

Bachelorarbeit

Graph Exploration and Analysis on Business Graphs

Graphexploration und -analyse auf Unternehmensgraphen

Milan Gruner

`milangruner@gmail.com`

Eingereicht am <TBD>

Fachgebiet Informationssysteme

Betreuung: Prof. Dr. Felix Naumann, Michael Loster, Toni Gruetze

Abstract

Costructing a graph made up of thousands of businesses may be hard, but actually making sense of it is a lot harder. With huge amounts of data potentially being integrated into the data lake every day, automatic methods for finding interesting spots in the graph are needed. This paper discusses different approaches that can be taken to extract useful knowledge from such a graph.

Contents

1	Introduction	4
1.1	Technologies used in this paper	4
1.2	Usage Scenarios of Ingestion, Curation and GraphXplore	4
2	Data structures for business entities	6
2.1	The <i>Subject</i> data structure	6
2.2	A versioning scheme that stands the test of time	7
3	Using Apache Spark and GraphFrames for Graph Analysis	9
3.1	Why Spark, Cassandra and GraphFrames?	9
3.2	Constructing a GraphFrame from subject data	9
4	Approaches for subgraph extraction	11
4.1	Finding company groups in the business graph	11
4.2	Using PageRank to find relevant companies	12
4.3	Finding important connections	13
5	Motif Search	14
5.1	Pattern types and their applications	14
5.2	Related work	14
6	Pattern Analysis	15
6.1	User-aided approaches for Pattern Categorization	15
6.2	Pattern importance measures	15
6.3	Machine Learning Models for analyzing user feedback	15
6.4	Related work	15
7	Benchmarks and Experiments	16
7.1	GraphX vs. Plain RDDs	16
7.2	Design decisions and trade-offs	16
7.3	Technical challenges	16
8	Conclusion and outlook	17

1 Introduction

This paper describes the basic approaches that were taken for analyzing the graphs generated as a part of the Ingestion bachelor project at HPI Potsdam. It deals with the experiments I performed to evaluate certain algorithms and approaches to summarize automatically generated graphs in the ball park of tens of millions of nodes in a large and mostly unconnected and sparse graph. This graph contains businesses, places and persons and can be augmented freely by adding additional data sources (or manual data input) of your own. The data is visualized and controllable from the Curation interface, which was a project developed alongside the Ingestion pipeline. Curation generates graphs on-the-fly that contain the statistics data that was written as a side effect of the numerous Spark jobs in the Ingestion pipeline and gives a lot of insight into what's going on in the "data lake". All of the experiments contained in this paper were conducted using either the Curation interface, the Apache Zeppelin interactive web shell or the Spark shell.

1.1 Technologies used in this paper

Most of the topics in this paper deal with graph-related topics, so the GraphFrames framework (in Scala) will mostly be used to describe the algorithms that were employed, but it can be freely seen as functional pseudo code and adapted to other popular graph processing frameworks (such as GraphX, Giraph etc.) For UI-related or user-oriented algorithms, JavaScript (or JSX, so partly using React/ Redux etc.) is the listing language, as visual problems are more easily expressed in this tree-oriented but still functional style (and also because it's used in Curation). Whereever possible, ES2016 Syntax is used for the brevity and the similarity to the functional approach of Scala and the rest of the example code.

Also one of the purposes of the Ingestion and Curation projects was to enable the execution of modern text mining, deduplication and parallel data analysis techniques to the german business landscape. This means that some of the examples will be in German (but they will be translated if needed for understanding the technique).

1.2 Usage Scenarios of Ingestion, Curation and GraphXplore

The process of classic risk analysis which is still majorly employed in today's banks is mostly a model-oriented process that heavily relies on speculation or predictions. The techniques that Ingestion, Curation and this paper offer are universally usable to analyze huge amounts of structured or unstructured data in the form of WikiData, JSON,

CSV or similar data as well as newspaper articles, messages (eMail, instant or otherwise) or similar data sources.

The results of this can either be viewed as a graph or a relational model using the conversion algorithms briefly discussed later in this paper. Because this project consists of automatic processes that control the flow, categorization and extraction of all the data in the data lake, it may be effectively deployed in a bank or similar institution as a data collection, curation and analysis software stack.

