



Εξόρυξη από τα δεδομένα χρήσης του Web

Usage mining

Περιεχόμενα

- Εισαγωγή
- Web usage mining & Web Personalization
 - Συλλογή δεδομένων και προετοιμασία
 - Μοντελοποίηση των Web usage data
 - Ανάλυση των Web Usage Patterns

Εισαγωγή

- Συνεχής αύξηση του μεγέθους του web
- Τεράστιο οικονομικό ενδιαφέρον για την προσέλκυση πελατών και την προώθηση περιεχομένου σε αυτούς
- Στόχοι:
 - Αύξηση του αριθμού των επισκέψεων σε ένα server ώστε να αυξήσουμε τα κέρδη από διαφημίσεις
 - Να πείσουμε τους χρήστες να αγοράσουν προϊόντα και υπηρεσίες
- Ανάγκη να προβλέψουμε τις ανάγκες των χρηστών και να βελτιώσουμε την ευχρηστία του web site και να αυξήσουμε τη διάρκεια παραμονής των χρηστών σε αυτό

Web Mining for Personalization

Πως ελκύουμε πελάτες;

Προσαρμογή...

Στα ιδιαίτερα
χαρακτηριστικά
και προτιμήσεις
κάθε χρήστη

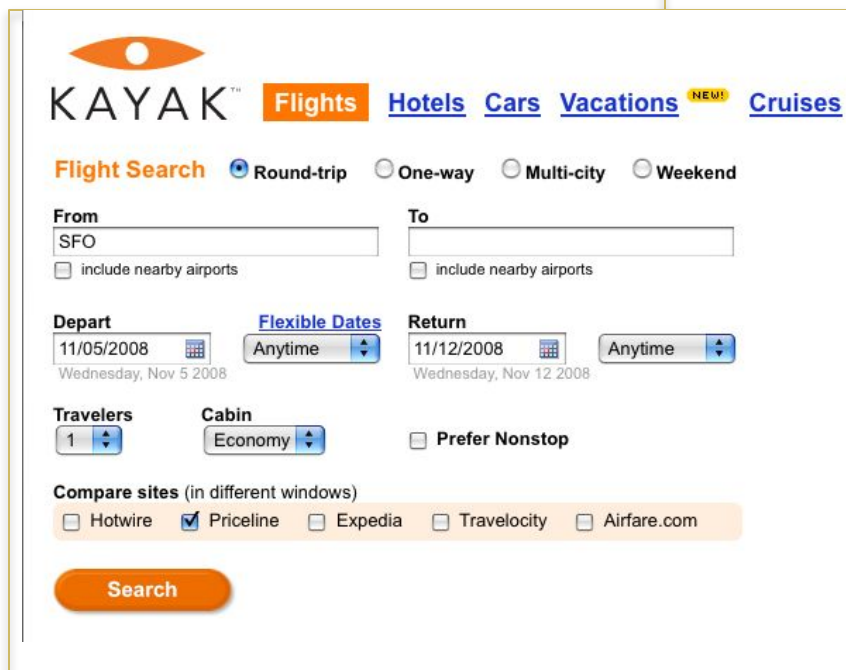
The image shows a screenshot of the Google News homepage. At the top, the Google News logo is visible, along with search bars for "Search News" and "Search the Web". Below the logo, there's a section for "Top Stories" with a dropdown menu set to "U.S.". The main content area displays several news articles, including "Greenspan Concedes to 'Flaw' in His Market Ideology (Update2)" and "Inflated Treasuries And Deflated Stockholders". On the right side, there's a "Personalize this page" sidebar with a "Drag to rearrange page. Click to edit." instruction. This sidebar contains a "Top Stories" section with buttons for various categories like World, U.S., Business, Elections, Sci/Tech, Entertainment, Sports, and Health. Below this, there are buttons to "Add a standard section", "Add a custom section", "Add a local section", and "Add a recommended section". At the bottom of the sidebar, there are links for "Show headlines only" and "Reset page to default".

The image shows a screenshot of the Yahoo! homepage. A "Choose your Interests" dialog box is prominently displayed in the center. The dialog box has a title bar and a close button. Inside, it says "Tell us what you want." and lists various interests with checkboxes. The interests are organized into two columns: "Santa Cruz, CA Content*", "News*", "Shopping*", "Science*", "Fashion*", "Photography*", "Politics/government news*", and "Arts*". There are also "Back" and "Next" buttons at the bottom of the dialog box. The background shows the Yahoo! homepage with a "Web Search" bar, a "Weather" widget for Sunnyvale, CA, and a "Fashion Photos" section.

Πως δημιουργούμε πελάτες;

Ανίχνευση και Προσαρμογή...

- Με βάση το IP, με χρήση cookies, με εγγραφή κλπ.



The screenshot shows the KAYAK website interface. At the top, there's a navigation bar with links for Flights, Hotels, Cars, Vacations, and Cruises. Below this, the 'Flight Search' section is active, showing options for Round-trip, One-way, Multi-city, and Weekend. The search form includes fields for 'From' (SFO) and 'To', with checkboxes for 'include nearby airports'. The 'Depart' date is set to 11/05/2008, and the 'Return' date is 11/12/2008. There are also options for 'Flexible Dates', 'Travelers' (1), 'Cabin' (Economy), and a 'Prefer Nonstop' checkbox. At the bottom, there's a 'Compare sites' section with checkboxes for Hotwire, Priceline, Expedia, Travelocity, and Airfare.com. A large orange 'Search' button is at the bottom.



The screenshot shows the Google Greece homepage. The address bar displays 'http://www.google.com.gr/'. The Google logo is prominently displayed with 'Ελλάς' (Greece) underneath. Below the logo, there are links for 'Ιστός', 'Εικόνες', 'Ομάδες', and 'Κατάλογος'. A search bar is present with a 'Google Αναζήτηση' button and a 'Γνωστικά Εργαλεία' button. There are also links for 'Προχωρημένη Αναζήτηση', 'Προσμήσεις', and 'Γλωσσικά Εργαλεία'. A section for 'στον ιστό' and 'σελίδες γραμμένες στα Ελληνικά' is visible. At the bottom, there's a footer with '©2005 Google - Ψάχνοντας σε 8.058.044.851 σελίδες'.

