



## Semantic-based Querying of Tree-Structured Data

Δημήτρης Θεοδωράτος (New Jersey Institute of Technology, ΗΠΑ)

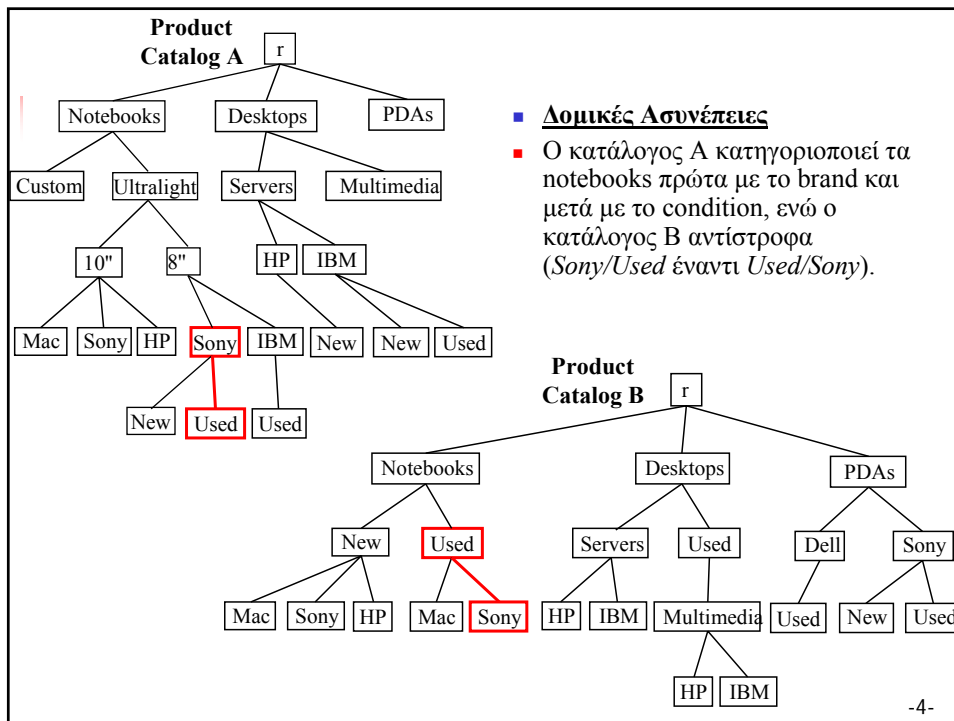
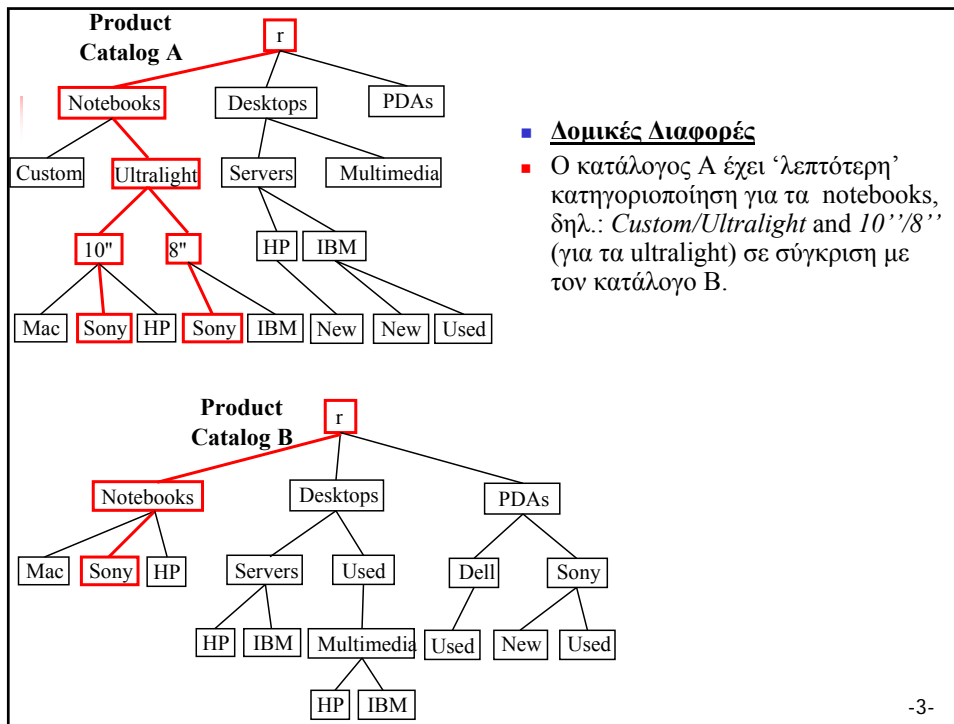
Θοδωρής Δαλαμάγκας (Εθνικό Μετσόβιο Πολυτεχνείο)

