Analyzing a Large Citation Network

Analyzing a Large Citation Network

Overview

Predicting the citations count of unpublished papers by a supervised learning approach from a highly-connected citation network.

Previous work

Most of the studies until now consider the number of citations a paper gets to rank its importance.

Approach:

- Study Design as corresponding factors (sample size, controls…), *Callaham, M., Wears, R.L., Weber, E. (2002)* [1]
- Topic analysis: more topics > more citations
- Author-related analysis: co-authors count, author's citations count

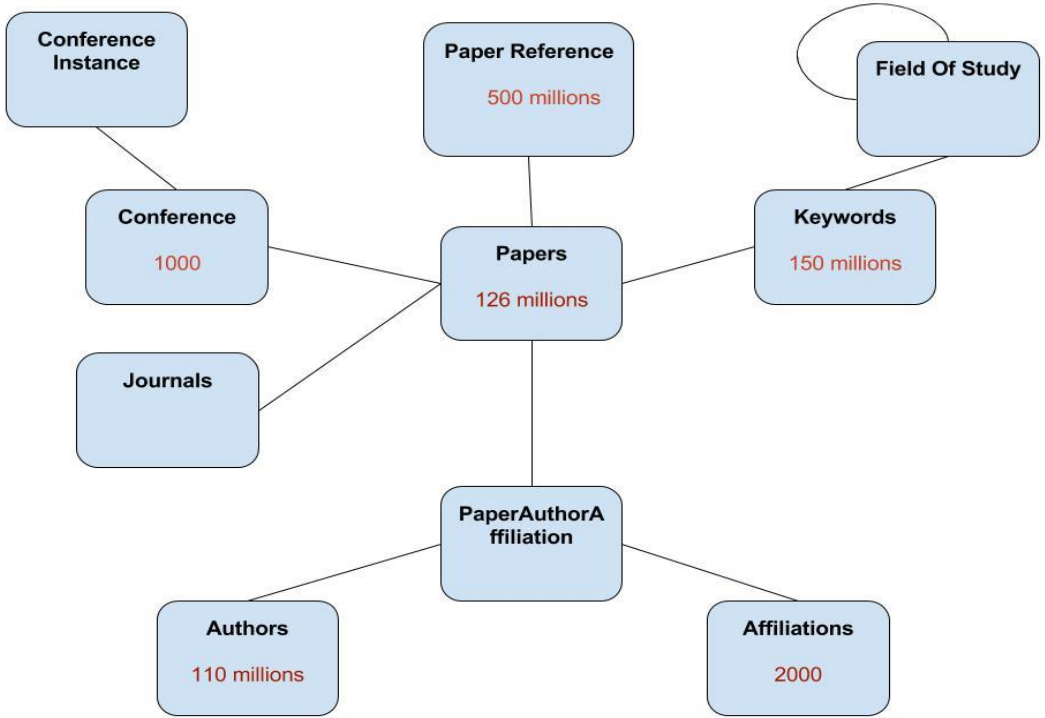
(1) Callaham, M., Wears, R.L., Weber, E. (2002) "Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals", JAMA, Vol. 287, pp.2847-50

Analyzing a Large Citation Network

Dataset

Microsoft Academic Service (MAS)

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june (Paul) Hsu, Kuansan Wang, 2015, http://dl.acm.org/citation.cfm?id=2742839



