(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105279274 B
(45)授权公告日 2018.11.02

(21)申请号 201510729331.3

(22)申请日 2015.10.30

(65)同一申请的已公布的文献号
申请公布号 CN 105279274 A

(43)申请公布日 2016.01.27

(73)专利权人 北京京东尚科信息技术有限公司
地址 100080 北京市海淀区杏石口路65号西杉创意园四区11C楼东段1-4层西段1-4层
专利权人 北京京东世纪贸易有限公司

(72)发明人 黄靖锋

(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 李宝泉 周亚荣

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
CN 101178711 A,2008.05.14,
CN 101377777 A,2009.03.04,
US 2010/0235164 A1,2010.09.16,
US 2009/0070311 A1,2009.03.12,
CN 102184225 A,2011.09.14,
CN 103902652 A,2014.07.02,
CN 104050256 A,2014.09.17,
胡海峰.用户生成答案质量评价中的特征表示及融合研究.《中国优秀硕士学位论文全文数据库 信息科技辑》.2014,(第03期),I138-1187.
周志敏.非事实类问题问答模型和特征的研究.《中国优秀硕士学位论文全文数据库 信息科技辑》.2013,(第03期),I138-1800.

审查员 李梦颖

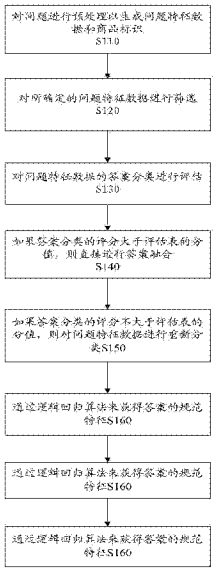
权利要求书1页 说明书6页 附图2页

(54)发明名称

基于自然语义问答系统的答案合成与匹配的方法和系统

(57)摘要

提供了一种基于自然语义问答系统的答案合成与匹配的方法，包括：从问题确定问题特征数据；对所确定的问题特征数据进行筛选；对问题特征数据中的答案分类进行评估；如果答案分类的评分大于评估表的分值，则直接进行答案融合；如果答案分类的评分不大于评估表的分值，则对问题特征数据进行分类；通过逻辑回归算法来获得答案的规范特征；以及对所获得的答案进行评分和数据处理。



CN 105279274 B

1.一种基于自然语义问答系统的答案合成与匹配的方法,包括:

从问题确定问题特征数据;

对所确定的问题特征数据进行筛选;

对问题特征数据中的答案分类进行评估;

如果答案分类的评分大于评估表的分值,则直接进行答案融合;

如果答案分类的评分不大于评估表的分值,则对问题特征数据进行重新分类;根据重新分类后的问题特征数据抽取相关的答案特征,然后通过逻辑回归算法来获得答案的规范特征,以得到答案;以及

对所获得的答案进行评分和数据处理。

2.根据权利要求1所述的方法,其中,通过确定所确定的问题特征数据的支持度和置信度来进行筛选。

3.根据权利要求1所述的方法,其中,评估表初始地被设置为空,并且在训练或使用期间不断更新。

4.根据权利要求1所述的方法,其中,直接进行答案融合包括:首先确定与答案类型相对应的所有答案,对所确定的答案进行排序,然后选择出最适当的答案。

5.根据权利要求1所述的方法,其中,通过决策树来对问题特征数据进行重新分类。

6.一种基于自然语义问答系统的答案合成与匹配的系统,包括:

用于从问题确定问题特征数据的装置;

用于对所确定的问题特征数据进行筛选的装置;

用于对问题特征数据中的答案分类进行评估的装置;

用于如果答案分类的评分大于评估表的分值,则直接进行答案融合的装置;

用于如果答案分类的评分不大于评估表的分值,则对问题特征数据进行重新分类的装置;

