基于父子文档实现的ES关联查询

黄耀鸿 @数说故事

背景

- 社会化媒体越来越多，越来越具有价值
- 社会化媒体中，一般包含用户、内容两种类型
- 实时返回对社会化媒体的分析结果非常有意义
- 数据量已经不太适合关系型数据库进行处理
- 如何使用Elasticsearch来进行数据分析？

概览

- Elasticsearch VS SQL
- 父子文档的概念
- 父子文档的使用方式
- 实际应用场景
- 实践经验

Elasticsearch VS SQL

select * from weibo where text = '宋仲基'

name	sex	text
A小姐	女	宋仲基
A小姐	女	宋仲基
B先生	男	宋仲基
C先生	男	宋
D先生	男	仲

转化为ES语句

```
{
    "query":{
        "term":{
            "text":"宋仲基"
        }
    }
}
```

Elasticsearch VS SQL

select sex, count(sex) from weibo group by sex

转化为ES语句

sex	count
女	2
男	3

```
"aggs": {
    "sex": {
        "terms": {
            "field": "sex"
        }
    }
}
```

Elasticsearch VS SQL

select sex, count(distinct nickname)
from weibo group by sex

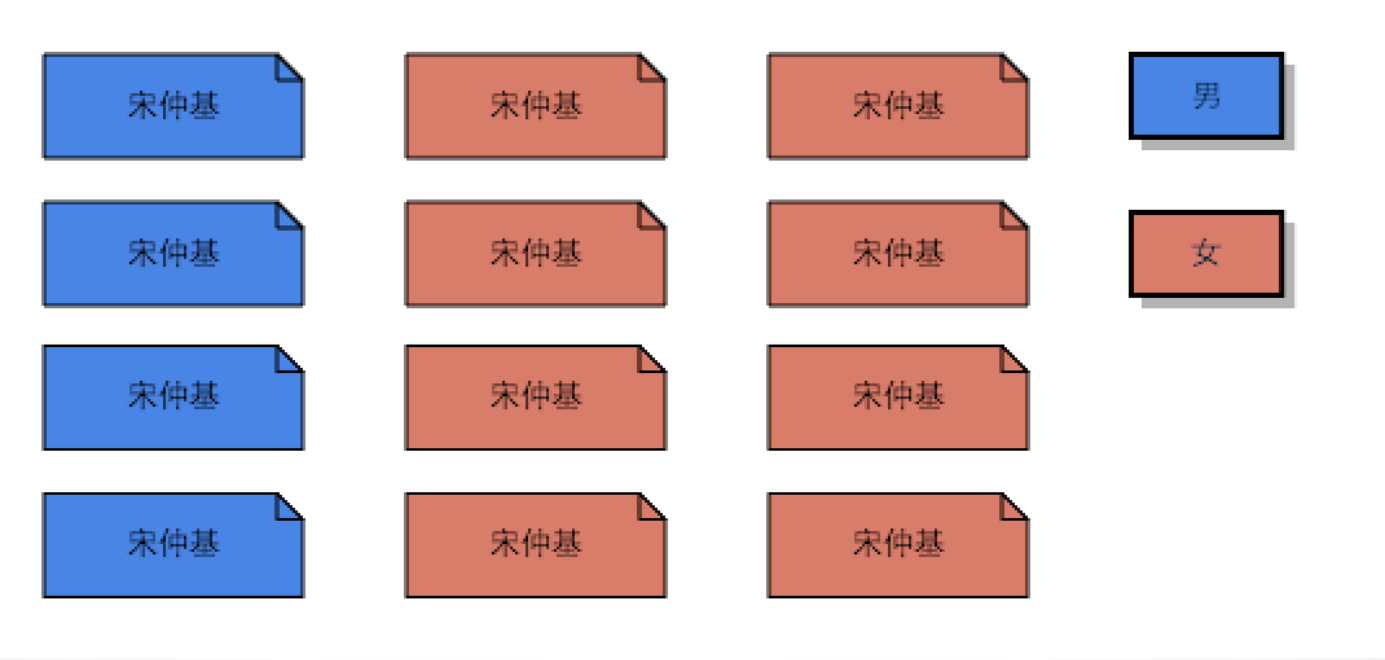
转化为ES语句

sex	count
女	1
男	3

?

