Similarity Measures for Multidimensional Data

Eftychia Baikousi
Georgios Rogkakos
Panos Vassiliadis

University of Ioannina
Dept. of Computer Science

Cube 1 or Cube 2 most Similar to Cube 0 ?

Cube 0

Date →

Product ↓

	2009	2010
Cola	10	12
Fanta	5	8
Chips	5	5
Popcorn	10	15

Cube 1

	2009	2010
Drinks	15	20
Snack	15	20

Cube 2

	2008	2009
Cola	7	10
Fanta	4	5
Chips	6	5
Popcorn	10	10

Motivating Example

Cube 1
Average sales by year

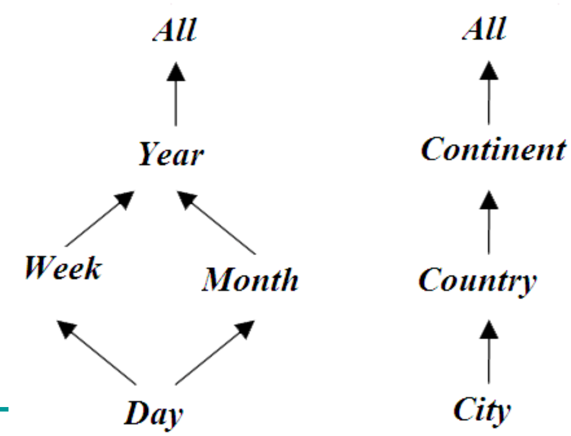
Location \ Date	2000	2001	2002
Rome	4	6	7
Athens	50	60	10
Paris	9	20	30

Cell

City	Year	Average Sales
Paris	2002	30

Cube 2
Average sales by year

Location \ Date	2002	2003	2004
London	9	8	12
Berlin	6	21	34
Madrid	53	28	5



Time Hierarchy **Location Hierarchy**

Contents

- **Background & Related Work**
- Distance Functions
  - between 2 values of a dimension
  - between 2 points in the multidimensional space
  - between 2 sets of points in the multidimensional space
- User Study Experiments
  - User study between 2 values of a dimension
  - User study between 2 sets of points in m/d space (cubes)

Background

- Fundamentals
  - Distance Measures
  - Hausdorff
  - Controversy on Metric Axioms

- Distances on Graphs
  - Highway Hierarchies
  - Semantic Similarity between Words

Distance Measures

- A distance measure is called a ***metric*** when :
  - $d(i,j) \geq 0$ & $d(i,j) = d(j,i)$ & $d(i,i) = 0$ & $d(i,j) \leq d(i,k)+d(j,k)$

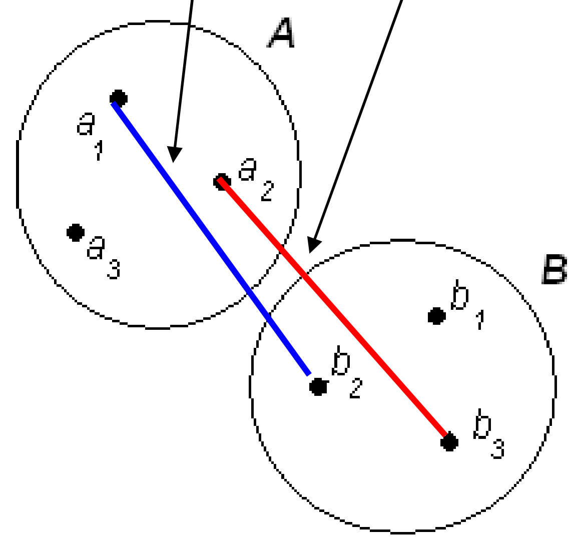
- Categorization
  - interval-scaled variables (Euclidean, Minkowski, Manhattan)
  - binary variables (Jaccard)
  - categorical variables

Hausdorff distance

- Example:

$$d_H(A,B)=max\{d_s(A,B),d_s(B,A)\}= max\{\ \mathbf{d_e(a_1, b_2)},\ \mathbf{d_e(b_3, a_2)}\}$$

  - d_e denotes the Euclidian distance
  - d_s denotes the max distance of the set of minimum distances.



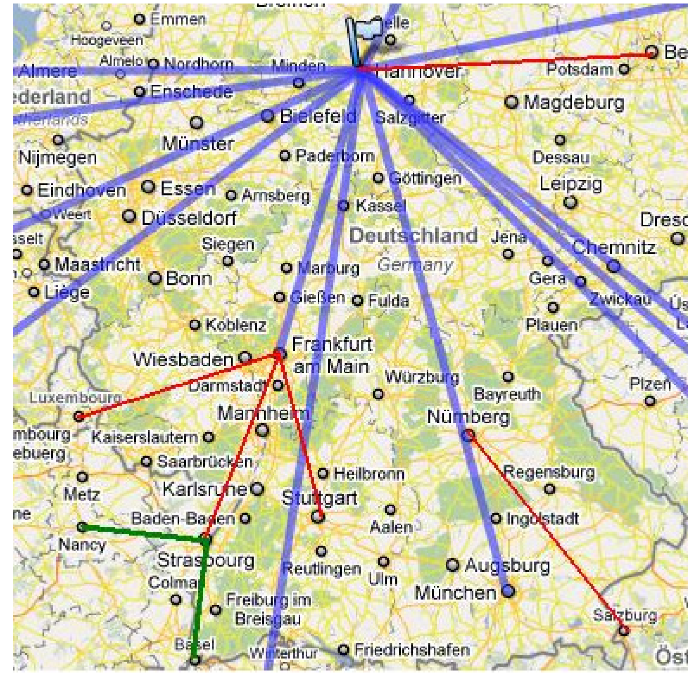
Controversy on Metric Axioms

- Properties of metrics are convenient for Mathematicians/Computer Scientists

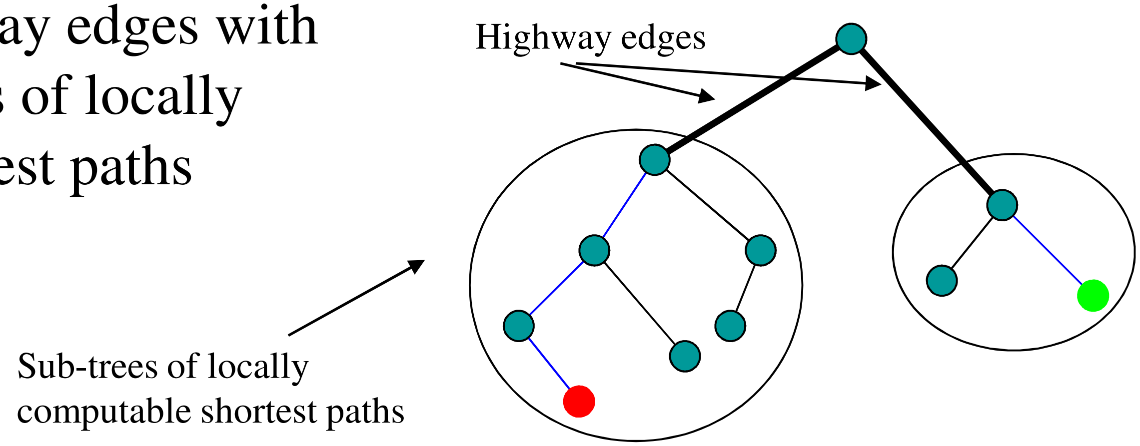
However

- Human perception does not comply with properties of metrics

Highway Hierarchies



- highways in road maps
  - The shortest paths among 2 points in a road network consists of
    - small roads locally
    - a highway road
  - Hierarchy: highway edges with attached sub-trees of locally computable shortest paths



