# 基于知识图谱的图书推荐系统

项目宣言：用科技重建精神家园

组长：章磊

组员：肖龙、张蓝天、郑好

## 调研部分

**背景**

图书的选择对于每个人都是一个问题，当遇到不熟悉的知识时，人们一般会通过网络搜索相关的问题。而互联网上的资源信息不断增长，仅仅包含了传统文档形式的网页以及网页的链接，很难满足快速获取信息的需求。

面对互联网上冗杂的信息，学者们希望能够合理组织互联网的资源，使互联网成为信息交换的媒介，把知识转换为计算机易于处理和理解的语义，以一种更加智能化的方式获取所需信息，清晰且有逻辑地呈现各种概念之间的关系，由此，知识图谱[1]应运而生。知识图谱是一种非常有效的知识组织方式, 在许多领域得到广泛应用。

近年来深度学习在自然语言处理上的应用使得知识图谱迅猛发展，深度学习方法广泛应用于问答系统[2,3]等任务。知识图谱源于语义网络，以语义网的资源描述框架规范形式表示实例和关系，对于特定领域而言，计算机更易于理解知识图谱中实体及其关系，更智能化处理信息。同时，知识图谱以三元组[4]的形式存储实体以及实体之间的关系，语义网的发展促进了利用知识图谱进行知识问答。

问答系统相较于传统基于关键字的查询更能精确检索信息[5]。作为一种新的搜索方式[6]可以与用户进行交互，无需用户浏览整个文档自行阅读，能提供良好的检索体验，满足用户精确信息需求。基于知识图谱的问答系统依赖实体和关系链接，因此能够从语义角度提供智能化问答。自然语言处理技术的发展，使得问答系统能够处理分析用户提问，将用户需要的确切答案以自然语言的形式输出。

对于图书的选择停留于作者、流派、书名，因此会通过上网查阅相关信息，但是搜索引擎的返回结果只是一些相关文档集合的网页，不能直接给出精确的答案，而且需要自己阅读这些信息网页，并进而分析获取自己需要的信息。包含大量无关信息的网页会增加新生获取自己所需内容的时间，不能有针对性的给出结果，缺乏良好的交互性。因此，这次创新是基于知识图谱设计一种。在对图书信息作出常见问题数据进行处理后，构建图书信息知识图谱，将有望为图书信息提供知识检索、知识可视化等支撑。搜索某个问题时，不再是传统上的搜索结果，不需要搜索网页，而是直接呈现出与问题相匹配的结果，从而为人们提供了便捷化查询方式，使得人们对图书的选择和对阅读的兴趣有很强的提高。

此外我做了相关的调查报告，虽然数据量不是很大，但是也从某种程度上反应了现代人的读书问题。详细见附件一

## 创新产品描述

* 1. **国内外研究综述**

根据领域和用途，知识图谱可以大致分为常识知识图谱、百科知识图谱、领域知识图谱、语言知识图谱等，其中领域知识图谱专门服务于特定领域，近年来在各自的领域有着广泛应用，如医学知识图谱[7]、生物学知识图谱[8]、地理知识图谱[9]、产品知识图谱[10]等。深度学习推动了知识图谱的构建、推理及应用，采用深度神经网络模型在知识图谱上取得了显著效果[11]，并且知识图谱能够将抽象的知识符号化，实现机器的认知智能，这使得知识图谱成为人工智能领域应用场景的重要技术。

* + 1. 国外研究综述

为实现计算机阅读分析和理解海量数据，提供精准服务，万维网之父Tim Berners-Lee提出了构建能够将知识提供给机器程序处理的万维网[12]。维基百科极大地推动了知识资源增长，奠定了知识图谱的资源基础。Google首次将“知识图谱”应用于智能化检索服务，实现文本内容转化为图谱的节点结构，使计算机理解内容。随着Web2.0的兴起，大规模知识图谱主要包括FreeBase、DBpedia、YAGO[13]等构建了起来。

IBM公司的Watson问答系统依托于DBpedia和YAGO等知识库，在竞赛中战胜了人类选手。Siri语音助手使用的是以Wikipedia为代表的知识库。麻省理工学院的Start是第一个基于网络的问答系统，从本地托管或Internet远程访问中的一组信息资源抽取出信息，通过呈现这些文本和多媒体信息组件回答自然语言问题，Start使用自然语言注释的机制将问题与候选答案相匹配，目前可以回答数百个英语的地理问题[15]。AnswerBus是一个基于句子级别网络信息检索的开放域问题回答系统，支持多种语言[16]，AnswerBus与普通搜索引擎不同，在处理来自Web文档的句子时进行动态命名实体提取，只提取与问题类型匹配的实体。

