

Πανεπιστήμιο Πειραιώς Τμήμα Πληροφορικής



Αποθήκες Δεδομένων και Εξόρυξη Γνώσης (Data Warehousing & Data Mining)

Εξόρυξη Χρονικής Γνώσης (temporal data mining)

Γιάννης Θεοδωρίδης, Νίκος Πελέκης

Ομάδα Διαχείρισης Δεδομένων
Εργαστήριο Πληροφοριακών Συστημάτων

<http://isl.cs.unipi.gr/db>

Περιεχόμενα



- Χρονολογικές σειρές (time series)
 - Ομοιότητα
 - Πρόβλεψη (πρόγνωση) μελλοντικών τιμών
- Χρονική ανάλυση καλαθιού αγορών
 - Ανακάλυψη ακολουθιακών προτύπων (sequential patterns)
 - Κανόνες συσχέτισης ακολουθιών (sequence association rules)
 - Εξαρτήσεις τάσεων
 - Επεισόδια (episodes)

Περιεχόμενα



- Χρονολογικές σειρές (time series)
 - Ομοιότητα
 - Πρόβλεψη (πρόγνωση) μελλοντικών τιμών
- Χρονική ανάλυση καλαθιού αγορών
 - Ανακάλυψη ακολουθιακών προτύπων (sequential patterns)
 - Κανόνες συσχέτισης ακολουθιών (sequence association rules)
 - Εξαρτήσεις τάσεων
 - Επεισόδια (episodes)

3

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Χρονολογικές σειρές (time series)



- Σύνολο τιμών χαρακτηριστικών στον άξονα του χρόνου
 - $\{ \langle t_1, a_1 \rangle, \langle t_2, a_2 \rangle, \dots, \langle t_n, a_n \rangle \}$
- Ανάλυση χρονολογικών σειρών: η ανακάλυψη προτύπων στις τιμές.
- Είδη προτύπων:
 - Τάση : συστηματική μη επαναλαμβανόμενη αλλαγή στο χρόνο (π.χ. σταθερή αύξηση)
 - Κύκλος : επαναλαμβανόμενη συμπεριφορά
 - Εποχιακό πρότυπο : παρόμοιες τιμές σε συγκεκριμένες περιόδους – μέρα, μήνα, έτος κλπ. - του χρόνου
 - Μη φυσιολογικά δεδομένα (outliers) : στοιχεία που δεν μπορούν να κατηγοριοποιηθούν

4

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Τεχνικές Ανάλυσης Χρονολογικών Σειρών

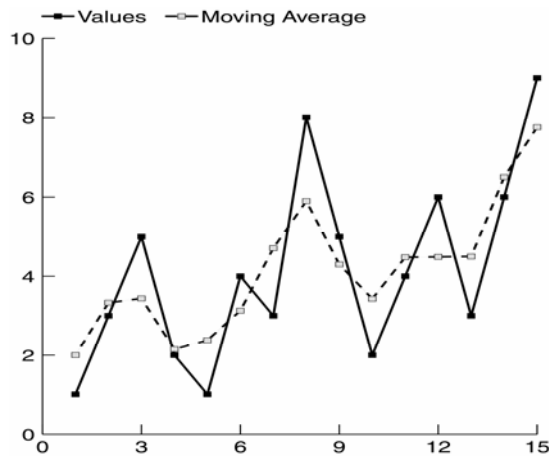


Εξομάλυνση (smoothing)

- Μια δημοφιλή τεχνική: **κινούμενος μέσος όρος** των τιμών (moving average)

Εφαρμογές:

- Ανακάλυψη τάσεων
- Εντοπισμός μη φυσιολογικών συμπεριφορών (outliers)