Τι κερδίζουμε;

Στοχευμένες διαφημίσεις/emails

- ο Με βάση
προηγούμενες
επισκέψεις/αγορές

☆ Amazon.com to me

[show details](#) Nov 6 (7 days ago)

[Reply](#)

To ensure receipt of our e-mails, please add store-news@amazon.com to your address book today.

amazon.com

[E-mail or print for immediate delivery](#)

[Amazon.com Gift Cards](#)

[Shoes](#)

[Your Amazon.com](#)

[Today's Deals](#)

[See All Departments](#)

Dear Amazon.com Customer,

Based on your interest in PUMA, you might like to know that you can save \$20 when you spend \$100 on PUMA new arrivals direct from Amazon.com. Shop [men's](#) and [women's](#) and enjoy FREE Super Saver Shipping (restrictions apply).

[Save \\$20 on PUMA New Arrivals, Plus Free Shipping](#)



PUMA Men's Jago
V Suede Sneaker



PUMA Men's Cell
Voltra Sneaker



PUMA Men's First
Round S Mix High
Sneaker



PUMA Men's TX
301 Sneaker



Amazon.com Invites You Back for New Fall Baby Apparel

Inbox | X



Amazon.com to me

[show details](#) Sep 30

[Reply](#) | ▼

amazon.com

Let them choose from millions of items

[Amazon.com Gift Cards](#)

[Baby](#)

[Your Amazon.com](#)

[Today's Deals](#)

[See All Departments](#)

Dear Amazon.com Customer,

As someone who's recently visited our [Baby Store](#) at Amazon.com, you might be interested in checking out the new [Baby Apparel Store](#). See new fall apparel from popular name brands like [The Children's Place](#) and [Greendog](#). Plus get FREE Super Saver Shipping on purchases of \$25 or more (restrictions apply).

► [Shop the Baby Apparel Store now](#)



Apparel for a
Beautiful Baby

► [Shop Amazon.com](#)

[New window](#)

[Print all](#)

[Turn off highlighting](#)

Would you like to...

[Map this](#)

1200 12th Ave S
Seattle, WA 98144

Sponsored Links

[Sky Online Direct](#)

Buy Hottest New Sky Clothing Direct
Free Shipping! Use Code: 000
[www.UnicaMelrose.com](#)

[Kids' Designers 70% Off](#)

Sign up to get kids designer labels
at sample sale prices.
[www.gilt.com](#)

[Baby Leg Warmers](#)

Hip Baby Leg Warmers in Many
Styles Great Gifts - Free Ship & No
Tax!
[www.jaminga.com](#)

Δυναμικές προτάσεις

amazon.com Hello, Magdalini Eirinaki. We have [recommendations](#) for you. (Not Magdalini?) FREE 2-Day Shipping, No Minimum Purchase

Magdalini's Amazon.com Today's Deals Gifts & Wish Lists Gift Cards Your Account | Help

Shop All Departments Search Amazon.com GO Cart Your Lists

Magdalini's Amazon.com Your Browsing History Recommended For You Rate These Items Improve Your Recommendations Your Profile Your Communities Learn More

Magdalini, Welcome to Your Amazon.com (If you're not Magdalini Eirinaki, click here.)

Today's Recommendations For You

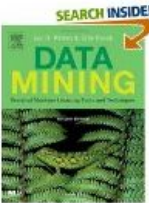
Here's a daily sample of items recommended for you. Click here to [see all recommendations](#)



[The First Years Newborn Soothie Pacifiers 2-Pack](#) ☒
★★★★☆ (31) \$4.97



[A Semantic Web Primer \(Cooperative Informatics\)](#) ☒ by Grigoris Antoniou
★★★★★ (13) \$33.60



[Data Mining: Practical Learning To...](#) ☒ by
★★★★★ (25) \$41.50

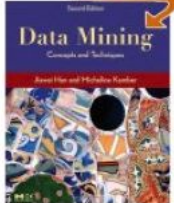
Any Category Artificial Intelligence Baby Girls Body Suits Britpop **Education** Dance Pop Europe Health, Mind & Body Infants Mattress P Shirts & Body Suits Sleepwear Software D

Amazon.com: Recommended for You

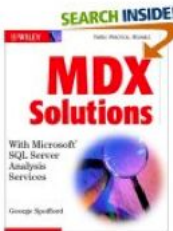
<http://www.amazon.com/gp/yourstore/home/n> Google

SJSU forms OnTeaching.html Blog Miner - CaseWiki www.paramithas.gr KandyKandy

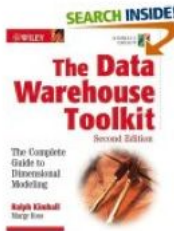
Customers Who Bought Items in Your Recent History Also Bought:



[Data Mining, Second Edition, Second Edition](#) Hardcover by Jiawei Han (Nov 3, 2005)



[MDX Solutions: With Microsoft SQL Server Analysis Services](#) Paperback by George Spofford (Jul 20, 2001)



[The Data Warehouse Toolkit](#) Paperback by Ralph Kimball (April 26, 2002)

Προσαρμογή ή Εξατομίκευση;

- Προσαρμογή

- Με βάση τις δηλωμένες προτιμήσεις του χρήστη
- Με χρήση ευριστικών τεχνικών

- Εξατομίκευση

- Συνδυασμός δηλωμένων και υποκρυπτόμενων στοιχείων (προτιμήσεων κλπ)
- Χρήση τεχνικών web mining

Περιεχόμενα

- Εισαγωγή
- Web usage mining & Web Personalization
 - Συλλογή δεδομένων και προετοιμασία
 - Μοντελοποίηση των Web usage data
 - Ανάλυση των Web Usage Patterns