Distances on Graphs

- Semantic Similarity between Words
  - Word similarity measures
  - Semantic hierarchies
  - 2 datasets (pairs of words)
    - One for constructing their method
    - The other to test it

Distances for Collections of Structured Data

- Relax operator
- Diff operator
- Distance between two relational databases under the same schema

Contents

- Background & Related Work
- **Distance Functions**
  - between 2 values of a dimension
  - between 2 points in the multidimensional space
  - between 2 sets of points in the multidimensional space
- User Study Experiments
  - User study between 2 values of a dimension
  - User study between 2 sets of points in m/d space (cubes)

Contents

- Background & Related Work
- Distance Functions
  - **between 2 values of a dimension**
  - between 2 points in the multidimensional space
  - between 2 sets of points in the multidimensional space
- User Study Experiments
  - User study between 2 values of a dimension
  - User study between 2 sets of points in m/d space (cubes)

Distance functions between 2 values of a dimension

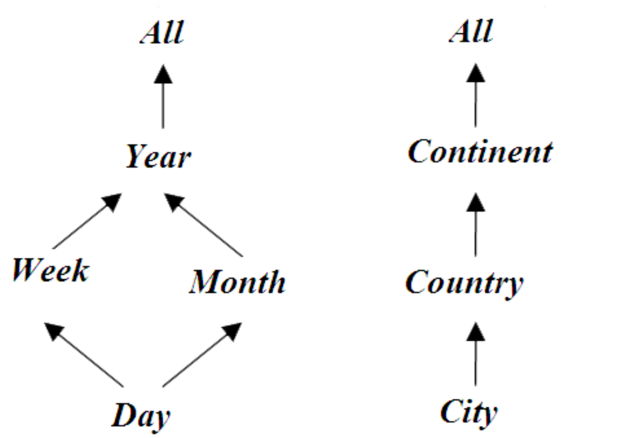
- *Locally* computable
- *Hierarchical*
- *Highway*

Distance functions between 2 values of a dimension

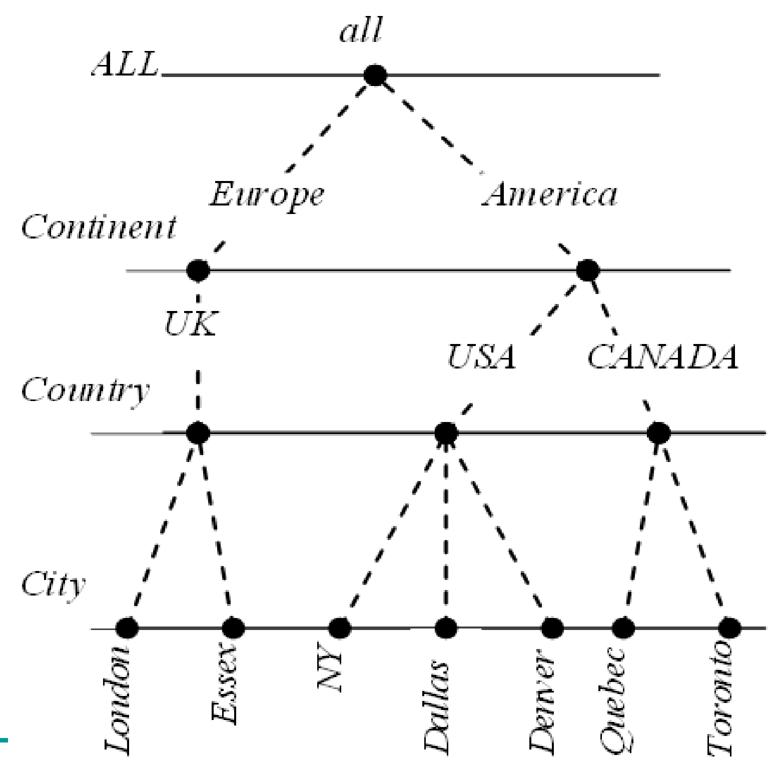
- Locally computable distance functions
  - explicit assignment
  - based on the values x and y
  - based on Attribute values

Hierarchical distance functions

- *W.r.t. an aggregation function*
- *W.r.t. Hierarchy Path*
- *Percentage distance functions*
- *Highway distance functions*

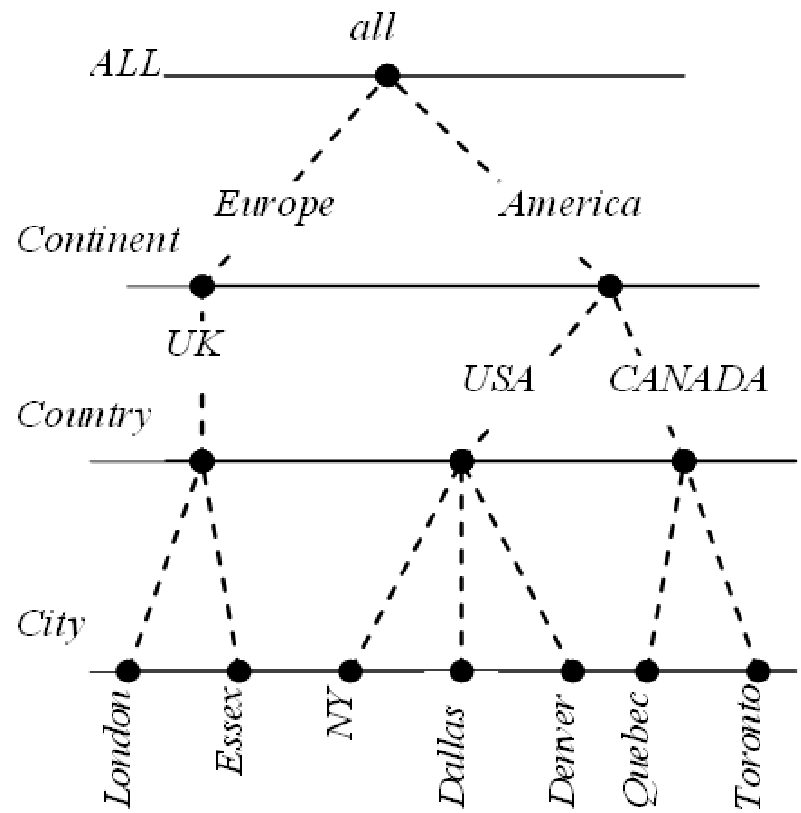


Time Hierarchy **Location Hierarchy**



Distance functions w.r.t. an aggregation function

- $x \in L_i$, $L_L \prec L_i$
- $desc_{L_L}^{L_i}(x)$ set of its descendants
- $$x_{aggr} = f_{aggr}(desc_{L_L}^{L_i}(x))$$
  $$y_{aggr} = f_{aggr}(desc_{L_L}^{L_j}(y))$$
  - f_{aggr} : *count*, *min*, *max*, *avg*, *sum*
- $\boxed{dist(x, y) = g(x_{aggr}, y_{aggr})}$
  - g can be from the locally computable functions



