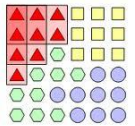
Data Mining and Data Warehousing

Chapter 6

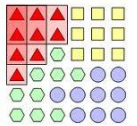
Mining Complex Data Mining

Instructor: Suresh Pokharel



What is IR?

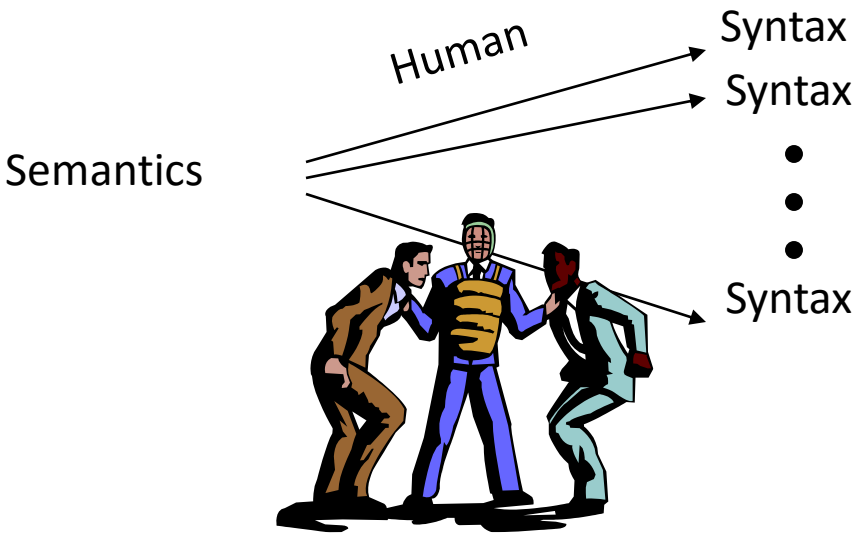
What is IR?



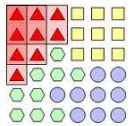
- Broad sense: Finding needed information
  - Information: textual, audio, image, video,…
  - Retrieval: search, browse, re-organize, …
- Narrow sense: Search in text database
  - There exists a collection of text documents
  - User gives a query to express the information need
  - A retrieval system returns relevant documents to users
- Also called "text/document retrieval"

Why the difficulty…?

Language: Many ways to express the same meaning.



Conversely…

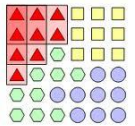


And, conversely, reading/hearing a phrase or sentence or question how to interpret the meaning?

Syntax —Computer - extremely hard problem→ Semantics

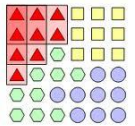
In fact, how do we define or represent meaning?!

What kinds of information are there?



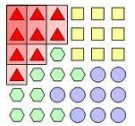
- Text
  - books, periodicals, WWW, memos, ads
  - published
- Photos, other Images
- Broadcast TV, Radio
- Telephone Conversations
- Databases
- Web pages

How much information is there?



Gigabyte	10^9 bytes	1000 megabytes
Terabyte	10^12 bytes	1000 gigabytes
Petabyte	10^15 bytes	1000 terabytes
Exabyte	10^18 bytes	1000 petabytes

Data Retrieval – Databases.



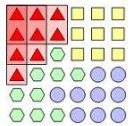
Database – structured repository of data.

Therefore, structured queries with *exact* response.

select name from students where score > 60;

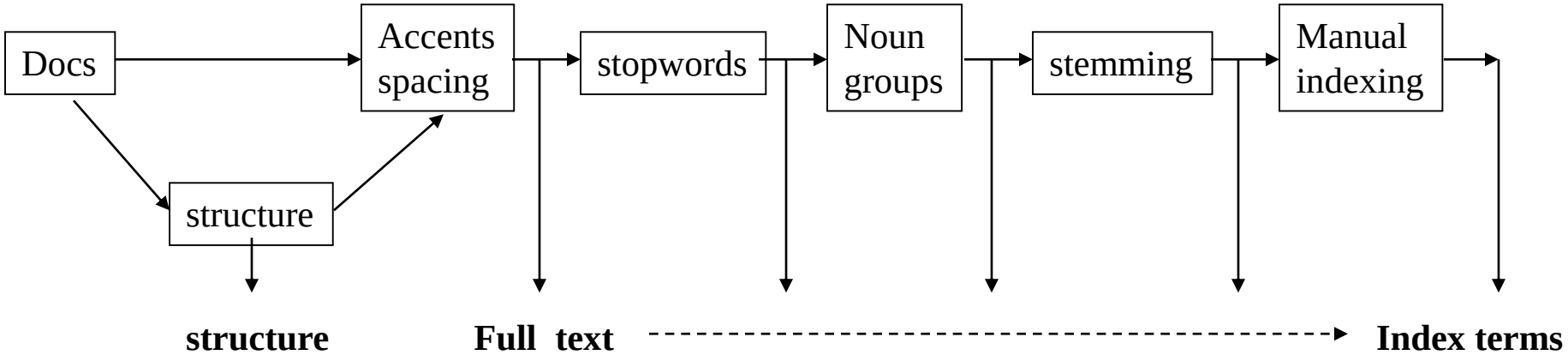
name	id	score
Joe Smith	3456	57
Nancy Parks	1244	63

Goal of IR



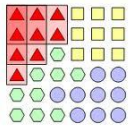
Retrieve exactly the set of all documents relevant to the user query.

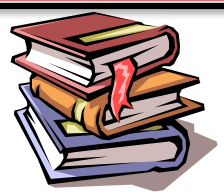
Logical view of the documents



- Document representation viewed as a continuum: logical view of docs might shift

Relevance Feedback Architecture





Query String

Document corpus

Revised Query

IR System

ReRanked Documents

1. Doc2
2. Doc4
3. Doc5
.
.

Query Reformulation

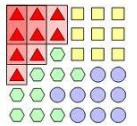
Ranked Documents

1. Doc1
2. Doc2
3. Doc3
.
.

1. Doc1 ⇓
2. Doc2 ⇑
3. Doc3 ⇓
.
.

Feedback

The Vector-Space Model

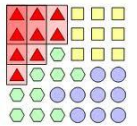


- Assume *t* distinct terms remain after preprocessing; call them index terms or the vocabulary.
- These "orthogonal" terms form a vector space.

