Management of scientific documents and visualization of citation relationships using weighted key scientific terms

Presenter: Shaopeng Wu

University of Bedfordshire

DATA 2016 Lisbon

Outline

- Introduction
- Data management
- Text Processing
- Data mining
- Visualisation
- Conclusion
- Acknowledgement

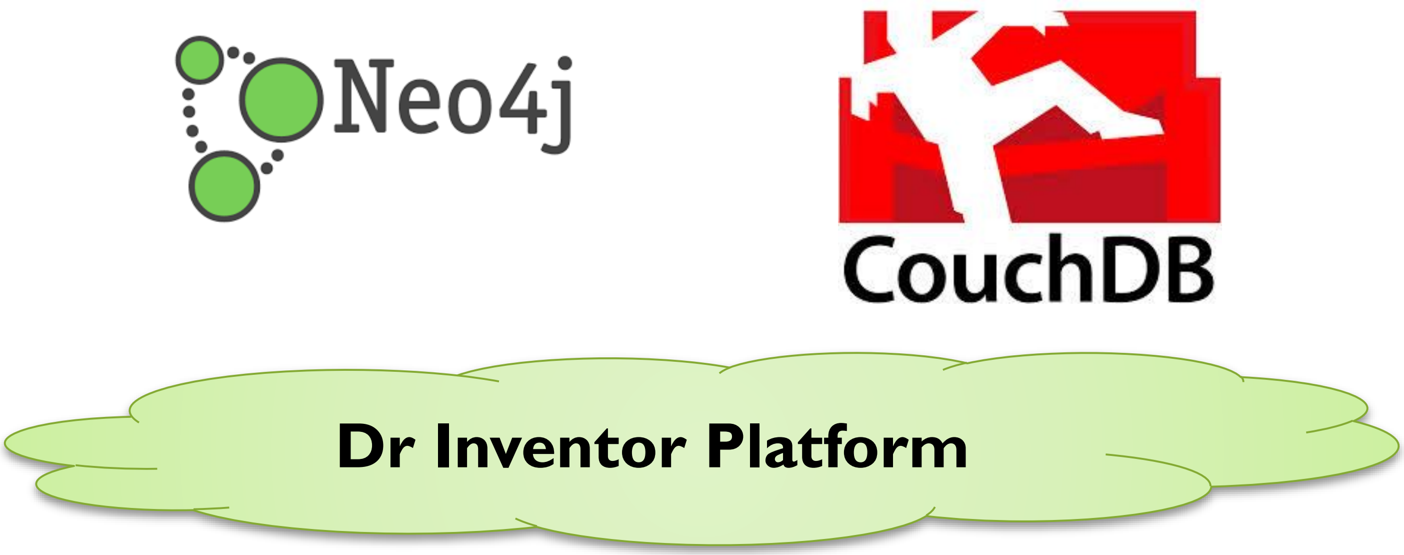
Objectives

- To manage scientific documents in big data platform
- To establish the citation paths among the documents in the repository
- To visualise the customised citation paths in directed graphs

