|  |  |  |
| --- | --- | --- |
| 文件状态：  [ ] 草稿  [√] 正式发布  []正在修改 | 文件标识： | SYZH1104-TYPT-04 |
| 当前版本： | V1.0 |
| 作者： | 邓虹玉 |
| 完成日期： | 2018/1/28 |

基于语义网的站内搜索引擎模块改造项目

概要设计

**北京慧博科技有限公司**

**2016年1月**

修订历史版本

|  |  |  |
| --- | --- | --- |
| **日期** | **修订内容** | **修订人** |
| 2018-01-28 | 创建文档版本。 | 邓虹玉 |
| 2016-02-04 | 审核文档 | 张亚情 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

目录

第一章 引言 1

1.1 文档目的 1

1.2 预期读者 1

1.3 术语解释 1

1.4 参考资料 1

第二章 概述 2

2.1 系统概述 2

2.2 数据规范 3

第三章 总体结构说明 5

3.1 系统总体框架 5

3.1.1 业务架构 5

3.1.2 应用架构 5

3.1.3 数据架构 6

3.1.4 技术架构 7

3.2 系统数据视图 8

3.2.1 概念数据模型 8

3.2.1.1 智能搜索概念数据模型 9

3.2.1.2 智能提示概念数据模型 9

3.2.1.3 分类导航概念数据模型 10

3.2.1.4 知识图谱概念数据模型 10

3.2.1.5 用户日志概念数据模型 11

3.2.1.6 数据抽取概念数据模型 11

3.2.2 逻辑数据模型 12

3.2.2.1 智能搜索逻辑数据模型 12

3.2.2.2 智能提示逻辑数据模型 13

3.2.2.3 分类导航逻辑数据模型 14

3.2.2.4 知识图谱逻辑数据模型 15

3.2.2.5 用户日志逻辑数据模型 15

3.2.2.6 数据抽取逻辑数据模型 16

3.2.3 数据分类 16

3.2.4 数据流转 16

3.2.5 数据存储与分布 18

3.3 系统组件视图 19

3.3.1 系统逻辑分层 19

3.3.2 组件关联设计 23

3.3.3 组件时序设计 24

3.3.4 功能组件设计 25

3.3.5 接口组件设计 26

3.3.6 公共组件设计 27

第四章 系统界面设计 28

4.1 界面框架设计 28

4.2 系统空控件选择 29

第五章 系统功能实现说明 31

5.1 智能提示 31

5.1.1 搜索词匹配补全 31

5.1.2 搜索词语义补全 31

5.1.3 搜索词语义消歧 32

5.1.4 搜索词全拼提示 33

5.1.5 搜索词首字母提示 33

5.1.6 搜索词混合拼音提示 34

5.1.7 统计提示 34

5.2 智能搜索 35

5.2.1 搜索意图识别 37

5.2.2 搜索语义扩展 38

5.2.3 商品过滤筛选 39

5.2.4 搜索拼音及错词纠正 39

5.2.5 搜索模型 40

5.3 智能推荐 40

5.3.1 基于内容的推荐 40

5.3.2 用户行为相关推荐模型介绍 41

1. 引言
   1. 文档目的

本概要设计的依据是基于语义网的站内搜索引擎模块改造项目的可研、业务需求规格说明书和软件需求规格说明书，编写目的是为详细设计人员及开发人员和功能开发人员、测试人员等，在进行详细设计、开发、测试等过程提供指导和依据。

* 1. 预期读者
* 研发组、测试组、设计组等相关项目干系人；
* 基于语义网的站内搜索引擎模块改造项目项目相关成员。
  1. 术语解释
* 业务架构：是信息化总体架构的起点，业务架构定义了在总体企业战略指导下希望实现的业务能力及其关系。
* 主要以架构图的方式描述系统的组成和框架，一般从系统功能和系统技术层次两个架构视角进行设计。
  1. 参考资料
* 基于语义网的站内搜索系统建设方案；
* 基于语义网的站内搜索系统需求规格说明书；

1. 概述
   1. 系统概述

站内搜索引擎功能模块是电商类网站用户通往商品页面的重要渠道，模块功能是否满足用户需求，直接决定电商商品的销量，是电商平台可持续发展的关键因素。

站内搜索引擎功能模块旨在帮助用户对店铺、商品等信息进行精确检索和信息挖掘，提供全面、准确、快速的信息检索服务，为用户第一时间找到目标店铺和商品。同时，用户使用站内搜索引擎模块功能，产生大量的搜索日志和浏览历史等数据，数据真实的反馈用户的浏览习惯、行为以及偏好，是网站进行网站架构优化、提升用户体验、大数据应用建设等其他应用的重要底层数据。

基于语义网的站内搜索引擎的优化改造，一是对前端体验优化，改善前端搜索功能、提供搜索增项数据、筛选条件排序优化；二是搜索引擎模块功能的改造，在Solr的基础上对搜索引擎进行二次改造，研究利用爬虫技术，抓取采购专区商品页面目标数据，建设语义网概念的智能索引库，优化中文分词和搜索结果排序算法；三是开发数据存储功能，存储底层的用户数据，以供数据分析和其他项目应用拓展。

通过上述改造，结合爬虫技术、中文分词、反序索引库等技术的引入，基本实现前端体验合理化、系统功能智能化、数据应用简单化的总体目标。对于用户，站内搜索的优化可以改善用户搜索的准确率，快速找到用户想要的商品，避免过多操作，提升体验；对于运营，站内搜索引擎模块的价值巨大，可以了解用户的意图，改善站点的用户体验，优化网站的结构，可以收集用户的画像相关数据，作为用户浏览相关推荐的参数，另外，用户数据还可以应用在大数据和人工智能上，助力其他产品。

* 1. 数据规范

|  |  |  |
| --- | --- | --- |
| **数据类型名称** | **说明** | **默认** |
| 文本 | 单行文本，限制文本长度，默认可输入任何值 | 20字符 |
| 大文本 | 多行类型文本，限制文本长度，默认可输入任何值 | 500字符 |
| 整数 | 限制填写内容为整数值，不能填写数字以外内容，不能输入带小数点的数字 | 10位 |
| 双精度书 | 限制填写内容为带小数点的数字，默认为小数点后两位，不能填写数字以外内容，如输入数字小数点为3位或以上或整数时，页面显示两位小数点。（默认小数点3位以上时，采用四舍五入的方法） | 12位 |
| 关联 | 指关联其它模块中字段的。比如企业账户和企业角色两个业务模块，在创建企业账户时需要选择该账户具有那个企业角色中的权限。 | 无限制 |
| 下拉框 | 指多个数据项，用户可点击出现下拉列表框，默认为单选 |  |
| 日期 | 指默认该字段呈现为yyyy-mm-dd。如2016-09-09 |  |
| 时间 | 指默认该字段呈现为hh:mm:ss。如18:00:00 |  |
| 短时间 | 指默认该字段呈现为hh:mm。如18:00 |  |
| 长时间 | 指默认该字段呈现为yyyy-mm-dd hh:mm:ss。如  2016-09-09 18:00:00 |  |
| 布尔 | 只能为二选一，允许该字段是空值，选择后则只能二选一，如性别选择为男或女 |  |
| 金额 | 指只能输入整数或带小数点的数字，其他内容无法填写，默认为小数点后两位，超过两位或输入整数时，显示两位小数点，显示方法为：10，000.00（默认小数点3位以上时，采用四舍五入的方法展现） | 12位 |
| 附件 | 指可上传任意类型的文件，不限制上传文件的格式和上传文件的大小，默认只可上传一个文件 | 一个文件 |
| 图片 | 指只可上传图片格式类型的文件，具体可根据系统设置限制上传的图片格式和上传图片的大小，默认只可上传一个图片 | 一个图片 |
| 链接 | 指输入的网址，在查看时可以通过系统自带的浏览器打开 | 50字符 |
| 地址 | 指由选择省、市、区三级联动的下拉框（可数据字典维护），详细地址（文本框，限制100字符）组成的组件 |  |
| 固话 | 指由区号（文本框）、电话号码（文本框）、分机号（文本框）三个文本框组成的字段，系统检测只能输入正确区号（最多输入4位正整数），电话号码格式的整数（最多输入8位正整数），分机号只检测为整数即可（最多输入5位整数），该字段在查看时显示为：010-1234567-801 |  |

