



Personalization

Τεχνολογίες & Υπηρεσίες [Part II]

Πανεπιστήμιο Πατρών
Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Μαρία Ρήγκου, ΠΔ407/80



Δομή Μαθήματος

- ✘ Τι είναι observational personalization?
- ✘ Τι είναι web usage mining
- ✘ Διαδικασία εξατομίκευσης με βάση web usage mining
 - Στάδια προεπεξεργασίας δεδομένων
 - Αναγνώριση χρήστη
 - Αναγνώριση pageview
 - Συμπλήρωση μονοπατιού
 - Ανακάλυψη προτύπων (pattern discovery)
 - Clustering
 - Classification
 - Association rule mining
 - Sequential pattern discovery
- ✘ Περιορισμοί και νέες προοπτικές



Πού οφείλεται το ενδιαφέρον για observational personalization

- ✖ Τα δεδομένα που χρησιμοποιούνται σαν είσοδος δεν είναι **υποκειμενικές** περιγραφές από πλευράς χρηστών
 - δεν περιέχουν προκατάληψη
- ✖ Το web usage mining μειώνει την ανάγκη για συγκέντρωση υποκειμενικών βαθμολογήσεων από πλευράς χρηστών καθώς και την καταγραφή προσωπικών προτιμήσεων κατά την εγγραφή (registration)
- ✖ Τα προφίλ εξαγονται (και ενημερώνονται) **δυναμικά** από τα patterns των χρηστών
 - δεν μειώνεται η απόδοση του συστήματος με το χρόνο λόγω μη ενημέρωσης των προφίλ



Observational Personalization

- ✖ Εξατομίκευση που βασίζεται στην παρατήρηση της πλοηγητικής συμπεριφοράς του χρήστη
- ✖ Στηρίζεται στη **μελέτη της καταγεγραμμένης πλοηγητικής συμπεριφοράς προηγούμενων χρηστών** με σκοπό να εντοπιστούν στοιχεία που θα καθορίσουν το πώς θα πρέπει να εξατομικευτούν οι πληροφορίες, οι υπηρεσίες ή τα προϊόντα που προσφέρει μια web εφαρμογή
- ✖ Χρησιμοποιεί τεχνικές από το χώρο του **web usage mining**
 - : Εφαρμογή μεθόδων στατιστικής και data mining σε **δεδομένα web log**, ώστε να προκύψει ένα σύνολο χρήσιμων μοτίβων (patterns) που αναπαριστούν την πλοηγητική συμπεριφορά των χρηστών



Λίγα λόγια για το web mining

✦ Το web mining ορίζεται ως “η χρήση τεχνικών ανάκτησης δεδομένων (*data mining*) για την ανακάλυψη και εξαγωγή πληροφοριών από έγγραφα και υπηρεσίες Ιστού” και διακρίνεται με βάση το κομμάτι του Ιστού που εξετάζει σε (Kosala and Blockeel, 2000):

- **web content mining** (ανάκτηση πληροφοριών από δεδομένα περιεχομένου Ιστού)
- **web structure mining** (ανάκτηση πληροφοριών από δεδομένα δόμησης Ιστού)
- **web usage mining** (ανάκτηση πληροφοριών από δεδομένα χρήσης Ιστού)