*Αντώνης Κουφόπουλος* (Εθνικό Μετσόβιο Πολυτεχνείο)



## Οργάνωση Δεδομένων με Δενδρικές Δομές

- Δεδομένα σε Δενδρικές Δομές (**tree-structured data**): ένας τρόπος να οργανώνουμε τις πληροφορίες στο Web (πχ. θεματικές κατηγορίες, κατάλογοι προϊόντων, κλπ.), κυρίως με την χρήση της γλώσσας XML.
- Οι ερωτήσεις σε tree-structured data γίνονται με την χρήση ερωτήσεων μονοπατιών (πχ. XPath και XQuery).
- Όμως, όταν εφαρμόζουμε ερωτήσεις σε **tree-structured data**, συναντάμε κάποια εμπόδια, όπως:
  - Την ημι-δομημένη μορφή των δεδομένων, δηλ. δομικές διαφορές και ασυνέπειες (**structural differences and inconsistencies**),
  - Την έλλειψη σημασιολογίας.





## Ημι-δομημένη μορφή των Tree-structured Data

- Πως επηρεάζουν οι **δομικές διαφορές** και **ασυνέπειες** την διαδικασία εφαρμογής ερωτήσεων;
  - **Ο χρήστης πρέπει να τις λάβει υπ'όψιν του στον ορισμό της ερώτησης**
  - Πρέπει να ορίσει ρητά τις διαζεύξεις όλων των δυνατών περιπτώσεων αλληλουχίας κόμβων, πχ:  
*/Notebooks/Sony/Used[price<900] OR*  
*/Notebooks/Used/Sony[price<900] OR*  
*/Notebooks/Ultralight/Sony/Used[price<900] OR ...*
- Οι χρήστες πρέπει να μπορούν να εφαρμόσουν ερωτήσεις ακόμη και **αν δεν ξέρουν** (ή **δεν ενδιαφέρονται**) για την **ακριβή** μορφή των tree-structured πηγών δεδομένων.

-5-



## Έλλειψη σημασιολογίας σε Tree-structured Data

- Τα tree-structured data παρέχουν κυρίως **συντακτική** και όχι **σημασιολογική** πληροφορία.
- Παρ'όλα αυτά, μπορεί να εμπεριέχεται κάποια σημασιολογία
  - Κάποιοι κόμβοι σχετίζονται σημασιολογικά, πχ. τα *Mac*, *HP*, *Sony* αναφέρονται στην μάρκα (brand).
  - Αυτή η πληροφορία μπορεί να γίνει μέρος της ερώτησης, και να χρησιμοποιηθεί για βελτιστοποίηση.

-6-



## Η Μέθοδός μας

- Ορίζουμε την έννοια των **γράφων διαστάσεων (dimension graphs)** για να αποτυπώσουμε την σημασιολογική πληροφορία των tree-structured data.
- Ορίζουμε μια γλώσσα ερωτήσεων για tree-structured data που δεν θα εφαρμόζεται στην δομή τους, και θα χειρίζεται επιτυχώς τις δομικές διαφορές και ασυνέπειες.
- Συζητούμε θέματα αποτίμησης των ερωτήσεων.
- Θα δείξουμε πως οι dimension graphs μπορούν να χρησιμοποιηθούν για να εφαρμόσουμε ερωτήσεις σε πολλαπλές πηγές από tree-structured data.