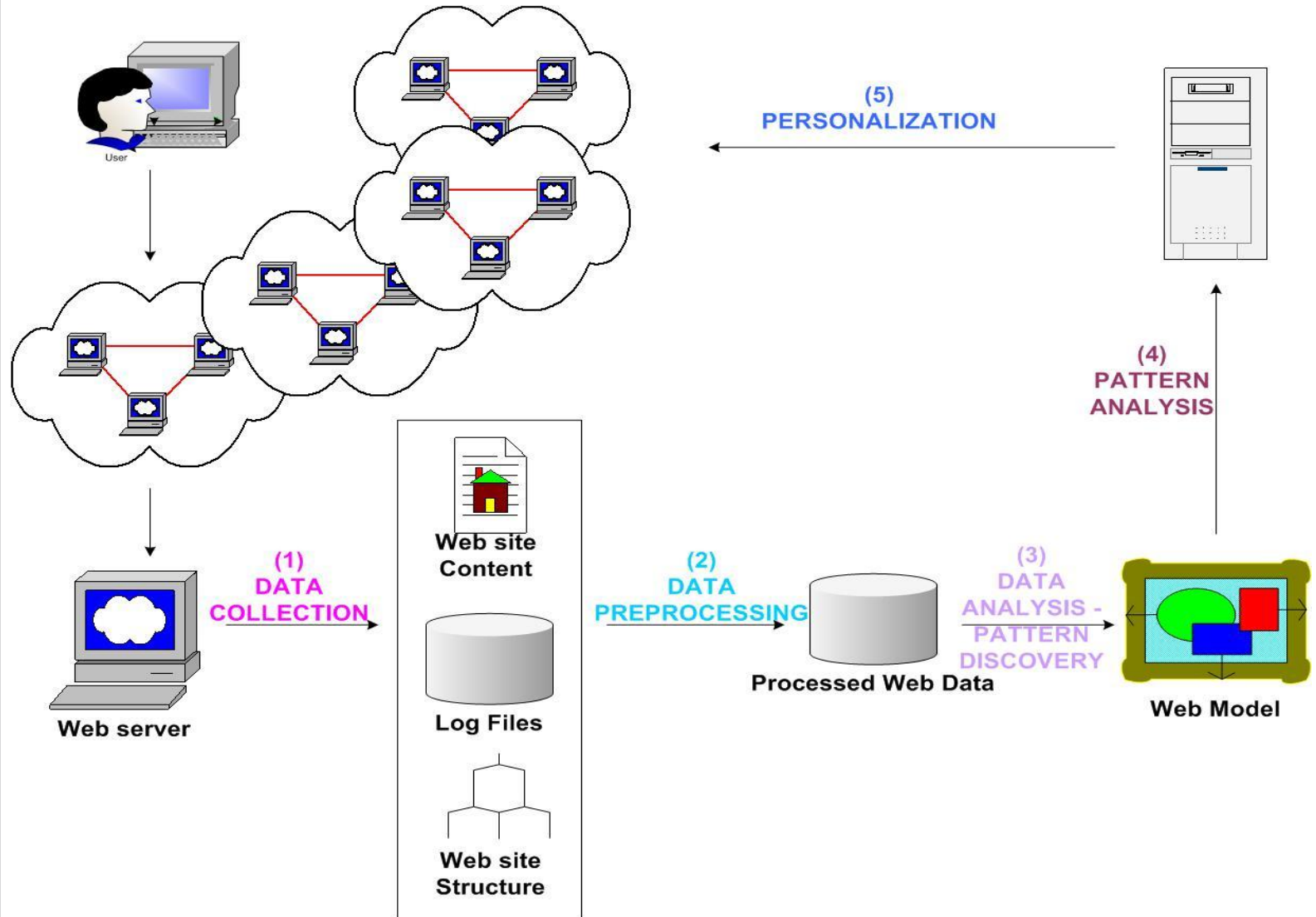
Εξατομίκευση στο Web - Ορισμός

Κάθε ενέργεια που προσαρμόζει πληροφορίες ή υπηρεσίες ενός a Web site στις ανάγκες ενός χρήστη/ή μιας ομάδας, εκμεταλλευόμενη τη γνώση που υπάρχει στη συμπεριφορά περιήγησης του χρήστη

Βήματα:

- συλλογή δεδομένων περιήγησης (usage data)
- κατηγοριοποίηση των δεδομένων (ανά χρήστη ή ομάδα)
- ανάλυση δεδομένων και δημιουργία μοντέλου (χρήση content-based filtering, collaborative filtering, Web usage mining, page ranking)
- Καθορισμών των ενεργειών που απαιτούνται για την εξατομίκευση με βάση το μοντέλο

Web Personalization



Περιεχόμενα

- Εισαγωγή
- Web usage mining & Web Personalization
 - Συλλογή δεδομένων και προετοιμασία
 - Μοντελοποίηση των Web usage data
 - Ανάλυση των Web Usage Patterns

Προετοιμασία δεδομένων

- Η πιο απαιτητική, υπολογιστικά, διαδικασία
- Απαιτεί συχνά ειδικούς αλγορίθμους και ευριστικά
- Περιλαμβάνει
 - Προ-επεξεργασία των δεδομένων
 - Ολοκλήρωση των δεδομένων από πολλαπλές πηγές
 - Μετατροπή των δεδομένων σε κατάλληλη μορφή

Δεδομένα χρήσης – Usage data

- Η βασική πηγή για αυτά που αποκαλούνται *clickstream data*
- Τα αρχεία καταγραφής (web log) ακολουθούν κάποιο πρότυπο
- Extended Log Format (W3C)
 - Prefixes: *c* (client), *s* (server), *r*(remote), *cs* (client to server), *sc*, *sr* (server to remote server-proxies), *rs*
 - Identifiers: *date, time, ip, bytes, cached, status, comment, method, uri, uri-stem, uri-query*

Ένα τυπικό αρχείο

```
<date><time><c-ip><cs-method><cs-uri-stem><sc-status><cs-host>  
><cs(User-Agent)><cs(Referer)>
```

2002-01-28 14:31:33 66.77.73.87 GET /people/mhalk.html 404 www.db-net.aueb.gr
FAST-WebCrawler/3.3+(crawler@fast.no;+http://fast.no/support.php?c=faqs/crawler) -

2002-01-28 14:36:26 192.93.2.2 GET /projects.html 200 www.db-net.aueb.gr
Mozilla/4.75+[en;+X11;+U;+Linux+2.4.16+i686)
<http://www.google.com/search?hl=en&q=DBGlobe+>

2002-01-28 14:36:31 192.93.2.2 GET /index.php 200 www.db-net.aueb.gr
Mozilla/4.75+[en;+X11;+U;+Linux+2.4.16+i686) <http://www.db-net.aueb.gr/projects.html>

