基于遗传算法的问答系统优化方法

一 背景

1.1 问答系统

近年来，随着自然语言处理技术的发展，智能问答系统受到了极大的关注，从聊天软件‘小黄鸡’的风靡，到流行于各大网络平台的应答机器人，智能问答系统在众多领域得到应用。一个优质的问答系统解决客户常见的问题，降低人工开销，并能提供 24 小时无间断服务。

在限定领域问答系统中(如政府事务问答系统)，通常采用基于问题答案对的问答系统。如图 1，基于问题答案对的问答系统首先对用户输入的问题进行分析，包括关键词提取，关键词扩展等一系列的预处理。然后将预处理的问题输入问题检索模块进行问题匹对，包括 lucene 检索，模糊匹配等文本相似度计算方法。最后对检索出来的匹对问题进行筛选和选择，选择出最优的答案。基于问题答案对的问答系统由于高效和简单，已经广泛应用在限定领域和社区问答中，如百度知道，智能客服等具体的应用。

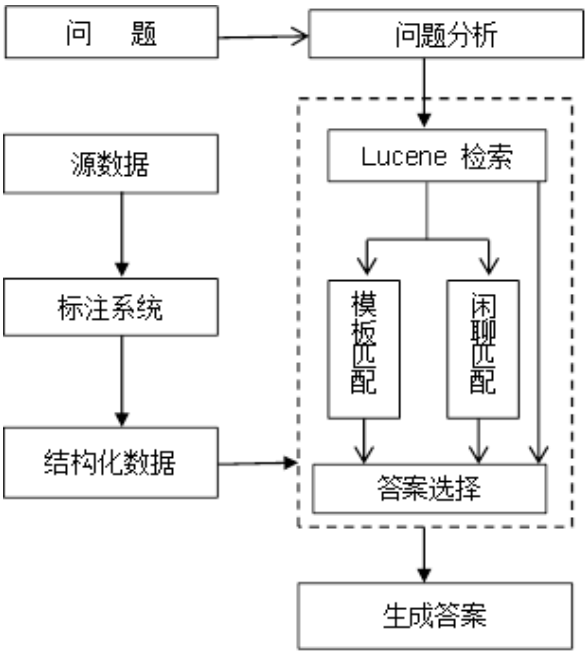


图 1 问答系统流程

1.2 遗传算法

遗传算法（Genetic Algorithms ,GA）是一类借鉴生物界自然选择和自然遗传机制的随机化搜索算法。它模拟自然选择和自然遗传过程中发生的繁殖、交叉和基因突变现象，在每次迭代中都保留一组候选解，并按适度值评估函数从解群中选取较优的个体，利用遗传算子（选择、交叉和变异）对这些个体进行组合，产生新一代的候选解群，重复此过程，直到满足某种收敛指标为止如图 2。遗传算法作为一种优化参数的方法，已经广泛应用与各种领域。