  Dimension = *t* = |vocabulary|
- Each term, *i*, in a document or query, *j*, is given a real-valued weight, $w_{ij.}$
- Both documents and queries are expressed as t-dimensional vectors:

  $d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$

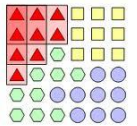
Document Collection



- A collection of *n* documents can be represented in the vector space model by a term-document matrix.
- An entry in the matrix corresponds to the "weight" of a term in the document; zero means the term has no significance in the document or it simply doesn't exist in the document.

$$\begin{array}{c|cccc} & T_1 & T_2 & \ldots & T_t \\ D_1 & w_{11} & w_{21} & \ldots & w_{t1} \\ D_2 & w_{12} & w_{22} & \ldots & w_{t2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & w_{1n} & w_{2n} & \ldots & w_{tn} \end{array}$$

Term Weights: Term Frequency



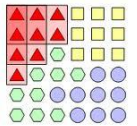
- More frequent terms in a document are more important, i.e. more indicative of the topic.

  f_{ij} = frequency of term *i* in document *j*

- May want to normalize *term frequency* (*tf*) by dividing by the frequency of the most common term in the document:

  $tf_{ij} = f_{ij} / max_i\{f_{ij}\}$

Term Weights: Inverse Document Frequency



- Terms that appear in many *different* documents are *less* indicative of overall topic.

  df_i = document frequency of term i

  = number of documents containing term i

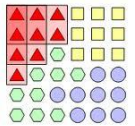
  idf_i = inverse document frequency of term i,

  $= \log_2 (N/ df_i)$

  (N: total number of documents)

- An indication of a term's *discrimination* power.
- Log used to dampen the effect relative to *tf*.

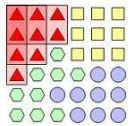
TF-IDF Weighting



- A typical combined term importance indicator is *tf-idf weighting*:

$$w_{ij} = tf_{ij}\, idf_i = tf_{ij} \log_2 (N/ df_i)$$

- A term occurring frequently in the document but rarely in the rest of the collection is given high weight.
- Many other ways of determining term weights have been proposed.
- Experimentally, *tf-idf* has been found to work well.



Key words are generated using *tf-idf on BOW*

$$tf \times idf = \text{tf} \times \log \frac{\text{N}}{\text{n}_\text{i}}$$

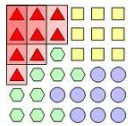
Where,

tf = Term Frequency

idf = Inverse Document Frequency

N = Total number of documents

n_i= (df_i) no of documents in which the term occur



e.g.

- ✓ *Number of words in document doc-1 = 100*
- ✓ *Number of word "cow" = 3*
- ✓ *tf = 3/100 = 0.03*
- ✓ *Number of documents = 1000*
- ✓ *Number of documents containing word "cow"= 200*
- ✓ *idf = log(1000/200) = 0.699*
- ✓ ***tf x idf = 0.03 * 0.699 = 0.021***

Basic Measures for Text Retrieval: Precision and Recall

Precision: This is the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses). It is formally defined as

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}.$$

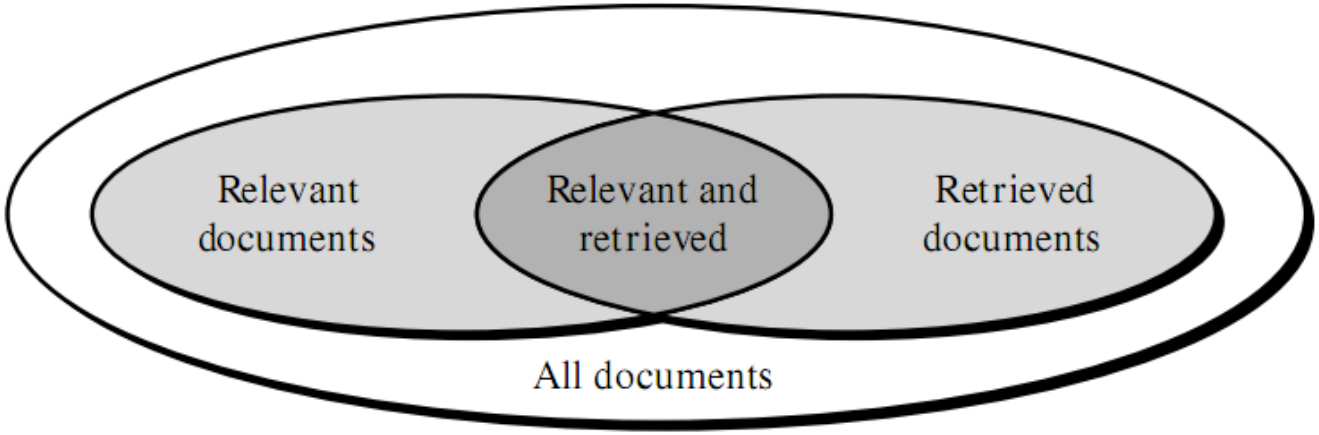
Recall: This is the percentage of documents that are relevant to the query and were, in fact, retrieved. It is formally defined as

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}.$$

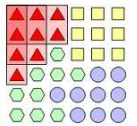
Basic Measures for Text Retrieval: Precision and Recall

F-Score: An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used trade-off is the F-score, which is defined as the harmonic mean of recall and precision:

$$F\_score = \frac{recall \times precision}{(recall + precision)/2}.$$

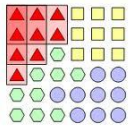


Relationship between the set of relevant documents and the set of retrieved documents



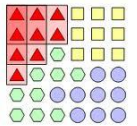
Web Mining

Web Mining – The Idea



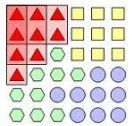
- In recent years the growth of the World Wide Web exceeded all expectations. Today there are several billions of HTML documents, pictures and other multimedia files available via internet and the number is still rising. But considering the impressive variety of the web, retrieving interesting content has become a very difficult task.

