DistSim - Scalable Distributed in-Memory Semantic Similarity Estimation for RDF Knowledge Graphs

Carsten Felix Draschner  
carsten.draschner@uni-bonn.de  
University of Bonn  
Bonn, Germany

Jens Lehmann  
jens.lehmann@cs.uni-bonn.de  
University of Bonn  
Bonn, Germany

Hajira Jabeen  
hajira.jabeen@uni-koeln.de  
University of Cologne  
Cologne, Germany

Abstract— In this paper, we present DistSim, a Scalable Distributed in-Memory Semantic Similarity Estimation framework for Knowledge Graphs. DistSim provides a multitude of state-of-the-art similarity estimators. We have developed the Similarity Estimation Pipeline by combining generic software modules. For large scale RDF data, DistSim proposes MinHash with locality sensitivity hashing to achieve better scalability over all-pair similarity estimations. The modules of DistSim can be set up using a multitude of (hyper-)parameters allowing to adjust the trade-off between information taken into account, and processing time. Furthermore, the output of the Similarity Estimation Pipeline is native RDF. DistSim is integrated into the SANSA stack, documented in scala-docs, and covered by unit tests. Additionally, the variables and provided methods follow the Apache Spark MLlib name-space conventions. The performance of DistSim was tested over a distributed cluster, for the dimensions of data set size and processing power versus processing time, which shows the scalability of DistSim w.r.t. increasing data set sizes and processing power. DistSim is already in use for solving several RDF data analytics related use cases. Additionally, DistSim is available and integrated into the open-source GitHub project SANSA.

I. INTRODUCTION

Various information domains are modeled using knowledge graphs. A popular standardized format to encode knowledge graphs as linked data is RDF. For optimizing available RDF data, we perform algorithms like Entity Resolution, Entity Linking, and Classification. Additionally, methods are needed to gain insights into the data and to find relevant entities for advanced analytics like Recommendation Systems, Clustering, and Anomaly Detection. The commonality in all of these methods is that they rely on computing similarities or distances between entities. RDF data can have arbitrary sizes up to billions of triples and gigabytes of volume.\(^1\) This large scale data indicates the need for distributed computing because it is cheaper and sometimes more convenient to scale horizontally rather than vertically. Processing large scale data can lead to extensive processing times, especially when the algorithms have non-linear complexity. For some of the non-linear complexity algorithms, there are probabilistic alternatives that reduce the complexity.\(^1\) This reduction in complexity offers a trade-off between run time and result quality. However, whether a reduction in quality is acceptable depends on the use case. Therefore, it is desired to develop algorithms by providing an appropriate trade-off between processing time and quality of results. The contributions of this work are:

- A scalable DistSim in-Memory Semantic Similarity Estimation Framework for RDF Knowledge Graphs
- Integration of DistSim into the holistic SANSA stack over a set of generic modules
- Representation of Semantic Similarity Estimation Experiments and their results in native RDF format
- Evaluation of scalability of different distributed similarity estimation approaches

II. PRELIMINARIES

A. Resource Description Framework

RDF is the W3C standard to represent semantic linked data. RDF data is, in principle, built-up by triples. Each Triple has a Subject, a Predicate, and an Object. The elements can be IRI, Blank Node, or Literals. Subjects can be IRI or Blank Node, Predicates are IRI and Objects can be IRI or Literal. IRIs are “Internationalized Resource Identifier”. Blank nodes are nodes in the Knowledge Graph without explicit given IRI. Literals are leaves in an RDF graph and represent explicit values like Strings, Integers or date-time information.

B. Apache Spark

Apache Spark is a framework for cluster computing, available in scala, python, and java. The building block components of Apache Spark are Spark SQL, Spark Streaming, MLlib, and GraphX. Large scale data sets can be processed in memory when a sufficient cluster hardware setup is available. Apache Spark MLlib is Spark’s scalable machine learning library consisting of standard learning algorithms and utilities, including feature transformation, clustering, hashing, and classification.

C. SANSA

The open-source SANSA stack\(^2\) uses Apache Spark and Apache Flink, which offer fault-tolerant, highly available, and scalable approaches to efficiently process massive sized data sets. SANSA provides various layers containing modules for semantic data representation, querying, inference, and analytics. The SANSA stack is available over GitHub.

