Rheinische Friedrich-Wilhelms-Universität Bonn

Scalable Entity Resolution

Amrit Kaur (Matriculation Number - 3055863)
Supervisor: Dr. Hajira Jabeen
First Examiner: Prof. Dr. Jens Lehmann
Second Examiner: Dr. Kuldeep Singh

Smart Data Analytics Group • University of Bonn • August 23, 2019

Contents

Introduction

Motivation

- Knowledge Graphs
- Entity Resolution task is Quadratic
- Existing approaches use Blocking Techniques for efficiency
- Need a generic approach that can perform Entity Resolution in a scalable manner in RDF data

Introduction

Problem Description

Given two different datasets:

∗ Find entities that point to the same real- world data

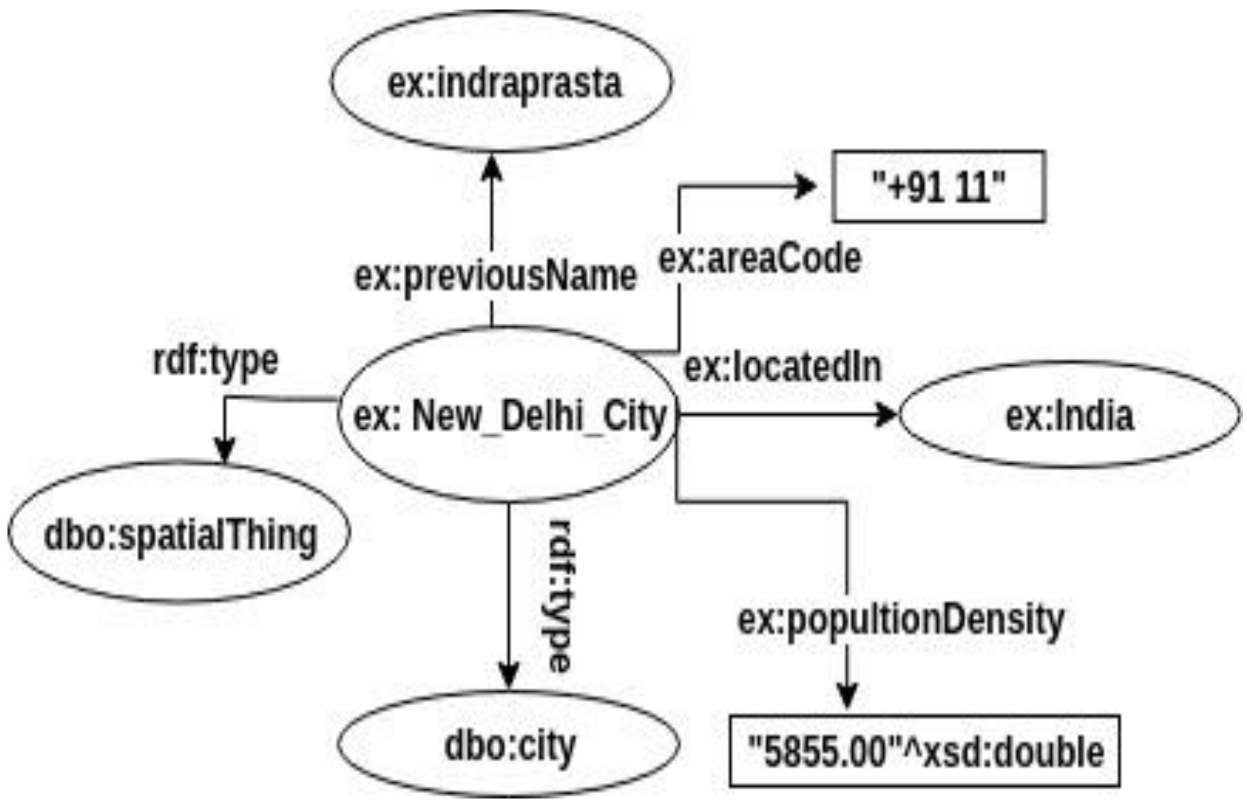


Fig 1. Entity in Dataset1

Introduction

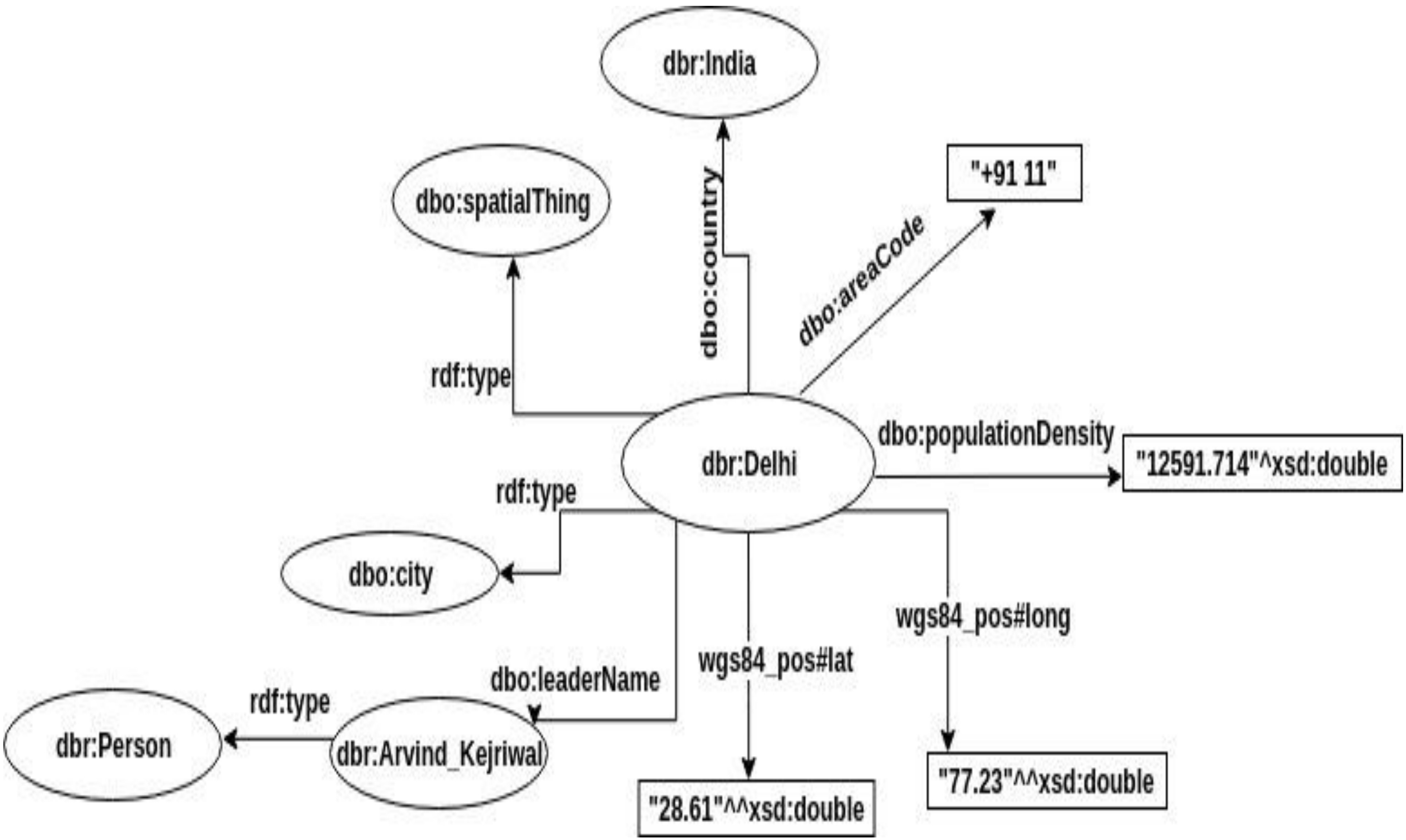


Fig 2. Entity in Dataset2

Existing Methodologies

Existing approaches perform a two step process :-

- Block Building
  - Attribute based Blocking
  - Attribute agnostic Blocking

- Block Processing
  - Non-Learning based approaches
  - Learning based approaches

