# 组合查询

**摘要：**本文档主要介绍组合查询实现方式，及对实现方式进行验证。组合查询根据查询语句关键词个数主要分为单关键词查询和多关键词查询，

组合查询为比财2.0中语音搜索功能实现的第二阶段的目标，第二阶段的目标为实现特定条件的组合查询，包含收益率、锁定期、理财期限、起购金额和银行属性等查询，如：对用户输入的“收益大于6%的产品”的查询语句，需要返回满足查询要求的金融产品，并对查询到的结果按一定属性进行排序展示。

组合查询根据用户输入查询语句中查询关键词的个数，分为单关键词查询和多关键词查询。本文档将分别介绍两种形式的实现方法。

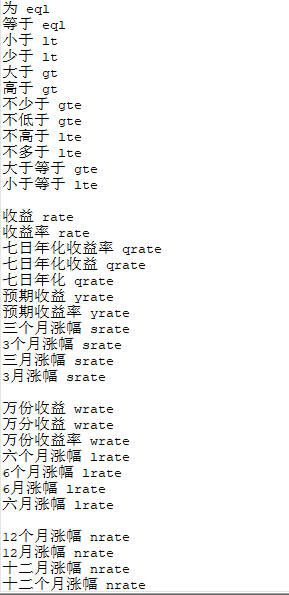
## 准备工作

在实现组合查询之前，需要做一些准备工作，主要包含配置自定义词典，对组合查询语句的解析结果输出格式进行定义等。

### 配置自定义词典

为适应比财2.0数据库的内容，为了解析查询语句的内容，加入了字段名称、逻辑关键词等词及其标注，同时添加产品和机构等词的自定义词典，是由于有些产品名称中有数字，如“非凡资产管理126天增利”，如果不添加产品名称词典，将会把该查询解析为组合查询，而本查询的意图是查询此产品，添加自定义词典后，才能正确解析。

自定义词典格式如下：



### 定义输出格式

为方便elasticsearch调用解析的结果，将解析的输出结果定义为json结构，参考elasticsearch查询时的json结构体格式。单关键词查询与多关键词查询的返回结果不同，下面分别介绍介绍。

#### 单关键词返回结果

根据产品属性，单关键词的查询语句主要包含四个维度：产品类型，逻辑关系词，字段名称，字段值四个，具体结构如下：

{

 “att\_value”:\*\*

“att\_name”:[

\*\*,

\*\*,

…..

]

“relat\_sign”:\*\*

“prd\_type”:[

\*\*,

\*\*,

……

]

}

#### 多关键词返回结果

借鉴elasticsearch中bool过滤的形式，通过操作符来表示返回结果的布尔逻辑，它包含以下操作符：

* must :: 多个查询条件的完全匹配,相当于 and。
* must\_not :: 多个查询条件的相反匹配，相当于 not。
* should :: 至少有一个查询条件匹配, 相当于 or。

如下

{   
           "should": [   
                    {

  “att\_value”:\*\*

“att\_name”:[

\*\*,

\*\*,

…..

]

“relat\_sign”:\*\*

“prd\_type”:[

\*\*,

\*\*,

……

]

},

 {

  “att\_value”:\*\*

“att\_name”:[

\*\*,

\*\*,

…..

]

“relat\_sign”:\*\*

“prd\_type”:[

\*\*,

\*\*,

……

]

}

                    ]

}

说明：must操作符内的产品类型必须相同，should不需要。

#### 字段说明

**relat\_sign**代表含义及表示方法

* 等于：eql
* 大于：gt
* 小于：lt
* 大于等于：gte
* 小于等于：lte

**prd\_type**代表含义及表示方法

* 货币基金：1
* 理财产品：2
* 纯债基金：3

**attr\_name**代表含义及表示方法

* 收益率
  + - * 七日年化收益率、预期收益率、3个月涨幅：RATE
      * 万份收益、六个月涨幅：RATE2
      * 12个月涨幅：RATE3
* 理财期限：PERIOD
* 锁定期：LOCKPERIOD
* 起购金额：MIN\_AMOUNT