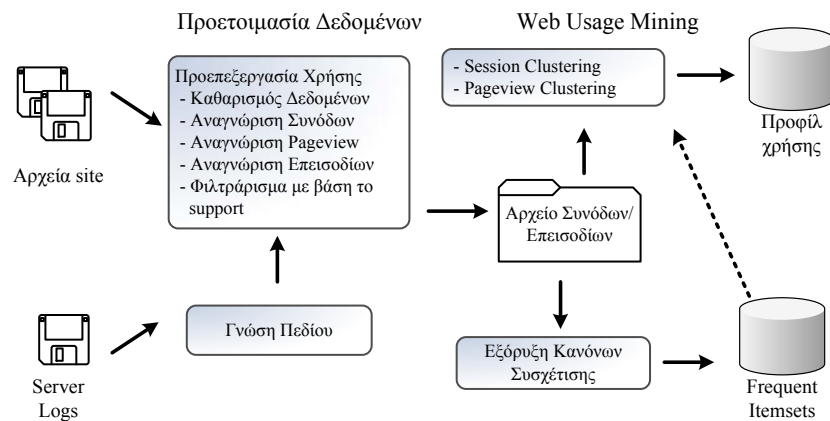


Η διαδικασία εξατομίκευσης με βάση web usage mining

✦ Είναι στην ουσία μια διαδικασία data mining

- Off-line
 - **Data Collection:** συγκεντρώνονται δεδομένα χρήσης από διάφορες πηγές (web servers, clients, proxy servers) λαμβάνοντας υπόψη τη δομή και το περιεχόμενό τους.
 - **Data Pre-processing:** καθαρίζονται τα δεδομένα από θόρυβο, επιλύονται ασυμβατότητες, ολοκληρώνονται και ενοποιούνται ώστε να χρησιμοποιηθούν σαν είσοδος στο στάδιο που ακολουθεί.
 - **Pattern Discovery:** ανακαλύπτεται νέα γνώση με την εφαρμογή τεχνικών από το χώρο του machine learning και της στατιστικής, όπως το clustering, το classification, η ανακάλυψη association rules και sequential patterns. Στόχος είναι να αυτοματοποιηθεί η κατασκευή των μοντέλων χρηστών
- On-line
 - **Knowledge Post-processing:** η νεοαποκτηθείσα γνώση χρησιμοποιείται στην παραγωγή και αποστολή εξατομικευμένων προσαρμογών (με τη μορφή π.χ. recommendations).

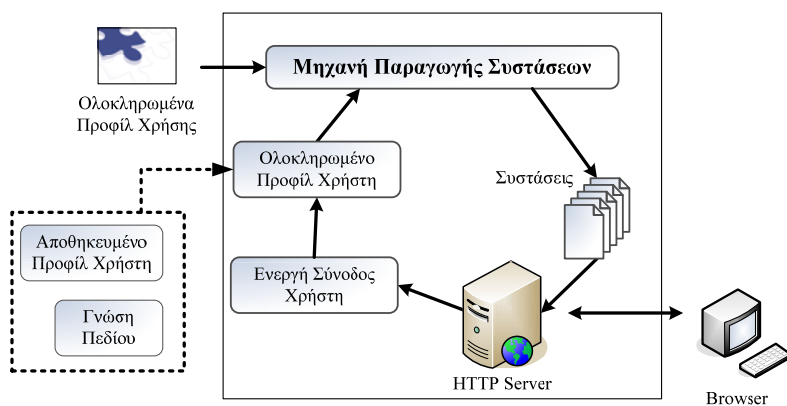
Η off-line φάση της διαδικασίας εξατομίκευσης



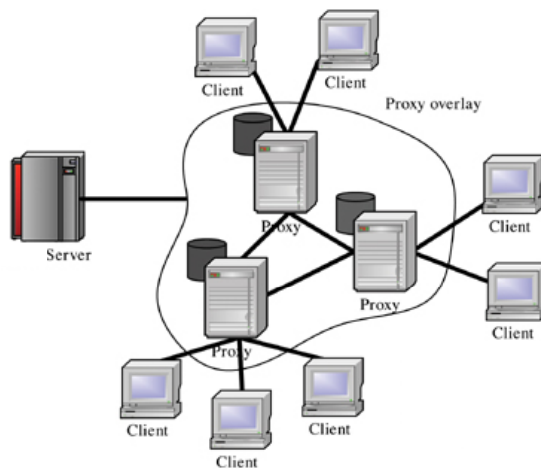
Η on-line φάση της διαδικασίας εξατομίκευσης

- ✦ Ενώ ο browser του χρήστη στέλνει HTTP requests στο server, ο server κρατάει **ιστορικό για το τρέχον session**
- ✦ Η μηχανή recommendations εξετάζει το τρέχον server session σε συνδυασμό με τα **μοτίβα** που έχουν εντοπιστεί
- ✦ Τα μοτίβα που εντοπίζονται χρησιμοποιούνται στην on-line φάση ώστε να παρέχουν στους χρήστες εξατομικευμένο περιεχόμενο με βάση την τρέχουσα συμπεριφορά τους κατά την πλοήγηση
 - προτεινόμενοι σύνδεσμοι ή προϊόντα,
 - στοχευμένες διαφημίσεις (targeted advertisements),
 - κείμενα και γραφικά που να ικανοποιούν τις προτιμήσεις των χρηστών.