Distance Functions w.r.t. Hierarchy Path

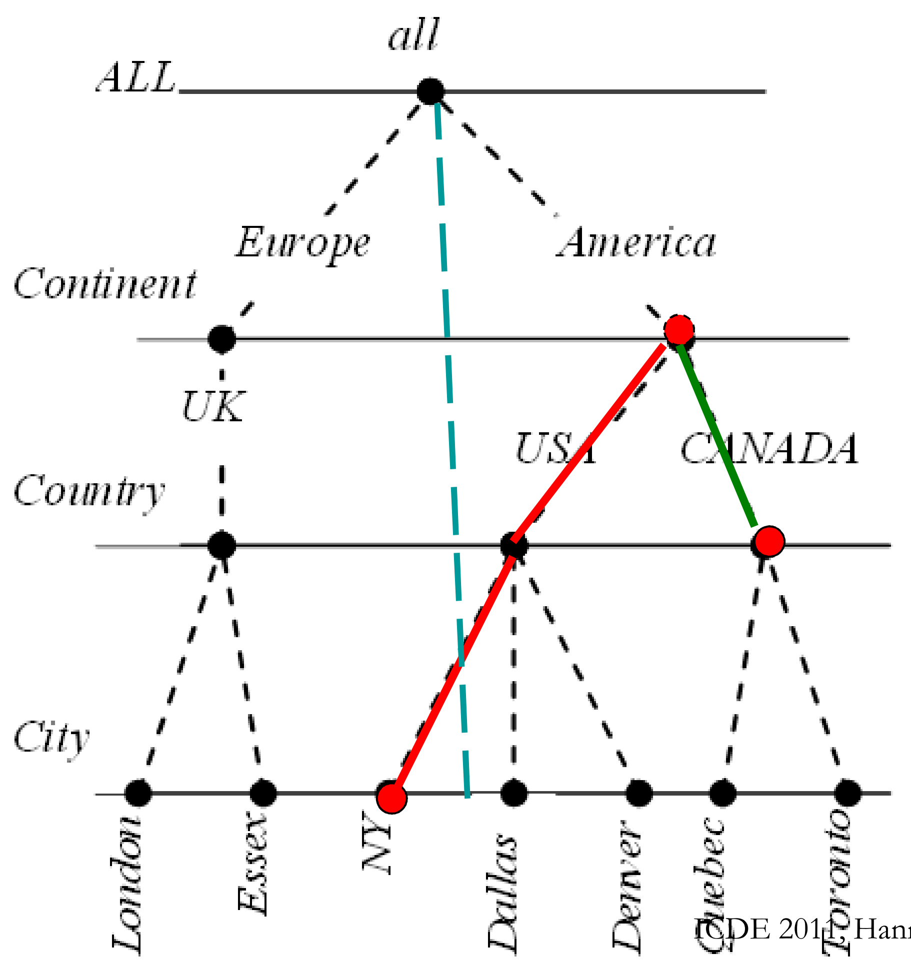
- Assume 2 values *x* and *y* s.t.
  - $x \in L_x$ and $y \in L_y$

- *lca* (*x*,*y*) : the *Lowest Common Ancestor* of *x* and *y*

  - $d_{\text{path}}(x, y) = \left(\dfrac{w_x * |path(x, lca)| + w_y * |path(y, lca)|}{(w_x + w_y) * |path(ALL, L_1)|} \right)$
  - $d_{\text{depth}}(x, y) = \left(\dfrac{|path(lca, L_1)|}{|path(ALL, L_1)|} \right)$

Example w.r.t. Hierarchy Path

- x=‘NY’, y=‘Canada’ $lca(x,y)$ = ‘America’



- $f_{\text{path}} = \dfrac{\boxed{|\,path(x, lca)\,|} + \boxed{|\,path(y, lca)\,|}}{2 * \boxed{|\,path(ALL, L_1)\,|}}$

  $= (2+1)/2*3 = 0.5$

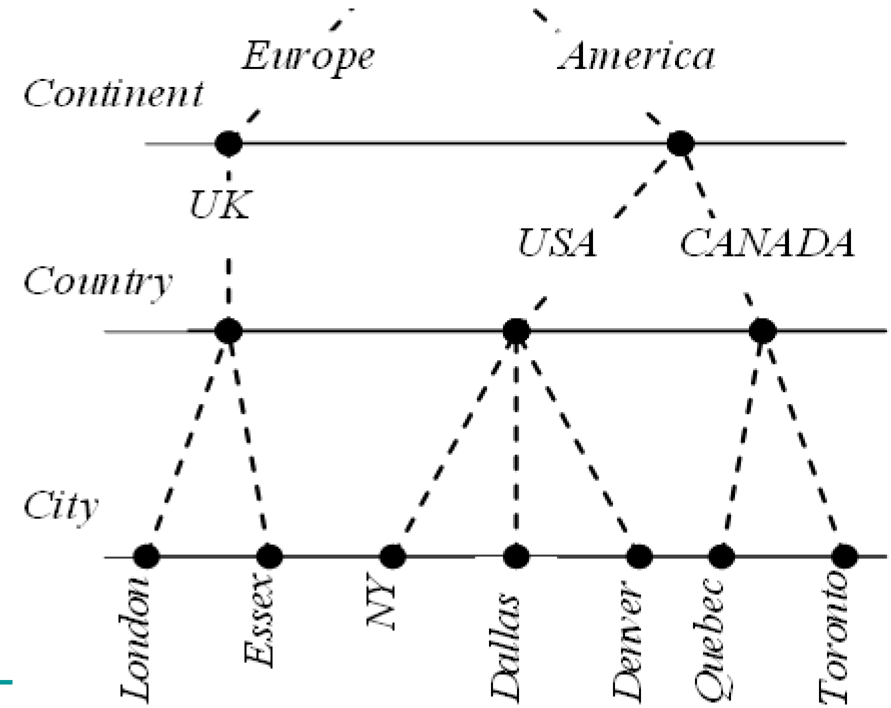
- $f_{\text{depth}} = \dfrac{|\,path(lca, L_1)\,|}{\boxed{|\,path(ALL, L_1)\,|}} = 2/3$

Percentage distance functions

- $\boxed{dist(x, y) = 1 - \frac{|desc_{L_i}^{L_x}(x)|}{|desc_{L_i}^{L_y}(y)|}}$, only when y is an ancestor of x
- the percentage of occurrences over the values of the hierarchy
- Example: $dist(\text{'USA'}, \text{'America'})$

$$1 - \frac{|desc_{City}^{Country}(\text{'USA'})|}{|desc_{City}^{Continent}(\text{'America'})|} = \frac{2}{5}$$

where L_i is the detailed level L_{city}



Highway Distance Functions

- Every level L grouped into k groups,
- $\boldsymbol{r}_{\mathbf{k}}$ the representative
  - distance between two representatives can be thought of as a highway

$$\boxed{d(x, y) = d(x, \boldsymbol{r}_{\mathbf{x}}) + d(\boldsymbol{r}_{\mathbf{x}}, \boldsymbol{r}_{\mathbf{y}}) + d(y, \boldsymbol{r}_{\mathbf{y}})}$$

