*Computer Science and Application* 计算机科学与应用, 2011, 1, 134-139
http://dx.doi.org/10.12677/csa.2011.13027 Published Online December 2011 (http://www.hanspub.org/journal/csa)

# A Classification of Chinese Questions in the Domain of Stocks

**Mengfei Ren[1], Peng Wang[1], Hengjin Cai[1], Zhenyu Zhang[1], Pingli Wang[2], Meng Meng[2], Zhengye Bian[2]**
[1]International School of Software, Wuhan University, Wuhan
[2]Economics and Management School, Wuhan University, Wuhan
Email: {renmengfei, wangpengtoo, stevenzzy9}@gmail.com; hjcai@whu.edu.cn
Received: Oct. 8th, 2011; revised: Nov. 2nd, 2011; accepted: Nov. 20th, 2011.

**Abstract:** Question classification is a necessary part of Q & A system, and it determines the accuracy of answer extracting in some degree. This paper introduces a classification that depends on the sample set including questions from stock forum. We divide questions into three main parts: conception, investors and stocks. And we also divide conception class into seven parts: definition I, definition II, comparison, description, reason, enumeration, and attribute. The test results show that this approach did good performance in stock domain.

**Keywords:** Chinese Question Classification; Stock Domain; Hand-Crafted Rule

# 股票领域中的一种中文问句分类方法

**任梦菲[1]，王　鹏[1]，蔡恒进[1]，张震宇[1]，王萍莉[2]，孟　蒙[2]，卞正野[2]**
[1]武汉大学国际软件学院，武汉
[2]武汉大学经济与管理学院，武汉
Email: {renmengfei, wangpengtoo, stevenzzy9}@gmail.com; hjcai@whu.edu.cn
收稿日期：2011 年 10 月 8 日；修回日期：2011 年 11 月 2 日；录用日期：2011 年 11 月 20 日

**摘　要：**问句分类是问答系统的重要组成部分，分类的准确性在一定层面上决定了答案抽取的策略和准确性。本文所介绍的分类方法是以网络论坛中的用户提问作为样本集，并分析每种问题所含的比例，有侧重地提出一套适合于股票领域中问句分类的规则。系统将该领域问题分为概念、投资家、个股三大类。其中，概念类中又包括定义类 1、定义类 2、比较类、描述类、原因类、列举类、属性类等七类。初步测试表明这种分类方法有效地提高了股票领域问句分类的准确性。

**关键词：**问句分类；股票领域；规则

## 1. 引言

自然语言问答系统(简称问答系统或 QA 系统)，是近年来自然语言处理技术应用研究中受到广泛关注的热门课题之一。它允许用户以自然语言进行提问，而系统通过推理分析，从一定的信息来源提取文本信息，得到有效答案。

目前，英文的问答系统取得了不错的成绩。Text Retrieval Conference(TREC)会议中涌现了大批高性能、检索方法具有代表性的问答系统[1]。

中文在词语切分方面已经比较成熟，比如 LTP 是哈工大社会计算与信息检索研究中心开发的一套中文语言处理系统，于 2011 年 6 月 1 日开源[2]。中国科学研究院的 ICTCLAS 等汉语分词系统。但与英文相比，中文句子表达形式多样，语法更加多变复杂，规律难于琢磨，增大了系统分析难度。

由于中文问答系统缺乏相应的语料库、中文信息处理的许多基础性技术还没有突破，国内在这方面所投入的人力物力和重视程度不够等原因[3]，中文问答系统的发展比较缓慢。

北京理工大学自然语言处理实验室开发的银行领域自动问答系统 BAQS[4]是一个面向银行领域的中文问答系统，目前已经实现了面向某商业银行的个人业

 

务自动问答。BAQS 的问句分类方式更加接近于领域知识分类，着眼于运用本体技术。哈尔滨工业大学研究生开发的面向金融的问答系统[5]，根据问句的语义对问句进行分类，系统将金融领域问题分为 14 类，利用问题-答案对和 FAQ 库进行回答。山西大学所开发的面向农民的问答系统[6]，首先按照是否包含疑问词进行划分。含有疑问词的问句，以疑问词作为分类标准；不含疑问词的问句，采用规则表进行分类。

问句分类是问答系统实现的关键技术。问句分类的方法在一定程度上决定了答案的提取策略以及最后问题回答的准确率，对问答系统的错误结果进行分析证明，有 36.4%的错误是由于问句分类系统造成的[7]。针对问句分类，目前的研究主要集中于两个方面，一是基于规则的方法，二是基于统计的机器学习的方法。