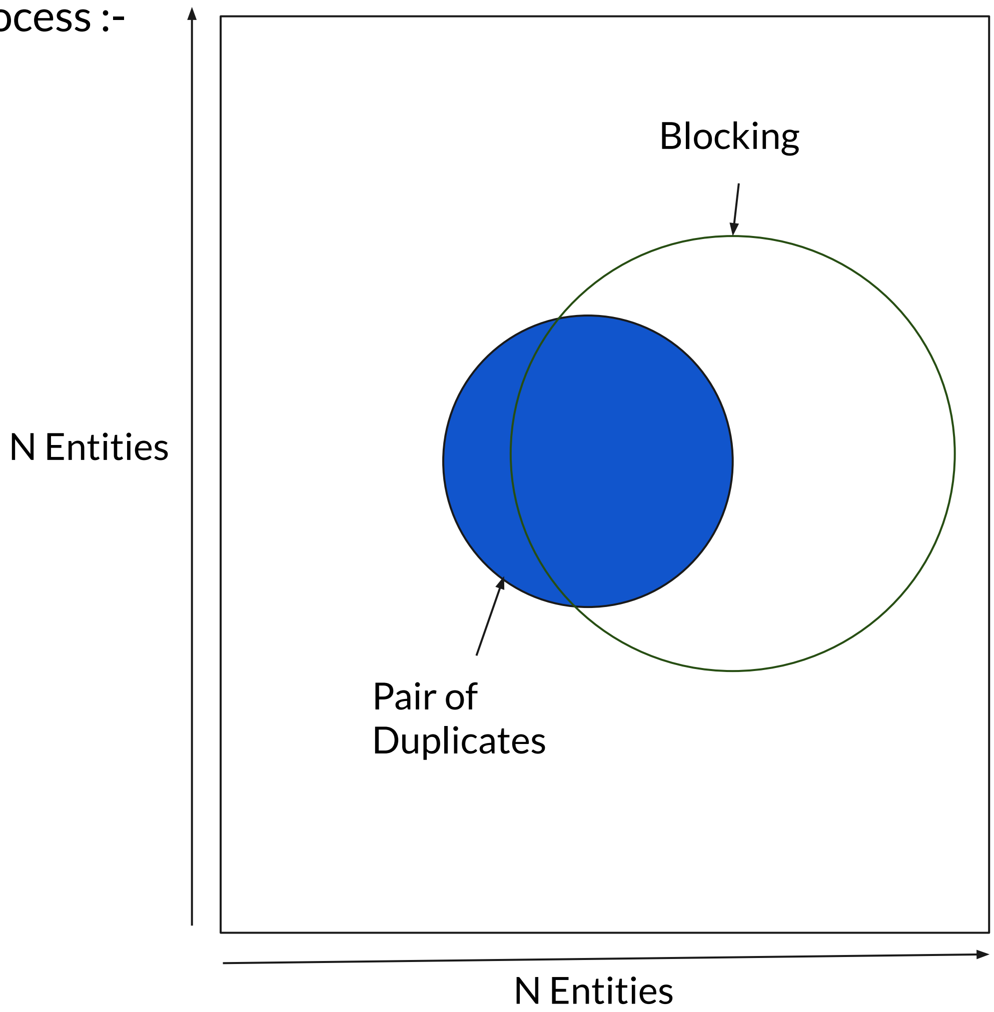


Fig 3. Brute Force Approach

Approach

Focus :-

- Remove block building stage
- Removing multiple iterations used for learning based approaches
- Effort of labelling data for learning based approaches

Approach - minHash LSH method (considering all attributes)

Input:-

Dataframe x

Entity	**Attributes**
Subject1	Set(subject,predicates,objects)
Subject2	Set(subject,predicates,objects)

Dataframe y

Entity	**Attributes**
Subject1	Set(subject,predicates,objects)
Subject2	Set(subject,predicates,objects)

Highlight linear time!

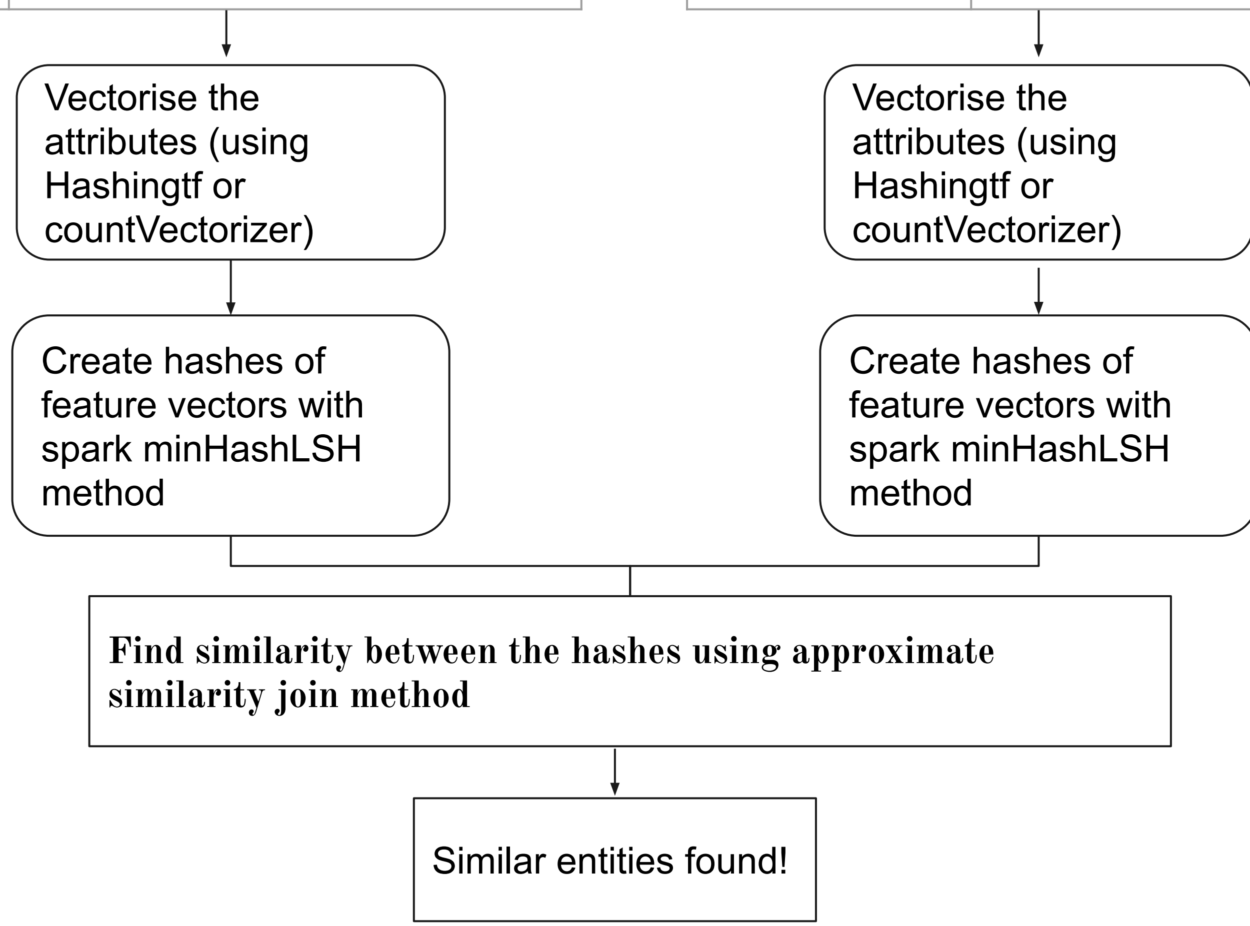


Fig 4. Spark minHash LSH method with all attributes

Challenges

- Incomplete knowledge graphs
- We do not get good results

Approach - minHash LSH method (1 or 2 attribute)

