ADAMiSS: Universal system for data analysis

Jakub Peschel

Masaryk University

October 3, 2019

Overview

1. Introduction
   - Motivation
   - Goal
2. Architecture
   - ADAMiSS
   - Storage
   - Transaction DB
   - Analytical Operators
3. UseCases
4. Experiments
   - Frequent Item-set Mining
   - Sequence Mining
   - Similarity Searching
5. Summary
   - Summary

Motivation

- Large volumes of data

Motivation

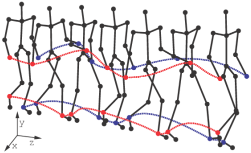
- Large volumes of data
- Lot of different types of data

Motivation

- Large volumes of data
- Lot of different types of data

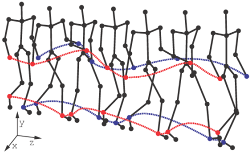
Motivation

- Large volumes of data
- Lot of different types of data



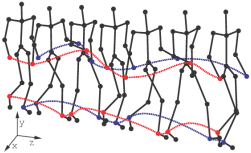
Motivation

- Large volumes of data
- Lot of different types of data



Motivation

- Large volumes of data
- Lot of different types of data
- Unstructured or semi-structured



Goal

- Universal analytical tool
    - Universality in type of input
- Unstructured data

- Test

Goal

- Universal analytical tool
  - Universality in type of input
- Unstructured data
- Analysis of data

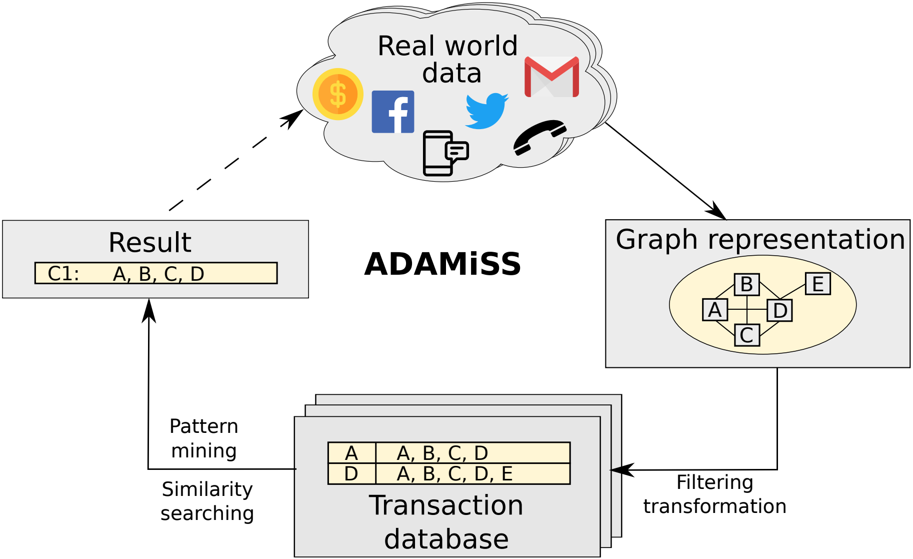
- Test

Goal

- Universal analytical tool
  - Universality in type of input
- Unstructured data
- Analysis of data
  - by similarity
  - by pattern mining

- Test

ADAMiSS: Overview



Real world data

Graph representation

Filtering transformation

Transaction database

A	A, B, C, D
D	A, B, C, D, E

Pattern mining

Similarity searching

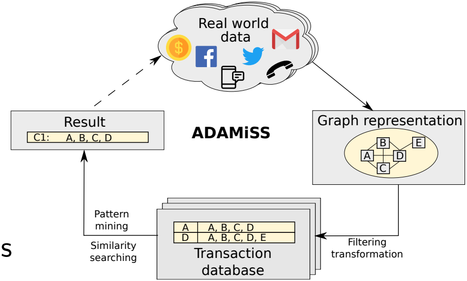
Result

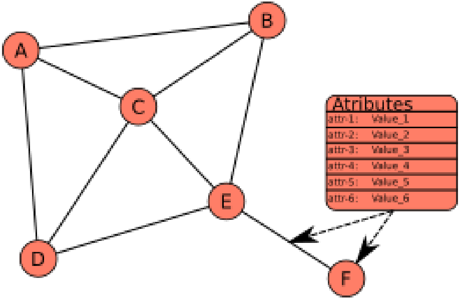
C1: A, B, C, D

ADAMiSS

Storage - Graph representation

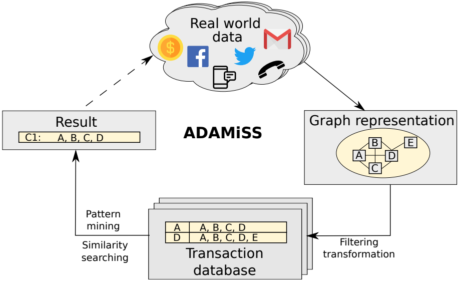
- Multigraph
- Nodes and edges have attributes
- Graph as unifying structure





