信息检索设计文档

小组分工：

崔一鸣： 负责爬虫模块，数据库设计，web 后端开发

邹雨婷： 负责 web 前端开发，搜索自动补全

张　滔： 负责通配符索引的构建

周文涛： 负责检索和网页排名

钱塘文： 负责倒排记录表的构建

1. 项目简介

What's News 是一个垂直的新闻内容检索网站，收录了网易，搜狐，凤凰，央视等 4 个新闻网站的 10W+条网页。能够对新闻内容，评论进行检索，排名。检索时间小于 2 秒。同时 What's News 还提供了热点新闻推荐，搜索自动补全，相关搜索推荐，Snippet 生成，搜索结果预览，自定义排序等内容。

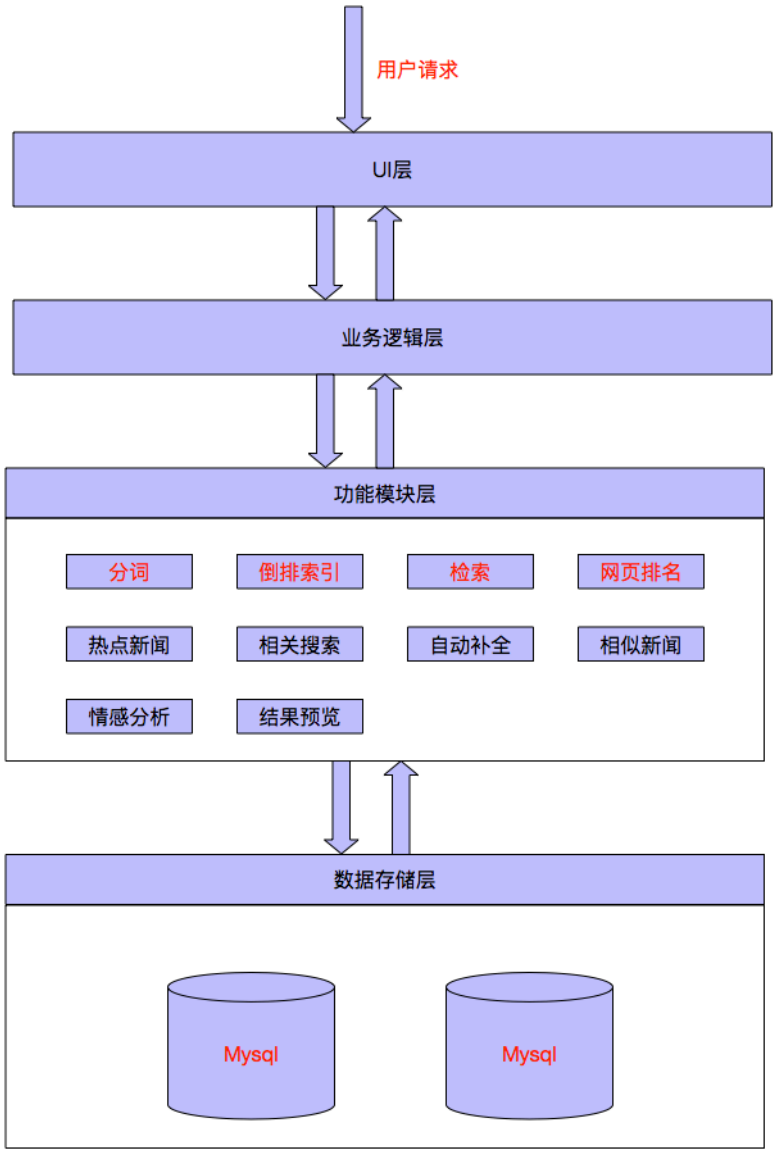
What's News 后端使用 Python 的 Django 框架，核心模块使用 java 开发。总代码量 4k+行。