Input:-

Dataframe x

Entity	Attributes
Subject1	Set (subject, p1, o1, p2, o2)
Subject2	Set (subject, p1, o1, p2, o2)

Dataframe y

Entity	Attributes
Subject1	Set (subject, p1, o1, p2, o2)
Subject2	Set (subject, p1, o1, p2, o2)

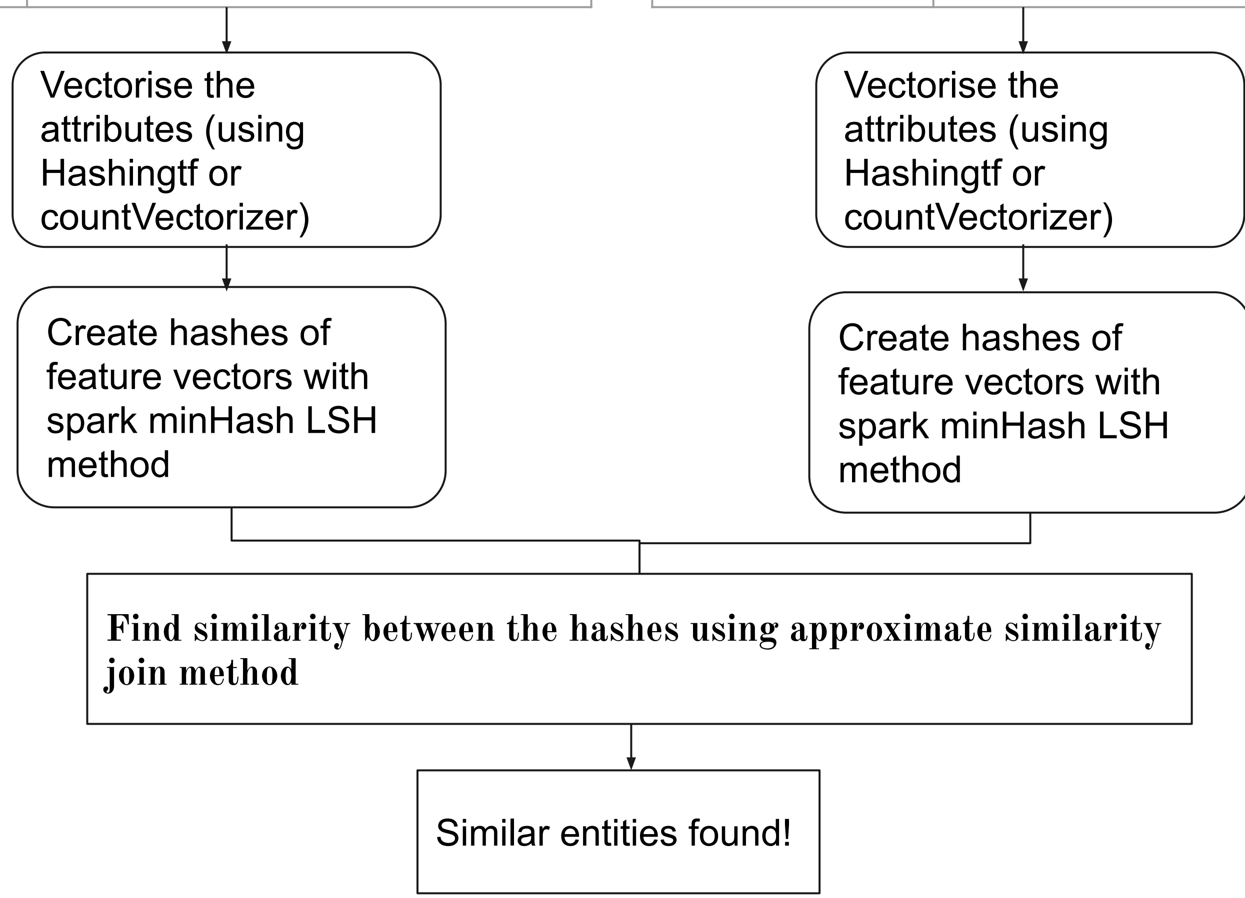


Fig 5. Spark minHash LSH method with 1 or 2 attribute

Approach - minHash LSH method (1 or 2 attribute)

Inspiration : -

Köpcke, H., Thor, A., and Rahm, E.: Learning-Based Approaches for Matching Web Data Entities. IEEE Internet Computing, pp. 23-31, July/August, 2010