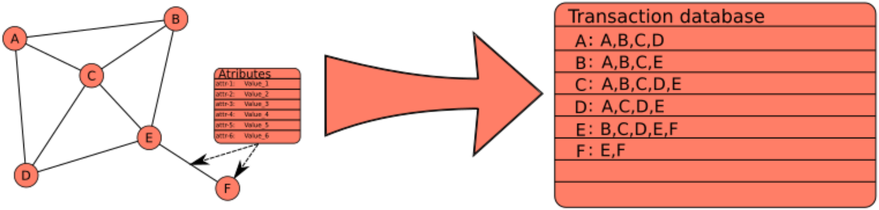
Transaction DB

- Flat structure
- Unified for analytical operators
- Transformed from graph
- Filtration based on
  - Properties of graph
  - Attributes



C1:	A, B, C, D

A	A, B, C, D
D	A, B, C, D, E



Transaction database

A: A,B,C,D

B: A,B,C,E

C: A,B,C,D,E

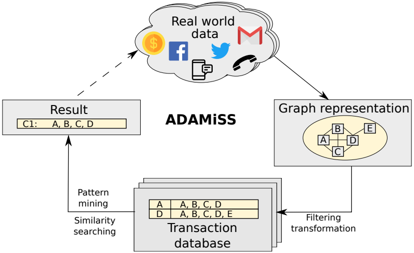
D: A,C,D,E

E: B,C,D,E,F

F: E,F

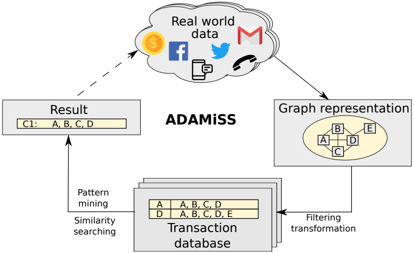
Analytical Operators

- Input: transaction database
- Pattern mining and similarity search
  - Strong analytical tools
  - Pattern mining discover unknown
  - Similarity analysis looks for known



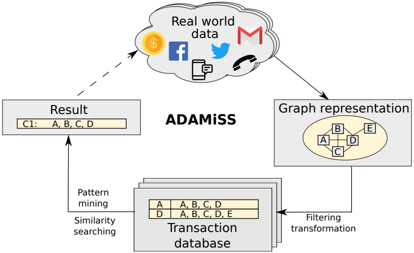
Analytical Operators

- Input: transaction database
- Pattern mining and similarity search
  - Strong analytical tools
  - Pattern mining discover unknown
  - Similarity analysis looks for known
- Pattern mining
  - Frequent itemset mining
  - Sequence mining
  - Association rule mining



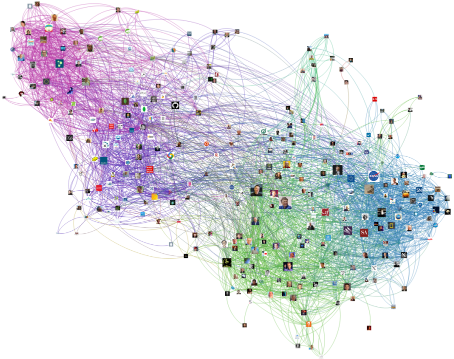
Analytical Operators

- Input: transaction database
- Pattern mining and similarity search
    - Strong analytical tools
    - Pattern mining discover unknown
    - Similarity analysis looks for known
- Pattern mining
    - Frequent itemset mining
    - Sequence mining
    - Association rule mining
- Similarity in metric space
    - K-nn query
    - Range query
    - Similarity join



