Wikipedia PageRank

INFM2600 - Big Data- Wintersemester 2020/2021

Autoren: Andreas Voigt, Ricardo Lissner, Rituraj Singh
andreas.voigt@hochschule-stralsund.de
ricardo.lissner@hochschule-stralsund.de
rituraj.singh@hochschule-stralsund.de
Matrikel-Nr. (Andreas Voigt): 19398
Matrikel-Nr. (Ricardo Lissner): 19134
Matrikel-Nr. (Rituraj Singh): 19539

Eingereicht am: 10.02.2021

Eingereicht bei: Sebastian Plotz

Zusammenfassung

Dieses Dokument beschreibt das Vorgehen und die Lösung zur Aufgabe *„Wikipedia PageRank“*. Ziel ist es, die Beitragsseiten der Wikipedia Deutschland mit Hilfe des PageRank Algorithmus zu bewerten und abzuspeichern. Dies ist mit Hilfe einer Java-Implementation und der Apache Spark Engine und Hadoop erfolgt. Der Quellcode befindet in der separaten *„WikipediaPageRank.java“* Datei. Eine alternative Lösung befindet sich in der Datei *„WikipediaPageRankWithoutThirdParty.java*“. In diesem Dokument wird auf die relevanten Codebausteine eingegangen und diese näher erläutert.

1 INHALTSVERZEICHNIS

2 KURZE THEORETISCHE GRUNDLAGEN PAGERANK UND WIKIPEDIA

2.1 PAGERANK ALGORITHMUS

„Der PageRank-Algorithmus ist ein Verfahren, eine Menge verlinkter Dokumente, [..], anhand ihrer Struktur zu bewerten und zu gewichten. Dabei wird jedem Element ein Gewicht, der PageRank, aufgrund seiner Verlinkungsstruktur zugeordnet.“[1]

Mit Bezug auf Wikipedia bedeutet das, dass jedes Element im PageRank hierbei einer Wikipedia-Beitragsseite entspricht. Die Verlinkungen der Beitragsseiten erfolgen über so genannte Wiki-Links. Die Berechnung des Algorithmus erfolgt mit der Formel:

$$PR(p_i) = (1-d) + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

- **PR(pi)** – PageRank der Webseite pi (Initial bei 1)
- **D** – Dämpfungsfaktor (zwischen 0 und 1, hier 0,85)
- **M(pi)** – Menge der Webseiten, die auf pi verlinken
- **L(pj)** – Anzahl der von pj ausgehenden Links

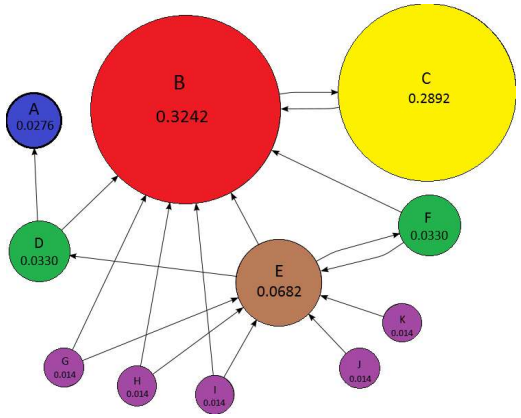

*Abbildung 1*https://de.wikipedia.org/wiki/Datei:PageRank-Beispiel.png (Stand: 22. Juli 2012)*

[1] https://de.wikipedia.org/wiki/PageRank (Stand: 10. Dezember 2020 um 16:47 Uhr)