Analyzing a Large Citation Network

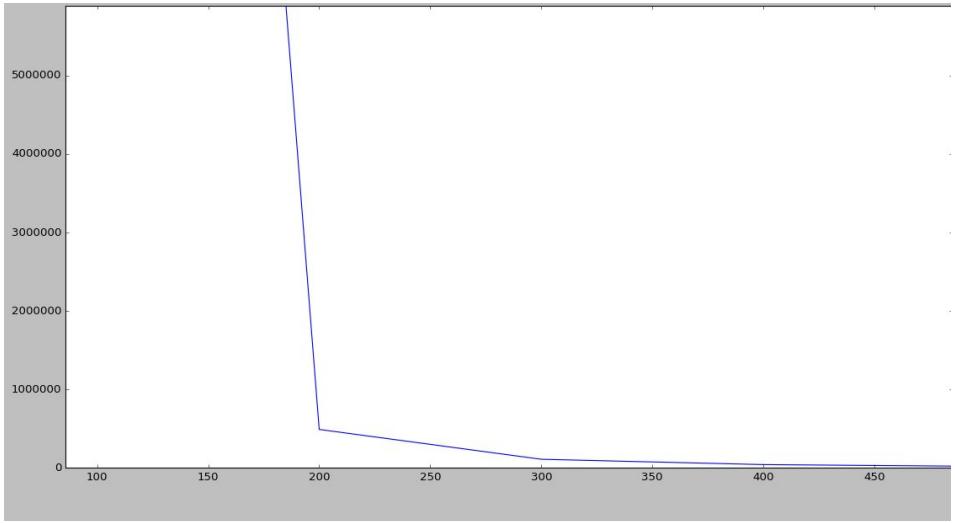
Approach

- **Paper importance is by number of citations**
- **Features to investigate**
  - Author Rank: number of citations an author gets from his published papers
  - Affiliation Rank: number of citations that an affiliation gets from its published papers
  - Conferences Rank: number of citations that an affiliation gets from its published papers
  - Fields of Study Popularity: number of citations that a field of study has

Analyzing a Large Citation Network

Pre-processing

Cleaning the data



The distribution of citations

X-axis: number of citations
Y-axis: number of papers

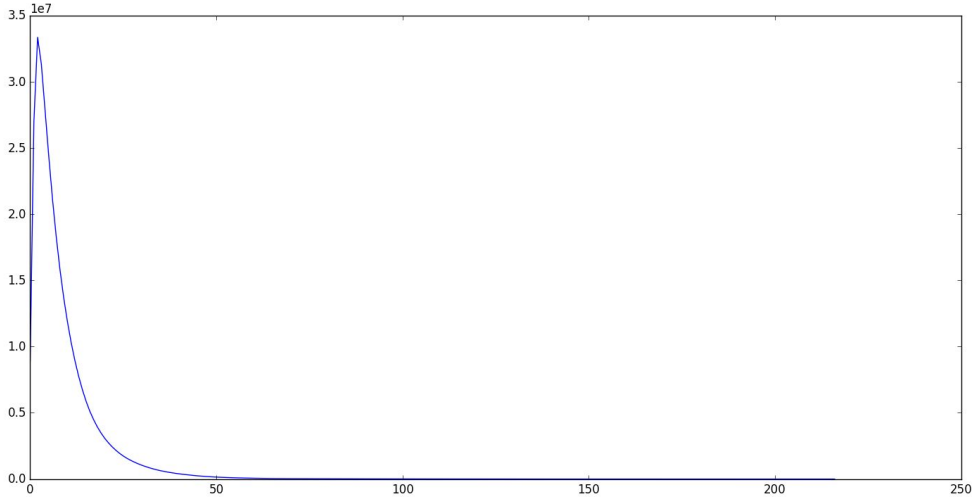
Pre-processing:

Take papers with citations < 200.

Analyzing a Large Citation Network

Pre-processing

Cleaning the data



The distribution of citations age

X-axis: citation age in years
Y-axis: number of papers

Pre-processing:

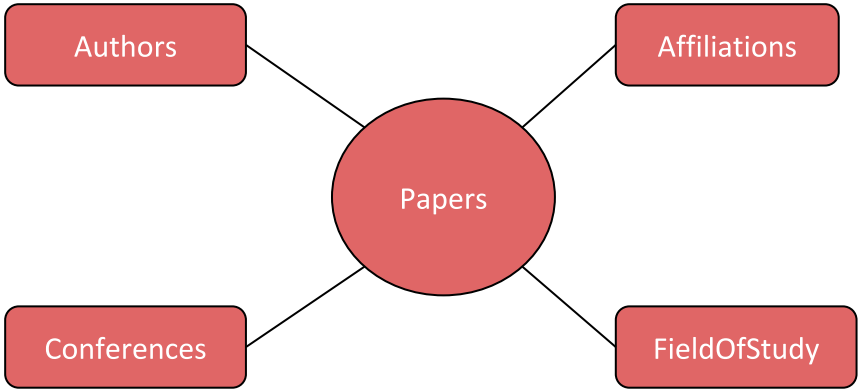
Take citations whose age < 3 years to balance old and new papers.