## 单关键词查询

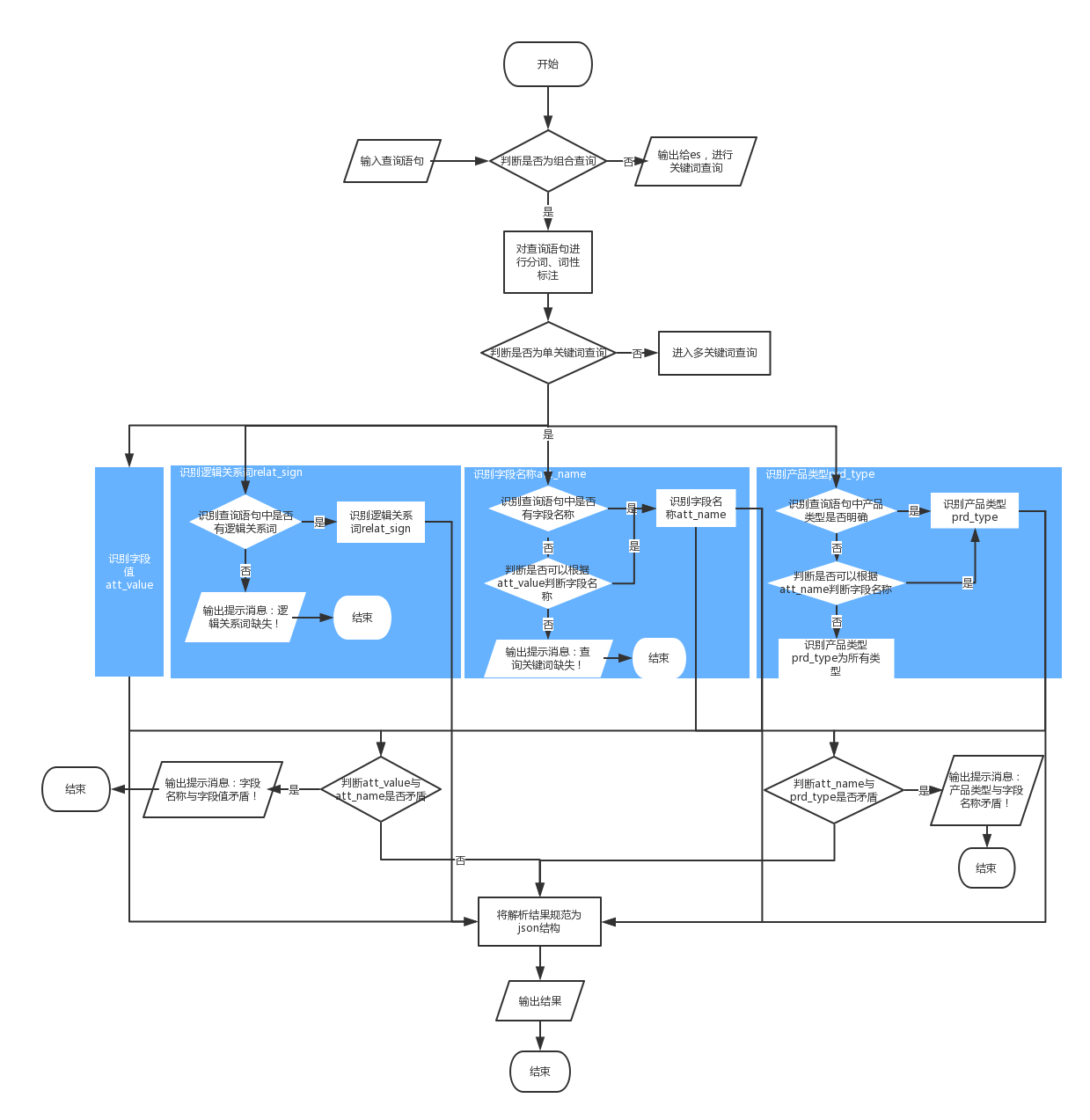
单关键词查询指针对产品的一个字段进行查询，重点是从查询语句中解析出查询的四个维度的信息，对无法解析的语句给出用户相应的提示信息。

### 实现方法

单关键词查询的实现主要包含以下几个步骤：

1. 识别查询语句中的字段值att\_value；
2. 识别逻辑关系词relat\_sign；
3. 识别字段名称att\_name；
4. 识别产品类别prd\_type；
5. 判断字段值与字段名称、字段名称与产品类别是否矛盾，若矛盾，输出提示信息，若不矛盾，转到下一步；
6. 组织输出结果。