数据分析场景

- 假设微博中，发表内容提及到“宋仲基”的人群有100万人，他们的男女分布是多少？



数据分析场景

select sex, count(sex) from weibo

where text = '宋仲基' group by sex X

- 由于一个人可能发表多条提及到“宋仲基”的微博。所以需要对微博用户进行去重统计，才能得到真正的男女比例。

父子文档的基本概念

- 在ES中，每一条记录都是一个文档。
- 在ES查询中，每次查询都是返回文档的集合
- 父子文档(parent-child)，可以让两种不同的文档类型形成一对多关系，从而实现关联查询
- 例如：一个微博用户可以对应多条微博内容

父子文档的Mappings

```
PUT /weibo-index
{
  "mappings": {
    "weibo": {
      "_parent": {
        "type": "user"
      },
      "properties": {
        "内容": {
          "type": "string"
        }
      }
    },
    "user": {
      "properties": {
      }
    }
  }
}
```

- 建立Mappings的时候，子文档需要指定_parent
- 可以有多种type的子文档指向同一type的父文档

父子文档的索引

```
POST /weibo-index/user/1855826357
{
  "nickname": "A小姐",
  "sex": "女"
}
```

- 更新父文档不会影响子文档
- 索引子文档的时候，一定要指定父文档的ID

```
POST /weibo-index/weibo/379832?
parent=1855826357
{
  "source": "iphone",
  "text": "ABC"
}
```

- 可以先索引父文档，也可以先索引子文档
- 子文档会索引到父文档所在的Shard

父子文档的查询

```
POST /weibo-index/user/_search/
{
  "query": {
    "has_child": {
      "type":      "weibo",
      "query": {
        "term": {
          "source": "iPhone"
        }
      },
      "inner_hits": {}
    }
  }
}
```

- 可以根据子文档的分数进行排序
- inner_hits可以关联显示文档

```
"hits": [
  {
    "_type": "user",
    "_id": "1855826356",
    "_source": {
      "nickname": "B先生",
      "sex": "男"
    },
    "inner_hits": {
      "weibo": {
        "hits": {
          "hits": [
            {
              "_type": "weibo",
              "_id":
"3798329022436050",
              "_source": {
                "text": "AAA",
                "source": "iPhone"
              }
            }
          ]}}}}]
```

父子文档的聚合

```
POST /weibo-index/user/_search
{
   "size": 0,
   "aggs": {
      "weibo": {
         "children": {
            "type": "weibo"
         },
         "aggs": {
            "source": {
               "terms":{
                  "field": "source"
               }
            }
         }
      }
   }
}
```

- 不支持reverse_nested

```
"aggregations": {
    "weibo": {
      "doc_count": 1,
      "source": {
        "doc_count_error_upper_bound": 0,
        "sum_other_doc_count": 0,
        "buckets": [
          {
            "key": "iPhone",
            "doc_count": 1
          }
        ]
      }
    }
}
```

父子文档的遍历

```
POST /weibo-index/user/_search?scroll=1m&search_type=scan
{
   "query": {
      "match_all" : {}
   }
}

POST /weibo-index/weibo/_search?scroll=1m&search_type=scan
{
   "query": {
      "match_all" : {}
   }
}
```

- 父子文档分别进行遍历

父子文档的删除

DELETE /weibo-index/user/1855826357

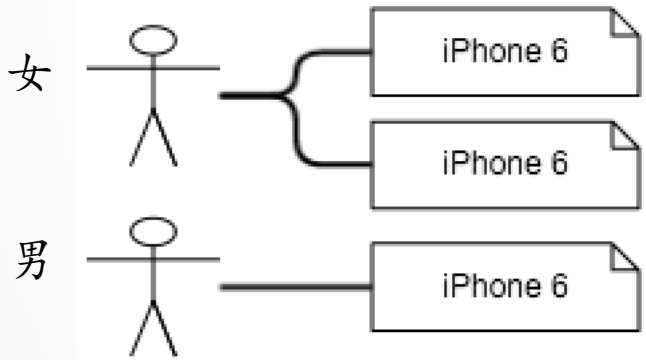
DELETE /weibo-index/weibo/3798329022436050