Intention:-

Maximum utilisation of data in knowledge graphs for entity comparison

Idea :-

Select common attributes based on subject similarity

Approach - minHash LSH subjects and jaccard similarity attributes

Step1: Find matching entities based on LSH subjects

Input:-

Dataframe x

Entity	Entity_sub_tokenised
Subject1 (Ex:- Budapest)	Set(Budapest)
Subject2 (Ex:- Apple_Store)	Set(Apple, Store)

Dataframe y

Entity	Entity_sub_tokenised
Subject1 (Ex:- Apple)	Set(Apple)
Subject2 (Ex:- Budapest_City)	Set(Budapest, City)

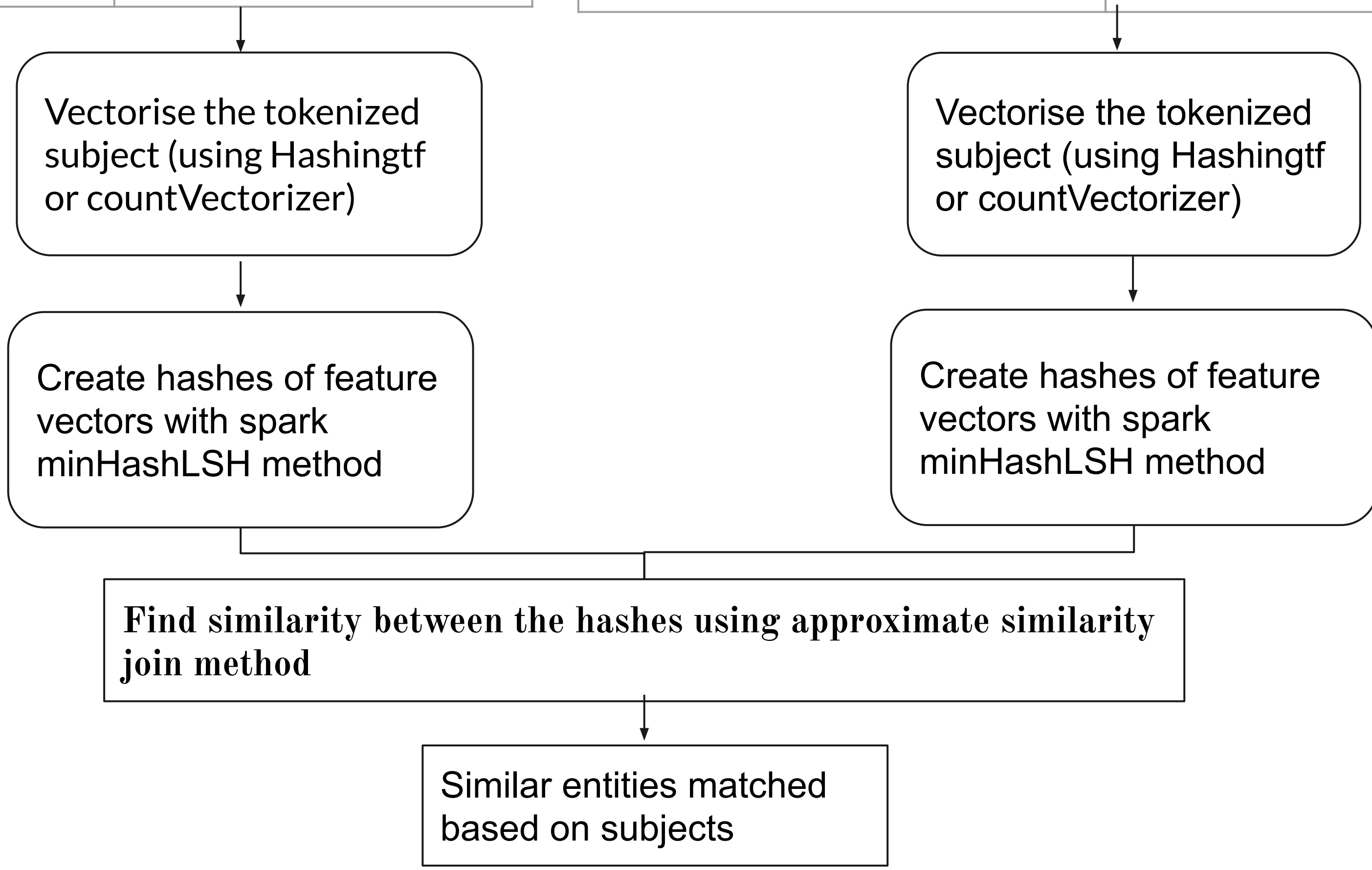


Fig 6. Spark minHash LSH subject

Approach - minHash LSH subjects and jaccard similarity attributes

Focus: Maximum utilisation of data

Step 2 : Compare the predicates for matched entities

Similar entities matched based on subjects

↓

Extract predicates for the similar matched entities and compare their respective predicates by jaccard similarity thresholds

Focus: Reduce False positives

Step 3 : Compare the objects for intersecting predicates in matched entities

Compare the objects of only intersecting predicates by jaccard similarity threshold for object

↓

Similar entities found!