1) 基于规则的方法

这种方法主要是通过提取各类问句类型的疑问词和相关词组的特征规则，通过规则来判断句子是所属哪种类型的[8]。此类方法相对简单，分类的速度较快，不需要大量的训练数据。但规则的整理及提取有很大的难度，而且不能灵活列举出所有的规则，扩展性不强。传统的基于规则分类主要移植了英语的问答技术。不过，在限定域问题的处理上，规则的适用性和适用性较强。

2) 基于统计的机器学习的方法

这种问题分类的方法通过对大量的问句进行统计学习，提取出能够代表各种问句类型的特征规则，并建立模型，实现对问题的分类。这种方法最具代表性的就是 SVM(支持向量机)算法进行英文问题分类。这种分类能保证问题分类的精度在 90%左右[9]。但是对于中文的问句分类，由于中文的提问方式灵活，分类特别复杂，就会产生很大的误差。

由于中文的灵活性远远大于英文，疑问词也种类繁多。传统的基于规则的问句的分类准确率只达到 57.57%[10]。而本论文所描述的规则主要结合了股票领域的知识结构和限定领域的特点，通过对样题进行分析统计，总结出一套有别于传统问句分类方法的方法，更适用于股票领域知识。同时我们对股票类问题进行了研究和实验，取得了不错的效果。

本系统采用哈尔滨工业大学的 LTP 进行中文分词和词性标注，利用基于规则的方法研究股票领域问句的处理。

## 2. 股票领域中的问句分类

本系统所面向的对象主要是拥有初级、中级股票知识的股民。作为限定域中文问答系统，本系统可以回答的问题，句式更加复杂多样，可以回答的问题种类更加丰富。

### 2.1. 股票领域中三大类问题

通常的中文问答系统会直接匹配语料库判断疑问词，进而对问句分类。但当我们深入到限定域知识中去研究，便发现并不是每类问题都有同样的重要等级。比如，在股票领域，用户很少提问关于地点的问题，而相对多的关注股票领域术语的定义、相似或相反术语的异同点比较，那么我们就无需花费过多精力在地点类问题上大做文章。针对股票领域知识的特点，本系统并没有首先直接根据疑问词对问句进行传统的分类，而是经过分析网上股票论坛和常见股票问题，统计潜在用户在股票领域的需求，从而将系统定位于三大类问题：基本概念类、个股信息类、投资家类。三类以外的其他类问题比例为 13.9%，包括一些判断句、复杂句式等，不在我们的研究范围之内。三类问题所占比例如图 1 所示。

为了实现类别的划分，系统需要建立三个对应的系统字典，通过将用户问句中的分词与字典词语一一比较匹配，将问句归入不同类别，并记录下匹配的专有名词 keyword(i)。对于查询字典的顺序有所要求，如果一个问句中同时出现基本概念字典和投资家字典中的词应该如何处理呢？按照三种分类所占比例不同，查询字典的顺序分别为：基本概念字典、上市公司和股票字典、投资家姓名字典。本系统不涉及复杂

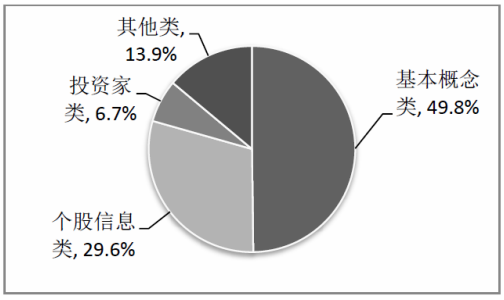


**Figure 1. The proportion of three types of questions**
**图 1. 三类问题所占比重**

 

的问句形式，因此不考虑问句中出现三个或三个以上专有名词的情况。

三大类问题的分类流程如图 2 所示。

经过初次匹配之后，我们首先将问句归入三大类中，进行进一步处理。

## 2.2. 基本概念类问题的类别细分

属于股票基本概念类问题范畴的问句繁多，根据中文词语的特征，我们将股票基本概念类问题进一步细分为：定义类 I、定义类 II、比较类、原因类、描述类、列举类、属性类等七类问题。相比于传统中文问答系统的问句分类，本系统增加了比较类类型，这符合用户对于股票对比的要求。同时，针对中文表达形式的复杂，我们制定了一套具有匹配优先级顺序的问句分类流程。本系统所使用的 431 道基本概念类问句的样题中，七类问题所占比重如图 3 所示。可以看出定义类、描述类问题较多，其他类问题相对较少。