1.1 搜索引擎地址：(在 UCAS 网络内访问)

http://123.57.35.217:8888/search

1.2 源代码地址： https://github.com/LeoCui/NIR

2. 总体框架



3 存储方案设计

3.1 mysql 数据库

系统主要采取 mysql 来进行数据存储方案。mysql 数据库中储存的信息包括文档内容信息，搜索历史记录信息，倒排记录表信息等。

注：在操作数据库时，要注意查询的列有没有索引，没有索引的查询会非常慢，但是索引会增加空间。后期视代码逻辑决定要不要给某个列加索引。

建表语句见 github 上 createTable.db 文件

3.1.1 news_info 表

news_info 表存储的是新闻的基本信息，表中字段包括：

id： 自增 id，主键

title: 新闻标题

url: 新闻的原始 url

url_hash: 对 url 进行 hash

pv: 新闻浏览量

category: 新闻分类

is_handled: 1 表示已经被倒排记录表处理，0 表示未被处理

comment_number: 新闻的评论数

publish_time: 新闻的发布时间

create_time: 该条记录的创建时间

update_time: 该条记录的更新时间

extra_info: 额外选项，方便以后拓展

索引项：pv，url_hash，comment_number，publish_time，这几项建索引是为了最热新闻推荐和 rank。

3.1.2 content_info 表

content_info 表存储的是新闻的内容部分（不含评论），单独拆开一张表，是

因为新闻内容的大小不确定，mysql 中变量只能定长。为了节省内存，可能需要把一条新闻内容存在多条记录中。

表中字段包括：

id: 自增 id，主键

news_id: 所在的新闻的 id

sequece_number: 该条记录在该条新闻内容中的序列号。

content: 内容

create_time: 该条记录中的创建时间

update_time: 该条记录的更新时间

extra_info: 额外选项，方便以后拓展

3.1.2 comment_info 表

comment_info 表存储的是新闻的评论，之所以要讲新闻和评论分开存储，原因如下：

a. 通常来说，一篇新闻的评论总是在更新（新增评论，删除评论）。但是新闻的内容是不变的，当评论更新时，我们就需要相应地更改倒排记录表，如果新闻内容和评论存储在一起，当评论更新时，我们需要对整个文档修改倒排记录表，分开的话只需要对评论部分修改倒排记录表。

b. 后续可能会有只针对评论信息的检索，那么对底层存储来说就需要将评论和新闻内容分开。

表中字段包括：

id: 自增 id，主键

news_id: 所在的新闻的 id

comment_number: 一条记录里的评论数

content: 评论内容，含有多条评论信息

create_time: 该条记录中的创建时间

update_time: 该条记录的更新时间

extra_info: 额外选项，方便以后拓展

索引：news_id

注：一条评论包括“日期 时间 用户名 内容”。评论之间用|分开

比如：2017-08-18 13:23:23 开飞机的贝塔 支持威武有希望了||2017-08-13 13：43:34 杨总 哈哈

为了防止干扰，爬虫程序在向数据库中插入评论时，需要预处理，保证用户名和内容中不含空格和|

3.1.3 search_history 表

主要用于自动补全和相关搜索推荐

表中字段包括：

id：自增 id，主键

content：搜索内容

user_id：用户 id，暂时不用

user_ip：用户 ip，暂时不用

create_time：该条记录创建时间

update_time：该条记录更新时间

extra_info： 额外选项，方便以后拓展

3.1.4 posting_list 表

倒排记录表，主要是为了将内存中建立的倒排记录表持久化存储起来，以便查询时使用。

表中字段如下：

id：自增 id，主键

term：单词

term_hash：该单词对应的 hash 值，（crc32）

df：出现该单词的文档的数目

sequence：该条记录在该 term 对应的记录的序号

content： 倒排记录表项

create_time：该条记录创建的时间

update_time: 该条记录更新的时间

extra_info: 额外选项，方便以后拓展

由于 term 对应的文档数不定，无法确定 posting_list 大小，所以可能一个单词对应的文档存储在多条记录中。

posting_list 中一个文档信息的结构如下： 0/1 docID tf <position1,position2>

不同文档之间用|分开其中:

0 表示新闻（contetn_info 表），1 表示评论（comment_info 表）

docID 对应的 news_id

tf 表示 term 在 docID 中出现的次数

<>中保存的是位置信息，用于位置索引

如： 0 34 3 <1,5,8>|1 45 2 <3,4>

注：当 content 的内容不够一个容纳一个文档的位置信息时，会把该文档的位置信息拆成多份存储。后面用到的时候需要注意。

3.1.5 hot_news 表

热点新闻表，主要存储的是最近的最热新闻。

表中字段如下：

id：自增 id，主键

news_id：新闻在 news_info 表中的 id

rank：该条新闻的热度排名

create_time：该条记录创建的时间

update_time：该条记录更新的时间

extra_info: 额外选项，方便以后拓展

3.1.6 dictionary 表

词典表，存储的是所有的 term

表中字段如下：

id: 自增 id，主键

term: 词项

term_hash: 词项的 hash（crc32）

create_time：该条记录创建的时间

update_time：该条记录更新的时间

extra_info: 额外选项，方便以后拓展

3.1.7 kgram_index 表

存储的是 kgram 索引，用来进行通配符查询

表中字段如下：

id: 自增 id，主键

kgram: kgram 索引

kgram_hash：kgram 的 hash 值（crc32）

number: 该条记录中 content 中 termId 的个数

content: content 里面存的是该 kgram 对应的在 dictionary 表中的 term 的 id 的集合。格式为<termId1, termId2>，一条记录插不下需要拆成多条插入。

create_time：该条记录创建的时间

update_time：该条记录更新的时间

extra_info: 额外选项，方便以后拓展

4. 模块设计和实现

4.1 Web 服务模块

Web 服务负责接收用户的 Http 请求，然后调用各种服务，进行相应的逻辑处理，最后返回给用户正确的结果。

Web 服务分为前端和后端两个部分，前后端是完全分离的。

4.1.1 Web 后端

web 后端使用了 Python 的 Django 框架，开发语言为 Python，web 后端主要

负责处理用户的请求，然后得到检索结果，并且将检索结果传递给前端。web 后端包含以下几个部分：

4.1.1.1 路由分发

路由分发采用的是 django 框架的路由分发机制，主要负责**正则匹配**用户的 url，然后交给相应的模块去处理。

该部分位于 app/search/url.py 文件中。

4.1.1.2 逻辑处理部分

这部分主要负责将用户输入的参数封装成 Json 的格式，然后通过调用各种接口(检索接口，相关搜索接口，历史记录接口，热点新闻接口，snippet 生成接口)，接口以 Json 的格式返回相关的内容，然后将内容返回给前端展示。

该部分位于 app/search/views.py

4.2.1 web 前端

web 前端使用了 jquery，javascript，ajax，模板等技术，主要负责 UI 和展示。这里不再具体阐述。

该部分位于 app/search/templates 文件夹中。

4.2 爬虫模块

爬虫模块负责爬取网站新闻存入数据库中，包括 4 个爬虫：网易新闻 APP，新浪新闻 App，搜狐新闻 App，央视新闻 App。爬取了 10w 个网页

爬虫模块使用 python 编写。

为了防止多次爬取到相同的新闻，在 news_info 表中设置了 url_hash 这个字段，保证相同 url 的新闻只被加入数据库中一次。

该部分位于 crawler 文件夹中

4.2.1 抓包

因为移动端的数据格式清晰，所以爬虫选择了移动端新闻 App。利用 fidder 抓包拿到移动端请求 url，然后利用 requests 模拟请求。

4.2.2 爬取新闻

这部分负责爬取部分新闻，解析返回的 Json 字符串，然后按照指定的格式存入数据库中的 news_info，content_info，comment_info 表中，供建立倒排记录表使用。

为了保持新闻的时效性，爬虫的脚本被加入了系统的定时任务，每天早上 8 点执行。

4.2.3 网页更新

这部分指的是网页内容有变化（评论增加，评论删除）的时候，用来更新数据库（news_info，comment_info 表）。

这部分和上个模块不同，上个模块指的是爬取了新的网页，本模块指的是已经爬取的网页的内容改变。

当网页内容改变时（比如该网页被删除），除了更改 news_info，comment_info 表之外，还需要删除倒排记录表中含有原有 docID 的记录。

4.3 倒排索引模块

这部分指的是读取数据库中的新闻，然后建立倒排记录表存入数据库中的 posting_list 表中。

该部分用 Java 编写

4.3.1 分词 segmentation()　<不对外接口>

利用 HanLP 外部工具进行分词，传入参数待分词的句子或文章，返回分词结果 arraylist< String>，为后续建立词典表做前期准备.