5

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

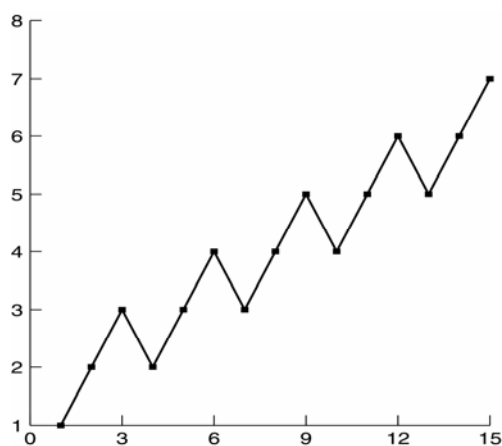
Τεχνικές Ανάλυσης Χρονολογικών Σειρών



Αυτο-συσχέτιση (autocorrelation) :

συσχέτιση μεταξύ διαφορετικών υποακολουθιών της χρονολογικής σειράς

- Για ανακάλυψη κύκλων ή εποχιακών (π.χ. ετήσιων) προτύπων
- Χρονική υστέρηση (lag)** Χρόνος μεταξύ συσχετισμένων υποσειρών.



6

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Ομοιότητα (Similarity)



- $\text{sim}(X, Y)$: ομοιότητα μεταξύ δύο χρονολογικών σειρών, $X = \langle x_1, x_2, \dots, x_n \rangle$ και $Y = \langle y_1, y_2, \dots, y_m \rangle$
- Ζητήματα:
 - Διαφορά μήκους (στη γενική περίπτωση, $n \neq m$)
 - Κλίμακα (π.χ. Fahrenheit vs. Celsius)
 - Παρουσία outliers (που δεν μπορούν να κατηγοριοποιηθούν)
- Παράδειγμα ομοιότητας: longest common subsequence – LCS
 - Εύρεση της μεγαλύτερης (δηλαδή, με το μεγαλύτερο μήκος) κοινής υποακολουθίας.

$$X = \langle 10, 5, 6, 9, 22, 15, 4, 2 \rangle$$

$$Y = \langle 6, 9, 10, 5, 6, 22, 15, 4, 2, 8 \rangle$$

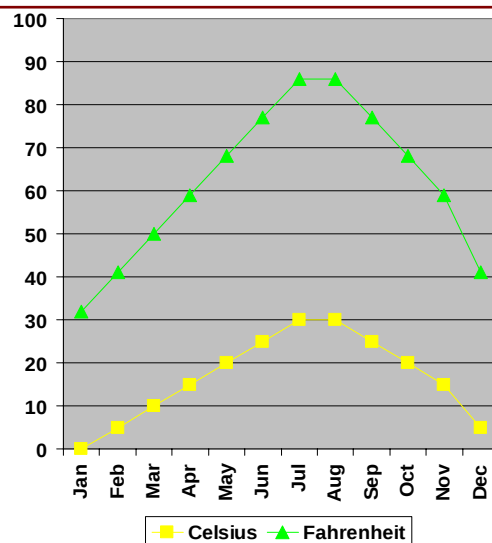
$$\text{output} = \langle 22, 15, 4, 2 \rangle$$

$$\text{sim}(X, Y) = \text{length}(\text{output}) / \max(\text{length}(X), \text{length}(Y)) = 4/10$$

7

ΠΑ.ΠΕΙ. – Γιάννης Θεοδοριάδης

Παράδειγμα ομοιότητας (;)



Παράδειγμα:
μέση μηνιαία
θερμοκρασία σε 2
πόλεις

Κι όμως, οι 2 σειρές
είναι πρακτικά
ταυτόσημες!

$$F = C * 1.8 + 32$$

8

ΠΑ.ΠΕΙ. – Γιάννης Θεοδοριάδης

Ομοιότητα βασισμένη σε γραμμικό μετασχηματισμό



- Ψάχνουμε για μια συνάρτηση γραμμικού μετασχηματισμού f
 - που να μετατρέπει μια τιμή της πρώτης σειράς (κοντά) σε μια τιμή της δεύτερης σειράς
- ϵ_f – Ανεκτή διαφορά στα αποτελέσματα
- δ – Επιτρεπόμενη διαφορά χρονικής τιμής

