SAMSTAR

A ***S***emi-***A***utomated lexical ***M***ethod for generating ***STAR*** schemas using an ER diagram

Il-Yeol Song, *Ritu Khare,* and Bing Dai
The iSchool at Drexel
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104
United States of America

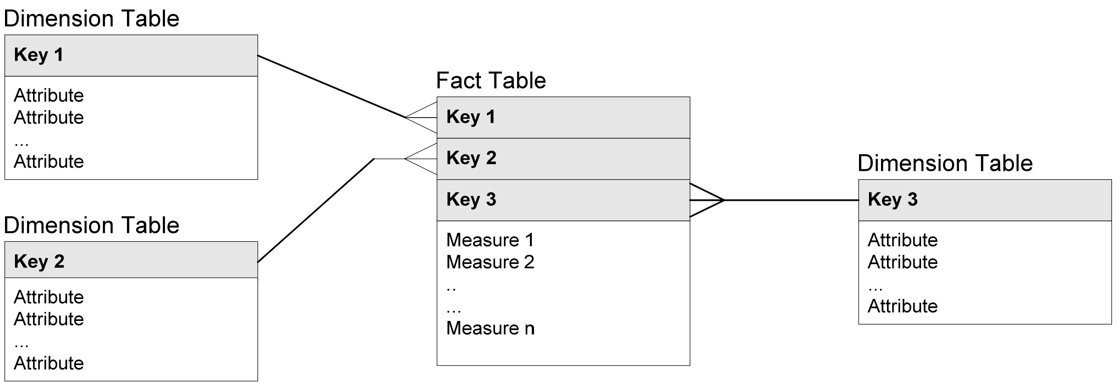
The ACM 10th International Workshop on Data Warehousing and OLAP (DOLAP 2007), Lisbon, Portugal, Nov. 9, 2007

Order of the Presentation

1. Problem Statement
2. Existing Methods: Types
3. SAMSTAR: Key features
4. Key Concepts and Ideas used
5. SAMSTAR Algorithm
6. Case Studies
7. Appraisal
8. Future Research Possibilities

1. Problem Statement

- Research Goal:
  - Semi-Automatic method to generate Star Schemas



Structure of Star Schema

  - Save a great deal of time for expert designers
  - Give a smooth head-start to novices

2. Existing Methods: Types

Diverse Range of approaches to design star schema

- **User or Demand driven** : Highest priority to the needs of users.
- **Supply/Data/Source driven**: Source data(such as ER diagram) is used as an input to build data warehouse.
- **Goal driven**: Entire focus on business goals
- **Hybrid driven**: two or more of these factors are suitably blended to give rise to hybrid methods.

Giorgini, Rizzi, and Garzetti (2005)

Phipps and Davis (2002)

Prat, Akoka, and Comyn-Watttiau (2006)

Guo et al. (2006)

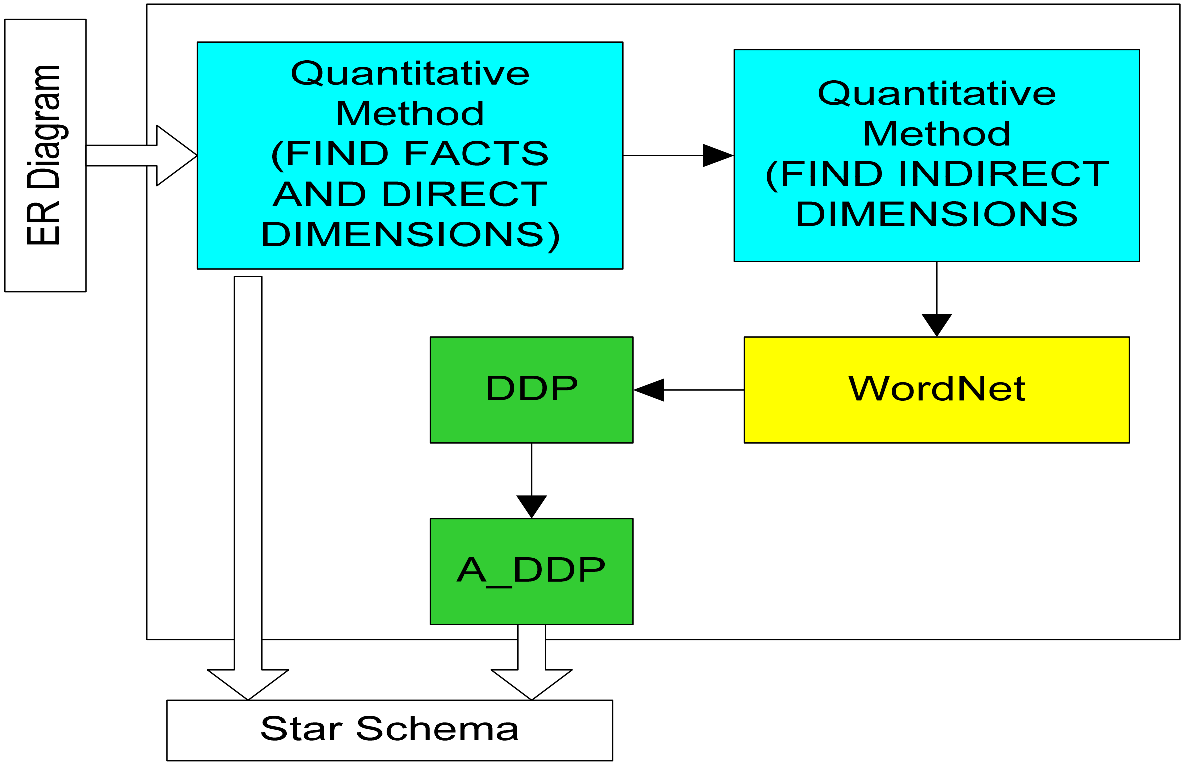
Chen and Hsu (2005)