单关键词查询实现流程如下：



对实现中的一些细节做简单说明：

要求查询语句至少要包含字段值和关系词，若不包含字段值，将查询语句抛给es做关键词查询，若有字段值，没有逻辑关系词，则提示用户补全信息。

不同字段名称的解析方法如下：

1. 收益率：
2. 字段名称、产品类型、关系词、字段值均包含，且之间不互相矛盾
3. 不包含产品类型，但可通过字段名称确定产品类型

如：不包含产品类型或是产品类型不明确，但字段名称为七日年化收益或万份收益，则可判断产品类型为货币基金；字段名称为3个月涨幅、6个月涨幅或12个月涨幅，则可判断产品类型为纯债基金；字段名称为预期收益率，则可判断产品类型为理财产品。

1. 不包含产品类型和字段名称的情形

如：大于5%，则返回所有产品类型的收益率

1. 锁定期：
   1. 字段名称、产品类型、关系词、字段值均包含，且之间不互相矛盾
   2. 不包含产品类型的情形，产品类型直接返回货币基金
2. 理财期限：
   1. 字段名称、产品类型、关系词、字段值均包含，且之间不互相矛盾
   2. 不包含产品类型的情形，产品类型直接返回理财产品
3. 起购金额：
   1. 字段名称、产品类型、关系词、字段值均包含，且之间不互相矛盾
   2. 不包含产品类型，其他字段包含，产品类型返回所有产品类型
   3. 不包含产品类型和字段名称的情形

如：小于5万的产品，解析为产品类型为3种，字段名称解析为起购金额，具体返回结果见测试结果

**说明**：暂不支持单关键字的多条件查询，如收益率大于4%小于6%的产品，锁定期大于1天小于10天的产品

对查询语句表达不清晰的各种情况进行了总结，如下：

1、字段名与产品类型矛盾：提示用户矛盾信息

如：锁定期小于2天的理财产品

解析：锁定期只有货币基金类有，而理财产品没有，锁定期与理财产品矛盾

2、字段名与字段值矛盾

如：收益率大于30天的产品

解析：30天为锁定期或理财期限的值，不是收益率的值

3、省略关系词：提示用户输入查询逻辑关系词

“大于”、“小于”等关系词缺省

4、省略字段名称，且无法根据字段值解析

如：大于4的产品

### 实例验证

本文档给出了26条查询样例，如下，来验证上述实现方法的识别准确率。

1. 收益大于6%的产品

2、收益率大于6%的理财产品

3、收益率大于6%的货币基金

4、收益率大于6%的宝宝类产品

5、收益率大于6%的纯债基金

6、预期收益率大于6%的产品

7、七日年化收益率大于6%的产品

8、收益率不低于6%

9、收益率大于等于百分之六（修改结巴原始词库后可以识别）

10、大于6%的产品

11、大于6.2%的理财产品

12、大于6%的货币基金

13、万份收益大于1.2的产品

14、六个月涨幅大于6%的产品

15、收益率大于5%的货币基金、纯债基金

16、十二个月涨幅不少于5.4%的产品

17、12个月涨幅大于5.4%的产品

18、锁定期小于1天的宝宝类产品

19、锁定期小于1天的货币基金

20、锁定期小于一天的产品

21、锁定期小于1天

22、期限小于30天的理财产品

23、理财期限小于等于三十天的产品

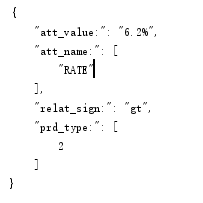
24、起购金额小于5万的理财产品

25、起购金额小于六万的产品

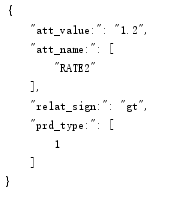
26、小于5万的产品

具体例子

1、大于6.2%的理财产品



2、万份收益大于1.2的产品



3、大于1.2的产品



通过对上述26条样例的验证，单关键词查询的识别正确率为100%，读者可提供其他样例来验证单关键词查询的实现方法。

## 多关键词查询

多关键词查询指针对产品的多个字段进行查询，重点是将多关键词查询切分为单关键词查询，并用单关键词查询的解析方法来解析每个单关键词查询，最后再做合并处理。

### 实现方法

由于多关键词查询语意复杂性，多关键词查询尝试了两种实现方法，一个是依存句法分析，一个是根据具体业务场景和业务规则来实现。

#### 依存句法分析

根据依存句法分析，判断词与词之间的依赖关系，据此解析查询语句。

尝试使用了三种依存句法分析工具，分别为deepnlp、huannlp和哈工大语言云(LTP)。三种工具分别使用不同的方法实现依存句法分析，其中deepnlp利用深度学习的方法，Huannlp利用crf方法，LTP将依存分析问题归结为在一个有向图中寻找最大生成树（Maximum Spanning Tree）的问题。在依存句法分析模块中，LTP分别实现了