Approach - minHash LSH subjects and jaccard similarity attributes

Step1: Find matching entities based on LSH subjects

Input:- Dataframe x

Entity	Entity_sub_tokenised
Subject1 (Ex:- Budapest)	Set(Budapest)
Subject2 (Ex:- Apple_Store)	Set(Apple, Store)

Dataframe y

Entity	Entity_sub_tokenised
Subject1 (Ex:- Apple)	Set(Apple)
Subject2 (Ex:- Budapest_City)	Set(Budapest, City)

Output :- Similar entities matched based on subjects

Entity1	**Entity 2**
Budapest	Budapest_City
Apple_Store	Apple

Approach - minHash LSH subjects and jaccard similarity attributes

Step 2 : Compare the predicates for matched entities

Entity1	Entity1_predicates	Entity2	Entity2_predicates
Budapest	Set(areaCode, rdf:type, country, timezone, populationDensity, postalCode, utcoffset, humidity)	Budapest_City	Set(areaCode, country, elevation, foundingDate, areaTotal, governingBody, populationDensity, timezone, rdf:type)
Apple_Store	Set(foundedBy, industry, keyPerson, product, parentCompany, rdf:type, subject, numberofLocations, logo, rdf:label , foaf:name)	Apple	Set (genus, rdf:type, division, source, kingdom, class, carbs, fat, fibre, sugar, rdf:label, foaf:name, calcium, phosphorus, potassium, protein)

Output :- Similar entites matched with Jaccard Similarity predicates

Entity1	Entity2	Intersecting_predicates
Budapest	Budapest_City	Set(areaCode, rdf:type, country, populationDensity, timezone)

Approach - minHash LSH subjects and jaccard similarity attributes

Step3: Compare the objects for intersecting predicates in matched entities

i.e Set(areaCode, rdf:type, country, populationDensity, timezone)

Entity1	**Entity1_objects**	**Entity2**	**Entity2_objects**
Budapest	Set (1, City, Place, Location, PopulatedCity, Hungary, 1558465, Central_European_Time)	Budapest_City	Set(1, City, Place, Location, PopulatedCity, Hungary, 1759407, Central_European_TIme, Central_European_SummerTime)

Output:- Similar entities found

Entity1	**Entity2**
Budapest	Budapest_City

Evaluation - minHash LSH method (1 or 2 attribute)

Datasets:-

Match Task		Source size (Number of entities)	
Attributes	**Sources**	**Source1**	**Source2**
Title Authors Venue Year	DBLP-ACM	2,616	2,294
Title Authors Venue Year	DBLP-Scholar	2,616	64.263
Name Description Manufacturer Price	Abt-Buy	1,081	1,092

Table1 . Evaluation dataset description

Evaluation

DBLP-ACM dataset

	1- attribute (Title)		2- attribute (Title, Authors)	
Number of attributes	State of the art	minHash LSH	State of the art	minHash LSH
Precision (%)	94.9	85.88	96.9	92.05
Recall (%)	97.3	95	87.8	93
F-measure (%)	96.1	90.21	92.1	92.52

Table2 . Evaluation Results for DBLP-ACM dataset

A threshold of 0.25 is considered for 1-attribute and 0.28 for 2-attribute

Evaluation

DBLP-Scholar dataset

	1- attribute (Title)		2- attribute (Title, Authors)	
Number of attributes	**State of the art**	**minHash LSH method**	**State of the art**	**minHash LSH method**
Precision (%)	74.1	79.0	77.5	79.86
Recall (%)	91.7	87.0	84.8	83
F-measure (%)	82.0	82.31	81.0	81.40