4.3.2 初始化倒排记录表 buildPostingList()　<对外接口>

读取 news_info 中的 is_handled 为 0 的新闻 ID，到 content_info 表与 comment_info 表中获取相关的新闻正文、评论内容后建立倒排记录表，并按照数据表的设计规则、停用表的规则进行处理、存入 posting_list 表中

4.3.2 更新倒排记录表 updatePostingList()　<对外接口>

直接在原有的倒排记录表 posting_list 的基础上更新，更新的具体步骤与上面一致

4.3.4 数据库相关操作，诸如连接、批处理、更新、关闭等 mysqlDBInterface.java

将数据库的相关操作进行了泛化封装，使得数据库的操作在前几个功能中变得简单；将数据库的插入、更新操作进行了批处理，加快了数据库执行部分的速度；将数据库的查询操作进行了泛化封装处理，使得前几个功能可以重复使用该功能；

4.3.5 配置文件的生成

将诸如数据库的 ip 地址、端口号、相关数据表名等写入配置文件中，再将配置文件读入程序中进行操作，使得该程序具有一定的泛化性。当诸如数据库的相关信息发生改变时，无需重新打包可执行 jar 包。

4.3.6 停用词表位于配置文件夹中

将停用词表作为配置文件，使得当停用词表更新或者暂停使用时，无需重新打包可执行 jar 包，一定程度提高了程序的泛化性。

4.4 通配符索引模块

这部分主要负责建立通配符索引并且存入数据库中的 kgarm_index 表中，具体地，采用 2-gram 索引

4.4.1 更新词典表 putDictionary()　<不对外接口>

在初始化倒排记录表、或更新倒排记录表的程序中内部调用该函数，存储倒排记录表 posting_list 中已处理新闻中出现的、所有的词项 term

4.3.2 更新 k-gram 表 handleKGram()　<不对外接口>

在更新词典表的程序中内部调用该函数。对词典表中尚未进行 k-gram 索引处理的词项，建立其 k-gram 索引，便于进行通配符查询，并按照数据表的设计规则、停用表的规则进行处理、存入 kgram_index 表中。

4.4 检索模块

该部分主要接受用户的 query，然后返回对应的 docID 的集合。检索使用 hash 表的方式。在倒排记录中有字段 hash，表示某个 term 的 hash 值。

技术：java

该部分需要对外提供一个接口：

query 函数：

输入：json 格式，设为 input

input['query']: 查询的单词

input['page']；第几页

input['category']:　分类

input['source'] : 来源

input['from']: 开始时间，如 2017-12-09 12:00:00

input['to'] : 结束时间

input['sort']：排序方式： 0 表示按照相关度排序，1 表示按照时间排序

输出： json 格式，设为 output

output['resultCount']：所有相关的文档的数量

output[' keywords']： 查询的关键词，是一个 list

output['docList']: 相关的文档的信息，是一个 list，list 成员为 doc

doc 的结构如下：

doc['id']：文档的 id

doc['relationship']：文档的相关度，如 95.32%

4.5 网页 rank 模块

这部分主要负责对检索到的文档打分，排序并输出。

技术：java

该部分需要对外提供一个接口：

rank(str，docIDSet，maxNumber): str 表示查询的字符串，docIDSet 表示相关的文档集合。maxNumber 表示返回的文档数目，-1 表示不限，返回值是排好序的文档集合。

