TU Berlin

Advanced Information Management

Homework assignment 3

Recommender Systems, stream Processing, Network Analysis, and Log Analysis

Author:
Ward Schodts

Supervisor:
Juan Soto

July 16, 2016

1 Recommender Systems: Item-based Collaborative Filtering

See hand-written solution.

2 Collaborative Filtering

See hand-written solution.

3 Stream Processing

3.1 Foundations

What are the key differences between batch and stream data sources?

- Size: stream data sources can be ongoing forever while batch has limited size.
- Data availability: if you have a fixed batch dataset you know you can acces all the items in this batch. In streams this is not always true, you can get data for a while after which it stops again for while.

What problem will we encounter, if blocking operators, such as aggregations are used in streaming queries?

For these operators you need to see the whole dataset before you can give an answer. With streaming you never see the whole dataset and thus there will be no output.

How do streaming engines address the problem referenced just above? Hint: Examine the aforementioned Word Count examples.

They use methods such as maps to produce an answer for the already seen data. For example if you want the average of a stream they keep the total sum of already seen items and the amount of already seen items. So you can then return whenever you want the average up to that point.

Batch processing engines like Apache Spark simulate streaming behavior using discretized streams (D-Streams). What is the main idea behind D-Streams?

"The idea in D-Streams is to structure a streaming computation as a series of stateless, deterministic batch computations on small time intervals".[1]

What are the disadvantages attributed to using D-Streams on a batch processing engine, as opposed to using a runtime engine that offers native streaming support?

- Computational overhad due to spiltting up in DStreams
- Some application are time critical and need action immediately, they can't even wait until the micro batch is processed.

4 The PageRank Algorithm

See hand-written solutions.

5 Weblog Analysis

A description of what the original WebLogAnalysis.java class does (i.e. asking the question that is answered by it)

```
SELECT
       r.pageURL,
       r.pageRank,
       r.avgDuration
FROM documents d JOIN rankings r
                  ON d.url = r.pageURL
WHERE CONTAINS(d.contents, [keywords])
       AND r.pageRank > [rank]
       AND NOT EXISTS
           (
              SELECT * FROM Visits v
              WHERE v.destURL = d.url
                    AND v.visitDate = [date]
           );
```

It asks:
Which are the pages (URL, pageRank, avgDuration) that contain the words oscillation and editors and a pagerank higher than 40 that are not visited in 2007?

The answer/output is:

```
(69,url_16,34)
(41,url_39,33)
(48,url_46,37)
```

The answers to the above asked questions, including the output of your jobs.

Question 1: "Which IPs visit what webpages with a rank bigger than 90 that utilize the term 'Lorem' ?"

```
SELECT v.sourceIP, v.destURL
FROM visits v JOIN
      (
      SELECT
          r.pageURL
      FROM documents d JOIN rankings r
                  ON d.url = r.pageURL
      WHERE CONTAINS(d.contents, [keywords])
             AND r.pageRank > [rank]
      ) ON v.destURL = pageURL
```

The answer/output is:

```
(ip_5,url_30)
(ip_10,url_30)
```

Question 2: "What is the rank for the most visited website based on IPs that do not originate in Germany?"

I assumed with this question that most visited meant the amount of visits and not the total spend time for "most visited".

```
SELECT
      r.pageRank,
FROM rankings r JOIN
     (
```

```
SELECT destURL, count(destURL) AS pageHits
FROM Visits v
WHERE
    v.languageCode <> ['DE']
GROUP BY destURL
ORDER BY pageHits DESC
LIMIT 1
)
ON destURL = r.pageURL
```

The answer/output is:

```
(9)
```

This is the score from URL_9. As the query asked only the pageRank this is the only output.

Two ideas for other questions that can be answered using the provided data set.

1. What is the most popular (most visited) URL per country?
2. Does the page with highest pagerank equal the page with the most visits?

References

[1] Matei Zaharia et al. "Discretized streams: an efficient and fault-tolerant model for stream processing on large clusters". In: *Presented as part of the*. 2012.