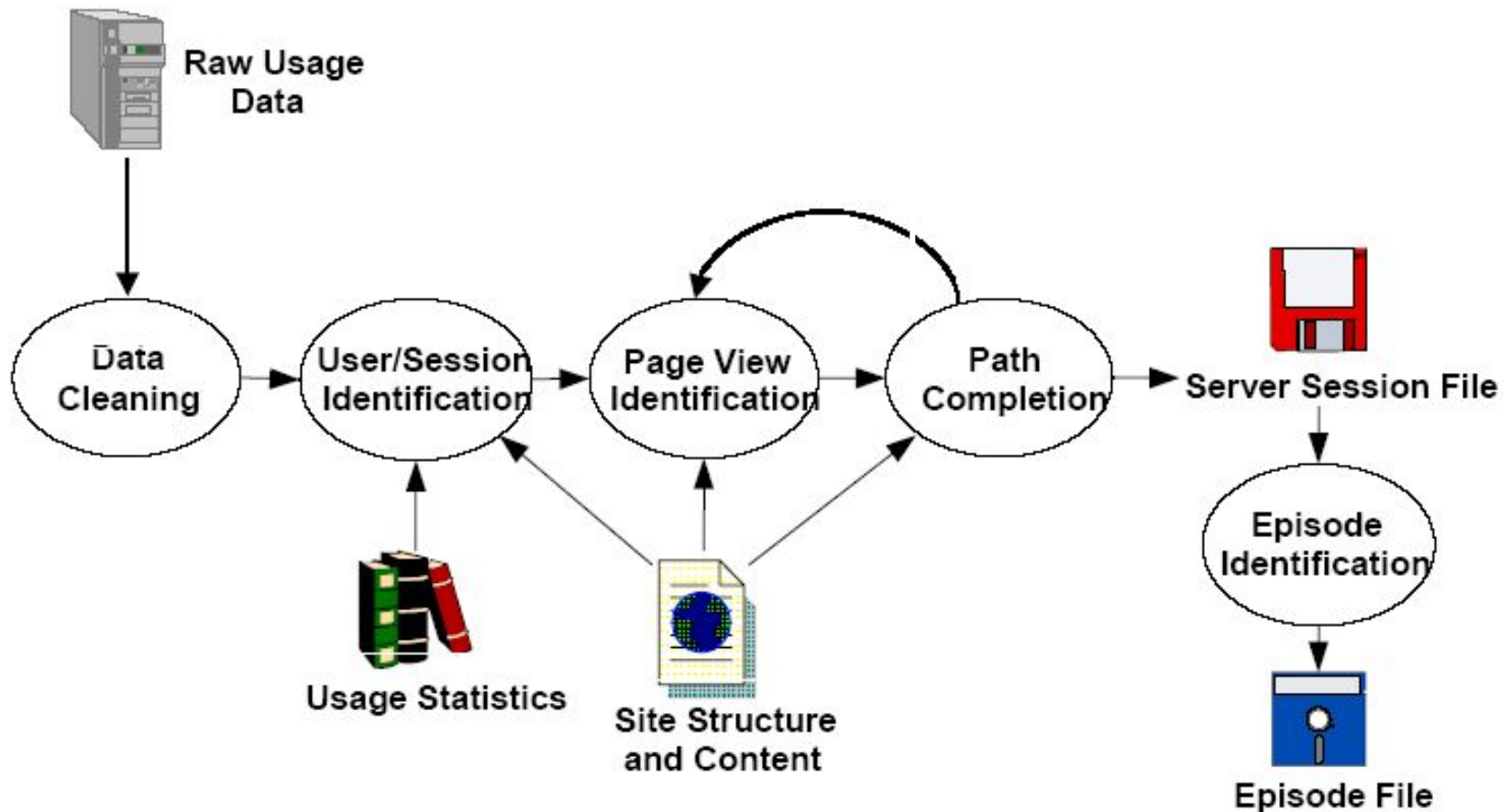
2002-01-28 14:36:38 192.93.2.2 GET /jobs.htm 200 www.db-net.aueb.gr
Mozilla/4.75+[en;+X11;+U;+Linux+2.4.16+i686) <http://www.db-net.aueb.gr/index.php>

2002-01-28 14:36:38 192.93.2.2 GET /images/dbnetOnlyMicro.gif 200 www.db-net.aueb.gr
Mozilla/4.75+[en;+X11;+U;+Linux+2.4.16+i686) <http://www.db-net.aueb.gr/index.php>

Δεδομένα χρήστη - User Data

- Δημογραφικά δεδομένα για τους χρήστες
- Βαθμολογίες που δίνουν οι χρήστες (User ratings)
- Προηγούμενες αγορές ή επισκέψεις
- Άλλες δηλωμένες ή υποκρυπτόμενες προτιμήσεις (π.χ. θέματα ενδιαφέροντος)

Προεπεξεργασία δεδομένων



[B.Mobasher,R.Cooley, J.Srivastava. Automatic personalization based on Web Usage Mining. Communications of the ACM 43(1) (Aug. 2000)]

Συγχώνευση & καθαρισμός δεδομένων

● Συγχώνευση δεδομένων

- Συγχώνευση αρχείων καταγραφής από πολλούς web & application servers
- Τι γίνεται αν δεν έχουμε session-ids;
 - Χρησιμοποιούμε παρόμοιες τεχνικές ταυτοποίησης, όπως στην αναγνώριση χρήστη ή συνεδρίας (session)
- Πολύ χρήσιμη για WUM σε πολλαπλά site

● Καθαρισμός δεδομένων

- Προσαρμοσμένη σε κάθε site. Περιλαμβάνει:
 - Αφαίρεση άχρηστων attributes από τα log files
 - Αφαίρεση άχρηστων εγγραφών από τα log files (π.χ. ??)