It is meant to be heavily extensible and allows the simple integration of new data sources as well as algorithms and data structures in general. The pipeline itself is very modular and is composed of many stand-alone Spark jobs that mostly read and write from or to the Cassandra database.

2 Data structures for business entities

2.1 The *Subject* data structure

The central data structure that exists inside the datalake is called *subject*. It contains all the final businesses, industry sectors, countries or persons that are the result of the computations in the Ingestion project.

This is what it looks like in Scala:

Listing 1: Subject

```

1  case class Subject(
2      var id: UUID = UUID.randomUUID(),
3      var master: UUID,
4      var datasource: String,
5      var name: Option[String] = None,
6      var aliases: List[String] = Nil,
7      var category: Option[String] = None,
8      var properties: Map[String, List[String]] = Map(),
9      var relations: Map[UUID, Map[String, String]] = Map(),
10
11     var master_history: List[Version] = Nil,
12     var name_history: List[Version] = Nil,
13     var aliases_history: List[Version] = Nil,
14     var category_history: List[Version] = Nil,
15     var properties_history: Map[String, List[Version]] = Map(),
16     var relations_history: Map[UUID, Map[String, List[Version]]] = Map()
17 )

```

The first three attributes (*id*, *master*, *datasource*) constitute the primary key in the Cassandra databases. The partitioning key that is used to decide on which Cassandra instance to store a given *Subject* entry is the *master* ID, which makes sure that a master node and all its data source slave nodes are stored on the same machine. *id* and *datasource* are used as clustering keys, which control the sorting inside the data partitions produced by database queries. They are also easily queryable when specific IDs or data source entries for a given master ID are needed.

When loading Cassandra data into a Spark job, simple fields (e.g. Strings) need to be wrapped in an Option in case that the field contains no value (otherwise there would be an error), which isn't necessary for the fields containing Maps or Lists, which would just contain an empty data structure in that case.

The *properties* field contains all the extracted data from the various data source extraction and normalization steps in the Ingestion pipeline. It is structured in a way to make

storing arbitrary data possible, by allowing to store a List of Strings for each property key. This way, a list of eMail addresses can be stored as well as a single postal code and they can be treated the same way.

The *relations* field is structured in a two-dimensional map. The first level is addressed by the target *subject* ID, while the second level's keys are the specific relation type. The value that is returned when the field is queried by the target ID and relation type is the relation attribute, which is mostly used to store a confidence measure (a value from 0 to 1) or the count of the relation type between the current *subject* and the target.

2.2 A versioning scheme that stands the test of time

Versioning is needed for the automatic system that includes various pipelines that write to lots of different tables in Cassandra during their runtime, but at the end it is expected that all data will flow back into the central 'data lake' that is the *subject* table, described earlier in this chapter. This means that fully automated processes may and will write millions of data entries to the datalake at an hourly basis (or even more frequent), so if any breaking changes are introduced by these, they need to be rerolled quickly. This means that each version isn't saved as a diff like in version control systems (e.g. git), but rather every single version has a copy of all the attributes, a timestamp and the name and version of the program that modified it. As our whole pipeline is made up of individual small Spark jobs that write mostly into their own Cassandra tables, the error in the data can be quickly located and traced to the corresponding source code.

Because our data domain included events, attributes and relations that had temporal validity parameters attached to them, we had to realize a versioning scheme that was compatible with the column-family oriented storage of Cassandra. As everything is distributed across the cluster here, a diff-oriented versioning would be hard to keep up, so our team decided on a deeply structured deep copy backup system for our versions.

The version data structures are stores for each attribute and each individual key of the subject entries, which makes restoring individual columns or single data entries efficient and easily parallelizable.

The version data structure that was used is structured like this:

Listing 2: Version

```
1  case class Version(  
2    version: UUID = UUIDs.timeBased(),  
3    var program: String,  
4    var value: List[String] = Nil,  
5    var validity: Map[String, String] = Map(),
```

```
6     var datasources: List[String] = Nil,  
7     var timestamp: Date = new Date()  
8 )
```

It is stored inside the collections of the *history* fields of the *Subject* data structure and contains useful meta information for restoring old versions, modeling temporal validity (attributes only being valid for a certain timespan) or filtering out data from certain data sources. They also contain the Spark program's name that created the relevant changes and the time they were made.