* 一阶解码(1o)
* 二阶利用子孙信息解码(2o-sib)
* 二阶利用子孙和父子信息(2o-carreras)

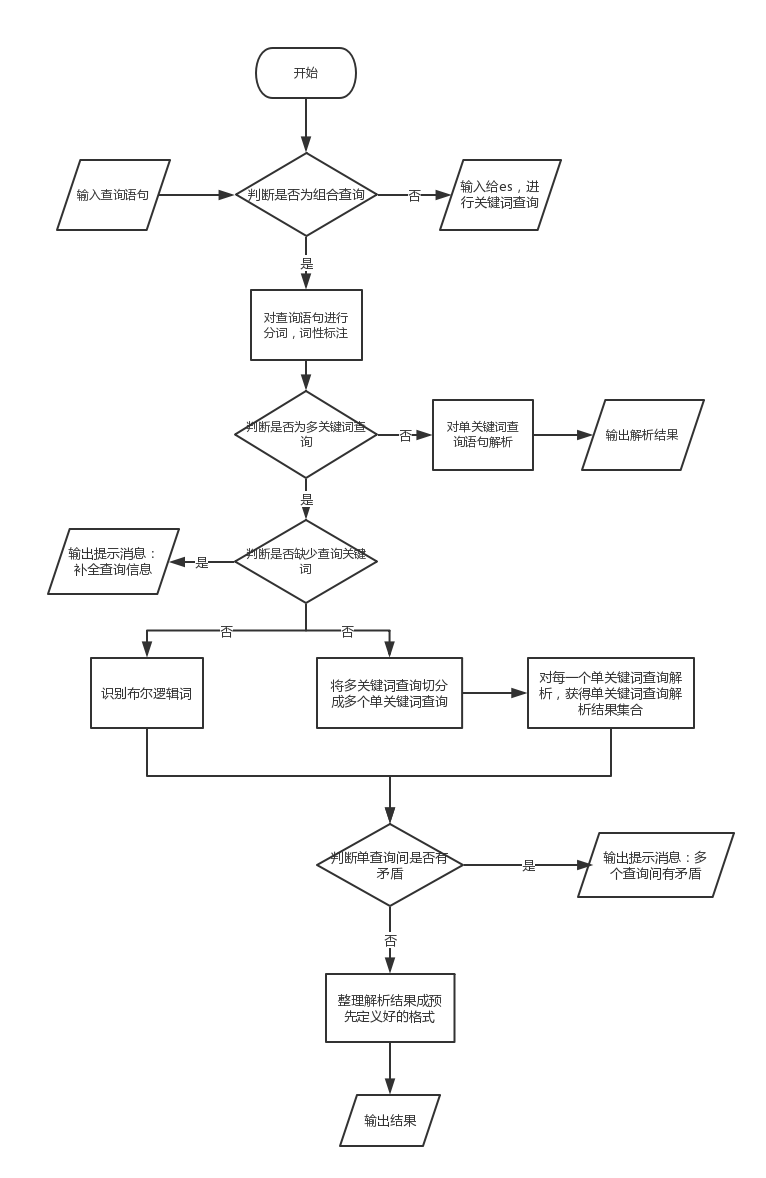
依存句法分析不能处理省略关键词的情况。

#### 根据业务规则实现方法

根据业务规则的实现方法主要分为以下几个步骤：

1. 判断是否为多关键词查询，若否，转到2，若是，转到3；
2. 单关键词查询解析；
3. 判断关键词是否省略，若否，转到5，若是，转到4；
4. 输出提示信息：请补全关键词信息；
5. 将多关键词查询语句切分成多个单关键词查询，并转到2；
6. 合并多个单关键词查询的结果。

流程图如下：



下面对多关键词查询实现过程中的一些细节做简要说明：

对判断是否为多关键词查询的方法说明：

根据查询语句中的字段名称个数和字段值个数来判定，若count(att\_name)<=1 且

count(att\_value)=1，为单关键词查询；若count(att\_nam)> 1或count(att\_value)>1，且count(att\_nam)<=count(att\_value)，则为多关键词查询。