2.2 WIKIPEDIA

Die Wikipedia Online Enzyklopädie umfasst eine Vielzahl an Sprachen. Für die Anwendung des Algorithmus wird sich auf die deutschsprachige Wikipedia beschränkt. Die Rohdaten der deutschen Wikipedia lassen sich als Komplettpaket in XML-Form herunterladen. Die Struktur der XML wird ebenfalls auf Wikipedia selbst erläutert.[2]

```
...
<page>
        <title>PageRank</title>
        <ns>0</ns>
        <id>963351</id>
        <revision>
                <id>206412979</id>
                <parentid>205799606</parentid>
                <timestamp>2020-12-10T15:47:32Z</timestamp>
                <contributor>
                        <username>Aka</username>
                        <id>568</id>
                </contributor>
                <minor/>
                <comment>/* Geschichte */ [[Benutzer:Aka/Tippfehler entfernt|Tippfehler
entfernt]]</comment>
                <model>wikitext</model>
                <format>text/x-wiki</format>
                <textbytes="15131" xml:space="preserve">Der '''PageRank'''-[[Algorithmus
]] ist ein Verfahren, eine Menge [[Hyperlink|verlinkter]] Dokumente,
...</text>
                <sha1>7ldudscuwewsl27ikqstocwo44g7s3a</sha1>
        </revision>
</page>
...
```

Auszug 1 Wikipedia Komplett-XML

Für den PageRank sind folgende XML-Knoten relevant:

- **page**: Enthält alle Daten einer Seite
- **title**: Titel der Seite
- **redirect**: Weiterleitung auf den Artikel der die vollständige Beschreibung enthält
- **ns**: Namespace der Seite. wird zur Gruppierung der Seiten verwendet. Hier wird nur der Namespace „0“ (Artikel)[3] betrachtet.
- **text**: Enthält den Inhalt der Seite. Dies umfasst auch die Wikilinks auf andere Artikel, welche für die Berechnung des PageRanks benötigt werden.

Wiki-Links:

Die Links auf andere Artikel folgen einem definierten Format:

[[Zielseite#Abschnitt|alternativer Text]]

Wobei „Abschnitt“ und „alternativer Text“ optional sind und für den PageRank nicht relevant. Wiki-Links die auf noch nicht vorhandene Seiten verlinken werden ebenfalls vom PageRank Algorithmus ausgeschlossen.

Redirect-Artikel:

Einige Artikel sind lediglich Weiterleitungen auf andere Artikel. Oft entsteht dies durch die Existenz mehrerer Synonyme für den gleichen Inhalt. Solche Weiterleitungenwerden aufgelöst, so dass Weiterleitungsartikel nicht im PageRank gewichtet werden.

[2]https://de.wikipedia.org/wiki/Wikipedia:Technik/Datenbank/Download
[3] Wikipedia, Hilfe:Namensräume", https://de.wikipedia.org/wiki/Hilfe:Namensr%C3%A4ume

3 IMPLEMENTATION WIKIPEDIA PAGERANK

Die Implementation erfolgte vollständig in Java. Für die Berechnung des PageRank-Wertes der Artikelseiten wurde die Lösung in einzelne Bereiche aufgegliedert die dann im Folgenden Erläutert werden.

1. Einlesen der XML-Datei
2. Filtern und Parsen der relevanten Daten aus den Artikeln
3. Lösen der Weiterleitungen
4. Links auf nicht existente Artikel herausfiltern
5. Element-Verlinkungen aufbereiten für den PageRank Algorithmus
6. Anwendung des PageRank Algorithmus
7. Speichern der Ergebnisse

3.1 ELEMENT-VERLINKUNG AUFBEREITEN UND ANWENDUNG PAGERANK ALGORITHMUS

Die Erläuterung beginnt hier mit beiden letzten Teilen, der Aufbereitung und dem PageRank Algorithmus, weil hier eine vorgefertigte Implementation aus einer früheren Arbeit genutzt wird. Es wird an dieser Stelle nicht näher auf den Algorithmus an sich eingegangen, sondern lediglich auf das Eingabeformat. Dieses stellt gleichzeitig das Ziel-Dataset für die vorherigen Teilschritte da. Der PageRank benötigt folgendes Eingabeformat für die Verlinkungen:

Element A(source) **-> Element B**(target)

Das bedeutet ein Datensatz besteht aus zwei Werten. Der erste Wert (source) gibt den Titel einer Wikipedia Seite an und der zweite Wert (target) ist der Link zu einer anderen Wikipedia Seite ausgehend von der ersten. Somit können viele Datensätze wie z.B. A -> B und A -> C enthalten sein. Diese Linkliste wird mit der Gesamtliste aller Artikel in den PageRank Algorithmus gegeben. In der Gesamtliste sind dann auch die Artikel enthalten die auf keine weiteren Artikel verlinken.