Ταυτοποίηση χρήστη

- Απαραίτητη για τη διάκριση μεταξύ των χρηστών
 - Εύκολη όταν έχουμε εγγραφή χρήστη (login)
- Διαφορετικά;
 - **Cookies** στη μεριά του χρήστη
 - η πιο συνηθισμένη προσέγγιση
 - Πρόβλημα: Μπορεί να τα απενεργοποιήσει ο χρήστης
 - Ταυτοποίηση μέσω IP address
 - Πρόβλημα: Δεν υπάρχει αντιστοιχία 1 προς 1
 - Ευριστικά: **IP & κατώφλι χρόνου σύνδεσης**

Ταυτοποίηση συνεδρίας

- Αφορά την κατάτμηση της δραστηριότητας του χρήστη σε επιμέρους συνεδρίες (sessions)
 - Κάθε session αναπαριστά μια μοναδική επίσκεψη στο site
- Χρησιμοποιούνται ευριστικές τεχνικές αν δεν μπορεί να γίνει αυθεντικοποίηση σύνδεσης (μέσω login) ή αν δεν υποστηρίζονται τα session ids
 - Αναδομούμε την ακολουθία ενεργειών του χρήστη μέσα στη συνεδρία, αναλύοντας τα δεδομένα χρήσης (clickstream data)

Ευριστικά για Sessionization

- Τεχνικές χρόνου

- Χρήση καθολικών ή τοπικών διαστημάτων αδράνειας για το διαχωρισμό των συνεδριών

- Τεχνική: Η συνολική διάρκεια ενός session δεν μπορεί να ξεπερνά ένα όριο θ

- Αν t_o η χρονική στιγμή του πρώτου αιτήματος μιας συνεδρίας S , κάθε αίτημα με χρονική σφραγίδα t ανατίθεται στο S αν $t - t_o \leq \theta$

- Τεχνική: Ο συνολικός χρόνος σε μια σελίδα δεν μπορεί να υπερβαίνει ένα όριο δ

- Αν t_1 η χρονική στιγμή ενός αιτήματος μιας συνεδρίας S , το επόμενο αίτημα με χρονική σφραγίδα t_2 ανήκει στο S αν $t_2 - t_1 \leq \delta$

Ευριστικά για Sessionization

- Τεχνικές δομής

- Χρησιμοποιούν είτε τη στατική δομή του δικτυακού τόπου, είτε την υποκρυπτόμενη σύνδεση που προκύπτει από το πεδίο referrer
- Referrer: ένα πεδίο που μας λέει ποια σελίδα οδήγησε το χρήστη στην τρέχουσα σελίδα

- H-Ref: Ένα αίτημα q προστίθεται σε μια υπάρχουσα συνεδρία S αν ο referrer για το q ανήκει στην συνεδρία S . Διαφορετικά το q είναι η αρχή μιας νέας συνεδρίας

Ολοκλήρωση μονοπατιού

- Τα Web logs μπορεί να μην καταγράψουν όλα τα αιτήματα
 - Λόγω caching στους clients ή στους proxies
 - π.χ. A->B->C->B->D
- Οι αναφορές που λείπουν προστίθενται με ευριστικές τεχνικές
 - Πρέπει να γνωρίζουμε τη δομή του web site και να εξετάζουμε και την πληροφορία του referrer

Ταυτοποίηση προβολής σελίδων (Pageview)

- Εξαρτάται από
 - Τη διασύνδεση των σελίδων του site
 - Τα περιεχόμενα των σελίδων
 - Την γνώση για τη δομή του site
- Εύκολη για static HTML σελίδες
- Πιο δύσκολη για σύνθετες, δυναμικές σελίδες με πλαίσια κλπ
- Μπορεί να θέλουμε να ταυτοποιήσουμε πολλές σελίδες μαζί ως μία (π.χ. η διαδικασία του “check-out” σε ένα site αγορών)
 - Χρειάζεται να ορίσουμε εκ των προτέρων ένα μοντέλο γεγονότων (“event model”)

Προεπεξεργασία - Παράδειγμα

<date><time><c-ip><cs-method><cs-uri-stem><sc-status><cs-host><cs(User-Agent)><cs(Referer)>

2002-01-28 14:31:33 66.77.73.87 GET /people/mhalk.html 404 www.db-net.aueb.gr
FAST-WebCrawler/3.3+(crawler@fast.no;+http://fast.no/support.php?c=faq/crawler) -

2002-01-28 14:36:26 192.93.2.2 GET /projects.html 200 www.db-net.aueb.gr
Mozilla/4.75+[en]+(X11;+U;+Linux+2.4.16+i686)
<http://www.google.com/search?hl=en&q=DBGlobe+>

2002-01-28 14:36:31 192.93.2.2 GET /index.php 200 www.db-net.aueb.gr
Mozilla/4.75+[en]+(X11;+U;+Linux+2.4.16+i686) http://www.db-net.aueb.gr/projects.html

2002-01-28 14:36:38 192.93.2.2 GET /jobs.htm 200 www.db-net.aueb.gr
Mozilla/4.75+[en]+(X11;+U;+Linux+2.4.16+i686) http://www.db-net.aueb.gr/index.php

2002-01-28 14:36:38 192.93.2.2 GET /images/dbnetOnlyMicro.gif 200 www.db-net.aueb.gr
Mozilla/4.75+[en]+(X11;+U;+Linux+2.4.16+i686) http://www.db-net.aueb.gr/index.php

Αναφορές

- R. Cooley, B. Mobasher, J. Srivastava, *Data Preparation for Mining World Wide Web Browsing Patterns*, in Knowledge and Information Systems, 1(1), 1999
- Πολύ ενδιαφέρον και με μεγάλη επιρροή στη μετέπειτα έρευνα
- Κάπως ξεπερασμένο
 - Δεν υπήρχαν τόσες πολλές δυναμικές σελίδες τότε

Ολοκλήρωση δεδομένων

- Ολοκλήρωση δεδομένων χρήσης / περιεχομένου / χρήστη
- Άλλα σημαντικά γεγονότα για το e-commerce:
 - Αλλαγές στα καλάθια αγορών
 - Εντύπωση (ικανοποίηση/απογοήτευση) του χρήστη
 - Click-throughs (στην περίπτωση των διαφημίσεων: πόσες φορές την επέλεξαν/πόσες φορές την είδαν)
- Χρήσιμη για τον ορισμό μετρικών όπως το **customer conversion ratios** (πόσοι αγόρασαν/πόσοι επισκέφθηκαν)
- Χρειάζονται συγκεκριμένο μοντέλο γεγονότων για κάθε site
- Google Analytics

Περιεχόμενα

- Εισαγωγή
- Web usage mining & Web Personalization
 - Συλλογή δεδομένων και προετοιμασία
 - Μοντελοποίηση των Web usage data
 - Ανάλυση των Web Usage Patterns

