ANALYSIS OF LARGE GRAPH DATA WITH GRADOOP AND KNIME

ALEXANDER KIPP (ROBERT BOSCH GMBH), STEFFEN DIENST, STEFAN KÜHNE (UNIVERSITÄT LEIPZIG), TOBIAS KÖTTER (KNIME)
Bosch Smart Semantics

Application fields

1) Expertise Finder

- Identifying Experts based on their real expertise
- Model trained with respective project documentation

2) Customer Insights

- Identification of relevant topics for users and the voice of customers
- Model trained with respective customer interactions (e.g. product reviews, Social Media Monitoring, ...)

3) Patent Analysis

- Identification of topic-related clusters and relations
- Model trained with patents
Bosch Smart Semantics

Application fields


4) Trendscouting

- Identifying relevant trends and mapping these to internal (technical) domains
- Model trained with external and internal data sources

5) Content as a Service (CaaS)

- Extracting content-based semantic footprints and mapping these to respective customer footprints
- Model trained with content and customer interactions
- Identifying user needs and respective, best suitable content / answer
- Model trained with user support interactions

6) Chatbot knowledge bases
Gradoop

An end-to-end framework and research platform for efficient, distributed and domain independent graph data management and analytics.
Gradoop

Graph Databases  
Graph Dataflow Systems  
Graph Processing Systems

Ease-of-use

Data Volume and Problem Complexity

© 2018 KNIME AG. All Rights Reserved.
KNIME® Analytics Platform
Over 2000 native and embedded nodes included:
BIGGR

Visual Workflows
ETL, Data Source-Adapter

+ Efficient distributed execution
of Graph-Analytics Workflows

= BIGGR
Big Graph Data Analytics Workflows

IKT 2020 - Forschung für Innovation: Förderkennzeichen 01IS16030
Patent Analysis: The Data

- Patent data from the US Patent and Trademark Office
- More than 6 million patents and 90 million citations
- Patents and assignees are represented as nodes
- Citations and authorship are represented as directed edges
Patent Analysis: Graph Centralities

**Top 10 Patents by Indegree**

- Process for amplifying nucleic acid sequences
- Mutant dwarfism gene of petunia
- Process for amplifying, detecting, and/or-cloning nucleic acid sequences
- Expandable intraluminal graft, and method and apparatus for implanting an expandable intraluminal graft
- Systems and methods for secure transaction management and electronic rights protection

**Top 10 Patents by PageRank**

- Method of transmitting information and multiplexing device for executing the method
- Process for producing biologically functional molecular chimeras
- Data sending and receiving system for packet switching network
- Method of automatically evaluating source language logic condition sets and of compiling machine executable instructions directly therefrom
- Transaction execution system with secure data storage and communications

Before 1980

- Node size: cites IPC / total cites
- Node color: % within IPC cites
- Link width: cites IPC / total outgoing cites
- Link color: % change w.r.t. prev. time frame

1980-1990

- **Node size:** cites IPC / total cites
  - Small: <50%
  - Medium: 50%-70%
  - Large: >70%

- **Node color:** % within IPC cites
  - Red: <50%
  - Orange: 50%-70%
  - Blue: >70%

- **Link width:** cites IPC / total outgoing cites
  - Narrow: <5%
  - Medium: 5%-15%
  - Broad: >15%

- **Link color:** % change w.r.t. prev. time frame
  - Red: <5%
  - Purple: 5%-10%
  - Green: >10%

- Key Patent Classes:
  - Textiles, Paper
  - Fixed Constructions
  - Mechanical Engineering, Lighting, Heating, Weapons
  - Physics
  - Electricity
  - Performing Operations, Transporting
  - Human Necessities
  - Chemistry, Metallurgy

1990-2000

- **Node size:** cites IPC / total cites
- **Node color:** % within IPC cites
  - <50%
  - [50%, 70%]
  - >70%
- **Link width:** cites IPC / total outgoing cites
- **Link color:** % change w.r.t. prev. time frame
  - <5%
  - [-5%, 5%]
  - >5%

2000-2010

2010-2017

Node size: cites IPC / total cites

Node color: % within IPC cites
<50% [50%,70%] >70%

Link width: cites IPC / total outgoing cites

Link color: % change w.r.t. prev. time frame
<5% [-5%,5%] >5%

Electricity Deep Dive

- **Node size:** cites IPC / total cites
  - Small: <50%
  - Medium: [50%, 70%]
  - Large: >70%

- **Node color:** % within IPC cites
  - Red: <50%
  - Orange: [50%, 70%]
  - Blue: >70%

- **Link width:** cites IPC / total outgoing cites
  - Thin: <5%
  - Wide: [5%, 15%]
  - Thick: >15%

- **Link color:** % change w.r.t. prev. time frame
  - Red: <5%
  - Orange: [5%, 15%]
  - Green: >15%
Cypher query:
MATCH (p1:PATENT)<-[e1: citation]-(p2:PATENT)
(p1)-[: assignedBy]->(a1:ASSIGNEE)<-[:assignedBy]-(p2)
WHERE e1.years_difference = 0

- Headset
- AliphCom
- Expanding speaker base
- Speaker grill
- Portion of an electronic device
Summary

• Graphs can be applied to many different use cases
• Gradoop is a powerful distributed graph processing framework
• KNIME provides an easy to use workflow based user interface
• BIGGR combines them to a powerful and easy to use graph analysis solution