**UDH Search去重与分组**

作者： 费英林

目录

[1. 索引数据去重 2](#_Toc421699126)

[1.1. 修改solrconfig.xml 2](#_Toc421699127)

[1.2. 修改schema.xml 3](#_Toc421699128)

[1.3. 验证 3](#_Toc421699129)

[2. 索引数据分组 3](#_Toc421699130)

[2.1. 修改schema.xml 3](#_Toc421699131)

[2.2. 分组示例 3](#_Toc421699132)

# 索引数据去重

去重的目的是阻止重复或近似重复的文档进入索引数据。

## 修改solrconfig.xml

下面是一个配置实例。字段signatureID用做一个文档的Unique ID，它是一个虚拟字段，代表字段data\_f1与data\_f2的组合，data\_f1与data\_f2是实际存在的字段。即data\_f1与data\_f2组合的值必须是唯一的，如果重复，已在索引中的数据会被覆盖。

<updateRequestProcessorChain name="**dedupe**">

<processor class="solr.processor.SignatureUpdateProcessorFactory">

<bool name="enabled">true</bool>

**<str name="signatureField">signatureID</str>**

<bool name="overwriteDupes">false</bool>

**<str name="fields">data\_f1,data\_f2</str>**

<str name="signatureClass">solr.processor.Lookup3Signature</str>

</processor>

<processor class="solr.LogUpdateProcessorFactory" />

<processor class="solr.RunUpdateProcessorFactory" />

</updateRequestProcessorChain>

启用去重：

<requestHandler name="/update" class="solr.UpdateRequestHandler">

<lst name="defaults">

<str name="update.chain">**dedupe**</str>

</lst>

</requestHandler>

## 修改schema.xml

将uniqueKey设置为signatureID（原值是ID，需要替换掉）：

<uniqueKey>signatureID</uniqueKey>

在fields属性组中添加属性signatureID：

<field name="signatureID" type="string" stored="true" indexed="true" required="true" multiValued="false" />

修改fields属性组中的属性ID，将required由true改为false：

<field name="id" type="string" indexed="true" stored="true" required="false" multiValued="false" />

## 验证

重建索引，查看结果，确认：

* 结果中是否存在重复数据
* 后面的重复数据会覆盖前面的数据

# 索引数据分组

## 修改schema.xml

分组的字段要求不能是tokenized的字段，需要定义为string类型。假如我们需要依据other\_articles\_title进行分组，由于other\_articles\_title是定义为中文分词类型的，直接使用这个字段无法实现分组的效果。我们需要另建一个字段other\_articles\_title\_nt用于分组：

<field name="other\_articles\_title" type="text\_ik" indexed="true" stored="true"/>

<field name="other\_articles\_title\_nt" type="**string**" indexed="true" stored="true"/>

## 分组示例

select?q=\*%3A\*&group=true&group.field=other\_articles\_title\_nt

查询条件为任意字段的任意值，设置group为true，分组字段为other\_articles\_title\_nt，每组默认显示分值最高的1条记录，默认显示10个分组。

select?q=\*%3A\*&group=true&group.field=other\_articles\_title\_nt&group.limit=3

同上。每个分组内显示前3条分值最高的记录。

select?q=\*%3A\*&group=true&group.field=other\_articles\_title\_nt&rows=100

同上。每个分组内显示分值最高的1条记录，显示100个分组。

select?q=\*%3A\*&group=true&group.field=other\_articles\_title\_nt&group.limit=3&rows=100

同上。每个分组内显示分值最高的3条记录，显示100个分组。

select?q=other\_articles\_sitename:"sina"&group=true&group.field=other\_articles\_title\_nt&group.limit=3&rows=100

查询条件为site name是新浪，设置group为true，分组字段为other\_articles\_title\_nt，每组默认显示分值最高的3条记录，显示100个分组。

select?q=other\_articles\_sitename:"sina"&group=true&group.field=other\_articles\_title\_nt&group.limit=3&rows=100&group.ngroups=true

同上。显示分组个数。

select?q=other\_articles\_sitename:"sina" AND other\_articles\_source:"新浪科技"&group=true&group.field=other\_articles\_title\_nt&group.limit=3&rows=100

（或者

select?q=other\_articles\_sitename:%22sina%22%20AND%20other\_articles\_source:%22%E6%96%B0%E6%B5%AA%E7%A7%91%E6%8A%80%22&group=true&group.field=other\_articles\_title\_nt&group.limit=3&rows=100）

查询条件为site name是新浪并且来源是新浪科技，设置group为true，分组字段为other\_articles\_title\_nt，每组默认显示分值最高的3条记录，显示100个分组。