Week 3

April 2016

wk	M	T	W	T	F	S	S
14					1	2	3
15	4	5	6	7	8	9	10
16	11	12	13	14	15	16	17
17	18	19	20	21	22	23	24
18	25	26	27	28	29	30	

Big data Integration and processing

2016 Monday February 15

046-320 • WK 08

Video 1.

Overview of Information Integration

It refers to the problem of using many different info. sources to accomplish a task.

10.00 → Explain the data integration problem

→ Define integrated views and schema mapping

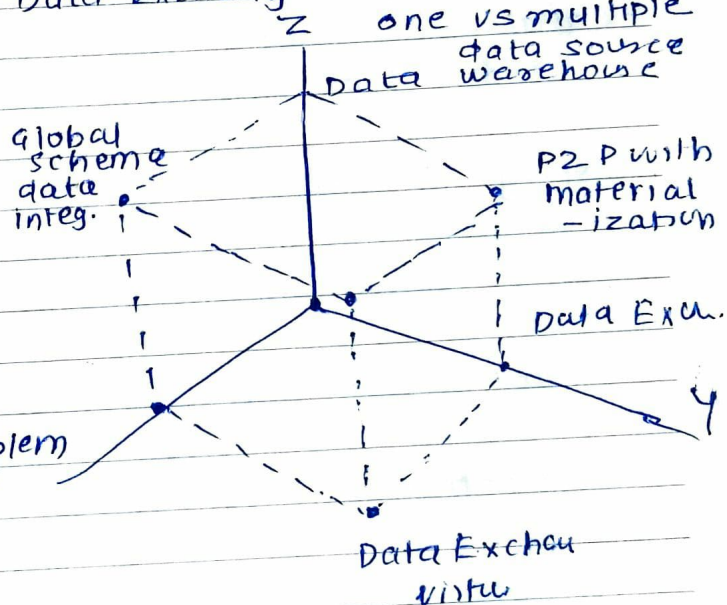
11.00 → Describe the impact of increasing the no. of data source

→ Appreciate the need to use data compression.

12.00 → Describe Record Linking, Data Exchange and Data fusion task.

1.00

Record linking



2.00

3.00 The "Big data" problem

→ Many sources-

4.00 ① Hundreds of tables

② schema mapping problem is a combinational challenge.

5.00

6.00 → Pay-as-you-go model

① Only integrate sources that are needed when needed

7.00 → probabilistic schema mapping

Design mediated schema

Attribute Grouping

NOTES

Data Integration Scenarios

→ 4 data sources each with one relation.

Data Exchange :- Given a source database with a finite no. of relations, a set of schema mappings, and a set of constraints that the target schema must satisfy, the data exchange problem is to find a finite target database such that both the schema mappings and the target constraints are satisfied.

Using Codebooks :- Logical observations Identifies Name and codes is a database.

Using Compressed data.

Compression :- Encoded representation of data so that it uses less space.

→ Dictionary encoding

Ontological Data :-

Ontology queries are graph queries.

Ontology :- A set of terms of a domain
- Relationship b/n the terms.

Snomed :- A medical ontology used for clinical data.

NOTES

Data Fusion :- • Data sources
• Data Items
→ A product
→ A part of product
→ A feature of a product
... } value

• Using data from a subset of sources find the true value or a true value distribution of a data item.
• Assemble all such values for the real-world entity represented by the data items.

Too many sources :-
→ Too many sources = too many values.
→ Voting to select the "right" value.
• Simple voting can be problematic –
Veracity Problem
• source Reliability
• Copy Detection
• Statistical techiques to estimate
• Trustworthiness of sources.
• Bias introduced by copies.
• True distribution of values for data items

Source Selection :-

The problem :- 1) choose only useful sources.

2) Adding sources first improves integration accuracy then reduces it.

The soln :- Order candidate source based on a measure of "goodness"

→ Add sources until the marginal benefit is less than the marginal cost

→ Current techniques scale well.

Notes

Quiz 3 :-
Information Integration

1. What is the main problem with big data information integration ?
→ Many sources.

2. Two possible sol^{n}s associated with "big data" info. integration as mentioned in lect ?
→ • Probabilistic Schema mapping
• Pay-as-you-go Model.

3. Mediated schemas ?
→ Schemas created from integrating 2 or more schems

4. In attribute grouping, how would one evaluate if two attributes should go together ?
→ • probability of Two attributes co-occuring
• similarity of Attribute.

5. What is data item ?
→ Data that represents an aspect of a real-world entity.

6. What is data fusion ?
→ Extracting the true value of a data item

7. What is a potential problem of having too many data sources as mentioned in text?

→ Too may data value.

8. What do we mean by the true value of a data item?

→ Extrapolated data from a data item that represent the worth of that item.

9. potential method to deal with too may data source

→ compare and weigh each source by their trustworthiness.

otes