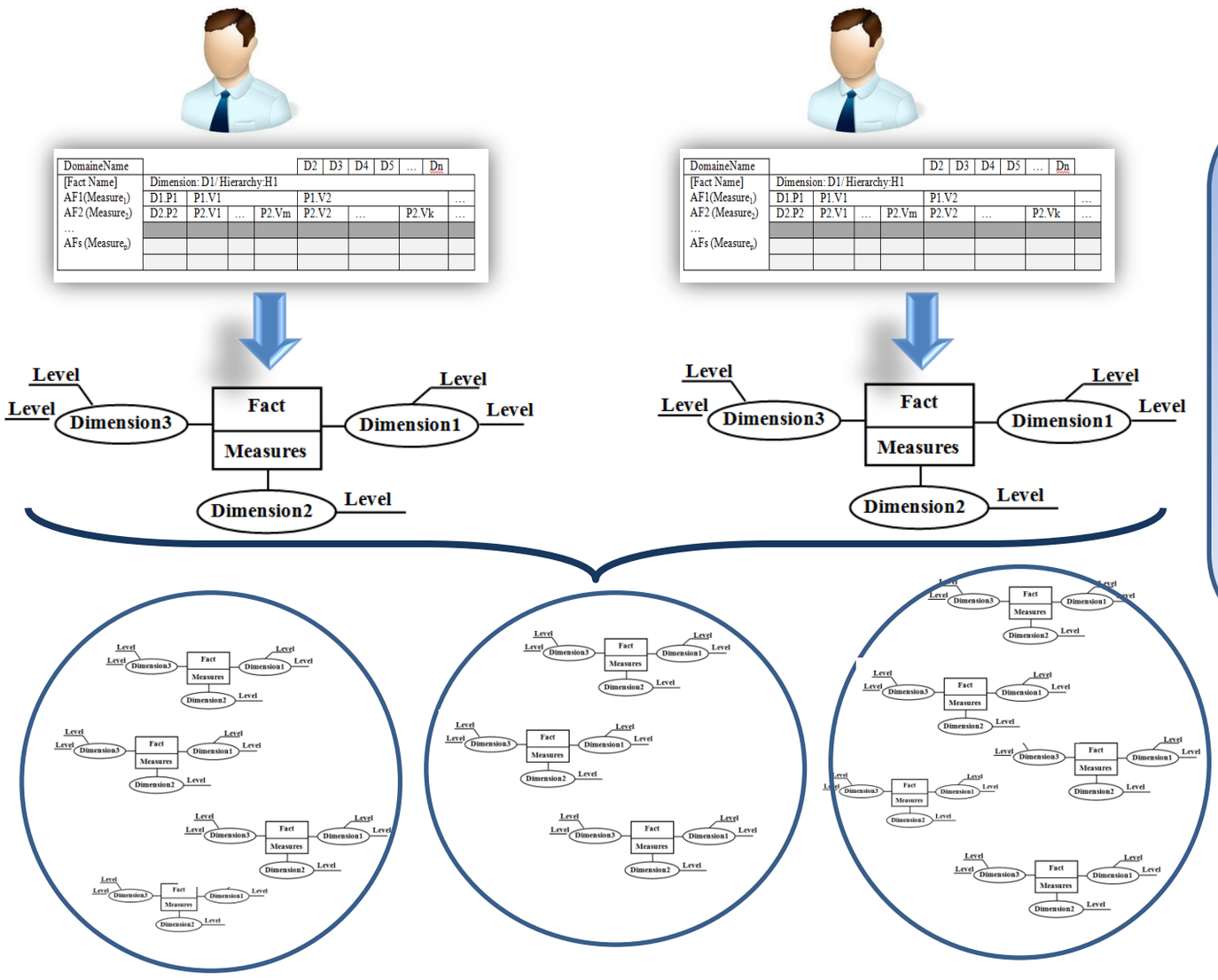
Clustering OLAP Requirements using aK-Mode

Nouha Arfaoui arfaoui.nouha@yahoo.fr
Jalel Akaichi jalel.akaichi@isg.rnu.tn
Institut Supérieur de Gestion Tunis
Tunisie

Introduction: The data warehousing is becoming increasingly important in terms of strategic decision making through their capacity to integrate heterogeneous data from multiple information sources in a common storage space, for querying and analysis. Since its design is not an easy task, we propose exploiting the OLAP requirements to construct the schemas of data marts which will be used next to build the schema of the Data Warehouse.
In this work we will focus on the clustering of OLAP requirements using a new algorithm aK-Mode.



1. OLAP Requirement

The n-dimensional table (Multidimensional Table (MT)) is a way to express the needs of the decision makers (they can find difficulties to express their needs using the SQL queries (especially when using the GROUP BY and/or HAVING clauses)). The MT is a tabular representation that can show the fact of the decision maker. This fact and its measures can be analyzed according to dimensions and their granularity levels. This choice is done because: i) It is easy to be used by a non computer scientist user, ii) It allows seeing values of certain attributes as a function of others, iii) The representation is close to decision makers' vision of data .

2. Clustering

Clustering is the unsupervised classification of patterns into groups called Clusters , it involves dividing a set of data points into non-overlapping groups, or cluster of points . The schemas of one cluster are more similar to those of another one, so the clustering aims at maximizing the homogeneity within the same group.
The base of each algorithm is the use of coefficients to calculate the similarity/dissimilarity measure between schemas
The problem: the schemas provide additional information that can influence the result of clustering. Using the traditional version of k-mode, this level is ignored.
The solution: we propose "aK-Mode" that extends Simple Matching (SM) dissimilarity measure by adding the ontology.

Algorithm	Type	Complexity	Coefficient
K-MODE	Partitioning	O (n)	Simple Matching
ROCK	Hierarchical	$O(kn^2)$	Links
QROCK	Hierarchical	$O(n^2)$	Threshold
CACTUS	Hierarchical	Scalable	Support
COOLCAT	Hierarchical	$O(n^2)$	Entropy
CLICK	Partitioning	Scalable	Co-occurrence
LIMBO	Hierarchical	O (nLogn)	Information Bottleneck
MULIC	Partitioning	$O(n^2)$	Hamming measure
HIERDENC	Hierarchical	O (n)	Simple Matching

3. The ontology

The ontology is used to resolve the heterogeneity problem in information sharing environments, but in our case, it can be used to improve the document quality using the hierarchical knowledge

3.1. The domain ontology

It contains information about different classes as well as the relationships between them.

3.2. The Wordnet

It is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations

4. The algorithm

Algorithm NewSimpleMatching

Input: S1, Mode // *two schemas*
Output: CoefSM
Begin
CoefD = SimilarityFunctionD (S1, Mode)
CoefM = SimilarityFunctionM (S1, Mode)
CoefL = SimilarityFunctionL (S1, Mode)

CoefSM=[(MaxD – CoefD)/MaxD]+(MaxM – CoefM)/MaxM]+
(MaxL – CoefL) /MaxL]
End

Algorithm SimilarityFunctionD

Input: S1, Mode // *two schemas*
Output: CoefD
Begin
For (each S1.Dimension)
if(Mode.Dimension.equals(OntoTerm(S1.Dimension))
coefSD++
End If
End For
End

Domain: it indicates the domain to which the schema belongs.
Schema: it is used to group the different facts, dimensions, measures, hierarchies and levels that belong to the same schema
Fact: it corresponds to the subject of analysis.
Dimension: it corresponds to the axe of analysis.
Measure: every fact has one or more measures that are numerical.
Hierarchy: it is a logical structure used to order levels as a means of organization data.
Level: it represents a position in a hierarchy.

is-Schema (Si, Dj): "Si" is a schema that belongs to the domain "Dj".
is-Fact (Fi, Sj): "Fi" is a fact that belongs to the schema "Sj".
is-Dimension (Di, Fj): "Di" is a dimension that belongs to the fact "Fj".
is-Measure (Mi, Fj): "Mi" is a measure that belongs to the fact "Fj".
is-Hierarchy (Hi, Dj): "Hi" is a hierarchy that characterizes the dimension "Dj".
is-Level (Li, Hj): "Li" is a level that exists into the hierarchy "Hj".

-**Partial-Name**: It takes into consideration the elements that have been pre- or post-fixed.
- **Levenshtein-Name**: it takes into consideration the misspellings or the use of different legal spelling variants.
- **Synonym**: it takes into consideration the use of different but synonymous words for the same thing.