Η on-line φάση της διαδικασίας εξατομίκευσης



Συλλογή δεδομένων clickstream





Συλλογή δεδομένων clickstream

■ Σε επίπεδο **web server**

- Είναι η κατ' εξοχή πηγή δεδομένων clickstream
- Όλοι οι web servers μπορούν να καταγράφουν σε πραγματικό χρόνο την αλληλεπίδρασή τους με τους clients με τη μορφή αρχείων log.
- Ένα αρχείο log καταγράφει την κίνηση και τη συμπεριφορά πλοήγησης των χρηστών στις σελίδες ενός δικτυακού τόπου.
- Κάθε φορά που ο web server εξυπηρετεί μια αίτηση, μια νέα εγγραφή προστίθεται στο αρχείο log με λεπτομερείς πληροφορίες σχετικά με
 - τι ζητήθηκε, **πότε**, **από ποιόν**, **σε ποια σελίδα** βρισκόταν ο χρήστης όταν το ζήτησε, το **εάν η αίτηση εξυπηρετήθηκε** με επιτυχία ή υπήρχαν προβλήματα από την πλευρά του server, και ένα σύνολο από επιπλέον στοιχεία που εξαρτώνται από τις ρυθμίσεις του κάθε server και τη μορφοποίηση των αρχείων log που κρατάει



Συλλογή δεδομένων clickstream

■ Σε επίπεδο **client**

- Η συλλογή δεδομένων σε αυτό το επίπεδο επιτυγχάνεται είτε με τη χρήση ενός απομακρυσμένου agent, είτε με την τροποποίηση του πηγαίου κώδικα του browser στον client, εφόσον βέβαια έχει εξασφαλισθεί η συναίνεση του χρήστη.
- Μας απαλλάσσει από το πρόβλημα της αναγνώρισης και πιθανής ανακατασκευής των sessions του χρήστη
- Μειονεκτήματα:
 - παρέχει στοιχεία που αφορούν στη συμπεριφορά **ενός χρήστη** (ή στην καλύτερη περίπτωση ενός μικρού αριθμού χρηστών που χρησιμοποιούν το ίδιο PC και το συγκεκριμένο τροποποιημένο browser)
 - Χρειάζεται η αποδοχή και **συγκατάθεση** του χρήστη να χρησιμοποιήσει έναν browser που καταγράφει κάθε πράξη του

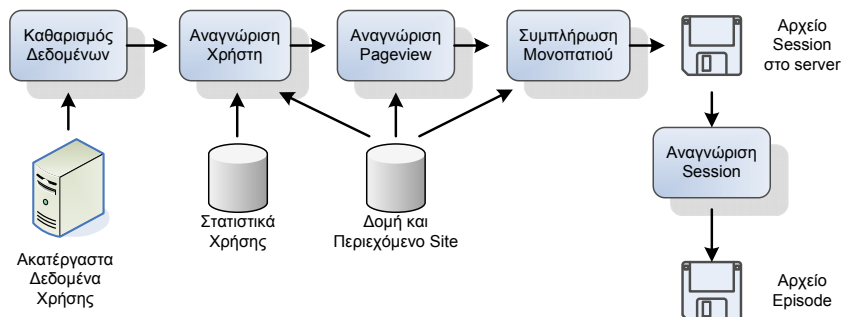
Συλλογή δεδομένων clickstream

✦ Σε επίπεδο proxy server

- Καταγραφή της συμπεριφοράς μιας ομάδας χρηστών που συνδέονται με ένα συγκεκριμένο proxy server, στο σύνολο των web sites που επισκέπτονται
- Προβλήματα λόγω caching
- Προβλήματα συγκέντρωσης και ενοποίησης των δεδομένων clickstream που αφορούν ένα συγκεκριμένο χρήστη, σε ένα web site, κατά τη διάρκεια μιας επίσκεψης
 - Για παράδειγμα, όταν ο proxy server δεν διαθέτει τοπικά μια σελίδα που του ζητήθηκε, τη ζητάει με τη σειρά του από τον αντίστοιχο web server. Στο αρχείο log του web server η αίτηση 'χρεώνεται' στον proxy server και απαιτούνται επιπλέον συνεννοήσεις, επεξεργασία και χρόνος για να εντοπιστεί (εφόσον είναι δυνατό) ο client που ζήτησε τη σελίδα από τον proxy

Προ-επεξεργασία Δεδομένων

- ✦ Για οποιοδήποτε τύπο usage mining πρέπει να προηγηθεί η αναγνώριση ενός συνόλου από **server sessions** στα ακατέργαστα δεδομένα χρήσης
- ✦ Στην ιδανική περίπτωση, κάθε server session παρέχει ακριβή απολογισμό του ποιος είχε πρόσβαση στο web site, ποιες σελίδες ζητήθηκαν και με ποια σειρά, καθώς και για πόσο χρόνο παρέμεινε σε κάθε σελίδα.