用于根据重新分类后的问题特征数据抽取相关的答案特征,然后通过逻辑回归算法来获得答案的规范特征,以得到答案的装置;以及

用于对所获得的答案进行评分和数据处理的装置。

7.根据权利要求6所述的系统,其中,通过确定所确定的问题特征数据的支持度和置信度来进行筛选。

8.根据权利要求6所述的系统,其中,评估表初始地被设置为空,并且在训练或使用期间不断更新。

9.根据权利要求6所述的系统,其中,用于直接进行答案融合的装置包括:用于首先确定与答案类型相对应的所有答案,对所确定的答案进行排序,然后选择出最适当的答案的装置。

10.根据权利要求6所述的系统,其中,通过决策树来对问题特征数据进行重新分类。

基于自然语义问答系统的答案合成与匹配的方法和系统

技术领域

[0001] 本发明涉及一种基于自然语义问答系统的答案合成与匹配的方法和系统。

背景技术

[0002] 答案合成是人工智能问答系统中的关键。在传统的人工智能问答系统中，通常只能针对限定的问题范围做出指定的固定回答。然而，对于快速变化的电子商务领域，人工智能问答系统所要服务的对象大多是购买商品的消费者，普通消费者大多希望对商品信息有更全面的了解，如果回答的信息不准确，则可能会导致销售商品失败。普通消费者在问答系统中提出的问题，是更接近于自然语言的。因此，能否准确的回答来应对多变的自然语言，这就给自然语义问答系统的答案的合成准确性带来了巨大的挑战。

[0003] 对问题的回答，现有的技术方案通常利用对问题提取关键字和分类后直接进行答案的生成。

[0004] 例如，在图1中示出了现有技术的典型系统的框图。

[0005] 图1的系统主要由问题分析模块、答案处理模块组成，其利用从问题分析出的分类直接来进行答案的生成。

[0006] 问题分析模块负责对语句进行分类、实体提取等操作。当问话被输入问答系统后，首先会对问题进行一次分类，分类是在系统预先设定好的一些固定的类别，用以回答用户可能提出的这一类别的问题，然后生成需要的实体特征和分类信息，最后再保存商品里提取出来的 信息，并提供给答案处理模块处理。

[0007] 答案处理模块负责搜索当前保存的分类信息，找到对应的处理模块，根据设定好的规则直接生成答案并返回。

[0008] 直接合成答案方法通过问题分析得到特定的一个分类，针对这个分类回答的都是固定的一个或多个答案。

[0009] 现有技术首先由问题分析模块进行分类并提取的分类传给答案处理模块，答案处理模块在答案的合成上只是机械的一一对应，一旦问题分析模块出现分类错误时，便只能回答错误的答案，对于问题表现形式不同而意思相同的问题，答案处理模块无法进行优化反馈。

[0010] 现有技术方案的主要缺点是当有提问方式发生变化时，现有的回答不能自动地进行变化和调整，也不能检测出分类错误，回答准确度不理想。

[0011] 自然语言的提问表现形式多种多样，可能同一个问题有多种含义，或者不同的问题有同一个含义，传统的答案合成技术通常只是提取出问题的关键字来进行答案的匹配，而没有对答案本身的特征进行分析，答案与问题的匹配度也没有进行过优化调整，只是线性地从问题出发，然后回答对应答案。这样会导致问题一旦发生改变以后，用户真实的意图不能通过多次的交互回答得到反馈积累，从而也不能提高回答的准确性。

[0012] 人类自然语言的多变性必然会导致问题特征的频繁变化，加之电子商务行业商品数量动辄以百万计数，越来越多的用户在购买前和购买后都需要咨询商品相关信息。因此，

问答系统采用单一静态的答案合成方法必然造成问题分类偏差增大，引发答案回答准确性的降低，降低用户满意度。

[0013] 因此，期望提供一种基于自然语义问答系统的答案合成与匹配的方法和系统。

发明内容

[0014] 为了解决现有技术中的上述缺点和问题中的至少一个而提出本发明。基于现有技术存在的缺点，本发明提供了一种的方法和系统。

[0015] 基于现有技术存在的缺点，本发明提出了一种改进的基于自然语义问答系统的答案合成与匹配的方法和系统，一方面可以在问题特征发生变化时进行分类的动态适配，而不是在分类错误已经发生时才进行分类选择和答案特征参数调优，减小答案选择的滞后性；另一方面，通过对当前的问题特征匹配最合适的答案分类特征，也可以大大提高答案合成结果的精度。本发明弥补了基于自然语义分析中答案合成方法的不足，针对电子商务行业的特点更准确的回答用户提问的相关信息，对于降低人工服务的成本也起到了促进作用。