Μοντελοποίηση δεδομένων για WUM

- Προ-επεξεργασία των στοιχείων για
 - Ένα σύνολο P από n pageviews
 - Ένα σύνολο T από m user transactions
- Κάθε συναλλαγή είναι μια ακολουθία (μήκους λ) από διατεταγμένα ζεύγη:
 - $t = \langle (p_1, w_1), (p_2, w_2), \dots, (p_\lambda, w_\lambda) \rangle$
- Ο ορισμός του “pageview” αλλάζει ανάλογα με την εφαρμογή
- Τα βάρη ορίζονται με διάφορους τρόπους
 - Χρόνος παραμονής στη σελίδα
 - Προβολή ή όχι της σελίδας (binary)
 - Βαθμολόγηση προϊόντος
- Το αποτέλεσμα είναι ένας πίνακας **user-pageview** (ή πίνακας συναλλαγών - **transaction matrix**)

Χρήση του περιεχομένου

- Ενσωμάτωση του περιεχομένου στη διαδικασία εξόρυξης γνώσης από τα δεδομένα χρήσης
- Δημιουργία ενός πίνακα
pageview-category
- Αυτόματη κατηγοριοποίηση των
σελίδων σε θεματικές ιεραρχίες
κατηγοριών
 - Δημιουργία πιο σύνθετου πίνακα
προβολής κατηγοριών και ειδικών
αλγορίθμων εξόρυξης γνώσης

Περιεχόμενα

- Εισαγωγή
- Web usage mining & Web Personalization
 - Συλλογή δεδομένων και προετοιμασία
 - Μοντελοποίηση των Web usage data
 - Ανάλυση των Web Usage Patterns

Ανίχνευση και Ανάλυση Προτύπων Χρήσης

1. Ανάλυση συνεδριών και επισκεπτών
 2. Ανάλυση συστάδων
 3. Ανάλυση συσχετίσεων (association & correlation)
 4. Ανάλυση ακολουθιακών προτύπων και προτύπων πλοήγησης
 5. Κατηγοριοποίηση και Πρόβλεψη
- Χρησιμοποιούνται συνήθως για την παραγωγή προτάσεων προς τους χρήστες

1. Στατιστική ανάλυση του Log

- Ανάλυση κίνησης:
 - Στατιστικά διαγνωστικά
 - Στατιστικά του server
 - Στατιστικά για Referrers
 - Δημογραφικά χρηστών
 - Στατιστικά πελατών
 - ...
- Χρησιμοποιείται για να βελτιώσει την απόδοση του site, για να παράγει προτάσεις για καλύτερη δομή, για να υποστηρίξει αποφάσεις σχετικές με προώθηση προϊόντων
- Εμπορικά προγράμματα ανάλυσης χρησιμοποιούν αλγορίθμους εξόρυξης γνώσης

- 
- Site activity:(total # of visits, average # of hits, successful/failed/redirected/cached hits, average view time, average length of a path through a site)
 - Diagnostic statistics (server errors and page not found errors)
 - Server statistics (top pages visited, entry/exit pages, single access pages)
 - Referrers statistics (top referring sites, search engines, keywords)
 - User demographics (top geographical location, most active countries/cities/organizations)
 - Client statistics (visitors' web browser, operating system, cookies)

Ανάλυση OLAP

- Αποθήκευση πληροφορίας σε data warehouses
- Χρήση τεχνικών OLAP (Online Analytical Processing) για την ανάλυση των δεδομένων
- Προβολή των δεδομένων σε διαφορετικά επίπεδα ολοκλήρωσης (ανά session, χρήστη, χώρα κλπ)
- Πολυδιάστατη ανάλυση (ο χρόνος είναι μια διάσταση)
- Η έξοδος των OLAP queries γίνεται είσοδος σε πολλούς αλγορίθμους data mining & visualization

Εργαλεία ανάλυσης

- Web Analytics Software:
<http://www.kdnuggets.com/software/web-analytics.html>
- Άλλα:
 - <http://www.chiark.greenend.org.uk/~sret1/stats/>
(Analog -free)
 - <http://www.123logalyzer.com/sample/>
(123LogAnalyzer)
 - <http://www.google.com/analytics/tour.html>
(Google Analytics)

- by Time Increment
- by Day of the Week
- by Hour of the Day
- by PageViews per day
- by Visitor Stay Length
- Visitor/IP Information
 - by Access Time
 - by Hits
 - by Bandwidth
 - by Stay Length
 - by Authenticated Visitors
- Resource Accessed
 - by Page Views
 - by Browsing Sequences
 - by Downloaded Files
 - by Accessed Directories
 - by Accessed Images
 - Top Entry Pages/Files
 - Top Exit Pages/Files
- Referrer Statistics
 - by Referring Domains
 - by Referring URLs
- Search Engine Statistics
 - Top Search Engines
 - by Keywords
 - by Each Search Engine
- Geographic Region
 - Top Countries
 - Top US States
 - Top US Cities
- Browsers and Platforms
 - Browsers
 - Operating Systems
 - Visiting Spiders
- Technical Information
 - Missing Files (404)
 - Miscellaneous Errors
- Marketing Performance Report
 - General Marketing Report
 - By Target Page Views
 - By Target File Downloads

Rank#	Directory	Hits	Data Transferred(KB)
1	/images/	343187	771525
2	/	37873	576798
3	/download/	24907	12675792
4	/se/	77	209
5	//	24	487
6	http://www.zy2000.com/images/	21	56
7	http://www.zy2000.com/	9	21
8	/pad/	7	31
9	/https://www.qwerks.com/order/	4	0
10	http://www.zy2000.com:80/download/	3	0

Top Entry Pages and Files

Rank#	Page/File	Count
1	/	12296
2	/download/mp3cd141.exe	4373
3	/down.htm	1494
4	/download/mp3cd120.exe	827
5	/FAQ.htm	678
6	/index.html?ZYYZSW	653
7	/news.htm	506
8	/device.htm	452
9	/download/mp3cd130.exe	433
10	/download/mp3cd140.exe	420

Top Exit Pages and Files

Rank#	Page/File	Count
1	/download/mp3cd141.exe	8037
2	/	5901
3	/down.htm	2532
4	/FAQ.htm	1388
5	/download/mp3cd120.exe	857
6	/device.htm	616
7	/news.htm	544
8	/index.html?ZYYZSW	474
9	/download/mp3cd130.exe	454
10	/download/mp3cd140.exe	444