Additionally, the *version* table contains a list of all changes to the datalake, making it traceable which programs were ran at what time, which table they wrote to and what data sources they processed. Using this, the Curation interface can display the annotated version list, and run Spark jobs that restore the whole data lake to a previous version or compute the changes that a version made in the *subject* table.

3 Using Apache Spark and GraphFrames for Graph Analysis

This chapter will briefly discuss the reasons for using Spark’s GraphX API alongside the Apache Cassandra database. Then the transformations that are necessary to extract GraphFrames-compatible data from the data lake will be detailed. The following chapters will build on the techniques described in this one, so that only the relevant graph processing sections need to be shown later on.

3.1 Why Spark, Cassandra and GraphFrames?

The subject table contains millions of entries. In order to efficiently extract a graph and calculate the results of complex graph algorithms, a parallel processing framework like Apache Spark is a really important measure to take into consideration when runtimes of multiple hours (or even days) aren’t desired. Cassandra is a database that is up to the challenge of delivering the input for and storing the results of these computations without much hassle (like serializing to CSV or Parquet files would involve). It is also well integrated into the Spark API and can take care of complex nested data structures like e.g. the properties and relations attributes of the *Subject* case class.

TODO continue

3.2 Constructing a GraphFrame from subject data

The core class of the GraphFrame API is the class GraphFrame itself, which is basically a combination of a vertex and relation dataframe as well as a lot of operations which can be performed on these. It is a useful tool for easily handling the graph data structure and interacting with it in a functional manner whilst still being able to interact with the relational API of Spark’s dataframes.

The case class entries that are queried from Cassandra are transformed into vertex and relation data using the Spark RDD API. The results are then translated into data frames, which GraphFrames expects when creating a new graph instance.

The *extractGraph* function from *GraphExtractor* takes care of this process:

Listing 3: *extractGraph* in *GraphExtractor*

```

1  def extractGraph(subjects: RDD[Subject], sc: SparkContext): GraphFrame = {
2      val spark = SparkSession
3          .builder()
4          .appName(appName)
```

```

5      .config(sc.getConf)
6      .getOrCreate()
7  import spark.implicits._
8
9  val subjectVertices = subjects
10     .map(subject => (subject.id.toString, subject.name))
11     .toDF("id", "name")
12  val subjectRelations = subjects.flatMap(subject => {
13     subject.masterRelations.flatMap { case (id, relTypes) =>
14         relTypes.map { case (relType, value) =>
15             (subject.id.toString, id.toString, relType, value)
16         }
17     }
18  }).toDF("src", "dst", "relationship", "value")
19
20  GraphFrame(subjectVertices, subjectRelations)
21  }

```

A spark session has to be manually created in the beginning because the implicit `.toDF()` method needed for transforming the RDD into a `GraphFrame` is contained in either a Spark SQL Context or a session object, the former being the deprecated way of gaining access to this method.

The three nested *map* statements make every relation type to every target subject contained into a new tuple in the resulting RDD. Because of the outer *flatMap* calls, the algorithm gets rid off all of the nested structure of the *relations* attribute, adding the source node to every given triple of the target, relation type, and relation value. It then names them in the scheme that the *GraphFrame* constructor expects.

4 Approaches for subgraph extraction

This chapter deals with various graph processing jobs developed using the GraphFrames framework that extract useful subsets of the full graph in the data lake. These are some examples of simple graph processing that lay the cornerstone for the remaining graph evaluation techniques discussed in later chapters.

4.1 Finding company groups in the business graph

As a first experiment with real-world applications, company group extraction was a good starting algorithm that could be adapted from the connected component tagging which was already present in the GraphFrames API. Here, all businesses that were connected via "owns" or "owned by" relations were considered as a company group.

