4ο πανελλήνιο συνέδριο
ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

13+14+15
Σεπτεμβρίου/**07**

Πρόγραμμα Συνεδρίου

Παρασκευή (απόγευμα): 16:00 - 20:30

Ανάλυση Δεδομένων και Πληροφορική

Πρόεδρος: Σ. Αναστασιάδου

Αίθουσα Β

18:30 - 20:30

Ανακοινώσεις

Ανδρεάδης Γιάννης, Χατζηπαντελής Θόδωρος:	*Η χρήση της R στην Ανάλυση Πολυμεταβλητών Δεδομένων Κοινωνικών Επιστημών.*
Καράκος Αλέξανδρος, Σταθάκη Μαρία:	*Υλοποίηση Διαδικτυακής υπηρεσίας - Ανάλυση Δεδομένων.*
Κουτσουπιάς Νίκος:	*Ταξινόμηση Δεδομένων Ροής Επιλογών Ιστοσελίδων.*
Μανιτσάρης Αθανάσιος, Μαυρίδης Ιωάννης, Μοσχίδης Οδυσσέας:	*Αξιολόγηση ευχρηστίας γραφικών διεπιφανειών διαδραστικών συστημάτων.*
Μάρκος Άγγελος, Μενεξές Γεώργιος, Παπαδημητρίου Γιάννης:	*Το Λογισμικό CHIC Analysis v1.0.*
Ζάμπογλου Μάρκος, Παπαδημητρίου Θεόφιλος, Παπαδημητρίου Γιάννης:	*Ιεραρχική Ταξινόμηση Χρονοσειρών Βάσει του Χρώματος.*

21:30 Δείπνο

4ο Πανελλήνιο Συνέδριο Ανάλυσης Δεδομένων

Ταξινόμηση Δεδομένων Ροής Επιλογών Ιστοσελίδων

Νίκος Κουτσουπιάς,
Πανεπιστήμιο Δυτικής Μακεδονίας

Περιεχόμενα

Μέθοδος - Δεδομένα

- Εφαρμογή με την υλοποίηση S-Pro v2.0
- Χρησιμοποιήθηκαν 7000 εγγραφές «ροής επιλογών» (click streams) www.msnbc.com
- Μιας ημερας (28.9.1998)
- Οι εγγραφές έγιναν με χρονική σειρά

Μέθοδος - Δεδομένα

ΔΕΔΟΜΕΝΑ & ΚΩΔΙΚΟΠΟΙΗΣΗ

- Κατηγορίες - Κωδικοί

Code	Category	New Code	New Category
1	Frontpage	2	FP
2	News	6	NW
3	Tech	10	TE
4	Local	4	LO
5	Opinion	8	OP
6	On-air	7	OA
7	Misc	5	MI
8	Weather	4	LO
9	Health	3	LI
10	Living	3	LI
11	Business	1	BU
12	Sports	9	SP
13	Summary	2	NW
14	bbs	2	NW
15	Travel	3	LI
16	msnnews	2	NW
17	msnsports.	9	SP

Μέθοδος - Δεδομένα

ΔΕΔΟΜΕΝΑ & ΚΩΔΙΚΟΠΟΙΗΣΗ

- **Χρησιμοποιηθήκαν οι 10 πρώτες επιλογές**
- **Κάθε γραμμή ροής επιλογών αντιπροσωπεύει ένα χρήστη**

Visitor Count	ClickStream
1	1 1
2	2
3	3 2 2 4 2 2 2 3
4	3
5	5
6	1
7	6
8	1 1
9	6
10	6 7 7 7 6 6 8
:	:
:	:

Μέθοδος - Δεδομένα

> Δεδομένα –1η φάση προ-επεξεργασίας

- Ο πίνακας ροης επιλογών μετασχήματιζεται σε πινακα διαστασεων 12X7000

- Vn = {Fn, Ln, BUn, FPn, LIn, LOn, MIn, NWn, OAn, OPn, SPn, TEn}

με:

n =1 … 7000,

Fn, Ln οι κατηγορίες της πρώτης (entry) και της τελευταίας (exit) σελίδας επίσκεψης της ροής n, και

BUn, ως TEn το πλήθος επισκέψεων ανα κατηγορία κατά τη διάρκεια της ίδιας ροής επιλογών (stream n).