4.6 其他模块

4.6.1 热点新闻推荐

该部分主要负责进行首页的热点新闻推荐。

该模块用 Python 编写，包括一个定时脚本，每隔

4.6.2 搜索自动补全

这部分主要负责根据查询的 query 给出相应的自动补全提示，调用的是百度的 api：http://suggestion.baidu.com/su

主要是前端的一段 javascript 代码来完成。

4.6.3 结果预览

该部分主要用于在检索页面预览内容，而不用点进去。当鼠标悬浮在预览按钮时，会发送一个 ajax 请求获取该新闻内容，然后以悬浮窗的形式显示出来。

该部分主要由前端的一段 javascipt 代码完成。

4.6.4 获取历史记录

该部分主要负责查询 search_history 表，然后获取最新的 4 条所有记录，返回给前端。

4.6.5 snippet 生成

该部分负责生成新闻的摘要，因为摘要要尽可能多地包含查询的关键词，所以该部分设计了一个滑动窗口（大小为 200 个字），遍历该文档，找到一个包含关键词最多的滑动窗口即为 snippet，并且将关键字标红显示。

4.6.6 相关搜索推荐

该部分主要根据搜索的 query 获取相关的搜索，该部分调用的是 Bing 的 API: api.bing.com/osjson.aspx?query=，然后将结果返回给前端

