

#29



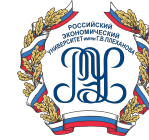
Complex Objects Identification and Analysis Mechanisms

Mikhail Ulizko ¹, Larisa Pronicheva ², Alexey Artamonov ², Rufina Tukumbetova ¹, Evheniy Tretyakov ³

¹ Plekhanov Russian University of Economics, Stremyanny Pereulok, 36, 115093 Moscow, Russia

² National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashira Hwy, 31, 115409 Moscow, Russia

³ Moscow Center of Fundamental and Applied Mathematics, Leninskie Gory, 1, Moscow, 119991, Russian Federation



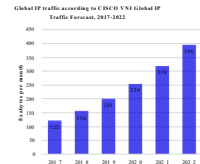
ABSTRACT

Currently the volume of information on the Internet global network is inclined to increase. It affects various areas of activity. The paper presents the issue of identifying and analyzing complex objects in the context of information sources devoted to project activity of the US National Institutes of Health (NIH). As part of the solution to the problem, information sources relevant to the provided topic were found, and the data were downloaded with the use of agent techniques. The methods of primary analysis and data processing were developed to create a data storage structures in SQL and NoSQL models. The analysis of the presented database models was conducted, their advantages and disadvantages were revealed. As a result software tools have been developed that provide data representation of a complex object and organization of work with it by web interface.

INTRODUCTION

Properties of the information on the Internet:

- heterogeneity,
- distribution,
- growing amounts.

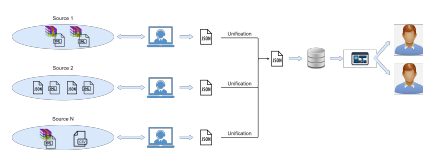


Complex object of the Internet is an entity, which possess a large number of properties, which are also heterogeneous.

Target sources



ALGORITHM



1. Getting data using agent technologies
2. Unification data into JSON format
3. Uploading data into data storage (non-relational database)
4. Data visualization and analysis

USED SOFTWARE:

- Python to scrape data
- Python to upload data to database
- Elasticsearch to store data
- Kibana & Elasticsearch to analyze and visualize data

GETTING DATA ON THE INTERNET

Web scraping is one of the ways to extract data on the Internet. It is often used with HTTP and the GET method. The agent goes on site pages with the use of unique URLs and collect the content of these HTML pages.



Agents can perform tasks periodically on schedule, by this solving the problem of continuous updating of information on the Internet.

A significant number of articles are devoted to solving the tasks software implementation of agents. Currently, one of the languages predisposed for the rapid writing of agents and solving data analysis problems is Python. To interact with Internet resources, a driver is also required, that provides remote behavior by a web browser. Selenium WebDriver is the driver. Using it with the help of Python makes it possible to control browser actions and collect data from web pages on schedule.

UNIFICATION DATA



JSON file example

```
{
  "Project": {
    "ProjectID": "...",
    "ProjectName": "...",
    "CoreProjectNumber": "...",
    "ProjectTitle": "...",
    "ContactPI": "...",
    "AdditionalPINames": null,
    "PIOrgName": null,
    ...
  }
}
```

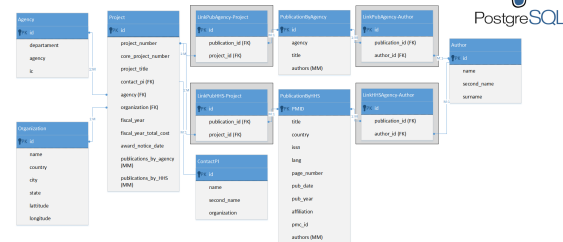
The JSON data format advantages:

- it is a light weighted format;
- it is easy for humans to read and write and for machines to understand and generate;
- it is often used as a format for data view in NoSQL database.

SQL vs NoSQL DATABASES

To design a database schema, the following steps are due to be taken:

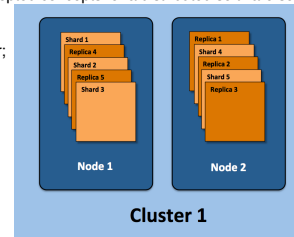
- identify objects, information of which is displayed in the database;
- determine the relationship between the objects;
- determine the objects attributes;
- normalize relationships.



One example of NoSQL database is ElasticSearch, an open source distributed software search engine that has a JSON REST API for interacting with downloaded data.

Main accepted concepts for a distributed software search system:

- Index;
- node;
- cluster;
- shard.

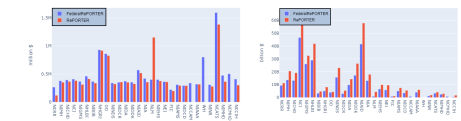


Settings for index

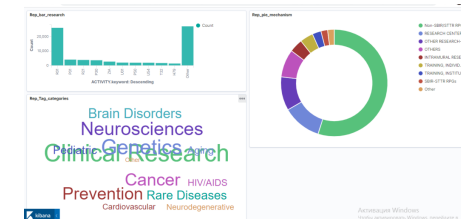
```
{
  "mappings": {
    "project": {
      "type": "text",
      "ProjectName": {
        "type": "text",
        "CoreProjectNumber": {
          "type": "text",
          "ProjectTitle": {
            "type": "text",
            "ContactPI": {
              "type": "text",
              "Agency": {
                "type": "text",
                "Department": {
                  "type": "text",
                  "PI": {
                    "type": "text",
                    ...
                  }
                }
              }
            }
          }
        }
      }
    }
  }
}
```

DATA ANALYSIS

The information obtained is interesting because it can be analyzed. The relational database is accessed using SQL queries. Since the initial data for queries is a collection of tables, the return value will also be a table (or a number). An important role in the analysis is the clarity in data representation, so developed systems with databases often use visualization tools such as Plotly, ChartJS, etc.



To visualize data in Elasticsearch, there is Kibana, an open source data visualization and exploration platform, which is a powerful analysis tool. An example of analysis is given below.



The analysis of project activities was carried out in four ways: conducting initial analysis using scripts in Python, building interactive Plotly graphs, building a standard information retrieval system (backend SQL query), building a search database using Elasticsearch and Kibana.

CONCLUSIONS

Within the framework of this problem, on the basis of the proposed methodology, the task of collecting and analyzing data on the National Institutes of Health has been solved. It is concluded that scientific research is mostly devoted to genetics, cancer, clinical research and is implemented in the form of grants.

This approach can also be used to analyze complex objects such as organizations involved in project activities.

The study was carried out at the expense of the Russian Science Foundation grant (project No. 19-71-30008)