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

Neo4j

CouchDB

Dr Inventor Platform
Introduction Text Processing

Computer Graphics, SIGGRAPH

MAS
I/24
Domains

Phrase list
Root list

Words extraction

Vector Space Model
Data Management: Concepts

- Citation
- References
- Keywords
- Corpus
Repositories

Virtual tables, docType Validation
Reduce function for data aggregation
Elasticsearch for the full text search in a doc

Graph Repository

- 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
Acknowledgement

Dr Inventor
(611383) FP7

CARRE
(611140) FP7