Neo4j Assignment

Mining Big Datasets

7/8/2023

Christos Vlassis

Table of Contents

Data insertion and database building

We added manually the column names to each excel file.

For importing the Authors we used the following code in the Neo4jBrowser:

```
LOAD CSV WITH HEADERS FROM "file:///AuthorNodes.csv" AS csvLine
CREATE (p:Author {article_id: toInteger(csvLine.Author_id), author_name: csvLine.author_name})
```

We used the following code to create the relationship between the Author nodes and Article nodes:

```
MATCH (a:Author), (ar:Article)
WHERE a.article_id = ar.article_id
CREATE (a)-[r:WROTE]->(ar)
RETURN a, r, ar
```

For importing the Articles csv:

```
LOAD CSV WITH HEADERS FROM "file:///ArticleNodes.csv" AS csvLine
CREATE (a:Article {article_id: toInteger(csvLine.article_id), article_title: csvLine.article_name, year_re
leased: toInteger(csvLine.year_released), abstract: csvLine.abstract, journal_name:csvLine.journal_p
ublished, journal_name:csvLine.journal_published })
```

For importing and creating the citations relationships between the articles:

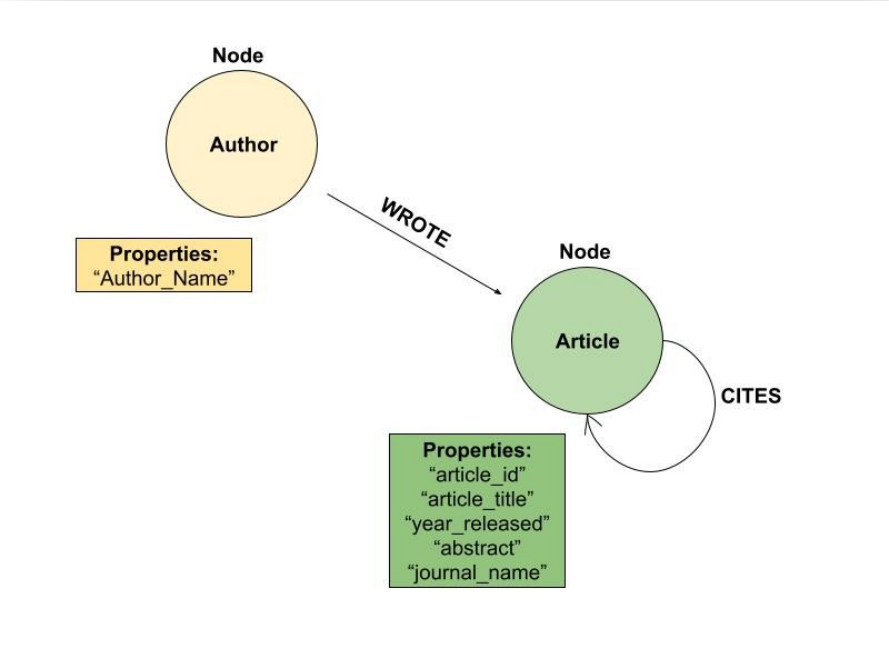
```
LOAD CSV WITH HEADERS FROM 'file:///Citations.csv' AS row
MATCH (a:Article { article_id: toInteger(row.article_id1) }), (b:Article { article_id: toIntege
r(row.article_id2 )})
CREATE (a)-[:CITES]->(b)
```

We used this command to remove the article_ids from the nodes with label Author. We did this because we already have the article_id in the nodes with label Article.

```
match (n:Author) remove n.article_id
```

Database Schema

Bellow follows the schema of the database built:



Questions - Queries

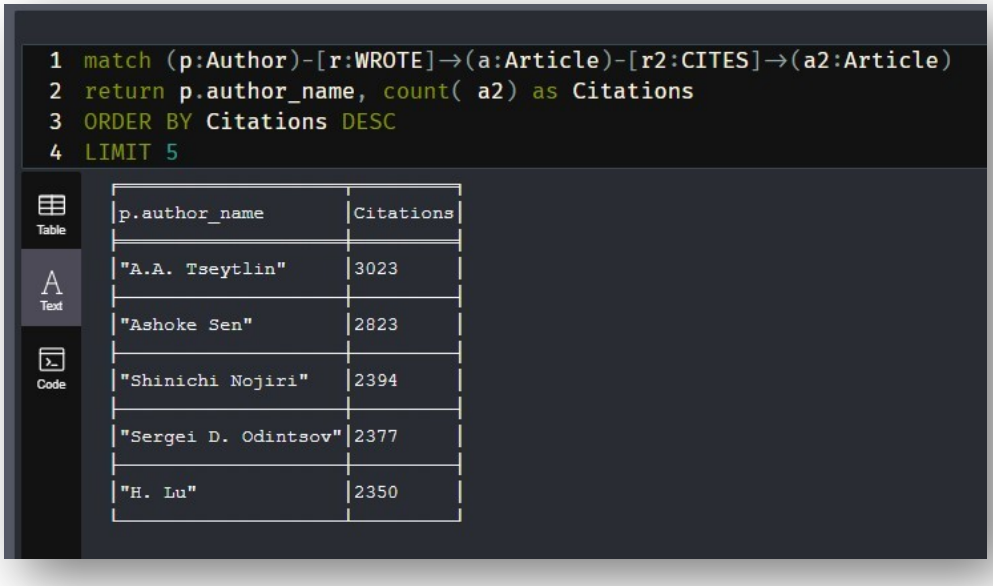
We will use the Browser to answer the queries:

