IIT KHARAGPUR

NPTEL ONLINE
CERTIFICATION COURSES

E-BUSINESS

PROF. MAMATA JENAMANI
DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING
IIT KHARAGPUR

Week 10: Lecture 1

USER BEHAVIOR MODELING FROM WEB LOG-II

We are going to learn

- A model of browsing behaviour
- Interpreting the model outcome

Properties of the transition probability matrix of a CBMG

- $p_{i1} = 0 \quad 2 \leq i \leq n-1$
  - No transition can be made to the *Entry* state from any state other than the *Exit* state.
- $p_{1n} = 0$
  - No transition can be made from the *Entry* state to the *Exit* state.
- $p_{nj} = 0 \quad 2 \leq j \leq n-1$
  - No transition can be made from the *Exit* state to any state other than the *Entry* state.
- $p_{nn} + p_{n1} = 1$
  - A transition from the *Exit* state to itself or to the *Entry* state.

Statistics that can be derived from the model

- Mean holding time matrix ($Z=[z_{ij}]$)
- Mean time spent in a state

$$\bar{t}_i = \sum_{j=1}^{n} p_{ij} \bar{z}_{ij}$$

  - in the context of the present problem represents the expected time a visitor spends in a state before making a transition
- Limiting state probabilities
  - Probability of customer visiting a state $\lim_{r \to \infty} P^r$

Statistics that can be derived from the model

- Mean number of transitions in the Process

$$\bar{v} = [I - P']^{-1}$$

where,

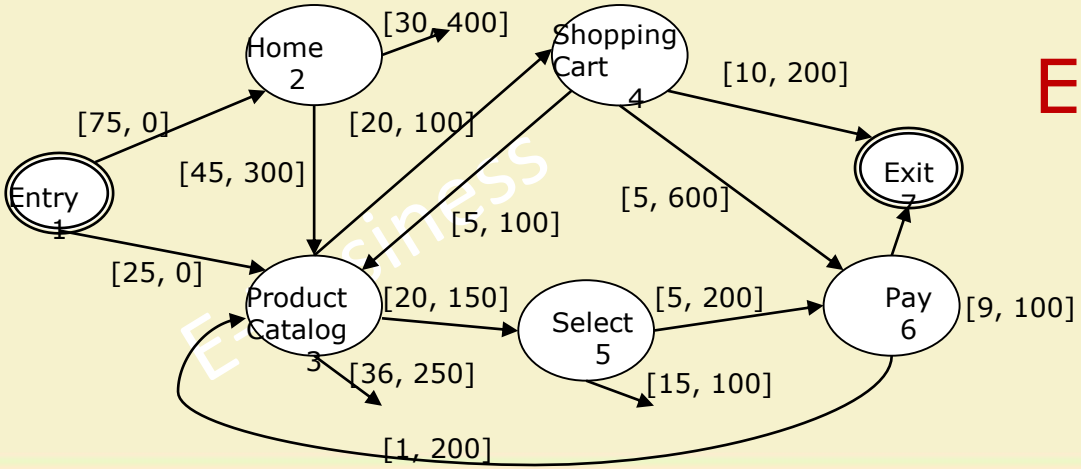
$\bar{v}$ is a set of row vectors with $\bar{v}_i$ indicating mean number of transitions from state *i* to other transient states,

I is the identity matrix, and

P' is the portion of the transition matrix after deleting the row and column associated with the trapping state.

- Average session length= $\sum_{j=1}^{n} \bar{v}_{1j}$
- Mean time spent in the session $\tau_i = \sum_{j=2}^{n} \bar{v}_{ij} \bar{t}_j$

Consider the following graph corresponding to users' navigational pattern in a site. Each node represents a state consisting of one or more pages. *Entry* and *Exit* are two dummy states. A dangling link from a state represents the connection to the *Exit* state. The numbers associated with each branch corresponds to the transition count and accumulated think time (in seconds) between the states *i* and *j* .



Example-I

You can compute the following:

1. Transition probability matrix and mean holding time matrix
2. Mean time spent in a state.
3. Average session length.
4. Average time spend per session.
5. Migration rate from the state *Shopping Cart*.
6. Buy-to-visit ratio

Transition probability matrix

	1	2	3	4	5	6	7	
1		75/100	25/100					1
2			45/75				30/75	1
3				20/76	20/76		36/76	1
4			5/20			5/20	10/20	1
5						5/20	15/20	1
6			1/10				9/10	1
7							1	1

Average holding time matrix

	1	2	3	4	5	6	7	
1								
2			300				400	
3				100	150		250	
4			100			600	200	
5						200	100	
6			200				100	
7								

Grouping the customers

- Clustering the sessions
  - Clustering algorithms (Ex. K-means)
- Understanding user behaviour in each cluster

