某个问题,根据一定的算法判断出该问题属于哪一个类型。目前没有统一或权威的分类标准,因此定义的类型集也是多种多样,如在美国TREC QA-Track中,Ittycheriah等人定义了含有31个基本类的双层问题类型库[4];Hovy等人定义了141个类型的多层分类方法[5]。在中文问答系统研究中,哈尔滨工业大学的问答系统基于问题焦点和答案内容,定义的问题类型集具有一定的代表性[6],如表1所示,分为6大类,65子类。类型分层即子类的划分有利于对问题更精确地把握,从而在答案抽取中定位得更精确。

在早期的问答系统研究中,包括TREC QA-Track中,大多数的QA系统都是采用基于规则的方法进行问题分类。规则的优点是准确、算法简单,缺点是需要花费大量的人力和时间,而且运行的效率不高。更重要的是,由于自然语言的复杂多变性,规则不可能穷尽。近几年,有些系统将文本分类技术应用到问题分类,取得了一定的效果,如Dell Zhang用支持向量机的方法[7]、Wei Li用一元和二元语言模型的统计方法[8]处理英语问题分类;Jun Suzuki等在依存句法的基础上采用

*)基金项目:国家自然科学基金资助项目(60272088)。张 亮 博士生,主要研究方向为自然语言处理、自动问答系统。

一种多层有向图的机器学习方法[9]处理日语问题分类等。

表1 一个有代表性的问题类型划分

Types	Sub-types
HUM	HUM_DESCRIPTION, HUM_ALIAS, HUM_PERSON, HUM_ORGANIZATION, HUM_LIST, HUM_OTHER
LOC	LOC_ADDRESS, LOC_LAKE, LOC_CITY, LOC_CONTINENT, LOC_COUNTRY…
NUM	NUM_AGE, NUM_PRICE, NUM_COUNT, NUM_DISTANCE, NUM_FREQUENCY…
TIME	TIME_YEAR, TIME_MONTH, TIME_DAY, TIME_LIST, TIME_RANGE, TIME_OTHER
OBJ	OBJ_ANIMAL, OBJ_FOOD, OBJ_COLOR, OBJ_CURRENCY, OBJ_ENTERTAIN…
DES	DES_ABBR, DES_DEFINITION, DES_REASON, DES_MEANING, DES_MANNER, DES_OTHER

我们将贝叶斯理论与中文问句的属性特征相结合，在一定的标注语料的基础上，通过对问句统计计算，进行问题归类，取得了较好的效果。

3 问题分类的计算模型

3.1 贝叶斯模型理论

在文本分类中，基于贝叶斯模型的技术是比较成熟的一种，优点是算法简单高效。其基本思想是：利用类别的先验概率和词的分布对于类别的条件概率来计算未知文本属于某一类别的概率。在假设文本中词的分布相互独立，即忽略上下文的 Unigram 模型中，贝叶斯分类模型可由公式(1)、(2)表示：

$$P(C/D)=\frac{P(C)\prod_{F_j\in V}P(F_j/C)^{TF(F_j,D)}}{\sum_i P(C_i)\prod_{F_j\in V}P(F_j/C_i)^{TF(F_j,D)}} \tag{1}$$

$$P(F_j/C)=\frac{1+TF(F_j,C)}{|V|+\sum_{C_l\in C}TF(F_j,C_l)} \tag{2}$$

其中，$P(C/D)$即为文档 D 属于 C 类文档的概率，$P(C)$为类型 C 的概率，$P(F_j/C)$是对在 C 类文档中特征 F_j 出现的条件概率的拉普拉斯概率估计，$TF(F_j,C)$是 C 类文档中特征 F_j 出现的频率；$|V|$为特征项词典集的大小，等于文档表示中所包含的不同特征的总数目。

3.2 基于贝叶斯的问题分类计算模型

将问答系统中的问句看作一种特殊的文本，则可以基于贝叶斯模型对问句进行分类。在进行分类计算之前，需做必要的预处理，主要工作为对问句分词和去停用词以及特征项扩展。所谓特征项是指问句中影响问题分类的最小语义单位，本文以对问句分词和去停用词后的词作为特征项。问句在处理中以特征向量序列表示，如对于问句"加勒比海的飓风季节是什么时候?"进入分类计算的向量序列即为("加勒比海","飓风","季节","什么","时候")。由于语料库的规模较小，而问句往往较短，为了提高匹配的精确度，需对特征项做同义词和近义词扩展。如飓风的扩展：〈飓风〉→〈台风〉|〈强风〉|〈季风〉…。系统扩展处理基于语言学资源《知网》[10]，这方面的具体工作可以参考文[11]。