1) Which are the top 5 authors with the most citations (from other papers). Return author names and number of citations.

Query used to answer question:

```
MATCH (p:Author)-[r:WROTE]->(a:Article)-[r2:CITES]->(a2:Article)
RETURN p.author_name, count(a2) as Citations
ORDER BY Citations DESC
LIMIT 5
```

Screenshot of result:



2) Which are the top 5 authors with the most collaborations (with different authors). Return author names and number of collaborations.

Query used to answer question:

MATCH (a1:Author)-[:WROTE]->(ar:Article)<- [:WROTE]-(a2:Author)

WHERE id(a1) < id(a2)

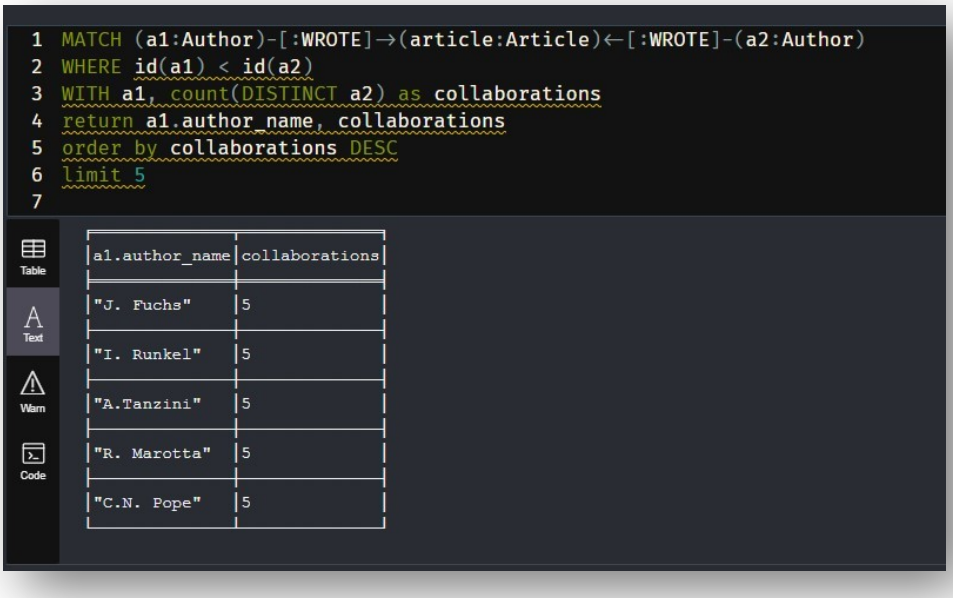
WITH a1, count(DISTINCT a2) AS collaborations

RETURN a1.author_name, collaborations

ORDER BY collaborations DESC

LIMIT 5

Screenshot of result:



```
MATCH (a1:Author)-[:WROTE]→(article:Article)←[:WROTE]-(a2:Author)
WHERE id(a1) < id(a2)
WITH a1, count(DISTINCT a2) as collaborations
return a1.author_name, collaborations
order by collaborations DESC
limit 5
```

a1.author_name	collaborations
"J. Fuchs"	5
"I. Runkel"	5
"A.Tanzini"	5
"R. Marotta"	5
"C.N. Pope"	5

3) Which is the author who has written the most papers without collaborations. Return author name and number of papers.