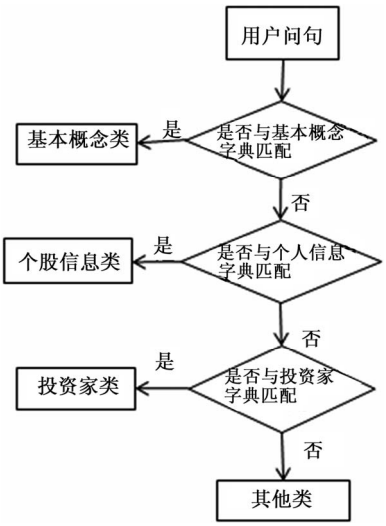


**Figure 2. The matching order of three types of questions**
**图 2. 三类问题的匹配顺序**

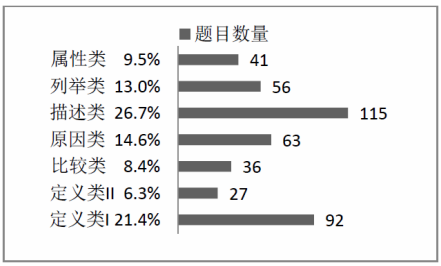


**Figure 3. Number comparison of seven types of basic conception class questions**
**图 3. 七类基本概念类样题数目对比**

考虑到中英文的差异，中文用词十分庞杂。比如，提问人物型问题，英文中疑问词 who 使用的频率极高，而中文的表达方式则包括“谁”“什么人”“何人”“哪个人”等众多词语，因此对特征词表的要求也相对严苛。利用特征词字典进行基本概念类问题的细分时，查询顺序的确定非常重要。例如，当出现“如何”一词，我们首先想到的大多是描述类问题，但少数问句如“如何理解”却又偏向于定义类型，所以对比查询特征词字典时，需要先查询“如何理解”一词，再查询“如何”一词，这也是定义类高、低优先级字典的由来。

分类方法通过将用户问句同特征表进行比较，找出问句所属类别，各分类的特殊词表见表 1。当用户问句被归入股票基本概念分支后，首先进行特殊词语的查找。查询顺序如下：

1) 若问句的分词中包含高优先级定义类字典中的词语，则将问句标记为“定义类 I”；

2) 若问句的分词中包含比较类字典中的词语，则将问句标记为“比较类”；

3) 若问句中包括原因类字典中的词语，去掉这些疑问词，将问句变为陈述句，并则将问句标记为“原因类”；

4) 若问句中包括描述类字典中的词语，则将问句标记为“描述类”；

5) 若问句中包括列举类字典中的词语，则将问句标记为“列举类”；

6) 若问句中包括低优先级定义类词典中的词语，则将问句标记为“定义类 II”；

7) 若问句中的名词数量小于等于两个则认为问句属于属性类；

8) 若问句不属于以上任意一个类别，则属于其他类别。

以上问句类别判断的优先级为：定义类 I，比较类，原因类，描述类，列举类，定义类 II，属性类，其他类。

## 2.3. 个股信息类和投资家类问句

经调查统计，个股信息类问句和投资家问句的句子形式相似，而且相对单一，几乎全部问题都可以基本上理解为属性类问题，按照属性类问题的处理方法进行以后的答案抽取。主要表达形式有两种：

 

**Table 1. Special words list**
**表 1. 特殊词表**

| 问句类型 | 特征词 | 例句 |
|---|---|---|
| 比较类 | “区别”“区分”“分辨”“划分”“不同”<br>“不一样”“一致”“相同点”“异同”“相对于”<br>“关系”“和”“还是”“关系” | 洗盘与出货如何区别；<br>要约收购与协议收购有何异同； |
| 原因类 | “为什么”“为何”“为啥”“原因” | 为什么会产生量价背离的现象；<br>中国为何没有做市商； |
| 描述类 | “怎么”“如何”“怎样”“怎么样” | 如何看 k 线图；<br>怎么计算除权除息价； |
| 列举类 | “哪些”“几”“哪个”“列举” | 哪些现象是庄家出货；<br>股票量价背离有几种情况； |
| 定义类 I<br>(高优先级) | “如何理解”“如何认识”“如看看待”“咋回事”<br>“怎么理解”“怎么认识”“怎么看待”“怎么回事” | 如何理解空头市场；<br>股权分置改革是怎么回事； |
| 定义类 II<br>(低优先级) | “定义”“意思”“什么”“何为” | 什么是量价指标；<br>机构投资的定义是什么； |
| 属性类 | 问句不属于以上类别并且问句中名词数量 $\leq 2$ 时 | 洗盘的目的；<br>洗盘的特征； |

① 具体的某个公司/投资家的属性信息
XX 公司的 XX(所属行业、分红等)是多少？
XX 投资家的 XX(投资理念、生平等)?