Καθαρισμός δεδομένων

- * Αφορά εργασίες όπως τη συγχώνευση των logs από διάφορους servers και το parsing του log ώστε να μετατραπεί σε πεδία δεδομένων
- * Εντοπίζονται requests για αρχεία γραφικών και συνήθως απομακρύνονται, κάτι που μπορεί να γίνει εύκολα ελέγχοντας για καταλήξεις αρχείων όπως "gif" ή "jpg".
- * Εντοπίζεται η κίνηση που οφείλεται σε agents ή spiders
 - δηλώνουν την ταυτότητά τους στο πεδίο user-agent
 - ζητούν πρόσβαση στο αρχείο robots.txt
- * Ομογενοποιούνται τα δεδομένα π.χ. οι αιτήσεις για τις σελίδες sample.edu, www.sample.edu, www.sample.edu/ και www.sample.edu/index.html στην ουσία αναφέρονται στο ίδιο αρχείο
- * Όταν το request συνοδεύεται από δεδομένα CGI θα πρέπει να αναλυθεί σε ζευγάρια name/value <http://www.sample.edu/cgi-in/query? name=Rigou°ree=engineering>

Αναγνώριση του χρήστη

Μέθοδος	Περιγραφή	Θέμα Απορρήτου	(+)	(-)
IP address & Agent	Κάθε μοναδικό ζευγάρι IP/Agent αποτελεί έναν μοναδικό χρήστη.	Μη σημαντικό.	Συνεχής διαθεσιμότητα. Δεν απαιτείται επιπρόσθετη τεχνολογία.	Δεν είναι εγγυημένα μοναδικά. Πρόβλημα με τα random και rotating IPs.
Ενσωματωμένο Session ID	Χρησιμοποιεί δυναμικά παραγόμενες σελίδες για να εισάγει ID σε κάθε link.	Μη σημαντικό/ Μέτριο.	Συνεχής διαθεσιμότητα. Ανεξάρτητη από το IP address.	Δεν υπάρχει η έννοια της επαναληπτικής επίσκεψης. Απαιτεί πλήρως δυναμικό site.
Registration	Οι χρήστες «υπογράφουν» ρητά κατά την είσοδό τους στο site.	Μέτριο.	Μπορεί να καταγράψει μοναδικούς χρήστες όχι απλώς μοναδικούς browsers.	Μπορεί να μην δεχτούν όλοι οι χρήστες να εγγραφούν.
Cookie	Αποθηκεύει ένα αναγνωριστικό στο μηχανήμα client.	Μέτριο/ Σοβαρό.	Μπορεί να καταγράψει επαναλαμβανόμενες επισκέψεις.	Μπορεί να απεργασποιηθεί. Αρνητική δημόσια εικόνα.
SW Agent	Πρόγραμμα που φορτώνεται στον browser και επιστρέφει δεδομένα χρήσης.	Σοβαρό.	Ακριβή δεδομένα χρήσης για ένα συγκεκριμένο web site.	Ενδέχεται να μην γίνει αποδεκτό. Αρνητική δημόσια εικόνα.
Προσαρμοσμένος Browser	Ο browser καταγράφει δεδομένα χρήσης.	Πολύ Σοβαρό.	Ακριβή δεδομένα χρήσης για ολόκληρο το web.	Οι χρήστες πρέπει να ζητήσουν ρητά και να εγκαταστήσουν το λογισμικό.