Introduction

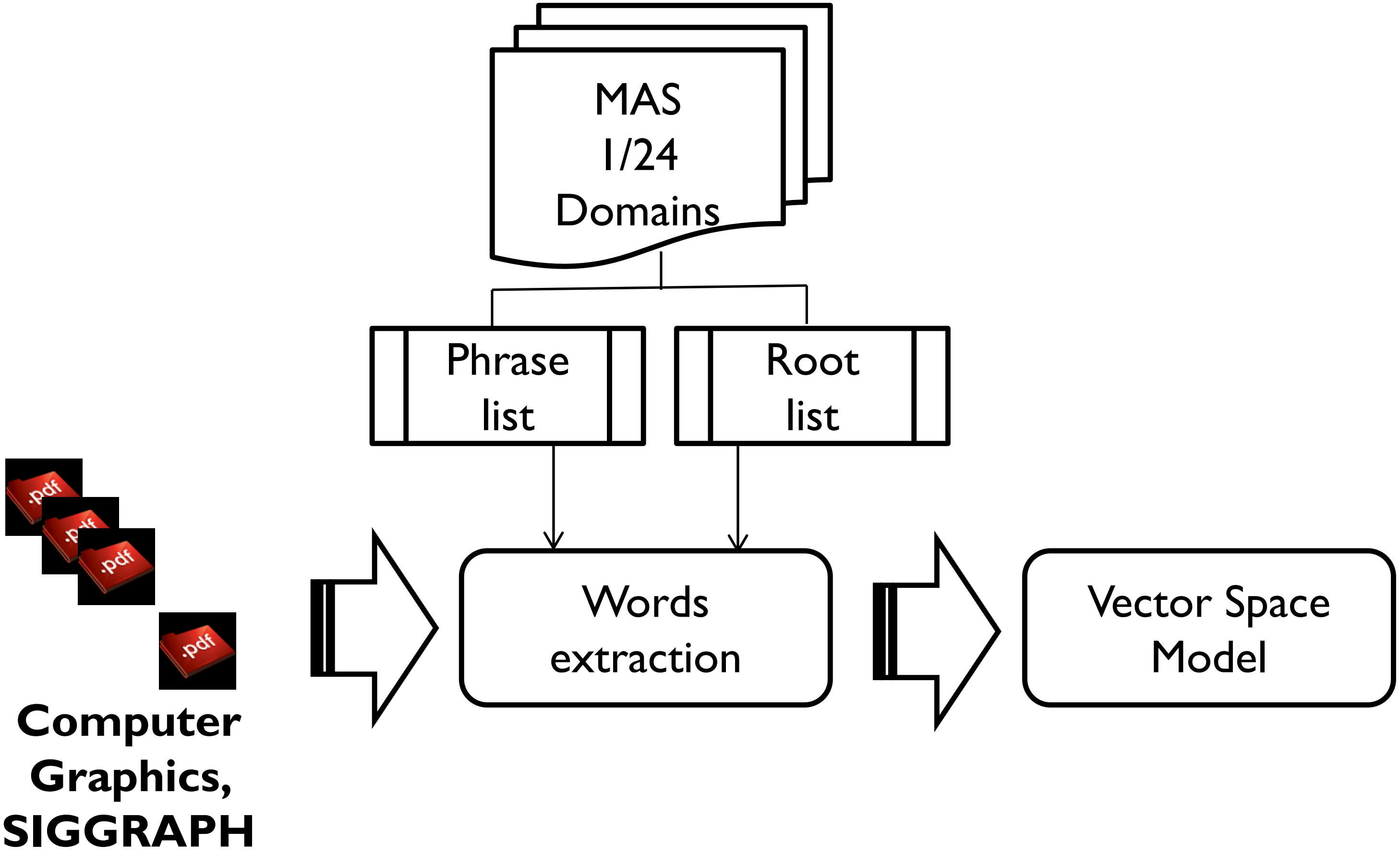
- Scientific documents are managed by big data platform Dr Inventor, by NoSQL database CouchDB, and graphic database Neo4J
- Topics are processed according to