* + 1. 国内研究综述

随着深度学习及知识图谱等相关技术的发展[17]，问答系统有了更高的分析推理能力和交互性。面向知识图谱问答系统从技术方法可以分为基于语义解析的方法和基于搜索排序的方法。前者把自然语言问句解析为结构化的查询语句，然后通过检索知识图谱从而获得答案。后者将问句与从候选答案中提取出来的特征进行比对、打分、排序得到最终答案。

目前大多数的知识图谱建立在开放领域的知识库上，Hao等[18]提出了一种端到端的神经网络模型，通过交叉关注机制根据候选答案的不同特征动态表示问句的语义，有效提升了问答匹配的效果。Qu等[19]利用RNN和CNN的优点提出了基于相似矩阵的卷积神经网络(AR-SMCNN)模型。为了克服传统知识图谱基于语义解析方法的缺点，Zhang[20]等提出了端到端的知识图谱问答模型。

国内的知名互联网公司也投入了对知识图谱的研究和构建，出现了百度知心、搜狗知立方等知识图谱，另外金融、医疗、教育、司法等各个行业也开始探索建立垂直领域知识图谱。以问答和对话为核心的聊天机器人已在国内应用起来，取得了不错的效果，如微软“小冰”、百度“度秘”等。

狗尾巴草智能科技使用了致力于构建中文链接开放数据的Zhishi.me[20]，其涵盖了三大中文百科知识，用于研发服务机器人、AI虚拟生命、语音交互技术等。

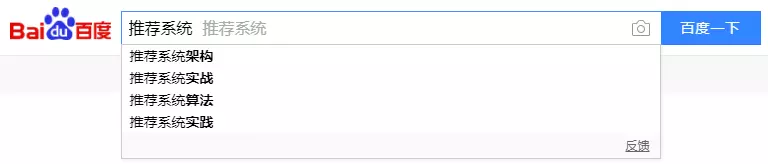
北京大学知识库问答系统[21]从图数据驱动的角度，提出了一种语义查询图，对用户问句中的查询意图以结构化的方式进行建模，从而简化为子图匹配的问题，使用关系优先和节点优先两个框架来构建语义查询图，根据实体和谓词匹配在执行查询语句时进行语义消歧，提高了系统精度和系统性能。

复旦大学知识工场实验室从大规模在线百科中获取知识库，构建了一个能够自动生成不断增长和更新知识库的中文知识抽取系统CN-DBpedia，其已经在问答、医疗、智能设备等领域广泛应用，API调用量达到了6亿次。

东师范大学研发的农业知识图谱，从百度百科和互动百科中爬取农业知识，使用KNN算法筛选实体并对实体分类，关系抽取部分通过语句解析和关系抽取模型，结合远程监督最终提取出关系，可以实现信息检索、实体识别与查询、智能问答等功能。此外，采用了众包的方式收集带标注的训练集数据。

### 什么是推荐系统

      推荐系统，正如它的字面信息一样，就是通过推荐内容满足用户个性化的需求，解决信息过载的问题的系统。推荐系统根据形式的差异接入了不同的场景，在大家的日常生活中就无时不刻都在享受这它的便利，当你一大早打开淘宝，扫一眼“猜您喜欢”，发现一个自己喜欢的宝贝，直接添加进购物车；打开今日头条，看了下自己感兴趣的新闻，中间看到一个自己不了解的内容，打开百度输入后，输入框下面展示了几个相关内容...

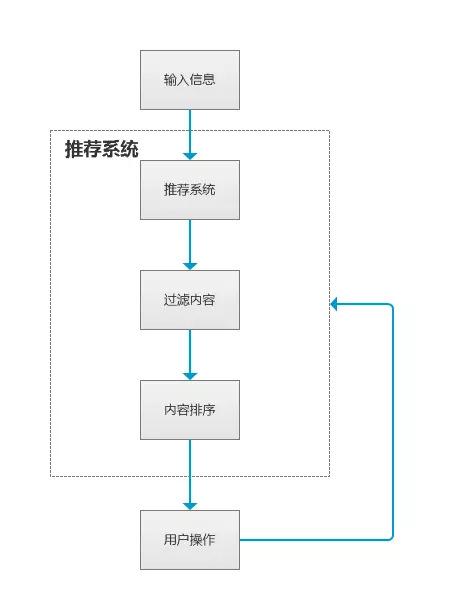


      根据推荐的形式能不能清晰地影响用户的操作可以把推荐划分为隐形推荐和显性推荐。隐形推荐不会对用户预期的操作产生影响，如：新闻排序，搜索结果排序等在用户不知不觉中给用户展现；显性推荐会改变用户预期的操作，如：输入联想、推荐问句等用户可以根据推荐的内容选择自己期望的内容。