Example-II

- A site contains five states: *Browse product (b), search product (s), select product, add to cart (a)*, and *pay*
- After clustering the sessions 6 clusters were found and CBMGs are created.
- Analysis of the CBMGs shows the following details.
- Study the details and provide conclusions on customer behavior

Cluster	1	2	3	4	5	6
% of the sessions	44.28	28	10.6	9.29	6.2	1.5
Buy to Visit Ratio	5.7	4.5	3.7	4	3.5	2
Session Length	5.6	15	27	28	50	81
V_a	11	15	21	20	32	50
V_b+V_s	3.6	11.4	20	23	39	70

Week 10: Lecture 2

E-BUSINESS CAPACITY PLANNING

We are going to learn

- Concept of capacity planning
- Steps involved in a typical capacity planning situation
- Understanding the concept through an example

What is Capacity Planning

- Capacity planning is the process of predicting when the future load level will saturate the system and determining the most cost-effective way of delaying system saturation as much as possible.
- Future workload is a function of a combination of three factors
  - Natural evolution of existing workload
  - Deployment of new application and services
  - Changes in customer behavior
- Prediction is the key to capacity planning

Defining the adequate capacity

- Adequate capacity is a function of three following elements
  - Service level agreement
    - Performance
    - Quality of service requirements
    - Scalability
    - Bottleneck analysis
  - Specified technologies and standards
    - Selection of E-Business Infrastructure
  - Cost Constraints
    - Budget constraints

Performance

- Deciding on Lower and upper bounds on performance parameters (e.g. response time, throughput) set by the management
  - Ex. Server side response time should be < 2 sec
  - Throughput > 30,000 requests per sec
- Customers expectation
  - 8 second rule
- Performance Modeling

Quality of service requirements

- Deciding on site availability
- Finding the cost of unavailability
- Determining single point failures
- Determining the minimum configuration possible
- Analysis of Mean time to failure (MTTF) and mean time to repair (MTTR)

Scalability and Bottleneck Analysis

- An infrastructure is said to be scalable if it provides adequate service level even when the work load increases above the expected level.
- Determining maximum throughput
  - Performance modeling
  - Finding the bottleneck resource that limits the performance
  - Workload forecasting
- Growing sites
  - Scaling up and scaling out

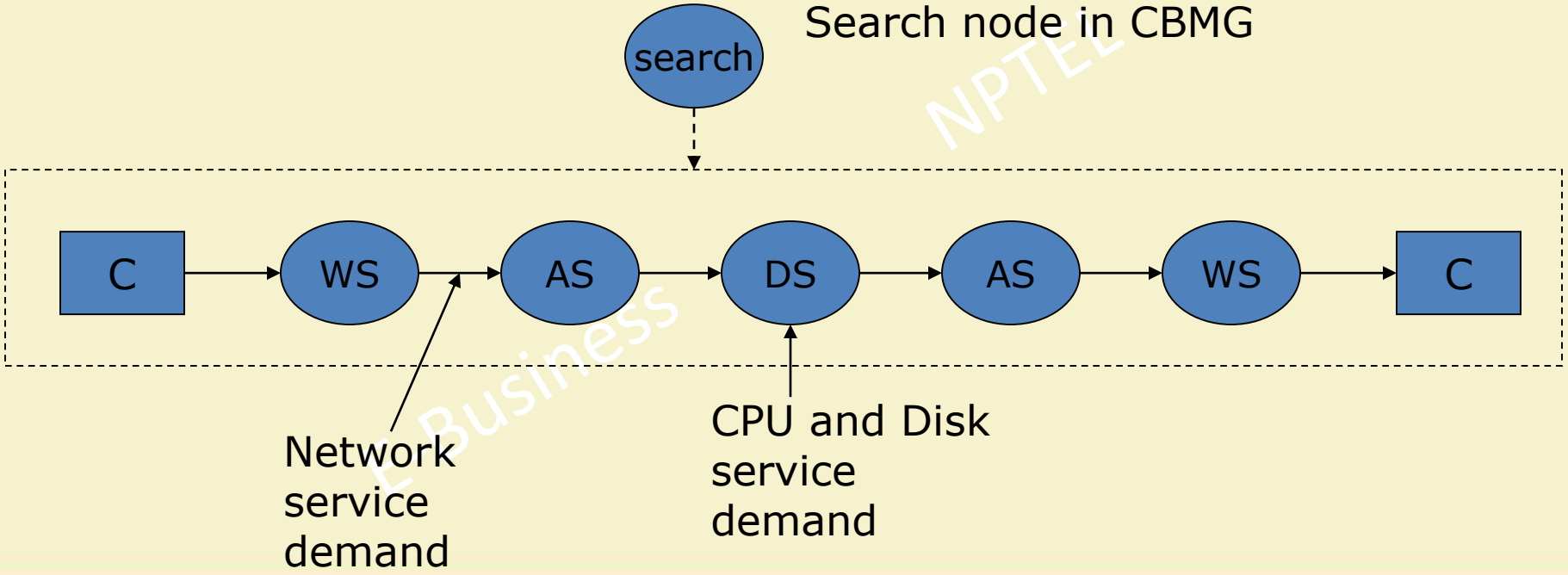
Steps in Capacity Planning

1. Characterizing Customer Behavior
2. Characterizing Site Work Load
3. Workload forecasting
4. Development of Performance Models
5. Obtaining Performance Parameters
6. Workload Forecasting