Query used to answer question:

MATCH (a:Author)-[:WROTE]->(ar:Article)

WITH ar, count(a) as filterr

WHERE filterr = 1

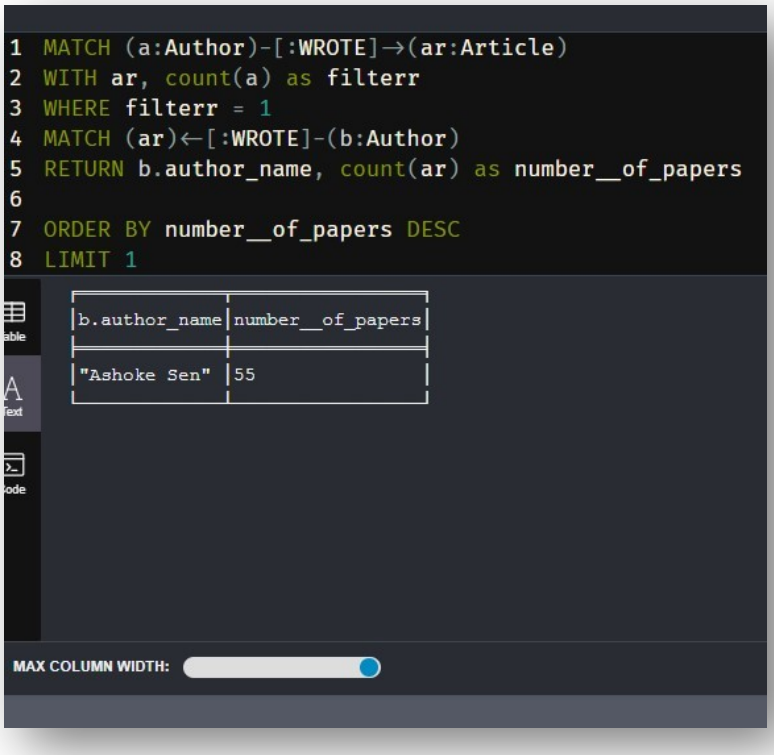
MATCH (ar)<-[:WROTE]-(b:Author)

RETURN b.author_name, count(ar) as number_of_papers

ORDER BY number_of_papers DESC

LIMIT 1

Screenshot of result:



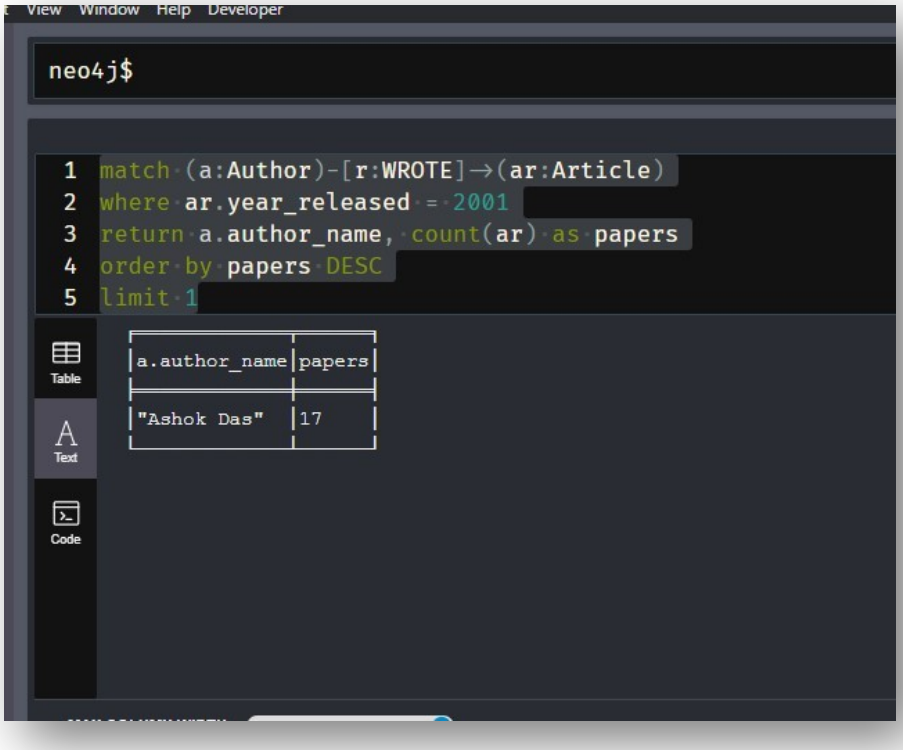
4) Which author published the most papers in 2001? Return author name and number of papers.

We used the WROTE not the property ‘published’.

