

# 实验 4：Spark-SQL 分布式数据库查询-推荐系统

22121630 汪江豪

**实验内容：**对于给定的数据集，查询任意两个网站之间的相似度。

**实验要求：**使用 5 种查询方法：连接查询、嵌套不相关子查询、嵌套相关子查询、intersect 操作，exists 操作。按相似度排序，返回相似度最高的前 20 个。并将最后 5 种方式耗时做成一个表。

**实验步骤：**

- 准备工作：在腾讯云上启动三个实验三实例；修改三台节点名称为 master, slave01, slave02，在/etc/hosts 中添加三台机器的内网 ip 和主机名，重启。
- 启动 hadoop 分布式文件系统：在 hadoop 目录下，执行 sbin/start-all.sh。
- 启动 spark-sql，设定 slave01, slave02 为工作节点：在 spark 目录下：  
bin/spark-sql –master yarn –num-executors 2  
在此之前，需要将 hadoop 目录下 etc/hadoop 添加到 spark/conf/spark-env.sh 中。
- 导入数据集，创建表：

```
create table relation(
    referrer int,
    referree int
)
using csv
options(
    path "./small_relation",
    delimiter ' ',
    header 'false'
);
```

- 执行 5 种查询操作：

- 思路：创建两个临时表，ref\_count 记录一个网站的引用标签数；common 记录任意两个网站的公共标签数，再根据这两个临时表进行查询，计算相似度。5 种 sql 可在 common 临时表或主查询中实现。

## 5.1 连接查询

- sql 语句：

```
25/02/25 14:21:34 INFO CliDriver: Time taken: 12.043 seconds, Fetched 20 row(s)
spark-sql> with ref_count as (select referrer, count(distinct referree) as web_count
>           from relation
>          group by referrer),
> common as (select a.referrer as web1, b.referrer as web2, count(*) as com_cnt
>            from relation a join relation b
>           on a.referree = b.referree
>          where a.referrer < b.referrer
>          group by a.referrer, b.referrer)
> select common.web1 as web1, common.web2 as web2,
> case when r1.web_count + r2.web_count - common.com_cnt > 0
> then common.com_cnt / (r1.web_count + r2.web_count - common.com_cnt)
> else 0
> end as similarity
> from common join ref_count r1 on common.web1 = r1.referrer
> join ref_count r2 on common.web2 = r2.referrer
> order by similarity desc limit 20;
```

- 运行结果：

```

25/02/25 14:26:19 INFO scheduler.DAGScheduler: Job 17 finished: processCmd at CliDriver.java:376, took 1.672420 s
3867 3366 1.0
327 334 1.0
788 814 1.0
3724 3926 1.0
3849 3808 1.0
668 878 1.0
3936 3944 1.0
1861 1866 1.0
3755 3891 1.0
3412 3420 1.0
3318 3391 1.0
1878 1887 1.0
2442 2549 1.0
3872 3913 1.0
4820 4822 1.0
1874 1885 1.0
190 246 1.0
1768 1882 1.0
887 889 1.0
1877 1985 1.0
Time taken: 10.802 seconds. Fetched 20 row(s)
25/02/25 14:26:19 INFO CliDriver: Time taken: 10.802 seconds, Fetched 20 row(s)
spark> |

```

耗时 10.802s

## 5.2 嵌套相关子查询:

- sql 语句:

```

25/02/25 14:26:18 INFO CliDriver: Time taken: 10.802 seconds, Fetched 20 row(s)
spark> WITH ref_count AS (
>     SELECTreferrer, COUNT(DISTINCT referree) AS web_count
>     FROM relation
>     GROUP BY referrer
> ),
> common AS (
>     SELECT
>         a.referrer AS web1,
>         b.referrer AS web2,
>         COUNT(*) AS com_cnt,
>         min((SELECT web_count FROM ref_count WHERE referrer = a.referrer)) AS web1_count,
>         min((SELECT web_count FROM ref_count WHERE referrer = b.referrer)) AS web2_count
>     FROM relation a
>     JOIN relation b ON a.referee = b.referee
>     WHERE a.referrer < b.referrer
>     GROUP BY a.referrer, b.referrer
> )
>     SELECT
>         common.web1,
>         common.web2,
>         CASE
>             WHEN common.web1_count + common.web2_count - common.com_cnt > 0
>                 THEN common.com_cnt / (common.web1_count + common.web2_count - common.com_cnt)
>             ELSE 0
>         END AS similarity
>     FROM common
>     ORDER BY similarity DESC
>     LIMIT 20;|