Web Mining



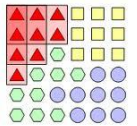
- Web is the single largest data source in the world
- Due to heterogeneity and lack of structure of web data, mining is a challenging task
- Multidisciplinary field:
  - data mining, machine learning, natural language processing, statistics, databases, information retrieval, multimedia, etc.

Web Mining



- Web mining is the use of data mining techniques to automatically discover and extract information from Web documents/services

Opportunities and Challenges

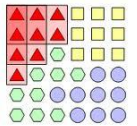


- Web offers an unprecedented opportunity and challenge to data mining
  - The amount of information on the Web is huge, and easily accessible.
  - The coverage of Web information is very wide and diverse. One can find information about almost anything.
  - Information/data of almost all types exist on the Web, e.g., structured tables, texts, multimedia data, etc.
  - Much of the Web information is semi-structured due to the nested structure of HTML code.
  - Much of the Web information is linked. There are hyperlinks among pages within a site, and across different sites.
  - Much of the Web information is redundant. The same piece of information or its variants may appear in many pages.

The 14th International World Wide Web Conference (WWW-2005), May 10-14, 2005, Chiba, Japan

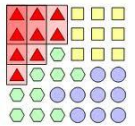
Web Content Mining

Opportunities and Challenges



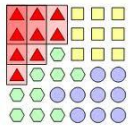
- The Web is noisy. A Web page typically contains a mixture of many kinds of information, e.g., main contents, advertisements, navigation panels, copyright notices, etc.
- The Web is also about services. Many Web sites and pages enable people to perform operations with input parameters, i.e., they provide services.
- The Web is dynamic. Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues.
- Above all, the Web is a virtual society. It is not only about data, information and services, but also about interactions among people, organizations and automatic systems, i.e., communities.

Web Mining



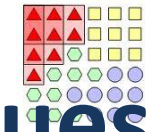
- The term created by Orem Etzioni (1996)
- Application of data mining techniques to automatically discover and extract information from *Web data*

Data Mining vs. Web Mining



- Traditional data mining
  - data is structured and relational
  - well-defined tables, columns, rows, keys, and constraints.
- Web data
  - Semi-structured and unstructured
  - readily available data
  - rich in features and patterns

Classification of Web Mining Techniques



- Web Content Mining
- Web-Structure Mining
- Web-Usage Mining

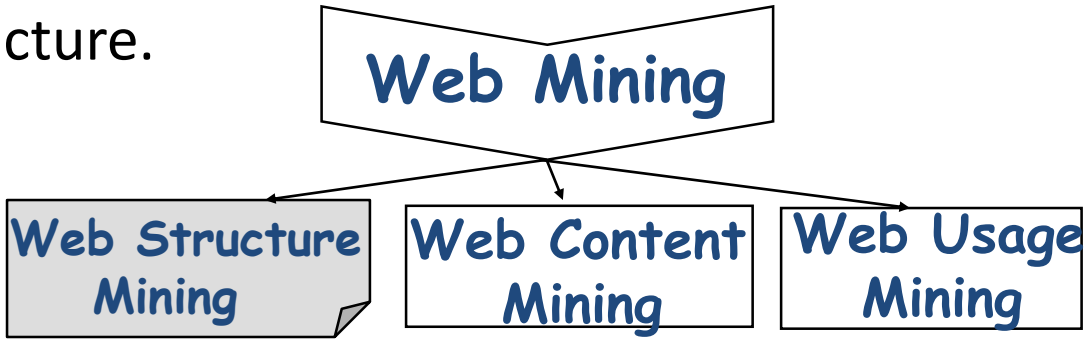
Web-Structure Mining

- Generate *structural summary* about the Web site and Web page

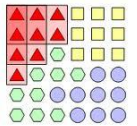
Depending upon the hyperlink, 'Categorizing the Web pages and the related Information @ inter domain level

Discovering the Web Page Structure.

Discovering the nature of the hierarchy of hyperlinks in the website and its structure.



Web-Structure Mining cont…



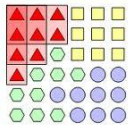
- Finding Information about web pages

  Retrieving information about the relevance and the quality of the web page.
  Finding the authoritative on the topic and content.

- Inference on Hyperlink

  The web page contains not only information but also hyperlinks, which contains huge amount of annotation.
  Hyperlink identifies author's endorsement of the other web page.

Web-Structure Mining cont…



- More Information on Web Structure Mining

Web Page Categorization. (Chakrabarti 1998)

Finding micro communities on the web
e.g. Google (Brin and Page, 1998)

Schema Discovery in Semi-Structured Environment.

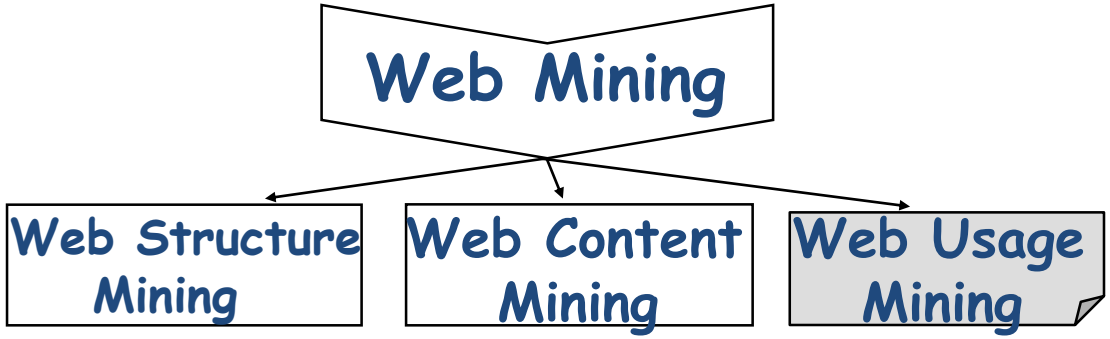
Web-Usage Mining

- What is Usage Mining?

Discovering user 'navigation patterns' from web data.

