基于知识图谱的问答系统研究报告

目录

基于知识图谱的问答系统研究报告 1

目录 2

一、 概念及发展历程 3

1.1 问答系统简介 3

1.2 问答系统的发展历程 4

二、 业界现状 5

2.1 问答系统分类 5

2.2 问答系统的核心技术 6

2.3 基于知识图谱问答的优势 7

三、 基于知识图谱的问答系统 7

3.1 知识图谱简介 7

3.2 知识图谱问答的核心技术 9

3.2.1 问题理解 10

（1）查询重写 10

（2）命名实体识别 11

3.2.2 构建知识图谱 11

3.2.3 搜索和排序 13

（1）索引和匹配 13

（2）候选答案重排序 14

3.2.4 模型评估 14

（1）分类指标评估 15

（2）排序指标评估 15

3.3 基于知识图谱构建问答系统原型 16

3.3.1 知识图谱原型构建 17

3.3.2 用户问题的理解 18

（1）中文分词 18

（2）命名实体识别 19

3.3.3 问答系统搜索排序 19

（1）知识图谱的反向索引 19

（2）知识图谱的检索排序 20

3.3.4 问答系统原型 21

四、 后续工作思考 22

4.1 智能客服 23

4.2 医疗助手 24

随着互联网的普及，网络上的信息越来越丰富，人们能通过信息检索得到自己想要信息。但信息检索存在很多不足，一是**返回结果太多**，用户很难快速定位到所需的信息；二是**结果缺乏推理逻辑和语义关系**，信息检索算法只关注语法形式和检索词语，而非语义逻辑。这些因素使得人们期望能用智能问答等新方式满足信息需求。

本报告从问答系统的概念和发展历程开始，介绍问答系统的分类和业界使用问答系统的现状。重点研究了问答系统的核心技术，包括问题理解、搜索排序等技术细节，并结合知识图谱逻辑推理和知识表示技术，构建了基于知识图谱的问答系统。基于部分银联银行卡、商户营销数据搭建业务知识图谱，构建银联业务知识的问答系统原型。针对今后智能客服、医疗助手、问答机器人等应用，提出了相关工作的开展策略和应用落地的建议。

# 概念及发展历程

## 问答系统简介

**问答系统**（Question Answering system, QA system）是用来回答人提出的自然语言问题的系统，通过输入完整和口语化的问句，经过自然语言问题解析、问题分类、命名实体识别、知识表示和知识推理、答案检索与重排序、候选答案生成等复杂的内部处理，直接返回明确的问题答案或者精确的文档结果。问答系统不仅仅需要计算机理解提问者的问题、问题的上下文信息、问题的类型等，也需要结合自然语言解析、知识推理等模块优化问答结果，问答系统相比传统的信息检索更为复杂且难以实现。现有的问答系统从答案生成方式来说可以分为封闭领域问答系统和开放领域问答系统。封闭领域的问答系统专注于回答特定领域的问题，通常根据答案语料结构化信息，或问题上下文来检索问题答案；开放领域问答系统通过构建序列化的模型，利用自然语言生成等方法实现问答系统。

## 问答系统的发展历程

问答系统是人工智能技术的核心研究领域，从早期基于专家系统的问答到基于计算语言学的问答，伴随着互联网、大数据、深度学习技术的应用，信息检索技术的成熟，问答系统也进入基于开放知识领域的问答系统时期。

20 世纪60年代随着人工智能的发展，科研人员试图建立一种能回答人们提问的智能系统。这个时期主要是限定领域、处理结构数据的问答系统，主要是**基于AI系统和专家系统**，代表系统有BASEBALL和LUNAR。在20世纪70年代到80年代，**基于计算语言学的问答系统**兴起，大量研究集中在如何利用计算语言学技术去降低构建问答系统的成本和难度，代表系统是Unix Consultant。进入20世纪90年代，**基于开放领域、知识图谱的问答系统**获得成功。由于互联网的飞速发展，产生了大量的电子文档，这为问答系统进入开放领域、基于文本的时期提供了客观条件。1999年TREC（text retrieval conference）设立了QA track ，极大地推动了问答系统的发展。随着2005 年末以来大量的社区问答（community based question answering）数据（如Yahoo！Answer）出现在网络上，有了大量的问答数据训练集，问答系统进入了开放领域、基于问题答案对时期。