  - $\boldsymbol{r}_{\mathbf{x}}$, $\boldsymbol{r}_{\mathbf{y}}$: representatives of the groups of x, y
  - representative selected w.r.t an ancestor or a descendant

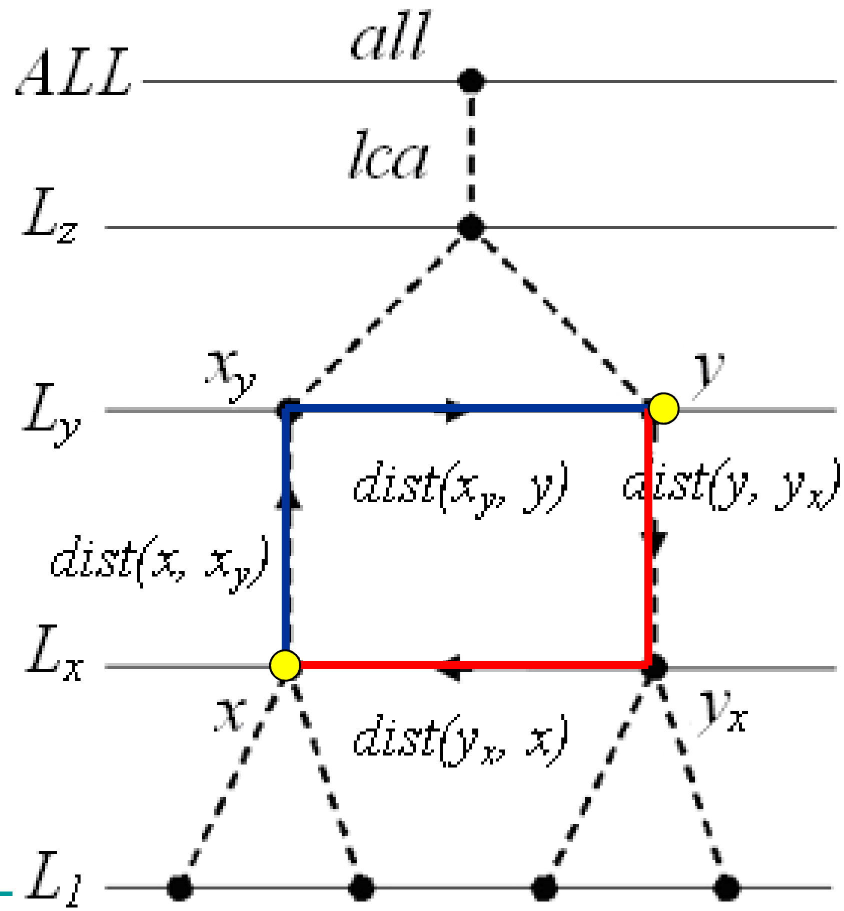
Highway Distance Functions

- **$r_{x,}$** : is an ancestor

$$d(x, y) = d(x, x_y) + d(x_y, y)$$

- **r_y** : is an descendant

$$d(x, y) = d(x, y_x) + d(y_x, x)$$



Contents

- Background & Related Work
- Distance Functions
  - between 2 values of a dimension
  - **between 2 points in the multidimensional space**
  - between 2 sets of points in the multidimensional space
- User Study Experiments
  - User study between 2 values of a dimension
  - User study between 2 sets of points in m/d space (cubes)

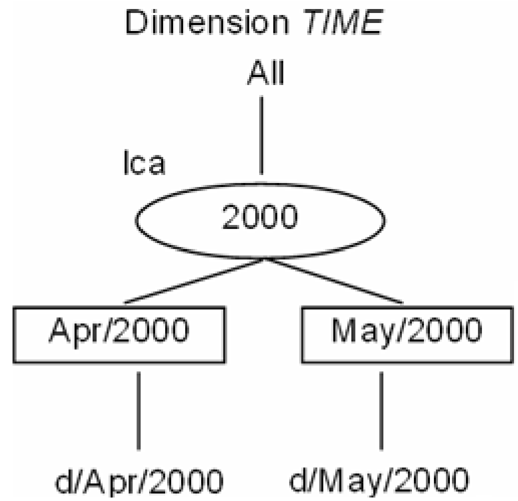
Distance functions between 2 points in the multidimensional space

- Assume two cells from a cube
  - $c_1 = (l_1^1, l_2^1, \ldots, l_n^1, m_1^1, m_2^1, \ldots, m_m^1)$
  - $c_2 = (l_1^2, l_2^2, \ldots, l_n^2, m_1^2, m_2^2, \ldots, m_m^2)$
- $dist(c_1,c_2)$ can be expressed w.r.t.
  - their level coordinates $d_i(L_i^1, L_i^2)$ and
  - their measure values $d_i(M_i^1, M_i^2)$

$$\boxed{dist(c_1, c_2) = f\,(d_i(L_i^1, L_i^2), d_i(M_i^1, M_i^2))}$$

Weighted Sum

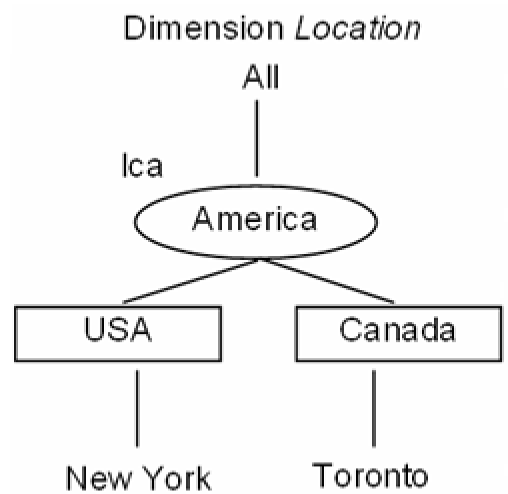
$$f: \frac{\sum_{i=1}^{n} w_i d_i(l_i^1, l_i^2)}{\sum_{i=1}^{n} w_i} + \frac{\sum_{i=1}^{m} w_i' d_i(m_i^1, m_i^2)}{\sum_{i=1}^{m} w_i'}$$



- **Example**

	Month	Country	Sales
c_1	*May/2000*	*USA*	*4*
c_2	*Apr/2000*	*Canada*	*3*

$$\frac{0.5 * (d(M_{c_1}, M_{c_2}) + d(C_{c_1}, C_{c_2}))}{0.5 + 0.5} + \frac{0.5 * d(S_{c_1}, S_{c_2})}{0.5}$$



- Minkowski

$$L_p = \sqrt[p]{\sum_{i=1}^{n} (d_i(l_i^{\,1}, l_i^{\,2}))^p} + \sqrt[p]{\sum_{i=1}^{m} (d_i(m_i^{\,1}, m_i^{\,2}))^p}$$

p-norm

- Minimum Partial distance
  - cells $c_1 = (l_1^{\,1}, l_2^{\,1}, \ldots, l_n^{\,1}, m_1^{\,1}, m_2^{\,1}, \ldots, m_m^{\,1})$
    $c_2 = (l_1^{\,2}, l_2^{\,2}, \ldots, l_n^{\,2}, m_1^{\,2}, m_2^{\,2}, \ldots, m_m^{\,2})$

$$dist(c_1, c_2) = \min_{d_i}\{d_i(l_i^{\,1}, l_i^{\,2})\} + \min_{d_i}\{d_i(m_i^{\,1}, m_i^{\,2})\}$$