定义1 设问句 $Q=\sum_{i=1}^{m}w_i$，w_i 是问句分词后的第 i 个词，(包括停用词)，m 是词的个数。S 是停用词集合，$n\in[0,m]$，则 $F=(F_1,F_2,\cdots,F_n)$是问句 Q 的特征向量，当且仅当 $F\subseteq Q$ 且 $F_i\notin S,F_i\in F$，n 是特征个数。

定义2 $F=(F_1,F_2,\cdots,F_n)$是问句 Q 的特征向量，$F'=(F'_{1,1},F'_{1,2},F'_{1,3},\cdots,F'_{n,1},F'_{n,2},F'_{n,3},\cdots)$是 F 的扩展向量，$F'_{i,1},F'_{i,2},F'_{i,3}\cdots$是特征项 F'_i 的同义近义词序列，$i\in[0,n]$。

由于问句一般较短，问句中的特征项很少，(2)式中的 $|V|$ 的绝大多数特征 F_j 不出现在待处理问句中，从而 $TF(F_j,D)=0$，且由于分母不变，我们要求的概率计算值最大的概率，所以在问题分类计算中，计算公式可由公式(1)转化为(3)。由于增加了特征项扩展，公式(2)转化为公式(4)：

$$P(C_i/Q)=\underset{C_i\in C}{\operatorname{argmax}}P(C_i)\prod_{F_j\in Q}P(F_j/C_i)^{TF(F_j,C_i)} \tag{3}$$

$$P(F_j/C_i)=\frac{1+TF(F_j,C_i)}{|V|+\sum_{C_l\in C}TF(F_j,C_l)}+\eta\times\sum_k\left(\frac{1+TF(F'_{j,k},C_i)}{|V|+\sum_{C_l\in C}TF(F'_{j,k},C_l)}\right) \tag{4}$$

其中，问句 Q 代替公式(1)中的文本 D，即处理的对象是问句。公式(4)中的 η 为特征项同义近义词扩展的影响系数，$\eta\in[0,1]$，$k\in[0,m]$，m 是特征项 F_j 扩展序列中词的个数，$F'_{j,k}$是特征项 F_j 的扩展序列中第 k 个词。

算法1 问题分类计算算法

相关基础：停用词库 S，分词软件，问句类型表 C，向量形式表示的问题标注语料库 D，同义词近义词扩展程序。

输入：待分析的问句 Q

输出：该问句的类型 T

1)对 Q 进行分词，得词序列$(w_1,w_2,\cdots,w_n)$。

2)依据定义1，去除$(w_1,w_2,\cdots,w_n)$中的停用词，获得特征向量$(F_1,F_2,\cdots,F_m)$，$m\leqslant n$。

3)建立变量 MaxValue，Sum_1，Sum_2(初值均为0)，Sum_3(初值为1)，$Count_1$，$Count_2$，Type。

4)依序从类型表 C 中读取一个类型 C_k，若读尽，则转17)。

5)依序从特征向量$(F_1,F_2,\cdots,F_m)$中读取 F_i，若读尽，则转16)。

6)依据定义2，生成特征向量 F_i 的同义近义词扩展序列$(F'_{i,1},F'_{i,2},F'_{i,3}\cdots)$。

7)依序从$(F'_{i,1},F'_{i,2},F'_{i,3}\cdots)$中读取一个 $F'_{i,k}$，若读尽，则转11)。

8)统计 $F'_{i,k}$在整个语料库中频率，存入 $Count_1$。

9)统计 $F'_{i,k}$在类型 C_k 中频率，存入 $Count_2$。

10)$Sum_1=Sum_1+(1+Count_2)/(|V|+Count_1)$，转7)。

11)$Sum_1=\eta\times Sum_1$(η为系数，这里取0.8)。

12)统计 F_i 在整个语料库中的频率，存于 $Count_1$。

13)统计 F_i 在类型 C_k 中的频率，存于 $Count_2$。

14)$Sum_2=Sum_1+(1+Count_2)/(|V|+Count_1)$

15)$Sum_3=Sum_3\times Sum_2^{count^2}$转5)。

16)若 $Sum_3>MaxValue$，则 $MaxValue=Sum_3$，且 $Type=k$(记录类型号)转4)。

17)$T=Type$(该问句的问题类型为 T)。

算法复杂性分析：在算法1中，影响算法复杂性的关键是2个循环：一个是对不同类型计算的循环，一个是特征向量及其同义近义词扩展 TF 计算的循环。所以其时间复杂度为 $O(m\times n\times k)$，这里 m 为问题类型的个数，n 为问句 Q 的特征向量的个数，k 为特征向量同义近义词扩展的个数。