Table3 . Evaluation Results for DBLP-Scholar dataset

A threshold of 0.15 is considered for 1-attribute and 0.42 for 2-attribute

Evaluation

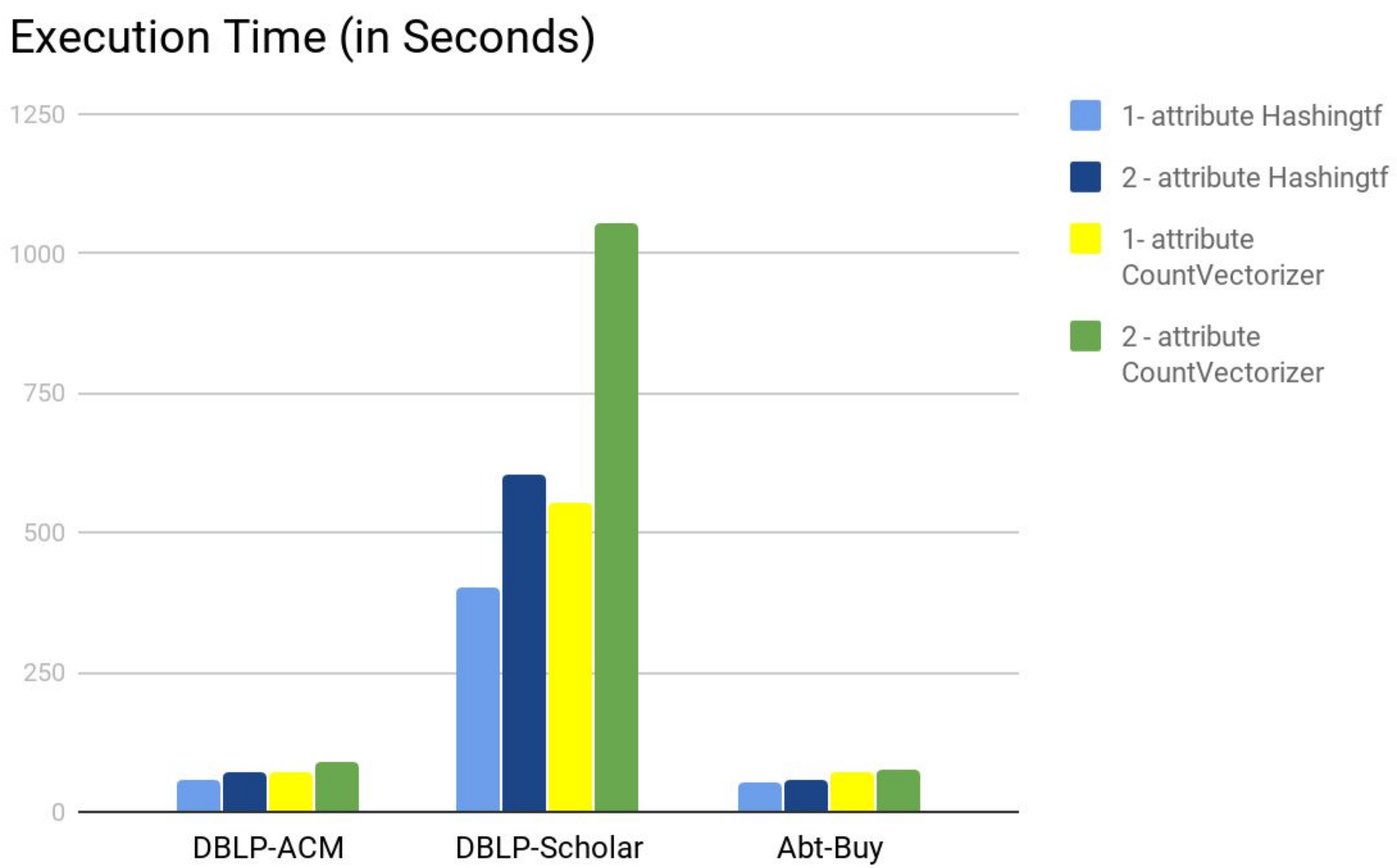
Abt-Buy dataset

	1- attribute (Name)		2- attribute (Name,Description)	
Number of attributes	**State of the art**	**minHash LSH method**	**State of the art**	**minHash LSH method**
Precision (%)	78.4	35.80	90.6	27.63
Recall (%)	36.4	48	17.6	31
F-measure (%)	49.7	41.01	29.5	29.21

Table 4 . Evaluation Results for Abt-Buy dataset

A threshold of 0.5 is considered for 1-attribute and 0.685 for 2-attribute

Evaluation


Execution Time (in Seconds)
1250
1000
750
500
250
0
DBLP-ACM
DBLP-Scholar
Abt-Buy
1- attribute Hashingtf
2 - attribute Hashingtf
1- attribute CountVectorizer
2 - attribute CountVectorizer

Evaluation- minHash LSH subjects and jaccard similarity attributes

Dbpedia Datasets			
Dataset Categorization	**Dataset Size**	**Entities in Dataset1**	**Entities in Dataset2**
Medium size	8.5 GB	Infobox 3.0rc (9,91,933 entities)	Infobox 3.4 (19,40,631 entities)
Large size	24.8 GB	DBpedia 3.0rc (47,81,249 entities)	Infobox 3.4 (19,40,631 entities)

Table 5 . Evaluation DBpedia datasets description

Evaluation

Dbpedia Datasets:-

- For medium size dataset :-
  The ground truth was constructed by considering the matches as entities with exactly same URL.
  Inspiration :-
  G. Papadakis, E. Ioannou, C. Niederée, P. Fankhauser, Efficient entity resolution for large heterogeneous information spaces, in: WSDM, 2011, pp. 535–544.

- For large size dataset :-
  The above created ground truth was considered in this case also.

Evaluation

Medium and Large size Datasets:-

Dataset1 - Infobox 3.0rc/Dbpedia 3.0rc and Dataset2 - Infobox 3.4		
	Jaccard Similarity Predicates	**Jaccard Similarity Objects**
Precision (%)	99.75	99.86
Recall (%)	86	85
F-measure(%)	92.36	91.83