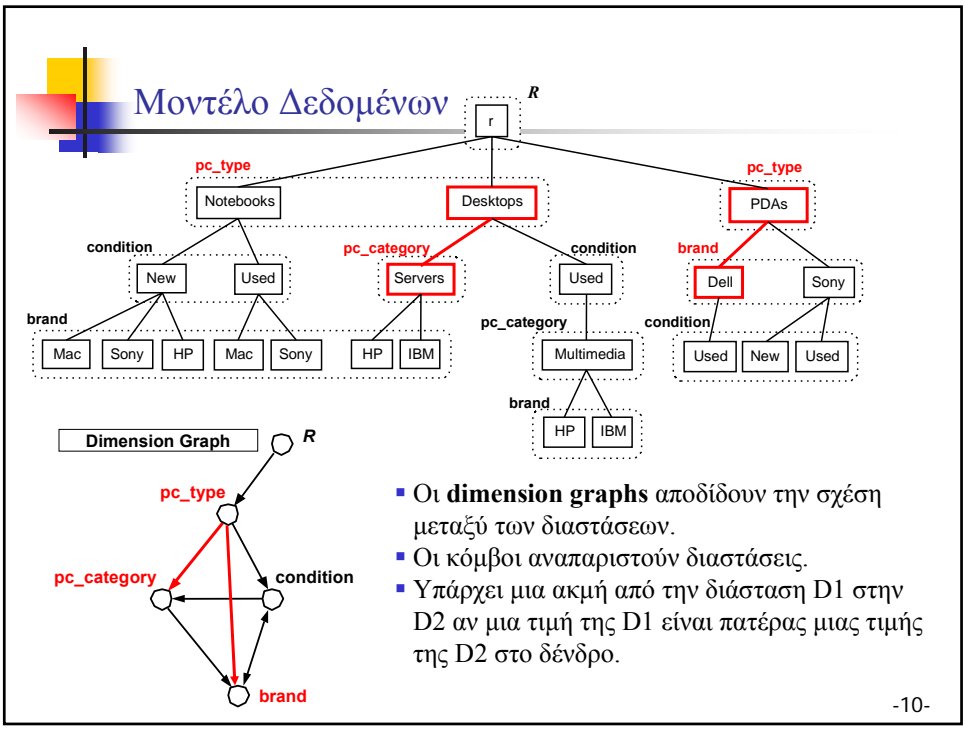
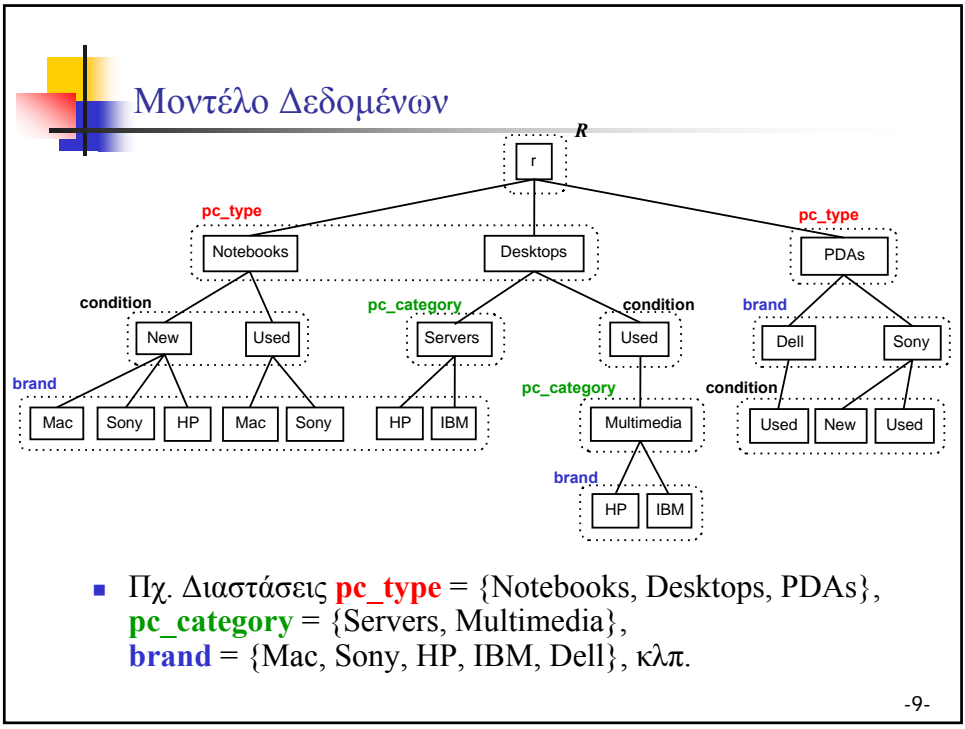
-7-