Characterizing and site workload

- Steps for workload characterization
  - For each of the e-business functions specified in the CBMG draw the CSID (client server interaction diagram)
  - Find out the service demand of each resource based on the service time (sum of service times for each references)
  - Identify the functions which are most demanding (in terms of service demand)
  - Find out the bottleneck resources in each case and the site throughput.
  - Find out the average throughput.

CBMG to CSID



Performance Modeling

- Each resource is to be modeled as a queue and all the resources can be modeled as a queuing network.
  - Analytical Model
  - Simulation Model
- Model parameters like arrival rate and service rate etc. can be obtained from the access log
- Performance Parameters like waiting time, response time and facility utilization etc., can be obtained from the model

Performance modeling concepts

- Single queue approach
  - λ: mean arrival rate (request/sec)
  - μ: mean service rate (request/sec)
  - Probability that zero requests served $p_0=1-(\lambda/\mu)$
  - Probability that k requests served $p_k=(1-(\lambda/\mu))(\lambda/\mu)^k$
  - Server utilization $U=1-p_0=(\lambda/\mu)$
  - Average no of requests per second $\overline{N}=\sum_{k=0}^{\infty} kp_k = U/(1-U)$
  - Average throughput of the server = $X=\mu.U+0.(1-U)=\lambda$
  - Average response time= $R=\overline{N}/X=(U/\lambda)/(1-U)=(1/\mu)/(1-U)=S/(1-U)$

where, $S=(1/\mu)$= *Service time*

An Example – Search function in a site

- At present
  - No of completed search requests = 18,000
  - Measurement interval = 3,600 sec
  - Average time to execute search = 48 ms
  - Utilization = $U=(\lambda/\mu)= \lambda S=(18,000/3600)(0.048)=0.24$
  - Response time = $S/(1-U)$
    $=0.048/(1-0.24)=0.063$ sec

- What is the response time if the number of requests during the peak hour grew by a factor of 4?
  - $\lambda_{new}=4*\lambda=20$ req/sec
  - $U= \lambda S=0.96$
  - R=1.2 sec
    - Increased by 1900%