```

- 运行结果:

```

25/02/25 14:27:12 INFO scheduler.DAGScheduler: Job 20 finished: processCmd at CliDriver.java:376, took 2.108434 s
1277 1353 1.0
3324 3387 1.0
2531 2544 1.0
1263 1560 1.0
789 792 1.0
218 239 1.0
3856 3907 1.0
1296 1501 1.0
3324 3346 1.0
3745 3766 1.0
3344 3397 1.0
2514 2516 1.0
1895 1908 1.0
3692 3703 1.0
449 457 1.0
2799 2898 1.0
277 384 1.0
1263 1678 1.0
1874 1885 1.0
4820 4822 1.0
Time taken: 7.789 seconds, Fetched 20 row(s)
25/02/25 14:27:12 INFO CliDriver: Time taken: 7.789 seconds, Fetched 20 row(s)
spark> |

```

耗时 7.789s

## 5.3 嵌套不相关子查询:

- sql 语句:

```

25/02/25 14:27:12 INFO CliDriver: Time taken: 7.789 seconds, Fetched 20 row(s)
spark-sql> WITH ref_count AS (
    >     SELECT referrer, COUNT(DISTINCT referree) AS web_count
    >     FROM relation
    >     GROUP BY referrer
    > ),
    > common AS (
    >     SELECT a.referrer AS web1, b.referrer AS web2, COUNT(*) AS com_cnt
    >     FROM relation a
    >     JOIN relation b ON a.referree = b.referree
    >     WHERE a.referrer < b.referrer
    >     GROUP BY a.referrer, b.referrer
    > )
    >     SELECT common.web1, common.web2,
    >     CASE WHEN r1.web_count + r2.web_count - common.com_cnt > 0
    >     THEN common.com_cnt / (r1.web_count + r2.web_count - common.com_cnt)
    >     ELSE 0 END AS similarity
    >     FROM common JOIN (
    >         SELECT referrer, web_count
    >         FROM ref_count
    >         WHERE referrer IN (SELECT web1 FROM common)
    >     ) r1 ON common.web1 = r1.referrer
    >     JOIN (
    >         SELECT referrer, web_count
    >         FROM ref_count
    >         WHERE referrer IN (SELECT web2 FROM common)
    >     ) r2 ON common.web2 = r2.referrer ORDER BY similarity DESC LIMIT 20;

```

- 运行结果:

```

25/02/25 14:28:05 INFO scheduler.DAGScheduler: Job 24 finished: processCmd at CliDriver.java:376, took 1.577346 s
3357 3360 1.0
327 334 1.0
708 814 1.0
3724 3926 1.0
3849 3880 1.0
868 878 1.0
3930 3944 1.0
1861 1866 1.0
3755 3891 1.0
3412 3420 1.0
3318 3391 1.0
1878 1887 1.0
2442 2549 1.0
3872 3913 1.0
4020 4022 1.0
1874 1885 1.0
194 246 1.0
1768 1862 1.0
887 889 1.0
1877 1985 1.0
Time taken: 12.874 seconds, Fetched 20 row(s)
25/02/25 14:28:05 INFO CliDriver: Time taken: 12.874 seconds, Fetched 20 row(s)
spark-sql> |