注：本文档中，所有数据显示，如超过数据列表的字符，自动截断，鼠标悬浮时，显示全部

1. 总体结构说明
   1. 系统总体框架
      1. 业务架构

基于语义网的站内搜索引擎模块改造项目系统在业务架构设计方面，强调以数据驱动为前提，以业务支持服务为目的，通过建设智能化电商搜索引擎，提升用户搜索体验，促进商城效益。如下图所示，业务框架主要包括资源抽取、资源加工、数据管理以及智能应用四大板块。

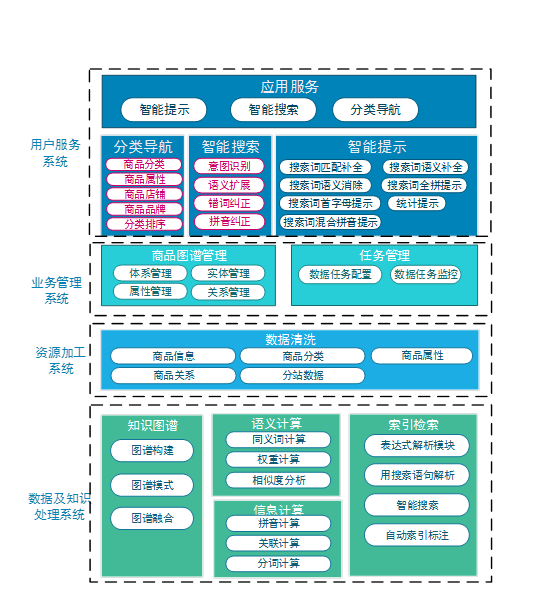


* + 1. 应用架构

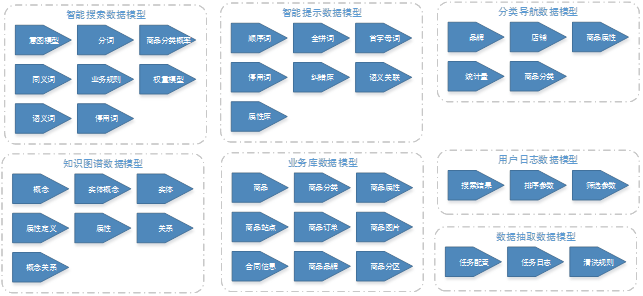
基于语义网的站内搜索引擎系统主要包含应用服务、业务管理、资源加工、数据及知识处理以及分析管理5个主要的业务应用。

用户服务主为内外部用户提供各种业务的入口和业务服务访问；业务管理主要为业务管理人员管理平台的各种支撑性服务；资源加工主要为商品及相关数据资源转化到知识图谱及索引库服务，语义计算、信息计算主要负责资源数据抽取后，清洗和计算出理等基础技术。

基于语义网的站内搜索引擎系统的应用架构如下图：



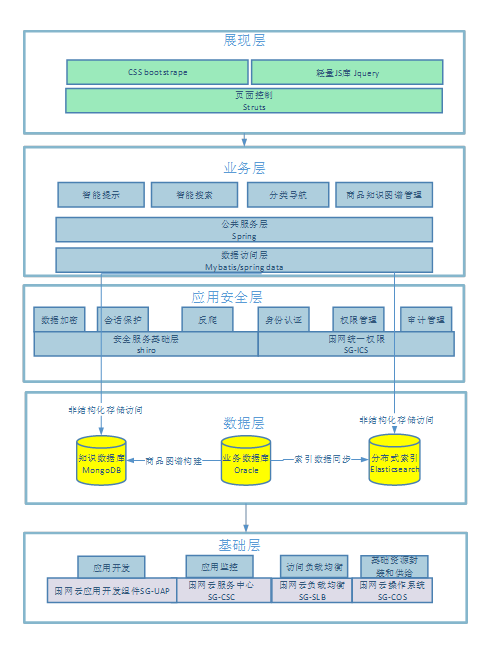
* + 1. 数据架构

如下图所示，基于语义网的站内搜索引擎系统主要设计数据模型包括智能搜索数据模型、智能提示数据模型、分类导航数据模型、知识图谱数据模型、业务库数据模型和以及用户日志数据模型。

* + 1. 技术架构

技术架构实现上，遵循SG-EA技术架构设计规范，采用组件化、动态化的软件技术，利用一致的可共享的数据模型，按照展现层、业务逻辑层、安全层、数据服务层、基础架构服务层进行划分，实现多层技术体系设计。通过集成平台进行界面，数据和服务集成，实现科技资源数据的各接口组件能够为内外部用户提供多层次上集成，实现重用，以满足公司对知识访问的服务需求，为各单位提供高效便捷的知识访问支撑。

技术架构如下图：



* 1. 系统数据视图
     1. 概念数据模型

基于语义网的站内搜索引擎系统概念数据模型主要包括智能搜索概念数据模型、智能提示概念数据模型、分类导航概念数据模型、知识图谱概念数据模型、用户日志概念数据模型以及数据抽取概念数据模型。

