Do-It-Yourself Analytics on Open Data

Katrin Braunschweig
Julian Eberius
Maik Thiele
The open society is a concept originally developed by philosopher Karl Popper.

In open societies, government is responsive and tolerant, and political mechanisms are transparent and flexible.

The state keeps no secrets from itself in the public sense.

It is a non-authoritarian society in which all are trusted with the knowledge of all.

KARL POPPER
THE OPEN SOCIETY AND ITS ENEMIES
Why Data Should be Open

- Many scientific data can be deemed to belong to the commons ("the human race"), e.g. the human genome, medical science, environmental data
- They have an infrastructural role essential for scientific endeavour (e.g. in Geographic Information Systems and maps)
- Data published in scientific articles are factual and therefore not copyrightable
- Public money was used to fund the work and so it should be universally available
- It was created by or at a government institution
Civic Applications based on Open Data

Database Technology Group

> Civic Applications based on Open Data

Civic Applications based on Open Data

© Prof. Dr.-Ing. Wolfgang Lehner

Do-It-Yourself Analytics on Open Data
Explore How U.S. Budget Proposal

Obama’s 2012 Budget Proposal: How $3.7 Trillion is Spent

Explore every nook and cranny of President Obama’s budget proposal.

Rectangles are sized according to the proposed spending. Color shows severity of cut or increase from 2010.

The president has proposed a five-year freeze of discretionary spending, excluding national security spending. This type spending accounts for about one-fifth of all spending.

Mandatory spending, which includes entitlement programs like Medicare and Social Security, is expected to continue to rise.

The proposal includes cuts to low income home energy assistance and community service block grants.

Cuts in the Environmental Protection Agency’s budget include reducing funds restore the Great Lakes’ environmental health.

Mr. Obama wants to spend more to train math, science and engineering teachers and to expand effective programs.

The Energy Department’s budget is

Published: February 14, 2011 | By SHAN CARTER and AMANDA COX | Source: Office of Management and Budget
FLUGLÄRM-KARTE BBI-AIRPORT BERLIN

VORSCHLAG 03/2011  VORSCHLAG 09/2010  IST-SITUATION

WESTWIND (65% aller Flüge)  OSTWIND (35% aller Flüge)

DAUERSCHALLPEGEL  MAX. SCHALLPEGEL  TAZ-ARTIKEL

Database Journalism
Explore the world

Gapminder World shows the world's most important trends.

- Wealth & Health of Nations
- CO₂ emissions since 1820
- Africa is not a country!
- Is child mortality falling?
- Where is HIV decreasing?

Load Gapminder World

Rosling’s Twitter

“Gender gap in years in school is 2-fold in poorest countries, and male gap between poorest & richest countries is 5-fold


Subscribe to newsletter

Subscribe to the newsletter for the latest Gapminder info.

http://www.gapminder.org/

Reducing child mortality – a moral and environmental imperative
September 27, 2010

Population growth explained with IKEA boxes
August 26, 2010

Gapminder Labs

Gapminder Agriculture
700 indicators from the Food and Agriculture Organization (FAO).

Gapminder USA
Compare Washington, Minnesota and Texas with countries of the World.

News

- Card Game on global gaps
- PowerPoint on life expectancy
- Finding data in Gapminder World just got easier...
- Teach with Gapminder World

More news

Do-It-Yourself Analytics on Open Data
Vision: making sense of the world by having fun with statistics!
  - Gapminder is a non-profit venture for development and provision of free software to visualize human development trends
  - Gapminder will ultimately be integrated into Google: this is the first time global datasets will be searchable over the Internet

Hans Rosling @ TED
  - TEDTalks: annual technology conference in California, USA
    http://www.ted.com/tedtalks/
  - Hans Rosling is a professor of global health at the Karolinska Institute, data visualization extraordinaire and the creator of the Gapminder tools
The Flux is a snapshot of the world population. Watch the wave of education sweep the world using the Year Slider. Some countries have done better than others. Click the colours to explore.

http://www.betterworldflux.com
Maps - find your way

A student project belonging to the Edyra research project at Database Technology Group of the TU Dresden.