```

耗时 12.874s

#### 5.4 intersect 操作:

- sql 语句:

```

25/02/25 14:28:05 INFO CliDriver: Time taken: 12.874 seconds, Fetched 20 row(s)
spark-sql> WITH ref_count AS (
    >     SELECT referrer, COUNT(DISTINCT referree) AS web_count
    >     FROM relation GROUP BY referrer),
    > common AS (SELECT a.referrer AS web1, b.referrer AS web2, COUNT(*) AS com_cnt
    >     FROM relation a JOIN relation b
    >     ON a.referree = b.referree
    >     WHERE a.referrer < b.referrer
    >     GROUP BY a.referrer, b.referrer)
    >     SELECT common.web1, common.web2,
    >     CASE WHEN r1.web_count + r2.web_count - common.com_cnt > 0
    >     THEN common.com_cnt / (r1.web_count + r2.web_count - common.com_cnt)
    >     ELSE 0 END AS similarity
    >     FROM common JOIN (
    >         SELECT referrer, web_count FROM ref_count
    >         WHERE referrer IN (
    >             SELECT web1 FROM common
    >             INTERSECT
    >             SELECT referrer FROM ref_count)
    >     ) r1 ON common.web1 = r1.referrer
    >     JOIN (
    >         SELECT referrer, web_count
    >         FROM ref_count
    >         WHERE referrer IN (
    >             SELECT web2 FROM common
    >             INTERSECT
    >             SELECT referrer FROM ref_count)
    >     ) r2 ON common.web2 = r2.referrer ORDER BY similarity DESC LIMIT 20;

```

- 运行结果:

```

25/02/25 14:29:03 INFO scheduler.DAGScheduler: Job 29 finished: processCmd at CliDriver.java:376, took 1.646812 s
3357 3360 1.0
327 334 1.0
700 814 1.0
3724 3926 1.0
3849 3880 1.0
868 878 1.0
3938 3944 1.0
1861 1866 1.0
3755 3891 1.0
3412 3420 1.0
3318 3391 1.0
1878 1887 1.0
2442 2549 1.0
3872 3913 1.0
4020 4022 1.0
1874 1885 1.0
194 246 1.0
1768 1882 1.0
887 889 1.0
1877 1985 1.0
Time taken: 14.002 seconds, Fetched 20 row(s)
25/02/25 14:29:03 INFO CliDriver: Time taken: 14.002 seconds, Fetched 20 row(s)
spark-sql> |

```

耗时 14.002s

5.5 exists 方式:

- sql 语句:

```

25/02/25 14:29:03 INFO CliDriver: Time taken: 14.002 seconds, Fetched 20 row(s)
spark-sql> WITH ref_count AS (
>     SELECT referrer, COUNT(DISTINCT referree) AS web_count
>     FROM relation
>     GROUP BY referrer
> ),
> common AS (
>     SELECT a.referrer AS web1, b.referrer AS web2, COUNT(*) AS com_cnt
>     FROM relation a
>     JOIN relation b ON a.referee = b.referee
>     WHERE a.referrer < b.referrer
>     GROUP BY a.referrer, b.referrer
> )
> SELECT
>     common.web1,
>     common.web2,
>     CASE
>         WHEN r1.web_count + r2.web_count - common.com_cnt > 0
>             THEN common.com_cnt / (r1.web_count + r2.web_count - common.com_cnt)
>         ELSE 0
>     END AS similarity
> FROM common JOIN (
>     SELECT referrer, web_count
>     FROM ref_count r
>     WHERE EXISTS (SELECT 1 FROM common c WHERE c.web1 = r.referrer)
> ) r1 ON common.web1 = r1.referrer
> JOIN (
>     SELECT referrer, web_count
>     FROM ref_count r
>     WHERE EXISTS (SELECT 1 FROM common c WHERE c.web2 = r.referrer)
> ) r2 ON common.web2 = r2.referrer
> ORDER BY similarity DESC LIMIT 20;

```

- 运行结果:

```

25/02/25 14:30:03 INFO scheduler.DMGScheduler: Job 33 finished: processCmd at CliDriver.java:376, took 1.543978 s
3357 3360 1.0
327 334 1.0
708 814 1.0
3724 3926 1.0
3849 3880 1.0
868 878 1.0
3930 3944 1.0
1861 1866 1.0
3755 3891 1.0
3812 3928 1.0
3318 3391 1.0
1878 1887 1.0
2442 2549 1.0
3872 3913 1.0
4620 4622 1.0
1874 1885 1.0
194 246 1.0
1768 1882 1.0
887 889 1.0
1877 1985 1.0
Time taken: 11.885 seconds, Fetched 28 row(s)
25/02/25 14:30:03 INFO CliDriver: Time taken: 11.885 seconds, Fetched 28 row(s)
spark> |

```

耗时 11.885s

结果比较：

方法	耗时/s
连接查询	10.802
嵌套相关子查询	7.789
嵌套不相关子查询	12.874
intersect 操作	14.002
exists 方式	11.885