Introduction Platform



Dr Inventor Platform

Introduction Text Processing



MAS
1/24
Domains

Phrase
list

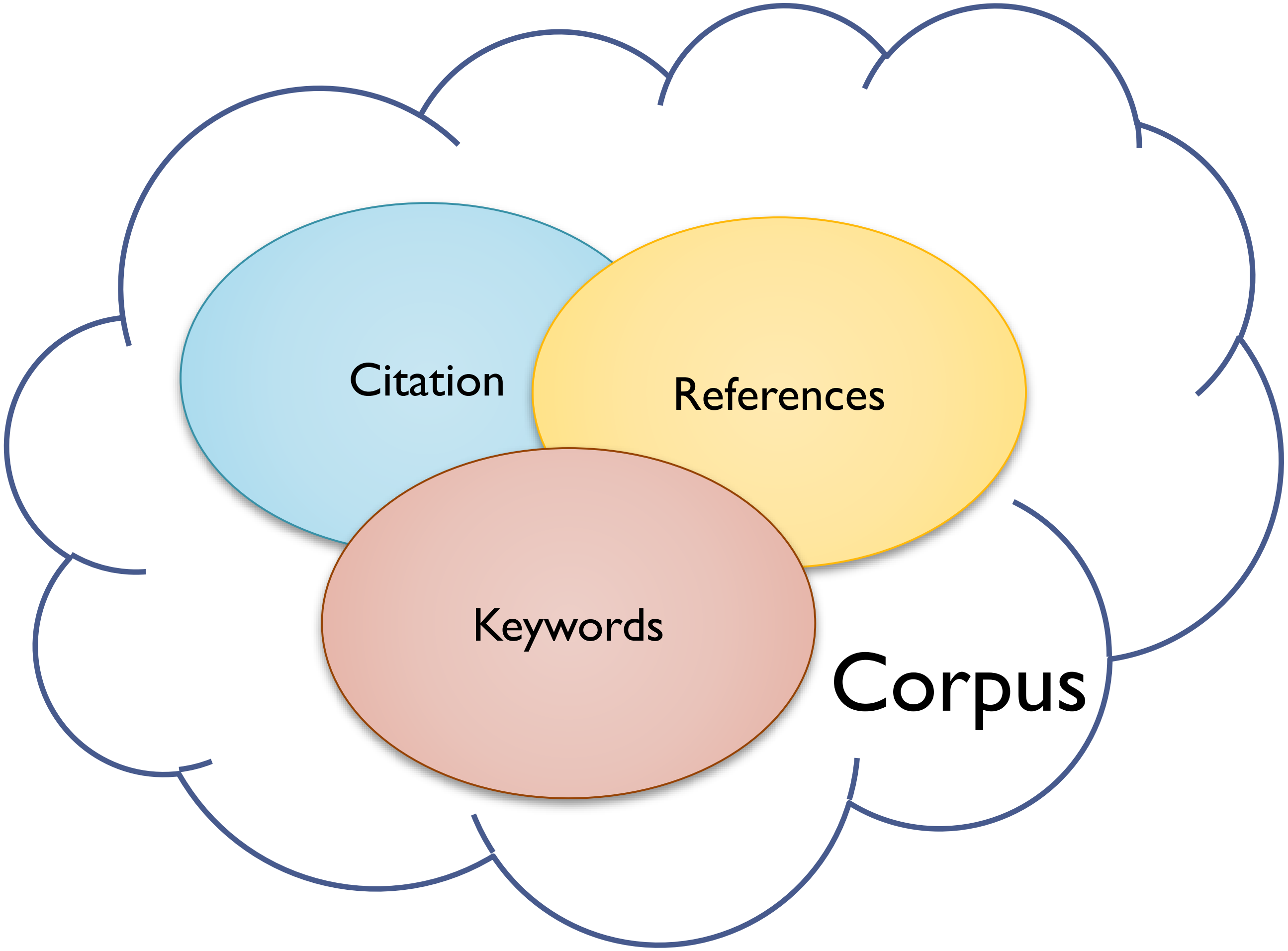
Root
list

Words
extraction

Vector Space
Model

Computer Graphics, SIGGRAPH

Data Management: Concepts



Repositories

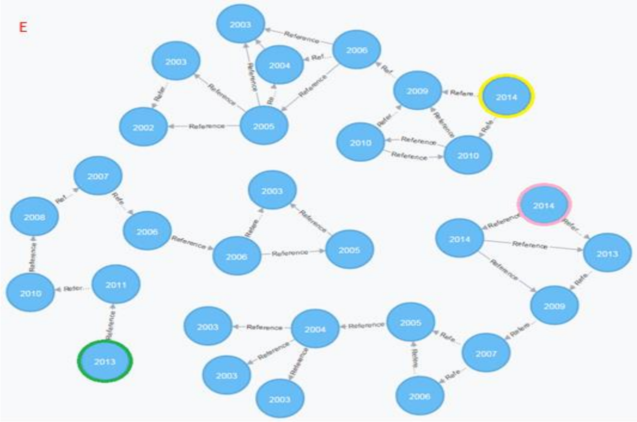
Virtual tables, docType

Validation

Reduce function for data aggregation

Elasticsearch for the full text search in a doc

Graph Repository



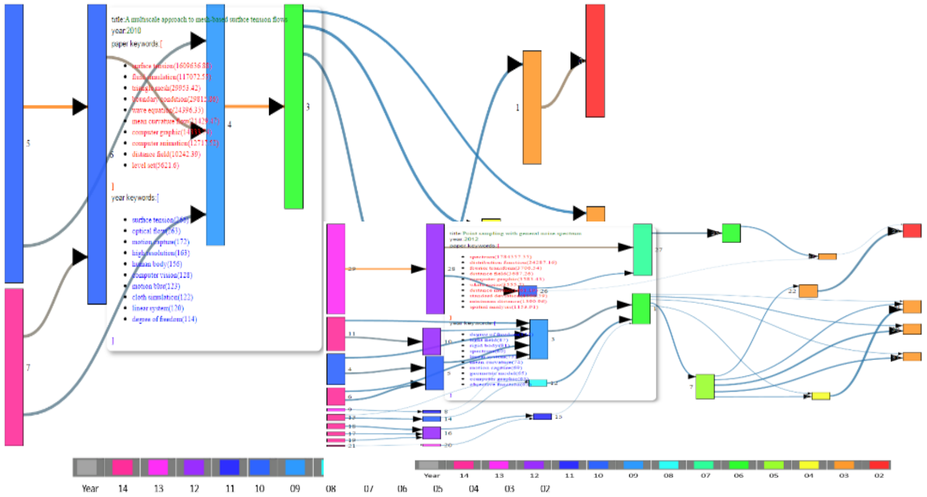
```
match p=(a:VolumePaper)-[r:Reference *3..]->(b:VolumePaper) RETURN Max(length( p))
match p=(a:VolumePaper)-[r:Reference*8..]->(b:VolumePaper) RETURN p,a.title
```

- Citation chain over years
- The length of the longest chain is 8
- Check and query the citations

Keyword term handling

- MAS API to obtain the keyword list
- Calculate the weight according to TF/IDF algorithm
  - Field term weighting
  - Citation term weighting
  - Term citation over years
  - Hierarchical word weighting
  - Citation distance

Visualisation of citation



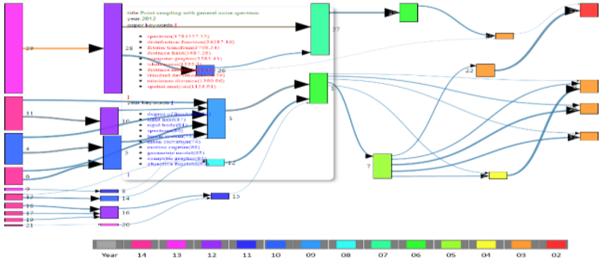
title:A multiscale approach to mesh-based surface tension flows
year:2010
paper keywords:[

- surface tension(1609636.88)