Choose the minutes in which you want to reach a point from your position.

Minutes: 15

moving possibilities

with public transport (DVG, VVO)
and
○ on foot or
○ with bike

help
Mobile Application

- Add new requests by guiding the user through a wizard-style input form
- Show (own) reports and there current rating and processing actual state
- Visualize all reports on a map
- Subscribe to a set of urban district and notify the user about news

Web Application

- Filter the requests by their category, their creation time (last 24 hours, last week, last month, all)
- Change the requests state (open, closed, closed) for authorized users
- Zoom in/out and adapt the type of visualization if the issue density gets very sparse
Open Civic Platform for Dresden (3)
What we are doing at EDYRA

http://wwwdb.inf.tu-dresden.de/edyra/
Open Data Survey
Open Data Survey

Why we did it?
- Because we build on it within our project
- Looking at real open data platforms is really disillusioning

Goal?
- Quantify the “open data trend“
- Find new challenges/argumentations for our project
- Determine a “definition of openness“ → See how a perfect open data repository would look like

How?
- Manual extraction of measures for about 50 open data repositories
- Looking at five repositories in detail by applying a set of crawlers → makes the survey repeatable
- Type of measures
  - Platform measure: number of datasets, meta data, data format, API‘s, licences, structure of the catalog, ...
  - Datasets: cardinality, number of attributes, quality, granularity, ...
Challenges

- Lots of contributors / maintainers
- Small information pieces → distributed, decentralised and very loosely coupled
- No standards, regarding file formats, meta data, data description, global tag sets, platform (CKAN is de facto standard)
- Different degree of schema information and meta data
- Innovative and unexpected reuse
- Facilitating non-expert users to analyze open data
- Open data suffer from poor quality → data quality awareness
- Business intelligence on Open Data requires data preprocessing → difficult task for a non-experts too
- Non-experts (as citizens) need guidance during data preprocessing/ integration and data exploitation
- Data provenance
Related Work

- Keyword Search Over Relational Databases / XML
- Query Segmentation
- Linked Open Data
- Web Search
- Public Data Integration
- (End-user) Visualization
- (End-user) Mashup Creation
- Web Data Platforms
- Natural Language Interfaces
- Graph Databases
- Information Retrieval Techniques
Our Approach – Big picture

Open Data Exploitation

Keyword phrase

SQL Dialect

Private Database

Schema and meta data catalog

Open Data Sources

RDF
XML
CSV
JSON

Open Data Integration

Semi-automatic Schema Annotation

Crowd-based Schema Augmentation

Do-It-Yourself Analytics on Open Data
Data Integration
Open Data – data.gov.uk

Opening up government

Looking for something specific, or just want to know more about how Government spends your money? You’ll find over 7,500 datasets to help you get answers.

List of Ministers’ Interests

Under the terms of the Ministerial Code, ministers must ensure that no conflict arises, or could reasonably be perceived to arise, between their ministerial position and their private interests, financial or otherwise. On appointment to each new office, ministers must provide their Permanent Secretary with a list in writing of all relevant interests known to them which might be thought to give rise to a conflict. The following list contains relevant interests of the ministers concerned current at the date of publication. Where a minister has disposed of a relevant interest, or where they did so before taking up ministerial office, it is not included in the List.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Department</td>
<td>Name</td>
<td>Role</td>
<td>Category</td>
</tr>
<tr>
<td>2</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>3</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>4</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>5</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>6</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>7</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>8</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>9</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>10</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>11</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>12</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>13</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>14</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>15</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>16</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>17</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
<tr>
<td>18</td>
<td>Prime Minister’s Office</td>
<td>Rt Hon David Cameron MP</td>
<td>Prime Minister</td>
<td>6. Charities</td>
</tr>
</tbody>
</table>
Open Data Problems

Meta data quality

- No standards (no defined set of categories or tags)
- Incompleteness (missing publisher, tags, ...)
- Too generic versus too specific

