Social Data Collection and Processing Framework

Dmitry Gushchanskiy
Alexander Bogdanov
Alexander Degtyarev
Saint-Petersburg State University

Political Psychology

P. Psych. {
- *Psychology*
- *Sociology*
- *Political Science*

P. Psych. & Social Networks

Actions to Simplify:

- Content-analysis automation
- Social network analysis

Arising Tasks

- Keywords Extraction
- Sentiment Analysis
- Graph Stats Calculation
- Graph Queries Execution

What Kind of Data Is Gathered

Source: vk.com

- Users
  - Sex
  - Birth date
  - City
- Groups
- Posts
- Comments

Data Acquisition

- Source – API VKontakte

Data Acquisition

- Source – API VKontakte
- API call restrictions

Data Acquisition

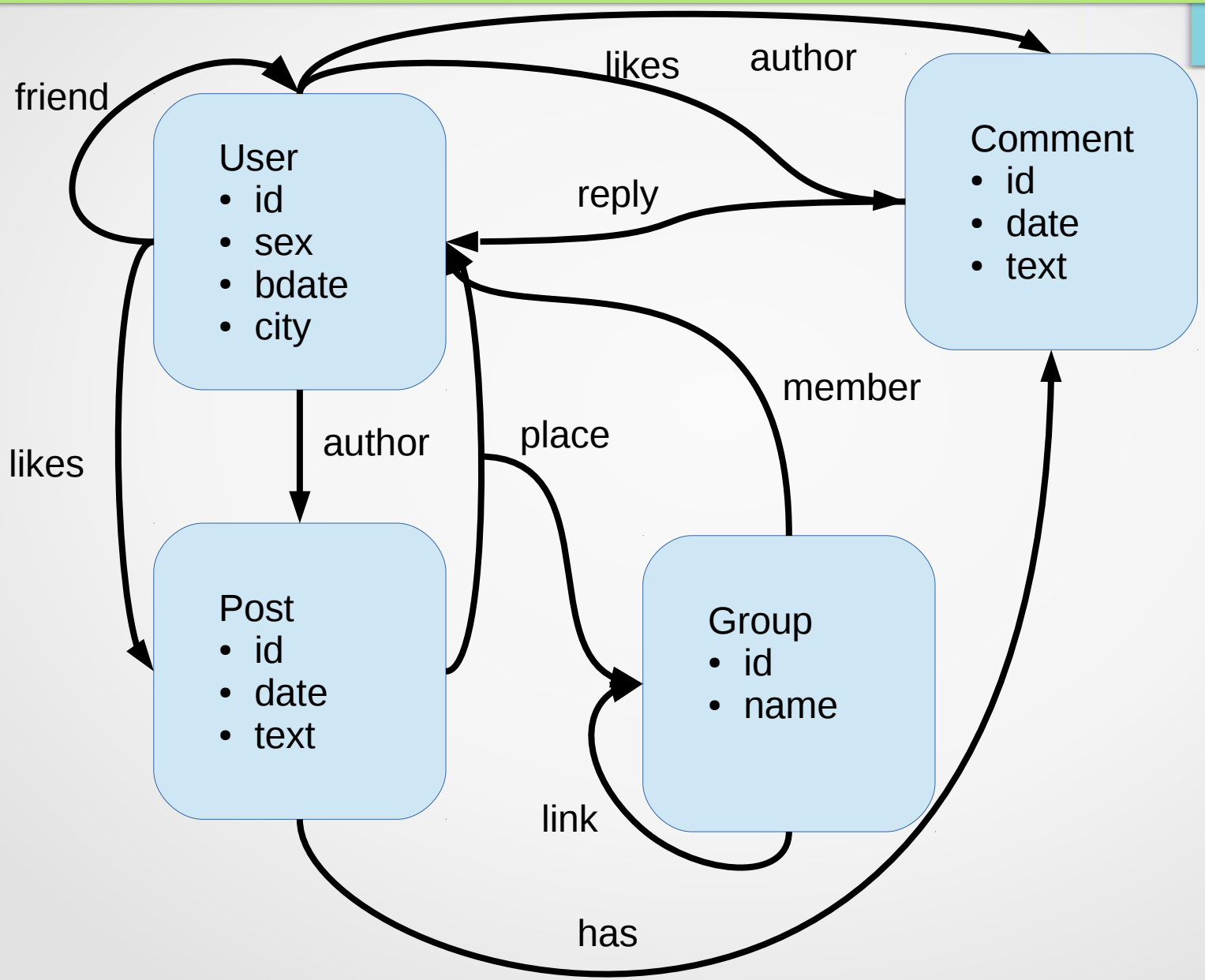
- Source – API VKontakte
- API call restrictions
- Easy manageable amount of data

Solution. Approach

- There is no Big Data here
- No need to worry about distribution, parallel computing and so on

Solution. Tools

Solution. Data



Solution. Initial Request

Group/User + (Time, Links, Depth, ...)

Solution. Processing Options

- Relation search via Neo4j pathfinding
- Sentiment analysis via I-Teco API (free version)
- Keyword extraction via AlchemyAPI (free version)

Test

Test data: the graph of 117 groups about Saint-Petersburg, especially its architectural and communal disputes.

Groups:

- “Красивый Петербург”
- “Скажем 'Нет!' Газпром-Сити!”
- “ВЦКП, квартплата, ЖКХ (Санкт-Петербург)”
- ...

Test Results

Posts from a group	Comments for a post	Gathered posts	Gathered comments	Users	Processing time (minutes)
100	100	9717	11263	76421	15,68
500	100	34826	41766	88067	155,78

Solution. Processing Options

- Relation search in no time
- Sentiment analysis is too slow for mass appliance
- Keyword extraction has tolerable speed, but limited free usage

Possible Improvements

- User interface, Graph visualization
- Integrated text processing options (Lucene)
- Further adaptation for psychologists' needs
- Adaptation for larger scale problems

Thank You!