自然语言处理编程大作业报告-5kg

1. 小组成员信息及分工

学号	姓名	分工
201928016029017	胡川	**组长**，知识获取、数据清洗、知识融合、智能问答、图谱可视化、报告撰写
201918018137010	陈楚珂	知识获取、数据清洗、知识融合、报告撰写
201928013329004	张祥祥	知识获取、报告撰写
201928000229002	张欢	知识获取、报告撰写
201928000243001	黄晓辉	知识获取、知识融合

2. 任务定义

知识图谱用节点和关系所组成的图谱，为知识建模，运用“图”这种基础性、通用性的“语言”，“高保真”地表达这个多姿多彩世界的各种关系，直观、自然、直接和高效。当今社会商品类型繁多，信息分散且价值密度低，如何准确快速地搜寻满足自己需求的产品成为了日常难题。小组旨在通过建立化妆品 (以口红和香水为主)的知识图谱，全面且精准地建立相关知识库；基于知识图谱设计智能问答系统，帮助科学选择符合需求的产品。根据需求，设定以下任务：(1)多源知识获取及融合；(2)图谱可视化；(3)基于图谱的问答系统开发。以 YSL 和纪梵希两个品牌口红为例，本系统应当能够回答“YSL 的口红系列一共有多少”、“300 左右的 YSL 口红有哪些”、“ 纪梵希禁忌之吻 N4 的价格区间是什么”、“和纪梵希禁忌之吻 N4 的价格区间相同的 YSL 口红有哪些”等问题。

3. 方法描述

图 1 研究方法

3.1 知识获取：分别从 APP 和电商官网两个数据源中，分别使用 Fiddler 和 Selenium 工具进行信息获取。(1) Fiddler 是位于客户端和服务器端的 HTTP 代理，也是目前最常用的 HTTP 抓包工具之一，它能够记录客户端和服务器之间的所有 HTTP 请求，可以针对特定的 HTTP 请求，分析请求数据、设置断点、调试 web 应用、修改请求的数据，甚至可以修改服务器返回的数据。Fiddler 要比其他的网络调试器要更加简单，因为它不仅仅暴露 http 通讯，还提供了一个用户友好的格式。Fiddle 包含一个简单却功能强大的基于 JScript .NET 事件脚本子系统，可以支持众多的 http 调试任务，并且能够使用.net 框架语言进行扩展(如图 2 所示)。(2) Selenium 是一个 Web 的自动化测试工具，最初是为网站自动化测试而开发的。Selenium 可以直接运行在浏览器上，它支持所有主流的浏览器，可以接收指令，让浏览器自动加载页面，获取需要的数据及实现页面截屏。(3)正则表达式匹配：使用单个字符串来描述和匹配一系列符合某个句法规则的字符串。通过正则表达式，检索、替换符合某个模式的文本，实现高效、方便地处理大量数据。

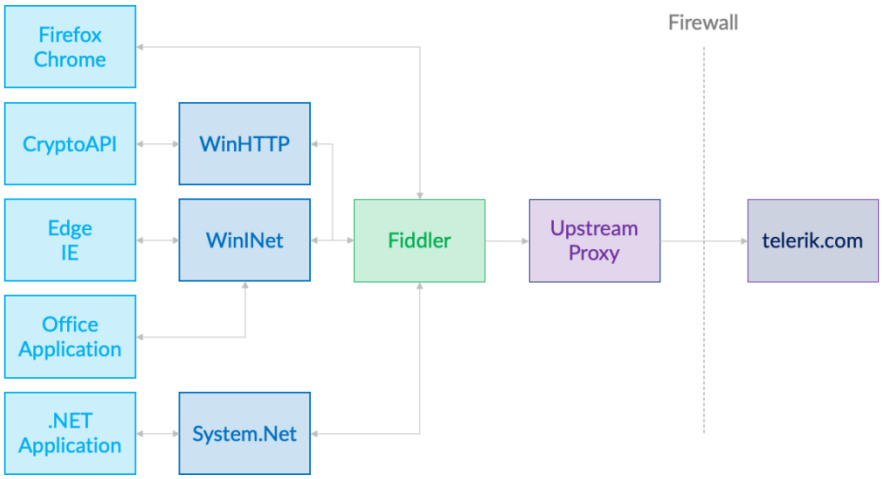

图 2 Fiddler 工作示意图

3.2 知识融合：对来源于 App 和网页的知识进行融合，实现对已有知识图谱进行补充、更新和去重。(1)利用实体的属性信息判定不同源实体是否可进行对齐；(2)使用 K-means 聚类方法，通过选取聚类中心，计算每个对象与聚类中心之间的欧式距离，以最小距离为原则分配对象，实现数据扩充。

3.3 知识存储：Neo4j 是一个目前最流行的、具有高性能的、成熟的图形数据库。基于图相关的概念描述其模型，数据将存为节点和关系，支持海量数据和迅速的

图查询特点。Neo4j 在进行数据存储时，通过自底向上构建的本地化的图数据库，能更快更好地管理节点与关系：支持大数据集合并且可以不断地扩展容量；声明式图查询语言的表现力丰富，查询效率高，同时有良好的扩展性；通过 ACID 事务保障数据安全。

3.4 知识应用：(1)知识图谱可视化：为屏蔽不同图数据管理系统、数据模型和协议之间的差异，简化图数据交互分析操作的复杂度并提高上层服务应用的通用性，使用基于现有图数据存储管理系统、查询协议和可视化分析工具的一种适合多元异构图数据管理系统的交互分析框架——InteractiveGraph(如图 3 所示)。其核心概念是图谱在线分析交互协议(InteractiveGraph Protocol, IGP)，IGP 协议用于定义图谱分析应用和数据服务端之间的交互协议和数据格式，定义了连接类、浏览类、探索类、实体匹配类、路径查询类等 5 类接口。InteractiveGraph 框架将前端的可视化需求转换为 IGP 的数据请求(Request)，在数据服务端通过软件模块将请求转义为对应图数据管理系统的图查询语句或者服务调用，并将查询的结果封装为 IGP 的数据响应(Response)，从而实现对不同图数据管理系统在查询语言、数据模型、服务结构等方面差异的屏蔽，增强交互分析应用的可扩展性和灵活性。(2)智能问答：问答系统使用一种基于规则和句法分析的问句转换方法，通过识别问题中的命名实体、疑问词、关键词等信息，将句法树转换为问题图并解析成查询语句，最终在图谱中查询并返回答案。