Week 10: Lecture 3

E-BUSINESS CAPACITY PLANNING-II

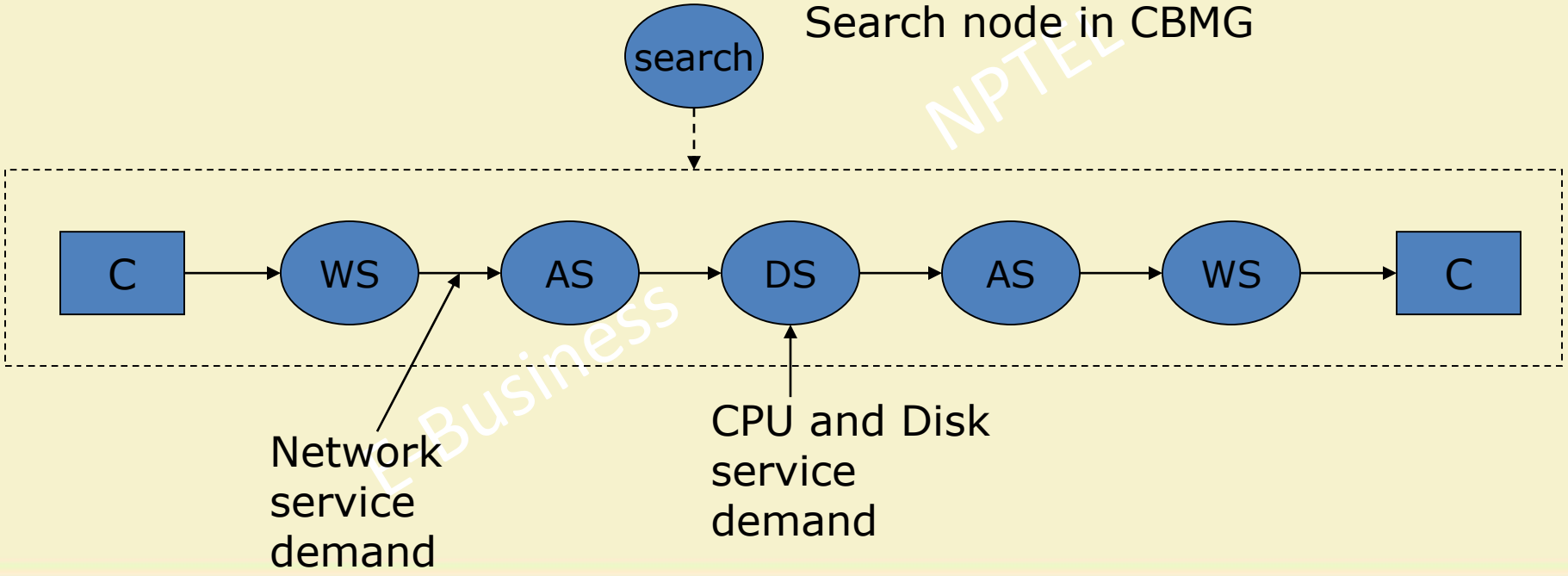
Steps in Capacity Planning

1. Characterizing Customer Behavior
2. Characterizing Site Work Load
3. Workload forecasting
4. Development of Performance Models
5. Obtaining Performance Parameters
6. Workload Forecasting

We are going to learn

- Concept of capacity planning
- Steps involved in a typical capacity planning situation
- Understanding the concept through an example

CBMG to CSID



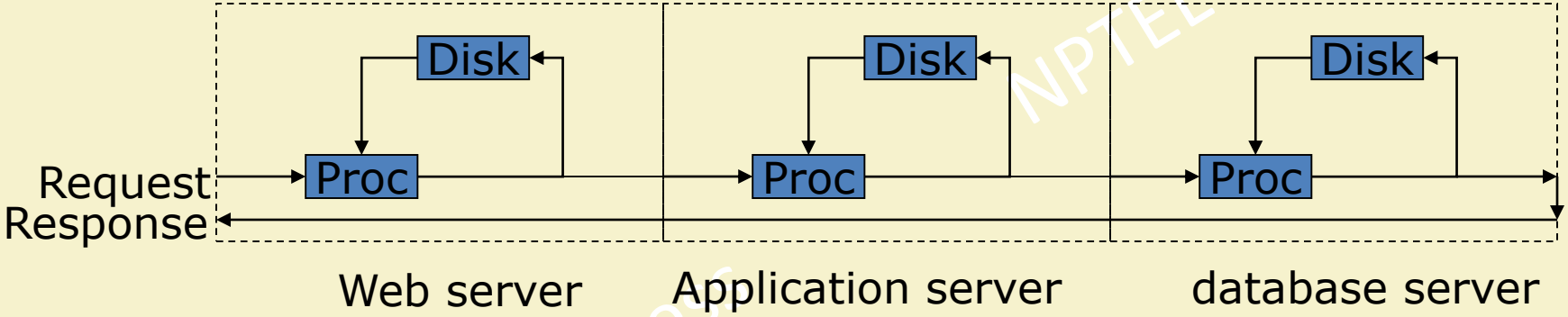
Performance Modeling

- Each resource is to be modeled as a queue and all the resources can be modeled as a queuing network.
  - Analytical Model
  - Simulation Model
- Model parameters like arrival rate and service rate etc. can be obtained from the access log
- Performance Parameters like waiting time, response time and facility utilization etc., can be obtained from the model

Queuing N/W Models

- More realistic
- Depends on the details up to which you would like to model
- Depends on availability of data.

Ex: Queuing N/W of a Site



Workload forecasting

- Visual inspection of data
- Choosing appropriate forecasting models
- Forecasting in terms of each of the business functions
- Using the forecasted values in performance models and workload characterization models
- Planning for the future waves of demand

A case of capacity planning

- A site sales computer components, other electronic products, software and gift items
- Store's revenue consists of merchandize revenue and banner ad revenue
  - 95% of the stores customers do not make any purchase
  - Last fiscal year generated a merchandize revenue of $94, 378, 000 and an ad revenue $900, 000
- During special events the traffic goes up to 400% of the ordinary days.

Planning Situations

- Assessing the Impact of the business goals
  - The board of directors set a goal for the next year: $130 million of merchandize revenue, $ 3 million for ad sales.
  - The company wants to know whether the site has an adequate infrastructure to support the goals without compromising the quality of service.
- Introducing a digitally downloadable product.
  - The company is planning to introduce a new product- digitally downloadable music in MP4 format.
  - Management wants to plan for adequate site capacity before launching the product

Assessing the Impact of the business goals

Metrics derived from CBMG

Metric	Value
$V_{welcome}$	1.172
V_{Browse}	2.583
V_{Search}	2.607
$V_{Register}$	0.115
$V_{Checkout}$	0.046

Metric	Value
$V_{special\_offer}$	0.013
$V_{Add\ to\ cart}$	0.304
V_{Select}	1.608
Avg session length	8.144
Buy to Visit ratio	4.6%

Statistics that can be derived from the model

- Mean number of transitions in the Process

$$V = [I - P']^{-1}$$

where,

V is a set of row vectors with V_i indicating mean number of transitions from state *i* to other transient states,

I is the identity matrix, and

P' is the portion of the transition matrix after deleting the row and column associated with the trapping state.