[0016] 根据一个方面，提供了一种基于自然语义问答系统的答案合成与匹配的方法，包括：从问题确定问题特征数据；对所确定的问题特征数据进行筛选；对问题特征数据中的答案分类进行评估；如果答案分类的评分大于评估表的分值，则直接进行答案融合；如果答案分类的评分不大于评估表的分值，则对问题特征数据进行分类；通过逻辑回归算法来获得答案的规范特征；以及对所获得的答案进行评分和数据处理。

[0017] 可选地，通过确定所确定的问题特征数据的支持度和置信度来进行筛选。

[0018] 可选地，评估表初始地被设置为空，并且在训练或使用期间不断更新。

[0019] 可选地，直接进行答案融合包括：首先确定与答案类型相对应的所有答案，对所确定的答案进行排序，然后选择出最适当的答案。

[0020] 可选地，通过决策树来对问题特征数据进行重新分类。

[0021] 根据另一个方面，提供了一种基于自然语义问答系统的答案合成与匹配的系统，包括：用于从问题确定问题特征数据的装置；用于对所确定的问题特征数据进行筛选的装置；用于对问题特征数据中的答案分类进行评估的装置；用于如果答案分类的评分大于评估表的分值，则直接进行答案融合的装置；用于如果答案分类的评分不大于评估表的分值，则对问题特征数据进行分类的装置；用于通过逻辑回归算法来获得答案的规范特征的装置；以及用于对所获得的答案进行评分和数据处理的装置。

[0022] 可选地，通过确定所确定的问题特征数据的支持度和置信度来进行筛选。

[0023] 可选地，评估表初始地被设置为空，并且在训练或使用期间不断更新。

[0024] 可选地，用于直接进行答案融合的装置包括：用于首先确定与答案类型相对应的所有答案，对所确定的答案进行排序，然后选择出最适当的答案的装置。

[0025] 可选地，通过决策树来对问题特征数据进行重新分类。

附图说明

[0026] 通过下面结合附图进行的描述，本发明一些示范性实施例的上述和其他方面、特征和优点对于本领域技术人员来说将变得显而易见，其中：

[0027] 图1示出了现有技术的典型系统的框图;以及

[0028] 图2是根据本发明的一个实施例的基于自然语义问答系统的答案合成与匹配的方法的流程图。

具体实施方式

[0029] 提供参考附图的下面描述以帮助全面理解本发明的示范性实施例。其包括各种细节以助于理解,而应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改而不会背离本发明的范围和精神。同样,为了清楚和简明,省略了对公知功能和结构的描述。

[0030] 图2是根据本发明的一个实施例的基于自然语义问答系统的答案合成与匹配的方法的流程图。

[0031] 在步骤S110中,对问题进行预处理以生成问题特征数据和商品标识。

[0032] 例如,当一用户输入的问题是:我想买三星s6手机。可以对该问题进行分词,例如可以将上面的问题分词为[我]-[想买]-[三星]-[S6]-[手机]。然后,可以通过所分的词确定商品标识或商品ID,例如可以将所分的词依次在数据库中(商品数据库中)进行匹配查找来确定商品标识或商品ID。例如,在上面示例的问题中,可以根据所分的词“S6”来确定商品标识或商品ID为35的商品。最后,可以以预定格式来输出预处理的结果。例如,可以格式{data:[我:0]-[想买:1]-[三星:2]-[S6:3]-[手机:2],id:35},其中,0代表指代词,1代表动词,3代表商品,2代表名词,id代表商品的ID号。当然,上面仅仅是一个示例,可以任何其它适当的格式来输出预处理的结果。