Before the ConnectedComponents algorithm is run, all ownership-related company relations are extracted from the *Subject* graph. This is done by whitelisting all relations (in *ownershipRelations*) that concern this particular subgraph, e.g. "owns" and "ownedBy". Then, after executing the search, all entries are grouped by the ID of the associated component and their names and UUIDs are saved to Cassandra as an entry of the *graphs* table. The following code snippet describes how this process is expressed using GraphFrames:

Listing 4: Company Group Extraction

```

1  val ownershipEdges = graph.edges.filter((edge) =>
2    ownershipRelations.contains(edge.getAs[String]("relationship")))
3  val ownershipGraph = GraphFrame(graph.vertices, ownershipEdges)
4
5  ownershipGraph.connectedComponents.run()
6    .select("id", "component", "name")
7    .rdd
8    .map(row => (
9      row.getAs[Long]("component"),
10     UUID.fromString(row.getAs[String]("id")),
11     row.getAs[String]("name")))
12  .groupBy(_._1) // component
13  .map(_._2.map(t => (t._2, t._3)).toList) // extract id and name
14  .filter(_._length >= minComponentSize)
15  .map(idNameTuples => ResultGraph(
16    outputGraphType,
17    idNameTuples.map(_._1),
18    idNameTuples.filter(_._2 != null).map(_._2)))

```

The result is a list of businesses like the following:

Oracle America, Hyperion Solutions, Sun Microsystems, Micros Systems, Oracle Financial Services Software, Oracle, PeopleSoft, Acme Packet, Art Technology Group

The corresponding IDs are also stored alongside the names and the graph type (here: "CompanyGroup"). This makes it easy to analyze the results without always joining all subject's names and IDs to this dataset. It also facilitates displaying a graph list in the Curation interface, because only one table's content needs to be fetched for it.

4.2 Using PageRank to find relevant companies

The PageRank algorithm by Page et al. [1] was originally intended to compute the likelihood that a user might arrive at a web page in a network of interconnected documents. It can however be adapted to most graph structures and also finds an application when dealing with business graphs, as the navigation between businesses over their outgoing relations is quite alike to the navigation between web pages over their hyper links. Therefore, the vertex output of the algorithm can be interpreted as the probability that a user might arrive at a certain business when freely navigating the graph, starting at any connected node. The relation output can be seen as the likelihood that the user travels along this relation to the next company.

Fortunately, PageRank is already implemented in the GraphFrames library. There are two different possibilities to call it on a given graph: it can either be executed for a given number of iterations, or it can be run until convergence is reached within a certain tolerance interval to keep the algorithm from running endlessly. Experience has shown that the former is quite a lot faster, but the number of iterations has to be manually adapted to the graph size and its degree of interconnectedness. Otherwise the results differ too much from the latter implementation. Here, the only parameter that has to be controlled is the tolerance threshold, which is independent of the input graph because it only depends on the accuracy that is desired.

After the PageRank algorithm was run, the results still have to be interpreted. This can be done using thresholding techniques and can be parallelized over the connected components of the business graph.

First, the individual *pagerank* values assigned to every node have to be normalized over the whole connected component, which is done by finding the maximum of this attribute in the component and dividing every node's attribute by it. Then, a threshold value has to be picked for the component. This can either be done using an absolute threshold that has to be determined for the whole graph, or using an adaptive thresholding technique like Otsu's method, which is normally used for image thresholding but can be applied here as well. It works by creating a histogram of the absolute probabilities of each *pagerank* value present in the component, which are grouped into bins. It then tries to separate

the values into two classes and minimize the variance inside these classes.

4.3 Finding important connections

5 Motif Search

5.1 Pattern types and their applications

5.2 Related work

6 Pattern Analysis

6.1 User-aided approaches for Pattern Categorization

6.2 Pattern importance measures

6.3 Machine Learning Models for analyzing user feedback

6.4 Related work

7 Benchmarks and Experiments

7.1 GraphX vs. Plain RDDs

7.2 Design decisions and trade-offs

7.3 Technical challenges

8 Conclusion and outlook

References

- [1] PAGE, Lawrence ; BRIN, Sergey ; MOTWANI, Rajeev ; WINOGRAD, Terry: The PageRank Citation Ranking: Bringing Order to the Web. / Stanford InfoLab. Version: November 1999. <http://ilpubs.stanford.edu:8090/422/>. Stanford InfoLab, November 1999 (1999-66). – Technical Report