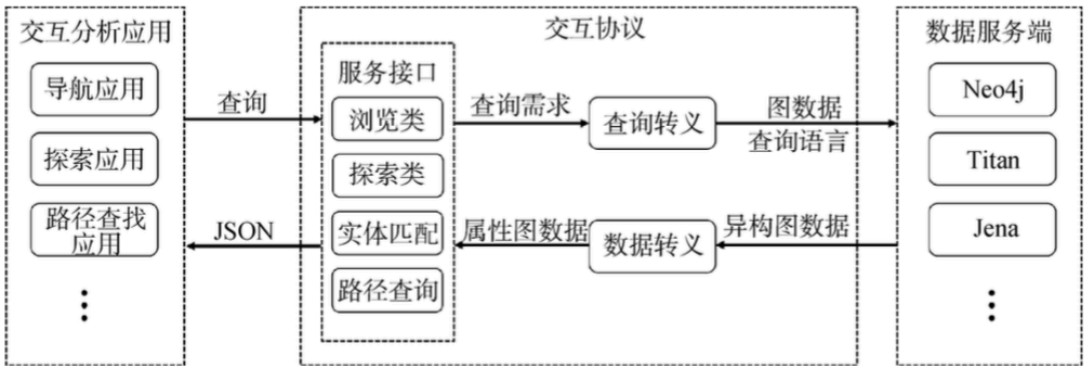

图 3　InteractiveGraph 结构示意图

4. 实现细节

通过抓包及爬虫技术从 APP 和网页端进行知识获取，基于正则表达式匹配方法对数据进行清洗，将数据持久化为.csv 文件。待数据准备完成后，将所需实

体及关系导入 Neo4j 数据库中，利用已创建的数据库，利用多元异构图数据管理系统的交互分析框架 InteractiveGraph 实现 Web 端数据可视化，运用基于规则和句法分析的问句-问题图的转换方法，有效地将自然语言问题解析为知识图谱查询语句，完成化妆品领域的问答系统开发。

4.1 知识获取

为了能够获取信息丰富且格式工整的半结构化数据，百科型和图鉴型的网站是最好的获取对象。然而，由于化妆品新品发售快、信息不一致和品类繁杂等特点，化妆品百科式的网站目前还没有在网络上出现。在经过一些调研工作之后，我们找到了一些符合条件的移动端应用程序。对移动端 APP 数据爬取主要采用抓包和分析 API 的方式。

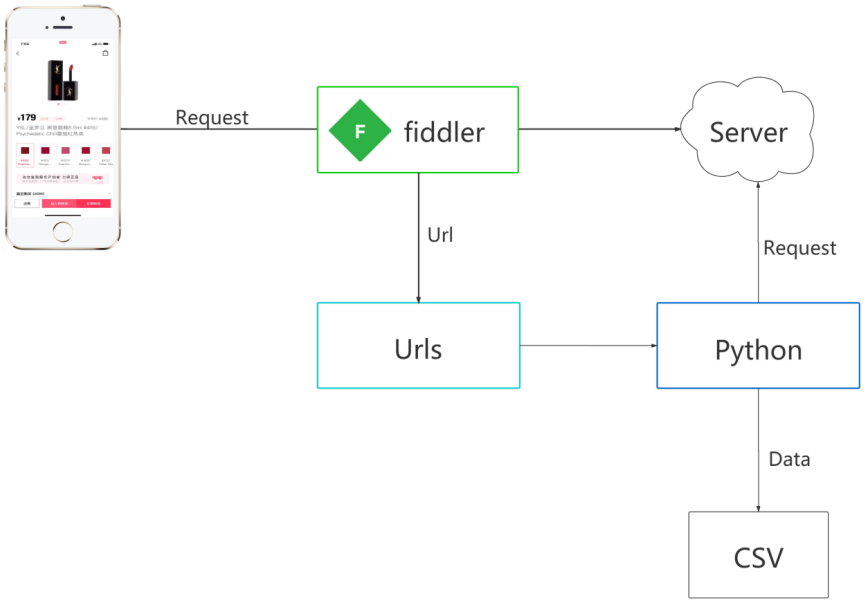

图 4　fiddler 流程框架

首先使用 Fiddler4 对 APP 发出的请求进行拦截分析。经过观察发现，该 APP 对服务端发送的请求除了类似类别、品牌、页数等检索信息外，还存在一个校验参数 token。Token 通过 APP 内置的加密算法对请求参数计算得来；也就是说，不同的检索请求对应不同的 token。这意味着，如果想通过 API 直接获取数据需要破解 APP 内置的加密算法。要想得到 APP 内置的加密算法，需要解开 APP、反编译并分析其源码。然而，由于该 APP 经过了严格的“加壳”处理，在多次尝试对 APP 反编译失败后，放弃了分析加密算法自动获取 API 数据的方法。最终，我们选择使用了手动发送请求、Fiddler4 拦截请求参数、Python 爬取数据的

策略。由于该策略部分步骤需要手动操作，因此在这一步花费了较长的时间。通过抓包，获取各类化妆品(如口红、香水、粉底液等等。由于时间有限，本项目关注口红和香水两类产品) 的产品信息。以口红为例，主要信息包括口红所属品牌、品牌所属国家、产品名、图片信息、价格等。