* + - 1. 智能搜索概念数据模型

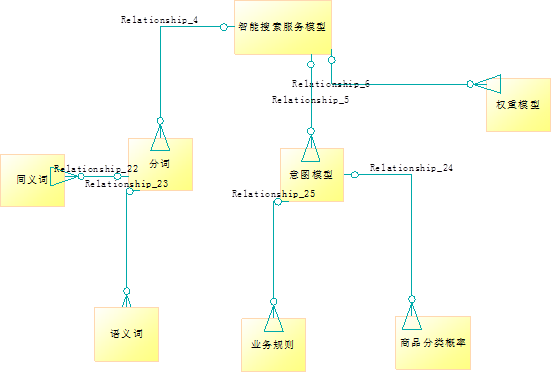


图 概念数据模型

* + - 1. 智能提示概念数据模型

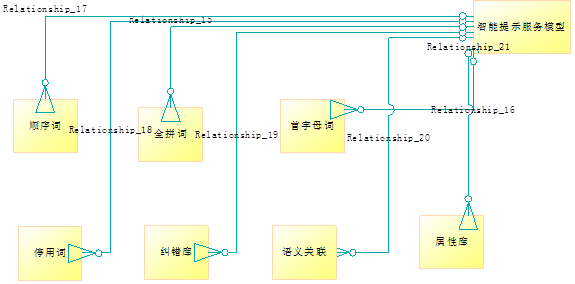


图 概念数据模型

* + - 1. 分类导航概念数据模型

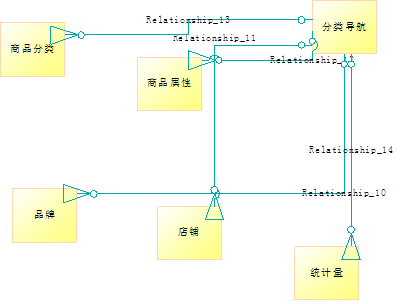


图 概念数据模型

* + - 1. 知识图谱概念数据模型

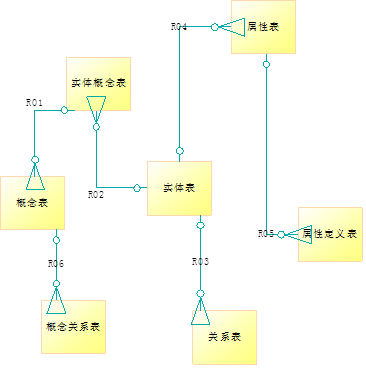


图 概念数据模型

* + - 1. 用户日志概念数据模型

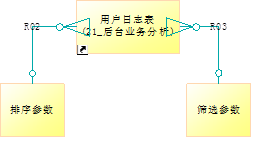


图 概念数据模型

* + - 1. 数据抽取概念数据模型

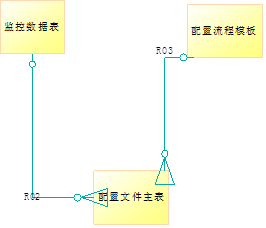


图 概念数据模型

* + 1. 逻辑数据模型
       1. 智能搜索逻辑数据模型

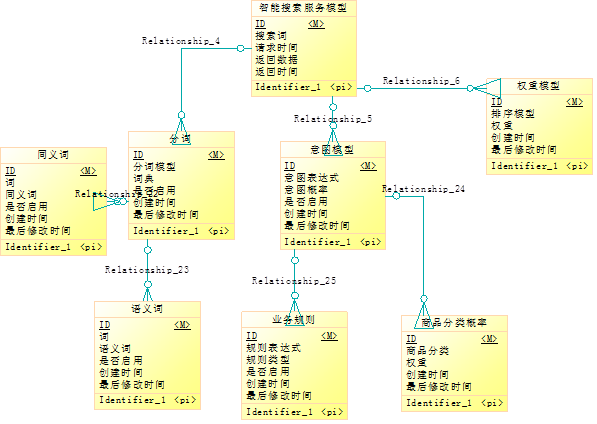


图 概念数据模型

* + - 1. 智能提示逻辑数据模型

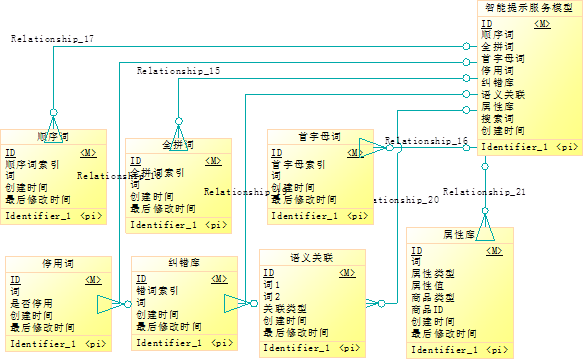


图 概念数据模型

* + - 1. 分类导航逻辑数据模型

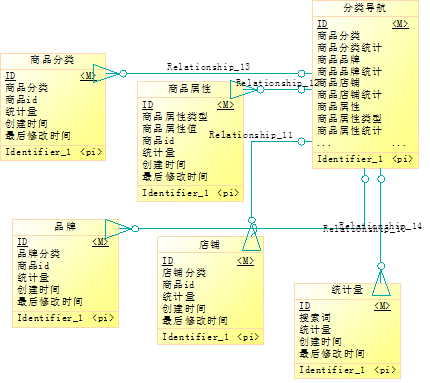


图 概念数据模型

* + - 1. 知识图谱逻辑数据模型

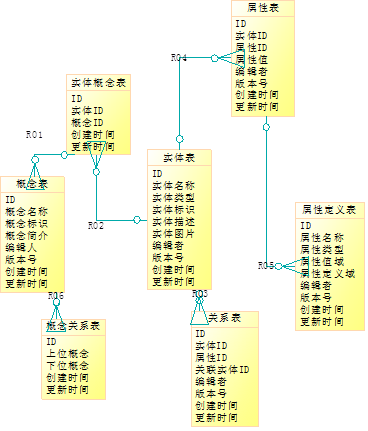


图 概念数据模型

* + - 1. 用户日志逻辑数据模型

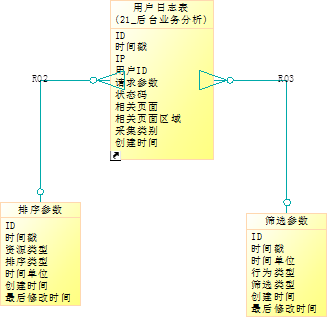


图 概念数据模型

* + - 1. 数据抽取逻辑数据模型

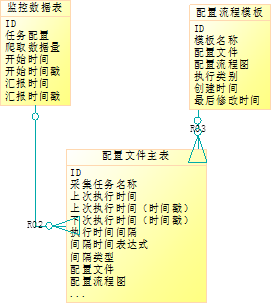


图 概念数据模型

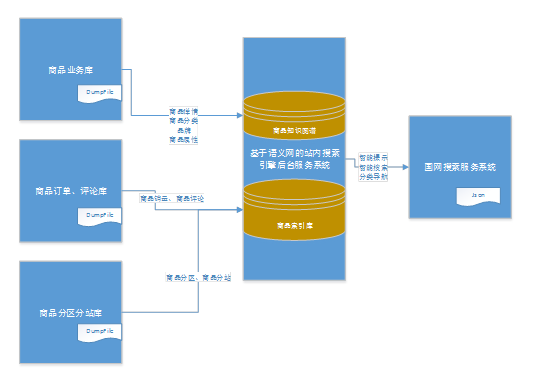
* + 1. 数据分类

数据分类清单表

|  |  |
| --- | --- |
| 数据实体 | 数据描述 |
| 智能搜索数据 | 结构化 |
| 智能提示数据 | 结构化 |
| 分类导航数据 | 结构化 |
| 知识图谱数据 | 结构化 |
| 业务库数据 | 结构化 |
| 用户日志数据 | 结构化 |

* + 1. 数据流转

如图所示，数据流转主要由商品业务库、商品订单评论库以及商品分区分站库经ETL流向基于语义网的站内搜索引擎后台系统，最后由国网搜索服务系统调用消费。



图数据流转图

数据流转清单表

|  |  |  |
| --- | --- | --- |
| **交换数据实体** | **源系统** | **目标系统** |
| 商品详情 | 商品业务库 | 基于语义网站内搜索后台服务系统 |
| 商品分类 | 商品业务库 | 基于语义网站内搜索后台服务系统 |
| 品牌 | 商品业务库 | 基于语义网站内搜索后台服务系统 |
| 商品属性 | 商品业务库 | 基于语义网站内搜索后台服务系统 |
| 商品销量 | 商品订单库 | 基于语义网站内搜索后台服务系统 |
| 商品评论 | 商品评论库 | 基于语义网站内搜索后台服务系统 |
| 商品分区 | 商品分区库 | 基于语义网站内搜索后台服务系统 |
| 商品分站 | 商品分站库 | 基于语义网站内搜索后台服务系统 |
| 智能提示结果 | 基于语义网站内搜索后台服务系统 | 国网搜索服务系统 |
| 智能搜索结果 | 基于语义网站内搜索后台服务系统 | 国网搜索服务系统 |
| 分类导航结果 | 基于语义网站内搜索后台服务系统 | 国网搜索服务系统 |

* + 1. 数据存储与分布

数据存储清单表

|  |  |
| --- | --- |
| 数据实体 | 存储系统名称 |
| 商品业务库 | 基于语义网站内搜索引擎后台服务系统 |
| 商品订单、评论库 | 基于语义网站内搜索引擎后台服务系统 |
| 商品分区、分站库 | 基于语义网站内搜索引擎后台服务系统 |
| 商品知识图谱 | 基于语义网站内搜索引擎后台服务系统 |
| 商品索引数据 | 基于语义网站内搜索引擎后台服务系统 |
| 同义语义词 | 基于语义网站内搜索引擎后台服务系统 |
| 用户日志 | 基于语义网站内搜索引擎后台服务系统 |
| 智能提示结果 | 基于语义网站内搜索引擎后台服务系统 |
| 智能搜索结果 | 基于语义网站内搜索引擎后台服务系统 |
| 分类导航结果 | 基于语义网站内搜索引擎后台服务系统 |