- Proportion of common coordinates

$$\frac{count(l_i^{\,1} = l_i^{\,2} \forall i \in \{1,2,\ldots,n\})}{n} + \frac{count(m_i^{\,1} = m_i^{\,2} \forall i \in \{1,2,\ldots,m\})}{m}$$

  - n: number of level values, m: number of measures
  - the number of level values same for both cells
  - the number of measures that have the same value for both cells

Contents

- Background & Related Work
- Distance Functions
  - between 2 values of a dimension
  - between 2 points in the multidimensional space
  - **between 2 sets of points in the m/d space**
- User Study Experiments
  - User study between 2 values of a dimension
  - User study between 2 sets of points in m/d space (cubes)

Distance functions between 2 sets of points in m/d space

- Cubes: C of l cells and C' of k cells
  - $c = (l_1, l_2, \ldots, l_n, m_1, m_2, \ldots, m_m)$
  - $c' = (l_1', l_2', \ldots, l_n', m_1', m_2', \ldots, m_m')$
- $dist(C, C') = f\,(dist(c, c'))$
  - f : a function of the partial distances $dist(c, c')$

$CUBE_1$

	Day	City	Sales
c_1	3/5/2000	London	5
c_2	3/5/2001	New York	6
c_3	4/5/2001	New York	7

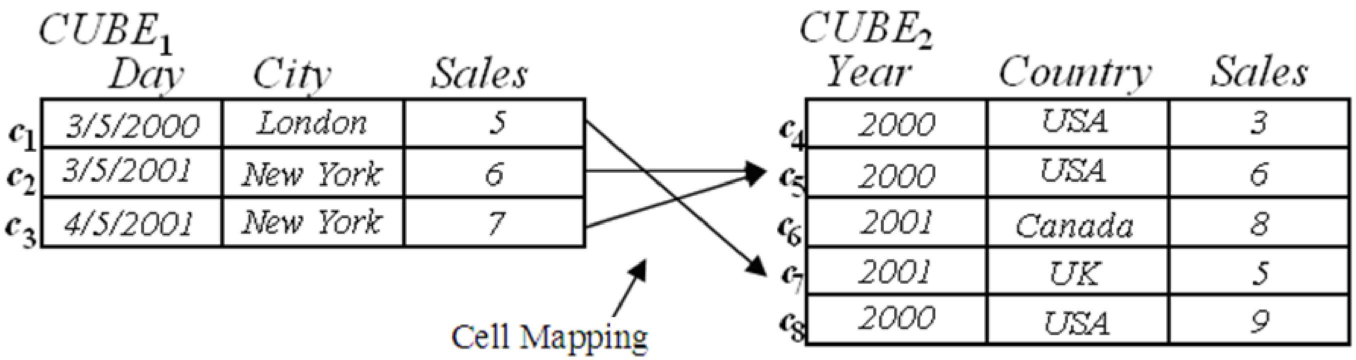
$CUBE_2$

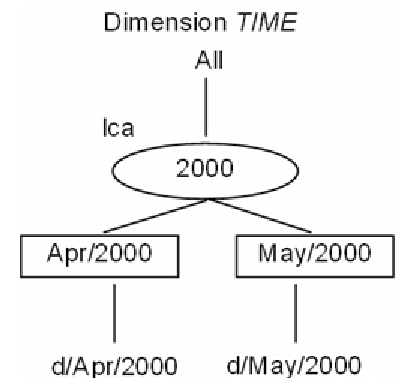
	Year	Country	Sales
c_4	2000	USA	3
c_5	2000	USA	6
c_6	2001	Canada	8
c_7	2001	UK	5
c_8	2000	USA	9

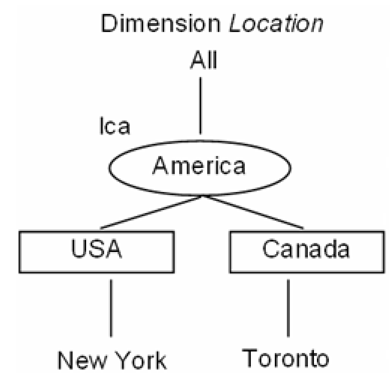
The *Cell Mapping* method

- Map a cell in a cube to the "closest possible representative" cell in another cube
- Compute all dimension value distances between every cell of 1st cube with every cell of 2nd cube
- The *Mapped* cell of 2nd cube: The cell with the less distance from a cell of 1st cube

The cell mapping method







- $d_{dim}(c_1,c_4)= \frac{1/3+5/6}{2} = \frac{7}{12}$
- $d_{dim}(c_1,c_5)= \frac{1/3+5/6}{2} = \frac{7}{12}$
- $d_{dim}(c_1,c_6)= \frac{2/3+5/6}{2} = \frac{9}{12}$
- $d_{dim}(c_1,c_7)= \frac{2/3+1/6}{2} = \frac{5}{12}$
- $d_{dim}(c_1,c_8)= \frac{2/3+5/6}{2} = \frac{9}{12}$

- $d_{dim}(c_2,c_4)= \frac{2/3+1/6}{2} = \frac{5}{12}$
- $d_{dim}(c_2,c_5)= \frac{2/3+1/6}{2} = \frac{5}{12}$
- $d_{dim}(c_2,c_6)= \frac{2/3+1/6}{2} = \frac{5}{12}$
- $d_{dim}(c_2,c_7)= \frac{2/6+5/6}{2} = \frac{7}{12}$
- $d_{dim}(c_2,c_8)= \frac{2/3+1/6}{2} = \frac{5}{12}$

- $d_{dim}(c_3,c_4)= \frac{2/3+1/6}{2} = \frac{5}{12}$
- $d_{dim}(c_3,c_5)= \frac{2/3+1/6}{2} = \frac{5}{12}$
- $d_{dim}(c_3,c_6)= \frac{2/3+1/6}{2} = \frac{5}{12}$
- $d_{dim}(c_3,c_7)= \frac{2/6+5/6}{2} = \frac{7}{12}$
- $d_{dim}(c_3,c_8)= \frac{2/3+1/6}{2} = \frac{5}{12}$

Closest Relative

Cell mapping part of the function

$$dist(C,C') = \frac{\sum_{i=1}^{k}(dist(c_i, c'))}{k} \quad \forall c' \mid dist_{dim}(c_i, c') = \min_{i}\{dist_{dim}(c_i, c')\}$$

- *dist*$_{\text{dim}}$: the distance between two cells according to their dimension values

- Each one of the *k* cells from cube *C* is mapped to the cell of the cube *C'* that has the minimum *dist*$_{\text{dim}}$ from it.