Analyzing a Large Citation Network

Feature investigation

Find feature weight vectors

First, we give each (author, conference, affiliation, field of study) in our dataset a weight based on the number of citations they have for their papers.



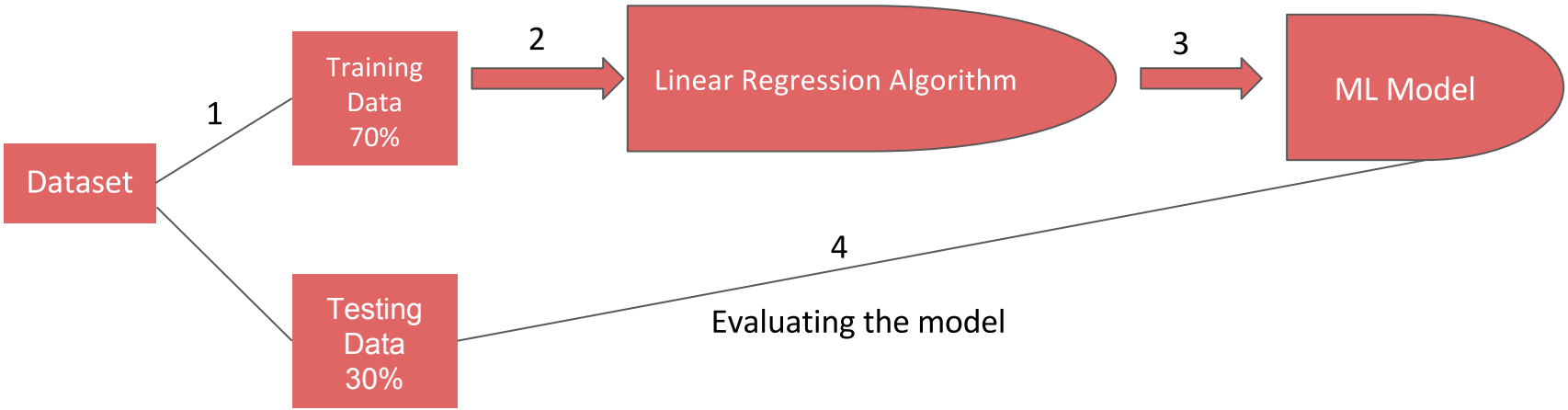
Example

paperCitations: how many times this paper was cited
authorSequence: position of the author in the paper (first, second, third, etc)

$$authorWeight = \sum_{authorPapers} (paperCitations * 1/authorSequence) / numberOfAuthorPapers$$

