IS5102
Database Management Systems

Lecture 17: Data Analytics

Alexander Konovalov

alexander.konovalov@st-andrews.ac.uk

(with thanks to Susmit Sarkar)

2021

NoSQL

Umbrella term for graph, key-value, document (and other) non-relational DBMS

Emphasise different query models

Analytics driving many of these models

Data Mining

What's in a name?

- Data Mining
- Data Science
- Data Analytics
- Machine Learning

Data Mining

What's in a name?

- Data Mining
- Data Science
- Data Analytics
- Machine Learning

Fuzzy distinctions . . .
Practically, aim at the same thing

Data Mining

Data Mining

The process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions. (Simoudis,1996).

Involves the analysis of data and the use of software techniques for finding hidden and unexpected patterns and relationships in sets of data.

Types of Data

Qualitative data

- Attributes, properties, category, . . .
- For instance Weather Descriptors (foggy, misty, cold, . . .)
- Sometimes numbers as labels (e.g. Likert Scale)

Quantitative data

- Can be Continuous (e.g. temperature)
- Can be Discrete (e.g. number of things, persons)

Link Analysis

Aims to establish links (associations) between records, or sets of records, in a database.

There are three specializations

- Associations discovery
- Sequential pattern discovery
- Similar time sequence discovery

Applications include product affinity analysis, direct marketing, and stock price movement.

Link Analysis

Associations Discovery: Finds items that imply the presence of other items in the same event.

e.g. ‘When a customer rents property for more than 2 years and is more than 25 years old, in 40% of cases, the customer will buy a property.’

Sequential Pattern Discovery: Finds patterns between events such that the presence of one set of items is followed by another set of items in a database of events over a period of time.

e.g. Used to understand long term customer buying behavior.

Similar Time Sequence Discovery: Finds links between two sets of data that are time-dependent, and is based on the degree of similarity between the patterns that both time series demonstrate.

e.g. ‘Within three months of buying property, new home owners will purchase goods such as cookers, freezers, and washing machines’.

Consistency

Databases have traditionally been very concerned with **consistency**

An issue in many concurrent accesses

Seemingly at odds with replicated and sharded data for analysis

. . . no longer so clear (if writes are infrequent)

Fast, Huge Data, Inconsistent?

Strengths lie in processing Big Data

Analytics, Web Applications

Designed for fast, flexible, complex queries

Designed for high availability

Often (not always) give up Strong Consistency

Interaction of Consistency, Availability, Partition Tolerance (Reliability) **CAP**

NewSQL

RDBMS are not going away

Neither are NoSQL databases

Lessons that transfer:

- high-scalability
- consistency

New branding: NewSQL (e.g. Google Spanner)

Reading and Consolidation

Chapter 11, *Database System Concepts*, Silbershatz, Korth and Sudarshan

Chapter 33, *Database Systems*, Connolly and Begg