        另外，根据推荐的阶段不同，也可以将推荐分为相关性推荐、预测式推荐、生成式推荐。相关性推荐根据用户当前信息，召回相似度较高的内容作为推荐的内容；预测式推荐为根据用户历史信息，可以是用户信息、操作记录、购买记录等，预测用户可能感兴趣的内容，作为推荐的内容；根据用户的信息推荐，不管是相关性还是预测式的都会导致推荐的内容随用户使用时长增加，变得内容单一，降低用户的新鲜感，因此还需要生成一些无关的内容作为推荐的补充，以满足用户的新鲜感，这就是生成式推荐。

### 推荐流程

      推荐的过程可以简单理解为三个步骤：召回、过滤、排序。首先系统根据获取到的信息，召回适合推荐内容，获取的信息可以是用户的搜索记录、购买记录、评论等。召回的内容中有的是这个用户不关注的，可能是他已经买过了的宝贝或者已经看过了的内容，这会儿就需要根据过滤的条件，将不需要的内容进行过滤。经过过滤产生的推荐集还需要根据内容的相关度进行排序，最后系统根据相关度的排序，将内容分配到对应的模块，这样用户就能看到自己感兴趣的内容了。



      有的系统也会将过滤放在第一步，先根据条件过滤一些输入信息，然后喂给推荐系统。这样能够减少推荐系统的计算量，缩短推荐系统处理时间，提高推荐系统的即时性，但是这么做也会存在一些问题：减少输入导致类别特征的内容丢失，影响推荐系统的内容数量与质量

### 推荐系统为什么要知识图谱

传统的推荐系统会根据用户的历史行为，为用户打上隐形标签，并为用户推荐相关的产品。假设用户订购了普吉岛的旅游产品，比如自由行，用户的原因可能是喜欢旅游，喜欢海岛，于是推荐系统为该用户打上了“喜欢海岛”的标签，为该用户推荐了很多海岛的相关产品；有些用户可能喜欢普吉岛的某些服务，例如露天泳池，某家酒店等，推荐系统则引入酒店等一些特征加入推荐模型中；某些用户因为去普吉岛只需要落地签，不需要额外的手续，于是推荐系统加入了签证相关的特征；某些用户因为去普吉岛的机票打折，喜欢泰国的一些旅游景点，喜欢海鲜等，如果将这些特征全加入推荐系统，会发现推荐系统变得很困难。  
酒店数据，机票数据，签证数据，景点数据等，每种数据都需要单独的数据库或者数据表去维护，将这些数据联合分析可能要做大量的BI工作，这些繁杂的分析都可以用知识图谱取代，这就是旅游业需要知识图谱的原因。

### 知识图谱怎么构建

1）定义schema，比如实体的类型，数据类型，属性类型，类别等  
2）做一些知识采集的操作，从文本中抽取关系，从外部知识库补充一些已有的三元组  
3）数据库迁移，从sql数据库中的数据迁移到知识库中，数据库备份，数据结构的转换  
4）实时更新就是检查数据的一致性，对重复的内容做知识融合，比如china和中国，尽量保证实体的唯一性。

实现层，大概分成六个步骤，分别是知识获取、知识抽取、知识融合、知识存储、知识推理、知识建模和知识发现

1. 知识获取是获取外部数据的方式，包括爬虫和实时入库的技术方法；
2. 知识抽取就是，对三元组进行知识的抽取，包括实体抽取、关系抽取和属性的抽取；
3. 知识融合就是，抽取出来之后，存在很多的数据冗余和噪声，要去做实体的消歧，数据的整合；
4. 知识存储，刚才讲了，实际是要构建一个三元组RDF的数据结构，如果把所有的顶点和边构造出来之后，要对他进行图数据库的存储；
5. 知识推理，刚才也讲到了，如果要做一些深层次的知识问答，就要做很多的训练，无论有监督的还是半监督的；知识建模更多的是去理解语义，涉及到属性的映射，实体的连接；
6. 知识发现，两大主要的应用是知识的检索和知识的问答。这些构建了知识图谱的实现层。