- Average session length= $\sum_{j=1}^{n} V_{1j}$
- Mean time spent in the session $\tau_i = \sum_{j=2}^{n} V_{ij}\,\overline{t}_j$

Finding the expected workload on the server (For Merchandise)

- How many purchases are required to generate the specified revenue (i.e., $130 million)?
- With this many number of purchase how many unique sessions are required?
  - Revenue throughput ($ earned / second) = $ 130 million/(second / year) = (13,000,000/31, 536,000)=4.122
  - Observed value of Average sales =$225
  - Revenue throughput ($ earned / second) = (Sessions/sec) * BV * Average sales
  - (Sessions/sec) = 4.122/(0.046*225)=0.398 sessions/sec
- How many requests will be made to the server (system throughput)
  - Expected system throughput = (number of sessions) * (average session length) = 0.398*8.144 = 3.241 transactions/sec
- Observed traffic burst = 20 times of the normal load
- Site has to serve 20*3.241 = 64.82 requests per second
- Is your infrastructure adequate?
- If not plan for the infrastructure before realizing the business goal.

Find out performance parameters

- Expected system throughput can be considered as the arrival rate for the performance model
  - queuing model obtained from the CSID
- With this arrival rate and keeping service times constant find performance parameters: waiting time, facility utilization, bottleneck element etc.
- Are these parameters satisfying the quality of service requirements?
- Now decide what measures should be taken to satisfy the quality of service requirement.

Week 10: Lecture4

INTRODUCTION TO RECOMMENDER SYSTEM

We are going to learn

- Introductory concepts
- Framework for understanding recommender system

What do recommender systems do?

- Recommender systems suggest items (information, products and services) that are of interest to the users based on customer demographic, features of the items, user preferences (e.g., rating or purchase history) etc.
- Ex. Book recommendation, movie recommendation

Advantages

- Effective Information management
  - Decreased information overload
- Better customer relationship management
  - Increased sales
  - Loyal customers
- Important in e-commerce sites
  - Effective Management of virtual customers
  - Personalized Service for the customers

