

Gooniu 微博搜索平台设计文档

信息检索大作业

组员：

郑喆君：201328015059079

王彦：201328015059074

崔艳玲：201318015029004

景丽莎：201328015029013

麻婧：201328015029023

2013 年 12 月 19 日

前言

你今天微博了吗？随着互联网的飞速发展，有越来越多的网民有自己的微博账号，甚至有的人不止一个账号，多达 10 个以上。自然，微博的数量也是数亿级别的，要想从这么多微博中找到自己感兴趣的信息可谓是非常困难。

本项的目的是为了实现一个微博搜索引擎，帮助人们快速找到所需。我们小组所做的工作以及结果如下。

微博来源

微博的下载是基于新浪微博开放平台。

微博字段说明：

返回值字段	字段类型	字段说明
created_at	string	微博创建时间
id	int64	微博 ID
mid	int64	微博 MID
idstr	string	字符串型的微博 ID
text	string	微博信息内容
source	string	微博来源
favorited	boolean	是否已收藏，true: 是，false: 否
truncated	boolean	是否被截断，true: 是，false: 否
in_reply_to_status_id	string	（暂未支持）回复 ID
in_reply_to_user_id	string	（暂未支持）回复人 UID
in_reply_to_screen_name	string	（暂未支持）回复人昵称
thumbnail_pic	string	缩略图片地址，没有时不返回此字段
bmiddle_pic	string	中等尺寸图片地址，没有时不返回此字段
original_pic	string	原始图片地址，没有时不返回此字段
geo	object	地理信息字段 详细
user	object	微博作者的用户信息字段 详细
retweeted_status	object	被转发的原微博信息字段，当该微博为转发微博时返回 详细
reposts_count	int	转发数
comments_count	int	评论数
attitudes_count	int	表态数
mlevel	int	暂未支持
visible	object	微博的可见性及指定可见分组信息。该 object 中 type 取值，0: 普通微博，1: 私密微博，3: 指定分组微博，4: 密友微博；list_id 为分组的组号

pic_urls	object	微博配图地址。多图时返回多图链接。无配图返回“[]”
ad	object array	微博流内的推广微博 ID

用户字段如下：

返回值字段	字段类型	字段说明
id	int64	用户 UID
idstr	string	字符串型的用户 UID
screen_name	string	用户昵称
name	string	友好显示名称
province	int	用户所在省级 ID
city	int	用户所在城市 ID
location	string	用户所在地
description	string	用户个人描述
url	string	用户博客地址
profile_image_url	string	用户头像地址（中图），50×50 像素
profile_url	string	用户的微博统一 URL 地址
domain	string	用户的个性化域名
weihao	string	用户的微号
gender	string	性别，m：男、f：女、n：未知
followers_count	int	粉丝数
friends_count	int	关注数
statuses_count	int	微博数
favourites_count	int	收藏数
created_at	string	用户创建（注册）时间
following	boolean	暂未支持
allow_all_act_msg	boolean	是否允许所有人给我发私信，true：是，false：否
geo_enabled	boolean	是否允许标识用户的地理位置，true：是，false：否
verified	boolean	是否是微博认证用户，即加 V 用户，true：是，false：否
verified_type	int	暂未支持
remark	string	用户备注信息，只有在查询用户关系时才返回此字段
status	object	用户的最近一条微博信息字段 详细
allow_all_comment	boolean	是否允许所有人对我的微博进行评论，true：是，false：否
avatar_large	string	用户头像地址（大图），180×180 像素
avatar_hd	string	用户头像地址（高清），高清头像原图
verified_reason	string	认证原因
follow_me	boolean	该用户是否关注当前登录用户，true：是，false：否
online_status	int	用户的在线状态，0：不在线、1：在线
bi_followers_count	int	用户的互粉数
lang	string	用户当前的语言版本，zh-cn：简体中文，zh-tw：繁体中文，en：英语