3.2 EINLESEN DER XML-DATEI

Apache Spark bietet nativ keine direkte Unterstützung für Textdateien mit einer XML-Struktur. Daher müssten XML-Dateien wohl als roher Text eingelesen und weiter aufgearbeitet werden. Für dieses Problem existiert jedoch die Third-Party Bibliothekspark-xml[4]. Der Einsatz einer solchen Bibliothek bedeutet eine Änderung der Abhängigkeiten in der „pom.xml“. Da die Nutzung von Third-Party Bibliotheken nicht ausgeschlossen wurde, wird diese folgend verwendet. Diese Nutzung solcher Bibliotheken bieten den Vorteil, dass sie in der Regel deutlich besser getestet sind und ihre korrekte Funktion schon in diversen Anwendungen unter Beweis stellen mussten.

```
<dependency>
<groupId>com.databricks</groupId>
<artifactId>spark-xml_2.12</artifactId>
<version>0.11.0</version>
</dependency>
```

Auszug 2pom.xml

An dieser Stelle wurde jedoch noch eine alternative Lösung erarbeitet die gänzlich mit den gelieferten Mitteln der Spark Engine auskommt. Die Lösung befindet sich in der separaten Datei *„WikipediaPageRankWithoutThirdParty.java“*. Die alternative Lösung wird jedoch im Folgenden nicht näher erläutert.

[4]https://github.com/databricks/spark-xml

Mit Hilfe der Bibliothek können nur auch XML formatierte Dateien gelesen werden. Sofern kein Schema definiert wurde, wird durch den internen XML-Parser ein bestmögliches zu erstellen. Hierbei wird die XML-Struktur in ein Dataset umgewandelt, wobei die Option „*rowTag*“ angibt bei welchem XML-Knoten ein neues Datensatz anfangen soll.

```
Dataset<Row>wikipediaDF = spark.read()
                                    .format("xml")
                                    .option("rowTag", "page")
                                    .option("mode", "FAILFAST")
                                    .load("wikipedia.xml");
```

Auszug 3Einlesen der Wikipedia XML-Datei

3.3 Filtern und Parsen der relevanten Daten aus den Artikeln

Der gelesen Datensatz wird zunächst nach dem relevanten Namespace „0“ geliefert und nur noch die relevanten Felder selektiert. Die Redirect-Artikel werden separat selektiert da sie für die Weiterleitungsauflösung benötigt werden.

```
Pattern wikilinksPattern = Pattern.compile("\\[\\[(.*?)[\\|#\\]].*?\\]?\\]");

Dataset<Row>redirections = wikipediaDF.filter("redirect._title IS NOTNULL")
.select(col("title"),col("redirect._title").as("redirectTo"));

Dataset<Row>articles = wikipediaDF.filter("ns == '0'").select(col("title"),col("revision.text").as("text"));

Dataset<Row> links = articles.flatMap((FlatMapFunction<Row, Row>) row -> {
                                    String text = (String) ((Row)row.get(1)).get(0);

                                    // find all linkingstootherpages
                                    List<Row>linksTo = newArrayList<>();
                                    MatcherwikilinksMatcher = wikilinksPattern.matcher(text);
                                    while(wikilinksMatcher.find()) {
                                             linksTo.add(RowFactory.create(row.get(0),
wikilinksMatcher.group(1)));
                                    }
                                    if(linksTo.isEmpty()) {
                                             // addtheactualpagewhich links tonoone
                                             linksTo.add(RowFactory.create(row.get(0), null));
                                    }

                           returnlinksTo.iterator();
                  }, RowEncoder.apply(linkSchema));
```

Auszug 4 Filtern und Parsen aller Page Blöcke

Jeder einzelne Artikeldatensatz wird nun mit Hilfe der flatmap()-Methode und regulären Ausdrücken analysiert. Hierbei werden die Verlinkungen von Artikeln zu anderen Artikeln extrahiert.