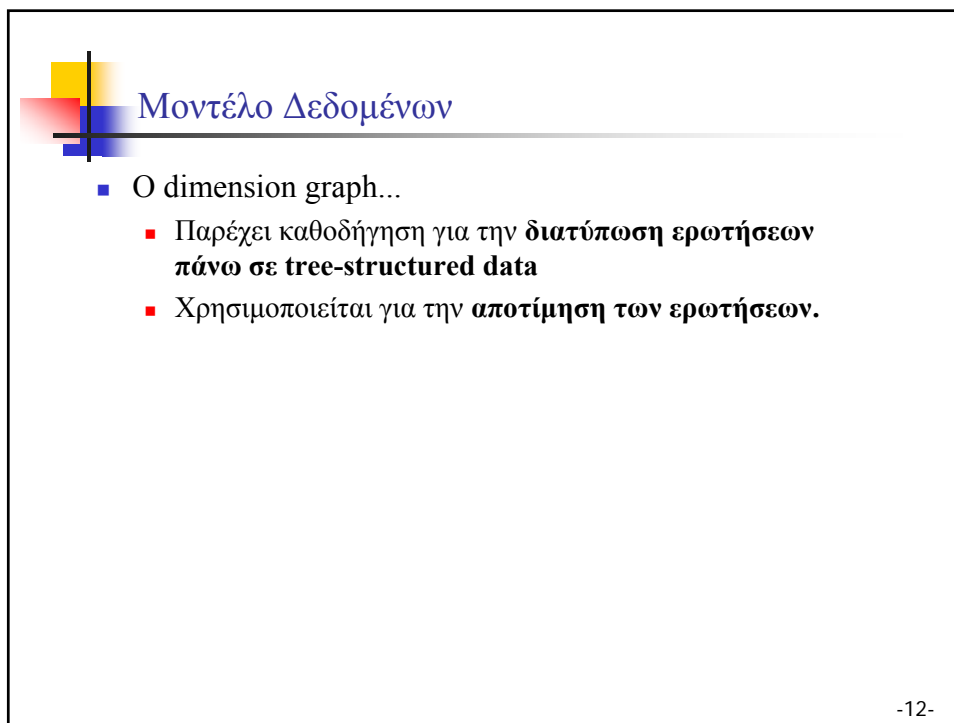
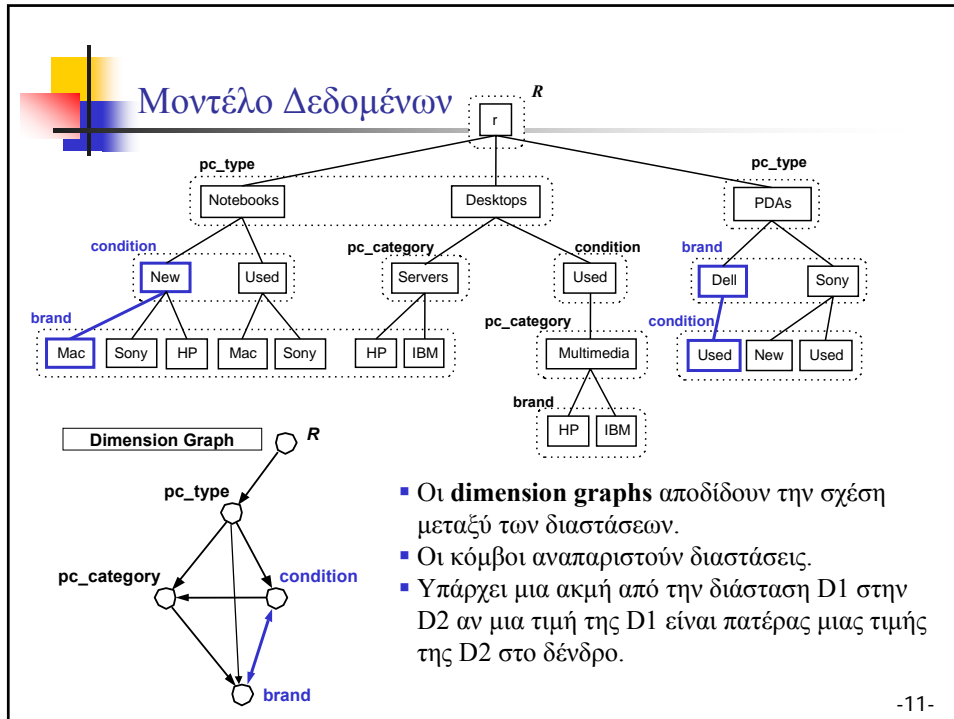