此外，选择使用 Selenium 爬取京东电商官网信息，作为口红信息补充。针对京东官网内所有主流品牌的口红，获取的数据包括口红对应的品牌、类型、色号、价格、图片、买家印象、评价人数、以及对应的 5 条评价。由于京东口红商品数据很多，为了保障爬取数据不重复、全面性、准确性，能包括所有主流口红信息，在爬取和对应的数据清洗上花费了两周时间，最终得到了较为准确、全面的口红信息，能一定程度上反应当前市场口红信息。

```
title = driver.find_element_by_xpath('/html/body/div[6]/div/div[2]/div[1]').text
color = driver.find_element_by_xpath('//div[@id="choose-attr-1"]/div[2]/div[{}]/a/i'.format(i)).text
price = driver.find_element_by_xpath(
'/html/body/div[6]/div/div[2]/div[4]/div/div[1]/div[2]/span[1]/span[2]').text
picture = driver.find_element_by_xpath('//div[@id="spec-n1"]/img').get_attribute('src')
impression = driver.find_element_by_xpath('//div[@id="comment"]/div[2]/div[1]/div[2]/div').text
comment_num = driver.find_element_by_xpath('//div[@id="comment"]/div[2]/div[2]/div[1]/ul/li[1]/a'
).text
for i in range(5):
comment = driver.find_element_by_xpath(
'//*[@id="comment-0"]/div[{}]/div[2]/p'.format(i + 1)).text.replace(',', ' ').replace('\n', ' ')
goods_list.append(comment)
```

基于正则表达式匹配方法，处理获取知识。结合正则表达式从文本字符串中获取我们感兴趣的内容，实验中，主要提取产品所属品牌的中英文名，网页色块图、产品名字、价格、销量和产品评价等，最终得到可用于构建图谱的结构化数据(.csv)。

```
def wash_kouhong(self):
    data = self.data['kouhong']
    data = data[~data["goods_name"].str.contains(r"\*")] #洗掉套装
    data.loc[:, 'goods_name'] = data['goods_name'].str.replace('N°', '#')
    data.loc[:, 'goods_name'] = data['goods_name'].str.replace(r'N\d', '#')
    data.loc[:, 'goods_name'] = data['goods_name'].str.replace('KIKO', 'KIKO/奇寇')
    patt = r'^(?P<engName>[^\u4E00-\u9FA5]*)/(?P<chiName>[ ]?[^ ]+) (?P<name>[^ ]+) (?P<Other>.+)'
    df_new = data['goods_name'].str.extract(patt, expand=True)
    df_new.loc[:,'name'] = df_new['name'].str.replace(r'[\dg\.ml]','') # 洗掉规格
```

```
    # 开始洗尾巴
    patt2 = r'(?P<code>#.*\d+)?(?P<sub_name>.+)?'
    df_new_new = df_new['Other'].str.extract(patt2, expand=True)
    df_new_new.loc[:, 'code'] = df_new_new['code'].str.replace(r'[^\d]+', '') #洗掉非数字
    df_new_new.loc[:, 'sub_name'] = df_new_new['sub_name'].str.replace(r'[^A-za-z\u4E00-\u9FA5 ]+', '') #洗掉非中英文字符或空格
```

4.2 知识融合

根据品牌、产品系列和产品三个属性信息，对来自 APP 和网页端的数据进行实体对齐。此外，通过口红的颜色信息提取 RGB 参数，基于常用口红色系划分，选取 14 个色系中心，计算采样数据与聚类中心的距离。基于最小距离分类原则，将样本划分为 14 类，将口红颜色大致分为 14 个色系(见图 5，如砖红色、豆沙色、珊瑚橘等等)。

```
def distance(x):
    dist = np.zeros([1,len(center)])
    for i in range(len(center)):
        d = np.sqrt((x[0]-center[i][0])**2+(x[1]-center[i][1])**2+(x[2]-center[i][2])**2)
        dist[:,i] = d
    index = np.where(dist == dist.min())[1][0] # minimum distance label...
    return index
```

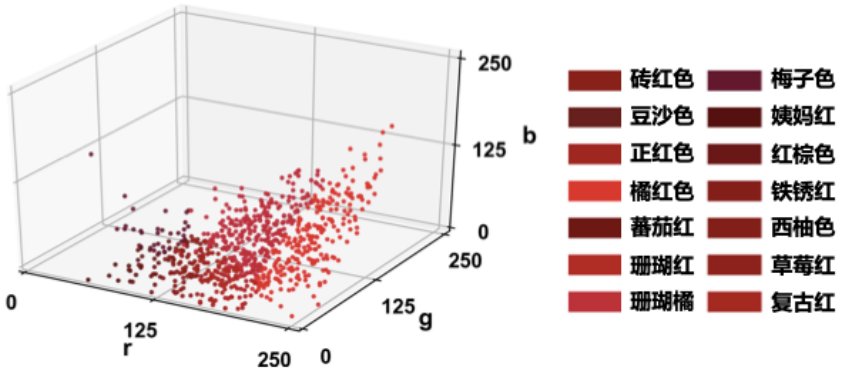

图 5 口红颜色分类

4.3 基于 Neo4j 图数据库的数据存储

利用获取知识信息，基于 Neo4j 图数据模型，建立起化妆品知识图谱(图 6)。图谱实体包括品牌(Brand)、品牌所属国家(Country)、香水产品(Perfume)、口红系列(LipstickType)、口红产品(Lipstick)、价格区间(PriceRange)、评论(Comment)和

颜色(Color)共 8 个节点；定义 6 类关系，如口红产品—[have comment]—评论、口红产品—[one of]—口红系列等等。

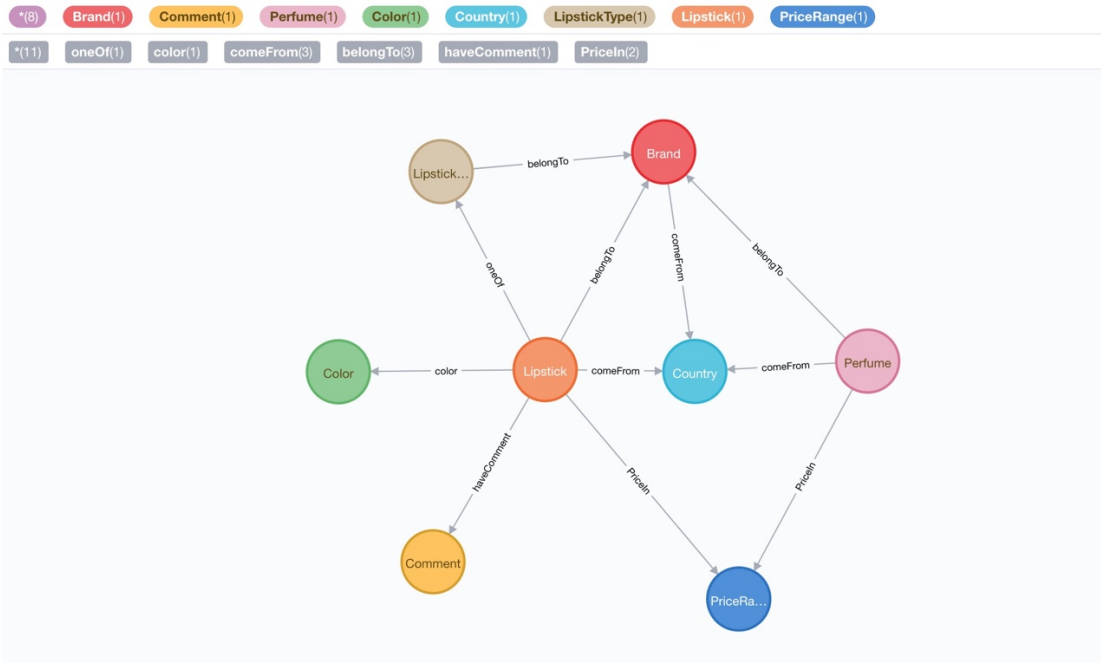

图 6　化妆品知识图谱

特以雅诗兰黛花漾倾慕爱心限定唇膏 333 为例(图 7)，根据图谱描述，它是属于美国雅诗兰黛品牌下花漾倾慕爱心限定系列的产品，价格在 200 元左右，所属色系为珊瑚红。根据网上评论，大多数人认为它很显气质、嫩滑水润、款式漂亮、温和滋润、颜色美丽等等。

图 7　某产品的知识图谱

4.4 问答系统

基于知识图谱的问答系统分为两个步骤：首先把自然语言问题转化成对应的查询语句，然后使用查询语句检索知识库得到答案。其中，如何正确的将自然语言问题解析为查询语句是问答的关键。本系统运用一种基于规则和句法分析的问句-问题图的转换方法，能够有效地将自然语言问题解析为知识图谱查询语句(Cypher)，具体分为实体链接、句法分析、关键词识别、问题图转换、查询语句生成和知识查询和答案生成共 6 个步骤，下面进行详细叙述。

中文含义	问句举例
产品咨询	(1) 300 左右的 YSL 口红有哪些 (2) 有哪些 YSL 口红在 300 左右 (3) YSL 口红在 300 左右的有哪些 (4) 颜色漂亮的口红有哪些 (5) 和纪梵希小羊皮唇膏 307 的颜色相同的迪奥口红有哪些
数量查询	(1) YSL 的口红系列一共有多少 (2) YSL 的口红一共有多少
价格查询	(1)纪梵希禁忌之吻 N4 的价格是什么 (2)纪梵希禁忌之吻 N4 的价格区间是什么
特征评价查询	(1) 小钢笔印迹唇釉 113 的品牌是什么 (2) 禁忌之吻大理石纹唇膏 25 的颜色是什么 (3) 黑管唇膏 225 的评价有哪些

(1)实体链接

实体链接的任务是识别出自然语言问句中涉及到的实体。在本系统中，实体链接使用字典匹配的方式完成。例如对于文本“YSL”和“圣罗兰”都指向知识图谱中的实体<label=“Brand”, id=1>。例句“和小羊皮唇膏 307 的颜色相同的迪奥口红有哪些？”的实体链接过程如下所示。

(2)句法分析

在实体链接之后，问答系统对问句进行句法分析得到问句的句法树。本系统句法分析使用 HanLP (https://github.com/hankcs/hanlp)自然语言处理工具包完成。例如对上述问句经过句法分析，可以得到图 8 句法树，其中蓝色节点为实体。

(3)关键词识别

在得到问句句法树后，需要对问句中的关键词进行识别。问句关键词包括：问题词（Question）、标签词（Label）、属性词（Attribute）和值（Value）等。然后，删去叶子节点上的未识别词语。例如，对上述句法树，关键词识别后的句法树如图 8(右)所示。

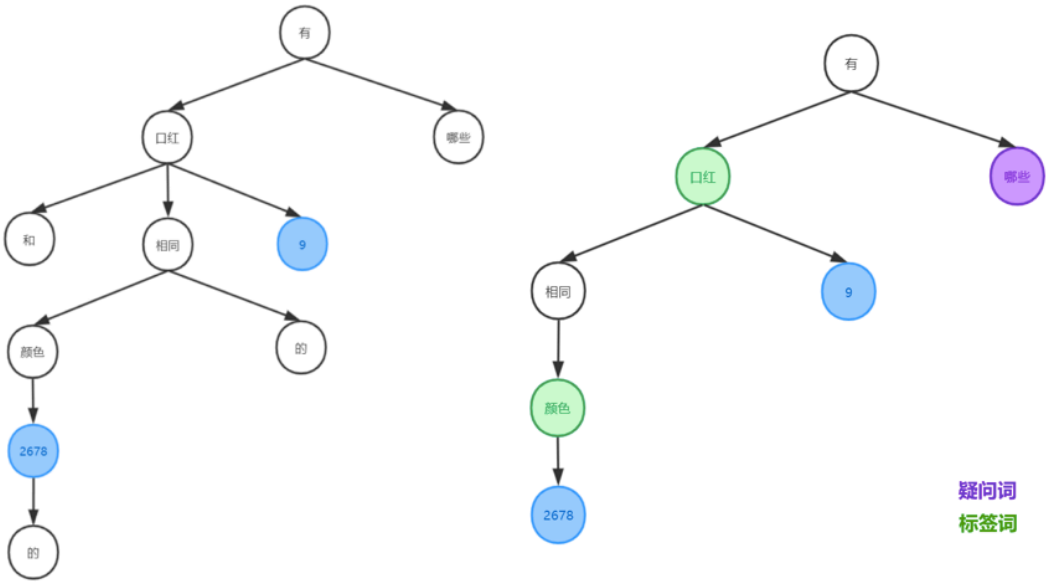

图 8　左：句法树；右：关键词识别后的句法树

(4)问题图转换

将标注了问题词后的句法树转换为问题图的主要原因是为了解决问句同义异构现象。例如问句“和小羊皮唇膏 307 的颜色相同的迪奥口红有哪些？”和“有哪些迪奥口红和小羊皮唇膏 307 的颜色相同？”以及“迪奥口红和小羊皮唇膏 307 的颜色相同的有哪些？”。它们所包含的中文含义相同，都是询问相同颜色的产品，但句法树却不一样(如图 9)。所以，问题图转换的关键就是要将一系列同义异构问句转换成相同的数据结构。

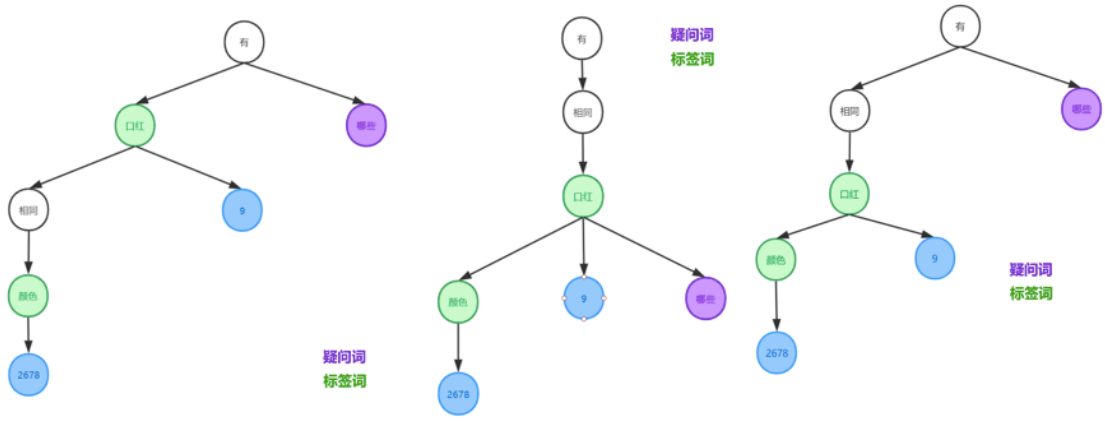

图 9　关键词识别后的句法树

问题图转换分为两个步骤，首先把句法树从树结构转换成无向图结构存储。随后对无向图中的未标注词进行删减操作(如图 10 所示)：

- 如果未标注词的度为 1，直接删除；
- 如果未标注词的度为 2，删除并连接左右节点；
- 如果未标注词的度大于 2，删除并将其他边连向距离疑问词最近的那个节点。