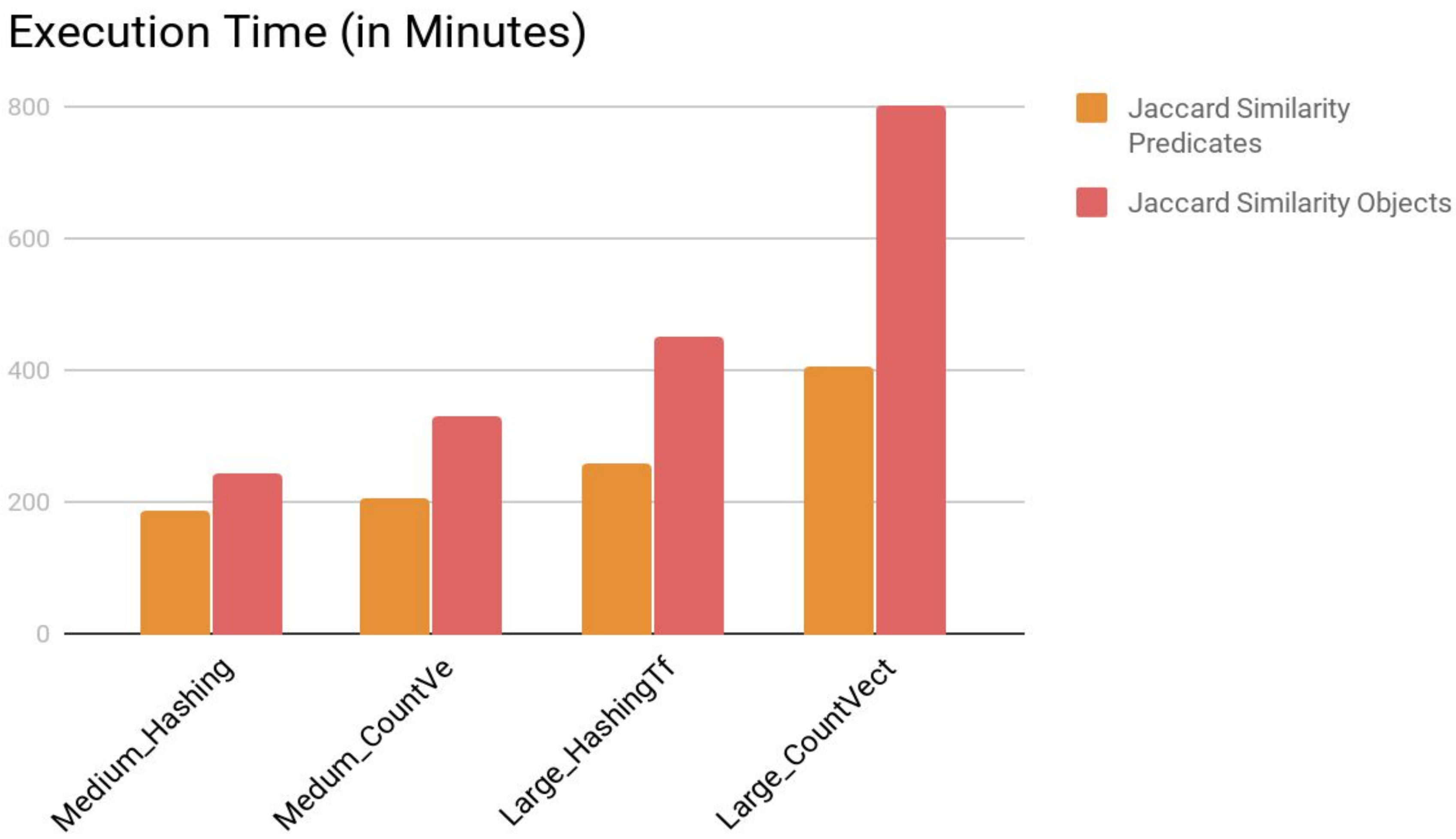
Table 6. Evaluation Results for Dbpedia datasets

A threshold of 0.10 is considered for LSH subjects (the lower the better)

A Jaccard Similarity of 0.15 is considered among the predicates.

A Jaccard Similarity of 0.25 is considered among the objects.

Evaluation- minHash LSH subjects and jaccard similarity attributes



Conclusion

- Scalable
- Efficient
- Deals with heterogeneity of data
- Considers structured as well as unstructured data
- Entity Resolution - An important contribution for SANSA ML upcoming release (integration task ongoing)

Future Work

- Datasets with id's as subjects instead of URL are difficult to compare with our approach.
- Creation of ground truth for larger datasets
- Extend to other data source of RDF data like Yago, Wikidata, etc.
- Perform entity resolution with different Dbpedia language datasets.

References I

- *G. Papadakis, E. Ioannou, C. Niederée, P. Fankhauser, Efficient entity resolution for large heterogeneous information spaces, in: WSDM, 2011, pp. 535–544.*

- *Köpcke, H., Thor, A., and Rahm, E.: Learning-Based Approaches for Matching Web Data Entities. IEEE Internet Computing, pp. 23-31, July/August, 2010*

- *G. Papadakis, G. Papastefanatos, T. Palpanas, M. Koubarakis, Scaling entity resolution to large, heterogeneous data with enhanced meta-blocking, in: EDBT, 2016.*

- *O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, J. Widom, Swoosh: a generic approach to entity resolution, VLDB J. 18 (1) (2009) 255–276.*

- *SANSA Stack Github repository. https://github.com/SANSA-Stack/.*

References II

- *SANSA Stack.*
  http://sansa-stack.net/faq/#what-does-SANSA-stand-for.

- *Jeffrey Fisher, Peter Christen, Qing Wang, Erhard Rahm, A clustering-based framework to control block sizes for entity resolution. in: KDD, 2015.*

- *Köpcke, H., Thor, A., and Rahm, E.: Comparative evaluation of entity resolution approaches with FEVER. In Proc. of VLDB, 200*

Thank you !!!