Data quality

- No standards, no global schema (file formats, calendar definitions, naming conventions, currencies, data type e.g. NULL)
- Structured, semi-structured and unstructured data
- Incompleteness (empty fields, row, columns)
- Often denormalized

Search over open data is restricted on meta-data only
Semi-automatic Schema Annotation

Database Technology Group

Semi-automatic Schema/Semantic Annotation

Data Cleansing

Entity Extraction

Dedublication

Relationship Detection

Annotation

Indexing

Precision Model

Collaborative Schema Augmentation

Crowd

Do-It-Yourself Analytics on Open Data
Semi-automatic Schema Annotation

Motivation

- Due to the heterogeneous and error-prone character of Open Data it is often not possible to automatically augment it.
- Instead there are a lot of ambiguities within which need to be decided through human intervention.

→ Crowd-sourcing approaches

- To solve ambiguities that could encounter during data integration, e.g. raising questions, such as "Is the attribute c_revenue of type currency?", "Does c_id match customer?" or "Is James Cameron an instance of UK Prime Minister?"
- Can be easily answered by non-expert users.

Related Work

- User Feedback as a First Class Citizen in Information Integration Systems, 2011.
- CrowdDB: answering queries with crowdsourcing. In SIGMOD '11.
- Human-assisted graph search: it's okay to ask questions. In VLDB‘11.
"For (m)any data integration problem, if you don’t involve human, then there is no hope."

AnHai Doan
**Company Name** | **Address** | **Market Cap**
--- | --- | ---
Google | Googleplex, Mnt. View CA | $210Bn
Intl. Business Machines | Armonk, NY | $200Bn
Microsoft | Redmond, WA | $250Bn

- **SELECT Market_Cap**
  - **FROM Companies**
  - **WHERE Company_Name=“IBM“**

  **Result:** 0 rows
  **Problem:** Entity Resolution

- **SELECT Market_Cap**
  - **FROM Companies**
  - **WHERE Company_Name=“Apple“**

  **Result:** 0 rows
  **Problem:** Closed World Assumption

---

*Hard queries for computers, very easy for humans*
Microtasking – Virtualized Humans

- Current Leader: Amazon Mechanical Turk ([https://www.mturk.com/mturk/welcome](https://www.mturk.com/mturk/welcome))
- Requestors place Human Intelligence Tasks (HITs)
  - Minimum price: $0.01
  - Other parameters: # of replicas (assignments), expiration, user interface
  - API-based: createHIT(), getAssignments(), approveAssignments(), enforceExpire()
- Requestors approve jobs and payment

Workers (a.k.a. Turkers) choose jobs, do them, get paid
Architecture

CrowdSQL

Metadata
- Parser
- Optimizer
- Executor
- Files Access Methods

Statistics

Results
- Turker Relationship Manager
- UI Creation
- Form Editor
- UI Template Manager
- HIT Manager

Disk 1

Disk 2
CrowdSQL

- CREATE TABLE Department ( 
  university STRING,
  name STRING,
  url CROWD STRING,
  phone STRING,
  PRIMARY KEY (university, name) );

- CREATE CROWD TABLE Professor ( 
  name STRING PRIMARY KEY,
  email STRING UNIQUE,
  university STRING,
  department STRING,
  FOREIGN KEY (university, department) 
  REF Department(university, name) )
Example User Interfaces Created by CrowdDB

(a) Crowd Column & Crowd Tables w/o Foreign Keys

(b) CROWDEQUAL

(c) CROWDORDER

(d) Foreign Key (normalized)

(e) Foreign Key (denormalized)
CrowdSQL Query Plan Generation

```sql
CREATE CROWD TABLE department (
    name STRING PRIMARY KEY,
    phone_no STRING);

CREATE CROWD TABLE professor (
    name STRING PRIMARY KEY,
    e-mail STRING,
    dep STRING,
    REF department(name);
);

SELECT *
FROM PROFESSOR p, DEPARTMENT d
WHERE d.name = p.dep
AND p.name = "Michael J. Carey"
```

(b) Logical plan before optimization
(c) Logical plan after optimization
(d) Physical plan
Games with a purpose

Motivation

- Many tasks are trivial for humans, but very challenging for computer programs (e.g. speech recognition, natural language understanding, computer graphics)
- People spend a lot of time playing games

Idea

- Computation + Game Play
- People playing GWAPs perform basic tasks that cannot be automated.
- While being entertained, people produce useful data as a side effect.