DELETE /weibo-index/weibo/3798329022436050?parent=1855826357

- 删除父文档不会同时删除子文档
- 不指定parent会删除所有该ID的子文档
- 如果想删除某个父文档下面的子文档，需要用delete by query实现

父子文档的应用

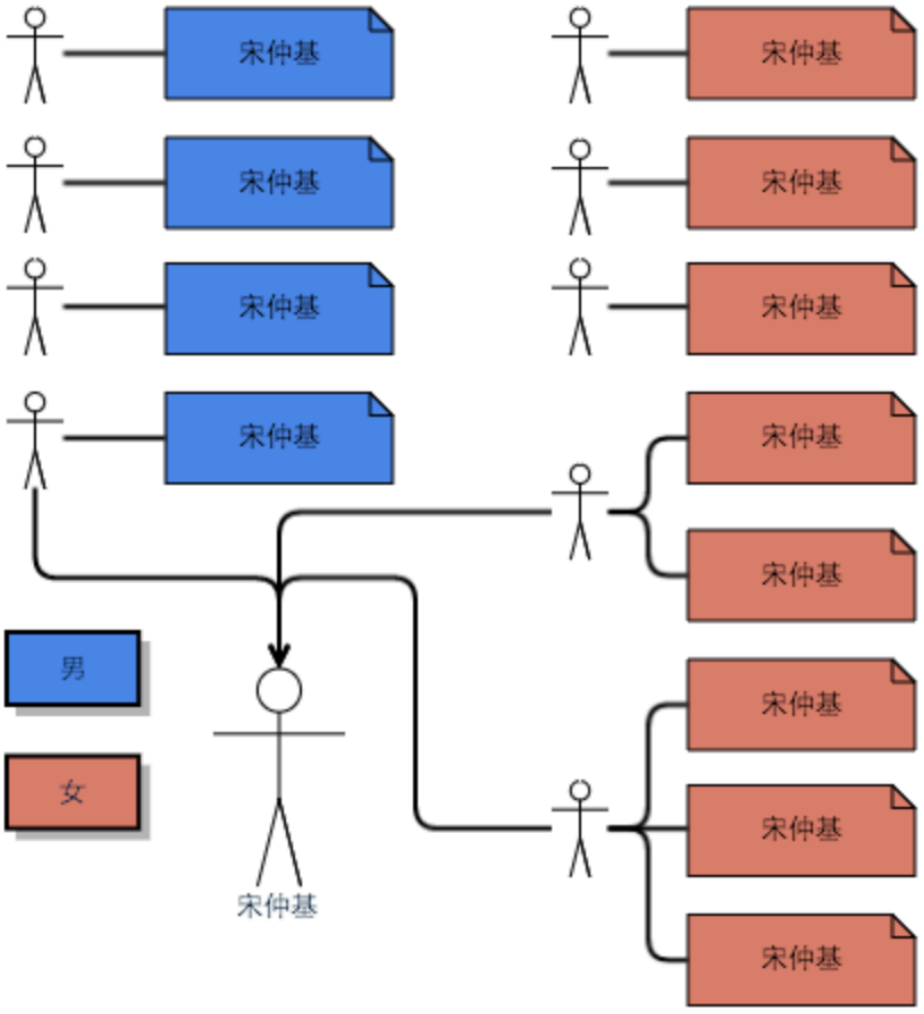
- 使用iPhone 6发布过微博的用户的男女比例分布



sex	count
女	1
男	1

```
POST /weibo-index/user/_search
{
 "query":{
  "has_child":{
   "type":     "weibo",
   "query":{
    "term":{
     "source": "iPhone 6"
    }
   }}
 },
"aggs":{
  “sex”:{
   "terms":{
    “field”: "sex"
   }
  }
 }}
```