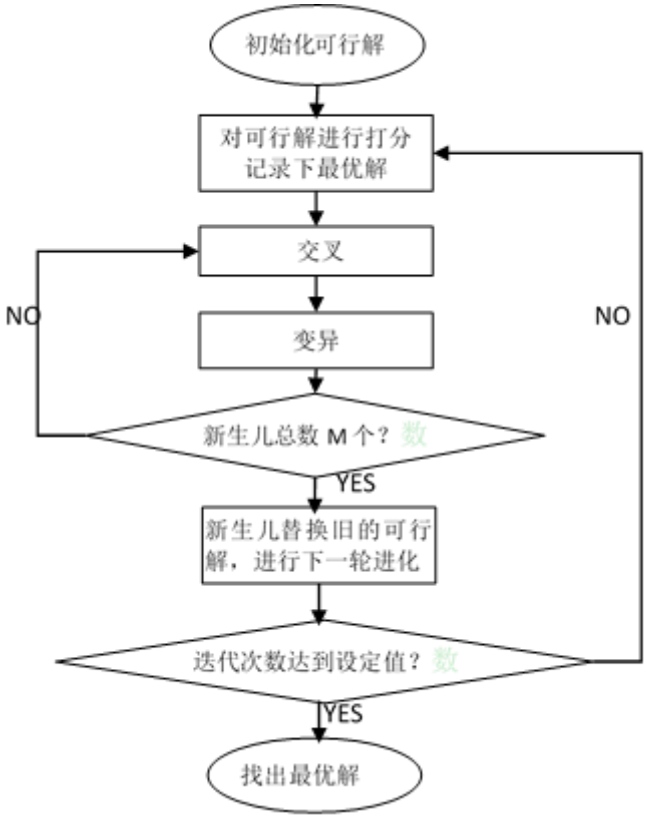


图 2. 遗传算法流程

二 问题描述

基于问题答案对的问答系统具有知识整理简单，搜索高效的特点，被广泛应用在智能问答系统中。然而，在限定领域的复杂搜索中，其简单高效带来了一定的问题。首先，在一般的系统中，通常采用一种信息检索的方法(vsm)进行信息检索，由于每一种方法都有其缺点（如 vsm 不能解决语义连续的信息检索），不能适应多种类型的问题检索，因此需要在问答系统中引入多种检索方法。但是，如果引入多种信息检索的方式，会造成检索结果的多样性且不同搜索方法的评价尺度不一，无法简单的进行排列比较。第二，在问答系统中往往会有不同问题类型的数据，在检索中需要同时对不同类型的问题进行检索。如在政府事务中，既有政府事务的问题，也有用户的一些闲聊类型的问题，在检索过程中其重要性不相同和问题格式不相同，因此无法采用统一尺度进行比较选择。

综上可以看出，在包含不同检索方法和不同类型的问答对的限定领域问答系统中，如何选择一个有效的且优化的评价因子，对不同检索方法和不同类型问题进行合理选择显得的尤为重要。并且，有效的因子不仅可以组合多种方法和数据类型，还可以提高问答系统的性能和用户的体验。

三 发明要点

本发明提出一种将遗传算法和问答系统进行有效结合的方案和系统。该方案首先利用多种信息检索的方法来优化和增强问答系统对问题的检索性能，提高查准率和查全率，并且对不同类型的问题进行区分检索，提高系统对限定领域的检

索能力。然后，利用遗传算法在开发集上对以上不同评价因子进行优化选择，选择系统最优的参数。具体而言，该发明主要包含以下内容：

1. 集成多种信息检索方法的问答系统。基于问答对的传统的问答系统一般采用单一的信息检索方法对相似问题进行检索。当问答对比较复杂时，单一的检索方法不能满足检索要求，如词袋模型对语序有要求的问题无法检索。本发明提出一种集成多种检索方法的问答系统，有效的提高了系统的查准率和查全率。

2. 区分不同类型问题的检索方法。在传统的限定领域问答系统中，一般是直接从所有的问答对中进行检索，不区分不同类型的问答类型。但当用户咨询特定领域的问题时（如政府事务），用户关心的往往是限定领域的问题，对于一些闲聊问题不是很关心。为了解决此问题，本发明提出了一种不同类型问题的区分检索方法，针对不同类型问题进行区别检索，然后根据领域进行特殊选择，从而有效提高系统限定领域内的查准率和查全率。

3. 基于遗传算法的评价因子优化方法。在以上的两个方法中，问答系统中集成了多种评价标准，包括多种信息检索方法的评价因子和检索不同类型问题的评价因子，因此需要选择一个最优的组合评价标准，提高系统的整体性能。本发明提出了一种利用遗传算法进行参数选择的方法，可以根据不同的领域开发集，自动选出系统最优评价因子。

四 发明内容和系统实现

4.1 集成多种信息检索方法的问答系统

图 1 为问答系统的构架示意图。该系统分为三个模块：问题预处理或问题分析，信息检索和答案筛选。在本发明中，将在信息检索模块和答案筛选模块中集成多种信息检索方法。如图 3 所示，首先利用 Lucene 工具从数据集上检索出