平滑处理

设 $TF(F_j,C_i)$为类型 C_i 中特征 F_j 出现的频率，有三种情况：case1：C_i 中包含 F_j；case2：C_i 中不包含 F_j，但包含 F_j 的扩展项；case3：C_i 中既不包含 F_j，也不包含 F_j 的扩展项。

如果是第三种情况，则 $TF(F_j,C_i)=0$，因此需要做相应的平滑处理，$TF'(F_j,C_i)$中融入了对这三种情况的考虑。

$$TF'(F_j,C_j)=\begin{cases}TF(F_j,C_i)+\lambda\times\sum_k TF(F'_{j,k},C_i) & \text{if case1 or case2}\\ \text{Total}_0(C) & \text{if case3}\end{cases}$$

其中，λ 是系数，$\lambda\in[0,1]$，$TF(F'_{j,k},Q)$是特征项 F_j 的扩展序列中第 k 个词在问句集中的频率，$\text{Total}_0(C)$是在类型 C 中特征项的数目。

3.3 计算模型的改进

问句具有自身的属性特征。如：问句一般都比文本短；问句在句法上有一定的结构形式；问句中疑问词以及疑问词与其他词的搭配等。这些特性都对问题的归类有重要的影响。因此，在基于贝叶斯的计算模型中，需要将这些影响因素体现出来，从而提高分类的准确率。

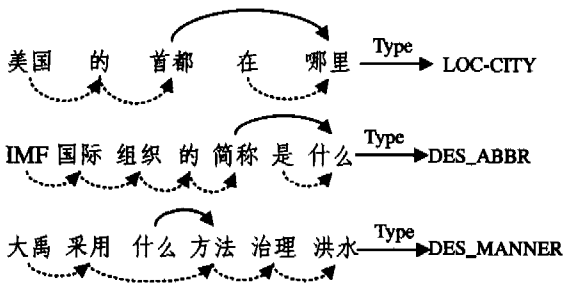

图 2 问句结构与类型关系分析

观察如下问句："美国的首都在哪里？""IMF 国际组织的简称是什么？""大禹采用了什么方法治理洪水？"等，如图 2 所示。通过对疑问结构的分析，发现在问句结构内部存在这样的内在联系，即子类型往往由疑问词和疑问焦点确定，如"首都"+"哪里"->LOC_CITY，"简称"+"什么"->DES_ABBR，"什么"+"方法"->DES_MANNER。有鉴于此，我们对以上计算模型做进一步的改进。首先提取问句中的疑问词，然后以疑问词为核心，与其他特征项组合，进行基于疑问词的 2-gram 模型计算。

定义 3 $F=(F_1,F_2,\cdots,F_n)$是问句 Q 的特征向量，Q_w 是疑问词集合，如果 $w\in F$，且 $w\in Q_w$，则称 w 为向量 F 的疑问词。

定义 4 设$\{F_1,F_2,\cdots,F_n\}$是问句 Q 的特征项的集合，w 是向量 F 的疑问词，$w\in\{F_1,F_2,\cdots,F_n\}$，$S=\{F_1,F_2,\cdots,F_n\}-\{w\}$，$\overline{F}=\{w\}\times S$，则称 $\overline{F}$ 为基于疑问词的特征项 2 元组合。

定义 5 $F=(F_1,F_2,\cdots,F_n)$是问句 Q 的特征向量，$F'=(F'_{1,1},F'_{1,2},F'_{1,3},\cdots,F'_{n,1},F'_{n,2},F'_{n,3},\cdots)$是 F 的扩展向量，w 是向量 F 的疑问词，$w\in\{F_1,F_2,\cdots,F_n\}$，$S'=F'-\{w\}$，$\overline{F}'=\{w\}\times S'$，则称 $\overline{F}'$为基于疑问词的特征项 2 元组合扩展。