Moody and Kortink (2000)

3. SAMSTAR: Key features

- Primarily Source and secondarily User and Goal driven.
- Focus on semantics and structure of ER diagram
- Automatically identifies a set of fact candidates from a complex and large Entity Relationship Diagram.
- Universal Approach in determining dimensions
  - Dimension Design Patterns (Jones and Song, 2006)
  - WordNet

3.1 SAMSTAR: Key features Architecture

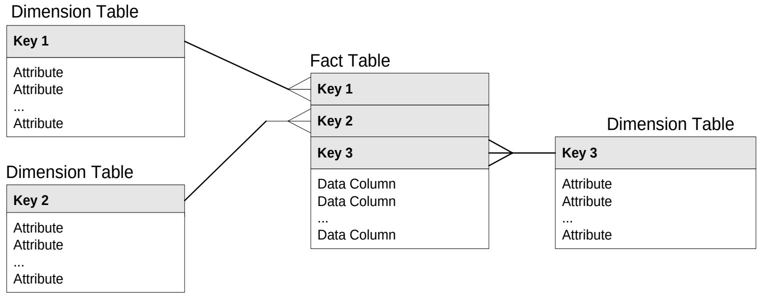

ER Diagram
Quantitative Method (FIND FACTS AND DIRECT DIMENSIONS)
Quantitative Method (FIND INDIRECT DIMENSIONS
WordNet
DDP
A_DDP
Star Schema

4. Key Concepts and Ideas Used

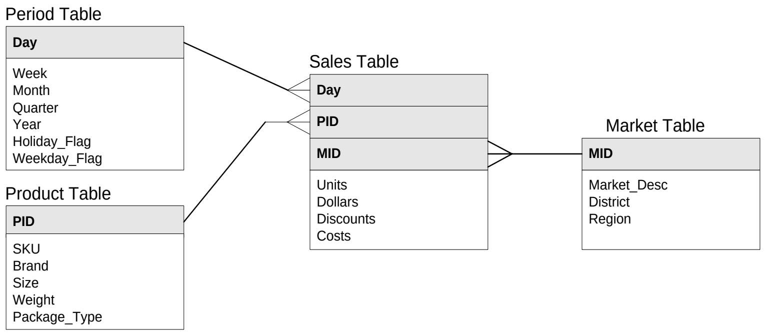
1. Facts and Dimensions: *Many-to-one* relationships
2. Direct and Indirect *Many-to-one*
3. Connection Topology Value (CTV)
4. High CTV and Threshold
5. Candidates of Dimensions
6. Annotated Dimension Design Patterns(A_DDP)

4.1 Key Concepts: Facts & Dimensions

- Ideas have been formulated on the basis of structure of star schema
- Facts are entities belonging to *Many side* of a *Many to One* relationship
- Dimensions are entities belonging to *One side* of a *many-to-one* relationships.