另外还需要做的是实体对齐和实体消歧。  
关于实体对齐。举例来说，比尔盖茨这四个字是中文名称，Bill Gates是他的英文名称，但其实这两个指的是同一个人。由于文本的不一样，开始的时候导致这是两个实体。这就需要我们对它进行实体对齐，把它统一化。  
另外是实体消歧。举例来说，苹果是一种水果，但是在某些上下文里面，它可能指的是苹果公司。这就是一个实体歧义，我们需要根据上下文对它进行实体消歧。

### 怎么跟推荐系统结合

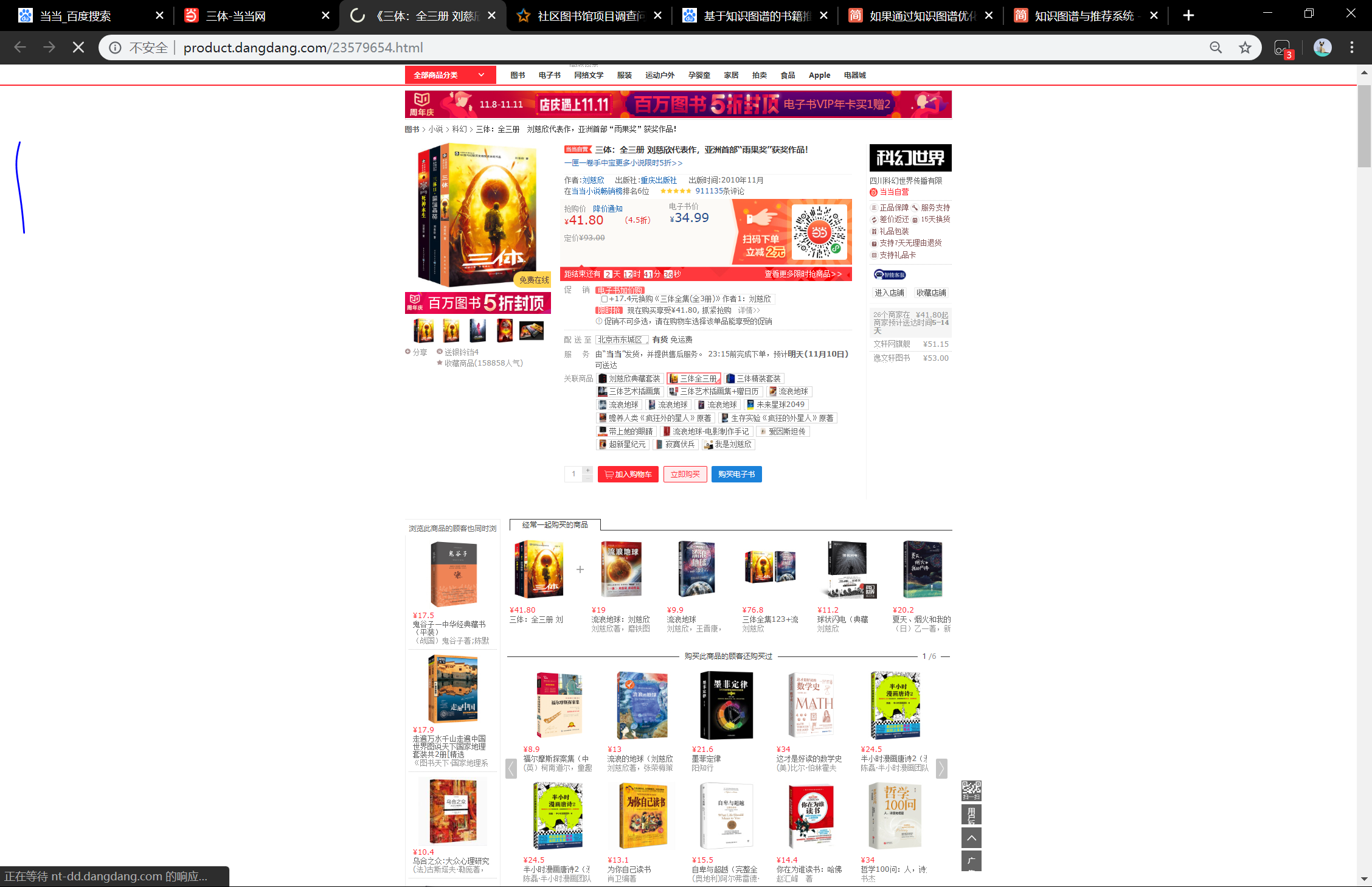
知识图谱一般不用于基于特征的推荐系统，一般用于基于路径的推荐，分为两种meta-path和meta-graph，用户喜欢普吉岛的酒店，喜欢spa服务和泳池，可以建立这样的一条路径作为推荐系统的特征，参与计算。缺点：需要提前设定这样的路径，不支持自动搜索路径。知识图谱特征学习，将特征转化为向量的形式，辅助推荐。应用于embedding，协同过滤中只考虑user特征和item特征，可以利用知识图谱作一些特征，embedding有很多方法，深度学习。