\(^1\)https://www.w3.org/wiki/DataSetRDFDump

\(^2\)SANSA.
III. RELATED WORK

A. Developments of Semantic Similarity Estimations

In recent years, many different semantic similarity estimators have been developed for various use cases [3, 4]. Three types of Semantic Similarity Estimation have been proposed [5, 6]. Structure/Path-based semantic similarity estimations assign the similarity over two entities’ distance in a knowledge graph. The smaller the distance is, the higher is the similarity (Shortest Path [7], Weighted Links [8], Wu and Palmer [9]). These and further approaches differ in how the knowledge graph structure path is deduced and weighted for the resulting similarity value. The second class of semantic similarity estimations is based on Information Content (Resnik et al. [10], Lin et al. [11]). These approaches assign the similarity by the highest information content of shared features. The information content is calculated based on inverse feature frequency. The rarer a feature is, the higher is its Information Content. The third group is feature-based semantic similarity measures. The feature-based methods differ in the normalization of the calculated overlap of features. The initial approach is Jaccard Index [12]. Jaccard defines the similarity as the cardinality of the intersection of features divided by the cardinality of the union of features. Based on these feature-based semantic similarity measures, probabilistic approaches like minHash [13] were developed. MinHash reduces the processing time in computing all pair similarity by representing the sparse hot encoded feature vector in a dense minHashed vector. Based on these min-hashed vectors, elements are grouped in buckets over Locality Sensitivity Hashing. This bucketing results in a much smaller search space for each entity to calculate similarity values.

B. Distributed Semantic Similarity Estimation Frameworks

In recent years frameworks were developed for semantic analytics like similarity estimation [14, 2, 3], but these developments are not optimized for large data, and scalability over distributed computing. Apache Spark is the state of the art, open Source Framework for distributed data analytics. Spark MLlib provides two similarity estimation modules (Bucketed Random Projection for Euclidean Distance and MinHash [15]) for Jaccard Distance. Apache Spark does not incorporate RDF data out of the box. On the other hand, Knowledge Graph and RDF do not have native feature set vectors, such that they can be used out of the box.

IV. DISTSIM

A. Pipeline Architecture

The Scalable Distributed in-Memory Semantic Similarity Estimation is implemented as a stacked pipeline aligned with the standards of Apache Spark MLlib. Figure 1 shows that the approach consists of six modules: ReadIn, Feature Extraction, Count Vectorization, Similarity Estimation, Metagraph Creation, and WriteOut. DistSim makes use of ReadIn and WriteOut software modules from the SANSA stack RDF Layer to read and write RDF data. For scalable similarity estimation, it uses Count Vectorizer and MinHash with Locality Sensitivity Hashing from Apache Spark MLlib. In addition, DistSim provides feature extraction, a set of similarity estimation models, and the meta graph creation as novel contributions. The pipeline for semantic similarity estimation reads-in and writes-out RDF data.

B. DistSim as Resource

DistSim is developed as open source and is fully integrated into the SANSA-Stack and documented with scala-docs. The estimators provide methods for nearestNeighors which estimates for one URI the k most similar elements represented by their URI. Alternatively, allPairSimilarity calculates the similarity of all pairs of URIs from the two DataFrames (of length n and m). DistSim provides the output RDF data enriched with similarity annotations and meta-information (see figure 2). The annotated meta information of the similarity estimation not only makes the results reproducible, but it also allows the possibility to comprehend the conditions and parameters used for the estimation. For every novel developed module, we provide unit tests.

C. DistSim Feature Extraction

The Semantic Similarity Estimations of DistSim operate on feature sets. These feature sets are derived from the reading the RDF data set using the Feature-Extractor Module, which is implemented as a Transformer. Developers can set the Feature-Extractor methodology using modes. The mode specifies how the information stored in triples for a specific URI is transformed into the assigned feature set. In this paper, we present two out of twelve available feature extraction modes, supplied as a parameter in Feature Extractor Transformer initialization. The corresponding figures 3, 4 show on the left-hand side the sample KG. with the entities in blue, the used triple information for features in green, and the ignored information in red. On the right-hand side, the corresponding schematic feature...
accurate functions only on promising candidates. Scalable processing time and call in a second step more models, such that we calculate a set of first estimates in addition, we can stack MinHash with different DistSim approximate method is optimal for large scale data scenarios. This scalable but MinHashLSH, can be used for the probabilistic approach in calculating Jaccard similarity. This scalable but MinHashLSH[15],[13] can be used for the probabilistic estimation (see table I). The scalable alternative D. Semantic Similarity Estimation Models

DistSim provides feature set based semantic similarity estimations (see table I). The scalable alternative MinHashLSH[15],[13] can be used for the probabilistic approach in calculating Jaccard similarity. This scalable but approximate method is optimal for large scale data scenarios. In addition, we can stack MinHash with different DistSim models, such that we calculate a set of first estimates in scalable processing time and call in a second step more accurate functions only on promising candidates.