Structure of Star Schema

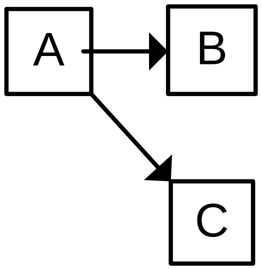


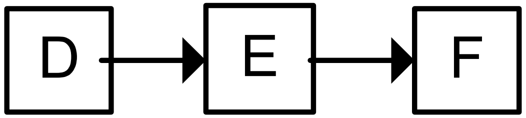
Example of Star Schema

4.2 Key Concepts: Direct and Indirect Relationships

Two types of M:1 relationships

- Direct & Indirect relationships





- Choosing fact entities,
  - Give a higher weight to entities having direct many to one relationships over entities having indirect many to one relationship

4.3 Key Concepts: Connection Topology Value

- It measures the degree to which an entity of ER Diagram is fit to be a fact entity.
- The total connection topology value (CTV) of an entity is a function of the topology value of direct relationships and the topology value of indirect relationships.

CTV(e) = weight_d*Count(Node(e)) + weight_i* ∑CTV(Node(e))

CTV(e) = the connection topology value of an entity e

Node(e) = an entity having direct *many-to-one* relationship with e, and lying on the *one* side.

Count(Node(e)) = total number of nodes of e.

Weight_d = Percentage of weight given to direct *many-to-one* relationship

Weight_i = Percentage of weight given to indirect *many-to-one* relationship

weight_d > weight_i

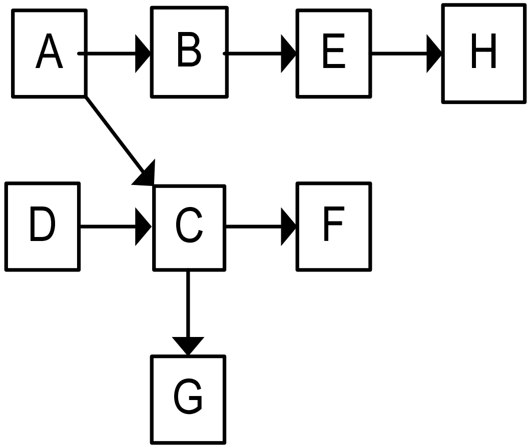
4.3.1 Key Concepts: CTV Example

weight_d=100%
weight_i=80%
The CTV for each entity is:

CTV (H) = 1* 0 + 0.8 * 0 = 0
CTV(F) = 0
CTV(G) = 0
CTV(E) = 1*1 + 0.8 * CTV(H) = 1
CTV(B)= 1*1 + 0.8 * (CTV(E))= 1.8
CTV(C)=1*2 + 0.8 * (CTV(G) + CTV(F)) = 2

CTV(D)=1*1 + 0.8 * (CTV(C)) = 1 + 0.8 * 2 = 2.6

CTV(A)= 1*2 + 0.8 * (CTV(B) + CTV(C)) = 2 + 0.8 * (1.8+2) = 5.04



4.4 Key Concepts: High CTV & Threshold

- High CTV value is identified by all values higher than threshold.

For an entity e,

CTV(e) > Th

Where **Th** is **threshold**

Threshold is calculated by following equation, adopted from research in power engineering (Christie, 2003)

Threshold = Mean + K* StandardDeviation

$$Th = X + K^* \sigma$$

$$\left[Th = \left(\sum_{i=1}^{N} CTV(i)\right) / N + K^* \sqrt{\left(\sum_{i=1}^{N} (CTV(i) - X)^2\right) / N} \right]$$

X = Mean

σ = Standard Deviation

N= total number of entities

k= variable parameter (The value of k is adjustable and can be varied accordingly to desired degree of rarity.)