# 业界现状



## 问答系统分类

如图 1 问答系统不同的分类图所示，问答系统可以有多种构建或分类方法。从信息系统建设方法来说，可以分为基于问答对、基于规则模式和基于机器学习的问答系统；从答案来源来说，可以分为基于专家知识、基于知识库和基于自由文本的问答系统；从用户问题领域来说，可以分为特定领域的问答系统和开放领域的问答系统；从答案生成方式来说，可以分为基于检索的问答系统和基于生成式的问答系统。

图 1 问答系统不同的分类

如图 2 不同答案来源的问答系统优势和缺点对比图。就答案来源来说，基于专家知识的问答，答案更为准确但覆盖面不够，基于自由文档的问答可以回答各个领域知识，但通常缺乏权威，而基于知识图谱的问答系统能很好结合答案权威性和多样性，是自然语言处理的明日之星。基于问答对和基于规则模式的问答系统通常易于实现，但难以拓展，回答问题较为有限，基于机器学习的问答通常能回答范围更广的问题，但答案的准确性严重依赖机器学习数据、算法和模型；就答案生成方式来说，基于搜索引擎方式的问答找到答案准确性更好，但通常难以实现问题推理，并不易组织答案，基于生成式的问答系统能更好生成人类的语言，但任处于初步探索阶段。



图 2 不同答案来源的问答系统优势和缺点

## 问答系统的核心技术

问答系统的核心技术主要分为三个部分：（1）**问题理解**，如何处理用户的提问，包括如何对问题分类、提取问题的关键词、对用户的提问进行分析理解；（2）**候选答案的检索和排序**，缩小答案的范围，提高答案的抽取的效率和精度，答案检索可以分为对文档的检索、对段落的检索和对知识图谱的检索。（3）**候选答案抽取和答案的生成**。在通过检索和排序系统获取候选答案集合后，需要生成问题答案。基于表层特征抽取、模板抽取、关系抽取及知识表示和推理可以将候选答案集合转化为问题答案。

尽管当前有很多较为成熟的问答系统，如苹果的Siri、微软小娜、IBM的Waston等，但问答系统也存在很大的提升空间。如微软小娜等问答系统通常是闲聊机器人，使用模板匹配、自然语言生成的方式进行问答对话，提供的答案并**没有语义和推理逻辑**，此外，聊天机器人并不能解决专业性问题，无法为客服、金融、电商等垂直领域提供专业聊天支持，也**无法回答领域知识**。

## 基于知识图谱问答的优势

由于现有问答系统无法理解问题语义关系，不能辅助用户进行推理逻辑，基于知识图谱的问答系统越来越流行，知识图谱提供理解问题逻辑的算法，利用知识图谱知识表示和逻辑推理技术，能更好从知识图谱中计算出最佳的结果作为问题答案。

知识图谱可以解决基于搜索引擎问答的很多不足。首先基于知识图谱的**知识表示能高效表征问题逻辑和候选答案**；其次知识图谱能**解决问答系统中的语义鸿沟问题**，基于关系抽取、知识推理等知识图谱相关算法，知识图谱很容易将用户提问映射到知识库中的查询中，并建立从问题到答案之间的语义匹配关系。

# 基于知识图谱的问答系统



## 知识图谱简介

**知识图谱**是为描述客观世界中的概念、实体、事件等之间关系的一种知识表示形式。知识图谱用有向图进行描述，图中的“节点”描述现实中存在的概念或对象，图中“边”描述这些节点之间的关系。除了“节点”和“边”，知识图谱中也存在“属性（值）”，用来描述“节点”存在的属性。如图 3 知识图谱的语义网络示意图展示了一个关于“中国银联”的知识图谱，知识图谱通常使用三元组进行物理存储，包括**实体关系型三元组**（中国银联，位于，上海）和**实体属性型三元组**（银联，成立于，2002年），实体关系型三元组描述实体之间的关系，实体属性型三元组描述实体具有的某些特征。除了实体、关系、属性外，知识图谱也包括文本、语义概念等信息。

**实体** 实体指独立存在且具有某种区别性的事物。如一个人、一种动物、一个国家、一种植物。具体的事物就是实体所代表的内容，实体是知识图谱中的基本元素，不同的实体间有不同的关系。

**语义概念** 语义概念具有同种特性的实体构成的集合，如人类、动物、国家、植物等。概念主要指集合、类别、对象类型、事物的种类，例如人物、地理等。

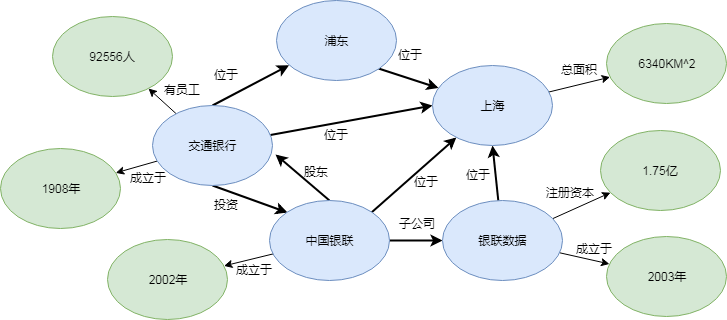


图 3 知识图谱的语义网络示意图

**内容** 内容通常是实体和语义类的名字、描述、解释等，展示形式一般有文本、图像、音视频等。

**属性值** 属性指对象具有的属性和相关的值，不同的属性类型对应于不同类型属性的边。

**关系** 在知识图谱上，关系的表现形式是一个将节点（实体、语义类、属性值）映射到布尔值的函数。

知识图谱数学上定义为，其中,表示知识图谱中实体的集合，表示关系的集合，，表示实体和关系组成的向量空间。

## 知识图谱问答的核心技术

如图 4 基于知识图谱表示的自动问答系统所示，基于知识图谱的问答系统本质来说，是一种基于语义匹配、知识表示等特征构建的信息检索系统。对于用户提出的问题，结合用户提问的上下文、对问题进行查询重写，词法分析和命名实体识别等处理后，在知识库和相关的文档中进行检索，基于历史问答对构建的排序模型对用户提问进行处理，获取最佳的问题答案，并将问题答案返回给用户。

对于离线构建训练搜索排序模型来说，将历史的问答对处理构建训练数据集，首先建立倒排索引并进行关键词检索和语义检索，筛选并减少不相关的检索结果；对于通过文档表示和文档检索获得的候选问答对进行知识匹配，并基于BM25、TF-IDF和知识图谱的语义网络表示构建粗匹配，获取知识匹配的候选结果；最后，对粗匹配获得的候选问答对，构建学习排序模型、梯度下降树、神经网络等排序算法等，将候选问答对重排序，获取最终的问题答案。

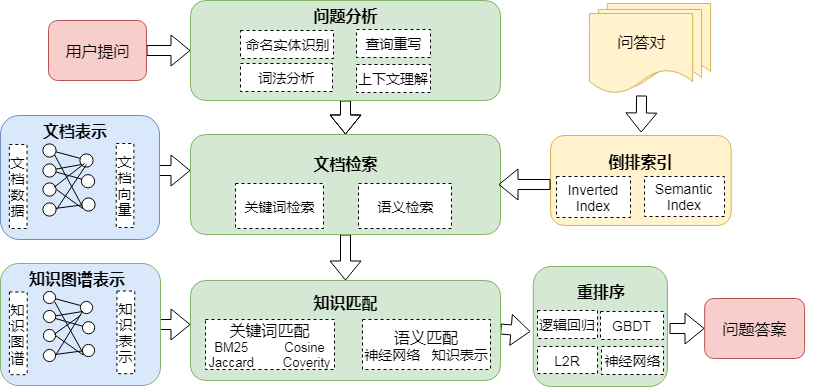


图 4 基于知识图谱表示的自动问答系统

### 问题理解

对问题的理解重点关注问题的分类，如何回答不同提问的类型；如何对一些用户提问进行查询重写，将用户的提问转化为计算机可处理的问题；对问题中的实体进行识别，发现问题中的关键实体。