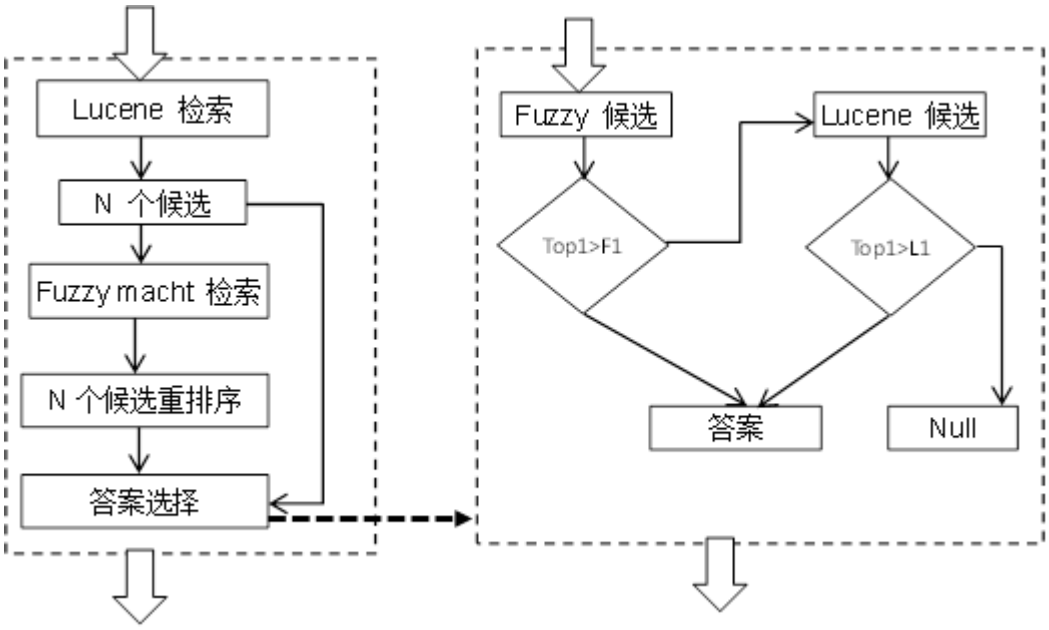


图 3 集成多种信息检索方法的问答系统

N 个候选，然后利用模糊匹配算法进行 N 个候选的重新打分排序。得到两个候选列表后，首先查看模糊匹配的候选列表总的 Top1 的分数是否大于阈值 F1，如果大于阈值则得到答案。否则查看 Lucene 的 top1 是否大于阈值 L1，如果大于阈值得到答案，否则得到的答案为空。经过多种搜索方式的检索，可以充分利用 Lucene 检索和模糊检索的优势。由于 lucene 检索具有快速高效的特点，在系统中利用模糊匹配的方法对 lucene 检索的候选进行重新排序，减少计算量从而提高效率。另外，模糊匹配可以对不同语序的问题进行检索，从而弥补 Lucene 检索的缺点。

4.2 区分不同类型问题的检索方法

在图 3 的系统构架中，将所有的问题类型统一检索处理，没有对类型问题进行区分检索。为了解决此问题，本发明提出一种区分不同类型问题的检索方法。如图 4，在 lucene 检索获得候选之后，将候选结果按照不同的类型：领域问题和闲聊问题，分别用模糊匹配的方法进行重新打分排序，获的不同类型问题的两个候选列表。由于特定领域的问答系统倾向回答用户领域内的问题，本发明制定了选择不同类型问题的评价方法。在图 4 的右半部分，首先查看领域模糊匹配的候选列表中的 Top1 的分数是否大于阈值 F1，如果大于阈值则得到答案。否则查看闲聊模糊匹配的候选列表中的 top1 是否大于阈值 F1，如果大于阈值得到答案，则查看 Lucene 的 top1 是否大于阈值 L1，如果大于阈值得到答案，否则得到的答案为空。在整个流程中，增加了领域内问题的查全率和查准率，同时也解决了闲聊问题的回答，提高了问答系统的性能和用户体验。

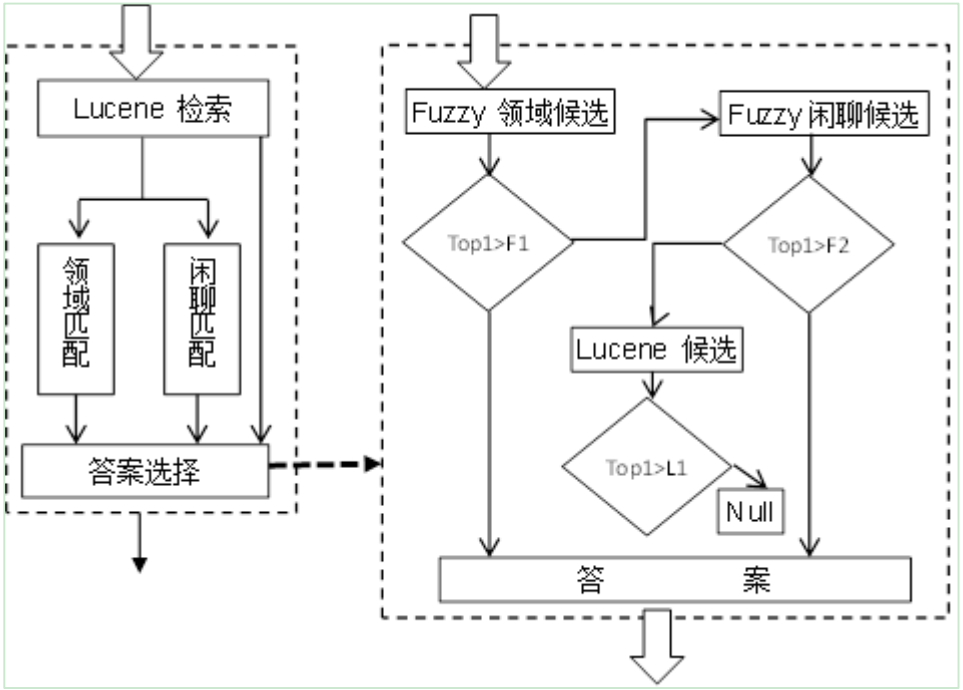


图 4 区分不同类型问题的检索方法的问答系统

4.3 基于遗传算法的评价因子优化方法

在以上的两个方法中，存在很多的评价因子，既上文中的 F1，F2 和 L1 阈值。如何有效的选择优化的评价因子，对问答系统的整体性能提升尤为重要。因此，

本发明提出利用遗传算法优化选择评价因子。具体流程如下：

1）**选择评价函数。**我们选择问答系统中的领域内的 F1 值和准确率之和作为评价函数，如 1 式。

$$\mathrm{Acc} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} + \frac{TR}{SUM} \quad (1)$$

其中，TP 为领域类被判定为正类，FP 闲聊类判定为领域类，FN 领域类判定为闲聊类，TR 为问答系统正确回答的问题个数，SUM 为问题总数。Acc 值越大，对应的系统性能越好。

2）算子选择

遗传算法的算子包括选择算子、交叉算子和变异算子。由于不同检索方法对应的打分大小尺度不一，对于模糊匹配的的阈值设定 0-1 之间变化，lucene 匹配的阈值设定在 0-3 之间变化。

1. 选择算子

为了保证算法的全局搜索能力，采用最优个体保存算子，即父代群体中的最优个体直接进入子代群体中，保证遗传过程中所得到的个体不会被交叉和变异操作所破坏。

2. 交叉算子

交叉算子是产生新个体的主要方法，决定了遗传算法的全局搜索能力，在遗传算法中起关键作用。由于参数不是很复杂，变换形式比较单一，所以选择简单有效的单点交换算子。

3. 变异算子

变异算子是产生新个体的辅助方法，它决定了遗传算法的局部搜素能力。变异算子和遗传算子相互配合，可以共同完成对搜索空间的全局搜索和局部搜索。为了快速的选择最优评价因子，在这里也引入简单的变异算子。

3） 算法流程

图 5 为利用遗传算法优化评价因子的流程示意图。首先对于给定的 F1，F2 和 L1 随机产生四个有理数，组成初始的父代序列。然后通过遗传算法的交叉和变异产生新的子代序列，将所有的子代和父代传入问答系统中，得出问答系统在开发集上的性能指数（Acc）。通过选择算子获取最优的四个组合序列，然后作为初始的父代进行迭代训练。详细流程如图 5 所示。