Μέθοδος - Δεδομένα

> Δεδομένα –2η φάση προ-επεξεργασίας

- Επιπλέον κατηγοριοποίηση του πλήθους επισκέψεων σε τρεις κλάσεις:

- Υψηλές τιμών αντιστοιχούν σε πλήθος επισκέψεων με >=3 επιλογές (clicks) στην ίδια ροή,
- Μέσες τιμές αντιστοιχούν σε 1-2 επιλογές
- Χαμηλές σε καθόλου (0) επιλογές

Μέθοδος - Δεδομένα

> Ανάλυση – CAH

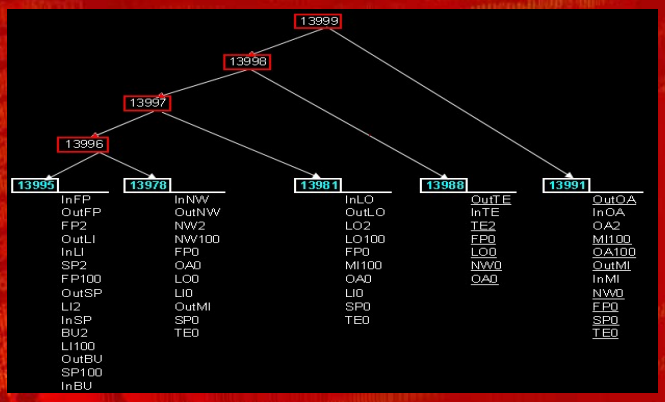
- Επιλέξαμε να ερευνήσουμε τις πέντε πρώτες ομάδες (λ=0,22)

Κόμβος	A	B	Βάρος	Αποστ(Δ)	Ιδ(Εσω)	Ιδ(Εξω)	λ(r)
:	:	:	:	:	:	:	:
13992	13987	13989	0,060857	0,131561	2,030647	1,136019	0,358743
13993	13992	13982	0,152429	0,138071	2,168718	0,997949	0,315142
13994	**13984**	**13993**	**0,372857**	**0,141161**	**2,309879**	**0,856788**	**0,270565**
13995	**13994**	**13985**	**0,455571**	**0,159129**	**2,469008**	**0,697659**	**0,220313**
13996	**13995**	**13978**	**0,604857**	**0,163374**	**2,632382**	**0,534285**	**0,168722**
13997	**13996**	**13981**	**0,725857**	**0,167672**	**2,800054**	**0,366613**	**0,115773**
13998	**13997**	**13988**	**0,815571**	**0,177343**	**2,977396**	**0,18927**	**0,05977**
13999	**13998**	**13991**	**1**	**0,18927**	**3,166667**	**1,36E-14**	**4,29E-15**

Μέθοδος - Δεδομένα

> Ανάλυση – CAH

- Προέκυψε το παρακάτω δενδρόγραμμα



Συμπερασματα

> Η CAH μπορεί να **αξιοποιηθεί** για την ανάλυση δεδομένων **ροών επιλογής**

> Η μεθοδολογία και ο τρόπος επεξεργασίας των διαθέσιμων δεδομένων είναι δυνατό να αποτελέσουν εργαλεία δουλειάς για **ερευνητές, διαχειριστές ιστοτόπων** και «**μαρκετίστες**» του διαδικτύου

> Με την έλευση του **σημασιολογικού ιστού** (semantic Web) οι μέθοδοι της ανάλυσης δεδομένων πρόκειται να αποτελέσουν εργαλεία άμεσης διαχείρισης και για τους portal editors


4ο Πανελλήνιο Συνέδριο Ανάλυσης Δεδομένων
Ταξινόμηση Δεδομένων Ροής Επιλογών Ιστοσελίδων
Νίκος Κουτσουπιάς,
Πανεπιστήμιο Δυτικής Μακεδονίας

Ταξινόμηση Δεδομένων Ροής Επιλογών Ιστοσελίδων

Νίκος Κουτσουπιάς
Πανεπιστήμιο Δυτ. Μακεδονίας

Η διερεύνηση της συμπεριφοράς επισκεπτών εμπορικών και μη δικτυακών προορισμών αποτελεί ένα από τα πιο σημαντικά αντικείμενα μελέτης της επιστήμης εξόρισης δικτυακών δεδομένων. Στην παρούσα εργασία προτείνεται η χρήση της Ιεραρχικής Ταξινόμησης για τη αναζήτηση ομάδων και σχηματισμών σε δεδομένα που προέρχονται από «ροή επιλογής ιστοσελίδων» (click-stream). Χρησιμοποιώντας δέκα χιλιάδες επιλογές επισκεπτών μιας δικτυακής πύλης ενημέρωσης, εξετάζουμε τις συνήθειες τους σε δέκα κατηγορίες ειδήσεων με σκοπό την καλύτερη κατανόηση και εξυπηρέτηση των αναγκών δικτυακών χρηστών και εφαρμογών.

Click Stream Cluster Analysis

Nikos Koutsoupias
University of W. Macedomia

Web portal visitor behavior is an important research field of web data mining. In this work Cluster Analysis is proposed as a tool for the discovery of groupings inside data based on web click-streams. Using a data set of ten thousand user traces on pages classified in ten news categories we aimed in investigating user visiting patterns in order to understand and better serve the needs both of internet users and web-based applications.