Most Path Throughs

Rank#	Path	Count	Percentage%
1	/	5231	27.38%
2	/download/mp3cd141.exe	4131	21.62%
3	/->/down.htm->/download/mp3cd141.exe	1954	10.23%
4	/->/down.htm	1038	5.43%
5	/down.htm	801	4.19%
6	/download/mp3cd120.exe	785	4.11%
7	/->/FAQ.htm	488	2.55%
8	/FAQ.htm	412	2.16%
9	/download/mp3cd130.exe	403	2.11%
10	/download/mp3cd140.exe	383	2.00%

Παράδειγμα – Στατιστική ανάλυση

- Τα πιο συχνά μονοπάτια που διάλεξαν οι χρήστες/Μέσο μήκος μονοπατιού
 - Μ. Μήκος μονοπατιού: 3 σελίδες
- Σελίδες εισόδου/εξόδου, σελίδες μοναδικής πρόσβασης
 - 65% των επισκεπτών έφυγαν από το site στη σελίδα `/home/misc/apage.htm`
- # επιτυχημένων / αποτυχημένων / αναδρομολογημένων / αποθηκευμένων αιτημάτων σελίδας
- Top referring sites, top keywords/phrases
 - Top keyword : “web mining”
- Πιο δραστήριες χώρες/πόλεις/οργανισμοί

2. Ανάλυση συστάδων

- Ομαδοποίηση στοιχείων με κοινά χαρακτηριστικά
 - Συστάδες χρηστών
 - Συστάδες σελίδων
- Χρήση γνωστών αλγορίθμων συσταδοποίησης:
 - π.χ. K-means
 - Ανάγκη ορισμού ενός μέτρου ομοιότητας (ή απόστασης)
- Το αποτέλεσμα της συσταδοποίησης δεν ερμηνεύεται άμεσα
 - Ανάλυση κάθε συστάδας και εξαγωγή επιχειρηματικής γνώσης
 - Χρήση συστάδων ως είσοδος στη διαδικασία εξατομίκευσης

Συστάδες χρηστών

- Βασίζονται στα δεδομένα χρήσης
- Ομάδες χρηστών με παρόμοια πρότυπα περιήγησης.
- Χρήσεις: Κατάτμηση της αγοράς, ανακάλυψη κοινοτήτων χρηστών
- Αντιπροσώπευση/περιγραφή των συστάδων
 - Υπολογισμών των centroids
 - Κάθε χρήστης αναπαρίσταται με ένα ζυγισμένο σύνολο από προβολές σελίδας/προβολές κατηγορίας/λέξεις κλειδιά
- Εξατομίκευση (Personalization):
 - Για κάθε νέο χρήστη υπολογίζουμε την ομοιότητά του με κάθε centroid και τον αναθέτουμε στην πιο όμοια συστάδα
 - Του προτείνουμε items/pageviews που έχουν προτιμήσει τα υπόλοιπα μέλη της συστάδας

Συσταδοποίηση σελίδων

- Με βάση τα δεδομένα χρήσης και τα χαρακτηριστικά περιεχομένου των σελίδων (π.χ. λέξεις/ιδιότητες προϊόντων)
 - Συστάδες με βάση τη χρήση: pages/items που αγοράζονται (επισκέπτονται) από πολλούς χρήστες
 - Συστάδες με βάση το περιεχόμενο: pages/items που σχετίζονται με το ίδιο θέμα
- Εξατομίκευση (Personalization):
 - Προτείνουμε στο χρήστη σελίδες (ή items) με βάση τις προηγούμενες επισκέψεις/αγορές του

3. Κανόνες συσχέτισης

- Εύρεση συχνών προτύπων / συσχετίσεων μεταξύ σελίδων/προϊόντων που επισκέφθηκαν/αγοράστηκαν μαζί
- Χρήση της παραγόμενης γνώσης για:
 - Καλύτερη οργάνωση του περιεχομένου του
 - Παροχή καλύτερων προτάσεων προϊόντων
- $A \square B (c, s)$
 - A, B: σύνολα σελίδων/προϊόντων
 - $c: \text{confidence} = P(B/A) = \#A,B / \#A$, $s: \text{support} = P(A \& B) = \#A,B / \#all$
- Χρήση **Apriori**
 - “perfume X, handbag Y $\square$ camera Z”
 - “special_promo/, products/gadgets/ $\square$ shopping_cart/”

4. Εξόρυξη ακολουθιακών προτύπων

- Εύρεση συχνών προτύπων μέσα στις συνεδρίες όπου ένα σύνολο από επισκέψεις σελίδων (αγορές προϊόντων) ακολουθείται από ένα άλλο σύνολο επισκέψεων (σε ένα χρονικά διατεταγμένο σύνολο συνεδριών)
- Χρήση για:
 - Ανάλυση τάσεων (trend analysis)
 - Πρόβλεψη μελλοντικών προτύπων επίσκεψης για τοποθέτηση στοχευμένων διαφημίσεων
 - Παραγωγή προτάσεων

Μοντέλα/αλγόριθμοι

- Συνήθως χρησιμοποιούνται μοντέλα Markov
 - Χρησιμοποιούν μέρος της ακολουθίας επισκέψεων
 - Προβλέπουν ποιες θα είναι οι επόμενες επισκέψεις
 - Προβλέπουν και παρέχουν προτάσεις περιεχομένου με τη μορφή συνδέσμων
 - Μεταφορτώνουν προκαταβολικά περιεχόμενο (Web prefetching)
- Τα μοντέλα Markov ορίζουν μια ακολουθία από καταστάσεις και μεταβάσεις με κάποια πιθανότητα
- Η πιθανότητα να μεταβούμε στην κατάσταση s_i από την κατάσταση s_j ακολουθώντας κάποιο μονοπάτι, ισούται με το γινόμενο των πιθανοτήτων των ενδιάμεσων μεταβάσεων
- Μπορούμε να υπολογίζουμε κάθε πιθανότητα μετάβασης (π.χ. σε μια κατάσταση γ) λαμβάνοντας υπόψη μας όχι μόνο την κατάσταση στην οποία βρισκόμαστε (π.χ. β) (Αλυσίδες Markov 1^{ης} τάξης) αλλά και την κατάσταση από την οποία ήρθαμε στην κατάσταση στην οποία βρισκόμαστε (π.χ. από την α ή την α') (Αλυσίδες Markov 2^{ης} τάξης)
- Όσο μεγαλύτερης τάξης είναι οι αλυσίδες τόσο μεγαλύτερη η ακρίβεια αλλά μικρότερη η κάλυψη του μοντέλου και μεγαλύτερη η πολυπλοκότητα