Given integer value $\delta > 0$, real number $\epsilon < 1$, and linear function f , and two time series X, Y with the longest one of length n . Let $X' = \langle x_{i_1}, x_{i_2}, \dots, x_{i_m} \rangle$ and $Y' = \langle y_{j_1}, y_{j_2}, \dots, y_{j_m} \rangle$ be the longest subseries in X and Y respectively where:

- $\forall 1 \leq k \leq m-1, |i_k - j_k| \leq \delta$ and
- $\forall 1 \leq k \leq m, \frac{y_{j_k}}{(1+\epsilon)} \leq f(x_{i_k}) \leq y_{j_k}(1+\epsilon)$.

Then $\text{sim}_{\epsilon, \delta}(X, Y) = \max_f(m/n)$.

9

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Πρόβλεψη



- Πρόβλεψη (ή πρόγνωση) μελλοντικών τιμών μιας χρονικής σειράς
- Η παλινδρόμηση (regression) μπορεί να μην είναι αρκετή
- Στατιστικές Τεχνικές
 - βασισμένες στη μέθοδο της αυτο-παλινδρόμησης (autoregression)
 - παραδοχή: η νέα τιμή εξαρτάται από τις προηγούμενες

Given a time series $X = \langle x_1, x_2, \dots, x_n \rangle$, a future value, x_{n+1} , can be found using

$$x_{n+1} = \xi + \phi_n x_n + \phi_{n-1} x_{n-1} + \dots + \phi_1 x_1 + \epsilon_{n+1}$$

where ϕ_i are the autoregressive parameters and ϵ_{n+1} represents a random error

10

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Περιεχόμενα



- Χρονολογικές σειρές (time series)
 - Ομοιότητα
 - Πρόβλεψη (πρόγνωση) μελλοντικών τιμών
- Χρονική ανάλυση καλαθιού αγορών
 - Ανακάλυψη ακολουθιακών προτύπων (sequential patterns)
 - Κανόνες συσχέτισης ακολουθιών (sequence association rules)
 - Εξαρτήσεις τάσεων
 - Επεισόδια (episodes)

11

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Ακολουθίες



- **Ακολουθία (sequence)** είναι μια διατεταγμένη λίστα από στοιχειοσύνολα (itemsets)
 - Παρόμοια με τη χρονολογική σειρά, με τη μόνη διαφορά ότι μπορεί να περιέχει itemsets αντί τιμών
- **Υπακολουθία (subsequence)** Τ μιας ακολουθίας S είναι μια ακολουθία τέτοια ώστε για κάθε itemset που ανήκει στην T υπάρχει αντίστοιχο υπερσύνολο itemset που ανήκει στην S, στην ίδια διάταξη. Παράδειγμα:
 - $S = \langle \{ABC\}, \{BDE\}, \{CD\}, \{ACE\} \rangle$
 - $T = \langle \{BD\}, \{CE\} \rangle$

12

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Το πρόβλημα της εύρεσης κανόνων συσχέτισης ακολουθιών



Let $I = \{I_1, I_2, \dots, I_m\}$ be a set of items. A sequence, S , is : $S = \langle s_1, s_2, \dots, s_n \rangle$ where $s_i \subseteq I$.

Let $I = \{I_1, I_2, \dots, I_m\}$ be a set of items. One sequence $T = \langle t_1, \dots, t_m \rangle$ is a subsequence of another $S = \langle s_1, \dots, s_n \rangle$ if $\forall 1 \leq j \leq m-1, i_j < i_{j+1}$ and $\forall 1 \leq j \leq m, \exists 1 \leq k \leq n$ such that $t_{i_j} \subseteq s_k$. In this case, S contains T .

Given a set of customers and transactions for each customer, the support of a sequence $s(S)$ is the percentage of total customers whose customer-sequence contains S .