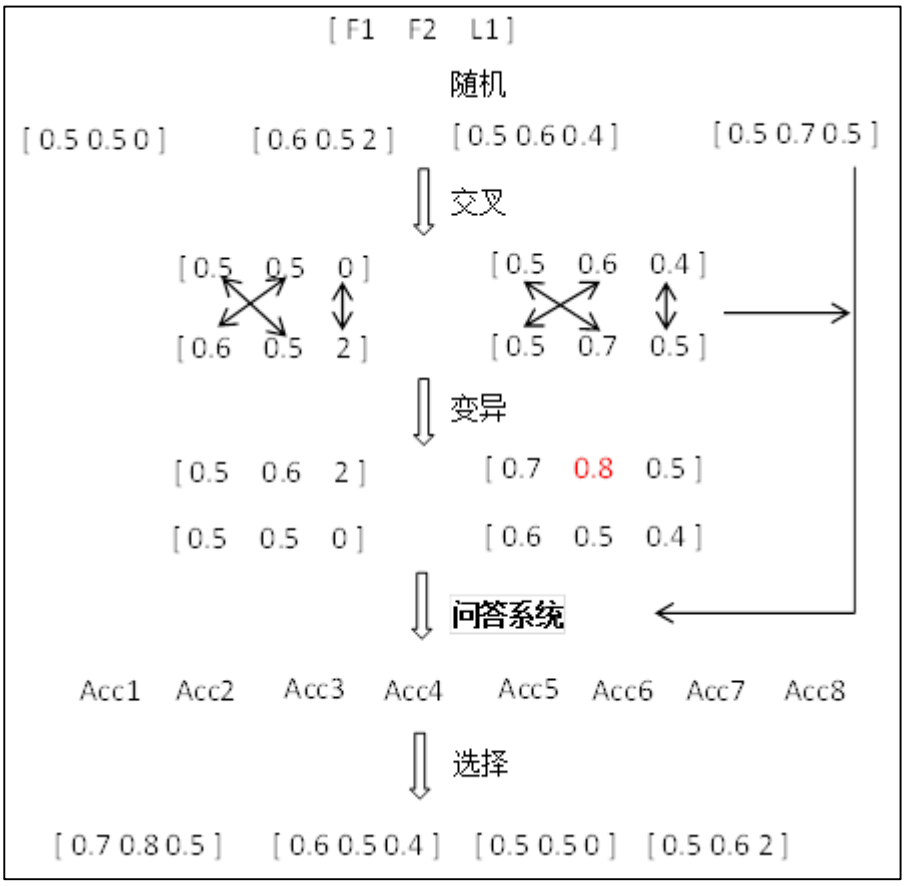
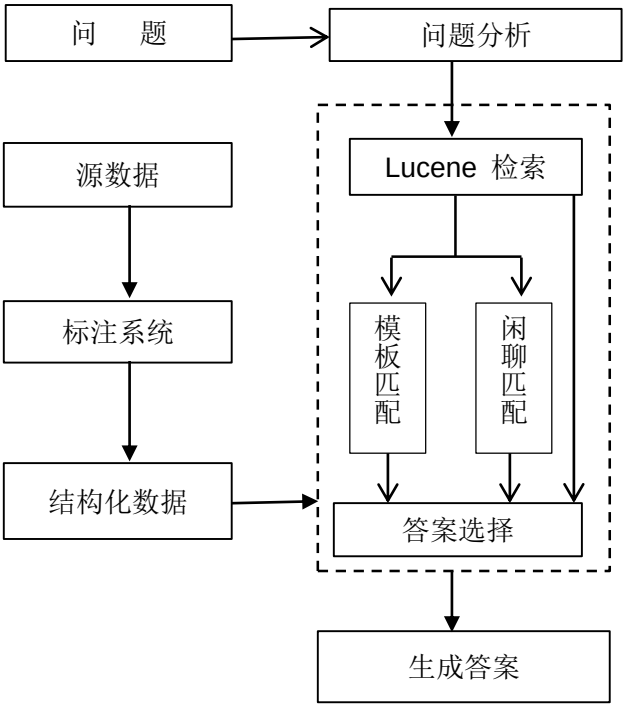