#### （1）查询重写

查询重写（Query Rewriting）通过对用户的提问进行改写，**提高查询过程中的召回率和命中率**。查询重写的首要目标是提高答案的覆盖范围，增加召回率通常是在返回更多相关答案和更精确答案的一个平衡，增加召回率可以基于拓展查询词和拓展查询范围。拓展查询词可以将查询词进行同义词拓展，基于word2vec等距离度量方法，获取更多的查询词，提高查询词的数量。另一种提高查询召回率的方法是拓展查询的范围，例如将一些查询的限定词去除，拓展查询的范围，如对于查询“黄色可爱的小猫”，可以放松查询范围，只检索“黄色的小猫”或者“可爱的小猫”这类相关的检索结果。增加查询的命中率，即增加查询的精确度可以从问题的切分入手，提高一个查询问题的词语切分效果和切分精度，有些情况下将一个问题中每个字进行切割作为查询的单位。

#### （2）命名实体识别

命名实体一般指的是文本中具有特定意义或者指代性强的实体，通常包括人名、地名、组织机构名、日期时间、专有名词等。命名实体识别（Named Entity Recognition）系统就是从非结构化的输入文本中抽取出上述实体，并可以按照业务需求识别出更多类别的实体，比如产品名称、型号、价格等。学术上涉及的命名实体一般包括三大类（实体类，时间类，数字类）和七小类（人名、地名、组织机构名、时间、日期、货币、百分比）。实际应用系统中，命名实体识别模型通常只要识别出人名、地名、组织机构名、日期时间即可。

### 构建知识图谱

如图 5构建知识图谱的核心技术所示，构建基于知识图谱的问答系统工程的核心流程包括：知识抽取、知识融合、知识表示、知识推理、知识检索和分析等核心环节，知识图谱系统是问答系统的答案来源。首先确定垂直领域的知识模型，然后根据数据来源选择不同的知识获取手段导入知识，接下来需要综合利用知识表示、知识融合、知识推理等技术对所构建的知识图谱进行质量提升，最后根据场景需求设计不同的知识访问与呈现方法。

**知识抽取**按任务可以分为概念抽取、实体识别、关系抽取、事件抽取和规则抽取等。传统专家系统时代的知识主要依靠专家手工录入，难以扩大规模。现代知识图谱的构建通常大多依靠已有的结构化数据资源进行转化形成基础数据集，再依靠自动化知识抽取和知识图谱补全技术从多种数据来源进一步扩展知识图谱，并通过人工众包来进一步提升知识图谱的质量。

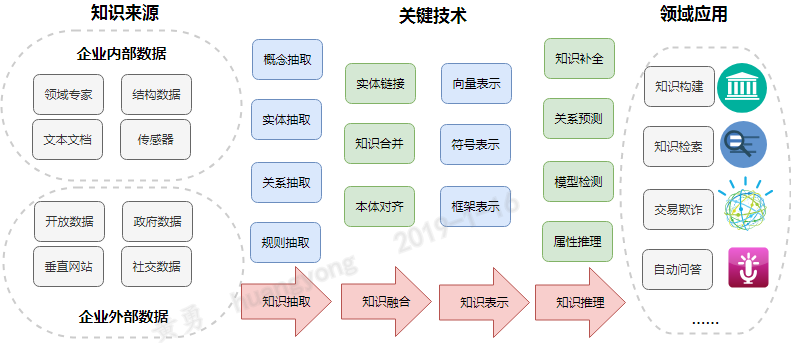


图 5构建知识图谱的核心技术

在构建知识图谱时，可以从第三方知识库产品或已有结构化数据获取多源知识，则需要**知识融合**。例如关联开放数据项目（Linked Open Data）会定期发布其经过积累和整理的语义知识数据，其中既包括通用知识库 DBpedia和 YAGO，也包括面向特定领域的知识库产品，如 MusicBrainz和DrugBank等。当多个知识图谱进行融合，或者将外部关系数据库合并到本体知识库时需要处理两个层面的问题：（1）通过模式层的融合，将新得到的本体融入已有的本体库中，以及新旧本体的融合；（2）数据层的融合，包括实体的指称、属性、关系以及所属类别等。

**知识表示**将知识图谱中的实体关系表示为低维度向量。基于三元组的知识表示形式比较直观，但是其在计算效率、数据稀疏性等方面面临着诸多问题，从而研究知识的表示学习和推理算法是问答系统的核心。知识图谱的表示和推理技术，结合一阶逻辑推理、知识表示、随机游走等不同算法，通过构建知识图谱自动推理技术，能有效帮助解决机器进行知识处理、语义匹配等相关问题。

**知识推理**（知识图谱推理）是人工智能系统的重要组成部分，包括归纳推理、演绎推理、确定性推理和不确定性推理等不同种类。智能系统的知识推理过程是通过推理机来完成的，推理机是智能系统中用来实现推理的程序。推理机的基本任务就是在一定控制策略指导下，搜索知识库中可用的知识，与数据库匹配，产生或论证新的事实，搜索和匹配是推理机的两大基本任务。

### 搜索和排序

对用户的一个提问，知识库中存在成千上万条候选的答案，为了能准确找到最佳答案，通常采用匹配-重排序的方法对问答系统进行处理。搜索排序的目标是将用户的提问特征，知识图谱的表示特征进行特征工程，通过检索、匹配和重排序，找到问题的最佳答案。

#### （1）索引和匹配

索引和匹配是对现有的训练数据进行筛选，获得TOP-K候选问答结果的过程。通常索引和匹配通过从上万条答案中筛选出数百条候选答案，并将这些候选答案通过重排序计算和问题的相关性。索引是对知识图谱、文本等数据建立快速查询的过程，通常对文档或者知识库进行分词后，对每个文档-词构建倒排索引，为对问题的快速检索进行。匹配使用相对简单的常用检索模型对用户问题从索引中快速检索出 Top-k 候选结果集。常用检索模型主要有向量空间模型 （Vector Space Model）、布尔模型 （Boolean Model）、概率检索模型如BM25 等，通常 Top-k 的候选集选取还结合离线计算质量分高的文档以排除掉文本相关但质量分太低的文档。

#### （2）候选答案重排序

重排序使用计算复杂的机器学习排序模型对 Top-k 候选结果集进行精确的重排序，获取问题对应的最佳答案。

基于机器学习的排序系统一般分为离线学习系统和在线预测排序系统。离线系统的设计需要靠特征的选择、训练集的标注、学习排序等算法的选定、确定损失函数、以最小化损失函数为目标进行优化，以获取排序模型的相关参数。在线预测排序系统将待预测结果输入到机器学习得到的排序模型，即可得到结果的相关性得分，进而依据相关性得分得到搜素结果的最终排序。给定一些历史的问答对，将其训练为机器学习排序模型，同构建在线预测模型，对每个用户的提问，实时返回用户期望的答案。

### 模型评估

问答系统的模型评估和采用的搜索排序模型相关，通常机器学习排序模型中采用三类排序算法：单文档排序，文档对方法，文档列表方法。单文档排序只考虑每个文档的打分和排序，而文档对方法不仅考虑单文档的打分情况，还考虑文档对之间的偏序关系，文档列表同时考虑不同的问题的文档排序和同一个问题中文档的偏序关系。这些不同的排序方法有不同的模型评估策略。

#### （1）分类指标评估

问答系统的评估采用分类相关的模型评价。通常可以采用精准率（Precision）、召回率（Recall）、F1、AUC等相关的模型评价指标。

其中Precision衡量检索到的文档有多少相关性，也被称为查准率，Recall衡量检索是否能查全，也被叫做查全率。F1是综合衡量一个问答系统是否能同时兼顾查全率和查准率的指标。

#### （2）排序指标评估

除了基于分类指标的模型评估，问答系统也可以基于排序指标进行模型评估。常见的排序评估指标包括平均准确率（MAP）和平均倒排秩序（MRR），可以综合衡量单文档排序、文档对排序方法和文档列表排序方法等不同的排序模型，相比分类指标更适合问答系统、检索系统等的模型评估。

其中，M表示有M个提问，每个提问中Q个问答对作为模型的训练集。MAP也是衡量模型排序准确性的指标，MRR是衡量模型排序查全率的指标。

## 基于知识图谱构建问答系统原型

基于前期调研和对知识图谱的研究工作，研究院开展了基于业务知识图谱的问答系统原型搭建，将银行卡数据、商户数据、营销活动数据相结合，构建银联知识图谱原型。

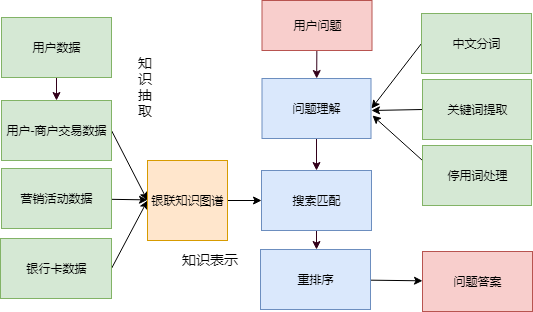


图 6 问答系统原型架构图

如图 6 问答系统原型架构图所示，整个原型分为三部分：（1）银联知识图谱原型的构建，基于用户数据，抽取用户-商户的交易数据，并对营销活动数据，银行卡数据进行结构化，将这些数据整合成银联知识图谱；（2）用户问题理解，对于用户提出的问题，进行中文分词、停用词处理和关键词提取等算法；（3）基于算法处理后的问题和知识图谱答案进行知识表示学习，通过搜索匹配获取候选答案集合，最后通过对候选答案重排序，计算得到最终的问题答案。

### 知识图谱原型构建

知识图谱构建主要分为三个步骤：（1）对银行卡数据的结构化，计算获得银联银行卡权益的三元组；（2）对商户数据的结构化，计算获取银联商户的三元组；（3）对营销活动数据进行结构化，计算获取银联营销活动的三元组。将上述的银行卡三元组、商户三元组、营销活动三元组进行知识融合，构建知识图谱原型。

如图 7 知识图谱的实体、关系和属性所示，对于银行卡数据来说，计算抽取银行卡的实体关系型三元组和实体属性型三元组。例如对于“广发唯品会联名卡金卡”实体和相关的文本信息分别抽取相关的实体关系型三元组和实体属性型三元组。通过基于模板的关系抽取算法，抽取实体关系型三元组：（广发唯品会联名卡金卡，卡等级，金卡）、（广发唯品会联名卡金卡， 发卡组织，银联）等，并抽取实体属性型三元组如：（广发唯品会联名卡金卡，年费，80元/卡/年）、（广发唯品会联名卡金卡，取现比例，50%）等。对于商户数据来说，抽取商户相关的实体属性三元组和实体关系三元组。使用模板关系抽取算法，抽取实体关系型三元组如：（俊文宝石店，位于，香港）、（位元堂，位于，中国）等，抽取实体属性型三元组如：（位元堂，电话，2727-8911）等信息。对于银联的营销活动数据来说，使用基于模板的关系抽取算法，抽取实体属性型三元组如（交通银行，其他活动，借记卡绑定云闪付送5元-500元红包）等。

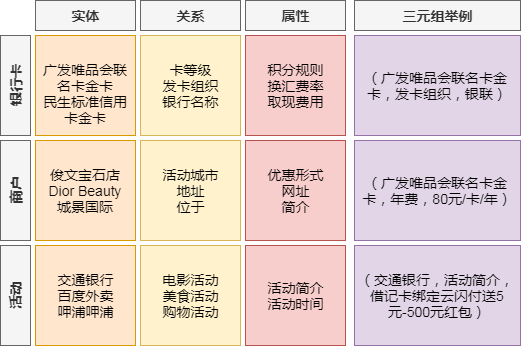


图 7 知识图谱的实体、关系和属性举例

在银联卡数据、营销活动数据、商户数据的三元组抽取，构建知识图谱的过程中，如何确定实体和实体之间的关系十分重要。我们队银联卡数据确定了关系类型包括：发卡组织、银行名称、年费、取现比例、积分规则等关系信息；对于营销活动数据，我们确定的关系类型包括：美食活动、电影活动、商超活动、活动规则、活动时间等关系类型。对于商户数据，我们确定的关系类型包括：商户地址、商户电话、商户类型、商户网站、商户活动等信息。通过上述关系结合构建了知识图谱原型。

### 用户问题的理解

#### （1）中文分词

对用户问题的理解，我们使用**基于统计的分词方法，**进行中文分词处理。给定一个用户的提问，将这个用户提问的汉字序列切分成一个一个单独的词。当前有三种基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。考虑到银联知识图谱建设中，需要使用很多的专用名词，例如：“广发银行”、“位元堂”等。

我们采用词语二次切分的统计分词模型进行中文分词。首先，我们构建了银联公司相关的语义词典，并提高这些词典在中文分词中的权重，基于N-gram的N元语法模型可以提高这些专有名词被分类正确的准确率。除此之外，考虑到不同类型的分词模式（精准分词、全模式和细分模式），我们选择细分模型，对其中一些长词进行二次切分，提高相关词语在问答系统中的查全率和召回率。

#### （2）命名实体识别

命名实体识别的主要思想在于根据现有实例的特征**总结识别和分类规则**。相关的方法可以被分为有监督（supervised）、半监督（semi-supervised）和无监督（unsupervised）三类。有监督学习包括隐形马科夫模型（HMM）、决策树、最大熵模型（ME）、支持向量机（SVM）和条件随机场（CRF）。

我们采用**基于规则和词典的方法**进行专有名词识别。在问题理解中使用命名实体识别，提高问题中关键词的权重，可以提高模型的查准率。考虑到训练语聊的不足，我们使用的这些方法主要是读取注释语料库，记忆实例并进行非监督学习，根据这些例子的特征生成针对某一种实例的识别规则，从而识别其中的专有名词。

### 问答系统搜索排序

#### （1）知识图谱的反向索引

我们将**知识图谱中的三元组作为问答系统中的答案，从而需要对知识图谱建立反向索引（Inverted Index）**。对于知识图谱中的实体关系三元组和实体属性三元组，和问题理解一样，都需要使用基于统计的分词算法，进行中文分词，并采用细粒度的中文分词，提高信息检索的精准度和查全率。除此之外，我们还需要对知识图谱建立索引系统，提高知识图谱的信息查找速度。

我们构建反向索引首先通过对三元组进行中文分词。将分词获取的“三元组-词语”索引进行反向索引，获取“词语-三元组”的索引模型，从而为后续建立检索和排序模型提供基础。倒排文档索引的优势不仅在于关键词个数少带来的检索效率提高，还在于其特别易于同信息检索技术结合。在实际应用中，查询中所包含的关键词往往是很少的，完全不包含查询中的所有关键词的文档，一般来说是不会被列入结果集的。因此，以关键词为主键进行索引，只需要用查询中包括的关键词，进行简单的查询就能够找出相关的文档。

#### （2）知识图谱的检索排序

常见的信息检索模型包括三类：向量空间模型、概率模型、推理网络模型。考虑到银联知识图谱的数据量较少，且很容易进行检索排序，我们的信息检索模型使用向量空间模型，基于TF-IDF和BM25算法对知识图谱进行检索和排序，其中TF-IDF是词频和倒排文档频率的乘机，BM25是对TF-IDF的改进，提高短文本在检索模型中的权重。

其中，TF描述第i篇文档中的第j个词的词频，IDF是词的倒排文档频率，描述词在不同文档中出现的频繁程度。BM25是对TFIDF在短文本中的改进，针对问题查询Q，是每个文档和问题Q的BM25得分，其中Q是包含n个关键词的查询问题，和是弹性系数，D是知识图谱的倒排文档集合，。

考虑到目前银联知识图谱原型中数据较少，在只采用BM25打分即可在简单的问答系统中获取较好的打分结果，后续随着预料数据的增大，文档和查询类型的增加，可以考虑使用深度神经网络和其他重排序技术对文档进行二次排序，提高模型的查全和查准率。

### 问答系统原型

基于我们对银联知识图谱的构建、用户问题理解和问答系统的搜索排序模型建设，我们构建了基于Lucene的问答检索原型系统，系统可以对用户的提问进行细粒度的分词和关键词识别，并对关键词提升权重，通过对知识图谱中三元组的建设，包括构建实体关系型三元组和实体属性型三元组，我们获取了大量关于银行卡、商户和营销活动的知识图谱，并对这些知识图谱建立索引和排序模型。考虑到银联的知识图谱存在大量的专有商户和银行名称，我们建立了银联的专有分词词典，并结合无监督模型进行名词识别。对于知识图谱三元组的问答结果，我们的排序模型基于BM25算法，通过问题中关键词的权重提高，答案的倒排文档停用词处理，提高排序效果。

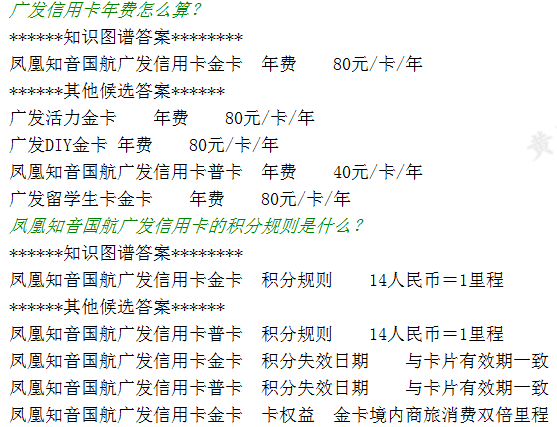


图 8基于知识图谱问答结果和相关推荐的图例

如图 8基于知识图谱问答结果和相关推荐的图例所示，基于知识图谱的问答系统可以自动回答用户的问题提问，并推荐一些相关的候选答案，从而提高用户查询信息的速度，满足用户需求。我们测试了数百条问答对，TOP-1的正确答案超过80%，TOP-5的正确答案比率超过85%，后续可以考虑结合深度学习、机器学习模型等不同的模型进行重排序，进一步提高模型的查全和查准率。

# 后续工作思考



## 智能客服

针对银联当前在智能客服领域的业务需求，如客服工作量大，找到持卡用户需要的服务信息慢，云闪付用户面对多个银行、多个商家优惠信息时，难以找到有效信息。基于银联小U问答、业务运营中心的商户服务场景下，我们可以构建结合营销活动、银联用户、商户和银行的知识图谱，并将这些知识融合，建立同时为商户、用户和内部客服坐席人员服务的智能客服系统。

一是**提高知识图谱基础技术的研究水平，增强技术储备能力**。当前银联的营销多种多样，商户和银行活动类别丰富，为了能处理银联的结构化数据、半结构化数据、非结构化数据、多媒体数据，需要提高银联在知识图谱基础技术上的研发能力。基于实体和关系抽取构建知识图谱，基于知识表示和知识推理挖掘业务逻辑，基于知识融合和知识问答来满足各类业务需求，并结合深度学习、搜索排序技术，增强银联在知识图谱领域的基础研发能力。

二是**对银联内部的客服、业务运营中心等业务知识进行梳理，构建基于知识图谱的问答系统**，服务云闪付用户和内部客服坐席人员。除了在问答系统中应用知识图谱外，将知识图谱用于客服工单系统中，为客服坐席人员提供信息查询服务，切切实实为客服提供检索服务。基于银联客服的问答数据、工单数据和银联业务数据，搭建智能客服系统， 帮助云闪付用户解决APP内问题，减轻客服工作量。

三是**结合研究院的国家工程实验室讲解等场景，构建基于知识图谱的问答机器人**。利用研究院当前的虚拟现实、手机POS、人工智能、风险防控和区块链等展厅场景，搭建基于知识图谱的智能机器人，让机器人能代替讲解员，为研究院的参观者进行业务介绍和疑问解答。

## 医疗助手

通过前期柴总和大专家等医疗领域公司接触，了解到现有很多药物、疾病、医院的数据，可以进行结构化知识抽取，构建医疗领域知识图谱，并搭建医药和疾病的领域问答系统，辅助医生进行疾病治疗，为患者普及医疗科普知识。

一是**构建医疗知识图谱，辅助医生进行疾病治疗，为患者普及医疗科普知识**。通过和外部医疗企业合作，构建医疗领域基于疾病、药物、医生和医院的知识图谱，对这些知识建立搜索排序模型，将各种疾病和相关病理、治疗方法综合起来，搭建疾病推理系统，为医生和患者提供相关的知识图谱搜索引擎和问答系统，并提供信息查询和疾病治疗服务。

二是**深化跨领域合作，深入探索基于知识图谱的医疗问答系统**。基于知识图谱的医疗问答系统是传统病例搜索的高级形式，能更好满足医生查找资料、医药的信息需求，研究院要深化和医疗领域的跨领域合作，并结合医疗领域的专家知识，应用知识图谱的表示和推理技术，构建医疗辅助系统、问答系统，同时深化和各类医疗机构、国家卫生健康机构的合作。除此之外，将银联用户的信息和医疗健康档案信息结合，构建标准化、信息化、便利化的智能系统，辅助就诊治疗和银联支付，则能形成银联在各类垂直行业的业界影响力。