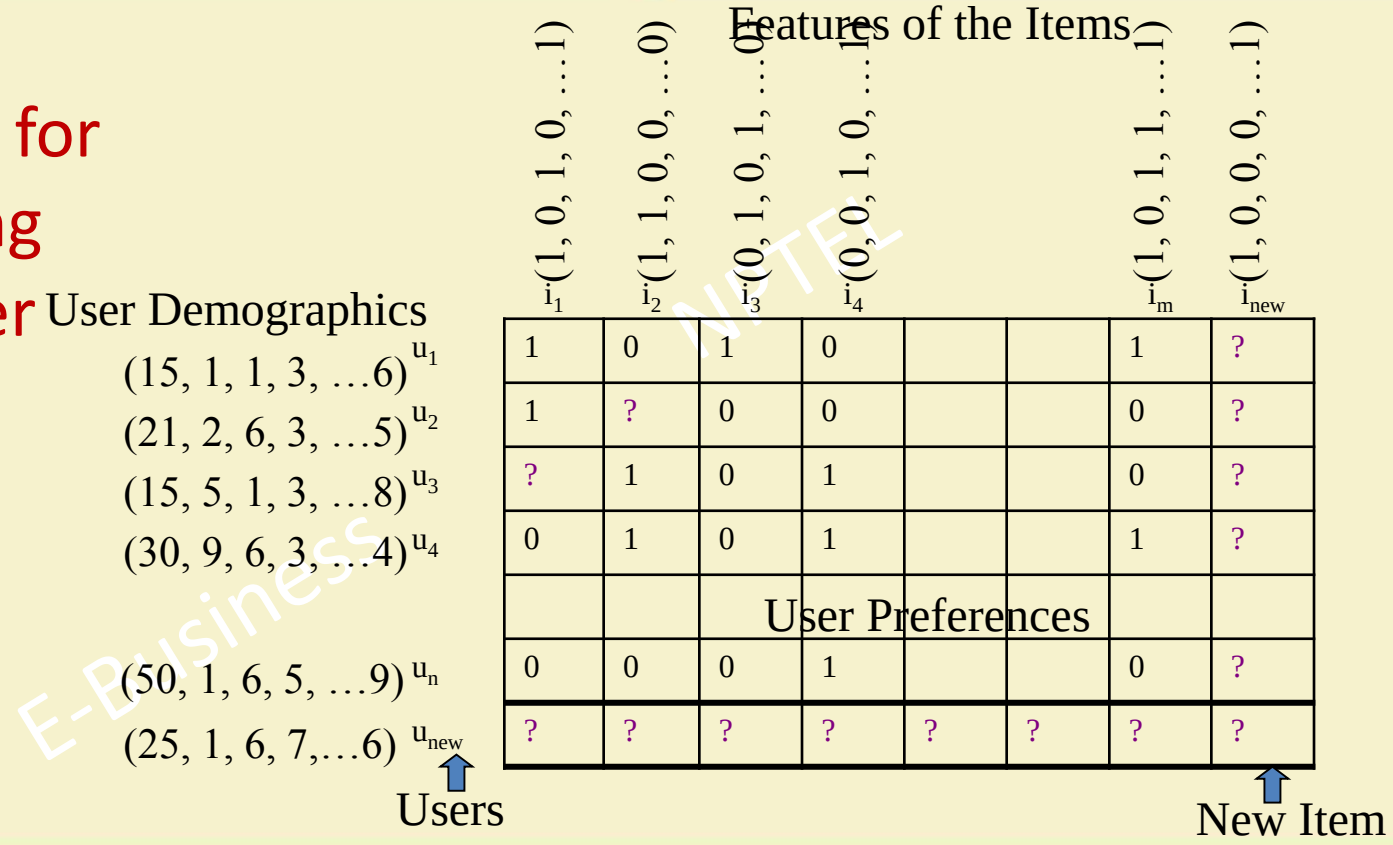
Some early research initiatives

- Social information filtering: algorithms for automating "word of mouth"
  - Recommending music albums
  - Upendra Shardanand and Pattie Maes
- GroupLens
  - Personalized recommendations for usenet news items
  - Resnick, Konstan
- Citeseer
  - Recommendation for relevant articles
  - Rucker and Pollano

Some early commercial implementations

- Double-click.com
  - Personalized banner ads
- CDnow.com
  - Music albums
- Amazon.com
  - Books

A framework for understanding recommender systems



Elements in a typical recommendation scenario

- Users U (u_1, u_2 … u_n)
  - Each user is associated with his demographic data
  - Each user has a list of items I_{ui} ($I_{ui} \in I$) on which a user has expressed his preferences
  - The preferences of user u_i on the item i_j (denoted as p_{ij}) is a subjective rating explicitly stated by the user or an implicit measure inferred from the purchase, navigation, browsing and searching pattern of the user.
- Items I (i_1, i_2 … i_n)
  - Each item is associated with a set of features

Types of recommendation decisions

- Prediction
  - Predicting preferences for an item $i_j \notin I_{ua}$ for an active user u_a
  - Can be further classified as personalized or non-personalized
- Top-N recommendations
  - Recommending a list of N items, $I_r \subset I$, that the active user u_a will like most. Recommended list must be on the items not already rated or chosen by u_a, that is $I_r \cap I_{ua} = \varnothing$
  - Can be further classified as personalized or non-personalized
- Top-M users
  - Recommending a list of M users for a newly available item i_{new} who will value i_{new} most.

Types of recommendation systems

- Popularity based
- Content based
- Collaborative filtering
- Association based
- Demographics based
- Reputation based
- Hybrid of the above

Popularity based recommendation system

- Recommending most popular items within a community
- Information used
  - User preferences
- Type of recommendation decision
  - Top-N recommendations
    - Popularity measures: Percentage of users who purchased the item, average rating for the item etc.
- Non-personalized
- Simple and efficient

Content based recommendation system

- Content based information filtering
  - Degree of relevance to a particular user of an item determined by its content
- Information used
  - Features of items
  - Individual user preferences
- Type of recommendation decision
  - Prediction
  - Top-N recommendations
  - Top-M users
- Personalized

Collaborative filtering recommendation system

- User-to-user correlation based on taste
  - Social information filtering
- Information used
  - User preferences
- Type of recommendation decision
  - Prediction
  - Top-N recommendations
- Personalized

Association based recommendation system

- Recommending the items that can be purchased with the items that a user has purchased in the past or shown interest to purchase
  - Co-occurrences of items that the users frequently preferred to purchase together
- Information used
  - Feature of items
- Type of recommendation decision
  - Prediction
  - Top-N recommendations
- Personalized

Demographics based recommendation system

- User-to-user correlation based on demographics
- Information used
  - Individual user preferences
  - Features of the items
- Type of recommendation decision
  - Prediction
  - Top-N recommendations
  - Top-M users
- Personalized