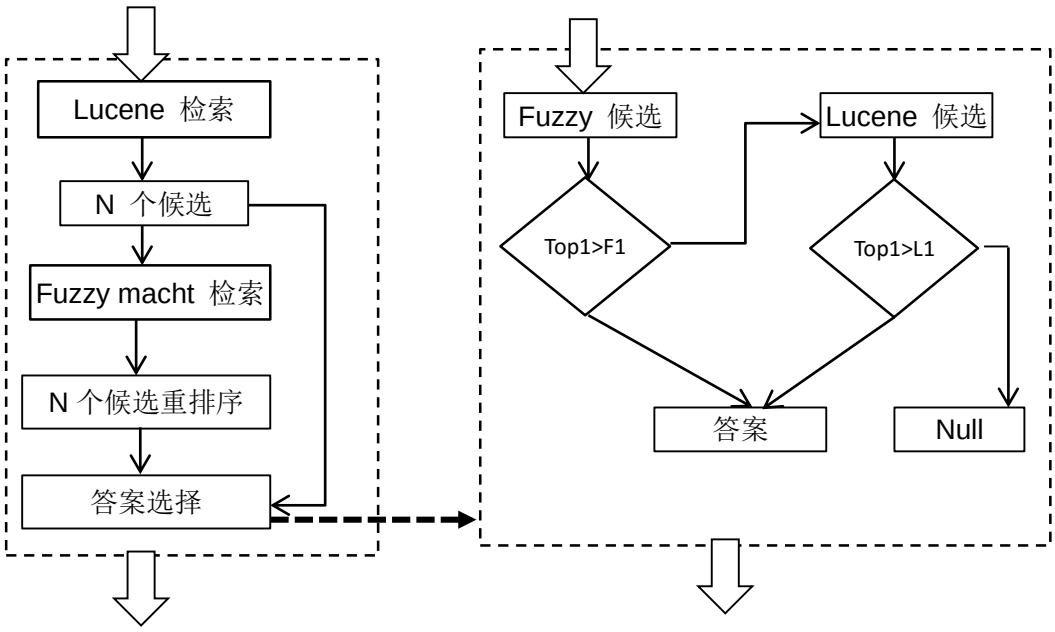