Το πρόβλημα αναγνώρισης χρήστη

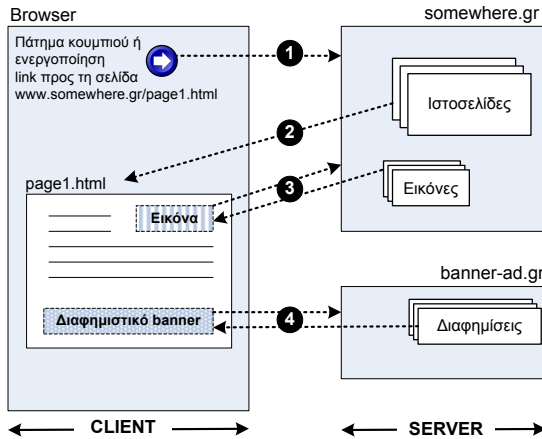
- ✘ Όταν η αναγνώριση πρέπει να στηριχθεί σε IP και agent
 - **Μοναδικό IP address / Πολλαπλά server sessions.**
Ένας proxy server μπορεί να εξυπηρετεί τις αιτήσεις πρόσβασης πολλών χρηστών, πιθανόν και κατά την ίδια χρονική περίοδο.
 - **Πολλαπλό IP address / Μοναδικό server session.**
Κάποιοι ISPs ή εργαλεία privacy αναθέτουν τυχαία κάθε αίτηση του χρήστη σε ένα από ένα IP address από ένα διαθέσιμο σύνολο διαφορετικών.
 - **Πολλαπλά IP address / Μοναδικός Χρήστης.**
Ένας χρήστης που έχει πρόσβαση στο web από διαφορετικά μηχανήματα.
 - **Πολλαπλά server sessions / Μοναδικός χρήστης.**
Ένας χρήστης ανοίγει περισσότερα από ένα παράθυρα browser και επισκέπτεται διαφορετικά τμήματα ενός web site ταυτόχρονα.
 - **Μοναδικός client / Πολλαπλοί χρήστες.**
Περισσότεροι από έναν χρήστες χρησιμοποιούν το ίδιο μηχανήμα (net café).



Αναγνώριση των sessions

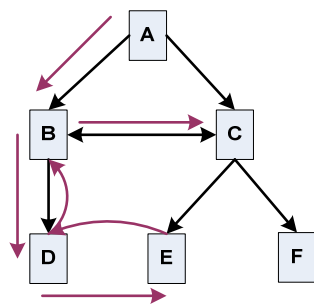
- ✘ Το session για τις ανάγκες της εξατομίκευσης ορίζεται σαν **μια ακολουθία από pageviews που ζητήθηκαν (μέσω διαδοχικών αιτήσεων HTTP) από ένα και μοναδικό χρήστη και εξυπηρετήθηκαν από ένα και μοναδικό server.**
- ✘ **Pageview** είναι αυτό που καταλήγει να βλέπει ο επισκέπτης ενός web site στο παράθυρο του browser
 - Συνδυασμός από html αρχεία, εικόνες, βίντεο, αρχεία ήχου, flash animations, κτλ. που όταν ζητείται ένα URL ζητούνται διαφανώς και καταφθάνουν στον client για να συμπληρώσουν τα περιεχόμενα του παραθύρου.
- ✘ Πότε θεωρούμε ότι **ολοκληρώθηκε** ένα session;
 - Προσαρμοσμένοι browsers που παρακολουθούν την κυκλοφορία σε ολόκληρο το web είναι σε θέση να γνωρίζουν το επόμενο click εκτός site και μπορούν γνωρίζουν τότε ολοκληρώνεται ένα session
 - Χρονικό όριο 30´
 - Αν η μόνη διαθέσιμη πηγή πληροφοριών είναι σε μορφή αρχεία ECLF log, οι χρήστες και τα sessions θα πρέπει να θεωρηθούν ταυτόσημες έννοιες

Αναγνώριση των pageviews



- * Στα sites που αποτελούνται από ένα frame η έννοια του pageview ταυτίζεται με αυτή της σελίδας (HTML αρχείο)
- * Στα multi-frame sites ένα δεδομένο pageview αποτελείται από πολλά αρχεία σελίδων

Συμπλήρωση μονοπατιού



Μονοπάτι Πλοήγησης:
A > B > D > E > D > B > C

URL	Referrer
A	-
B	A
D	B
E	D
C	B

- * Η αίτηση για τη σελίδα C, έχει σαν referrer τη σελίδα B ενώ η σελίδα που ζήτησε πριν τη C ο χρήστης δεν είναι η B αλλά η E.
- * Αυτό αποτελεί ένδειξη ότι στο σημείο αυτό παρεμβλήθηκαν αιτήσεις σελίδων που εξυπηρετήθηκαν μέσω caching και δεν καταγράφηκαν στο server.
- * Το πλήρες μονοπάτι θα μπορούσε να είναι -μεταξύ άλλων- το E > D > B > C ή το E > D > B > A > C.
- * Με βάση το πεδίο referrer μπορούμε να αποκλείσουμε τη δεύτερη περίπτωση, αλλά δε βοηθάει αν εξετάσουμε εναλλακτικά το μονοπάτι E > D > B > A > B > C.
- * Μια ευρετική λύση που χρησιμοποιείται πολύ συχνά είναι να επιλέγουμε το συντομότερο μονοπάτι.