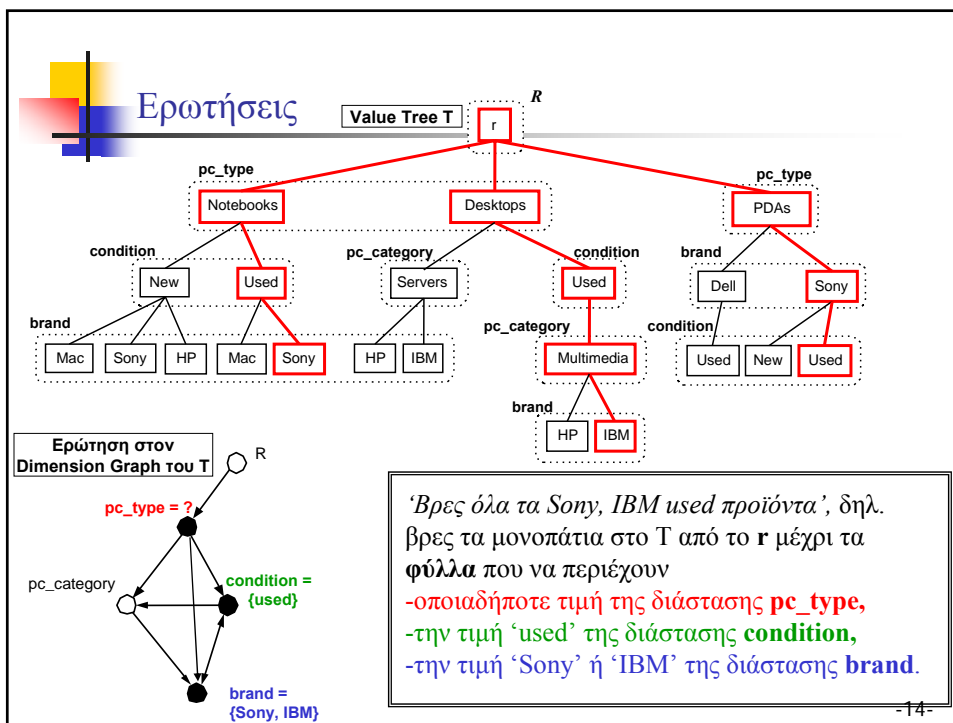
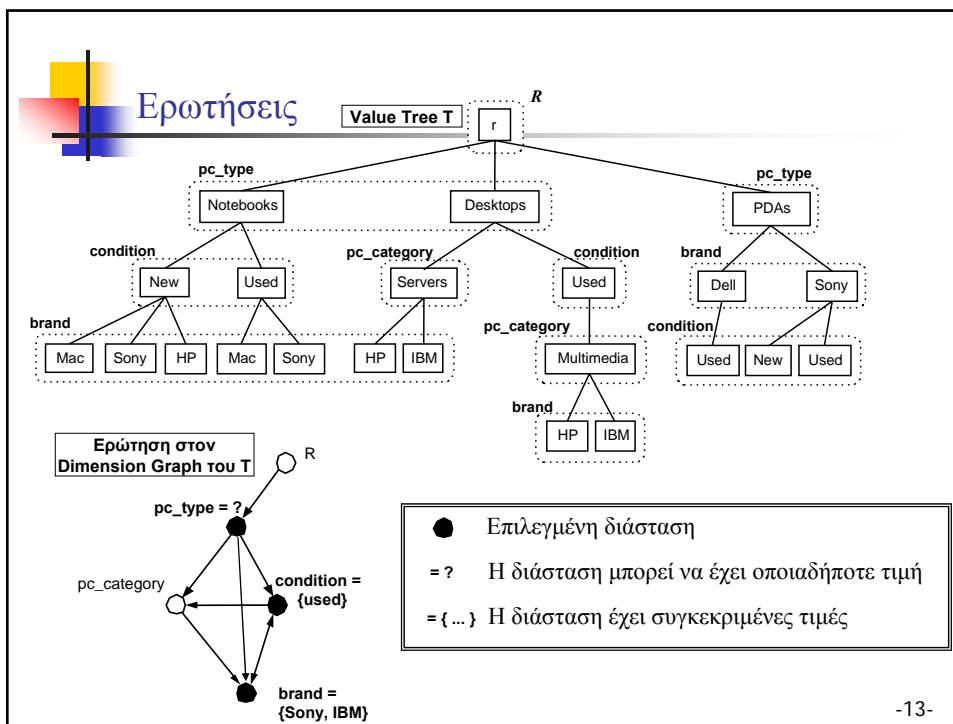
## Μοντέλο Δεδομένων