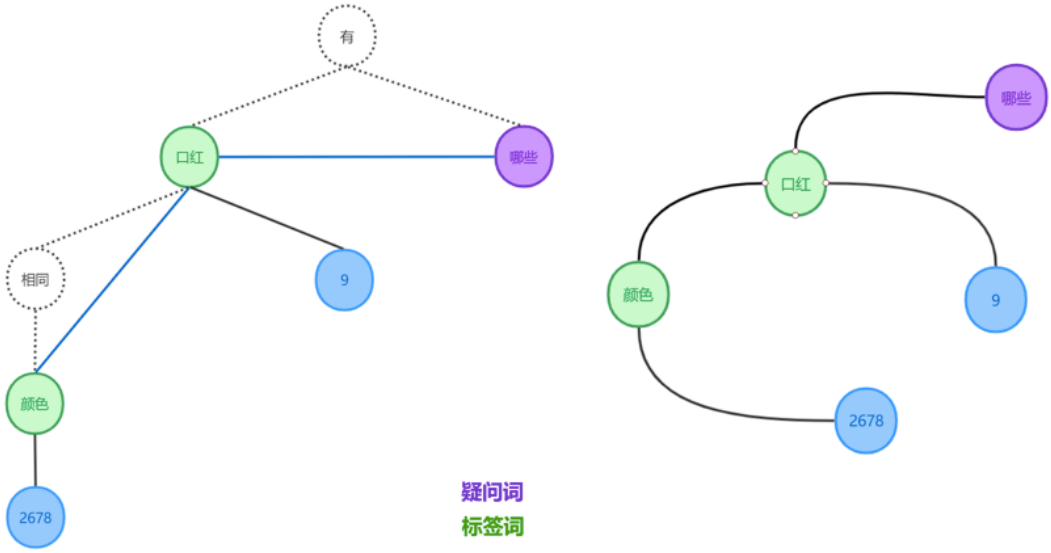

图 10　问题图生成

(5)查询语句生成

生成过程从问题图的疑问词节点开始，递归调用、逐层分析，如图 11 所示。

(6)知识查询和答案生成

解析出查询语句(Cypher)后，使用 Cypher 在 Neo4j 中检索结果并返回。

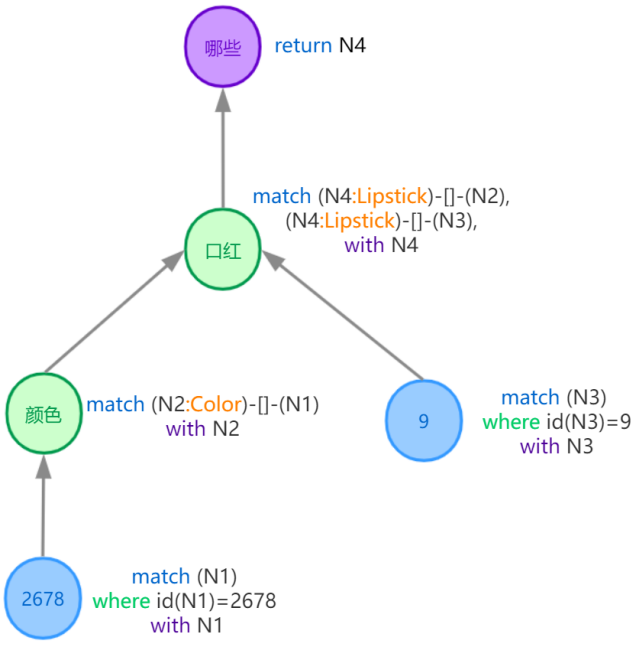

图 11　问题图生成

4.5 基于 Web 的交互式图谱可视化

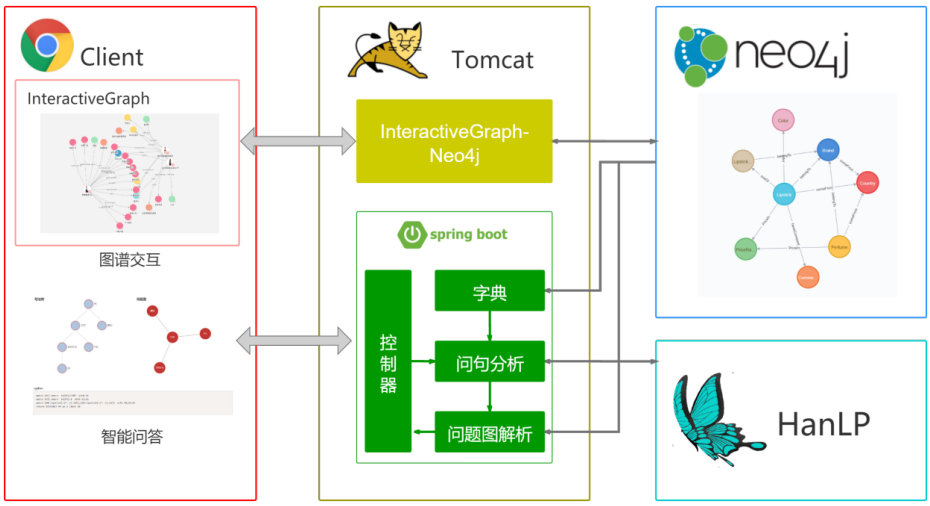

图 12　Web 系统架构

本系统实现了一个基于 Web 的交互式可视化系统以完成图谱交互和问答系统可视化。系统分为图谱交互和智能问答两大模块。前者实现了对知识图谱中的实体查询、实体信息浏览、实体展开等操作。后者可以让用户输入自然语言问题，提交到问答系统并展示返回的答案。知识图谱交互模块使用了 InteractiveGraph 工具(https://github.com/grapheco/InteractiveGraph)，该工具包括前端显示插件 InteractiveGraph 和数据库连接中间件 InteractiveGraph-Neo4j。用户通过搜索框搜

索到的化妆品会自动添加到图谱可视化控件中，供用户进行浏览、展开等操作。问答系统的后台由 SpringBoot 开发，解析前端发送的用户问题，查询数据并返回。此外，页面中还使用了 Echarts.js(https://www.echartsjs.com/en/index.html)展示句法树和问题图。

5. 实验结果

详见：http://60.205.227.226:8080/CosmeticsKG-Web/

(1)产品搜索

图 13　产品搜索示例

(2)产品参数对比

以下图为例：倾慕唇膏 333(记为 A)属于美国雅诗兰黛品牌下倾慕唇膏系列，全新挚爱哑光唇膏 400(记为 B)属于意大利品牌乔治阿玛尼品牌下全新挚爱哑光唇膏系列而莹亮纯魅圆管唇膏 84(记为 C)属于法国圣罗兰品牌下莹亮纯魅圆管唇膏系列。三者皆为复古红色系且价格在 200 元左右，且普遍认为 A 与 C 都拥有款式漂亮的特点。