The confidence (α) for a sequence association rule $S \Rightarrow T$, is the ratio of the number of customers (customer-sequences) which contains both sequences S and T to the number that contain S .

13

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Αναλογίες με το 'παραδοσιακό' πρόβλημα



παραδοσιακή ανάλυση

- Η βάση δεδομένων: σύνολο από «καλάθια αγορών»
- Το εργαλείο: συχνά itemsets
- Τα πρότυπα που προκύπτουν:
 - κανόνες συσχέτισης (μεταξύ itemsets)

χρονική ανάλυση

- Η βάση δεδομένων: ακολουθία από «καλάθια αγορών»
- Το εργαλείο: συχνές ακολουθίες από itemsets
- Τα πρότυπα που προκύπτουν:
 - κανόνες συσχέτισης ακολουθιών (από itemsets) και πολλά άλλα ...
 - εξαρτήσεις τάσεων
 - επεισόδια ...

14

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Παράδειγμα συχνής ακολουθίας



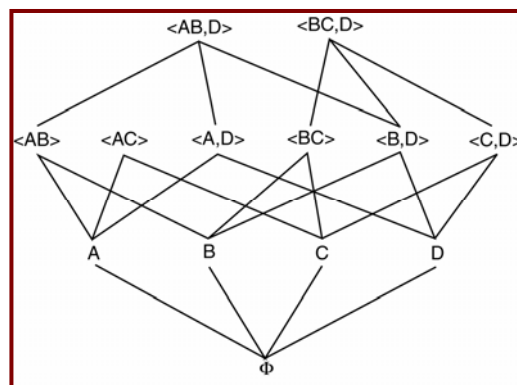
- Κατ' αναλογία με τα συχνά itemsets
- Αγορές που έγιναν από 3 πελάτες C_1, C_2, C_3
 - $s(\langle \{A\} \rangle) = 3/3$
 - $s(\langle \{A\}, \{C\} \rangle) = 1/3$
 - $s(\langle \{A\}, \{D\} \rangle) = 2/3$
 - $s(\langle \{BC\}, \{D\} \rangle) = 2/3$
 - $s(\langle \{AB\}, \{C\}, \{D\} \rangle) = 1/3$

Customer	Time	Itemset
C_1	10	AB
C_1	20	BC
C_1	30	D
C_2	15	ABC
C_2	20	D
C_3	15	ACD

15

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Δικτυωτό πλέγμα (lattice) συχνών ακολουθιών



Εμφανίζονται μόνο οι συχνές ακολουθίες του προηγούμενου παραδείγματος με $\text{support} \geq 2/3$

- **Ιδιότητα συχνών ακολουθιών (frequent sequence property):**
κάθε υπακολουθία μιας συχνής ακολουθίας είναι συχνή !

16

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

SPADE



- Sequential Pattern Discovery using Equivalence classes (Ανακάλυψη ακολουθιακών προτύπων με χρήση κλάσεων ισοδυναμίας)
- Αναγνωρίζει πρότυπα διασχίζοντας το lattice από πάνω προς τα κάτω.
- Διαιρεί το lattice σε κλάσεις ισοδυναμίας και κάνει αναζήτηση ξεχωριστά σε καθεμία.
- **ID-List:** Συσχετίζει πελάτες και συναλλαγές με κάθε στοιχείο.

17

ΠΑ.ΠΕΙ. – Γιάννης Θεοδοριάδης

Παράδειγμα SPADE



- ID-List για ακολουθίες με μήκος 1:

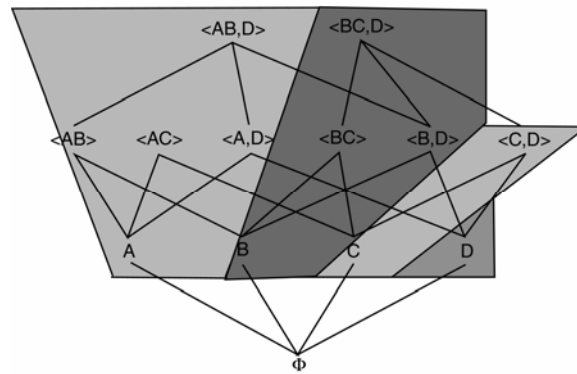
A		B		C		D	
Customer	Time	Customer	Time	Customer	Time	Customer	Time
C_1	10	C_1	10	C_1	20	C_1	30
C_2	15	C_1	20	C_2	15	C_2	20
C_3	15	C_2	15	C_3	15	C_3	15

- Το πλήθος εμφάνισης της ακολουθίας $\langle \{A\} \rangle$ είναι 3
- Το πλήθος εμφάνισης της ακολουθίας $\langle \{A\}, \{D\} \rangle$ είναι 2

18

ΠΑ.ΠΕΙ. – Γιάννης Θεοδοριάδης

Θ_1 Κλάσεις ισοδυναμίας



19

ΠΑ.ΠΕΛ – Γιάννης Θεοδοριάδης

Ο Αλγόριθμος SPADE



Input:

D //ID-Lists for customer transactions.
 s //Support.

Output:

F // Frequent sequences.

SPADE Algorithm

Determine frequent items, F_1 ;
 Determine frequent 2-sequences, F_2 ;
 Find equivalence classes ϵ for all 1-sequences $[S]_{\Theta_1}$;
 for each $[S] \in \epsilon$ do
 Find frequent sequences F ;

20

ΠΑ.ΠΕΛ – Γιάννης Θεοδοριάδης

Κανόνες συσχέτισης ακολουθιών



- **Sequence association rules:**

Κανόνες συσχέτισης που περιέχουν ακολουθίες

- Παραδείγματα:

$\langle \{A\}, \{C\} \rangle \Rightarrow \langle \{A\}, \{D\} \rangle$; $s = 1/3$; $c = 1$

$\langle \{BC\}, \{D\} \rangle \Rightarrow \langle \{A\}, \{C\} \rangle$; $s = 1/3$; $c = 1/2$

- Υλοποίηση σε 2 βήματα:

1. Εύρεση συχνών ακολουθιών, δηλαδή αυτών που ικανοποιούν το minimum support (π.χ. με τον αλγόριθμο SPADE)
2. Συνδυάζοντας τις συχνές ακολουθίες, καταγραφή πιθανών κανόνων και εύρεση αυτών που ικανοποιούν το minimum confidence

Customer	Time	Itemset
C_1	10	AB
C_1	20	BC
C_1	30	D
C_2	15	ABC
C_2	20	D
C_3	15	ACD

21

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Εξαρτήσεις Τάσεων



- **Εξάρτηση τάσης (trend dependency):** κανόνας συσχέτισης $X \Rightarrow Y$ μεταξύ δύο χρονικών καταστάσεων I_1, I_2 , μιας βάσης δεδομένων, όπου X και Y είναι πρότυπα της μορφής (χαρακτηριστικό, τελεστής)

- Παράδειγμα:

- Στην πορεία του χρόνου, ο μισθός ενός υπαλλήλου ακολουθεί αυξητική τάση
- $(SSN, =) \Rightarrow (Salary, \leq)$

- **Support:** το ποσοστό των πλειάδων του $I_1 \times I_2$ που ικανοποιούν και τα δύο πρότυπα, X και Y

- **Confidence:** η αναλογία των πλειάδων του $I_1 \times I_2$ που ικανοποιούν και τα δύο πρότυπα, X και Y , προς τις πλειάδες του $I_1 \times I_2$ που ικανοποιούν το πρότυπο X

22

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Παράδειγμα εξάρτησης τάσης



- $(SSN, =) \Rightarrow (Salary, \leq)$
- Given two tuples $t_1 \in I_1$ and $t_2 \in I_2$, if $t_1(SSN) = t_2(SSN)$ then $t_1(Salary) \leq t_2(Salary)$
- Support=4/36; Confidence=4/5