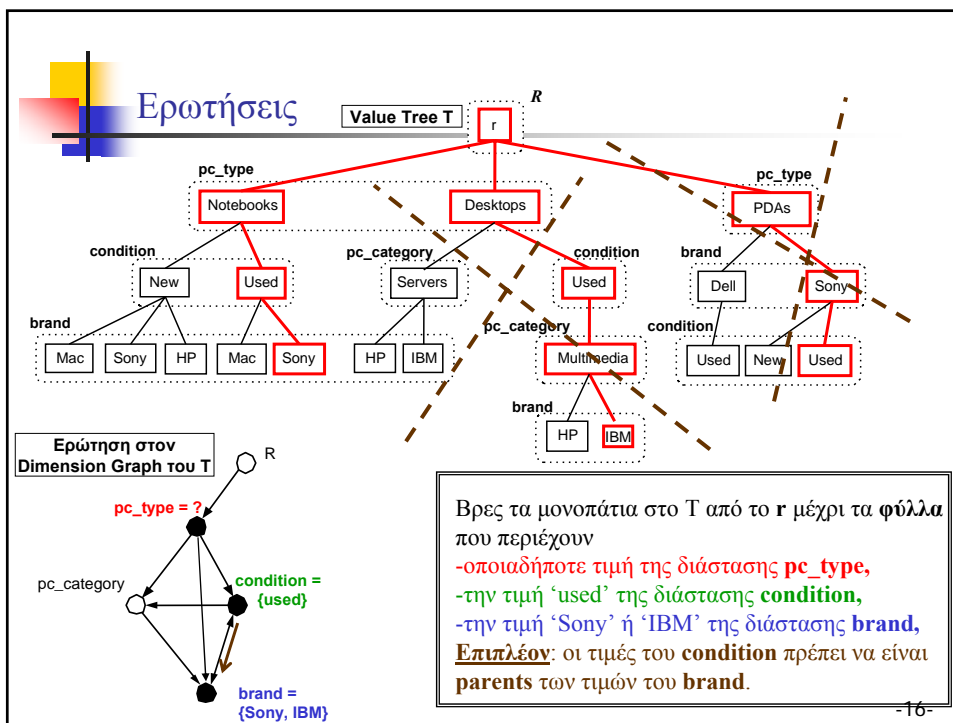
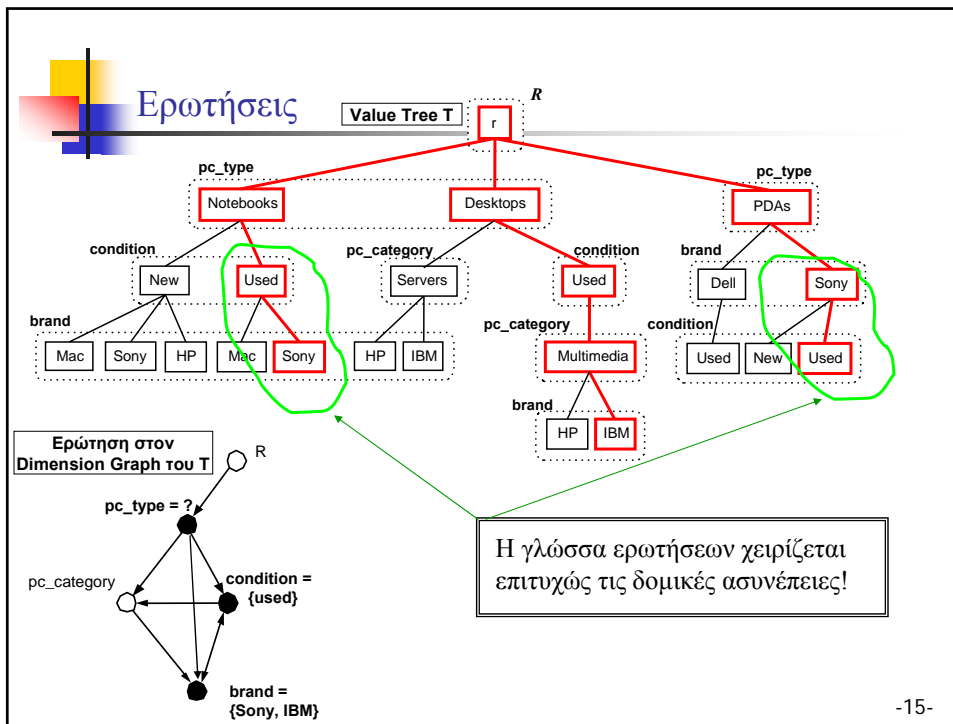
- Οι τιμές (δηλ. οι κόμβοι) στα δένδρα ομαδοποιούνται σε **διαστάσεις (dimensions)**.
- Μια διάσταση...
  - ...είναι ένα σύνολο από σημασιολογικά σχετιζόμενους κόμβους (δηλ. τιμές) του δένδρου.

-8-





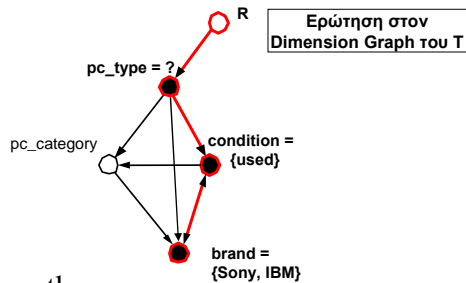






## Αποτίμηση Ερωτήσεων

- Για την αποτίμηση των ερωτήσεων χρησιμοποιούμε τον dimension graph για να βρούμε **answer paths**.
  - Ένα answer path είναι ένα **απλό μονοπάτι** στον dimension graph που περιέχει **όλες** τις επιλεγμένες διαστάσεις.

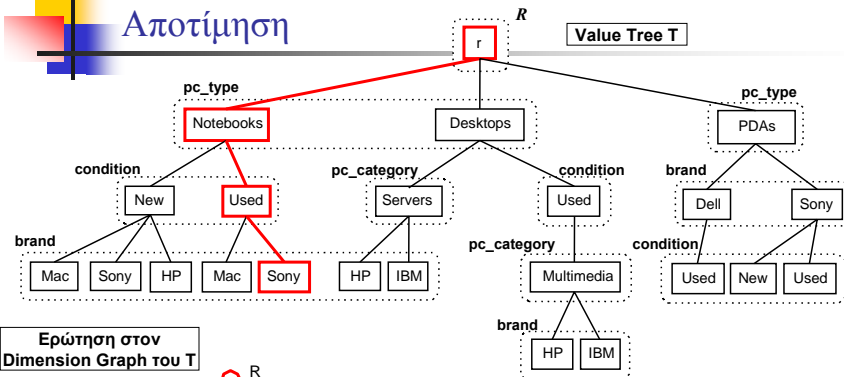


- Παραδείγματα answer paths:
  - `/R/pc_type/condition/brand,`
  - `/R/pc_type/pc_category/brand/condition, ....`

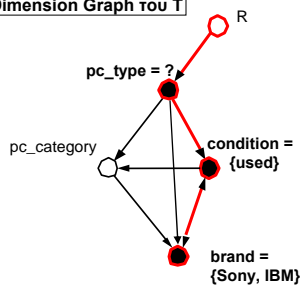
-17-



## Αποτίμηση



Ερώτηση στον Dimension Graph του T

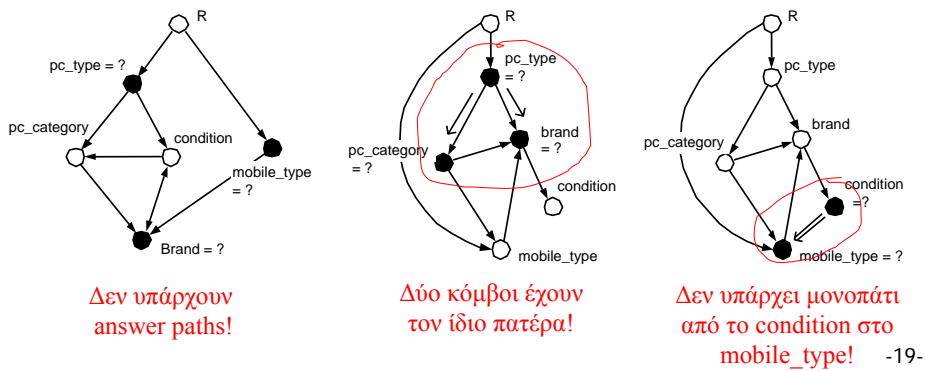


Τα answer paths χρησιμοποιούνται για την παραγωγή ερωτήσεων μονοπατιών που θα εφαρμοσθούν πχ. από ένα XQuery engine για να πάρουμε τις απαντήσεις από ένα δένδρο.  
 Πχ. `/R/pc_type/condition/brand` μας δίνει `/r/(Notebooks|Desktops|PDAs)/Used/(Sony|IBM)`

-18-

## Αποτίμηση Ερωτήσεων

- Χρησιμοποιούμε τους dimension graphs για να ανιχνεύσουμε **μη ικανοποιήσιμες ερωτήσεις**, δηλ. ερωτήσεις με κενή απάντηση σε κάθε δένδρο (**unsatisfiable queries**).
- Παραδείγματα μη ικανοποιήσιμων ερωτήσεων:



## Ερωτήσεις σε Πολλαπλές Πηγές

- Με τους dimension graphs μπορούμε να εφαρμόσουμε την ίδια ερώτηση σε πολλαπλές πηγές (data integration).
  - Έστω τα δένδρα  $T_1, T_2, \dots, T_n$  με ένα σύνολο διαστάσεων  $D$ .
  - Έστω  $G_1, G_2, \dots, G_n$  οι dimension graphs αυτών.
  - Κατασκευάζουμε έναν **global dimension graph G** συγχωνεύοντας τους  $G_1, G_2, \dots, G_n$ .
  - Κατασκευάζουμε τις ερωτήσεις μας στον  $G$ .
  - Οι επιλογές μεταφέρονται στους  $G_1, G_2, \dots, G_n$ .
  - Η αποτίμηση γίνεται όπως περιγράφηκε παραπάνω.