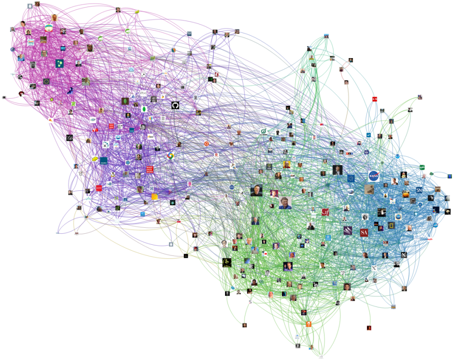
Use-cases

- Management of social network community



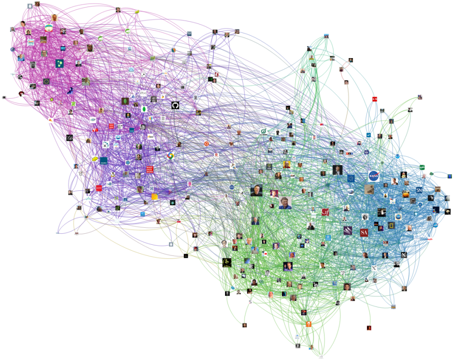
Use-cases

- Management of social network community
  - Group detection



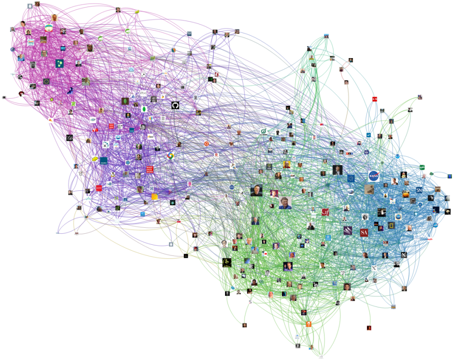
Use-cases

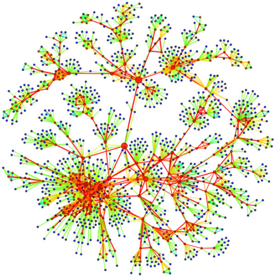
- Management of social network community
  - Group detection
  - Communication flows



Use-cases

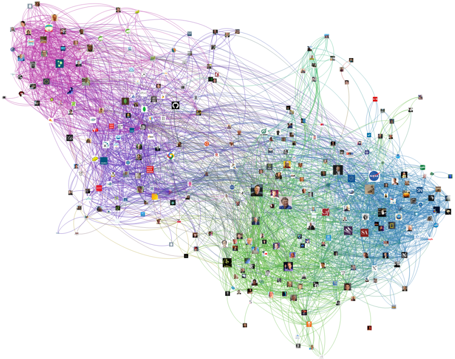
- Management of social network community
    - Group detection
    - Communication flows

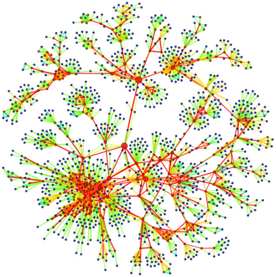




Use-cases

- Management of social network community
    - Group detection
    - Communication flows
    - Duplicate accounts detection





Frequent Item-set Mining

- **Data**: Twitter Higg's boson dataset
- **Size**: 304 691 interactions on Twitter
- **Process**:
  1. Creation of a graph
  2. Lists of neighbours as transactions
  3. Analytical method: frequent item-set mining (FP-Growth)
  4. Threshold for analysis: 11
- **Results**:
  - 7 communities of size 12
  - 94 communities of size 11

Sequence Mining

- **Data**: Kosarak dataset
- **Size**: 990 000 click-streams through Hungarian news web
- **Process**:
  1. Creation of a graph
  2. Stream by one user as a transaction
  3. Analytical method: sequence mining (GSP)
  4. Threshold: 1024
- **Results**:
  - Discovered 322 paths
  - 5 paths contained more than 4 nodes
  - Longest path has 16 nodes

Similarity Searching

- **Data**: Twitter Higg's boson dataset
- **Size**: 304 691 interactions on Twitter
- **Process**:
  1. Creation of a graph
  2. Lists of neighbours as transactions
  3. 12 nodes of randomly selected community as K-nn queries and range queries
  4. K for K-nn query is 10, distance for range query is 0.2 (M-index)
- **Results**:
  - Four nodes has most similar items inside community
  - One node has all ten outside of the community
  - Average amount of query nodes in range query results is 8.33 nodes

Summary

- What is goal?
  - Universal system for analysis of data
  - Analytical tools from area of pattern mining and similarity search
- What we propose?
  - Advanced Data Analysis by Mining and Searching System
  - Graph representation for capturing all the information
  - Transaction database as easily process-able format
  - Analytical operators: pattern mining, similarity search, etc.

- What it is for?
  - analysis of communities
  - analysis of sequences
  - exploration by similarity searching
- What has been done?
  - Datasets: Twitter Higg's boson, Kosarak
  - Analysis of communities
  - Analysis of sequences
  - Similarity of neighbourhood of community members

Acknowledgements

This work has been supported by the Ministry of the Interior of the Czech Republic under the "Security Research for the Needs of the State Program 2015-2020," through the Project No. VI20172020096, "Complex Analysis and Visualization of Large-scale Heterogeneous Data."