Αναγνώριση επεισοδίων

- ✘ Προαιρετικό βήμα της προ-επεξεργασίας
- ✘ Το **episode** (=επεισόδιο) ορίζεται από το W3C σαν έναν υποσύνολο ενός session χρήστη που έχει σημασιολογική αξία
 - Π.χ. υποσύνολο των pageviews σε ένα news portal που ανήκουν στο τμήμα των αθλητικών
- ✘ Με την αναγνώριση των episodes πετυχαίνουμε **μεγαλύτερο βαθμό διάσπασης μέσα στο session**

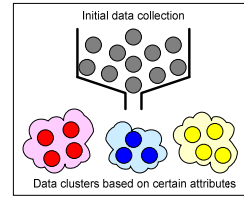


Pattern discovery

- ✘ Το αρχείο session που παράγεται κατά το στάδιο προεπεξεργασίας των δεδομένων που προηγήθηκε, διοχετεύεται σαν είσοδος σε μια ποικιλία αλγορίθμων και τεχνικών ανάκτησης δεδομένων:
 - Clustering
 - Classification
 - Association rule mining
 - Sequential pattern discovery



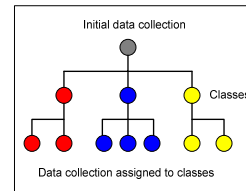
Clustering (Συσταδοποίηση)



- ✘ «Η διαδικασία ομαδοποίησης των δεδομένων ενός αρχικού συνόλου σε κλάσεις ή clusters με τρόπο ώστε τα αντικείμενα που ανήκουν στο ίδιο cluster να έχουν μεγάλη ομοιότητα μεταξύ τους αλλά να διαφέρουν σημαντικά από τα αντικείμενα που ανήκουν σε διαφορετικά clusters»
- ✘ Οι ομοιότητες υπολογίζονται με βάση τις τιμές των χαρακτηριστικών που χρησιμοποιούνται για να περιγράψουν τα αντικείμενα.
- ✘ Στα πλαίσια της εξατομίκευσης διακρίνουμε δύο περιπτώσεις χρήσης του clustering:
 - στο σχηματισμό clusters σελίδων
 - στο σχηματισμό clusters χρηστών
 - Σελίδες που αντιμετωπίζονται με παρόμοιο τρόπο από τους χρήστες ή περιέχουν παρόμοιο περιεχόμενο και αντίστοιχα, χρήστες που εμφανίζουν παρόμοια πλοηγητική συμπεριφορά ομαδοποιούνται στο ίδιο cluster.



Classification (Κατηγοριοποίηση)



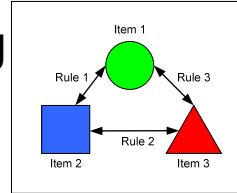
- ✘ Ο στόχος του classification είναι η **αντιστοίχιση αντικειμένων σε κάποια από τις κλάσεις ενός συνόλου προκαθορισμένων κλάσεων**
- ✘ Εντοπίζει τα ιδιαίτερα εκείνα χαρακτηριστικά που διαφοροποιούν τις κλάσεις με βάση ένα σύνολο από αντικείμενα που έχουν ήδη αντιστοιχηθεί στις κλάσεις και δίνονται σαν είσοδος στο classification
 - 'Αντικείμενα' μπορεί να είναι οι χρήστες ή οι σελίδες.
- ✘ Ένα παράδειγμα περιγραφής μιας κλάσης χρηστών θα μπορούσε να είναι το ακόλουθο :
- ✘ Περιοριστική τεχνική για το περιβάλλον του web καθώς χρειάζεται προκαθορισμένες κλάσεις και δεδομένα εκπαίδευσης.

```
sex = male
30 <= age <= 40
marital-status = single
number-of-children = 0
education = higher
```

Association rule mining

Οι χρήστες που επισκέπτονται τις σελίδες A.html και B.html τείνουν επίσης να επισκέπτονται (με confidence 75%) τη σελίδα C.html.

Η τιμή του support αντιπροσωπεύει το γεγονός ότι το itemset {A.html, B.html, C.html} ήταν παρόν στο 1% των καταγεγραμμένων sessions.