All entities are numbered 1,2,3,…N

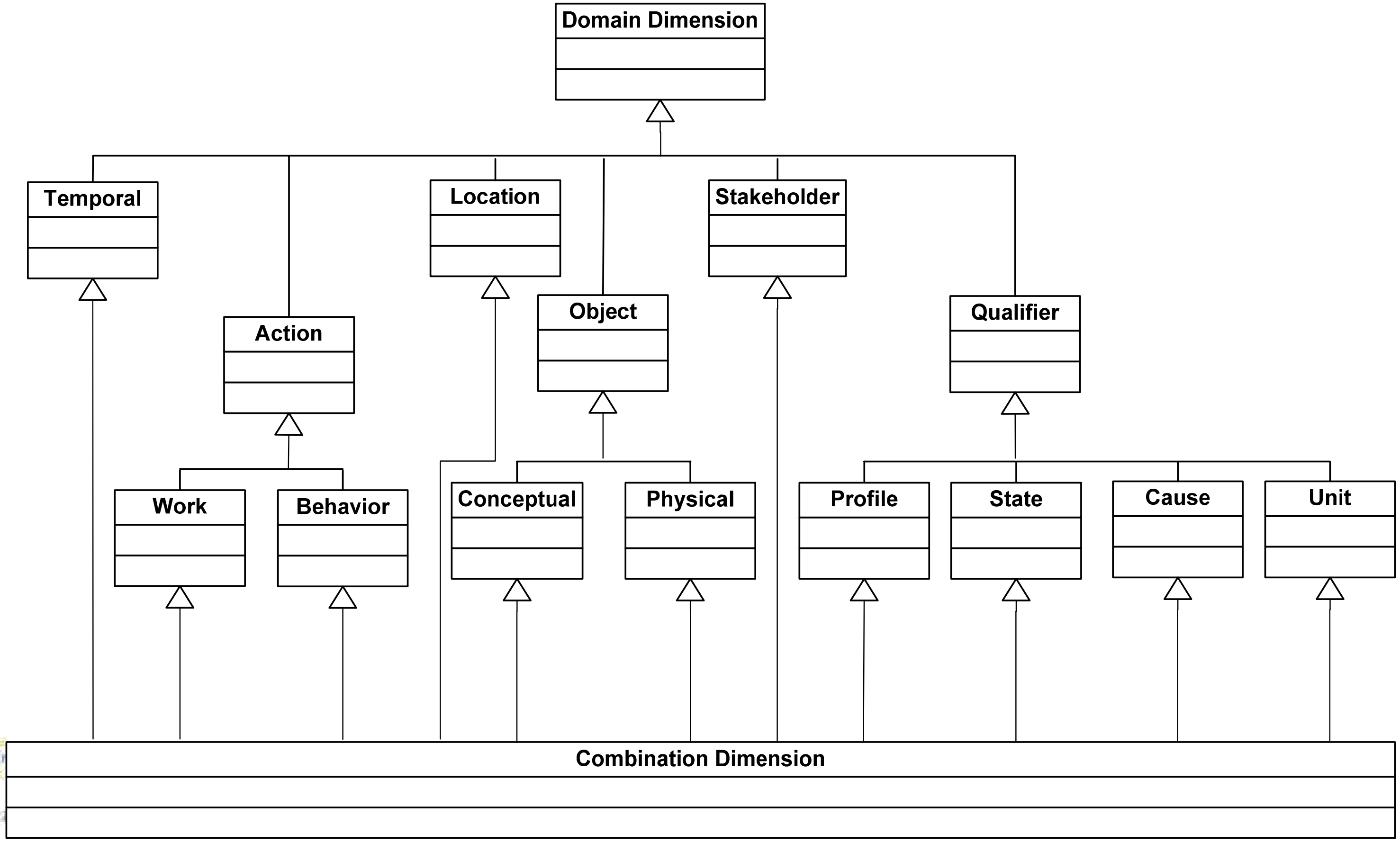
4.5 Candidates of Dimensions

An entity B is a dimension entity of a fact entity A, if it falls in one of the following two categories:

- B has a direct 1: M relationship with the A.
- B has an indirect 1: M relationship with A and either B or its one of its synonyms(WordNet) matches with one of the entities listed by Dimension Design Patterns (Jones and Song, 2006) or Annotated DDP (slide 14-15)

4.6 Annotated DDP

- Following the framework of Dimensional Design Patterns (DDP) (Jones and Song, 2006),