How to design these games such that...

- People enjoy playing them!
- They produce high quality outputs!

Related Work

- http://www.gwap.com
Games with a purpose – Example

Agreement on “flower”

Alice

shoe
rocks
flower

Bob

flower
Data Exploitation
Exploitation: Drill Beyond

„Drill Beyond“: Extending OLAP using Open Data

- “Open World“ approach: Include all external open data sets → open schema

„country with GDP greater X“
select * from nation where nation.gdp > X;

„sum of prices for EU customers“
select sum(o_total_price)
from orders, nation, organization
where ... and organization=‘EU’;

Main problems

- Query segmentation + mapping to external data sources
- Keyword-search over relational data + huge set of open data
- Ambiguities → top-k results + user feedback
- Integration of “open world“ results into local database
Motivation – Keyword Search

Querying structured data

- Relational databases
  - A repository for a significant amount of data (e.g., enterprise data)
  - RDBMS that manages an abstract view of underlying data
- Structured Query Language (SQL)
  - Precise and complete
  - Difficult for casual users

Querying unstructured data

- (Web) documents
  - Collection of unstructured (natural language) documents available online
  - Search engine: The most popular application for information discovery
- Keyword search
  - Simple and user-friendly
  - Approximating the precise results, in statistical and semantic ways
Enabling casual users to query relational databases with keywords

- “casual users”
  - Without any knowledge about the schema information
  - Without any knowledge of the query language (SQL)
  ⇒ Search system should have the knowledge in behalf of users

Challenges

- Inherent discrepancy of data between IR and DB
  - Information often splits across the tables (or tuples) in relational databases
Research Dimensions

Model
- Data Representation
- Query Representation

Indexing
- Efficient Processing
  - Top-k query processing
  - Indexing structure

Processing

Ranking
- Ranking
- Presentation
Related Work

Proximity
- [Goldman et al., VLDB, 1998] Proximity Search in Databases

DataSpot
- [Palmon et al., VLDB, 1998] DTL's DataSpot - Database Exploration Using Plain Language
- [Palmon et al., SIGMOD, 1998] DTL's DataSpot - database exploration as easy as browsing the Web

DBXplorer
- [Agrawal et al., 2002, ICDE] DBXplorer: a system for keyword-based search over relational databases

BANKS
- [Hulgeri et al., 2001, DEBU] Keyword Search in Databases
- [Hulgeri et al., 2002, ICDE] Keyword Searching and Browsing in Databases using BANKS
- [Kacholia et al., 2005, VLDB] Bidirectional Expansion For Keyword Search

DISCOVER
- [Hristidis et al., 2002, VLDB] DISCOVER: Keyword search in relational databases
- [Hristidis et al., 2003, VLDB] Efficient IR-Style Keyword Search over Relational Databases.
- [Liu et al., 2006 SIGMOD] Effective Keyword Search in Relational Databases

ObjectRank
- [Balmin and Hristidis et al., 2004, VLDB] ObjectRank: Authority-Based Keyword Search in Databases
- [Balmin and Hristidis et al., 2008, TODS] Authority-based search on databases
Vision: Do-It-Yourself Analytics

“number of cafes vs. age distribution per district of Dresden”

- Process Query
- Look up fitting data sets
- Compute suitable visualization
- Compute interaction / exploration features
Vision: Do-It-Yourself Analytics (2)

"number of cafes vs. age distribution per district of Dresden"

NLP techniques + Lookup services (e.g. GeoNames)

- number of cafes vs age distribution per district of Dresden
- value dimensions
- relations/operations
- natural geographic entity

Process Query
- Look up fitting data sets
- Compute suitable visualization
- Compute interaction / exploration features
"number of cafes vs. age distribution per district of Dresden"