- fluid simulation(117072.55)
- triangle mesh(29953.42)
- boundary condition(29815.86)
- wave equation(24396.35)
- mean curvature flow(21429.47)
- computer graphic(14333.74)
- computer animation(12717.52)
- distance field(10242.39)
- level set(5621.6)

]
year keywords:[

- surface tension(260)
- optical flow(263)
- motion capture(172)
- high resolution(163)
- human body(156)
- computer vision(128)
- motion blur(123)
- cloth simulation(122)
- linear system(120)
- degree of freedom(114)

]



title:Point sampling with general noise spectrum
year:2012
paper keywords:[

- spectrum(1784337.33)
- distribution function(24287.10)
- fourier transform(2700.24)
- distance field(2687.26)
- computer graphic(1583.45)
- white noise(2233.7)
- distance metric(1801.19)
- standard deviation(1390.59)
- minimum distance(1300.06)
- spatial analysis(1173.01)

]
year keywords:[

- degree of freedom(1[illegible])
- light field(87)
- rigid body(81)
- spectrum(80)
- linear system(75)
- mean curvature(74)
- motion capture(69)
- geometric model(65)
- computer graphic(63)
- objective function(63)

]

Acknowledgement

Dr Inventor
(611383) FP7

DRINVENTOR

CARRE
(611140) FP7

CARRE

European
Commission