



中国科学院
CHINESE ACADEMY OF SCIENCES



中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

技术方案

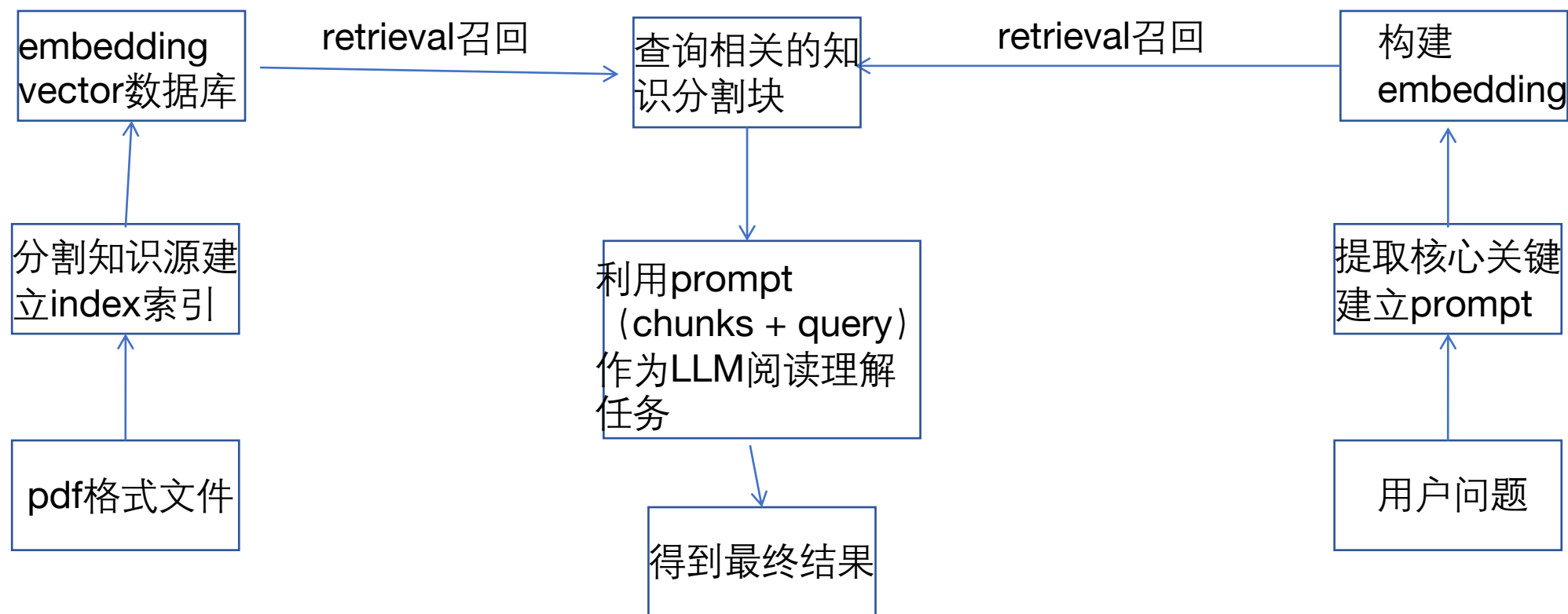
中国科学院大学邓绎如
2024年4月30日



技术方案

RAG (Retrieval Augmented Generation)

pipeline:



实验平台: NVIDIA TESLA V100 32G单卡

问答开源大模型: internlm/internlm2-chat-7b

<https://hf-mirror.com/internlm/internlm2-chat-7b/tree/main>

召回粗排模型: maidalun1020/bce-embedding-base_v1

https://huggingface.co/maidalun1020/bce-embedding-base_v1

检索精排模型: maidalun1020/bce-reranker-base_v1

https://huggingface.co/maidalun1020/bce-reranker-base_v1

建立本地检索向量库：

`service.feature_store.py --repo_dir fintech_data`

将pdf文件转化为txt文件，

再利用利langchain进行特征提取，

word embedding建立本地向量数据库，

利用BCE模型进行重排。

```
class FeatureStore:
    """Tokenize and extract features from the project's documents, for use in
    the reject pipeline and response pipeline."""

    def __init__(self,
                 embeddings: HuggingFaceEmbeddings,
                 reranker: BCERerank,
                 config_path: str = 'config.ini',
                 language: str = 'zh') -> None:
        """Init with model device type and config."""
        self.config_path = config_path
        self.reject_throttle = -1
        self.language = language
        with open(config_path, encoding='utf8') as f:
            config = pytoml.load(f)['feature_store']
            self.reject_throttle = config['reject_throttle']
```

建立本地检索向量库：

检索中，确定的回答的问题会利用基础模型提取关键词，将输入问题与向量库题在向量空间进行相似性比较，在知识库中检索 top 相似的 chunk，综合问题和检索到的 chunk 建立的本地数据库存储在workdir中。

```
def _split_text(self, text: str, separators: List[str]) -> List[str]:
    """Split incoming text and return chunks."""
    final_chunks = []
    # Get appropriate separator to use
    separator = separators[-1]
    new_separators = []
    for i, _s in enumerate(separators):
        _separator = _s if self._is_separator_regex else re.escape(_s)
        if _s == '':
            separator = _s
            break
        if re.search(_separator, text):
            separator = _s
            new_separators = separators[i + 1:]
            break

    _separator = separator if self._is_separator_regex else re.escape(
        separator)
    splits = _split_text_with_regex_from_end(text, _separator,
                                              self._keep_separator)
```


技术细节



中国科学院计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

```
python3 -m finfech.main --standalone
```

- 建立prompt对问题进行关键词获取:

```
self.TOPIC_TEMPLATE = '告诉我这句话的主题,
```

直接说主题不要解释: “{}”

- 对得到的关键词在知识库中检索召回:

```
sess.chunk, sess.knowledge, sess.references = self.retriever.query(sess.query,  
context_max_length=self.max_length)
```

```
if llm_config['enable_remote']:  
    self.context_max_length = llm_config['server']['remote_llm_max_text_length']  
if language == 'zh':  
    self.TOPIC_TEMPLATE = '告诉我这句话的主题, 直接说主题不要解释: "{}"  
    self.SCORING_QUESTION_TEMPLTE = "{}"\n请仔细阅读以上内容, 判断句子是否是个有主题的疑问句, 结果用 0~10  
    self.SCORING_RELAVANCE_TEMPLATE = '问题: "{}"\n材料: "{}"\n请仔细阅读以上内容, 判断问题和材料的关联度,  
    self.GENERATE_TEMPLATE = '请仔细阅读参考材料回答问题, 不需要解释答案, 问题: "{}" \n 材料: "{}"\n' # no  
else:  
    self.TOPIC_TEMPLATE = 'Tell me the theme of this sentence, just state the theme without explan  
    self.SCORING_QUESTION_TEMPLTE = "{}"\nPlease read the content above carefully and judge wheth  
    self.SCORING_RELAVANCE_TEMPLATE = 'Question: "{}", Background Information: "{}"\nPlease read  
    self.GENERATE_TEMPLATE = 'Background Information: "{}"\n Question: "{}"\n Please read the refe  
    self.max_length = self.context_max_length - 2 * len(self.GENERATE_TEMPLATE)  
f process(self, sess: Session):
```

```
prompt = self.TOPIC_TEMPLATE.format(sess.query)  
sess.topic = self.llm.generate_response(prompt)  
if len(sess.topic) < 2:  
    # topic too short, return  
    sess.code = ErrorCode.NO_TOPIC  
    return  
print(sess.query)  
print(sess.topic)  
#sess.topic = sess.query  
# retrieve from knowledge base  
sess.chunk, sess.knowledge, sess.references = self.retriever.query(sess.query, context_max_length=self.max_length)
```

- 得到相关chunks信息，结合此信息构建prompt让大模型进行阅读理解：

```
self.GENERATE_TEMPLATE = '请仔细阅读参考材料回答问题,不需要解释答案, 问题：“{}” \n 材料：“{}” \n'
```

```
response = self.llm.generate_response(prompt=prompt, history=sess.history, backend='remote')
```

```
prompt = self.GENERATE_TEMPLATE.format(sess.query, sess.knowledge)
response = self.llm.generate_response(prompt=prompt, history=sess.history, backend='remote')

sess.code = ErrorCode.SUCCESS
sess.response = response
return
```

'告诉我这句话的主题，直接说主题不要解释：“{query}”

'请仔细阅读参考材料回答问题,不需要解释答案, 问题：“{}” \n 材料：“{}” \n'

通过以上两个prompt得到从精排之后的knowledge和query得到答案

```
2024-04-30 19:57:21.015 | DEBUG | fintech.service.llm_server_hybrid:generate_response:523 - Q:告诉我这句话的主题，直接说主题不要解释：“请以json格式抽取2023年年度报告中，东方精选混合的债券投资明细，键名为债券名称，键值为比例，比例以百分数表示，保留2位有效数字 A:主题：2023年年度报告中，东方精选混合的债券投资明细。 backend local timecost 0.979518890380594
```

```
2024-04-30 19:53:35.337 | DEBUG | fintech.service.llm_server_hybrid:generate_response:523 - Q:收入
```

```
-  
股票投资收益—证券  
出借差价收入
```

```
-  
合计
```

```
-4,125,357,877.93
```

```
-2,382,633,948.82
```

```
7.4.7.11.2 股票投资收益—买卖股票差价收入
```

```
单位： A:2023年1月1日至2023年12月31日，活期存款利息收入为795,488.98元，2022年1月1日至2022年12月31日，活期存款利息收入为1,369,797.87元。2023年1月1日至2023年12月31日，活期存款利息收入对于2022年的增长比例为-42.8%。 backend local timecost 6.841302871704102
```

其中蓝色的为精排之后得到的材料（以便之后的阅读理解任务）



问题id	问题	答案	统计图
104	"请分析华泰柏瑞成长智选混合在2023年度四个季度报告期内基金总份额变化情况，并以季报时间为横轴、基金总份额为纵轴绘制折线统计图。"	"很抱歉，我无法直接绘制折线统计图，但我可以提供华泰柏瑞成长智选混合在2023年度四个季度报告期内基金总份额变化的数据，以供您进行分析和绘制。 根据您提供的数据，华泰柏瑞成长智选混合在2023年四个季度的基金总份额变化情况如下： - 第一季度：基金总份额为538,120,403.81元 - 第二季度：基金总份额为397,594,495.00元 - 第三季度：基金总份额为244,144,226.26元 - 第四季度：基金总份额为6,941,828.69元 您可以使用这些数据，以季报时间为横轴、基金总份额为纵轴，绘制折线统计图，以更直观地展示华泰柏瑞成长智选混合在2023年四个季度报告期内基金总份额的变化情况。 请注意，这只是一个简化的数据展示，实际的折线统计图可能需要更详细和准确的数据，并且可能需要使用专业的数据可视化工具来创建。"	暂无

由于没有引入图文模型，仅仅引入了文本模型，所以还没有生成图片的能力可以进一步增加图文模型对结果进行优化。

在很多测试结果中也可以给出了相应的图片生成的python代码，但由于模型限制，并不能很好的生成折线图饼状图等。

1.pdf提取，由于pdf提取成text格式文件，有些表格的数据横轴和纵轴不能很好的对应，导致了RAG模型不能很好的理解文本信息，由此可能带来一些结果上的生成偏差，这个可以通过优化pdf提取模块来解决。

2.参数调整问题，由于比赛时间较为紧张，排序的top k以及大模型的记忆聊天remember参数，以及大模型的温度以及top_p可以进一步进行调整。并且由于金融领域是很严谨的领域可以对于金融问答问题对于特定的大模型进行微调，这样可以使生成的答案更加严肃和一语中的。

3.有关图文问题可以通过外接画图模块来解决折线图饼状图的问题。设置prompt画图的参数让大模型生成相应的参数自动生成图片即可。