下面例举一条微博说明微博的字段格式：

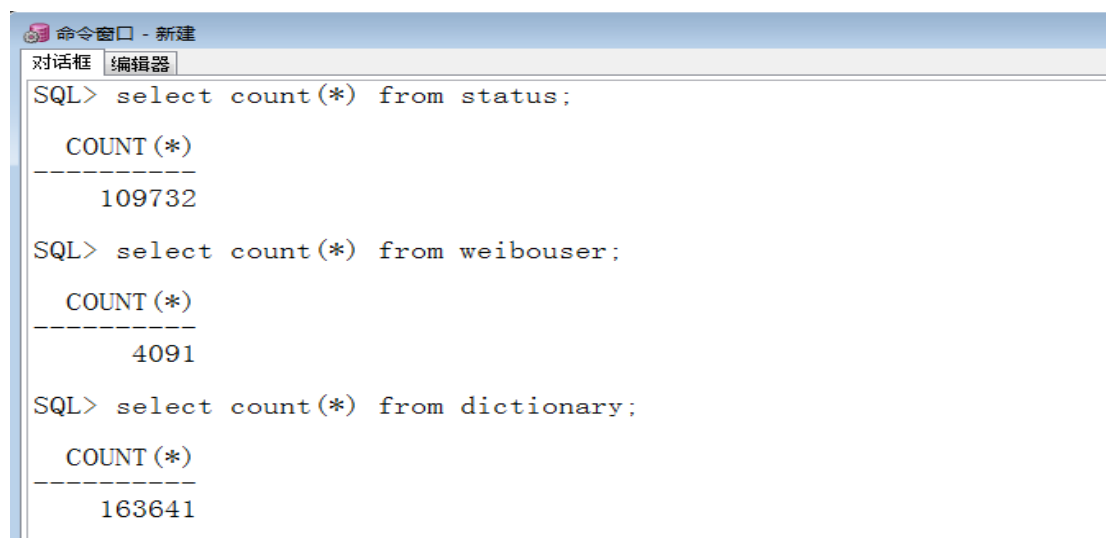
Status [user=User [id=2819027752, screenName=中国好声音, name=中国好声音, province=33, city=1, location=浙江 杭州, description=正版引进的大型励志类专业音乐评论节目《The Voice of China》由浙江卫视与星空旗下灿星制作联合出品！更有那英、庾澄庆、汪峰、张惠妹四位重量级…, url=, profileImageUrl=http://tp1.sinaimg.cn/2819027752/50/40021416681/1, userDomain=, gender=m, followersCount=1571784, friendsCount=289, statusesCount=7389, favouritesCount=8, createdAt=Fri Jun 01 15:15:45 CST 2012, following=true, verified=true, verifiedType=3, allowAllActMsg=false, allowAllComment=true, followMe=false, avatarLarge=http://tp1.sinaimg.cn/2819027752/180/40021416681/1, onlineStatus=1, status=null, biFollowersCount=227, remark=null, lang=zh-cn, verifiedReason=浙江卫视《中国好声音》官方微博, weihao=, statusId=], idstr=3628357360813072, createdAt=Mon Sep 30 22:27:49 CST 2013, id=3628357360813072, text=媒体评审依次投票, 双方票数你追我赶咬得很紧。101位媒体评审的票数结果为: 蘑菇兄弟 54 票, 苏梦玫 47 票, 仅差 7 票。蘑菇兄弟和苏梦玫, 谁将获得争夺哈林组冠军的机会? 哈林老师给出的分数配比会扭转当前的局势吗?, source=Source [url=http://weibo.com/, relationship=nofollow, name=新浪微博], favorited=false, truncated=false, inReplyToStatusId=-1, inReplyToUserId=-1, inReplyToScreenName=, thumbnailPic=http://ww1.sinaimg.cn/thumbnail/a806f328tw1e94w42lejhg20go0b40uz.jpg, bmiddlePic=http://ww1.sinaimg.cn/bmiddle/a806f328tw1e94w42lejhg20go0b40uz.jpg, originalPic=http://ww1.sinaimg.cn/large/a806f328tw1e94w42lejhg20go0b40uz.jpg, retweetedStatus=null, geo=null, latitude=-1.0, longitude=-1.0, repostsCount=0, commentsCount=0, mid=3628357360813072, annotations=, mlevel=0, visible=Visible [type=0, list_id=0]]

红色标注的是用户字段, 蓝色标注的是微博字段。可以看出一条微博维数较高, 一些信息对于我们微博检索没有用处, 所以我们进行了维规约, 抽取为如下信息。

用户字段保留了: id, screenName, url, profileImageUrl, followersCount=1571784, friendsCount=289, statusesCount=7389