Closest Relative

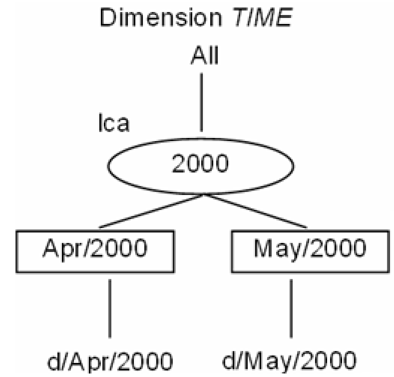
$CUBE_1$

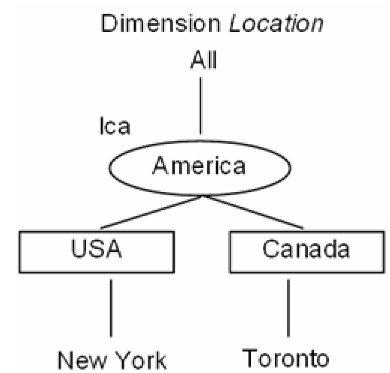
	Day	City	Sales
c_1	3/5/2000	London	5
c_2	3/5/2001	New York	6
c_3	4/5/2001	New York	7

$CUBE_2$

	Year	Country	Sales
c_4	2000	USA	3
c_5	2000	USA	6
c_6	2001	Canada	8
c_7	2001	UK	5
c_8	2000	USA	9

Cell Mapping





- cells c_1, c_2, c_3, mapped to cells c_7, c_5, and c_5

$d(c_1,c_7)= 5/12$, $d(c_2,c_5)= 5/12$, $d(c_3,c_5)= 5/12$

- Dimensions : f_{path}, cells: *weighted sum*,

$$d(CUBE_1,CUBE_2)= \frac{d(c_1,c_7)+ d(c_2,c_5)+ d(c_3,c_5)}{3}$$

Hausdorff

- $\boxed{H(C, C') = max(h(C,C'), h(C',C))}$
  - $h(C, C')$: directed Hausdorff
    - measures the max distance of a cube C to the "nearest" cell of the other cube C'
  - $h(C,C') = \max_{c \in C}\{\min_{c' \in C'}\{dist(c, c')\}\}$
  - $dist(c, c')$ distance between two cells c and c'
  - Includes bidirectional cell mapping method

Hausdorff computation

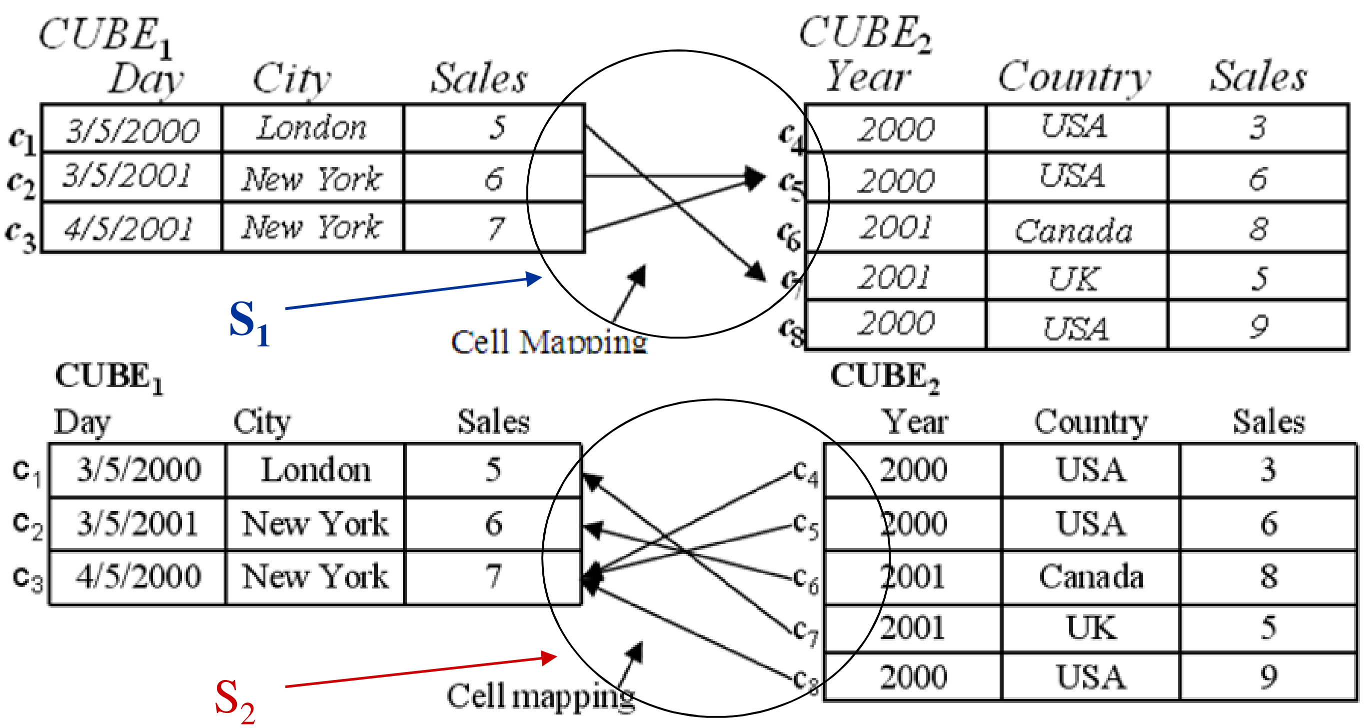
- Two sets of mapped cells
- For each set
  - for every pair of mapped cells
    - compute their distance considering their measures as well
- Obtain two sets of min distances between cells

  a) from *C* to *C*'

  b) from *C*' to *C*

  - For each set pick the greatest distance
- Pick the greater of the two greatest distances

Hausdorff



$S_1 : \{d(c_1,c_7), d(c_2,c_5), d(c_3,c_5)\}$

$S_2 : \{d(c_4,c_3), d(c_5,c_3), d(c_6,c_2), d(c_7,c_1), d(c_8,c_3)\}$

- $d(CUBE_1,CUBE_2) = max\{max\{S_1\},max\{S_2\}\} = max\{5/12,5/12\}=5/12$

Contents

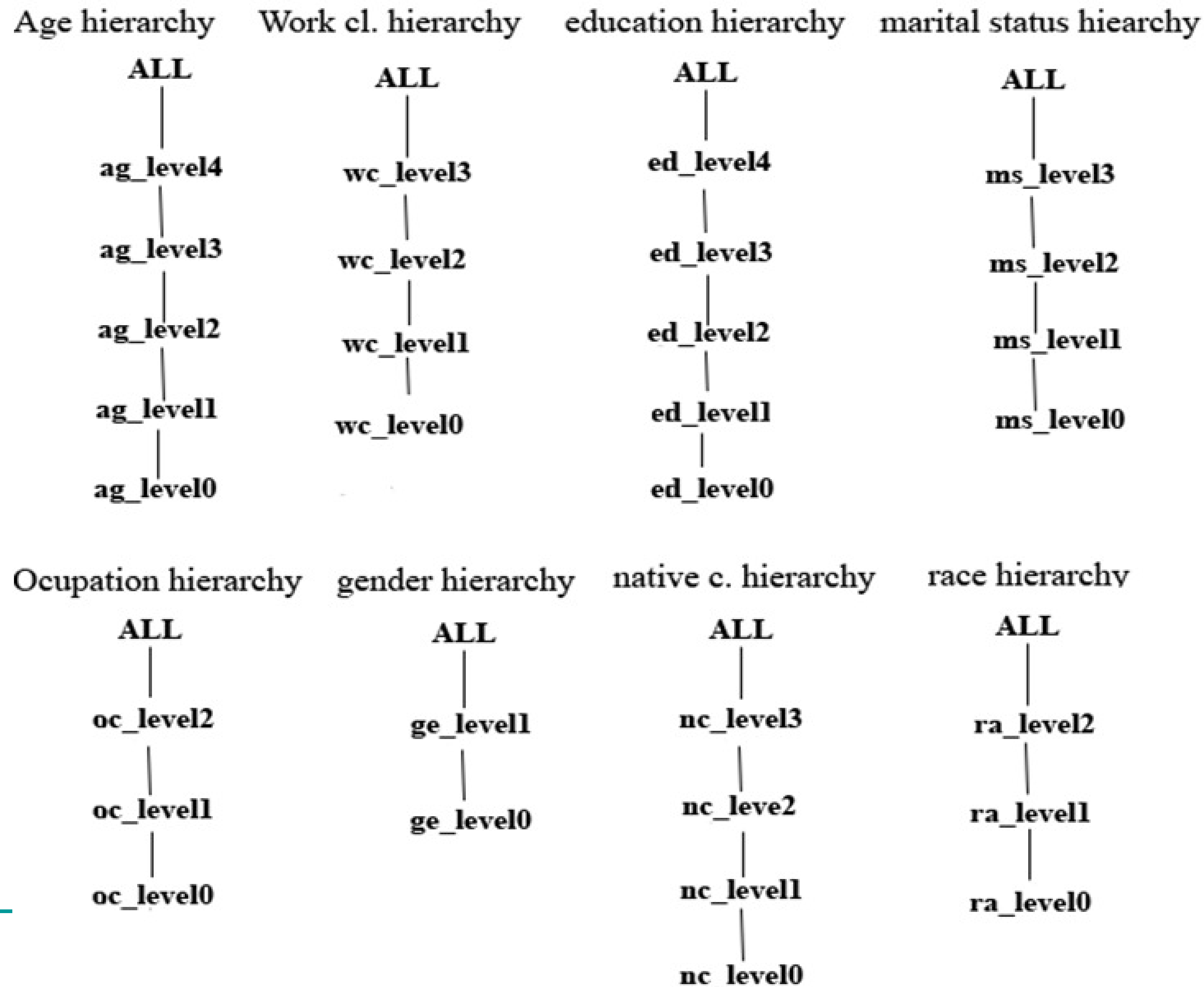
- Background & Related Work
- Distance Functions
  - between 2 values of a dimension
  - between 2 points in the multidimensional space
  - between 2 sets of points in the multidimensional space
- User Study Experiments
  - **<u>User study between 2 values of a dimension</u>**
  - User study between 2 sets of points in m/d space (cubes)

User study between 2 values of a dimension

- 15 users *users_all*
  - 10 *users_cs, 5 users_non*
- Dataset: '*Adult*'

Table	Value Type	# Tuples	# Dim. Levels
Adult fact		30418	-
Age Dim.	*Numeric*	72	5
Education Dim.	*Categorical*	16	5
Gender Dim.	*Categorical*	2	2
Marital Status Dim.	*Categorical*	7	4
Native Country Dim.	*Categorical*	41	4
Occupation Dim.	*Categorical*	14	3
Race Dim.	*Categorical*	5	3
Work Class Dim.	*Categorical*	7	4

Dimension Hierarchies of Adult



Experimental setting

- **Purpose** of the experiment:
  - which distance function between two values of a dimension is best in regards to the user preferences
- Each user was given 14 scenarios
- Each scenario contains:
  - a reference cube
  - a set of *variant* cubes
- variant cubes: slightly altering the reference cube
- The 14 scenarios included different kinds of cubes
  - value types, levels of granularity

Variant cubes

- altering
  - granularity level for one dimension
  - value range of the reference cube
- Example
  - reference cube
    - dimension levels *Age_level*1, *WorkClass_level*1
    - age interval [52, 56].
  - 1st type modification: change dimension level (e.g.,*age_level1* to *age_level*2)
  - 2nd type modification: change the age interval to [22, 26] or to [17, 26].

ag_level1	wc_level1
52-56	Gov
52-56	Private
52-56	Self-emp
52-56	Without-pay

ag_level2	wc_level1
47-56	Gov
47-56	Private
47-56	Self-emp
47-56	Without-pay

ag_level1	wc_level1
47-51	Gov
47-51	Private
47-51	Self-emp
47-51	Without-pay

Sample scenario

- Reference Cube

Cube4

ag_level1	wc_level1	ra_level1
52-56	Gov	White
52-56	Private	Colored
47-51	Self-emp	White
52-56	Without-pay	White

- Variant Cubes

Cube4_4

ag_level1	wc_level1	ra_level1
52-56	Gov	White
52-56	Private	Colored
52-56	Self-emp	White
52-56	Without-pay	White

Cube4_5

ag_level1	wc_level1	ra_level1
27-31	Gov	Colored
52-56	Private	Colored
47-51	Self-emp	White
52-56	Without-pay	White

Cube4_6

ag_level1	wc_level1	ra_level1
47-51	Self-emp	White
52-56	Without-pay	White

Cube4_8

ag_level2	wc_level1	ra_level1
47-56	Gov	White
47-56	Private	Colored
47-56	Self-emp	White
47-56	Without-pay	White

Cube4_1

ag_level1	wc_level1	ra_level1
37-41	Gov	White
37-41	Private	White
47-51	Self-emp	White
62-66	Without-pay	White

Cube4_2

ag_level1	wc_level1	ra_level1
52-56	Gov	White
47-51	Private	White
47-51	Self-emp	White
52-56	Without-pay	White

Cube4_3

ag_level1	wc_level1	ra_level1
37-41	Gov	White
37-41	Private	White
42-46	Self-emp	White
42-46	Without-pay	White

Cube4_7

ag_level1	wc_level2	ra_level1
52-56	With-Pay	White
52-56	With-Pay	Colored
47-51	With-Pay	White
52-56	Without-pay	White

Scenarios of User Study

- Each *variant cube*: most similar to the *reference cube* according to a distance function
- 14 scenarios organized as:
  - cubes with arithmetic type values (5 scenarios)
  - cubes with categorical type values (2 scenarios)
  - cubes with mixed type values (7 scenarios)

Notation of distance functions

Family	Abbr.	Distance function name
Local	δ_{M}	Manhattan
Aggregation	$\delta_{\mathrm{Low,c}}$	With respect to a lower level of hierarchy f_{aggr} =count
	$\delta_{\mathrm{Low,m}}$	With respect to a lower level of hierarchy f_{aggr} = max
Hierarchical Path	$\boldsymbol{\delta}_{\mathbf{LCA,P}}$	**Lowest common ancestor through $\boldsymbol{f}_{\mathbf{path}}$**
	$\delta_{\mathrm{LCA,D}}$	Lowest common ancestor through f_{depth}
Percentage	$\delta_{\%}$	Applying percentage function
Highway	$\boldsymbol{\delta}_{\mathbf{Anc}}$	**With respect to an ancestor $\boldsymbol{x}_{\mathbf{y}}$**
	δ_{Desc}	With respect to a descendant y_{x}
	$\boldsymbol{\delta}_{\mathbf{H,Desc}}$	**Highway, selecting the representative from a descendant**
	$\delta_{\mathrm{H,Anc}}$	Highway, selecting the representative from an ancestor

- *Top three most preferred distance functions*

	Users_all	Users_cs	Users_non
$\boldsymbol{\delta}_{\mathbf{LCA,P}}$	40.47%	38.57%	44.28%
$\boldsymbol{\delta}_{\mathbf{Anc}}$	18.09%	20%	14.28%
$\boldsymbol{\delta}_{\mathbf{H,Desc}}$	9.52%	10.71%	7.14%

- *Most preferred function by users w.r.t value type*

Value Type	Users_all	Users_cs	Users_non
Arithmetic	δ_{Anc}	$\delta_{LCA,P}$, $\delta_{H,Desc}$, δ_{Anc}	$\delta_{LCA,P}$
Categorical	$\delta_{LCA,P}$	$\delta_{LCA,P}$	$\delta_{LCA,P}$
Arithmetic & Categorical	δ_{Anc}	δ_{Anc}	$\delta_{LCA,P}$, δ_{Anc}

winner distance function **per scenario**

- *winner function*: is the most frequent function per scenario for all 15 users
- The most frequent winner function was $\delta_{\text{LCA,P}}$
- Percentages
  - 35.71% for the *Users_all* group
  - 35,71% for the *Users_cs* group
  - 57.14% for the *Users_non* group

Diversity and spread of user choices

- Two major findings
  - (a) All functions were picked by some user
  - (b) certain functions appeared as user choices for all users of a user group
    - $\delta_{\mathrm{LCA,P}}$, $\delta_{\mathrm{H,Desc}}$ and δ_{Anc} for *Users_cs*
    - $\delta_{\mathrm{LCA,P}}$, $\delta_{\mathrm{Low,m}}$ and δ_{Anc} for *Users_non*

most preferred family of functions

	Local	Aggregation	Hierarchy Path	Percentage	Highway
Users_cs	1	9	69	9	52
Users_non	2	5	34	5	24
Users_all	3	14	103	14	76

Selection stability of users

- 13th and 14th scenarios replicas of 3rd and 10th scenario
  - 4 out of 5 *Users_non* users
  - 6 out of 10 *Users_cs* users
- selected the same function for **both** of the two replicas scenarios
- The rest of the users selected the same function for only **one** replica

Contents

- Background & Related Work
- Distance Functions
  - between 2 values of a dimension
  - between 2 points in the multidimensional space
  - between 2 sets of points in the multidimensional space
- User Study Experiments
  - User study between 2 values of a dimension
  - **User study between 2 sets of points in m/d space (cubes)**

User study between 2 sets of points in M/D space

- which distance function between two cubes do the users prefer?
  - *Closest Relative*
  - *Hausdorff*

- Between dimensions $\delta_{\mathrm{LCA,P}}$
- Between cells *weighted sum*

Scenarios of User Study

- 14 scenarios
  - Each scenario contains 4 cubes (*A*, *B*, *C*, *D*)
    - Cube A: reference cube
    - *B*,*C*,*D*: variant cubes
      - one most similar to *A* according to the *Closest relative*
      - one most similar to *A* according to the *Hausdorff*
      - remaining less similar to *A* for both functions
    - Users were asked **to order** the three cubes from the most similar to the less similar when compared to the cube *A*

Sample scenario

A

ag_level1	wc_level1	AVG(hours_per_week)
27-31	Gov	41.636
27-31	Private	42.2742
27-31	Self-emp	46.3854
27-31	Without-pay	65

B

ag_level1	wc_level1	AVG(hours_per_week)
37-41	Private	40.2509
62-66	Without-pay	32.7143

C

ag_level1	wc_level1	AVG(hours_per_week)
22-26	Gov	36.5979
22-26	Private	38.602
22-26	Self-emp	43.6528
22-26	Without-pay	40

D

ag_level1	wc_level1	AVG(hours_per_week)
27-31	Gov	41.636
32-36	Private	42.8008

	d(A,B)	**d(A,C)**	**d(A,D)**
Closest Relative	0.34126	0.19812	0.10799
Hausdorff	0.38151	0.25170	0.30385

Scenario groups

- *no_measures*
  - Cube distances computed ignoring measures
- *not_equal*
  - Cube distances computed with different weights between k dimensions and l measures
    - $w_d = k/k+l$, $w_m = l/k+l$
- *equal*
  - Cube distances computed with equal weights between dimensions and measures
    - $w_d = w_m$

User Reliability & Stability

- User Reliability
  - 6th scenario has cube *B* identical to cube *A*
    - 2 out of 39 users answered wrong
    - 37 valid users
- User Stability
  - 13th and 14th scenario were replicas of the 5th and 9th scenario
  - ***User_ok*** : same ordering for one scenario
  - ***User_half_ok*** : same first choice
  - ***User_Stable*** : *User_ok* for both replicas
    or *User_ok* and *User_half_ok*

User Stability

	User_OK		User_Half_OK		User_Stable	
	Frequency	**Pct**	**Frequency**	**Pct**	**Frequency**	**Pct**
13th scenario	28	75%	5	13%	24	65%
14th scenario	19	51%	8	21%	24	65%

Most frequent distance function

- Most frequent function chosen as the first ordering in all scenarios

Over all scenarios	**Frequency**	**Percentage**
Hausdorff	154	**38%**
Closest relative	**232**	**57%**
Most distant cube	21	5%

Local scenario winner

- *Local scenario winner function*:
  - function that was mostly selected as the first choice from the users in each scenario

- *closest relative*: 6 scenarios
- *Hausdorff*: 5 scenarios

Group winner function

Scenario Group	Scenario	Winning function	Winner function
no_measures	Scenario1	*Closest relative*	*Closest relative*
	Scenario2	*Closest relative*	
	Scenario3	*Closest relative*	
	Scenario4	*Hausdorff*	
not_equal	Scenario5	*Hausdorff*	*Hausdorff*
	Scenario7	*Closest relative*	
	Scenario8	*Hausdorff*	
equal	Scenario9	*Hausdorff*	*Draw: both*
	Scenario10	*Hausdorff*	
	Scenario11	*Closest relative*	
	Scenario12	*Closest relative*	

Conclusions

- Taxonomy of distances
- Distance between values of a dimension:
  - Most preferred function according to the path of the lowest common ancestor
- Distance between sets of points in a m/d space
  - *Closest relative* and *Hausdorff*
- Future work
  - More user studies
  - Combine texts

Thank you for your attention!

User study Questionnaires & Results can be found :

http://www.cs.uoi.gr/~ebaikou/publications/2011_ICDE/

References

- Simone Santini and Ramesh Jain. Similarity measures.IEEE Trans. Pattern Anal. Mach. Intell., 21(9):871–883, 1999.
- P. Sanders and D. Schultes. Highway Hierarchies Hasten Exact Shortest Path Queries. In ESA, LNCS 3669, pages 568-579, Springer, 2005.
- Yuhua Li, Zuhair Bandar and David McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. In IEEE Trans. Knowl. Data Eng, pages 871-882, 2003
- Sunita Sarawagi. idiff: Informative summarization of differences in multidimensional aggregates. Data Min. Knowl. Dis- cov., 5(4):255–276, 2001.
- Sunita Sarawagi. User-adaptive exploration of multidimensional data. In VLDB, pages 307–316, 2000.
- Heiko Müller Johann-Christoph Freytag and Ulf Leser. Describing differences between databases. In CIKM, pages 612-621, 2006