基于知识图谱的出版物检索和推荐系统

设计说明书

项目成员：

马健

殷晓婷

雷一明

刘子晗

目录

[1.引言 3](#_Toc517252318)

[2.系统设计概述 3](#_Toc517252319)

[2.1系统总体设计 3](#_Toc517252320)

[2.2 DBpedia介绍 3](#_Toc517252321)

[2.3 echarts介绍 3](#_Toc517252322)

[3系统详细设计 4](#_Toc517252323)

[3.1 系统框架搭建 4](#_Toc517252324)

[3.2 系统数据处理 4](#_Toc517252325)

[3.3 结果可视化 5](#_Toc517252326)

## 1.引言

随着信息技术的发展，每年都有很多的出版物发行，包括电影、游戏以及图书等多个种类，如何找到用户感兴趣的出版物这一问题尚未得到十分有效的解决。基于知识图谱的出版物检索和推荐系统旨在基于链接开放数据库和知识图谱的相关内容，为用户提供出版物的检索、推荐以及检索结果可视化统计等功能。

## 2.系统设计概述

### 2.1系统总体设计

系统总体包括系统检索、推荐以及结果统计可视化几个功能，项目整体是一个基于javascript的网页，系统检索和推荐需要基于dbpedia数据库，通过sparql的查询语句完成相应功能，系统的结果可视化通过echarts插件完成。

### 2.2 DBpedia介绍

DBpedia 是一个特殊的语义网应用范例，它从维基百科(Wikipedia)的词条里撷取出结构化的资料，以强化维基百科的搜寻功能，并将其他资料集连结至维基百科。透过这样的语义化技术的介入，让维基百科的庞杂资讯有了许多创新而有趣的应用，DBpedia 同时也是世界上最大的多领域知识本体之一，也是 Linked Data 的一部分，美国科技媒体 ReadWriteWeb 也将 DBpedia 选为2009 年最佳的语义网应用服务。

DBpedia 2014 版的资料集拥有超过458万的物件，包括144万5000人、73万5000个地点、12万3000张唱片、8万7千部电影、1万9000种电脑游戏、24万1000个组织、25万1000种物种和6000个疾病。其资料不仅被BBC、路透社、纽约时报所采用，也是Google、Yahoo等搜寻引擎检索的对象。

### 2.3 echarts介绍

ECharts，缩写来自Enterprise Charts，商业级数据图表，一个纯Javascript的图表库，可以流畅的运行在PC和移动设备上，兼容当前绝大部分浏览器（IE6/7/8/9/10/11，chrome，firefox，Safari等），底层依赖轻量级的Canvas类库ZRender，提供直观，生动，可交互，可高度个性化定制的数据可视化图表。创新的拖拽重计算、数据视图、值域漫游等特性大大增强了用户体验，赋予了用户对数据进行挖掘、整合的能力。

支持折线图（区域图）、柱状图（条状图）、散点图（气泡图）、K线图、饼图（环形图）、雷达图（填充雷达图）、和弦图、力导向布局图、地图、仪表盘、漏斗图、事件河流图等12类图表，同时提供标题，详情气泡、图例、值域、数据区域、时间轴、工具箱等7个可交互组件，支持多图表、组件的联动和混搭展现。

## 3系统详细设计

### 3.1 系统框架搭建

系统是一个基于javascript的网页，网页主体包括检索、搜索结果及可视化展示、推荐结果及可视化展示几个部分。检索由界面logo、检索框、检索按钮、检索类别几个组件构成。



图1 系统检索部分设计图

检索结果包括检索结果主要内容展示，一个实体关键字检索之后可能会返回多个不同类别的出版物信息，这里选取了第一条检索结果进行展示，展示内容包括出版物的图片、语言、国家、作者以及内容简介。然后返回其他的检索结果列表，列表含有由10条检索结果信息，统计可视化部分会显示dbpedia数据库中检索到的所有对应的实体类别和数量。