② 列举拥有具体属性的公司/投资家有哪些？
经营 XX 行业的公司有哪些？
XX(投资理念)的投资家有哪些？

③ 当问句中包含 2 个或 2 个以上公司/股票名称时，为比较类。

①中情形，经 3.1.节的三大类初步划分，便可以直接归入个股信息类或投资家类问题中进行处理。

②中情形相对复杂，因为用户期待的是一系列公司、股票、投资家的列举，而问句中并不出现具体的公司名、股票名、投资家名，因此不能经过 3.1.节的初步划分，分入某一大类之中去。目前我们没有想到更好的解决办法。但此类问题用户关注度很高，针对此类问题，建议通过下拉列表的形式，让用户选择相关属性，通过数据库筛选，提取符合属性条件的公司、股票、投资家，虽然用户体验相对较差。

③中情形，可以直接归入个股信息类或投资家问题的比较类问题中。

## 3. 系统测试

在这个部分我们对两个方面做了测试，一方面是问题分类的准确率，另一方面是整个系统的准确率。

### 3.1. 测试标准

问句分类为本文的主要内容，测试集合选取来自互联网的 550 道股票相关问题，以保证测试结果的可靠性。

系统测试作为辅助测试(因为结果由问题分类和答案抽取两部分共同影响)，我们选择 100 道问题作为测试集进行测试。由于网络中的问题有大量的冗余(例如：“什么是利空”与“什么是庄家”对系统来说是同样地问题)，去除过度冗余我们选择了 100 道具有代表性的问题作为测试集。

#### 3.1.1. 分类测试方法

我们选取了百度知道、炒股吧、金融界等股票论坛中的 550 道问题进行分类测试。对于每个问句具体进行测试方法如下：

问句中存在表 1 中的特殊词，若该问句实际属于的类别与特殊词所在类别相同，我们记录该句的分类为正确，否则记录为错误。

问句中存在公司或者股票名称，若该问句实际属于个股信息类问题，即可用属性解释(如：武钢股份的配股)或者可用基本信息解释(如：武钢股份的简介)，则我们记录该句分类为正确，否则记录为错误。

问句中存在投资家的名称，如该问句实际属于投资家类问句，即可以用属性解释(如：巴菲特的投资经

 

历)或者可用基本信息解释(如：巴菲特的简介)，则我们记录该句分类为正确，否则记录为错误。

对于其他类问句的正确率我们不做统计。

#### 3.1.2. 系统测试方法

我们对系统进行测试，实验的问题样题选择论坛中用户的提问共 100 条，每类提问 10~20 条。测试使用的方法为计算准确率(P)、召回率(R)、效率($F_1$)。这三个指标是信息检索领域最广泛使用的三个参数。

计算方式如下[11]：

$$\lambda\text{召回率} = \frac{\text{系统检索到的相关文件数}}{\text{系统返回的文件总数}} \quad (1)$$

$$\sigma\text{准确率} = \frac{\text{系统检索到的相关文件数}}{\text{相关文件总数}} \quad (2)$$

$$F_1 = \frac{2*\sigma*\lambda}{\sigma+\lambda} \quad (3)$$

公式(3)中的$\sigma$表示准确率，$\lambda$表示召回率。其中，针对我们基于 FAQ 库之上的问答系统而言，“系统返回的文件总数”指的是首次搜索出的文本总段落数；“相关文件总数”是系统返回的总段落中，存在关键词的段落数目；“系统检索到的相关文件数”则是经过问句问类、分析符合正确答案的段落数。

召回率的局限性主要表现在：系统中相关信息量究竟有多少一般是不确知的；另外，召回率或多或少具有“假设”的局限性，这种“假设”是指检索出的相关信息对用户具有同等价值，但对用户来说，信息的相关程度在某种意义上也许比它的数量重要得多。

准确率的局限性主要表现在：如果检索结果是题录式而非全文式，由于题录的内容简单，用户很难判断检索到的信息是否与课题密切相关；同时，准确率中所讲的相关信息也具有“假设”的局限性。

### 3.2. 测试结果

分类测试与系统测试结果统计表分别在表 2 和表 3 中。

### 3.3. 测试分析

从表 2 中我们可以看出本文所使用的分类规则对基本概念类问题的分类结果是最为突出的，正确率达到 89.3%。相比之下个股信息和投资家的分类结果正确率稍微低一些，不过效果还是比较理想的。