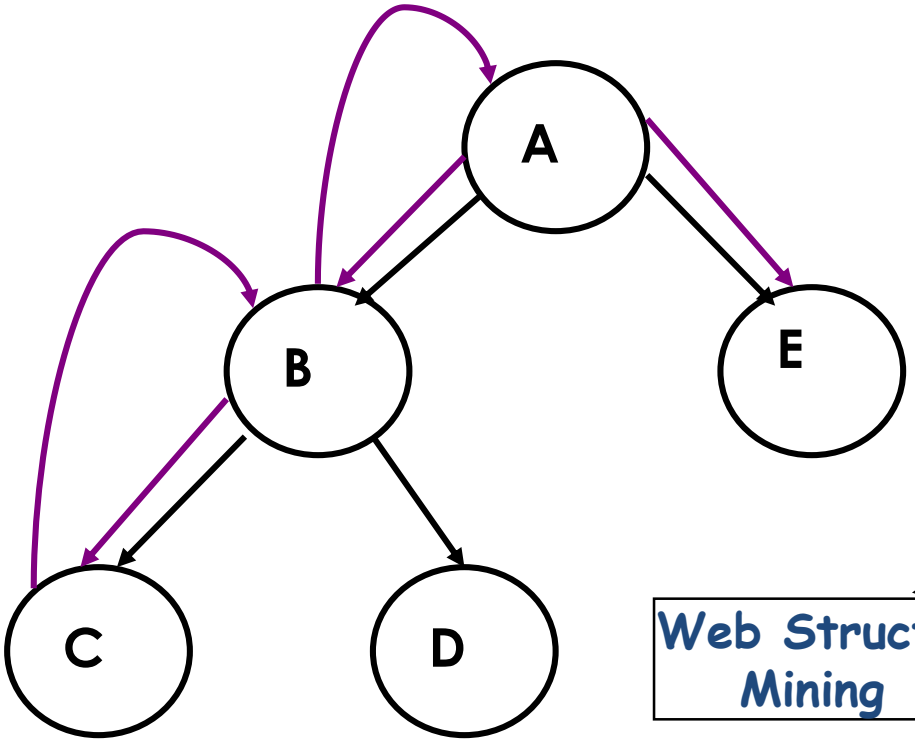
Prediction of user behavior while the user interacts with the web.

Helps to Improve large Collection of resources.

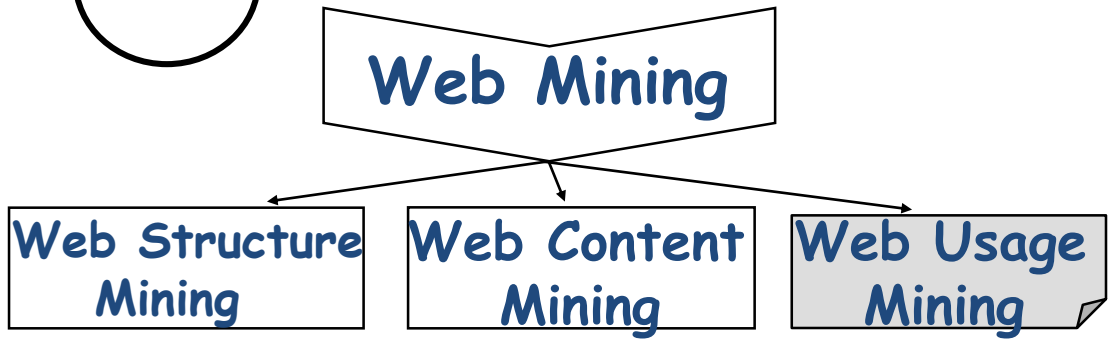


Web-Usage Mining cont…

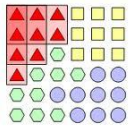
- Data Mining Techniques – Navigation Patterns



Web Page Hierarchy of a Web Site



Web-Usage Mining cont…



- Data Mining Techniques – Navigation Patterns

Analysis:

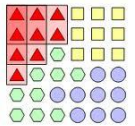
Example:

70% of users who accessed /company/product2 did so by starting at /company and proceeding through /company/new, /company/products and company/product1

80% of users who accessed the site started from /company/products

65% of users left the site after four or less page references

Web-Usage Mining cont…



- Data Mining Techniques – Sequential Patterns

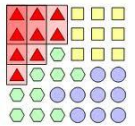
Example:

Supermarket

Cont…

Customer	Transaction Time	Purchased Items
John	6/21/05 5:30 pm	Beer
John	6/22/05 10:20 pm	Brandy
Frank	6/20/05 10:15 am	Juice, Coke
Frank	6/20/05 11:50 am	Beer
Frank	6/20/05 12:50 am	Wine, Cider
Mary	6/20/05 2:30 pm	Beer
Mary	6/21/05 6:17 pm	Wine, Cider
Mary	6/22/05 5:05 pm	Brandy

Web-Usage Mining cont…



- Data Mining Techniques – Sequential Patterns

Example:

Supermarket

Cont…

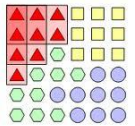
Customer Sequence

Customer	Customer Sequences
John	(Beer) (Brandy)
Frank	(Juice, Coke) (Beer) (Wine, Cider)
Mary	(Beer) (Wine, Cider) (Brandy)

Mining Result

Sequential Patterns with Support >= 40%	Supporting Customers
(Beer) (Brandy)	John, Frank
(Beer) (Wine, Cider)	Frank, Mary

Web-Usage Mining cont…



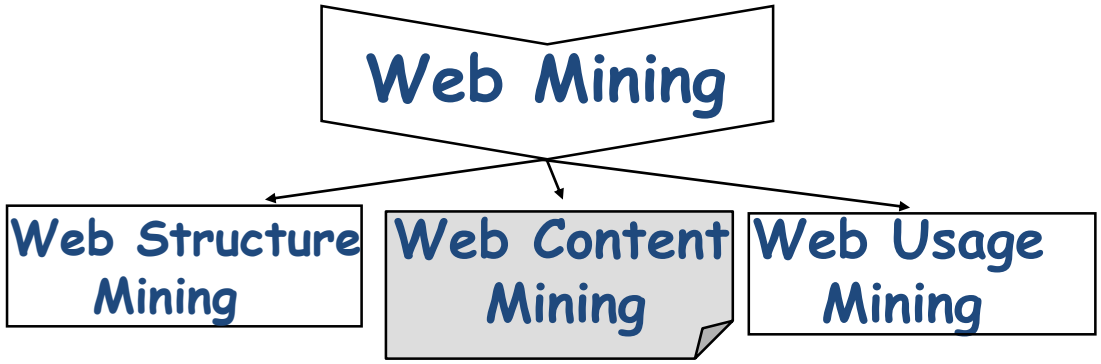
- Data Mining Techniques – Sequential Patterns