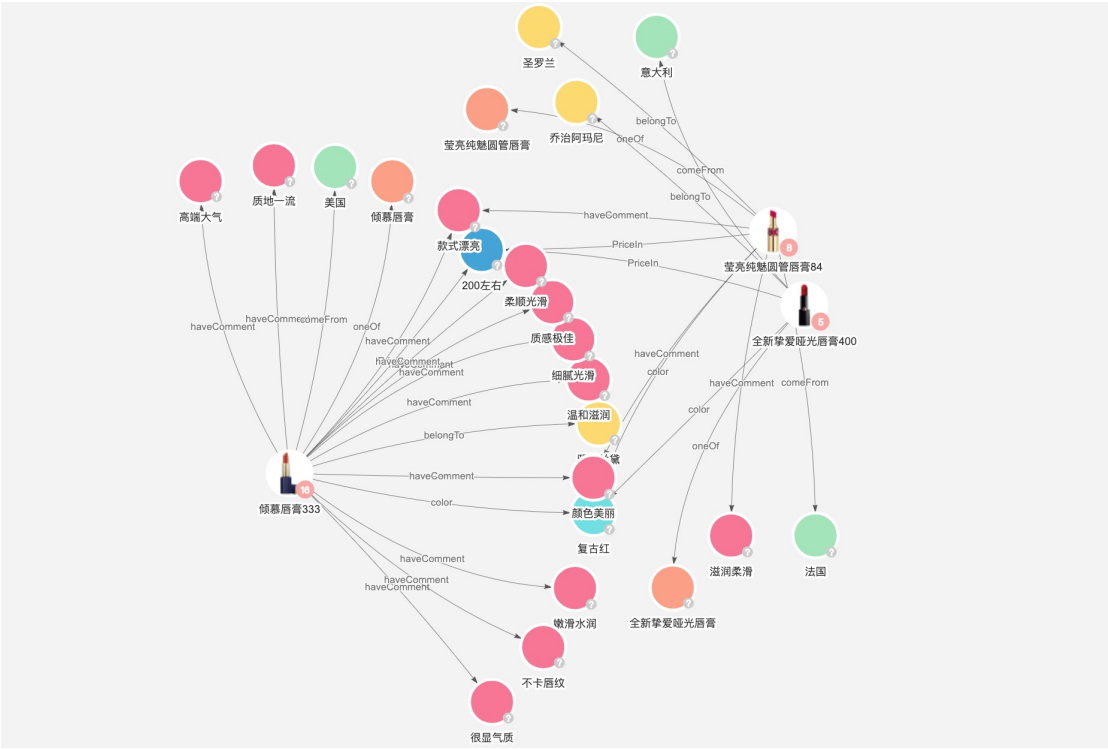

图 14　产品参数对比示例

(3) 产品问答

基本问答实现，见图 15、16 和 17。

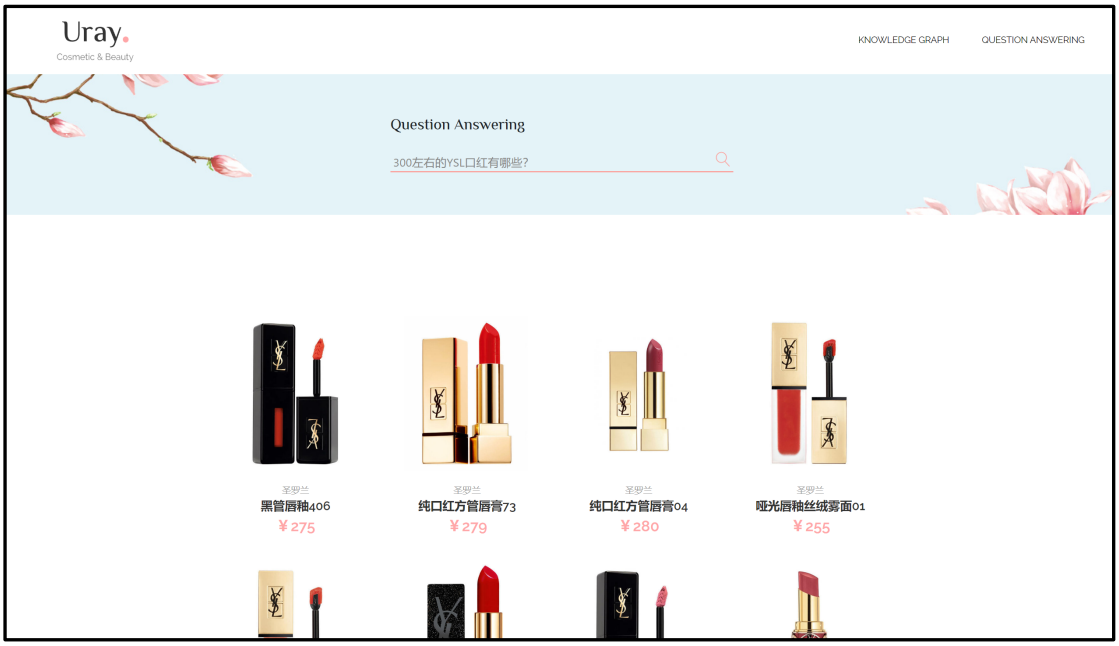

图 15　问答示例 1

查询“颜色美丽的复古红有哪些？”以及“水水嫩嫩的口红有哪些是复古红的？”。查询结果如下，结果表明系统能够正确分析句法树，正确转换问题图，

及正确解析问题。

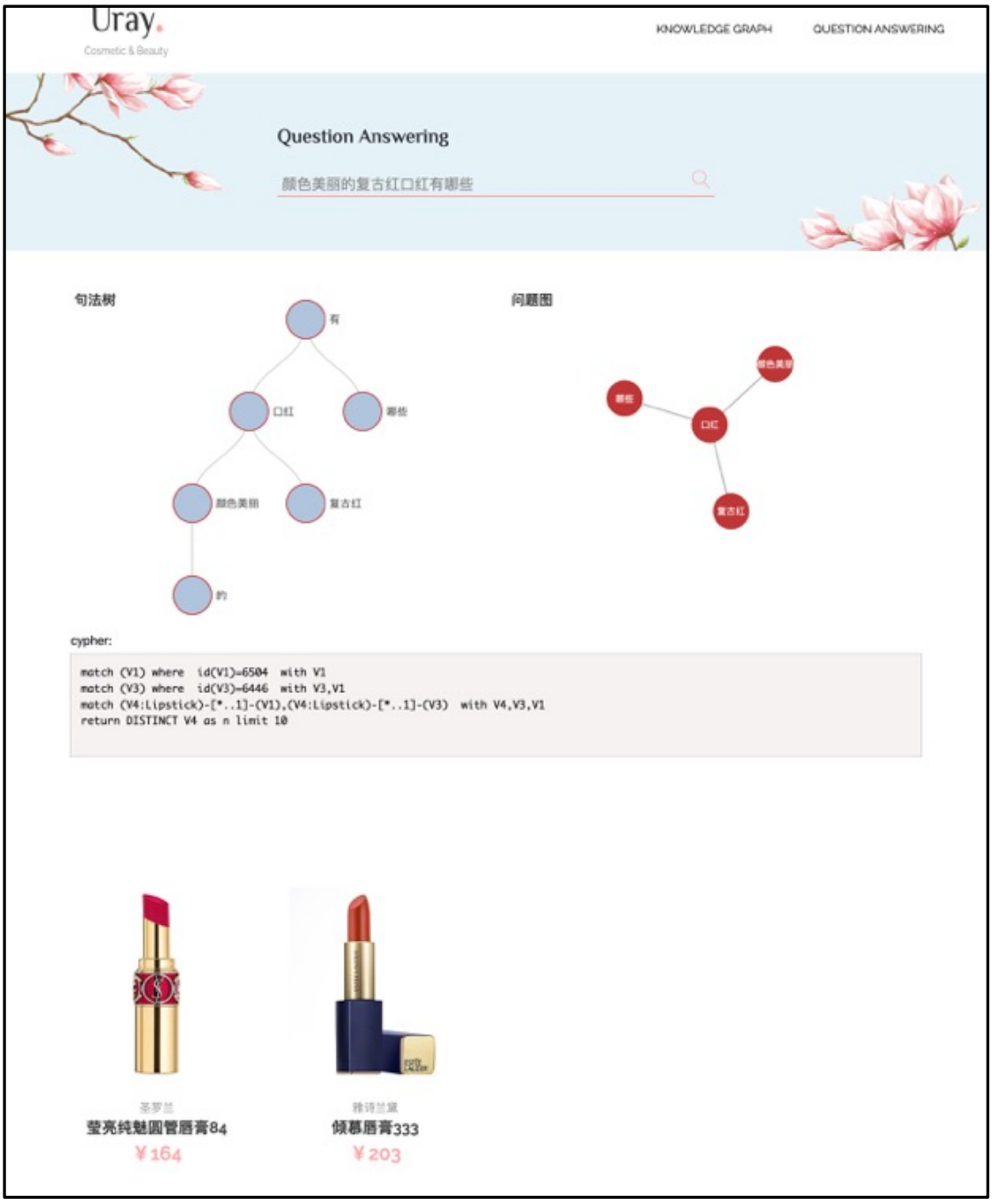

图 16　问答示例 2

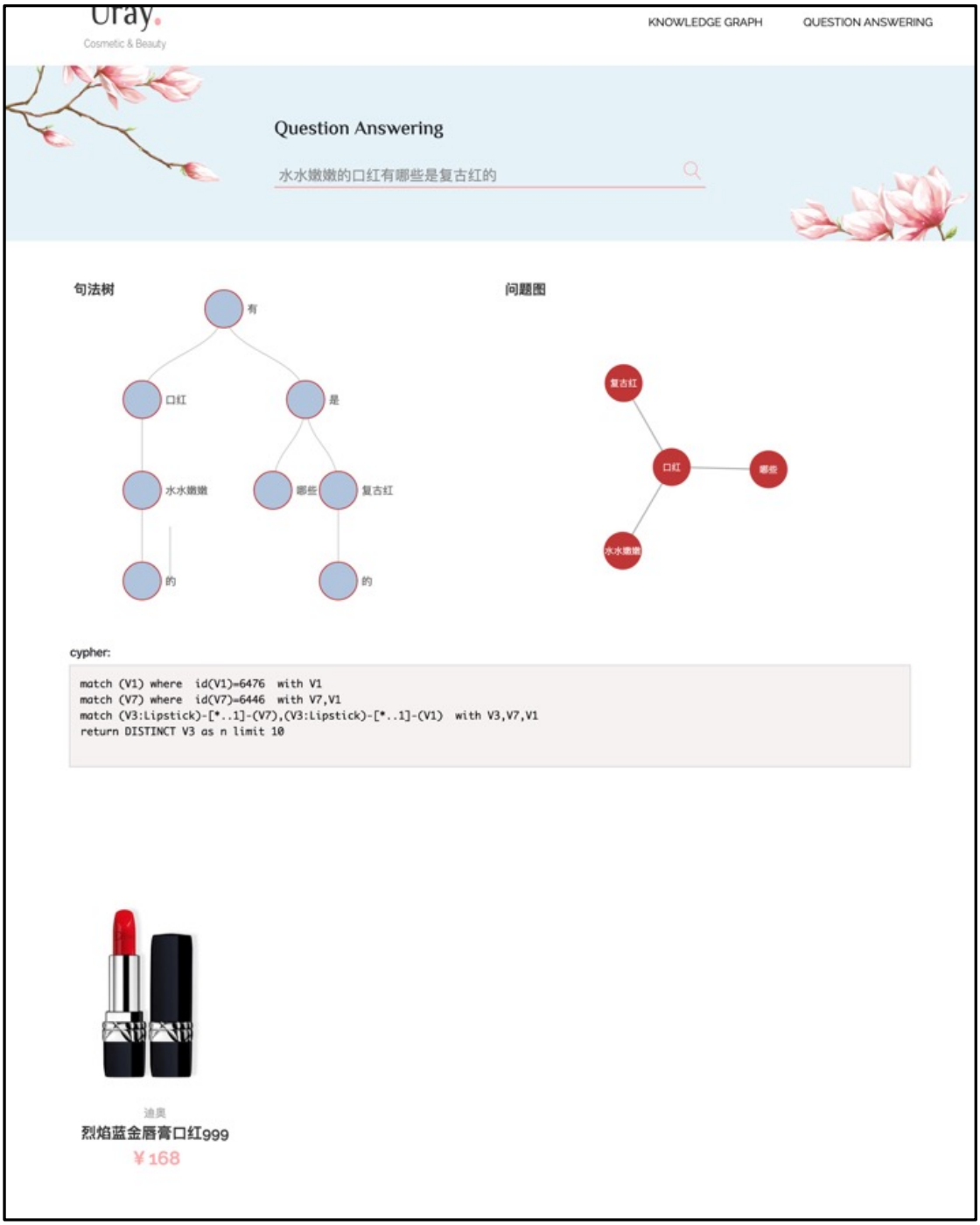
Uray.
Cosmetic & Beauty
KNOWLEDGE GRAPH
QUESTION ANSWERING
Question Answering
水水嫩嫩的口红有哪些是复古红的
句法树
有
口红
是
水水嫩嫩
哪些
复古红
的
的
问题图
复古红
口红
哪些
水水嫩嫩
cypher:
match (V1) where id(V1)=6476 with V1
match (V7) where id(V7)=6446 with V7,V1
match (V3:Lipstick)-[*..1]-(V7),(V3:Lipstick)-[*..1]-(V1) with V3,V7,V1
return DISTINCT V3 as n limit 10
迪奥
烈焰蓝金唇膏口红999
￥168

图 17　问答示例 3