Извлечение и интеграция информации из больших данных (Веб, социальные сети, тексты, …)

Information Extraction and Integration from Big Data (Web, Social Media, Texts, …)

проф. д.ф.-м.н. Л.А. Калиниченко (leonidk@synth.ipi.ac.ru)
к.т.н. А.Е. Вовченко (itsnein@gmail.com)

Аннотация курса

В настоящем курсе изучаются методы и инструменты извлечения (mining) и интеграции информации из различных источников больших данных (в масштабе Веба, социальных сред (Twitter, Linkedin, …), блогов, публикаций в средствах массовой информации, машинных логов, сенсорных данных, и пр. Большие данные обычно являются неструктурированными (чаще всего текстовыми), слабоструктурированными (например, в виде XML, JSON, баз данных NoSQL). Вместе с тем, образуются также и структурированные большие данные как, например, результат наблюдений (измерений) современными инструментами, накопления многочисленных таблиц в Вебе. Современные ИТ платформы включают распределенные инфраструктуры типа Hadoop, обеспечивающие параллельную обработку и анализ таких разноструктурированных больших данных на основе парадигмы Map/Reduce. Практическая часть предлагаемого курса ориентирована на подобную платформу на базе IBM BigInsights.

В курсе рассматриваются:

- программные способы сопряжения различных источников больших данных с кластерными платформами, что позволяет слушателям погрузиться в реальную среду больших данных;
- методы извлечения данных о сущностях (entities) реального мира (таких как личности, компании, продукты, разнообразные объекты исследования, и пр.) из текстов и способы программирования соответствующих экстракторов на алгебраическом языке AQL;
- методы извлечения, сопоставления и группирования (matching) и разбора (resolution) путем связывания (linking), устранения дублирования (deduplication) различных разноструктурированных представлений информации об одной и той же сущности реального мира (entity resolution);
- методы и операции слияния (интеграции) данных об одних и тех же сущностях реального мира и их связей, представленных в разных коллекциях, образованных в процессе разрешения сущностей (в частности, рассматриваются стратегии и операции устранения конфликтующих данных, операции поглощения и слияния данных);
- обзор методов и средств курирования данных, обеспечения качества данных.

Изучаемые методы и операции извлечения и интеграции информации о сущностях реального мира позволяют программировать интеграционные потоки вида ETL, образующие интегрированные структурированные данные, которые могут быть использованы в приложениях для дальнейшего анализа и обработки. Программирование изучаемых методов и операций извлечения и интеграции информации о сущностях реального мира осуществляется в курсе на декларативном языке HIL (Highlevel Integration Language, новом языке, разработанном IBM, ориентированным на разбор и интеграцию сущностей в среде Map/Reduce), используемом совместно с AQL, и отрабатывается на реальных данных.

Курс извлечения и интеграции данных связан с другими курсами Программы изучения платформ и аналитики больших данных на ВМК МГУ (http://synthesis.ipi.ac.ru/synthesis/student/BigData), включая ***«Управление разно-структурированными большими данными»*** и ***«Методы и платформы интеграции данных и организации хранилищ больших данных»***.