[0033] 在步骤S120中,对所确定的问题特征数据进行筛选。例如,可以 对在步骤S110中确定的问题特征数据和商品标识进行关联性规则检查,即采用关联规则算法,以获得置信度和支持度,关联规则算法的原理如下。设I＝{I1,I2…,Im}为所有问题特征的集合,设A是一个由问题特征构成的集合,称为问题特征集。事务T是一个问题特征子集,每一个事务具有唯一的事务标识Tid。事务T包含问题特征集A,当且仅当AT。如果问题特征集A中包含k个问题特征,则称其为k个问题特征集。D为事务数据库,问题特征集A在事务数据库D中出现的次数占D中总事务的百分比叫做问题特征集的支持度(support)。关联规则就是形如XY的逻辑蕴含关系,其中XI,YI且XY＝Φ,X称作规则的前件,Y是结果,对于关联规则XY,存在支持度和置信度。支持度是指规则中所出现模式的频率,如果事务数据库有s%的事务包含XY,则称关联规则XY在D中的支持度为s%,实际上,可以表示为概率P(XY),即support(XY)＝P(XY)。置信度是指蕴含的强度,即事务D中c%的包含X的交易同时包含XY。若X的支持度是support(x),规则的置信度为即为:support(XY)/support(X),这是一个条件概率P(Y|X),即confidence(XY)＝P(Y|X)。由于通过关联规则算法来计算支持度和置信度在本领域是已知的,在此不再进行详细描述。

[0034] 然后可以将支持度和置信度进行比较来对对所确定的问题特征数据进行筛选。例如,如果支持度和置信度都大于相应的阈值,则可以确定所确定的问题特征数据符合要求。如果支持度和置信度中的至少一个不大于相应的阈值,则可以确定所确定的问题特征数据不符合要求。阈值可以根据经验适当地进行设置。作为一个示例,支持度的阈值可以设为95%,置信度的阈值可以设为80%。当然可以将阈值设置为任意其它适当值。

[0035] 在步骤S130中，对问题特征数据的答案分类进行评估。例如，可以通过确定该答案分类的评分，然后将该评分与评估表的分值进行比较。

[0036] 评估表可以初始地被设置为空，并且在训练或使用期间不断更新。

[0037] 例如，如果所确定的评分大于评估表的分值，则可以输出{data:[我:0]-[想买:1]-[三星:2]-[S6:3]-[手机:2]，id:35-[80％,95％]，pg:true}，其中pg:true代表所确定的评分大于评估表的分值，如果不大于则为false。

[0038] 在步骤S140中，如果答案分类的评分大于评估表的分值，则直接进行答案融合。

[0039] 例如，首先确定与答案类型相对应的所有答案，对所确定的答案进行排序，然后选择出最适当的答案。

[0040] 例如，(1)可以通过直接融合生成的答案：{id:35-[80％,95％]，pg:true，answerid:80，answer:[这款:0][手机:1][有货4][哦:5][top:1]}，其中top:1代表排序结果；(2)通过重新分类获得的答案：{id:35-[80％,95％]，pg:false，newAnswer_id:100，answer:[这款:0][手机:1][有货4][哦:5][top:0]}，其中top0代表，新生成的答案；(3)没有找到答案：id:35-[80％,95％]，pg:true，defaultAnswer:yes，answer:[手机还没有找到，请联系客服MM哦][top:-1]}，其中top-1代表默认答案，或{id:35-[80％,95％]，pg:false，defaultAnswer:yes answer:[手机还没有找到，请联系客服MM哦][top:-1]}，其中top-1代表默认答案。

[0041] 在步骤S150中，如果答案分类的评分不大于评估表的分值，则对问题特征数据进行重新分类。例如，可以通过决策树来对问题特征数据进行重新分类。

[0042] 根据训练数据集，从根结点开始，递归地对每个结点进行以下操作，构建二叉决策树：

[0043] (1)设训练数据集为D，计算现有特征对该数据集的基尼指 数，对每一个特征A，对其可能取的每个值a，根据样本点对A＝a的测试为是或否将D分割成D1和D2两部分，