数据分布清单表

|  |  |  |  |
| --- | --- | --- | --- |
| 数据实体 | 站内搜索原系统 | 基于语义网站内搜索引擎后台服务系统 | 站内搜索可视化系统 |
| 商品业务库 | O | C |  |
| 商品订单、评论库 | O | C |  |
| 商品分区、分站库 | O | C |  |
| 商品知识图谱 |  | O |  |
| 商品索引数据 |  | O |  |
| 同义语义词 |  | O |  |
| 用户日志 |  | O |  |
| 智能提示结果 |  | O | C |
| 智能搜索结果 |  | O | C |
| 分类导航结果 |  | O | C |

说明：O-owner即所有者，C-copy即复制者

* 1. 系统组件视图
     1. 系统逻辑分层

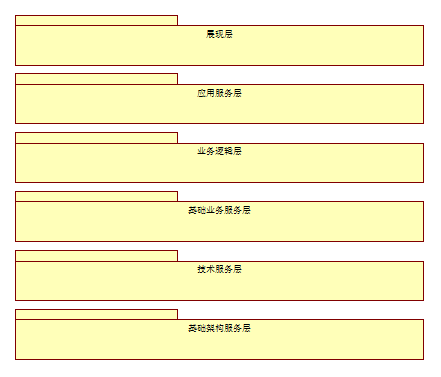
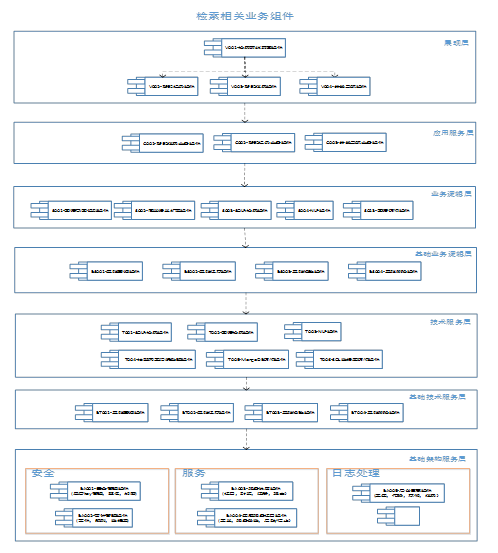


图 系统逻辑分层图

系统逻辑分层说明表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **逻辑层次** | **职责描述** | **技术实现** | **逻辑层次依赖** | **层间通信** |
| 展现层 | * 显示用户界面 * 接受用户输入 * 数据验证 * 呈现数据 * 向应用服务层发送请求 | * 使用JSP技术 * 使用JQUERY技术 * 使用AJAX技术使用 * 使用HTML5/DHTML技术 * 使用XML、JSON技术 * 使用Flex技术 * 使用WEBSOCKET技术 * 使用EASYUI技术 * 使用BOOTSTRAP技术 * 使用JQUERY.NICESCROLL.MIN.JS技术 * 使用ECHARTS3.0技术 * 使用JQGRID技术 * 使用JQUERY-SWIPE技术   使用JQUERY JBOX技术   * 使用MY97DATEPICKER技术 * 使用JQUERY SELECT2技术 * 使用JQUERY ZTREE技术 * 使用JQUERY.UPLOADIFY技术 * 使用JQUERYVALIDATION PLUGIN 1.11技术 | * 依赖应用服务层 | * 采用XML、JSON等数据载体 * 采用HTTP(S)协议进行通信 |
| 应用  服务层 | * 提供完整核心应用的服务 * 接收并处理展现层的数据请求 * 分析用户输入的请求数据 * 向业务服务层发起请求 | * 使用JAVA开发技术 * 使用SPRING进行IOC控制 * 使用SPRING MVC技术 | * 依赖业务逻辑层 | * 采用JAVA调用 * 采用共享缓存数据交换 * 采用HTTP(S)协议进行通信 |
| 业务  逻辑层 | * 接收应用服务层的请求 * 提供基础业务服务 | * 使用SPRING技术 * 使用LOG4J技术 * 使用JAVA开发技术 | * 依赖基础业务服务层 | * 采用JAVA调用 * 采用共享缓存数据交换 * 采用HTTP(S)协议进行通信 |
| 基础业务  逻辑层 | * 封装上层的公共基础业务逻辑处理服务 * 向技术服务层发起请求 * 模块服务调用 | * 使用SPRING技术 * 使用LOG4J技术 * 使用JAVA开发技术 | * 依赖技术服务层 | * 采用JAVA调用 * 采用共享缓存数据交换 * 采用HTTP(S)协议进行通信 |
| 技术服务层 | * 数据持久化操作 * 系统监控服务 * 基本接口服务 | * 使用SPRING技术 * 使用LOG4J技术 * 使用JAVA开发技术 | * 依赖基础技术服务层 | * 采用JAVA调用 * 采用共享缓存数据交换 * 采用HTTP(S)协议进行通信 |
| 基础技术  服务层 | * 封装上层服务使用的公共技术服务功能 * 数据持久化操作 * 数据缓存操作 | * 使用SPRING技术 * 使用STRUTS2技术 * 使用XFIRE技术 * 使用LOG4J技术 * 使用SPRINGDATA技术 * 使用MYBATIES/JDBC技术 * 使用MONGODB访问API技术国网大数据使用 * 使用Elasticsearch访问API技术 | * 依赖基础框架服务层 | * 采用Java调用 * 采用共享缓存数据交换 * 采用http(s)协议进行通信 |
| 基础架构  服务层 | * 支持系统整体基础框架的服务 * 服务器分发，请求服务级别分配 * 应用请求反向代理服务 * 静态元素请求响应服务 | * 使用nginx技术 |  |  |

* + 1. 组件关联设计



组件关联图

组件关联清单表

|  |  |  |  |
| --- | --- | --- | --- |
| 组件编号-名称 | 组件的方法 | 关联的组件编号-名称 | 关联的组件的方法 |
| V002-智能提示组件 | 搜索关联词 | C002-智能提示业务组件 | 提示关联词 |
| V003-智能搜索组件 | 搜索商品、店铺 | C001-智能搜索业务组件 | 搜索资源 |
| V004-分类导航组件 | 加载品类数据 | C003-分类导航业务组件 | 加载品类数据 |
| C002-智能提示业务组件 | 提示关联词 | S003-Solr检索组件 | 查询Solr |
| C001-智能搜索业务组件 | 搜索商品、店铺 | S001-图谱意图识别组件 | 意图识别 |
| S002-同义语义扩展组件 | 关键词语义扩展 |
| S004-NLP组件 | 分词、文本抽取、语义计算等 |
| S003-ES检索组件 | 查询Solr |
| C003-分类导航业务组件 | 加载品类数据 | S005-图谱访问组件 | 查看图谱 |