推荐部分根据检索内容的类别，在dbpedia中检索与待检索出版物类别相似的出版物加入到推荐列表并进行展示，并对相关的推荐指数进行显示。

### 3.2 系统数据处理

系统接收搜索框的输入内容后将其保存下来，然后通过查询语句进行检索，通过SPARQLWrapper工具完成sparql语句的查询，最后将查询结果保存至一个列表集合中，在页面上进行显示。其中检索的主要查询语句如下所示。

"prefix foaf: <http://xmlns.com/foaf/0.1/> "

+ "select distinct ?type ?value "

+ "where { "

+ "?s foaf:name ?sname . "

+ "?sname bif:contains '" + name + "'. "

+ "?s ?type ?value . "

+ "} ";

获取检索结果的语句部分如下所示。

text += "<img src = '" + data["http://xmlns.com/foaf/0.1/depiction"] + "' onload='changeSize(this)'/><br />";

text += "<h2>语言</h2>";

tmp = getContent(data["http://dbpedia.org/property/language"]);

text += "<p><a href='http://en.wikipedia.org/wiki/" + tmp + "'>" + tmp.replace(/\\_/, " ") + "</a></p>";

text += "<h2>国家</h2>";

tmp = getContent(data["http://dbpedia.org/property/country"]);

text += "<p><a href='http://en.wikipedia.org/wiki/" + tmp + "'>" + tmp.replace(/\\_/, " ") + "</a></p>";

text += "<h2>作者</h2>";

text += data["http://dbpedia.org/property/author"]+ "<br />";

text += "<h2>相关链接</h2>";

text += data["http://dbpedia.org/ontology/wikiPageExternalLink"] + "<br />";

search\_ting("result\_ting").innerHTML = text;

text = "<h2>简介</h2>";

text += data["http://dbpedia.org/ontology/abstract"] + "<br />";

推荐的相关语句如下所示。

"PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> "

+ "PREFIX owl: <http://www.w3.org/2002/07/owl#> "

+ "PREFIX xsd: <http://www.w3.org/2001/XMLSchema#> "

+ "PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> "

+ "SELECT COUNT(?movie) SAMPLE(?movie) "

+ "where "

+ "{ "

+ "dbr:" + name + " dct:subject ?o . "

+ "?movie dct:subject ?o . "

+ "FILTER (?movie != dbr:" + name + ") . "

+ "} GROUP BY ?movie ORDER BY DESC(COUNT(?movie)) limit 7";

### 3.3 结果可视化

结果可视化有两个部分，首先检索结果有三种类别，电影、图书以及游戏，三种类别出版物的检索结果分别以三个进度条的形式显示，选取echarts中的条形图进行展示，效果如下图所示。

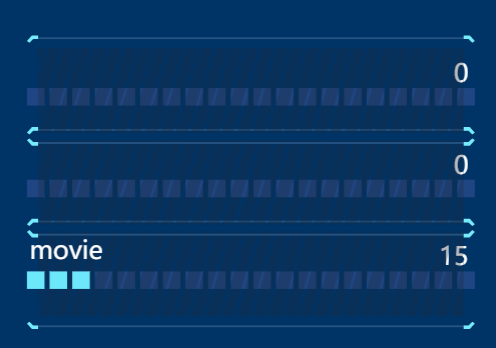


图2 出版物检索结果统计图

推荐结果列表中，每条结果对应着不同的推荐指数，每个出版物的指数分别通过查询结果获取，然后构成条形图的参数。

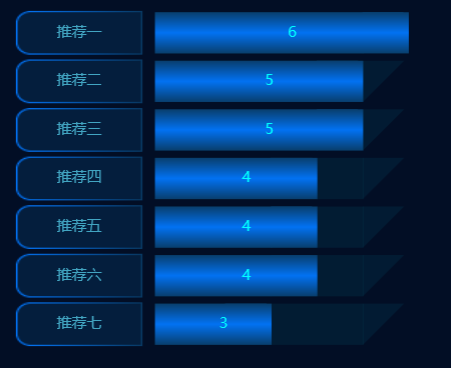


图3 出版物推荐结果统计图