Analyzing a Large Citation Network

Building machine learning model



Analyzing a Large Citation Network

Technical Issues

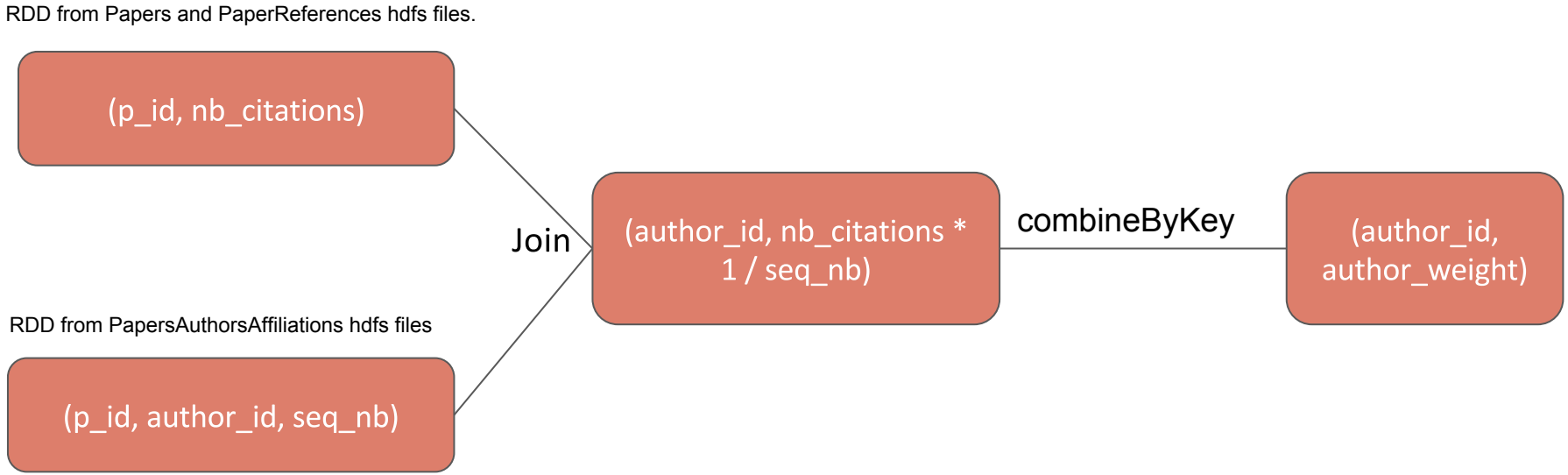
Using Apache Spark

- Configuration: 25GB memory for SparkContext
- Reading the data into Resilient Distributed Datasets (RDDs)
  - Another option was to use ApacheSpark DataFrames.
- Map/Reduce to process data
- Machine Learning Algorithm: LinearRegressionWithSGD

Analyzing a Large Citation Network

Feature Extraction

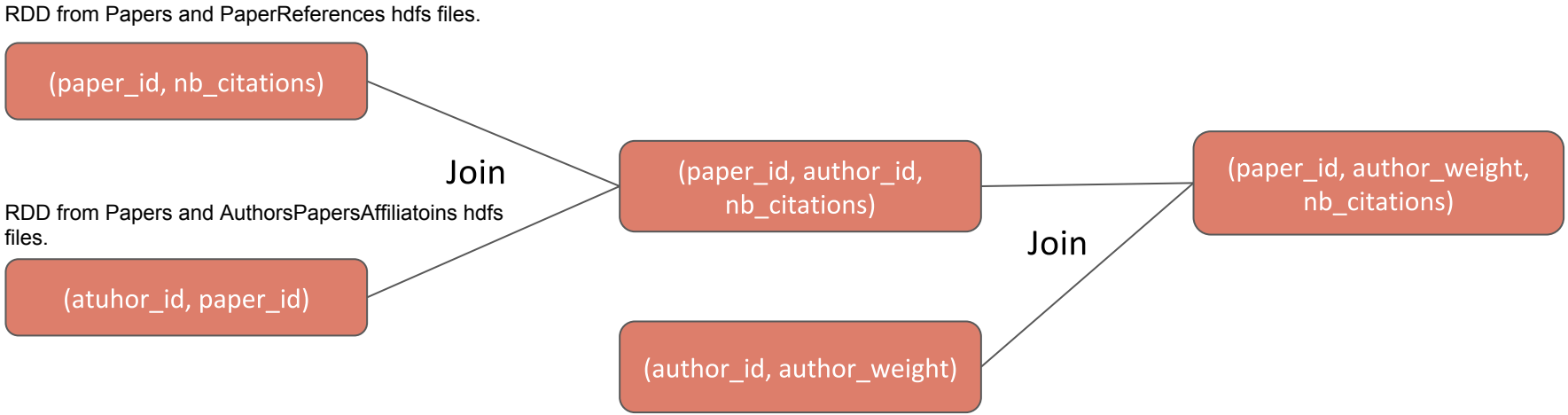
Example | Author feature (Step 1)



Analyzing a Large Citation Network

Feature File Construction

Example | Author feature (Step 2)



Analyzing a Large Citation Network

Results

Feature	Mean Square Error
Author	**549.694**
Affiliation	**639.167**
Field Of Study	**672.996**
Conference	**9.453**
(Author + affiliation + Affiliation + Field of Study + Conference) normalized	**417.55**

Analyzing a Large Citation Network

Future Work

- Investigating new features like **co-author** and **journal** or trying to do a weighted mix of these features

- Looking at the problem as a graph: citations from important papers should weigh more than other citations

- Building graph of field of studies and give the field of study a weight based on its **in-between degree**

- Tune the linear regression model parameters / Using other machine learning models

Analyzing a Large Citation Network

References

- **Microsoft Academic Service (MAS)** Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june (Paul) Hsu, Kuansan Wang, 2015

- **Predicting citation counts** Ron Daniel Jr, 2014

- **Mining of Massive Datasets** Anand Rajaraman and Jeffrey Ullman, 2011

- **Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals** Callaham, M., Wears, R.L., Weber, E., 2002