* + 1. 组件时序设计

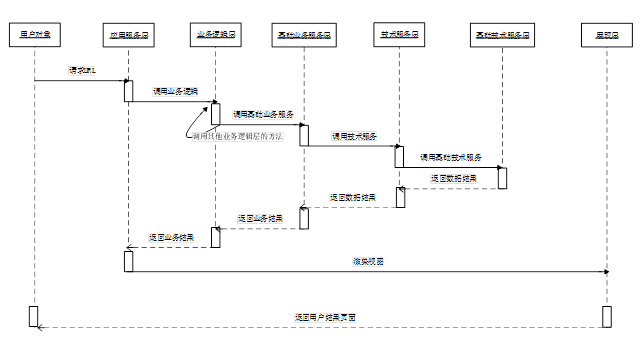


图 组件时序设计图

组件时序设计图的执行步骤如下：

1、用户通过浏览器向服务器发出请求

2、服务器将请求转发给应用服务层，服务层根据客户的请求方法名和参数，调用相应的业务逻辑层

3、业务逻辑层根据传过来的参数进行处理，调用相应的技术服务层方法

4、技术服务层返回给业务逻辑层处理结果

5、业务逻辑层将最终的处理结果返回给应用服务层

6、应用服务层将业务逻辑层返回的结果返回给展示层

7、展示层通过渲染视图将结果返回给用户

8、用户看到想要的结果页面

* + 1. 功能组件设计

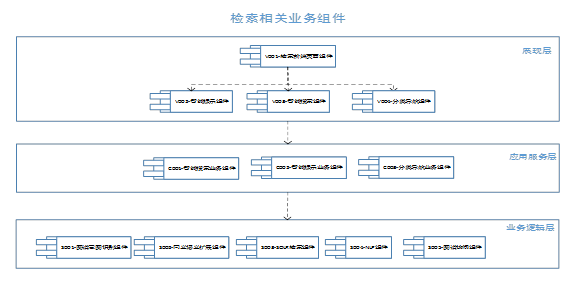


图 功能组件图

功能组件清单表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **功能组件编号-名称** | **功能编号-名称** | **系统逻辑分层** | **包含的组件编号-名称** | **是否接口组件** | **非功能性要求** |
| FUN15-信息检索管理 | FUN1501-信息检索管理 | 展现层 | V002-智能提示组件 | 否 | 易操作性 |
| V003-智能搜索组件 | 否 |
| V004-分类导航组件 | 否 |
| 应用服务层 | C002-智能提示业务组件 | 是 | 容错性  并发性  稳定性 |
| C001-智能搜索业务组件 | 是 |
| C003-分类导航业务组件 | 是 |
| 业务逻辑层 | S001-图谱意图识别组件 | 否 | 无 |
| S003-Solr检索组件 | 否 | 无 |
| S004-NLP组件 | 否 | 无 |
| S005-图谱访问组件 | 否 | 无 |

* + 1. 接口组件设计



图 接口组件图

接口组件清单

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 分类 | 接口组件编号-名称 | 职责描述 | 来源 | 系统逻辑分层 |
| 本系统接口组件 | C001-智能搜索业务组件 | 与知识图谱和Solr交互，提供商品、店铺搜索等服务 | 外部系统接口 | 技术服务层 |
| C002-智能提示业务组件 | 与知识图谱和Solr交互，提供用户搜索关键词联想词等服务 | 外部系统接口 | 技术服务层 |
| C003-分类导航业务组件 | 提供商品过滤筛选 | 外部系统接口 | 技术服务层 |

* + 1. 公共组件设计

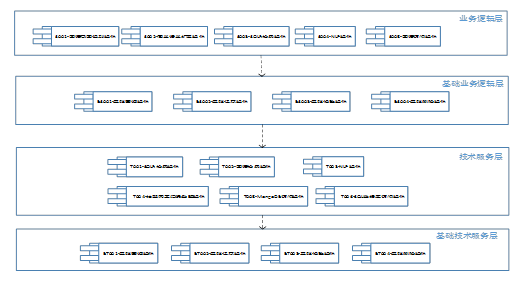


图 公共组件图

公共组件清单

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 分类 | 公共组件编号-名称 | 职责描述 | 来源 | 系统逻辑分层 |
| 本系统重用功能组件 | S001-图谱意图识别组件 | 负责用户搜索意图识别 | 外部系统接口 | 业务逻辑层 |
| S002-同义语义扩展组件 | 负责关键词同义词扩展 | 外部系统接口 | 业务逻辑层 |
| S003-SOLR检索组件 | 负责资源检索 | 外部系统接口 | 业务逻辑层 |
| S004-NLP组件 | 负责文本抽取、分词等NLP功能 | 外部系统接口 | 业务逻辑层 |
| S005-图谱访问组件 | 负责访问知识图谱 | 外部系统接口 | 业务逻辑层 |