Reputation based recommendation system

- Identifying the users that a user respects and then using the opinion of these selected individuals for generating recommendation
- Information used
  - User preferences
  - Reputation matrix
- Type of recommendation decision
  - Top-N recommendations
  - Prediction
- Personalized

Week 10: Lecture 5

CONTENT BASED RECOMMENDER SYSTEM

We are going to learn

- Content based recommender system

The approach

- Recommends the items similar to those liked by the user in the past
- Automatic learning and adaptively changing the user profile
  - Relevance feedback
- Particularly popular for document recommendation
- Selection of item features, and document classification based on the content
  - Typically if the items are textual documents the process belongs to a discipline called *text mining*

Phases of content based recommendation generation

- Feature extraction and selection
- Representation
- User profile learning
- Recommendation generation

Feature extraction and selection

- The method depends on the type of the item under consideration
  - Feature Specification: Extrinsic Features
    - Ex: For Movies: category, MPAA rating, Maltin rating, Academy award, length, Origin, Director etc.
  - Feature Extraction: Intrinsic Features
    - Ex: Text documents: extraction of nouns and noun phrases
- Choosing minimum number of features that are necessary and sufficient to describe the items
- Techniques used for extrinsic feature: statistical analysis (forward and backward stepwise multiple regression), genetic algorithm, attribute selection measures in decision trees induction etc.

Feature extraction and selection

- Intrinsic features from text documents
  - Hundreds or thousands of features
    - Other feature selection methods applicable to items with extrinsic features, become intractable hence infeasible
  - Use of an evaluation function that is applied to the features independently, for example
    - TF (within-document term frequency)
    - TF*IDF (within-document term frequency * Inverse-document frequency)
  - Top k features with highest score are used for the selection

Representation

- Preparing training data set
  - Each item in the training data set is labeled to indicate its preference (dependent variable) by a particular user and assigned a value to each feature (independent variable)
- Independent variables
  - Assigning values to each feature
    - Binary, continuous
    - deciding on scale
- Dependent variable
  - Preference rating by a particular user

User profile learning

- Establishing relationship between preference scores and item features
- Adaptive Learning algorithms
  - Multiple linear regression model
  - Decision tree induction algorithm
  - Back propagation neural networks
- Training the model

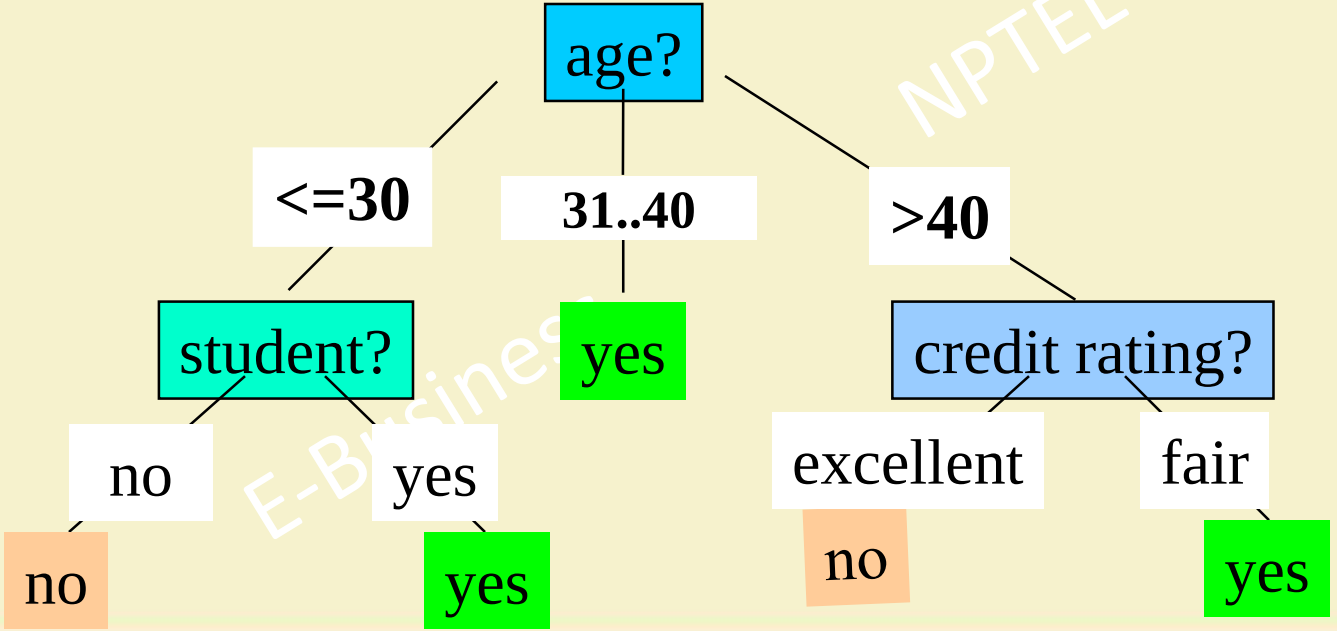
Recommendation generation

- Prediction and top-N recommendation
  - Use the user profile model for prediction
- Top-M users
  - For the new item find out the preference score of all the users and choose top-M users
- Observe the users action and retrain the learning algorithm if required

