An analysis of scholarly citations in Wikipedia

Alessio Bogon

University of Trento
Department of Information Engineering and Computer Science

Master Degree in Computer Science

March 24, 2016

Overview

Introduction

Wikipedia

- Most used encyclopedia
- Started in 2001
- Studied by many
  - Quality of citations [1]
  - Illness prediction [2]
  - Stock market moves strategy [3]

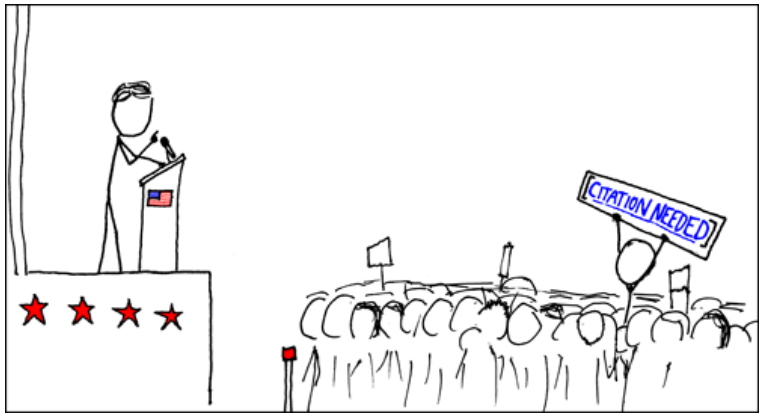

Figure: Wikipedian protester (xkcd.com/285)

Open questions

- Quality of papers in Wikipedia, in term of:
  - Incoming citations
  - Journals rank
  - Lifetime of citations
- Prediction systems
  - If a paper appears in Wikipedia, will it become more popular in the scientific community?
  - In another way, do researcher use Wikipedia as their primary data source?
  - Predict whether a publication is going to stay on a page

Available datasets

- Microsoft Academic Graph: a dataset containing all the papers every published with authors, references, journals, conferences, etc.
- Wikipedia dumps: text of all page revisions since the beginning
- Wikimedia hourly page view statistics

Microsoft Academic Graph

- Dataset size: 96 GB
- Contains over 120M papers (1800 – 2016)
- Information about authors, references, journals, conferences, keywords, etc.
- Problems
  - Only *computer science* conferences
  - Some of papers' publication dates are incomplete
  - Not all the papers have a DOI (32% of them)

The missing pieces

- History of papers appearing in Wikipedia (where and when)
- Usable/searchable page views dataset

Methodology?

Extracting citations from Wikipedia

Problems

- Citations can be structured: *wikimarkup* templates
  - Many different variants
  - Anybody can use custom macros
  - Different templates for each language
- or unstructured: plain text
  - Recognize substrings that appear to be citations
- Entity disambiguation
- Dataset size: 13,3 TB as of September 1st, 2015

Solution

- Focus on publication identifiers (*DOI*, *PMID*, *arXiv*, *ISBN*)
- The **wikidump** framework

Wikidump

- Facility framework to extract features from Wikipedia dumps
- Based on libraries by Aaron Halfaker
- Low memory consumption
- Highly parallelizable
- Written in Python

- Processed 445M page revisions (13,3 TB) in 21 hours

Type	Count
ISBN	1 153 330
DOI	651 199
PMID	372 939
PMC	79 841
arXiv	18 832

Table: Number of identifiers extracted

Wikimedia page views

Problems

- Dataset size: 23 TB (4,7 TB for the 2014)
- Aggregated and ordered by hour (8670 files per year)
- They need to be cleaned
- They need to be **reordered**
- Unfeasible on a single machine

Solution

- Exploit the UniTN Cisca Cluster

The Spark job

- UniTN Cisca Cluster
  - 125 workstations
  - 500 CPU cores in total
  - Available only at night and in the weekend
- The job
  - Normalize the content
  - Sample the first file
  - Repartition the keyspace
  - Sort each partition locally