Web usage examples

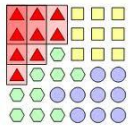
- In Google search, within past week 30% of users who visited /company/product/ had ‘camera’ as text.
- 60% of users who placed an online order in /company/product1 also placed an order in /company/product4 within 15 days

Web Content Mining

- ***'Process of information'*** or resource discovery from content of millions of sources across the World Wide Web
  - E.g. Web data contents: text, Image, audio, video, metadata and hyperlinks
- Goes beyond key word extraction, or some simple statistics of words and phrases in documents.



Document Clustering



- Unsupervised Learning : a data set of input objects is gathered
- Goal : Evolve measures of similarity to cluster a collection of documents/terms into groups within which similarity within a cluster is larger than across clusters.
- Hypothesis : Given a `suitable' clustering of a collection, if the user is interested in document/term *d/t*, he is likely to be interested in other members of the cluster to which *d/t* belongs.
- Hierarchical
  - Bottom-Up
  - Top-Down
- Partitional

Association

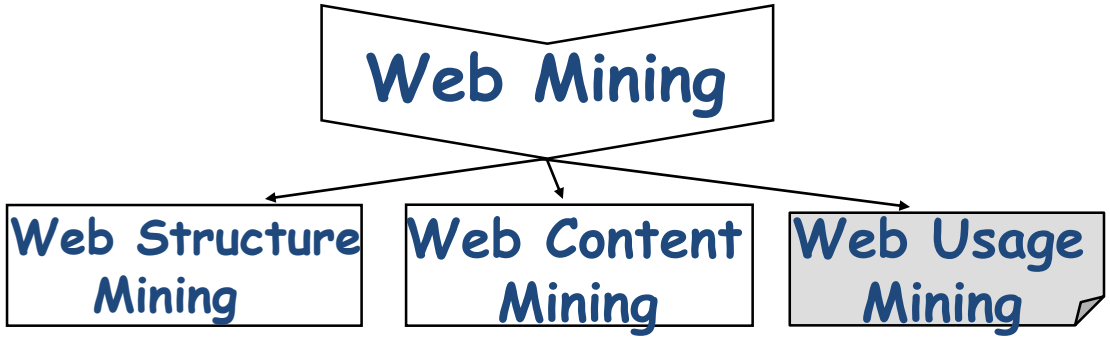
Example: Supermarket

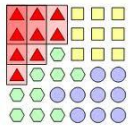
Transaction ID	Items Purchased
1	butter, bread, milk
2	bread, milk, beer, egg
3	diaper
…	……

- **An association rule can be**

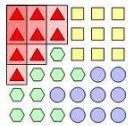
"If a customer buys milk, in 50% of cases, he/she also buys beers. This happens in 33% of all transactions.