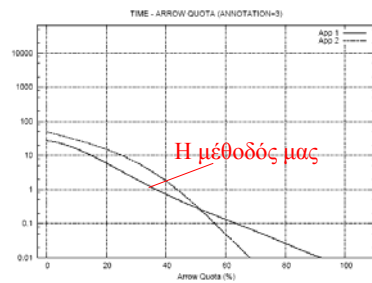
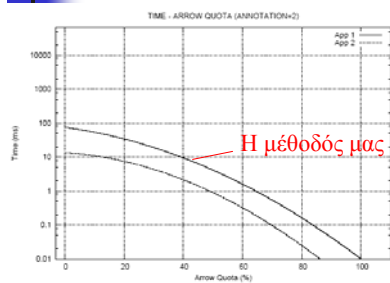
## Πειραματικά Αποτελέσματα

- Χρησιμοποιήσανε συνθετικά δένδρα κωδικοποιημένα ως XML αρχεία.
- Τυχαία παραγωγή ερωτήσεων.
- Συγκρίναμε την μεθοδό μας με μια παρόμοια μέθοδο που δεν χρησιμοποιεί dimension graphs

-21-



## Πειραματικά Αποτελέσματα

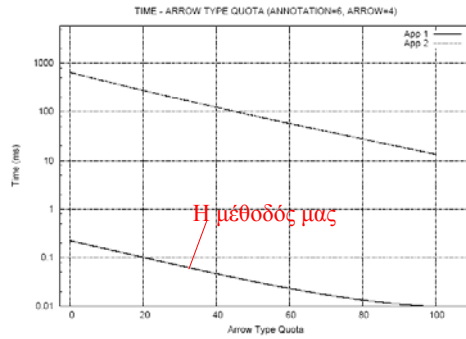


- Χρόνος εκτέλεσης σε σχέση με το ποσοστό των arrows για διάφορες τιμές επιλεγμένων διαστάσεων στην ερώτηση.

-22-



## Πειραματικά Αποτελέσματα

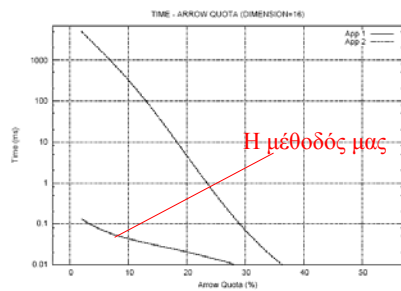
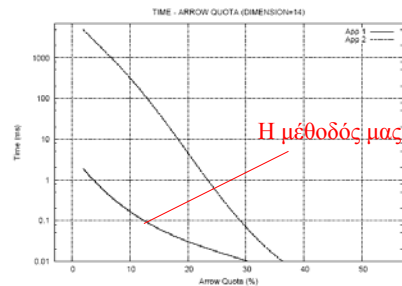
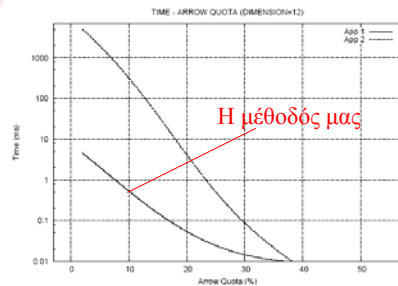


- Χρόνος εκτέλεσης σε σχέση με το ποσοστό των single arrows στην ερώτηση.

-23-



## Πειραματικά Αποτελέσματα



- Χρόνος εκτέλεσης για ερωτήσεις, καθώς μεγαλώνουμε τον dimension graph.

-24-



## Σύνοψη - Επίλογος

- Γλώσσα ερωτήσεων σε tree-structured data με την χρήση dimension graphs:
  - Η γλώσσα ερωτήσεων χειρίζεται αποτελεσματικά τις δομικές διαφορές και ασυνέπειες.
  - Οι dimension graphs αποτυπώνουν την σημασιολογική πληροφορία των tree-structured data.
  - Χρησιμοποιούνται για τον διατύπωση των ερωτήσεων και για την αποτίμηση τους.
  - Οι dimension graphs μπορούν να χρησιμοποιηθούν και για να εφαρμόσουμε ερωτήσεις σε πολλαπλές πηγές.
  - Η μέθοδός μας κερδίζει σε χρόνο μερικές τάξεις μεγέθους σε σχέση με μια μέθοδο που δεν χρησιμοποιεί dimension graphs.
  - Ερωτήσεις.....