微博字段保留了: id, text, repostsCount, commentsCount, createdAt

经过小组成员近三个月的微博抓取, 微博量达到 109732 条, 用户量达到 4091 位。词典中的词条数达到 163641 条。系统是靠 oracle 数据库完成对微博数据, 用户数据, 词典数据的存储。如下图所示:



```

命令窗口 - 新建
对话框 编辑器
SQL> select count(*) from status;

COUNT (*)
-----
109732

SQL> select count(*) from weibouser;

COUNT (*)
-----
4091

SQL> select count(*) from dictionary;

COUNT (*)
-----
163641

```

数据库中有三个表，分别为微博表，微博用户表，倒排索引表。每个表说明如下：

微博数据表 status:

Name	Type	Nullable	Default	Comments
ID	VARCHAR2(50)			
CREATEDDATE	VARCHAR2(50)	Y		
TEXT	VARCHAR2(1000)	Y		
RETWEETEDSTATUS	VARCHAR2(50)	Y		
ORIGINALPIC	VARCHAR2(100)	Y		
REPOSTCOUNT	NUMBER(10)	Y		
COMMENTSCOUNT	NUMBER(10)	Y		
USERID	VARCHAR2(50)	Y		

微博用户表 weibouser:

Name	Type	Nullable	Default	Comments
ID	VARCHAR2(50)			
SCREENNAME	VARCHAR2(50)	Y		
DESCRIPTION	VARCHAR2(500)	Y		
IMAGEURL	VARCHAR2(100)	Y		
GENDER	VARCHAR2(5)	Y		
FOLLOWERCOUNT	NUMBER(10)	Y		
FRIENDSCOUNT	NUMBER(10)	Y		
STATUSCOUNT	NUMBER(10)	Y		

词典表:

Name	Type	Nullable	Default	Comments
TERM	VARCHAR2(100)			
OBJECT	BLOB	Y		

倒排索引：

我们采用倒排索引的方法建立索引。对于抓取到的微博用 IKAnalyzer 分词器进行分词，建立词项-文档对，存入数据库。对于新抓取到的微博，支持增量式存数索引信息。一个词项的倒排索引没有按照微博号排序，因为经观察，微博号不是递增的。因此索引中的微博按照时间排序，这样做也是为了后面既可以按照查询词项为单位和也可以按照以文档为单位计算文档评分进行排名。同时为倒排索引的快速合并提供了条件。

由于数据库中词项是主键，在进行查找的时候可以利用数据库高效的进行。

微博查询：

下图是我们组的搜索引擎系统的界面。用户可以在这里输入自己想要查找的内容。

在内存中有一个高速缓存，缓存用户已经输入过的查询及其结果。如果用户输入的查询之前已经有历史记录，那么直接返回高速缓存中的结果，显著的提高系统性能。这样做的好处在于，对于结果较多的查询，比如“中国”，如果每次都去访问数据库返回查询结果，就会产生较长时间的延迟，采用缓存就可以解决这个问题。缓存采用了先进先出的算法进行调度，其依据为两点：第一，最长时间内未被访问的微博可能已经不是当前热点话题，所以要调出缓存；第二，缓存中的内容如果长时间放进内存，那么这个查询对应的结果就会得不到更新，所以将其调出缓存，下次再有用户查询的时候可以查到更新的结果。

如果用户输入的查询在缓存中没有历史记录，那么就先对用户输入的查询进行分词，把分词结果传入数据库得到对应的倒排记录。把倒排记录传递到排序功能模块。



转基因

谷妞一下!

☒ 根据时间排序 ☐ 根据影响力排序 ☐ 根据相关性排序[加入我们](#) | [About Gooniu](#) | [Mailto Gooniu](#) | [Teacher WangBin](#)

©2013_IR_HomeWork_Gooniu 客服电话: 13261723921

排序功能:

如果用户输入的查询包含多个词项,我们不是采用去交集的方法,而是取并集。因为有时候取交集的时候得不到返回的结果,显然返回弱相关的比什么都不返回要好。对于取并集的微博按照出现词项的多少进行初步排序。此外,我们实现了三种排序方法:

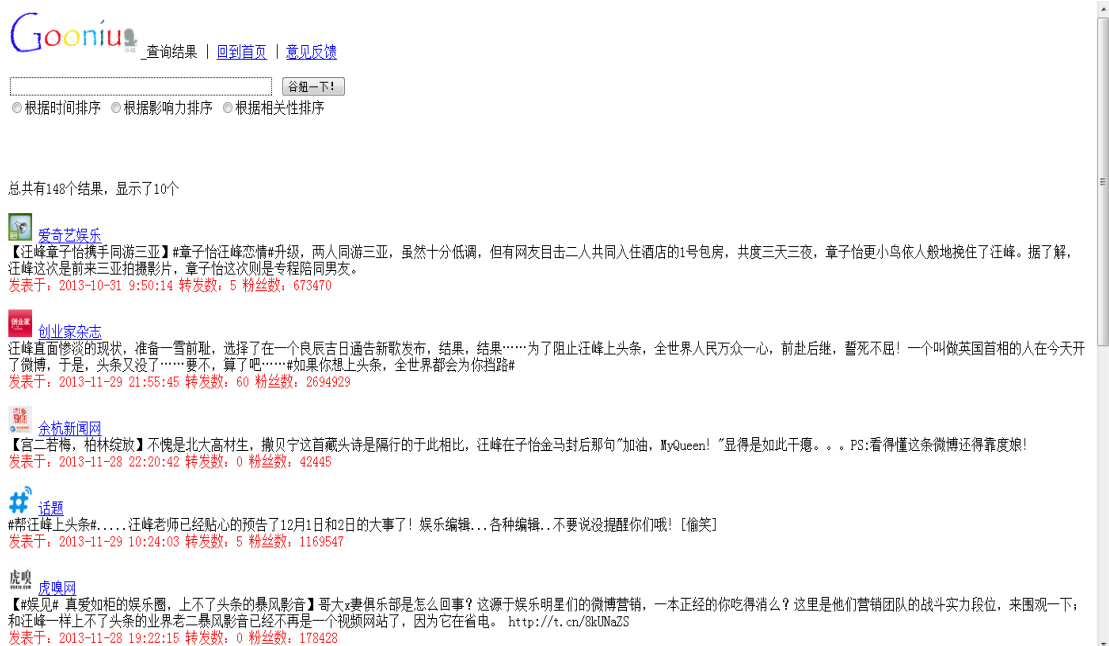
1. 按时间顺序:将最新发布的微博放在前面。微博具有时效性,大多数情况下人们喜欢看“新鲜事”,而不是已经过时的。此功能较容易实现。
下图为搜索关键字为“转基因”,按时间顺序,将最新发布的微博放在最前面的结果截图。

2. 按影响力排序:我们的影响力算法是权衡该微博的转发数和该微博发起人的粉丝数来排序。具体公式为 $(100000 * \text{转发数} + \text{粉丝数})$ 。其中 100000

这个参数是跟对一些（微博大 V 粉丝数/微博转发量）得到。对于返回的微博，按照两部分的得分，进行排序。下图为搜索关键字“宠物”，按影响力排行的结果截图。



3. 以相关度（tf-idf）排序：这是一种经典的权重评分办法。下图为搜索关键字为“汪峰”，以相关度（tf-idf）来得到的结果。



容错式检索：

这个模块包含两部分内容，查询扩展和拼写矫正。

查询扩展：基于用户不足导致的查询历史不够，我们不能采用基于查询日志进行查询扩展。实现了基于词典的简单用户查询扩展功能。对于词典我们在内存中建立了一个字典树，每次用户想查询文本框内输入内容的时间，系统会匹配后面可能的扩展。比如用户查询“担担面”，在只输入一个“担”的时候后面会匹

配成“担担面”以共用户选择。以提高查询的准确率。

拼写矫正：对于用户错误查询或者通配符查询，该系统提供了容错式检索功能。用户输入的查询词项如果在缓存和数据库中都找不到对应的倒排索引，那么就认为用户输入的内容可能打错字或者漏字了。我们采用了 **1-gram** 索引，对于词典中的每个词，分成字并建立索引。这样对于错误的查询词项中的每个字查找对应的索引，索引中出现频率最高的那个词最有可能是原查询正确的形式。

用户查询：

这是本微博检索系统的扩展功能。

测试：

经过小组成员测试数百次输入查询的测试，均能返回较好的检索结果。

创新点：

- 1、高速缓存机制加快系统检索速度，提高系统性能。
- 2、基于转发数和微博用户影响力的排序。

经验总结：

本系统还存在的不足：缺乏一个综合的排名，缺乏把微博转发数、发布者影响力、**tf-idf**、时间等因素综合起来的综合评分计算方法。由于时间紧迫未能实现。在今后空余时间实现。

经过这次微博系统的实现，本组成员不仅锻炼了项目整体架构的能力，团队协作的能力，代码实现的能力，更重要的是对信息检索课程的知识有了一个系统整体的认识。

附录：

源代码请见 `source` 文件夹