50%: confidence
33%: support

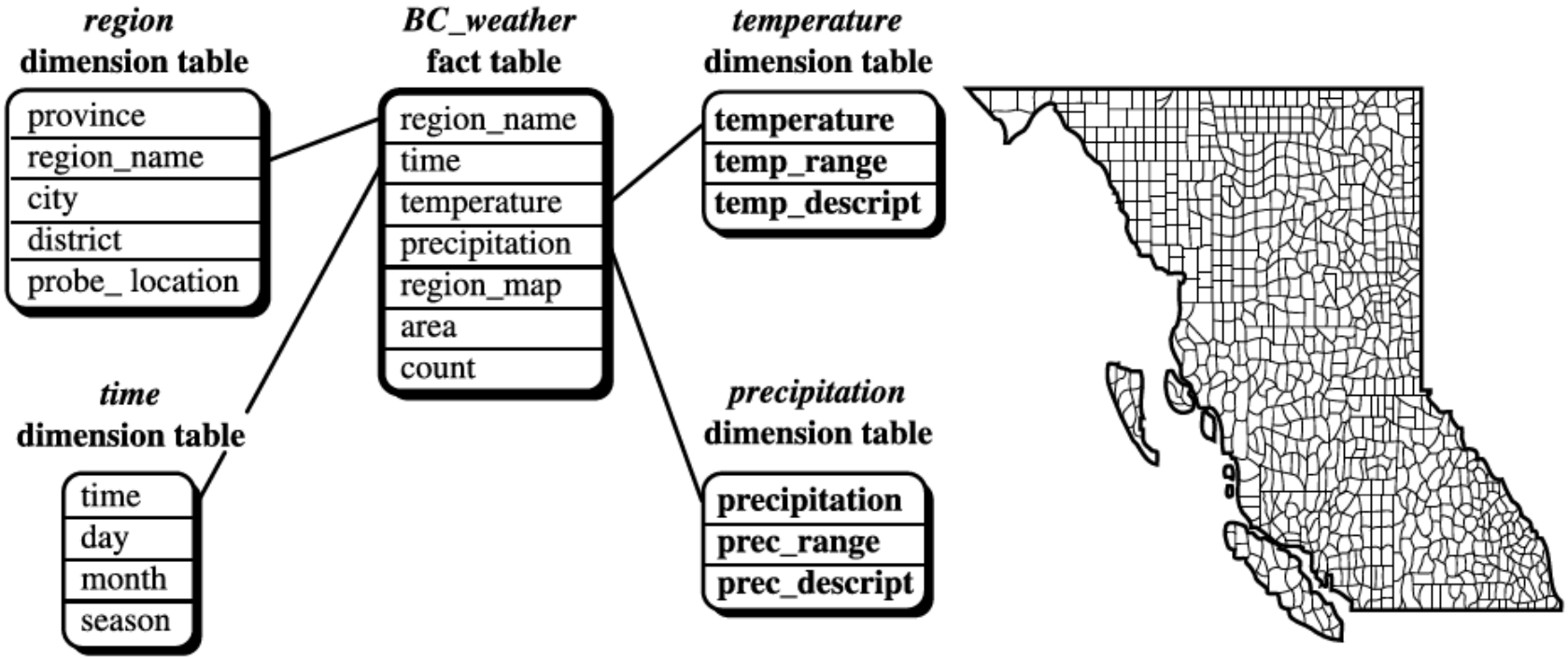




Spatial Data Mining



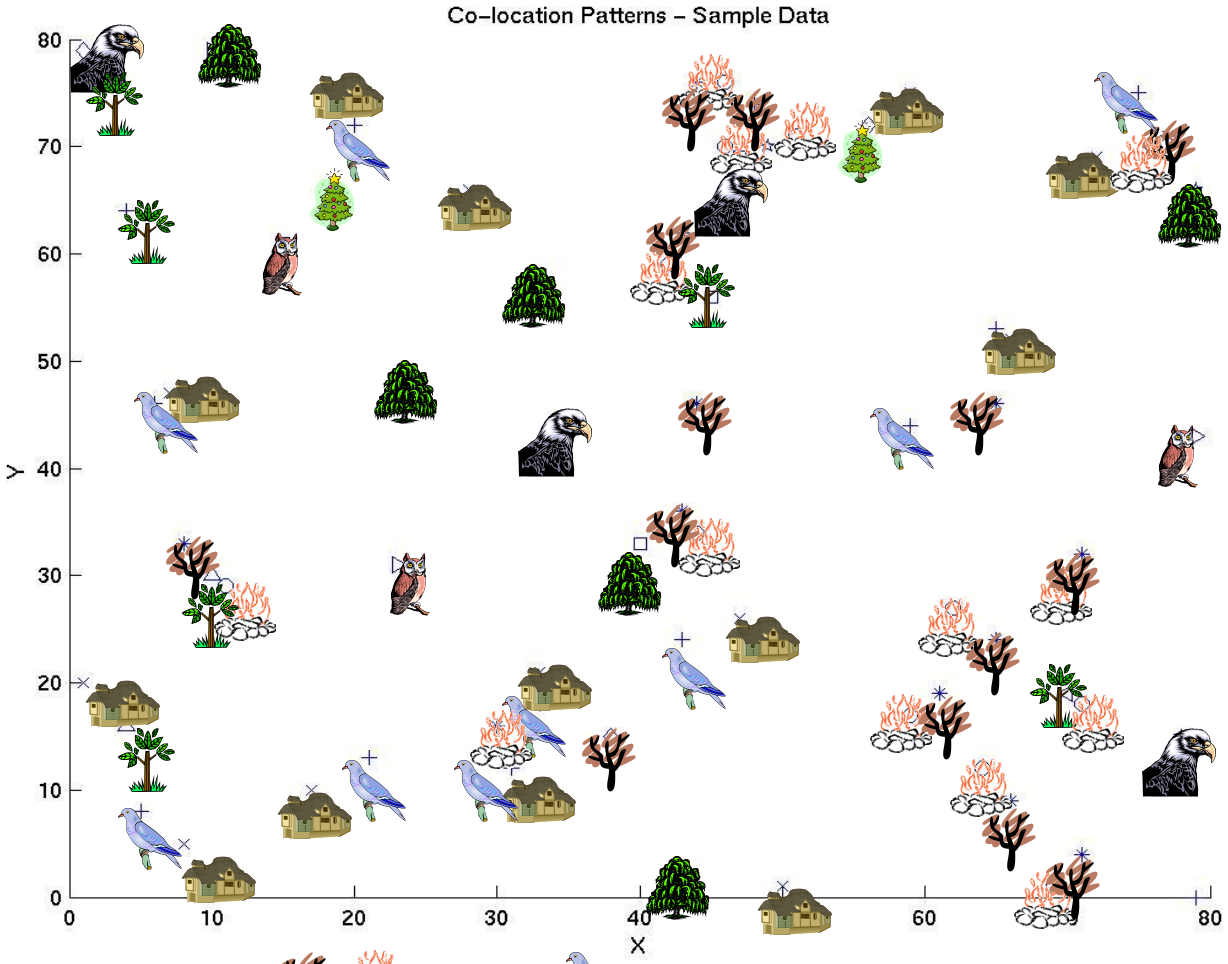
- A spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI chip layout data.
- Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases.
- **Spatial data mining** is the process of discovering interesting, useful, non-trivial patterns from large spatial datasets
- Such mining demands an integration of data mining with spatial database technologies. It can be used for understanding



A star schema of the *BC_weather* spatial data warehouse and corresponding BC weather probes map.