实践经验

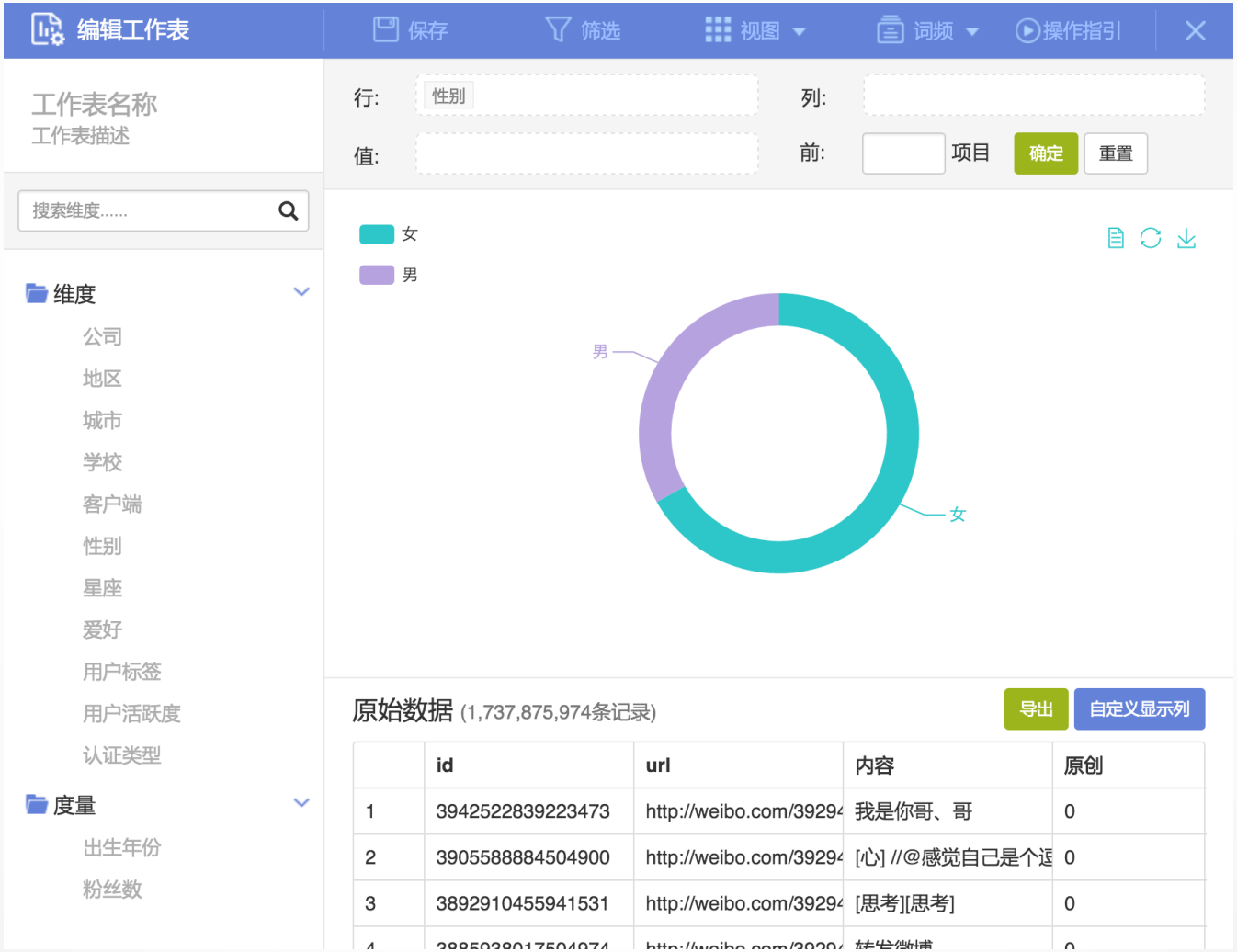


- 活动开始前，查询出提及宋仲基的用户性别比例，以及，过滤出关注宋仲基的用户，查询出其提及宋仲基的用户性别比例
- 活动结束后，再查询一次，只需要几秒钟，就可以快速得到活动效果，如果是男生比例上升了呢？

实践经验

- 明确是否需要父子文档，需要关联查询
- ES2.0存储父子文档关系改为doc values方式，降低内存使用
- 通用可视化插件一般不支持父子文档，不能快速看效果
- 子文档跟随父文档存储，Shard会出现不均衡的情况
- 减少了存储数据量，更容易更新父文档
- 和cardinality比较，更加麻烦，但是更加快速

数据可视化


编辑工作表
保存
筛选
视图
词频
操作指引
工作表名称
工作表描述
行:
性别
列:
值:
前:
项目
确定
重置
搜索维度.....
女
男
维度
公司
地区
城市
学校
客户端
性别
星座
爱好
用户标签
用户活跃度
认证类型
度量
出生年份
粉丝数
男
女
原始数据 (1,737,875,974条记录)
导出
自定义显示列
id
url
内容
原创
1
3942522839223473
http://weibo.com/3929
我是你哥、哥
0
2
3905588884504900
http://weibo.com/3929
[心] //@感觉自己是个逗
0
3
3892910455941531
http://weibo.com/3929
[思考][思考]
0

Q & A

Thank you