* **Οι κανόνες συσχέτισης συνδέουν μεταξύ τους ένα ή περισσότερα γεγονότα**

* Στόχος είναι να εντοπίσουμε 'εξαρτήσεις' μεταξύ διαφορετικών ειδών πληροφοριών που αρχικά φαίνεται να μη συνδέονται σημασιολογικά.

$\{A.html, B.html\} \rightarrow \{C.html\}$ [support = 0.01, confidence = 0.75]

* Στο χώρο του web personalization η τεχνική αυτή μπορεί να εντοπίσει συσχετίσεις μεταξύ σελίδων που δεν συνδέονται απευθείας, ή συσχετίσεις μεταξύ διαφορετικών ομάδων χρηστών με συγκεκριμένα ενδιαφέροντα

* Παραδείγματα κανόνων συσχέτισης:

- Το 20% αυτών που αγόρασαν το βιβλίο "Windows 2000" αγόρασαν και το "Word 2000"
- Το 30% αυτών που είδαν τη σελίδα "Special Offers", προχώρησαν σε παραγγελία του DVD "Lord of the Rings".
- Το 80% αυτών που επισκέπτονται το ηλεκτρονικό κατάστημα, μπαίνουν από τη σελίδα "Products".

Sequential pattern discovery (Ανακάλυψη σειριακών μοτίβων)

* Αποτελεί μια επέκταση του association rule mining που ενσωματώνει στον εντοπισμό των patterns **την έννοια του χρόνου**

- Π.χ. το sequential pattern (C) (A,B) δηλώνει ότι οι χρήστες που επισκέπτονται τη σελίδα C ζητούν στο μέλλον να δουν και τις A και B

* Παραδείγματα sequential pattern:

- Το 30% των πελατών που υποβάλλουν μια παραγγελία, επισκέπτονται εντός 10 ημερών τη σελίδα παρακολούθησης τρέχουσας κατάστασης παραγγελίας (order status update)
- Το 45% των νέων πελατών που αγοράζουν ένα κινητό τηλέφωνο, ξοδεύουν περισσότερα από 50 Ευros χρησιμοποιώντας το σε διάστημα 30 ημερών
- Με δεδομένες τις συναλλαγές ενός πελάτη που δεν έχει αγοράσει τίποτα τους τελευταίους 3 μήνες, βρες όλους τους πελάτες με παρόμοια συμπεριφορά



Εξατομίκευση: Περιορισμοί

- * Διασφάλιση προσωπικού **απορρήτου**
 - Ανησυχία για την εμπιστευτικότητα και τη διαφύλαξη των προσωπικών δεδομένων των χρηστών
- * Χαμηλή ανεκτικότητα σε **καθυστερήσεις**
 - Μικροί χρόνοι απόκρισης των διεργασιών mining
 - Μέρος της διαδικασίας πρέπει να γίνεται off-line
- * Αποτίμηση της **αποτελεσματικότητας** της εξατομίκευσης
 - Αξίζει η επένδυση σε χώρο και υπολογιστική ισχύ που απαιτεί η παροχή υπηρεσιών εξατομίκευσης; Βοήθησαν τους χρήστες και ήταν ακριβείς;
- * Πρόβλημα απώλειας ελέγχου / Ενόχληση χρήστη
 - **Μείωση ευχρηστίας**
 - Explicit profiling, λάθος τρόπος παρουσίασης εξατομίκευσης



Συμπεράσματα

- * Το observational personalization είναι ένα **data-intensive task**
- * Δεν υπάρχει έλλειψη δεδομένων
 - **click-stream** δεδομένα συσσωρεύονται με μεγάλους ρυθμούς,
 - **δημογραφικά** δεδομένα μπορούν να βρεθούν,
 - **προφίλ πελατών** είτε υπάρχουν διαθέσιμα είτε μπορούν να βρεθούν
- * Δεν υπάρχει έλλειψη μεθοδολογιών για data analysis
- * Η **ικανότητα εκμετάλλευσης των δεδομένων** αυξάνει με πολύ χαμηλότερο ρυθμό από το ρυθμό αύξησης των διαθέσιμων δεδομένων
- * Τα sites που υποστηρίζουν **personalization πραγματικού χρόνου** είναι ελάχιστα
- * Χαμηλή ανοχή σε καθυστέρηση μεταξύ acquisition και action