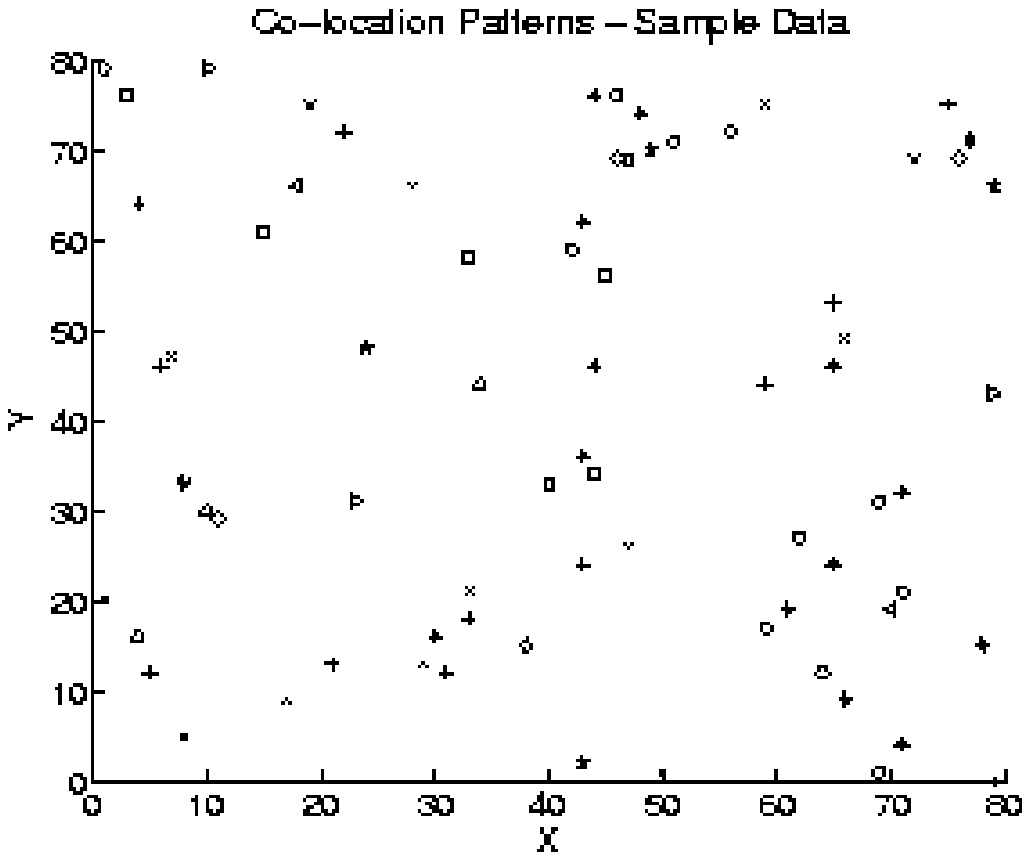
Associations, Spatial associations, Co-location

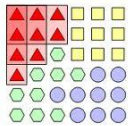
find patterns from the following sample dataset?



Answers: and

Co-location Rules – Spatial Interest Measures



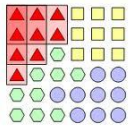


• A spatial association rule is of the form A ->B [s%;c%], where A and B are sets of spatial or nonspatial predicates, s% is the support of the rule, and c% is the confidence of the rule. For example, the following is a spatial association rule:

$$is\_a(X, \text{“school”}) \wedge close\_to(X, \text{“sports\_center”}) \Rightarrow close\_to(X, \text{“park”}) \quad [0.5\%, 80\%]$$

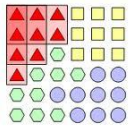
• This rule states that 80% of schools that are close to sports centers are also close to parks, and 0.5% of the data belongs to such a case.

Spatial Association Rule	Sup.	Conf.
$Stem\_height(x, high) \wedge Distance\_to\_edge(x, far) \rightarrow Vegetation\_Durability(x, moderate)$	0.1	0.94
$Vegetation\_Durability(x, moderate) \wedge Distance\_to\_water(x, close) \rightarrow Stem\_Height(x, high)$	0.05	0.95
$Distance\_towater(x, far) \wedge Water\_Depth(x, shallow) \rightarrow Stem\_Height(x, high)$	0.05	0.94



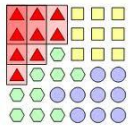
Time Series Modeling- Exponential Smoothing

Example



Estimation Year	Value
2010	25
2011	32
2012	24
2013	28
2014	26
2015	27
2016	?

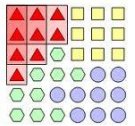
Example



Estimation Year	Value
2010	25
2011	32
2012	24
2013	28
2014	26
2015	27
2016	?

SN	Forecast	Explanation
1	27	Average
2	26.25	last 4 values and take average as other values are bit old and mayy not appropriate
3	28	See last two values and show the increase trend and conclude for 28
4	30	related to population, increase in area i.e. based on other factors
5	26	There is a trend of increase and derease and increase and decreas and so on and conclude that it will decreasae
6	27	take last 3 values and take average
7	26.833	(28*1+26*2+27*3)/6
8	26	Remove 32 and find average