相应地，公式(3)、(4)转化为公式(5)、(6)，其平滑处理参照 $TF'(F_j,Q)$。

$$P(C_i/Q)=\underset{C_i\in C}{\text{argmax}}P(C_i)\times\prod_{\overline{F}_j\in F}P(\overline{F}_j/C_i)^{TF'(F_j,C_i)} \quad (5)$$

$$P(\overline{F}_j/C)=\frac{1+TF''(\overline{F}_j,C_i)}{|V|+\sum_{C_l\in C}TF''(\overline{F}_j,C_l)}+\eta'\times\sum_k\left(\frac{1+TF''(\overline{F}'_{j,k},C_i)}{|V|+\sum_{C_l\in C}TF''(\overline{F}'_{j,k},C_l)}\right) \quad (6)$$

由于疑问词也存在同义或近义的问题，如〈哪里〉、〈哪儿〉、〈哪个地方〉、〈哪〉、〈什么地方〉，〈谁〉、〈何人〉、〈什么人〉、〈哪位〉等，这些疑问词在问句中语义相同，可以相互替换，因此在计算中，也需要做疑问词的同义扩展。

定义 6 $F=(F_1,F_2,\cdots,F_n)$是问句 Q 的特征向量，Q_w 是疑问词集合，w 为向量 F 中的疑问词，$w'=(w'_1,w'_2,\cdots,w'_n)$是 w 的疑问词扩展，当且仅当 $w'_i\in Q_w(i\leqslant n)$，且 w'_i 与 W 语义相等。

算法 2 基于疑问词的分类算法

在基础部分增加疑问词列表 Q_w，改进问题分类模型的关键是加入了对疑问词及其搭配的考虑。在具体的计算中，计算对象不再是单个的特征向量，而是疑问词与特征向量的组合。因此在计算前，问题语料中的特征向量需按定义 4 表示成$\{w\}\times S$ 的 2-gram 形式。

输入：待分析的问句 Q

输出：该问句的类型 T

1) 建立变量 MaxValue，Sum_1，Sum_2（初值均为 0），Sum_3（初值为 1），Count_1，Count_2，Type。

2)依序从类型表 C 中读取一个类型 C_k，若读尽则转结束处理。

3)对 Q 进行分词，得词序列$(w_1,w_2,\cdots,w_n)$。

4)去除$(w_1,w_2,\cdots,w_n)$中的停用词，获得特征向量$(F_1,F_2,\cdots,F_m)$，$m\leqslant n$。

5)依据定义 3，获取$(F_1,F_2,\cdots,F_m)$中的疑问词 w_f，并将其从$(F_1,F_2,\cdots,F_m)$去除。

6)依据定义 6，生成疑问词 w_f 的疑问词扩展$(w'_1,w'_2,\cdots,w'_n)$。

7)依据定义 4，生成$(w'_1,w'_2,\cdots,w'_n)\times(F_1,F_2,\cdots,F_m)$的 2-gram 形式。

8)……(其他步骤与算法 1 类似)

算法 2 中增加了对疑问词的判断和处理，其时间复杂度依然为 $O(m\times n\times k)$。

4 实验设计及结果分析

实验数据是在哈尔滨工业大学提供的 1890 句标注类别的问句库的基础上扩充的 3000 句标注句库，分类体系为 6 大类、65 小类，如表 1 所示，数据类型构成如表 2 所示。

表 2 测试语料构成情况

Number kind / Type	Sub-types number	Question number	Test number
DES	6	601	82
HUM	6	312	73
LOC	14	638	95
NUM	17	512	73
OBJ	13	550	77
TIME	7	388	40

其中，测试问句是从问题集里按 10%的比例随机抽取，再随机补充了部分类型一致的新造问句。因为在计算中存在大量的特征项和疑问词词频统计，且统计值是可复用的，因此

在后台处理中，统计语料库中所有特征项和疑问词在不同问题类型中出现的频率，并建立相应的索引。这样在具体计算时实时调用，提高了处理效率。表3即为语料库中疑问词频率索引形式。

表3 疑问词词频索引表

Type / Item	LOC			OBJ …
	LIST	LAKE	COUNTRY …	FOOD …
哪个	0	45	39 …	27 …
哪里	11	34	8 …	52 …
哪些	50	16	27 …	11 …
…	…	…	…	…