Name	SSN	Address	Salary
Joe Smith	123456789	10 Moss Haven	50,000
Mary Jones	111111111	10 Main	75,000
Bill Adams	222222222	215 North	100,000
Selena Shepherd	876543298	25 Georgetown	15,000
Paul Williams	908734124	13 East	250,000
Martha Laros	873659865	1010 Fox	150,000

I_1

Name	SSN	Address	Salary
Mary Jones	111111111	10 Main	85,000
Joe Smith	123456789	10 Moss Haven	52,000
Bill Adams	222222222	215 North	90,000
Selena Shepherd	876543298	25 Georgetown	15,000
Paul Williams	908734124	13 East	270,000
Bob Holder	888888888	22 South	20,000

I_2

23

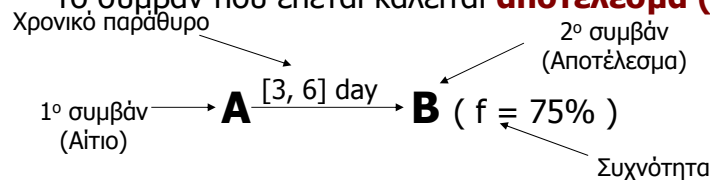
ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Επεισόδια



▪ (απλά) Επεισόδια (episodes):

- Συσχετίσεις μεταξύ δύο **συμβάντων (events)** όπου το ένα συνέβη **μετά** το άλλο μέσα σε ένα **χρονικό παράθυρο (timeframe)** με συγκεκριμένη **συχνότητα**
- το συμβάν που προηγείται καλείται **αίτιο (cause)** και το συμβάν που έπεται καλείται **αποτέλεσμα (effect)**



Στο **75%** των περιπτώσεων που συνέβη το **A**, ακολούθησε το **B** μέσα σε χρονικό διάστημα **3 έως 6 ημερών**

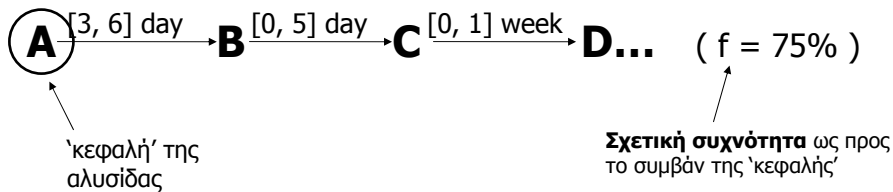
24

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Σύνθετα επεισόδια



- Σύνθετο επεισόδιο: αλυσίδα απλών επεισοδίων
 - αναπαράσταση ως γράφος
- Παράδειγμα σύνθετου επεισοδίου (αλυσίδα απλών)



Στο **75%** των περιπτώσεων που συνέβη το **A**, ακολούθησε το **B** μέσα σε διάστημα **3 έως 6 ημερών**, ακολουθούμενο από το **C** μέσα σε διάστημα **5 ημερών**, ακολουθούμενο από το **D** μέσα σε διάστημα **1 εβδομάδας**, ...

25

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Εφαρμογές



- **Χρονικά εξελισσόμενα δεδομένα:** π.χ. ανάλυση χρηματιστηριακών δεδομένων
 - παρακολούθηση των μεταπτώσεων της αξίας των μετοχών ως συμβάντα στην πορεία του χρόνου
- **Χωρο-χρονικά δεδομένα:** π.χ. ανάλυση σεισμικής δραστηριότητας
 - πέρα από τη χρονική διάσταση (χρόνος εμφάνισης φαινομένου), είναι κρίσιμη και η χωρική διάσταση (επίκεντρο)
- **Δεδομένα συμβολοσειρών:** π.χ. ακολουθίες DNA
 - μια ακολουθία DNA μπορεί να θεωρηθεί ως ακολουθία 'καλαθιών' συμβόλων

26

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης