EDUCATIONAL PROGRAM FOR “MATHEMATICAL INFORMATICS” SPECIALIZATION IN THE BELARUSIAN STATE UNIVERSITY

Perez Tchernov A., V. Romanchik

Belarusian State University, Minsk, Belarus

Intellectual data mining goes widespread both in research and commercial areas. A lot of internet-based companies had published their internal data analysis tools as open source software. Thus the barrier to entry to practical intellectual data mining area is rather low right now. There are a lot of outstanding examples of data mining usage in big internet companies. For example, Netflix and Twitter provide recommendation services. IBM is well known by its usage of innovative data mining technologies for question answering problems.

Let specify most popular tools and approaches for data analysis. Namely, different machine learning algorithms are used to generate recommendations. Specific clustering and search algorithms are popular to find out hidden patterns in data. Specialized linguistic analysis and graphs algorithms are used for social and communication data. Domain specific languages (DSL) allows data scientist to operate at a high level with base algorithms and data extraction queries.

There are a lot of specialized software tools nowadays, that implement aforementioned algorithms and approaches. We can use distributed batch and online data processing platforms (e.g. Hadoop, Storm), data classification and query tools (e.g. Mahout, Weka, Scalding), timeframe based event filtering (e.g. Esper), and specialized storage systems (e.g. Cassandra, HDFS).

It is necessary to use additional technologies for data retrieval from third-party services and sources. The list includes search crawlers, message queues, inter-component communication mechanisms and data serialization.

It is important challenge to follow the current trend of hybrid calculation (e.g. IBM Deep QA / Watson approach), when data mining technologies are combined with best semantic and linguistic approaches.

We formalized and propose the following vision on current data analysis trends in a new “Mathematical informatics” specialization program in the mechanic and mathematic faculty of the Belorussian state university. This program may be used as a basis for upcoming education courses and trainings in data mining area.

1. Data modelling:
   1.1. Data structures (34h);
   1.2. Relational databases and SQL (34h);
   1.3. NoSQL databases (with Cassandra, hbase), OLAP / MDX (with Pentaho) (34h);
   1.4. Capacity planning, data warehouse technologies, data archiving (34h);
   1.5. Semantic and SPARQL (with Virtuoso) (48h);
2. Information technologies
   2.1. Python, Matlab and R languages for data scientists (98h)
   2.2. Distributed algorithm and tools (with Mahout, Scalding, Mesos) (34h)
   2.3. Parallel algorithms (34h);
   2.4. Parallel programming languages (Erlang, MPI) (48h)
   2.5. Java online data processing (with Storm, Akka, Finagle) (48h)
   2.6. High scalable projects and their architectures (on Twitter example) (48h)
   2.7. Scala language (68h)
   2.8. Machine learning algorithms (34h);
3. Applied data analysis
   3.1. System and business – analyses (34h)
   3.2. Data mining (34h);

    3.3. Crawling, data retrieval and search technologies (34h);
    3.4. Computational linguistics (34h);
    3.5. Advanced graph algorithms (with Apache Giraph) (34h);
    3.6. Financial data analysis(34h)
    3.7. Bio informatics (34h);
4. Cloud and specific web /mobile technologies
    4.1. Cloud services, AWS (48h);
    4.2. Web – programming (34h);
    4.3. Mobile application and services (34h);
    4.4. Linux administration (34h)
    4.5. SEO (34h);
    4.6. Data visualization technologies (34h);
    4.7. Recommendation algorithm (on Netflix example) (34h);