[0044] 分类问题中，假设有k个类(例如，在上面的示例中，共有4个类，即代词、动词、商品、名词)，样本点属于第k类的概率为p_k，则概率分布的基尼指数定义为：

[0045] $$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$$

[0046] 利用上面公式，计算A＝a时的基尼指数。

[0047] (2)在所有可能的特征A以及它们所有可能的切分点a中，选择基尼指数最小的特征及其对应的切分点作为最优特征与最优切分点。例如，如果[我:0,0.9]-[想买:1,0.95]-[三星:2,1]-[S6:3,3]-[手机:2,1.2]，则[我:0,0.9]中前一个数字0就是最小的特征，0.9就是最优切分点。依最优特征与最优切分点。从现结点生成两个子结点，将训练数据集依特征分配到两个子结点中去。

[0048] (3)对两个自己点递归地调用(1)、(2)，直至满足停止条件([我:0,0]-[想买:1,1]-[三星:2,2]-[S6:3,3]-[手机:2,2])。

[0049] (4)生成决策树

[0050] 根据决策树得到重新分类结果。

[0051] 在步骤S160中，通过逻辑回归算法来获得答案的规范特征。例如，通过现有的问题

特征和标准特征进行对比，规范化为统一的问题特征，在答案特征库中找到对应的问题答案特征对应表，将所有相关的答案特征都抽取出来，这时开始细化答案特征，采用逻辑回归的算法，得到答案的规范特征。

[0052] 逻辑回归算法的原理是：

[0053] 对于二元变量的目标变量(答案特征正确还是不正确)来说，逻辑回归的目的就是要预测一组自变量(多个筛选后的特征)相对于目标变量的概率，这个概率P是介于[0,1]之间的，其计算公式如下：

[0054] $$P(y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}}$$

[0055] 其中，β_0是常数，β_1到β_k是自变量x_1到x_k各自所对应的系数。

[0056] 这里是为了将多个答案的特征向量统一起来，方便计算。

[0057] 在步骤S170中，对所获得的答案进行评分和数据处理。如果是直接融合的答案，则记录当前的评分结果；如果是重新生成特征的答案，则根据获得的评分来更新评分表，使得得出的匹配结果可以用于下一次的评估。

[0058] 模块具体实现的技术原理

[0059] 采用ROC曲线算法来评价正确答案和错误答案准确性(答案都会由人工来标记是正确的还是错误的)，算法原理是：

[0060] 对一个二分问题(正确和错误)来说，会出现四种情况。如果一个答案是正确的并且也被预测成正确，即为真正类(True Positive，缩写TP)，如果答案是错误的被预测成正确的，称之为假正类(False Positive，缩写FP)。相应地，如果答案是错误的被预测成错误的，称之为真负类(True Negative，缩写TN)，如果答案是正确的被预测成错误的则为假负类(false negative，缩写FN)。

[0061] 列联表如下表所示，1代表正类，0代表负类。

[0062]

		预测		合计
		1	0	
实际	1	TP	FN	TP+FN
	0	FP	TN	FP+TN
合计		TP+FP	FN+TN	TP+FP+FN+TN

[0063] 根据上面的分析方法，得到横坐标和纵坐标(x,y轴)

[0064] 正例分对的概率＝TP/(TP+FN)——x轴

[0065] 负例错分的概率＝FP/(FP+TN)——y轴

[0066] 需要乘以100，表示为[0,100]区间的值，这个值也就是每个答案的评分的值。

[0067] 将所有得到的取值分数的数据分为：N为正样本数(正例分对的概率的答案总数)，M为负样本数(负例错分的概率)，我们首先把所有样本按照得分排序，依次用rank表示他们，如最大得分的样本，rank＝n(n＝N+M)，其次为n-1。那么对于正样本中rank最大的样本，rank_max，有M-1个其他正样本比他得分小，那么就有(rank_max-1)-(M-1)个负样本比他得