识别布尔逻辑

若有明显表明and的词（如“和”，“且”，“而且”），则布尔逻辑为and

否则为or（可能的情形，包含明显表明or的词（如“或”），省略标志词，都默认为or类型）。

将查询语句切成多个单一关键词查询语句

几种情况：

1）有明显的分隔标识，如“，”、“或”、“和”、空格等，直接用分隔标识分隔

1. 没有明显的分隔标识

att\_name和att\_value都有

att\_name省略：根据att\_value识别

att\_value省略：无法继续，提示用户补全信息

多个单一查询结果产品类型是否矛盾判断

or：可以出现单一关键词查询结果中产品类型不同

and：不能出现上面or中的情况，返回结果为空

### 实例验证

选取20条样例，分别用上述两种方式进行解析。样例如下。

1. 万份收益大于1.3或锁定期大于3天的产品（1）
2. 万份收益大于1.3和锁定期大于3天的产品（1）
3. 锁定期大于1.3 收益率大于3天的产品（1）
4. 锁定期大于1.3，收益率大于3天的产品（1，3）
5. 万份收益大于1.3、锁定期大于3天的产品
6. 万份收益大于1.3，锁定期大于3天的产品
7. 万份收益大于1.3 锁定期大于3天的产品
8. 收益率大于5%或理财期限小于30天的产品
9. 万份收益大于1.3、锁定期大于3天、理财期限小于30天的产品
10. 收益率大于5%、起购金额小于5万
11. 收益率大于5%起购金额小于5万（1,2）
12. 收益率大于5%且起购金额小于5万（1,2）
13. 收益率大于5% 起购金额小于5万（1,2）
14. 大于5% 起购金额小于5万（1,2）
15. 收益率大于5%或起购金额小于5万（2）
16. 大于5%、理财期限小于30天（2）
17. 万份收益、起购金额大于1的产品（2）
18. 大于5%、小于30天的理财产品
19. 大于5%小于30天的产品（关键词省略 分割失败）
20. 锁定期和理财期限小于30天

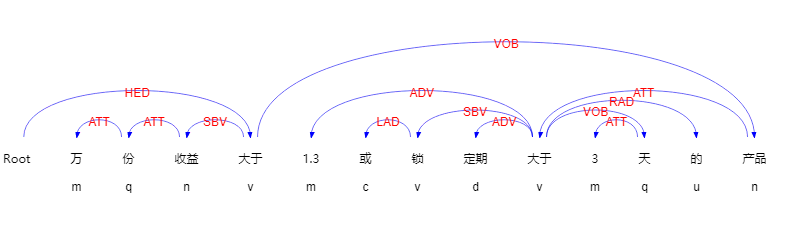
#### 依存句法分析

样例后面序号代表对应的工具可识别，序号含义为deep：1；huan：2；ltp：3

样例具体解析结果如下例子。

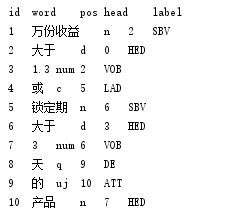
例子：万份收益大于1.3或锁定期大于3天的产品

哈工大语言云分析结果：



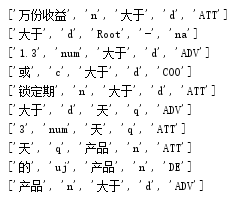
前半部分解析错误，“大于”宾语应为“1.3”

* deepnlp分析结果：



解析正确

* huannlp分析结果：



“或”后面部分解析错误

对20条多关键词查询语句进行验证，三种工具对其解析的正确率如下

|  |  |
| --- | --- |
| 依存句法工具 | 正确率 |
| 哈工大语言云 | 5% |
| deepnlp | 40% |
| huannlp | 35% |

综上，三种工具效果均不理想，主要表现在对关键词敏感，对标点符号不识别，和一些其他原因，导致解析结果不理想。

#### 根据业务规则实现方法

同样对上述20条样例进行测试，正确率为：80%（红色字体的样例为识别错误的）

#### 实验总结

在实现效率和准确率上，根据业务规则实现的方法更具优势。

### 说明

目前不能识别的类型：

单关键词多条件查询，如“收益率大于4%小于6%的产品”

多关键词单条件查询，如“万份收益、起购金额大于2的产品”

多关键词多布尔逻辑词查询，如“收益率大于5%且起购金额小于5万或锁定期小于3天的产品”

后面继续研究实现方法。