1. 系统界面设计
   1. 界面框架设计

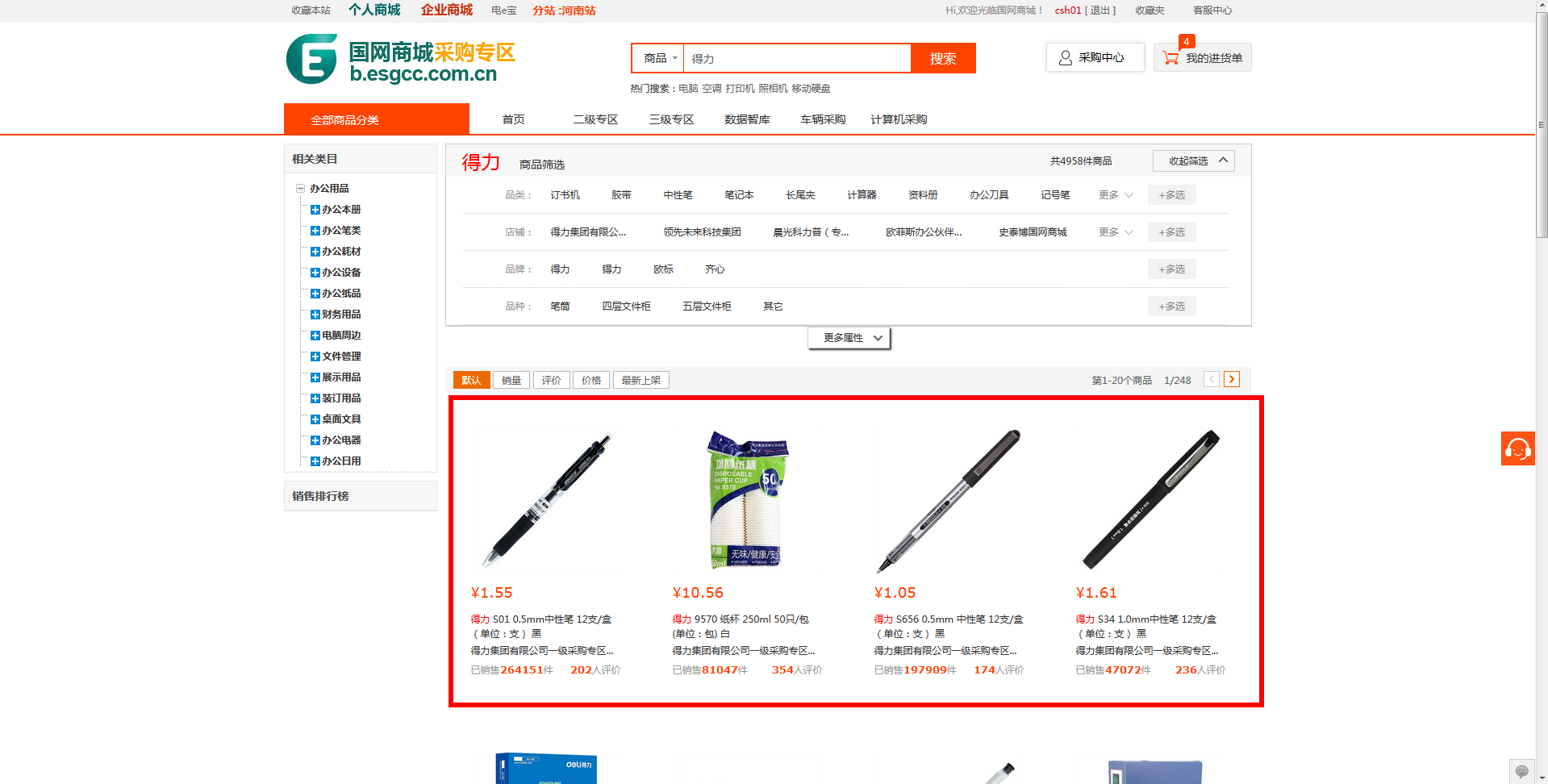
系统登录页设计：



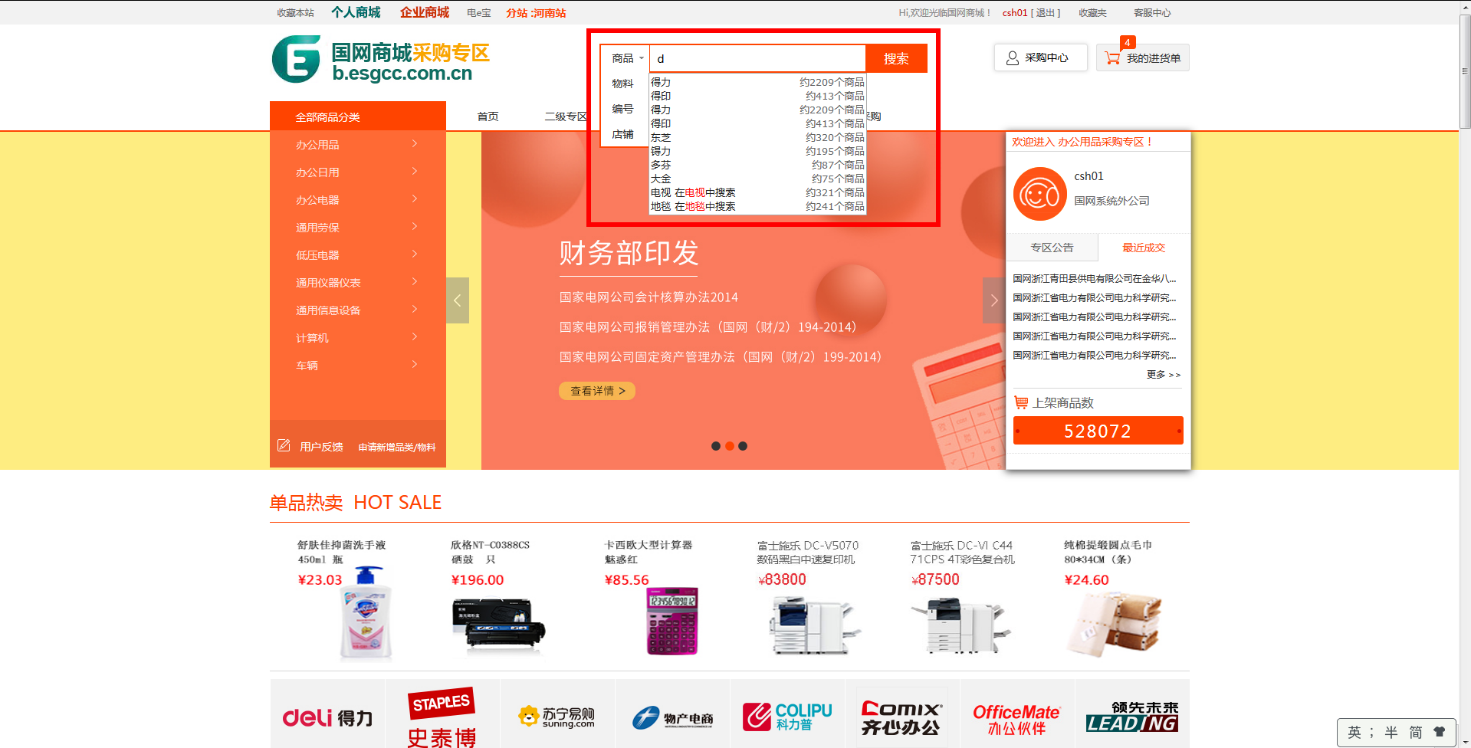
采购专区首页：



表格控件：



下拉控件：



* 1. 系统空控件选择

系统控件清单

|  |  |  |  |
| --- | --- | --- | --- |
| **业务场景** | **功能页面名称** | **选择控件** | **控件用途** |
| 商品/店铺搜索 | 首页及搜索列表页 | 下拉控件 | 通过下拉控件，以列表的形式展示商品提示、商品数量、意图识别提示。 |
| 表格控件 | 通过表格控件，以列表形式展示单据序号、单据状态、单据号单据类型、提交时间、单据金额、单据事由。其中单据状态、单据类型支持筛选，提交时间、单据金额支持排序功能。 |

1. 系统功能实现说明
   1. 智能提示

智能提示框广泛用于各大电商平台，对用户的输入或正在进行的输入，给予补全、较正或导向型的引导。帮助用户更加精确的描述搜索目标，提升用户体验，提高留存及转化。

* + 1. 搜索词匹配补全

搜索词匹配补全，是以前缀的方式，针对检索词进行补全。

本次poc中，输入 “笔”、“笔记”或“笔记本”都能提示出“笔记本”。

功能参见下图：

图 搜索词匹配补全

* + 1. 搜索词语义补全

搜索词语义补全，是在4.3.2的基础上，对搜索目标的关键信息予以提示的方式，对检索词进行补全。比如：商品的笔记本电脑，包含属性品牌、尺寸、屏幕、内存、硬盘等等。通过商品知识图谱，能够组合出很多的候选提示词如：