分小。其次为(rank_second-1)-(M-2)。最后我们得到正样本大于负样本的概率为

[0068] $$P=\frac{\sum_{i=1}^{n} rank_i-\frac{M(M+1)}{2}}{M\times N}$$

[0069] 得到的P值越大说明答案评分模型预测效果越好,这里计算出的值的区间是[0,1],值越大预测效果就越好。

[0070] 例如,如果pg=true,输出:{answerid:80,pf:75分}

[0071] 如果pg=false,输出:{newAnswer_id:100,pf:75分}

[0072] 在步骤S180中,对当前已生成的经过评价过的问题和答案发送到答案分类评价库中,使下一次的答案分类评价模块选用最新的特征来评估分类结果。例如,可以将经评分获得的结果{answerid:80,pf:75分}存储到答案分类评价库中。

[0073] 此外,本发明还可以涉及与上述方法相对应的用于基于自然语义问答系统的答案合成与匹配的系统。

[0074] 虽然本说明书包含许多特定实施方式细节,但是不应当将这些细节解释为对任何发明或可以主张的内容的范围的限制,而应当解释为对可以特定于特定发明的特定实施例的特征的描述。还可以将在本说明书中在分离的实施例的情境中描述的某些特征组合在单个实施例中实现。相反地,也可以将在单个实施方式的情境中描述的各个特征分离地在多个实施方式中实现或在任何适当的子组合中实现。此外,尽 管可能在上面将特征描述为在某些组合中起作用,甚至最初主张如此,但是可以在一些情况下将来自所主张的组合的一个或多个特征从组合中删去,并且可以将所主张的组合指向子组合或者子组合的变体。

[0075] 类似地,虽然在附图中以特定次序描绘了操作,但是不应当将这理解为需要以所示的特定次序或者以连续次序执行这样的操作、或者需要执行所有图示的操作才能达到期望的结果。在某些情况下,多任务以及并行处理可以是有利的。此外,不应当将在上述实施例中的各种系统组件的分离理解为在所有实施例中均需要这样的分离,而应当理解的是,通常可以将所描述的程序组件和系统集成到一起成为单个软件产品或封装为多个软件产品。

[0076] 计算机程序(也称作程序、软件、软件应用、脚本或代码)可以以任何形式的编程语言编写,所述编程语言包括编译或解释语言、或者说明性或过程语言,并且其可以以任何形式部署,包括作为独立程序或作为模块、组件、子程序或适于在计算环境中使用的其它单元。计算机程序没有必要对应于文件系统中的文件。可以将程序存储在保持其它程序或数据的文件(例如,存储在标记语言文档中的一个或多个脚本)的一部分、专用于讨论中的程序的单个文件或者多个协调文件(例如,存储一个或多个模块、子程序或部分代码的文件)中。

[0077] 上述具体实施方式,并不构成对本发明保护范围的限制。本领域技术人员应该明白的是,取决于设计要求和其他因素,可以发生各种各样的修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等,均应包含在本发明保护范围之内。