<table>
<thead>
<tr>
<th>Identified Dimension</th>
<th>Candidate Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of cafes</td>
<td>OpenStreetMap</td>
</tr>
<tr>
<td></td>
<td>OR</td>
</tr>
<tr>
<td></td>
<td>Recommendation service, e.g., Yelp</td>
</tr>
<tr>
<td>age distribution</td>
<td>Municipal Statistics Agency Dresden</td>
</tr>
<tr>
<td>district of Dresden</td>
<td>OpenStreetMap</td>
</tr>
</tbody>
</table>

Ambiguity → user feedback

Vision: Do-It-Yourself Analytics (3)
Vision: Do-It-Yourself Analytics (4)

- "number of cafés vs. age distribution per district of Dresden"

<table>
<thead>
<tr>
<th>Identified Dimension</th>
<th>Properties</th>
<th>Visualization Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of cafés</td>
<td>Number for each district</td>
<td>Bars per district OR color of each district</td>
</tr>
<tr>
<td>age distribution</td>
<td>Distribution for each district</td>
<td>Multiple histograms</td>
</tr>
<tr>
<td>district of Dresden</td>
<td>Polygon for each district</td>
<td>Map</td>
</tr>
</tbody>
</table>
Vision: Do-It-Yourself Analytics (5)

"number of cafes vs. age distribution per district of Dresden"

- Process Query
- Look up fitting data sets
- Compute suitable visualization
- Compute interaction / exploration features

Too much information for one visualization
→ enable exploration, e.g., clicking a district in the map opens histogram

number of cafes → age distribution
Topics for Bachelor Thesis
Extraktion semantischer Kontextinformationen aus Metadaten

Motivation:
Webdaten sind durch eine starke semantische und strukturelle Heterogenität gekennzeichnet, welche die Integration der Daten in ein einheitliches System erschweren. Zusätzlich verfügbare Informationen (Metadaten), die den Inhalt und den Kontext der Daten beschreiben, können ausgenutzt werden, um die Heterogenität zu überwinden.

Aufgabe:

Umsetzung: Python oder Java
Betreuer: Katrin Braunschweig
**Kontextbasiertes Clustering von Datensätzen**

**Motivation:**

Im Zuge der Open Data Initiative stellen viele Behörden ihre Datensätze in Datenkatalogen im Web zum Download zur Verfügung. Der Eintrag im Katalog folgt jedoch meist keiner sinnvollen Struktur, sondern lediglich der Reihenfolge, in der die Datensätze veröffentlicht wurden. Eine bessere Strukturierung könnte jedoch helfen, verwandte Datensätze leichter auffindbar zu machen.

**Aufgabe:**

Die Datensätze sollen mithilfe verschiedener Clustertechniken bezüglich unterschiedlicher Kategorien (z.B. Titel, Tags, Beschreibung, Format) geclustert werden. Die Wirkungsweisen unterschiedlicher Techniken sollen dabei gegenübergestellt und analysiert werden. (Zusatz: Wie können verschiedene Clusterings kombiniert werden um ggf. ein besseres Gesamtergebnis zu erhalten.)

**Umsetzung:** Python oder Java

**Betreuer:** Katrin Braunschweig
Schema-Extraktion aus Rohdaten

Motivation:

Die von öffentlichen Institutionen zur im Web zur Verfügung gestellten Daten liegen häufig in Rohdatenformaten ohne Schemabeschreibung, oder ohne maschinenlesbare Schemabeschreibung vor. Die automatische Weiterverarbeitung dieser Datensätze wird dadurch erschwert.

Aufgabe:


Betreuer: Julian Eberius
Techniken zur Automatischen Visualisierung für Endnutzer

Motivation:

Die meisten Open Data Plattformen stellen Rohdaten zur Verfügung, bieten jedoch keine oder nur eingeschränkte Möglichkeiten, die Daten zu visualisieren und sich so einen Überblick zu verschaffen. Selbst wenn solche Möglichkeiten vorhanden sind, sind diese oft generisch und passen die erzeugten Visualisierungen nicht den gewählten Datensätzen an.

Aufgabe:


Betreuer: Julian Eberius