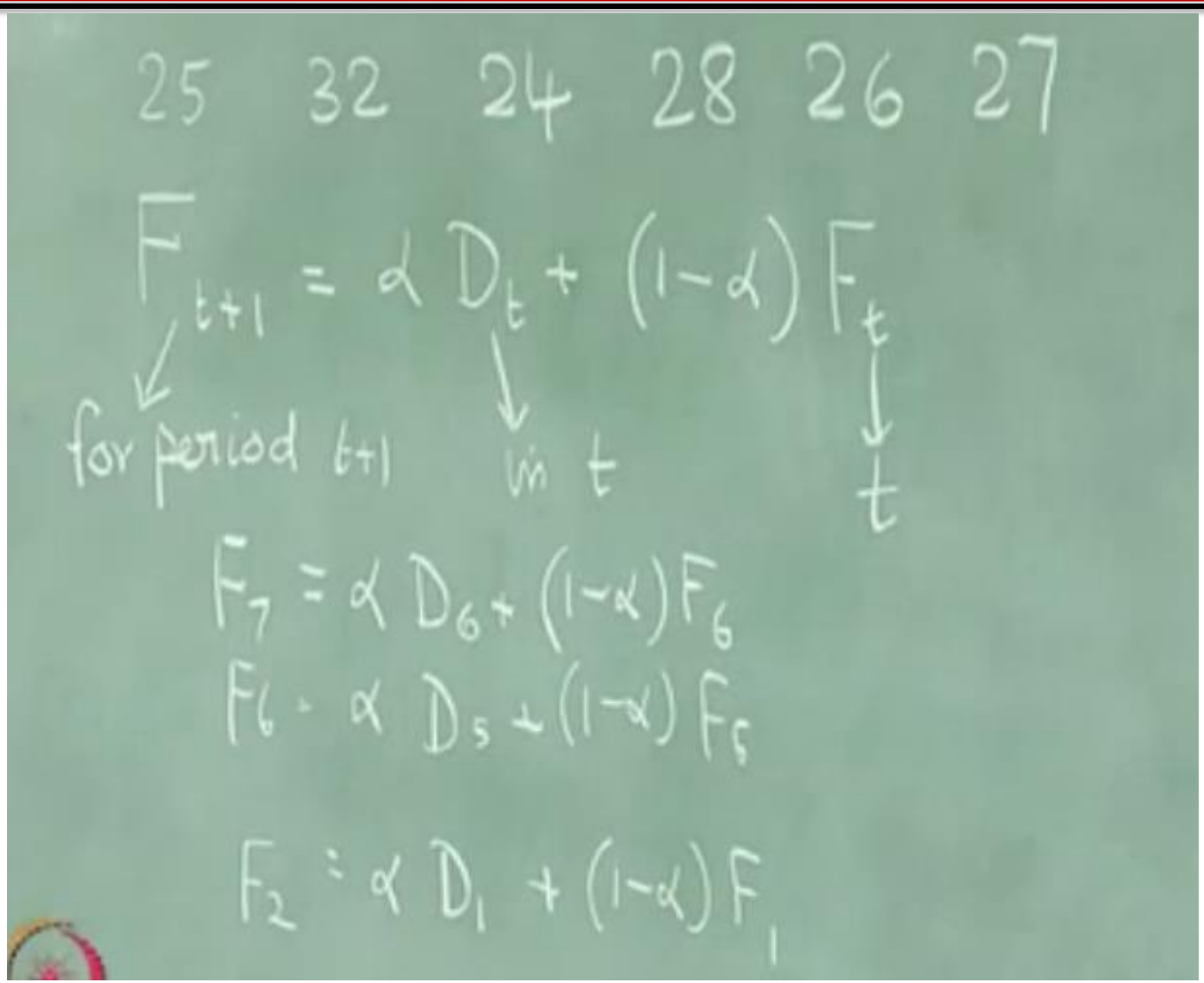
Example



$$F=a+\varepsilon$$

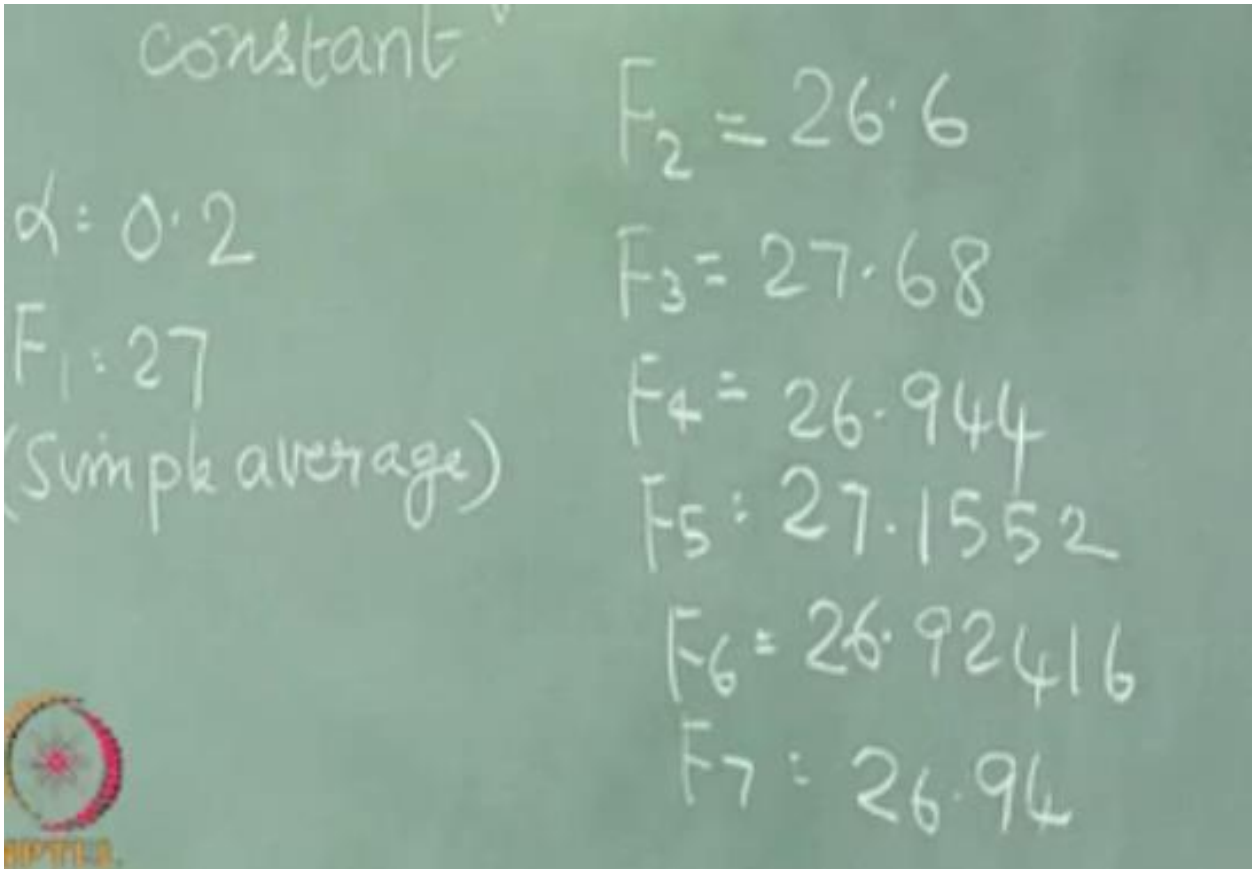
Where,
F = Forecast, a= average, ε=noise (0,δ2)

Exponential Smoothing



F: forecast
D: Demand
α: Smoothing constant

Exponential Smoothing



Exponential Smoothing

$$F_7 = \alpha D_6 + (1-\alpha) F_6$$

$$= \alpha D_6 + (1-\alpha)\left[\alpha D_5 + (1-\alpha) F_5\right]$$

$$= \alpha D_6 + \alpha(1-\alpha) D_5 + (1-\alpha)^2 F_5$$

$$= \alpha D_6 + \alpha(1-\alpha) D_5 + \alpha(1-\alpha)^2 D_4 + \cdots + \alpha(1-\alpha)^5 D_1 + (1-\alpha)^6 F_1$$