Example of a decision tree induction algorithm

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

The decision tree



Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)
- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

Attribute Selection Measure: Information Gain

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in data partition D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using attribute A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Attribute Selection: Information Gain

- Class P: buys_computer = "yes" (9 tuples)
- Class N: buys_computer = "no" (5 tuples)

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

$\frac{5}{14} I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's.

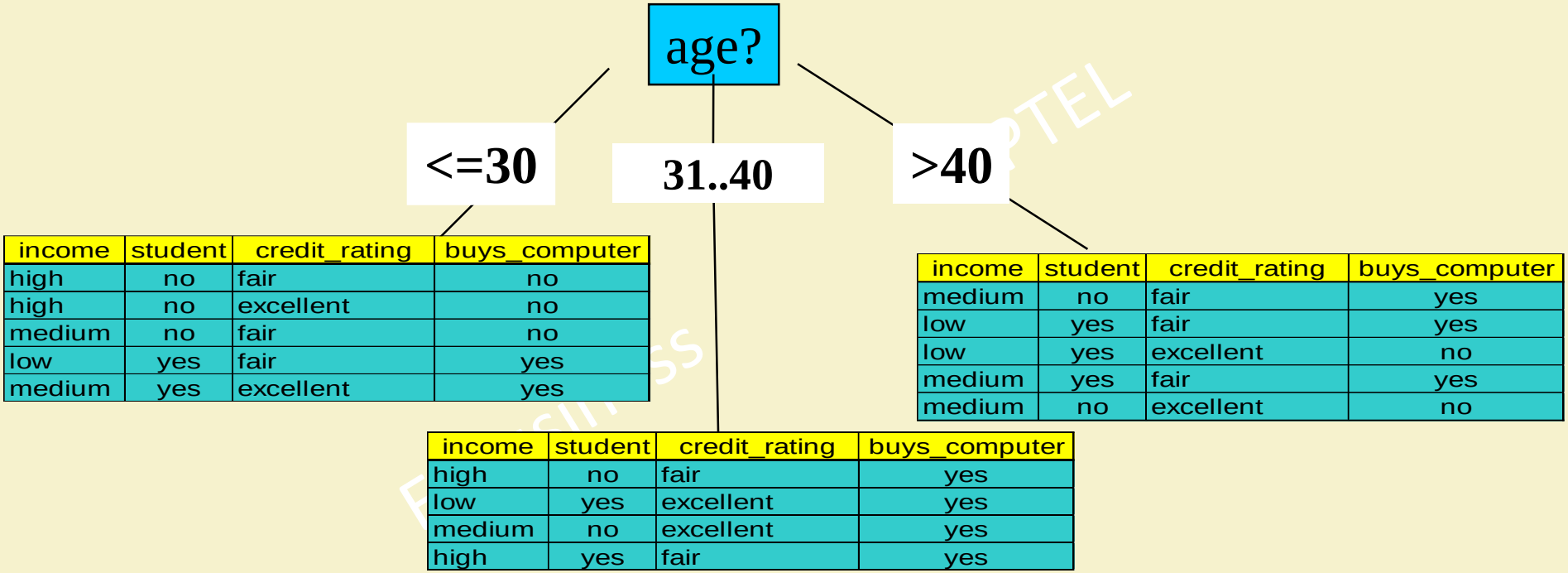
Hence $Gain(age) = Info(D) - Info_{age}(D) = 0.246$

Similarly, $Gain(income) = 0.029$

$Gain(student) = 0.151$

$Gain(credit\_rating) = 0.048$

Result of the partition based on the attribute “age”



income	student	credit_rating	buys_computer
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit_rating	buys_computer
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

income	student	credit_rating	buys_computer
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

Assignment

- Draw a decision tree corresponding to a customer of a online song store.
- A new song from an English album of classical songs has come. Should you include this customer in your target customer list?
- A new rock song from a Telugu movie has come. Should you suggest this too.

Category	Type	Language	Purchased
Movie	Classical	Hindi	Yes
Movie	Classical	Hindi	Yes
Album	Rock	English	No
Movie	Classical	English	Yes
Album	Classical	Hindi	Yes
Movie	Classical	Telugu	No
Album	Rock	Hindi	No
Movie	Rock	English	No
Movie	Classical	English	Yes

End of Week 10