### 图谱应用的难点

知识图谱虽然在推荐系统中应用存在优势，但是在实际应用中会因为它的种种难点被限制应用，下面和大家一起讲讲图谱应用的困难。

## 知识图谱schema维护

      在推荐系统中应用的图谱都是大规模的图谱，实体都是在万级的，像阿里的商品图谱甚至达到了十亿级。那么大的图谱完全由人工运营维护肯定是不现实的，实际上这些图谱也的确由系统自动进行维护，人工只是辅助进行运营。系统通过现成的表结构数据、机器阅读理解抽取的实体与关系自动构建知识图谱。例如：的商品图谱部分数据来源就是宝贝下面的商品详情：



        但是目前的技术还不能做到100%的自动构建准确，因此构建后如何筛选出有问题的关系就需要人工借助工具进行调整了，常见的需要人工纠正的有:

        1.同宝贝在不同商家出现不同详情的情况，需要人工二次确认

        2.宝贝详情变化后，需要人工更新

### 推荐的时效性差

    图谱的量级达到了一定，如何快速的万级亿级的实体和属性中找到对应的数据，对于模型来说是一个十分艰巨的工作。另外大规模的图谱，实体之间的关系密切，如果做到的二元遍历，那么延伸出的实体也是指数量级的，无法直接拿来做推荐。所以图谱推荐的时效性较差，不适合应用于需要实时返回推荐结果的场景，所以图谱推荐往往应用在用户使用的间隙生成推荐的内容。例如：资讯推荐、猜您喜欢等

### 在线维护困难

    图谱中实体的关系十分紧密，这就导致了修改一个实体或者一条边，那对应的变化可能是几百个实体和边，一个小小的改动可能就是蝴蝶效应，而且恢复困难。所以图谱服务都是通过本地数据应用于中台服务中，需要更新图谱时，再将本地运营的图谱发布到服务的本地数据库中应用。

### 如何避免脏数据对图谱更新的影响

      在购物网站推荐的场景中，图谱中实体与实体之间的关系是会有概率值来表示两个实体之间的相关度的。这个概率值会根据用户的购买操作记录自动评估实体与实体之间的概率。但是有时新店为了提高信誉和宝贝的评价，就会发生刷单的行为。刷榜单的行为会导致图谱中概率边的数值被影响，产生不准确的数值，导致推荐出相关度不足的结果。这种情况一般通过清洗订单数据，只将高置信的购买记录作为图谱更新的评估数据。

基于知识图谱的推荐只是推荐系统的一部分，推荐系统是一个复杂的系统。如今在互联网的下半场，推荐系统越来越被重视，图谱在推荐系统中的应用目前还比较浅，期望图谱的落地更加成熟。

## 市场效益估算

经济可行性

本课题的迎新问答系统使用的是Python语言，开发工具为PyCharm。用于训练的数据是公开的语料和爬取的淮阴工学院新生常见问题数据，操作系统是Windows 10，图谱数据库使用的是Neo4j，Neo4j社区版是开源的，具备Neo4j的全部功能而且适合用来构建小型项目以及前期的开发调试。因此系统在开发过程中所采用技术和软件均是开源的，不需要额外的费用。由于实体识别和关系抽取涉及到的模型已经训练好，而且本系统对硬件的要求较低，所以在配置较低的电脑上也能够运行。因此本课题研究的系统在经济上是可行的。

此外许多电商、相关微店的推荐方式可以运用这种模式。此外也可以运用到学校的图书馆中。