表 3 为我们将 100 道题分别在 3 个系统中搜索的结果进行的统计。从表中可以很明显的看出我们系统的优势。首先由于，我们采取了结构化数据库的策略，明显的提高了准确率，我们系统的准确率有效提高约 20%左右。除此之外本系统的召回率与效率也优于其他两个网站结果很多。由于我们所使用的服务器运行在普通 pc 机而非高速的服务器机上，所以求得结果的速度要慢一些。不过相信高性能的服务器能使我们系统的运行运行时间的缩短很多。

## 4. 总结

本系统结合对股票领域问题的统计分析，按照用户需求划分了三大类问题，在各类问题的细致划分中又提出适用于限定域的更加细致完善的分类规则，降低了对自然语言处理的难度，从而提高问答的准确程度，大大提高了问答系统的实用性。

哈尔滨工业大学的面向真实环境的金融问答系统，按照语义进行问句分类，虽然细致，但系统在分类标准和依据方面的说明略显模糊。北京理工大学自然语言处理实验室开发的 BAQS 系统采用基于规则的问句分类方式，虽然偏重了对限定域问题的分析，但划分中没有比较类，也没有像本文提到的，首先根据用户需求初步划分问句。山西大学开发的面向农民的问答系统，依然采取传统的基于规则方法，以疑问词作为分类标准。

**Table 2. Testing results of question classification**
**表 2. 分类测试结果统计表**

| 分类 | 总数 550(道) | 正确率 |
|---|---|---|
| 基本概念 | 274 | 89.3% |
| 个股信息 | 163 | 76.9% |
| 投资家 | 37 | 62.2% |
| 其他 | 76 | - |

**Table 3. Testing results of the system**
**表 3. 系统测试结果统计表**

| 系统 | 准确率 | 召回率 | 效率 |
|---|---|---|---|
| 本系统 | 56.98% | 61.78% | 59.28% |

 

经过随机抽取网络论坛中用户问题进行初步测试，系统问句分类部分的准确率、整个系统的召回率、准确率和效率均有不错的效果。本文所提出的问句分类方法在股票领域还是有一定的适用性。

系统需要进一步研究的是问句中动词的含义、深入的语义关系和命名实体之间关系，从而更灵活准确地实现问句分类。

## 参考文献 (References)


[1] E. Voorhees. The trec-8 question answering track report. Proceedings of TREC, 1999.

[2] W. X. Che, Z. H. Li and T. Liu. LTP: A Chinese language technology platform. Beijing: Proceedings of the Coling 2010: Demonstrations, 2010: 13-16.

[3] 李正华, 车万翔, 刘挺. 基于XML的语言技术平台[URL]. 第五届全国青年计算语言学研讨会(YWCL), 2010. http://ir.hit.edu.cn/ir_papers/2010/%E5%9F%BA%E4%BA%8E%20XML%20%E7%9A%84%E8%AF%AD%E8%A8%80%E6%8A%80%E6%9C%AF%E5%B9%B3%E5%8F%B0.pdf

[4] 樊孝忠, 李宏乔, 李良富等. 银行领域汉语自动问答系统BAQS 的研究与实现[J]. 北京理工大学学报, 2004, 24(6): 528-532.

[5] 辛霄. 面向真实环境的金融问答系统[M]. 哈尔滨: 哈尔滨工业大学, 2009.

[6] 贾君枝, 王永芳, 李婷. 面向农民的问答系统问句处理研究[J]. 现代图书情报技术, 2010, 5: 35-40.

[7] 张志昌, 张宇, 刘挺, 李生. 基于线索词识别和训练集扩展的中文问题分类[J]. 高科技通讯, 2009, 19(2): 111-118.

[8] 程显毅, 朱倩, 韩飞. 基于 HNC 和描述逻辑的问句语义块分析[J]. 广西师范大学学报(自然科学版), 2010, 28(3): 131-134.

[9] 文勖, 张宇, 刘挺, 马金山. 基于句法结构分析的中文问题分类[J]. 中文信息学报, 2006, 20(2): 33-39.

[10] 杜永萍, 黄萱菁. 开放领域的 QA 系统结构及性能分析[J]. 2009, 22(4): 527-531.

[11] 孙昂, 江铭虎, 贺一帆, 陈林, 袁保宗. 基于句法分析和答案分类的中文问答系统[J]. 电子学报, 2008, 36(5): 833-839.


 