<table>
<thead>
<tr>
<th>Similarity Coefficient</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bate[16] Distance</td>
<td>( \log_2 \left( 1 + \frac{</td>
</tr>
<tr>
<td>Braun-Blanquet[17] Similarity</td>
<td>( \frac{2XY}{(X+Y)X} )</td>
</tr>
<tr>
<td>Dice[13] Similarity</td>
<td>( \frac{</td>
</tr>
<tr>
<td>Jaccard[12] Similarity</td>
<td>( \frac{</td>
</tr>
<tr>
<td>Ochiai[19] Similarity</td>
<td>( \sqrt{\frac{</td>
</tr>
<tr>
<td>Simpson[20] Similarity</td>
<td>( \frac{2XY}{\min {X,Y} \left(</td>
</tr>
<tr>
<td>Tversky[21] Similarity</td>
<td>( \frac{</td>
</tr>
</tbody>
</table>

E. DistSim (Hyper-)Parameter Setup

The modular DistSim Pipeline allows a multitude of adjustments over (hyper-)parameters that can reduce the memory usage and the processing time. The trade-off comes with a loss of information. The CountVectorizer transforms a set of features into a vector with a fixed length. The length is adjustable by minimal document frequency (minDf) and upper bound vector size (maxVocabSize). Small feature vectors need less memory and can be processed faster but store less information. In Semantic Similarity Estimation over MinHashLSH, a higher number of hash tables (numHashTables) reduce the false-negative rate in detecting similar elements but increase the processing time and memory usage. The threshold on minimal similarity, respectively, minimal distance can minimize memory usage and processing time. If this threshold is more strict, fewer pairs of similar values have to be processed over a distributed system in allPairSimilarity.

F. DistSim Use Cases

Scalable distributed semantic similarity estimations are needed in several Use Cases. The extended SANSA stack is in use as a generic Big Data Analytics Toolbox of the Horizon 2020 Project PLATOON. SANSA, as an underlying toolbox for semantic analytics in Opertus Mundi[2] provides with the novel developed semantic Similarity Pipeline needed software modules. The project Simple-ML[3] provides an easy to use generic stackable machine learning framework which uses SANSA and DistSim as underlying Semantic Similarity Framework for RDF data.

V. EXPERIMENT AND EVALUATION

DistSim implements well-established similarity estimation functions for RDF data. The evaluation is presented for the performance assessment of DistSim on different data sizes and varying cluster processing setups. Here, the processing time is an indicator of DistSim’s distributed processing and scalability. The cluster processing power is adjusted over the spark-submit command, where the number of executor cores can be limited.

A. Data Sets

The evaluation of the scalability of DistSim is performed over multiple data sets of different sizes. The data set sizes are adjusted by creating synthetic data sets. We use synthetic data sets to ensure equally distributed graph density. In real-world graphs, cutting off fractions could lead to an unnatural graph appearance. Figure 4 shows on the left hand side the principle structure of the generated data set.

B. Scalability over increasing horizontal Cluster Computation

The processing power is regulated over the number of available cores (from \( 2^2 = 4 \) up to \( 2^7 = 128 \)). Table I shows the scalability over cluster setups. We see a clear decrease in processing time over increasing computation power.

---

2https://www.opertusmundi.eu

3https://simple-ml.de
### C. Scalability over Data Set Size

The use of probabilistic similarity estimator MinHashLSH allows scalable processing of large scale RDF data. Figure 5 shows for the all pair similarity estimation that MinHash (orange) scales better than the other approaches. The approaches Batet, Braun Blanquet, Dice, Jaccard, Simpson, and Tversky scale similar. For Nearest Neighbor Estimation, all approaches are on a similar scalable level, including MinHash, because Nearest Neighbor is a linear operation and not quadratic in complexity like All-Pair Similarity.

### VI. Conclusion and Future Work

DistSim integrated into the SANSA stack provides a scalable distributed open-source framework for semantic similarity estimation on RDF Knowledge Graphs. Multiple projects are already using DistSim modules. The community is actively using the SANSA stack for scalable distributed semantic analytics on large-scale RDF data. The availability of an easy to use evaluation pipeline shows clear infer-able effects of (hyper-)parameters to the corresponding processing times. The storage in a tabular format and semantic data representation (see figure 2) allows high reproducibility and understanding of the needed pipeline setup. The results are human and machine-readable. Using DistSim and the proposed analytic pipeline modules for RDF processing, additional RDF data analytic algorithms can be easily ported to distributed processing. The evaluation shows scalability of DistSim over different data set sizes and processing power. We are currently developing more approaches for feature extraction and semantic similarity estimations to cover additional semantic information.

### ACKNOWLEDGEMENT

This work was partly supported by the EU Horizon 2020 project PLATOON (Grant agreement ID: 872592).

### REFERENCES