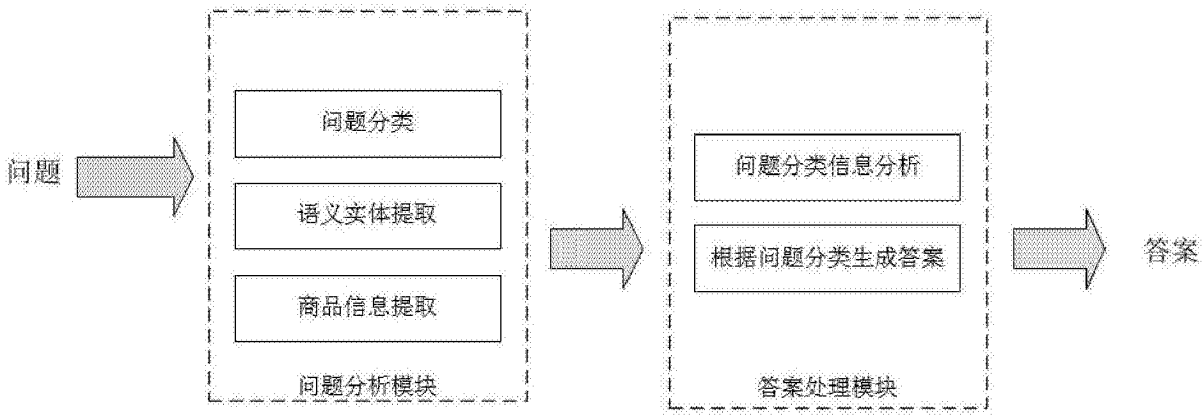

问题
问题分类
语义实体提取
商品信息提取
问题分析模块
问题分类信息分析
根据问题分类生成答案
答案处理模块
答案

图1

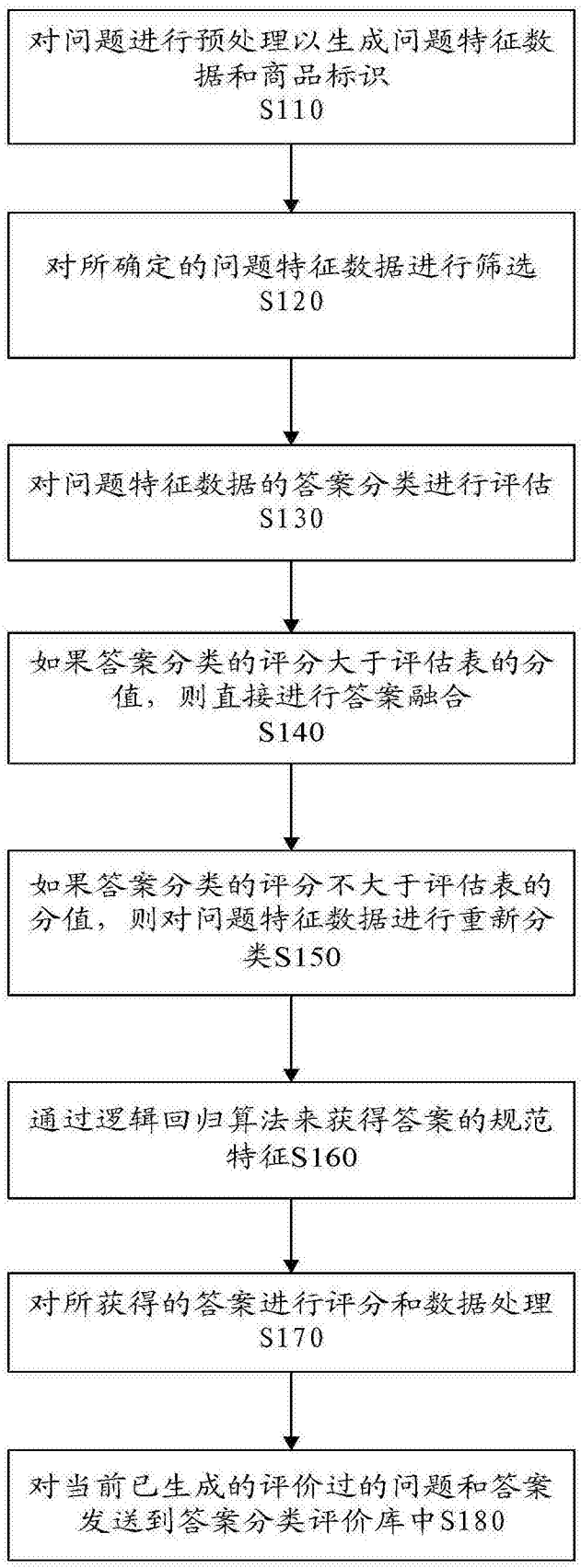

对问题进行预处理以生成问题特征数据和商品标识
S110
对所确定的问题特征数据进行筛选
S120
对问题特征数据的答案分类进行评估
S130
如果答案分类的评分大于评估表的分值，则直接进行答案融合
S140
如果答案分类的评分不大于评估表的分值，则对问题特征数据进行重新分类S150
通过逻辑回归算法来获得答案的规范特征S160
对所获得的答案进行评分和数据处理
S170
对当前已生成的评价过的问题和答案发送到答案分类评价库中S180

图2