4.6.7 自定义筛选

因为数据库中存储了新闻的来源，分类，时间字段，所以该部分支持用户对检索处理的内容进行二次筛选。

当用户点击相关的筛选按钮时，前端会将筛选条件发送给后端的检索模块，检索模块根据筛选条件进行相应的过滤操作，最后返回用户过滤后的结果。

4.6.8 自定义排序

支持按照相关度和发布时间进行排序。

5 效果展示

5.1 首页

5.2 热点新闻推荐

5.3 搜索自动补全


美国
NEWS
美国航空
美国大使馆签证中心
美国末日
美国众神
美国时间
美国亚马逊
美国签证
美国地图
美国神婆
美国严寒8人冻死

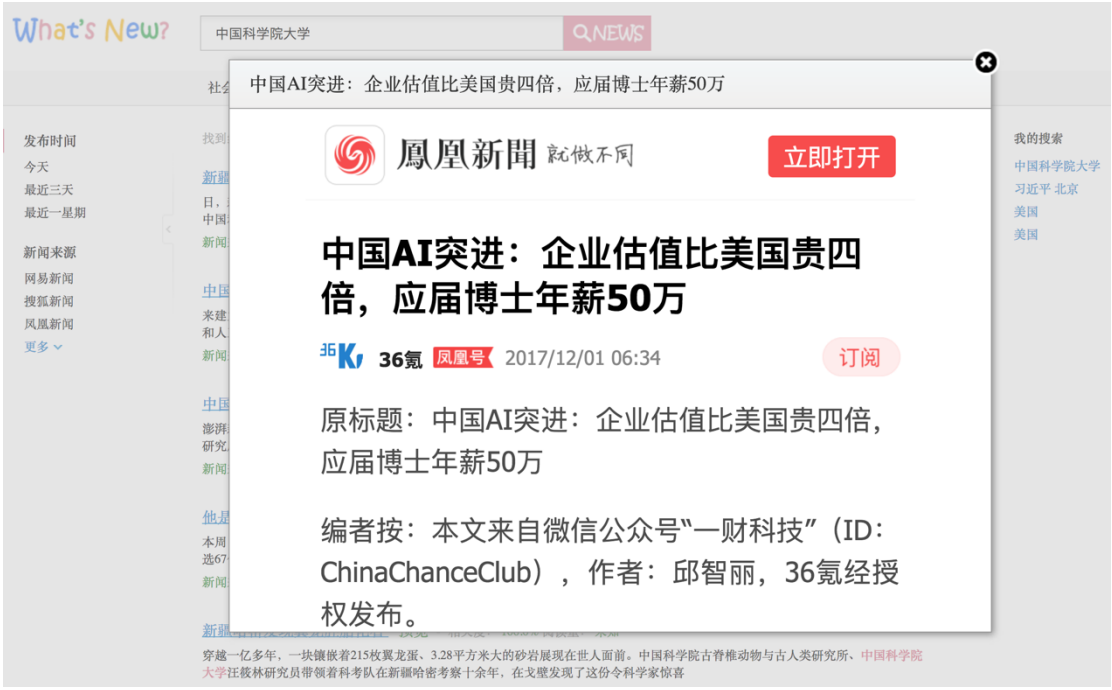
5.4 新闻搜索

5.5 评论搜索

新闻搜索支持评论搜索，当评论包含相关的内容也会被搜索出来

5.6 新闻预览

当鼠标悬浮在新闻标题后面的预览上时，会出现下面的预览浮层。



5.7 snippet 生成

在每条新闻的标题的下面有一小段相似新闻的摘要。

美国"第一女儿"极力夸奖"第一外孙女"说中文：真是太可爱了！ 预览 - 相关度：99.9% 阅读量：2193996

阿拉贝拉唱中文、念古诗和三字经的视频再次在中国圈粉。作为母亲，伊万卡自然也不会遮掩自己对阿拉贝拉的骄傲之情。据英国《每日邮报》报道，伊万卡·特朗普在13日的采访中极力夸赞了自己的女儿

摘要

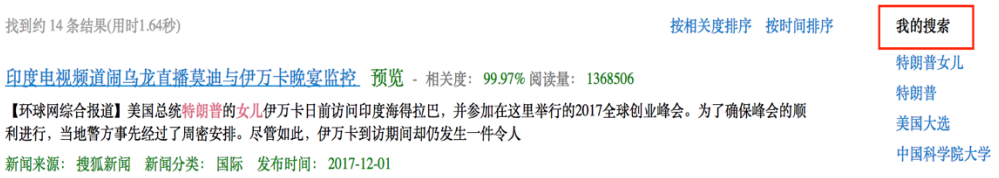
新闻来源：搜狐新闻 新闻分类：国际 发布时间：2017-11-15

5.8 相关搜索推荐

在结果页面的最下面有相关搜索推荐，个数限制在了 4 个

5.9 搜索历史推荐

在结果页的右侧有用户的搜索历史的推荐，个数限制在了 4 个



5.10 自定义筛选

在结果页，用户可以按照分类，来源，时间对检索出来的新闻进行二次筛选。

5.11 自定义排序

在结果页，用户可以按照相关度，发布时间进行排序，默认情况下按照相关度进行排序。

找到约 4 条结果(用时2.44秒)

按相关度排序　按时间排序

印度电视频道闹乌龙直播莫迪与伊万卡晚宴监控　预览 - 相关度：99.97% 阅读量：1368506

【环球网综合报道】美国总统特朗普的女儿伊万卡日前访问印度海得拉巴，并参加在这里举行的2017全球创业峰会。为了确保峰会的顺利进行，当地警方事先经过了周密安排。尽管如此，伊万卡到访期间却仍发生一件令人

新闻来源：搜狐新闻　新闻分类：国际　发布时间：2017-12-01

印度名城否认为伊万卡"清除乞丐"：没有直接关系　预览 - 相关度：99.97% 阅读量：1934531

据印度《教徒报》11日报道，海得拉巴已开启乞丐收容模式，在城市中沿街乞讨已被定为违法，大约有6000名乞丐将被强制性安置在收容所。由于海得拉巴即将在本月底举行"全球企业家峰会"，而美国总统特朗普的大女儿

新闻来源：搜狐新闻　新闻分类：国际　发布时间：2017-11-13

美国"第一女儿"极力夸奖"第一外孙女"说中文：真是太可爱了！　预览 - 相关度：99.9% 阅读量：2193996

阿拉贝拉唱中文、念古诗和三字经的视频再次在中国圈粉。作为母亲，伊万卡自然也不会遮掩自己对阿拉贝拉的骄傲之情。 据英国《每日邮报》报道，伊万卡·特朗普在13日的采访中极力夸赞了自己的女儿

新闻来源：搜狐新闻　新闻分类：国际　发布时间：2017-11-15

6 创新点

数据库用了事务
term_hash
评论和内容分离

7 经验教训