Query used to answer question:

```
MATCH (a:Author)-[r:WROTE]->(ar:Article)
        WHERE ar.year_released = 2001
        RETURN a.author_name, count(ar) AS papers
        ORDER BY papers DESC
        LIMIT 1
```

Screenshot of result:

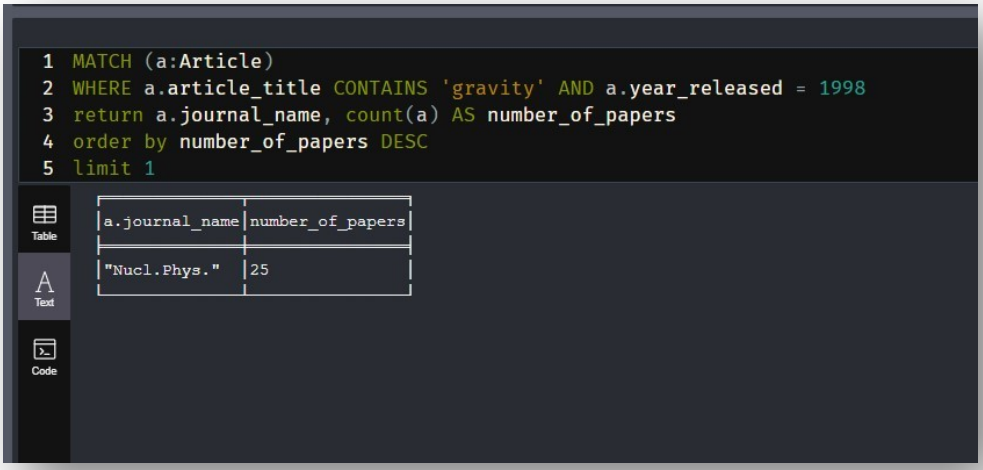


5) Which is the journal with the most papers about “gravity” (derived only from the paper title) in 1998. Return name of journal and number of papers.

Query used to answer question:

```
MATCH (a:Article)

WHERE a.article_title CONTAINS 'gravity' AND a.year_released = 1998

RETURN a.journal_name, count(a) as number_of_papers

ORDER BY number_of_papers DESC

LIMIT 1
```

Screenshot of result:



6) Which are the top 5 papers with the most citations? Return paper title and number of citations.

Query used to answer question:

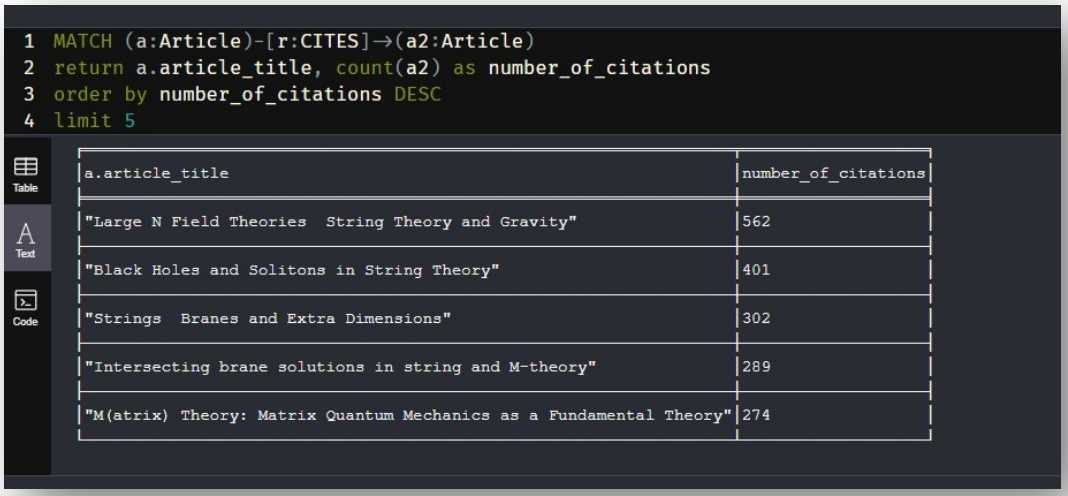
MATCH (a:Article)-[r:CITES]->(a2:Article)

RETURN a.article_title, count(a2) as number_of_citations

ORDER BY number_of_citations DESC

LIMIT 5

Screenshot of result:



```
MATCH (a:Article)-[r:CITES]->(a2:Article)
return a.article_title, count(a2) as number_of_citations
order by number_of_citations DESC
limit 5
```

a.article_title	number_of_citations
"Large N Field Theories String Theory and Gravity"	562
"Black Holes and Solitons in String Theory"	401
"Strings Branes and Extra Dimensions"	302
"Intersecting brane solutions in string and M-theory"	289
"M(atrix) Theory: Matrix Quantum Mechanics as a Fundamental Theory"	274

7) Which were the papers that use “holography” and “anti de sitter” (derived only from the paper abstract). Return authors and title.