- Took one night for the 2014 dataset
- Took many nights to get it to work

The Spark job — Workflow

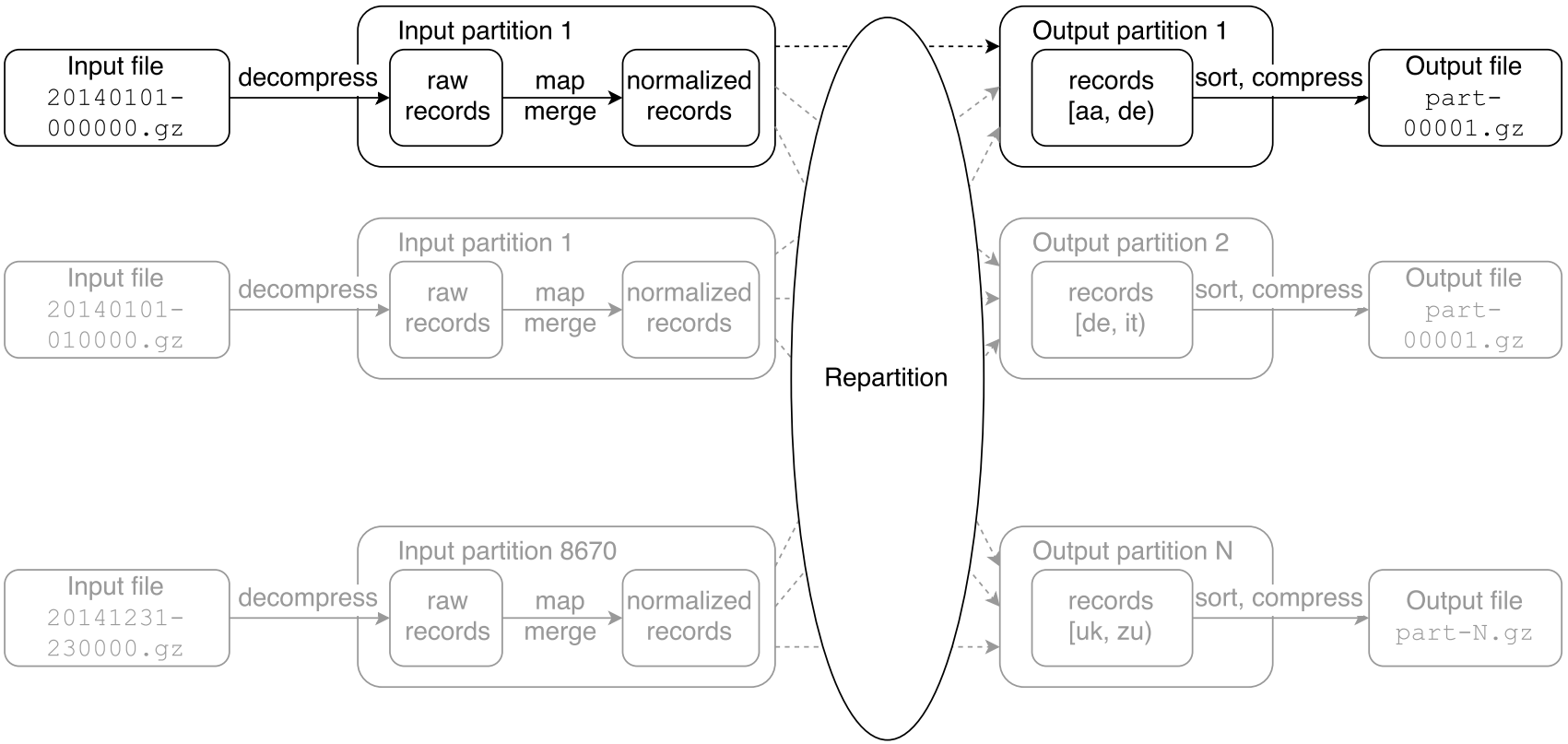

Figure: Workflow showing the processing of Wikimedia page views

Results

Conclusion


Finn Arup Nielsen.
Scientific citations in Wikipedia.
First Monday, 12(8):1–5, August 2007.


David J. McIver and John S. Brownstein.
Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time.
PLoS Computational Biology, 10(4), 2014.


Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis.
Quantifying Wikipedia Usage Patterns Before Stock Market Moves.
Scientific Reports, 3:1–5, 2013.