笔记本联想、笔记本苹果、笔记本14寸、笔记本15寸

可以根据属性的权重及运营用户数据的辅助，将语义相关且搜索频度高的词作为下拉提示候选集。

功能参见下图：

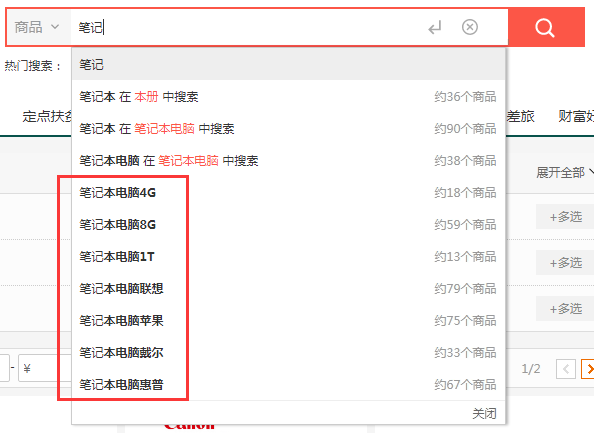


图 搜索词语义补全

* + 1. 搜索词语义消歧

搜索词的语义消歧，通常用于在搜索之前与用户的显示交互上，明确搜索目标。汉语中普遍存在歧义问题。比如：

搜索笔记本，到底是搜的电脑？还是办公用品？

下拉提示的关键字语义消歧，将知识图谱的同义关系及所属概念（类目）加入到待提示候选集。

功能参见下图：

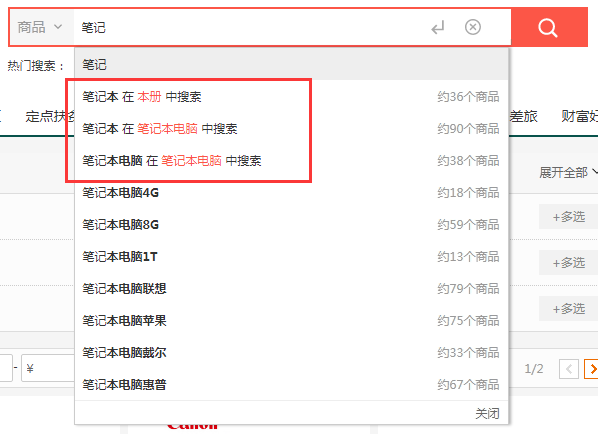


图 搜索词语义消歧

* + 1. 搜索词全拼提示

搜索词全拼提示，能够加快用户的输入。输入“bijiben”，能够识别出“笔记本”及其以“笔记本”开头的多种提示。

功能参见下图：

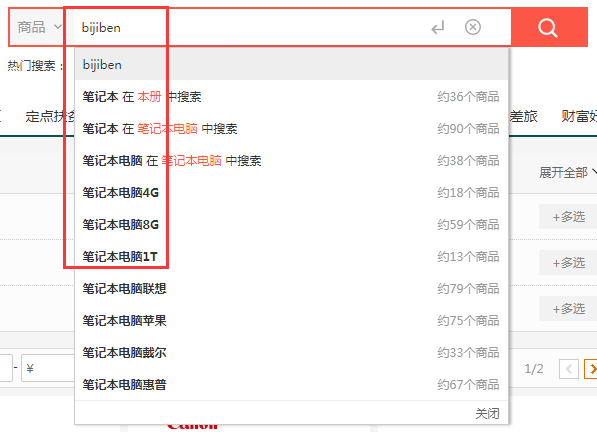


图 搜索词匹配补全

* + 1. 搜索词首字母提示

搜索词首字母提示，能够加快用户的输入。输入“bjb”，能够识别出“笔记本”及其以“笔记本”开头的多种提示。

功能参见下图：

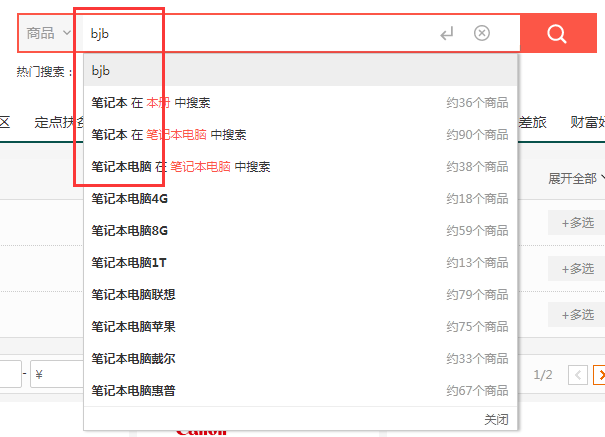


图 搜索词首字母提示

* + 1. 搜索词混合拼音提示

搜索词首字母提示，能够加快用户的输入。输入“b”、“bj”、“bij”、“bijb”、“bijib”等，都能够识别出“笔记本”及其以“笔记本”开头的多种提示。

功能参见下图：

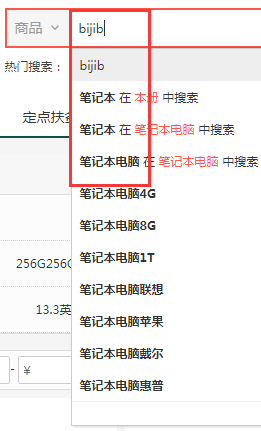
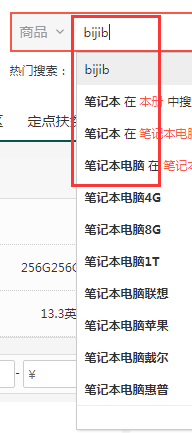
  

图 搜索词首字母提示

* + 1. 统计提示

搜索词统计提示，能够直观的看到，能够匹配搜索词的商品结果数目。

功能参见下图：

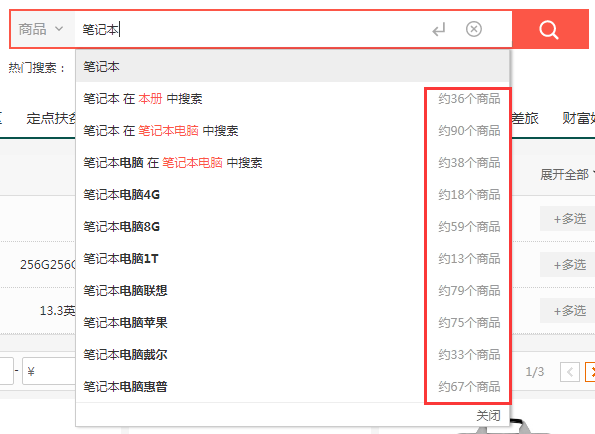


图 统计提示

* 1. 智能搜索

信息爆炸的时代大背景下，需要比以为更精准、智能的搜索。目前各大电商平台、搜索引擎（百度、谷歌、搜狗等），都将知识图谱应用到自己各个产品中，不断加强用户体验，提升产品价值。知识图谱最为突出的一个应用点，则是智能搜索。

以往以纯文本匹配的检索方式，会对搜索结果往往产生较大的不相关数据。举例来说：搜索“苹果”时，系统不知道用户找水果还是电脑，也不知道用户的倾向性偏好。

有别于文本检索，智能检索以识别用户意图的情况下进行检索，有以下优势：

1）能够解决歧义的问题。比如：用户最近搜了很多次电子产品，那搜“苹果”，大概率就买电脑。水果作为一个候选类目，可以去询问用户是否要找水果。

2）能够更加了解用户。通过对用户的建模（用户性别、年龄、消费层次、质量偏好等）能够结合搜索词展示最合适的商品。

3）能够将用户输入的错词进行智能纠正。



图 智能搜索执行流程图

上图 是本次POC的一个搜索流程图，其中分为三条路径：

1）意图识别命中

意图识别命中，智能搜索模块已经知道用户的意图。

此时，对搜索词进行语义相似词的扩展，这里的语义包含同义词和语义词。

接着，通过知识图谱，查询搜索目标所属于的概念有哪些属性。一方面做搜索属性域的扩展，另一方面生成过滤筛选项。搜索属性域扩展后，根据每个域的权重，选择部分属性域加入搜索。

然后，对生成查询条件集合的条件权重进行动态计算。

最后，执行搜索，并生成用户搜索的建议。

2）意图识别没命中、首次搜索有结果

意图识别未命中，退化成同义扩展后的关键词检索。

同义扩展的目标提升搜索的召回率，同义扩展后，对扩展词集合进行权重动态计算。

最后，并按搜索的结果属性聚合生成过滤筛选项，最后返回结果。

3）意图识别没命中、首次搜索无结果

意图识别没有命中且首次搜索也没有结果，有可能是搜索词本身就错了，尝试关键词纠错重搜。

纠错的主要方式为将搜索词转化成拼音，然后寻找拼音匹配下概率最大的。

于此同时，也可根据词的利文斯顿距离进行相似计算，根据得分和阈值来确定是否进行纠错。

通过拼音及错词纠错，重复情况2的步骤。

如果有数据返回，并告诉用户建议纠错的搜索关键词。如果没有数据，提示没有检索到数据。

* + 1. 搜索意图识别

意图识别是根据用户输入的文本词，识别用户真正想搜索的目标。

如果存在用户模型，可实现千人千面的搜索。举例来说：搜索“洗衣机”，如果用户为低消费群体，搜索结果侧重商品性价比；如果为高消费群体，搜索结果侧重展示质量、品牌、知名度。

在意图识别的过程中，知识图谱的语义结构数据至关重要。

下面以本次POC搜索的场景进行简要说明：

1）下图圆的节点表示一个小知识图谱，其中包含笔记本电脑、本册两个品类；两个品类下分别有联想、苹果；得力、晨光两个品牌。

笔记本电脑 -> 联想、苹果

本册 -> 得力、晨光

2）用户搜索“笔记本联想”，经过分词并识别后，发现用户更倾向于“笔记本电脑类目下品牌名为联想的笔记本”。此后，根据此意图信息进行加权检索，结果基本是关于笔记本电脑的，联想的排名靠前。



图4.4.1-1 意图识别过程示例

3）与此同时，用户也会疑惑，为什么没有出现本册。系统会构建用户搜索建议：“帮您找到笔记本电脑类目下的商品， 如果您想找本册，请点击。”





图 商城意图识别 用户搜索建议 示例

* + 1. 搜索语义扩展

语义扩展在搜索的场景中，一般以在不降低准确度的情况下，提高召回率（Recall）。语义扩展同时包含了语义与同义的扩展。

对于前者，也就是知识图谱里的关联关系。比如“笔记本”、“高配”、“轻薄”是语义相关的。

对于后者，也就知识图谱里的同义词。比如：“苹果”、“Apple”、“Mac”。

对于本次POC的场景，利用知识图谱的语义及同义关系进行搜索扩展和加权。在搜索“联想”时，能够自动的扩展“Lenovo”、“ThinkPad”等。

* + 1. 商品过滤筛选

过滤筛选的主要目标是让用户能够快速的筛选搜索返回结果。

在用户查询意图被正确识别时，通过查询知识图谱中属性域，得到商品过滤筛选项。

在用户查询意图未被正确识别时，通过返回结果数据指定属性的聚合，得到商品过滤筛选项。

对于知识图谱属性产生的商品过滤筛选项，由于是预先定义的语义结构数据，能够更加准确的控制数据质量；而根据聚合产生的商品过滤筛选项，由于数据的不可控性。相对前者，有可能出现数据的噪声。