因为疑问句归类的特殊性，我们仅需测试问题归类的准确率，计算公式如下：

$$Prec1=(a/b)\times 100\%$$

其中，a 为测试正确的数目，b 为测试总数，Prec1 用以评测问题子类型的准确率；

$$prec2=(\sum a_i/\sum b_i)\times 100\%$$

其中，$\sum a_i$ 是测试类型的子类型中测试正确问句数之和，$\sum b_i$ 是测试类型的测试问句数之和，Prec2 用以测评问题类型的准确率。

实验结果如表4和表5所示。表4是6大问题类型的测试结果，表5是HUM和DES中各子类型的测试情况。总的准确率达到81%，结果较为理想。如与规则方法相结合，预计准确率还可以有所提高。存在的问题是：①语料库规模较小，且不平衡，平均每个子类型才40几句。②分词不准确，对计算有负面影响，如“中科院计算所在哪里?”，分词软件切分的结果为“中科院/计算/所在/哪里”。③部分类型句式单一，如子类型 NUM_POSTCODE，数量少，多为“XXXX 的邮编是多少”这样的模式，所以准确率可达100%。④各子类型中 OTHER 型准确率较低，其主要原因是这一类型不能归类到其他类型的类型，其本身的聚类性较低，如 DES_OTHER 中的“冰淇淋用法语怎么说”?“IC卡和磁卡有什么不同”?“人民币单位是什么”? 这些问句中特征项和基于疑问词的2-gram组合的共性很低。

表4 6大问题类型的测试结果

	DES	HUM	LOC	NUM	OBJ	TIME
测试数	82	73	95	73	77	40
正确数	60	59	86	62	60	32
Prec2	0.73	0.81	0.91	0.85	0.78	0.80

表5 HUM和DES中各子类型的测试结果

	HUM						DES					
	DESC…	ALIAS	PERSON	ORGA…	LIST	OTHER	ABBR	REAS…	MEAN…	MANN…	DEFI…	OTHER
测试数	20	10	20	20	7	5	20	12	15	9	12	5
正确数	13	8	19	18	6	3	18	10	11	6	8	2
Prec1	0.65	0.80	0.95	0.90	0.85	0.60	0.90	0.83	0.73	0.67	0.67	0.4

结论 问题分类是问答系统技术处理的第一步，分类准确度对最终答案的正确抽取有极大的影响。本文在贝叶斯模型理论的基础上，分析问句结构特征，结合问题标注语料，研究了在模型中加入特征项扩展和基于疑问词的2-gram组合。实验表明，效果比较理想。今后将在现有的统计计算模型的基础上，深入挖掘问句的结构特征，引入词法分析和句法分析，使问题分类的准确率进一步提高。

致谢 在研究中使用了哈尔滨工业大学信息检索研究室提供的问题语料和语言学资源《知网》，在此表示诚挚的感谢。

参考文献

1 Voorhees E M. Overview of the TREC 2002 Question Answering Track. In: Proc. of the Eleventh Text REtrieval Conference (TREC2002), 2002

2 Lin J, Katz B. Question Answering Techniques for the World Wide Web. The 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL-2003), 2003

3 Chang Yi, Xu Hongbo, Bai Shuo. TREC 2003 Question Answering Track at CAS-ICT. The Twelfth Text REtrieval Conference (TREC2003), 2003

4 Ittycheriah A, Franz M, Roukos S. IBM's Statistical Question Answering System-TREC-10. Processing of TREC 2001, 2001

5 Hovy E H, Gerber L, Hermjakob U, et al. Toward Semantics-Based Answer Pinpointing. Proc. of the Human Language Technology Conference (HLT2001), 2001

6 http://ir.hit.edu.cn/

7 Zhang D, Lee Wee Sun. Question Classification using Support Vector Machines. The 26th Annual International ACM SIGIR Conference. Toronto, Canada, July, 2003

8 Li W. Question Classification Using Language Modeling: [CIIR Technical Report]. University of Massachusetts, Amherst, 2002

9 Suzuki J, et al. Question Classification using HDAG Kernel. In: Workshop on Multilingual Summarization and Question Answering 2003, (post-conference workshop in conjunction with ACL-2003), 2003. 61～68

10 董振东，董强. http://www.keenage.com

11 刘群，李素建. 基于《知网》的词汇语义相似度计算. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59～76