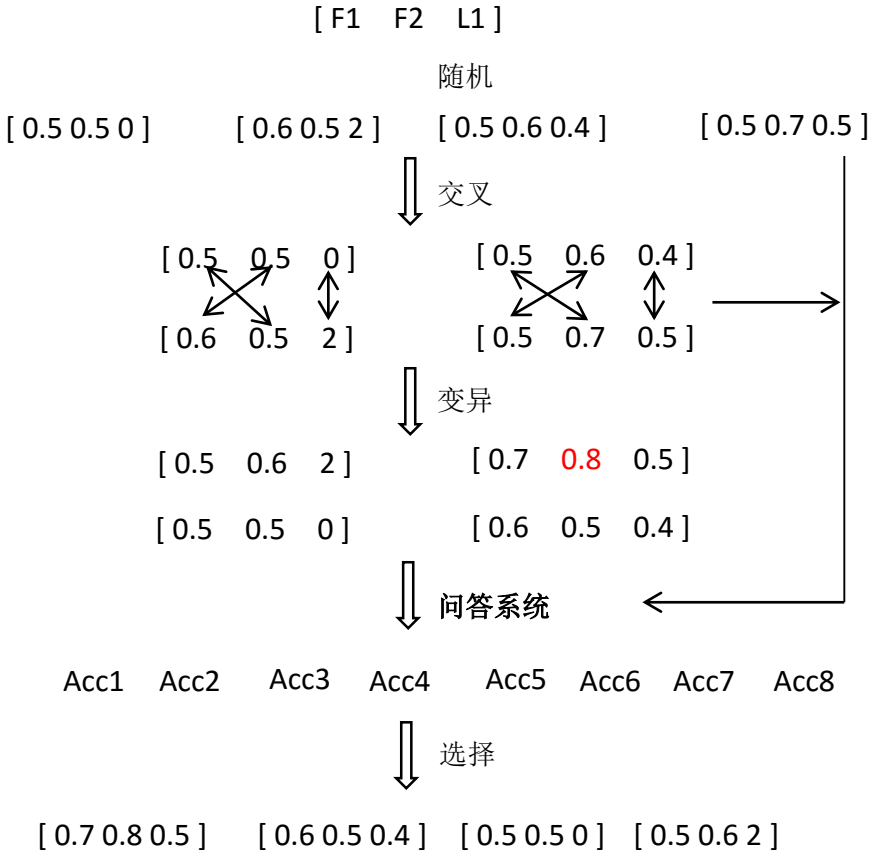
图 5 评价因子优化算法流程

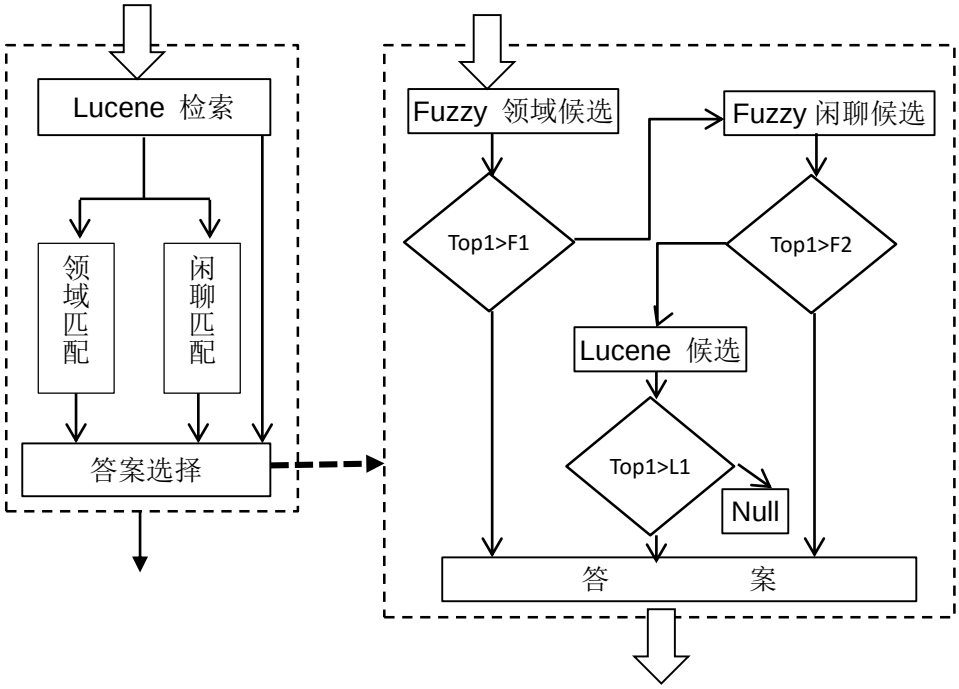
六 附图

6.1 问答系统


问 题
问题分析
源数据
Lucene 检索
标注系统
模板匹配
闲聊匹配
结构化数据
答案选择
生成答案


Lucene 检索
N 个候选
Fuzzy macht 检索
N 个候选重排序
答案选择
Fuzzy 候选
Lucene 候选
Top1>F1
Top1>L1
答案
Null


[F1 F2 L1]
随机
[0.5 0.5 0]
[0.6 0.5 2]
[0.5 0.6 0.4]
[0.5 0.7 0.5]
交叉
[0.5 0.5 0]
[0.6 0.5 2]
[0.5 0.6 0.4]
[0.5 0.7 0.5]
变异
[0.5 0.6 2]
[0.7 0.8 0.5]
[0.5 0.5 0]
[0.6 0.5 0.4]
问答系统
Acc1 Acc2 Acc3 Acc4 Acc5 Acc6 Acc7 Acc8
选择
[0.7 0.8 0.5]
[0.6 0.5 0.4]
[0.5 0.5 0]
[0.5 0.6 2]


Lucene 检索
领域匹配
闲聊匹配
答案选择
Fuzzy 领域候选
Fuzzy 闲聊候选
Top1>F1
Top1>F2
Lucene 候选
Top1>L1
Null
答 案