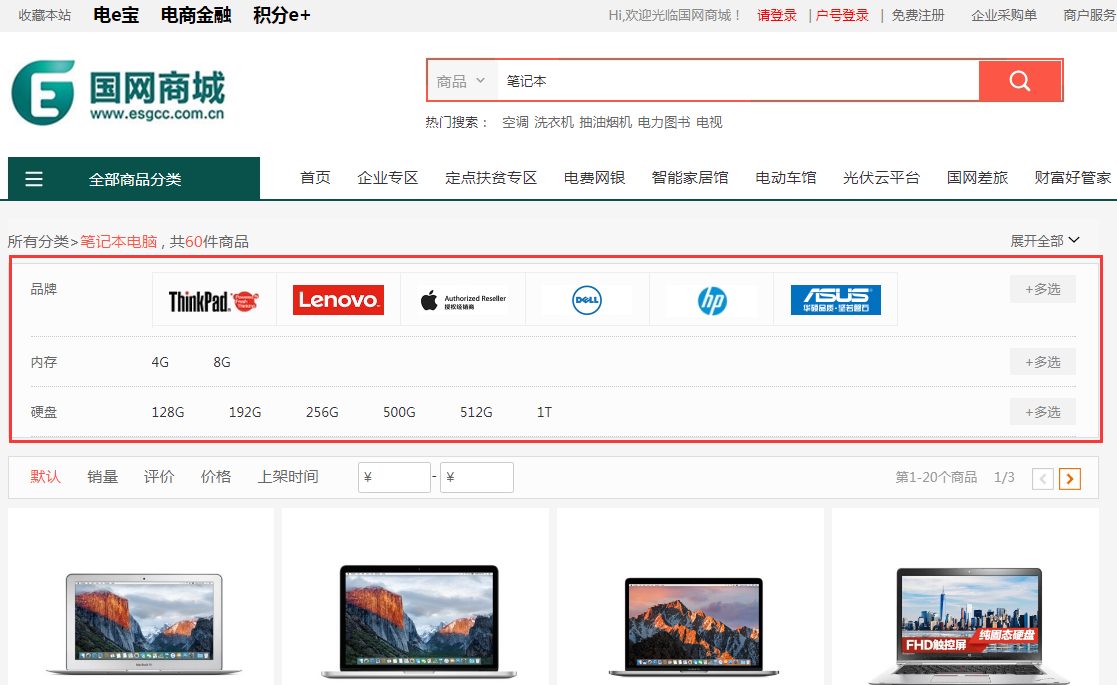


图 商品过滤筛选示例

* + 1. 搜索拼音及错词纠正

拼音及错词纠正解决用户输出失误的问题，提升用户体验，引导用户搜索，防止用户流失。

错词一般将词的谐音转化成拼音，再通过拼音去匹配命中且平台最希望用户搜索的词。

如图4.4.4-1，搜索拼音“bijiben”，能自动识别到搜索目标“笔记本”。

如图4.4.4-2，搜索错词“比集奔”，能自动识别到搜索目标“笔记本”。





图4.4.4-1 搜索拼音纠正示例





图4.4.4-2 搜索错词纠正示例

* + 1. 搜索模型

搜索模型的主要目的是解决搜索结果的排序问题。有很多种方式进行处理，比如：

1）利用意图识别的结果加权

2）利用同义词进行加权

3）利用搜索词的长度、位置信息进行加权

4）利用用户搜索日志的反馈信息进行加权

5）利用业务规则进行加权

本次poc主要使用1、2、3的加权方式，举例来说。搜索“笔记本联想”，

意图识别到搜索的是笔记本类目，那么此类目打分调高；识别到笔记本类目的品牌联想，此加权调高。

“联想”的同义词“ThinkPad”、“Lenovo”进行加权等。

* 1. 智能推荐
     1. 基于内容的推荐

早期的电商平台推荐系统往往会面临冷启动的问题（Cold Start，因数据缺失导致用户行为挖掘无法进行），此时结合业务进行内容的推荐将会有比较好的效果。

基于内容的推荐，通常电商平台的运营策略，一般使用方法为：

推荐用户同品类下，价格同一区间的商品。价格区间的计算方式，采用同品类商品拟合正态分布，去除最大最小的5%区间，然后按均匀分布切割价格区间，来进行推荐。

推荐用户热销商品。

根据业务，按条件加权推荐。比如，商城更希望某一商家的商品热卖，商城更希望某品牌单品热卖。

根据业务，设置规则。比如，买打印机的同时推荐打印纸等。

随机规则，部分电商（如：亚马逊）的“给我惊喜（Superise me）”功能，就是推荐用户意想不到的商品。此类推荐有时能打破常规，推荐用户自己都想不到的商品。

本次POC中混合使用方式1、2、5进行推荐，功能参见下图：



图4.5.1 基于内容的推荐示例

* + 1. 用户行为相关推荐模型介绍

由于本次poc未涉及用户行为数据，下面针对商城业务可以用到的推荐模型进行说明。

1）基于用户兴趣模型推荐

用户兴趣模型推荐，要从用户的行为和偏好中发现规律，并基于此给予推荐。如何收集用户的偏好信息成为系统推荐效果最基础的决定因素。用户有很多方式向系统提供自己的偏好信息，而且不同的应用也可能大不相同，详情可参看图4.4.2。用户兴趣模型建立后，推荐引擎设计人员可以根据自己应用的特点添加特殊的用户行为，并用他们表示用户对物品的喜好。



图4.5.2-1 用户兴趣模型示例

2）使用基于商品的协同过滤推荐

基于商品的协同顾虑推荐（CF）的原理是在计算邻居时采用物品本身，而不是从用户的角度，即基于用户对物品的偏好找到相似的物品，然后根据用户的历史偏好，推荐相似的物品给他。从计算的角度看，就是将所有用户对某个物品的偏好作为一个向量来计算物品之间的相似度，得到物品的相似物品后，根据用户历史的偏好预测当前用户还没有表示偏好的物品，计算得到一个排序的物品列表作为推荐。

下图给出了一个例子，对于物品 A，根据所有用户的历史偏好，喜欢物品 A 的用户都喜欢物品 C，得出物品 A 和物品 C 比较相似，而用户 C 喜欢物品 A，那么可以推断出用户 C 可能也喜欢物品 C。

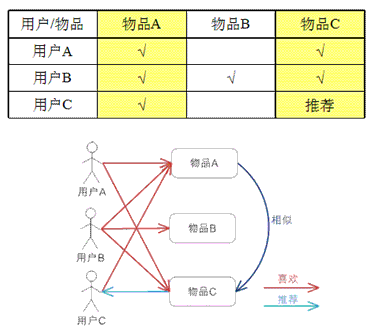


图 基于商品协同推荐示例图

3）基于商品的协同过滤推荐

基于用户的协同过滤推荐（CF）的基本思想为，基于用户对物品的偏好找到相邻邻居用户，然后将邻居用户喜欢的推荐给当前用户。计算上，就是将一个用户对所有物品的偏好作为一个向量来计算用户之间的相似度，找到 K 邻居后，根据邻居的相似度权重以及他们对物品的偏好，预测当前用户没有偏好的未涉及物品，计算得到一个排序的物品列表作为推荐。

下图给出了一个例子，对于用户 A，根据用户的历史偏好，这里只计算得到一个邻居 - 用户 C，然后将用户 C 喜欢的物品 D 推荐给用户 A。

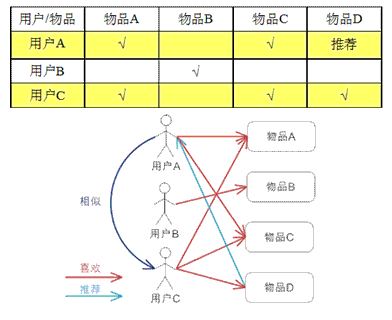


图 基于用户协同推荐示例图