5. Κατηγοριοποίηση

- Αναπτύσσουμε ένα προφίλ χρήστη για κάθε κατηγορία
 - Με χρήση: Δέντρων απόφασης, Κατηγοριοποιητών Naïve Bayesian, SVMs
 - Χρησιμοποιούμε συνεργατικό φιλτράρισμα (**collaborative filtering**): με βάση τους k-NN
- Collaborative Filtering
 - Οι χρήστες βαθμολογούν/αγοράζουν προϊόντα
 - Τα μοντελοποιούμε (ratings/purchases) ως διανύσματα
 - item vector $\langle i_1, i_2, \dots, i_n \rangle$
 - Π.χ. ο χρήστης A (binary): $\langle 1, 0, 0, 1, \dots, 0 \rangle$
 - Π.χ. ο χρήστης B (weighted): $\langle 0.6, 0, 0.1, \dots, 0.9 \rangle$
 - Λογική: Χρήστες με παρόμοιες βαθμολογήσεις/αγορές έχουν ανάλογα ενδιαφέροντα
 - Το amazon.com και άλλα e-commerce sites χρησιμοποιούν τέτοιους μηχανισμούς

★★★★★ COMEDY SUGGESTIONS (68) [See all 68 >](#)



Add



Stranger than Fiction

Because you enjoyed:
[Amelie](#)
[Lost in Translation](#)
[Big Fish](#)



Add



Grease

Because you enjoyed:
[Look Who's Talking](#)
[Pretty Woman](#)
[Sleepless in Seattle](#)



Add



This Is Spinal Tap

Because you enjoyed:
[The Big Lebowski](#)
[High Fidelity](#)
[Citizen Kane](#)

★★★★☆ DRAMA SUGGESTIONS (159) [See all 159 >](#)



Add



Babel

Because you enjoyed:
[Lost in Translation](#)
[Eternal Sunshine of the Spotless Mind](#)
[Sideways](#)



Add



Traffic

Because you enjoyed:
[Crash](#)
[Requiem for a Dream](#)
[Lost in Translation](#)



Add



Sophie's Choice

Because you enjoyed:
[Capote](#)

Collaborative filtering

- Δημιουργία «γειτονιάς»
 - Απεικόνιση κάθε χρήστη σε μια «γειτονιά» με χρήση μέτρων ομοιότητας διανυσμάτων
 - Υπολογισμός όλων των ομοιοτήτων μεταξύ ενός χρήστη και των προηγούμενων συναλλαγών των μελών της γειτονιάς
 - Επιλογή των top-k πιο όμοιων συναλλαγών με την τρέχουσα συναλλαγή του χρήστη
- Φάση προτάσεων
 - Με βάση τα «γειτονικά» διανύσματα συναλλαγών προτείνει στο χρήστη προϊόντα που δεν έχει αγοράσει αλλά έχουν αγοραστεί στις παρόμοιες συναλλαγές άλλων χρηστών

Παράδειγμα - Collaborative filtering

<i>Cit. Kane, Godfather, Scrubs, Lost, PrisonBreak, Friends</i>	<i>Sim (Jaccard)</i>
0 1 1 0 0 1	66%
0 0 1 1 1 0	0%
1 0 0 0 0 0	0%
1 1 0 0 0 1	66%
1 0 0 0 0 1	33%

Current user likes:
Godfather, Friends

0,1,0,0,0,1

1: likes
0: not rated/does
not like

System
recommends:

Citizen Kane,
Scrubs

Φιλτράρισμα βάση χρήστη

- n διανύσματα χρηστών U_i
- m αντικείμενα
- (Αραιός) πίνακας $m \times n$ με βαθμολογίες χρηστών για αντικείμενα: $r_{i,j}$
- Υπολογισμός της ομοιότητας μεταξύ χρηστών με βάση κάποιο μέτρο

- Cosine similarity $sim(U_i, U_j) = \cos(U_i, U_j) = \frac{U_i U_j}{\|U_i\|_2 \|U_j\|_2}$
- Pearson coefficient

- Πρόβλεψη των τιμών που λείπουν για ένα νέο χρήστη

$$pred(NU, m_i) = \frac{\sum_{m_i \in U_k} sim(NU, U_k) r_{k,i}}{\sum_{m_i \in U_k} sim(NU, U_k)}$$

Παράδειγμα

	m1	m2	m3	m4	m5
U1	5	4	3	3	3
U2	2	2	4	4	3
U3	1	3	4	5	5
U4	5	5	4	3	2
U5	3	4	3	2	3

NU	4	?	3	4	?
----	---	---	---	---	---

- Βαθμολογίες ταινιών: User-item matrix
- Πρόβλεψη για τις βαθμολογίες που λείπουν και προτάσεις για ταινίες

$$\text{sim}(NU, U_1) = \cos(NU, U_1) = \frac{NU \cdot U_1}{\|NU\|_2 \|U_1\|_2} = 0,9764$$

$$\text{pred}(NU, m_i) = \frac{(0,9764 * 4 + \dots)}{(0,9764 + \dots)}$$

Συνεργατικό φιλτράρισμα σε επίπεδο στοιχείου (Item-based)

- Η User-based προσέγγιση δεν έχει καλή κλιμάκωση
 - Κάθε φορά που προστίθεται ένας νέος χρήστης, πρέπει να υπολογιστεί η ομοιότητά του με όλους τους άλλους χρήστες
- Λύση: Item-based collaborative filtering
 - Σύγκριση στοιχείων με βάση τις βαθμολογίες που έχουν λάβει από τους χρήστες
 - Και πάλι χρησιμοποιούμε τους k-NN
- Βρίσκουμε τα top-k items που έχουν λάβει παρόμοια βαθμολογία από πολλούς χρήστες
 1. Υπολογισμός των ομοιοτήτων ανά δύο (item-item similarity)
 2. Χρήση των βαθμολογιών που έχουν δώσει ήδη οι χρήστες, πρόβλεψη των υπολοίπων βαθμολογιών
 3. Βαθμονόμηση των προϊόντων και πρόταση των top-n
- Τα προϊόντα που έχουν ίδιες βαθμολογίες είναι όμοια, άρα αν ένας χρήστης αγοράσει το ένα καλό είναι να του προτείνουμε το άλλο