4.6 Annotated DDP

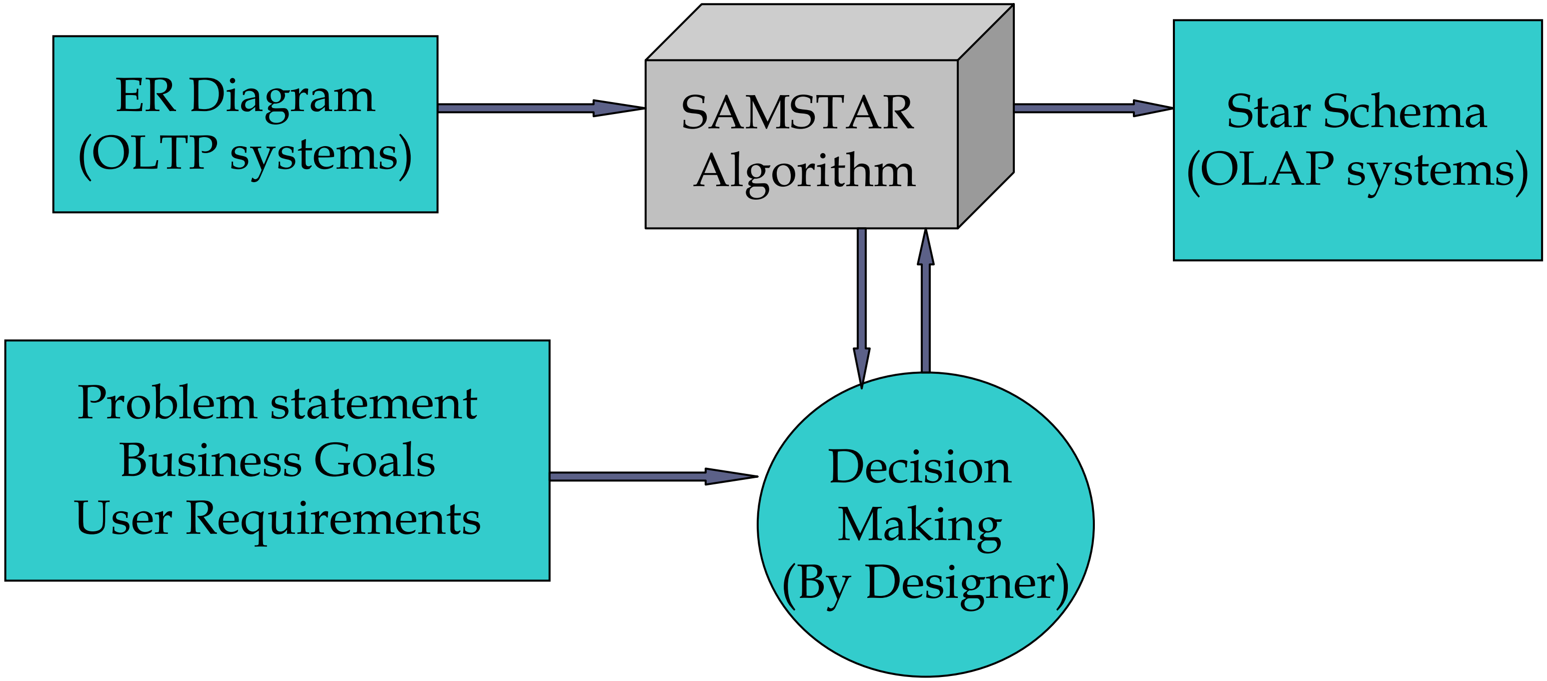
Created a list of dimensions of each of the six classes of DDP by referring to four sources

- Adamson and Venerable, 1998;
- Kimball and Merz, 2000;
- Kimball, 2002;
- Imhoff, Galemmo, and Geiger, 2003

The six classes of DDP have been instantiated to produce a list of 131 commonly used dimension entities. We refer to these entities as *Annotated DDP* (A_DDPs).

Examples of A_DDPs are: *account, account owner, activity, admissions decision, age group, agent, aircraft, airport, etc.*

5. The SAMSTAR Algorithm



5.1 SAMSTAR Assumptions

1. A structurally valid ERD (Dullea, Song, and Lamprou, 2003) is available .
2. A problem statement is available that clearly states the following:

- Business Goals: Business Process for which data warehouse has to be designed.
- User Requirements: Primary users of future system and the main measures they would be interested in.

5.2 SAMSTAR: Steps 1 to 7

Note: Steps in italics are manual, requiring designer's input.

1. Pre-process the input ERD to convert it into a Binary ERD.

2. Store Entities and Relationships

3. Let the designer choose weighting factors for direct and indirect relationships.

4. Calculate the connection topology value (CTV) for all entities

5. Calculate the threshold value, Th, for CTV.

6. Identify the entities having CTV higher than the threshold Th. These are the candidates for fact tables.

7. Decide and shortlist the fact entities based on the results from Step No. 6 and the problem statement to model. There could be more than one fact table for a business process.

5.2 SAMSTAR: Steps 8 to 11

Note: Steps in italics are manual, requiring designer's input.

8. For each fact entity, perform the following steps:

(i) Identify the entities having direct M:1 link with a fact entity

(ii) Identify entities having indirect M:1 link with the fact entity. Out of these entities, identify synonyms of entity names from WordNet. Extract the terms which match the Dimensional Design Pattern Entity list or the Annotated Dimensional Design Pattern List.

(iii) Combine the results to Steps 8(i) and 8(ii) to prepare a list of candidate dimensions for a given fact. Also, add a time dimension to a list of candidates of dimensions.

9. Decide the dimension entities based on problem statement and the result of Step 8

10. Let the designer post-process the Star Schemas:

(i) Check if 'time' is a redundant dimension.

(ii) Merge two or more related dimensions.

(iii) Rename the fact and dimension tables.

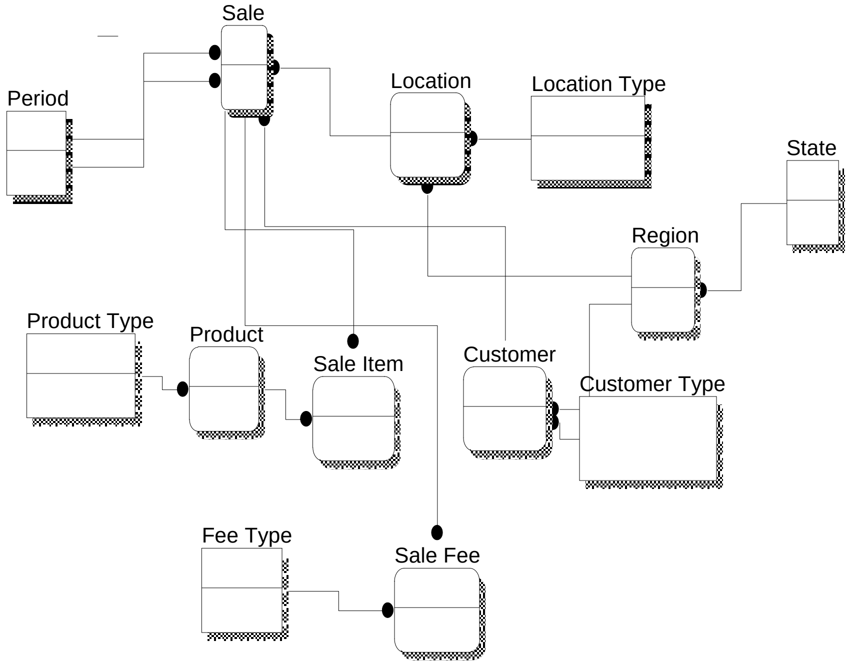
11. A star schema is generated.

6. Case Studies

- Three examples presented are extracted from the existing related literature.
- SAMSTAR algorithm was implemented in JAVA programming language.
- The values of direct and indirect weighting factors (weight_d and weight_i) were set as 1 and 0.8, respectively.
- The value of parameter k used in the calculation for threshold CTV was set to 1.5.
- Notation for ER diagrams is IDEF1X.

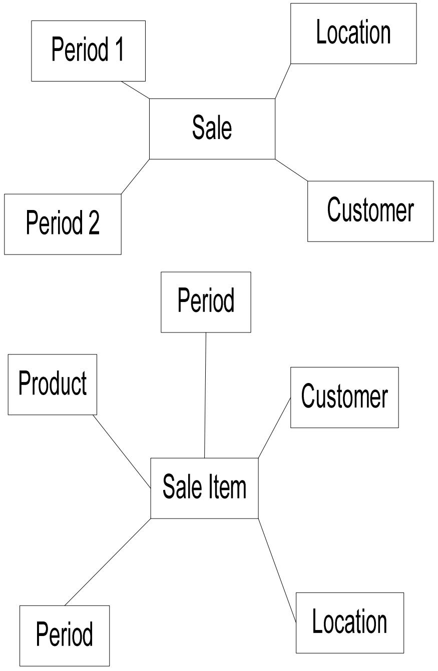
6.1 Case Study (Moody and Kortnik, 2000)

- SAMSTAR was tested on the following example Data Model given in this paper. Please note that ‘Sale’ entity has a *direct* M:1 relationship with ‘Location’ and an *indirect* M:1 relationship with ‘Region’.

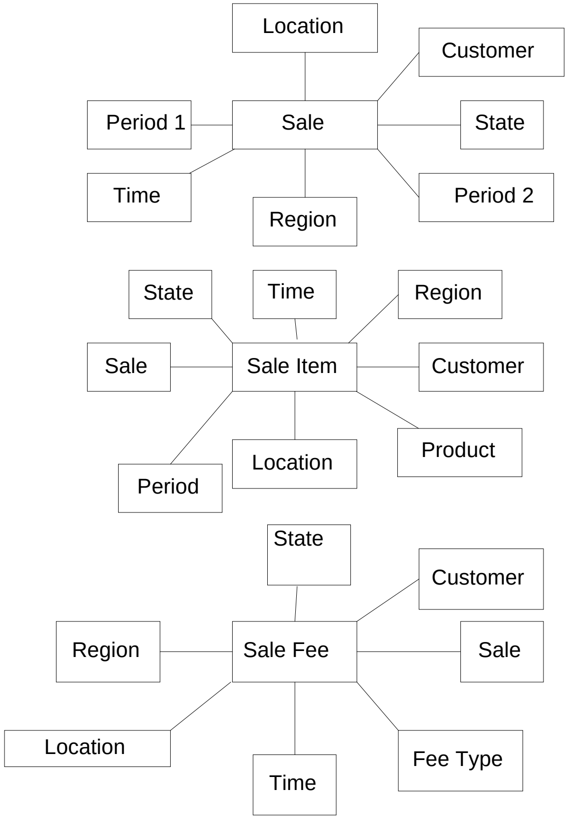


6.1 Output: More Facts, More Dimensions
More choices for the Designer
Discovery of Hidden facts and Dimensions

- The set of star schemas generated by the method stated in this paper is

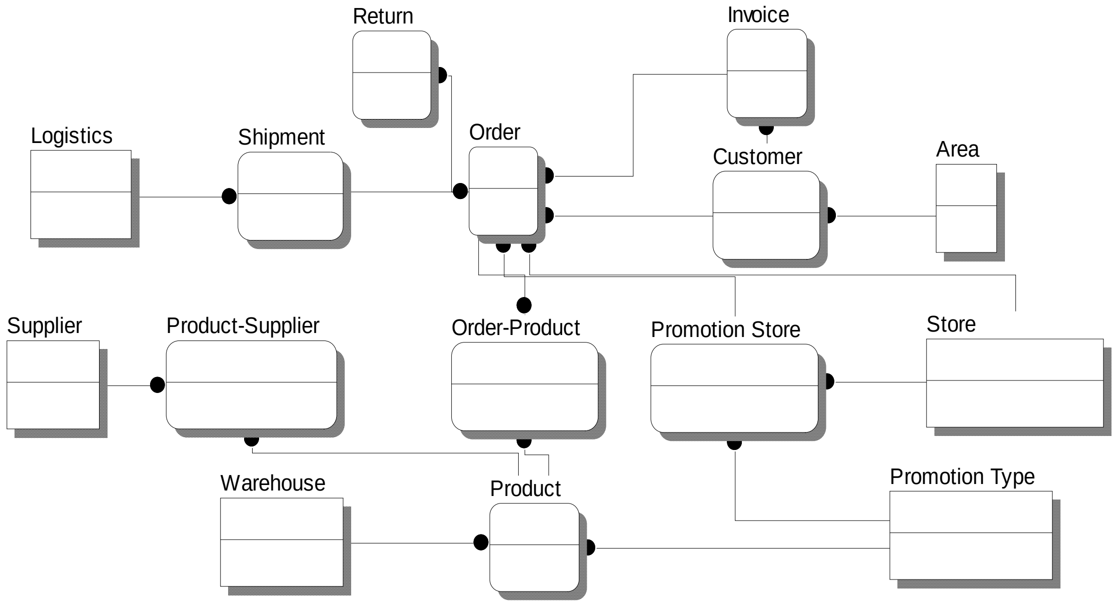


- The set of star schemas generated by the SAMSTAR method is



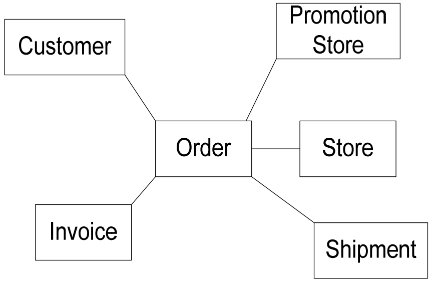
6.2 Case Study: (Chen and Hsu, 2005)

- SAMSTAR was tested on the following example Data Model given in this paper.

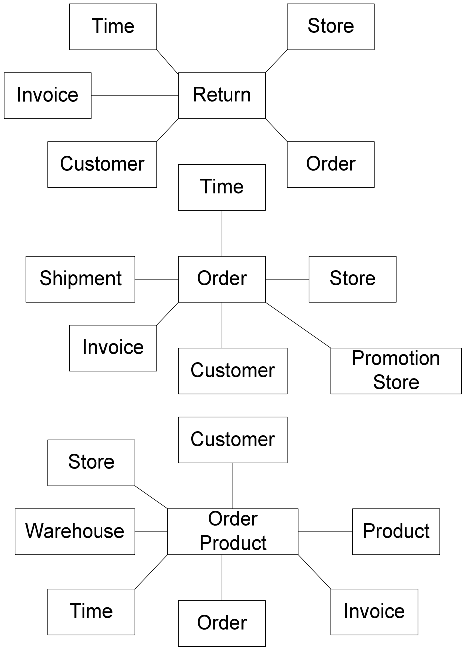


6.2 Output: More Facts, More Dimensions
More choices for the Designer
Discovery of Hidden facts and Dimensions

- The star schema generated by the method stated in this paper is



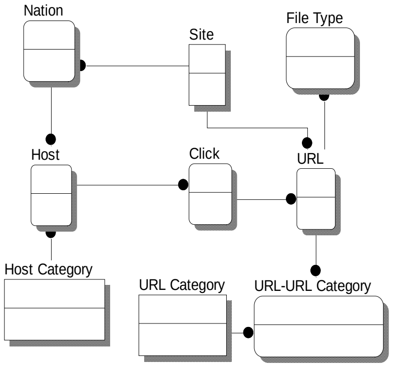
- The set of star schemas generated by the SAMSTAR method is

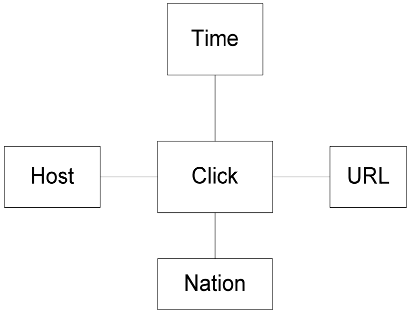


6.3 Case Study: (Golfarelli, Rizzi, and Vrdoljak, 2001)

- The example data model in the paper is

- SAMSTAR generates this schema for given ER





6.4 Case Studies Discussion

On the basis of the results obtained in afore-mentioned case studies, we infer that

- Schemas generated by SAMSTAR are similar to those generated by the manual steps in case study papers.
- SAMSTAR generated star schemas are the superset of the ones generated manually in the paper using the same ER diagrams, user needs and business goals.
- This shows our schemas are inclusive of all possible facts and dimensions.
- Our schemas have more facts and dimensions; this gives the designer a helpful aid and he/she could prune the schema as per the business and user requirements.

7. Appraisal

+It is a *universal method* to generate a star schema(s) in that we have used generalized DDPs and WordNet to identify dimensions of a fact table.

+It is *quantitative* in nature in that we analyze the structure of the ER diagram.

+It can be used to *automatically* identify a set of fact candidates from a large and complex ERD.

+It *simplifies* the work of experienced designers and gives a smooth head-start t novices.

-The manual portion of SAMSTAR might raise the probability of it being affected by human errors.

-Because it is primarily source driven; it might overlook some important user needs or important dimension entities if the problem statement is not comprehensive enough.

8. Future Research Topics

- Identify the appropriate relationship between the value of 'k' and the number of fact tables and the size of the ER diagram.
- SAMSTAR needs to be refined to better align with business goals and user needs.
- Complete rules and algorithm for attributes, dimension hierarchies, generalization, etc.
- Extend SAMSTAR for multiple ER diagrams to address data integration
- Posting SAMSTAR as a web service.

Thank You!!!

A ***S***emi-***A***utomated lexical ***M***ethod for generating ***STAR*** schemas using an ER diagram

Questions or comments?