3.4 Lösen der Weiterleitungen

Um die Weiterleitungen in den Verlinkungen aufzulösen wird der vorhandene Datensatz zu den Weiterleitungen genutzt.Diese werden mit den ermittelten Verlinkungen mehrfach verknüpft (*„left*“ und *„leftantijoin*“), so dass alle Weiterleitungen korrekt aufgelöst und die Weiterleitungen selbst entfernt wurden. Die Gesamtliste der Wikipedia Artikel wird ebenfalls bereinigt um die Weiterleitungsartikel per *„SemiJoin*“

```
links = links
         .join(redirections, col("title").equalTo(col("target")),"left")
         .select(col("source"),
(when(col("redirectTo").isNull(),col("target")).otherwise(col("redirectTo")).as("target")))
         .join(redirections, col("source").equalTo(col("title")),"left_anti");
// Weiterleitungsartikel entfernen
articles = articles.join(links, col("source").equalTo(col("title")), "semi");
```

Auszug 5 Auflösen der Weiterleitungen

Dieses Vorgehen ist so nur möglich, da wir davon ausgehen dürfen, dass keine mehrfachen oder zyklischen Weiterleitungen existieren.

3.5 LINKS AUF NICHT EXISTENTE ARTIKEL HERAUSFILTERN

In den Verlinkungen wird noch auf Seiten verlinkt die nicht in der Gesamtliste der Artikel existent sind. Daher werden die Links mit den Artikeln verknüpft (*„semijoin“*) um die Verlinkungen zu bereinigen.

```
// Nicht existente Seiten filtern.
links = links.join(articles, col("title").equalTo(links.col("target")), "semi");
```

Auszug 6 Links auf nicht existente Seiten herausfiltern

3.6 SPEICHERN DER ERGEBNISSE

Nach Abschluss der PageRank Berechnung werden die Daten wieder in das HDFS im CVS Format gespeichert. Die Datensätze sind hierbei absteigend sortiert und durch ein Tabulatorzeichen getrennt.

```
pages.sort(desc("rank"))
         .write()
         .option("sep", "\t")
         .csv("result");
```

Auszug 7Speichern der Ergebnisse

4 ERGEBNISSE

Die Ergebnisse stehen, nachdem sie persistiert wurden, auch nach der Berechnung wieder schnell zur Verfügung. Wikipedia unterliegt naturgemäß einer freien Online Enzyklopädie einer ständigen Anpassung. Die Downloadbaren XML-Dateien werden wöchentlich aktualisiert. Daher beziehen sich sämtliche Ergebnisse und Berechnungen auf einen Abzug der Daten vom 22.01.2012.

Nachfolgend werden nur die ersten 20 Seiten – also die Top 20 Wikipedia Artikel nach dem PageRank Algorithmus dargestellt.

Wikipedia Artikel	PageRank Punkte
Vereinigte Staaten	4824,62114742905
Deutschland	4192,77150013688
Frankreich	3199,45416122301
Zweiter Weltkrieg	2693,03001486309
Berlin	2563,09678744149
Schweiz	1889,59376564421
Vereinigtes Königreich	1775,46631010045
Paris	1710,39182654457
Wien	1541,39912410098
Erster Weltkrieg	1527,77260917044
Österreich	1481,80300433758
London	1474,58029488960
Englische Sprache	1460,38157027835
New York City	1340,30993828915
München	1333,75239532338
Italien	1302,97644735719
Hamburg	1208,80823295755
Deutsche Demokratische Republik	1205,18911927691
Russland	1104,51072476889
Polen	1067,17156318133

Tabelle 1 Ergebnisse PageRank Top 20