Query used to answer question:

MATCH (au:Author)-[r:WROTE]->(a:Article)

WHERE a.abstract CONTAINS 'holography' OR a.abstract CONTAINS 'anti de sitter'

RETURN au.author_name, a.article_title

Screenshot of result:

```
1 match (au:Author)-[r:WROTE]→(a:Article)
2 where a.abstract CONTAINS 'holography' OR a.abstract CONTAINS 'anti de sitter'
3 return au.author_name, a.article_title
```

Table | Text | Code

au.author_name	a.article_title
"Djordje Minic"	"Probable Values of the Cosmological Constant in a Holographic Theory"
"Petr Horava"	"Probable Values of the Cosmological Constant in a Holographic Theory"
"J. Bros"	"Decomposing Quantum Fields on Branes"
"M. Bertola"	"Decomposing Quantum Fields on Branes"
"R. Schaeffer"	"Decomposing Quantum Fields on Branes"
"U. Moschella"	"Decomposing Quantum Fields on Branes"
"V. Gorini"	"Decomposing Quantum Fields on Branes"
"Itzhak Bars"	"Two-Time Physics in Field Theory"
"A. Kehagias"	"Hyperbolic Spaces in String and M-Theory"
"J.G. Russo"	"Hyperbolic Spaces in String and M-Theory"

8) Find the shortest path between ‘C.N. Pope’ and ‘M. Schweda’ authors (use any type of edges). Return the path and the length of the path. Comment about the type of nodes and edges of the path.

Query used to answer question:

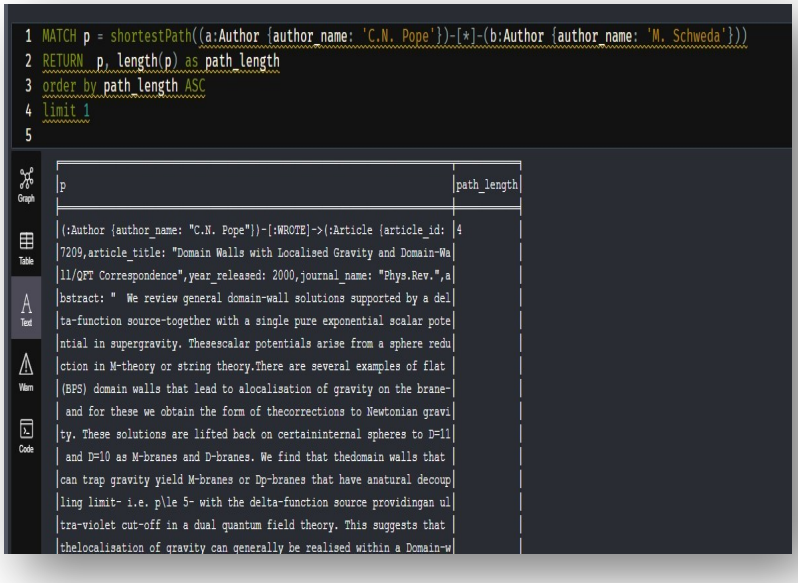
MATCH p = shortestPath((a:Author {author_name:'C.N. Pope'})-[*]-(b:Author {author_name: 'M. Schweda'}))

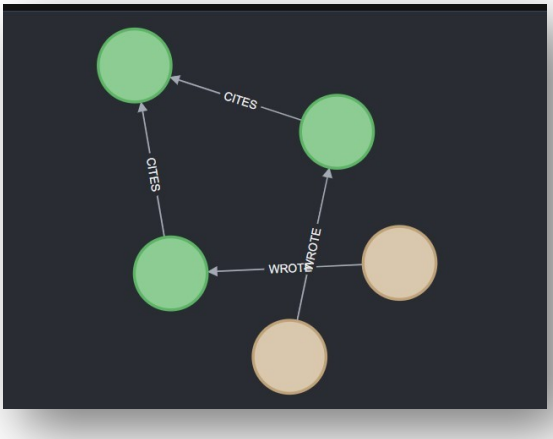
RETURN p, length(p) AS path_length

ORDER BY path_length ASC

LIMIT 1

Screenshot of result:





According to the output the faster way for between ‘C.N. Pope’ and ‘M. Schweda’ to be connected is the following: From node with id 119017, to node with id 140723, to node with id 119175. As we can see these nodes are articles that are related through the CITES relationship. Moreover, the top left node CITES the 2 other articles, creating a connection between the 2 authors.

1) Run again the previous query (8) but now use only edges between authors and papers. Comment about the type of nodes and edges of the path. Compare the results with query 8.

??/

10)Find all authors with shortest path lengths > 25 from author ‘Edward Witten’. The shortest paths will be calculated only on edges between authors and articles. Return author name, the length and the paper titles for each path.