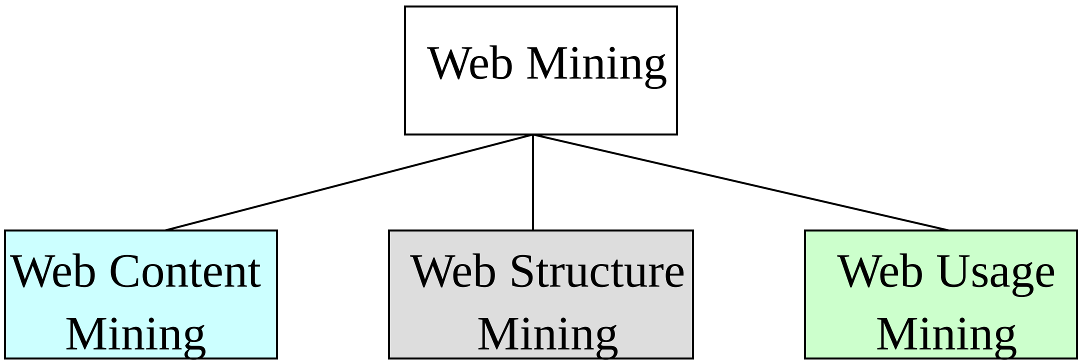
Mining the Web

- **The WWW is huge, widely distributed, global information service centre** for
  - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
  - Hyper-link information
  - Access and usage information
- **WWW provides rich sources for data mining**
- **Challenges**
  - Too huge for effective data warehousing and data mining
  - Too complex and heterogeneous: no standards and structure

Web Mining

Web mining is the **application of data mining** techniques to **extract knowledge from Web data**, i.e.
Web Content, Web Structure and Web Usage data



Web Content Mining

- **Extracting useful information** from **Web documents**.

- **Collection of facts a Web page was designed to convey to the users**.

- **Consist of text, images, audio, video, or structured records such as lists and tables**.

- **Issues addressed in text mining are**, **topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages.**

Web Usage Mining

- **Discover interesting usage patterns from Web data**

- **Identity or origin of Web users with their browsing behavior**

- **Web Server Data: User logs collected by Web server. e.g. IP address, page reference and access time etc.**

- **Application Server Data**: **Web logic/Story Server** have significant features to enable E-commerce applications e.g. **track and log business events**

- **Application Level Data**: **New kinds of events** logging and generating **histories of these specially defined events**.

Applications of Usage Mining

- Target potential customers for electronic commerce
- Enhance the quality and delivery of Internet information services to the end user
- Improve Web server system performance
- Identify potential prime advertisement locations

Web Usage Mining: Example

Statistics generated with http LogMiner version 0.1

General information

Information about analyzed log files

Generated: Wed Jan 17 04:30:41 2007

Number of entries processed 2406
Number of invalid entries 14
Processing time in seconds 0

Generated reports

Click on the report name you want to see

Number of reports generated 9

- Unique visitors in each day
- Unique visitors in each month
- Unique visitors from Google in each day
- Unique visitors from Google in each month
- Requested pages
- Requested images and CSS
- Referers
- Weekday distribution
- Hours distribution

Unique visitors in each day

Multiple hits with the same IP, user agent and access day, are considered a single visit

Number of unique visitors 233
Different days in logfile 4

14/Jan/2007	42 (18.0%)
15/Jan/2007	62 (26.6%)
16/Jan/2007	92 (39.5%)
17/Jan/2007	37 (15.9%)

Web Structure Mining

- Typical **structure: Web pages as nodes, and hyperlinks as edges connecting related pages**.
- Web Structure Mining: process of **discovering structure information from the Web**.

- **Hyperlinks**: **connects a location in a Web page to different location**
  - **Intra-Document hyperlink**
  - **Inter-Document hyperlink**.

- **Document Structure**: **Tree-structured format, HTML and XML tags**
  - **-Extract document object model structures**

PageRank- Introduction

- The heart of Google’s searching software is PageRank, a system for ranking web pages developed by Larry Page and Sergey Brin at Stanford University

- Essentially, Google interprets a link from page A to page B as a vote, by page A, for page B.

- But these votes doesn’t weigh the same, because Google also analyzes the page that casts the vote.

The original PageRank algorithm

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

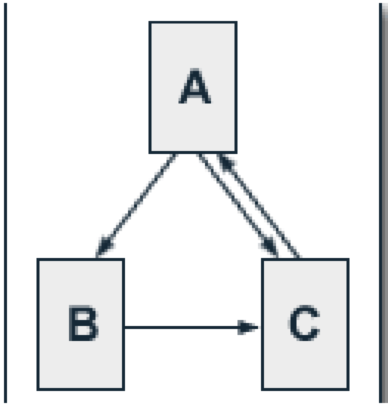
Where:

- PR(A) is the PageRank of page A,
- PR(Ti) is the PageRank of pages Ti which link to page A,
- C(Ti) is the number of outbound links on page Ti
- d is a damping factor which can be set between 0 and 1.
- PageRank of page A is recursively defined by the PageRank of those pages which link to page A

The Characteristics of PageRank

- We regard a small web consisting of three pages A, B and C, whereby page A links to the pages B and C, page B links to page C and page C links to page A. According to Page and Brin, the damping factor d is usually set to 0.85, but to keep the calculation simple we set it to 0.5.

- $PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$



$PR(A) = 0.5 + 0.5\ PR(C)$
$PR(B) = 0.5 + 0.5\ (PR(A) / 2)$
$PR(C) = 0.5 + 0.5\ (PR(A) / 2 + PR(B))$

We get the following PageRank values for the single pages:

$PR(A) = 14/13 = 1.07692308$
$PR(B) = 10/13 = 0.76923077$
$PR(C) = 15/13 = 1.15384615$

The sum of all pages' PageRanks is 3 and thus equals the total number of web pages.

The Iterative Computation of PageRank

- For the simple **three-page example it is easy to solve** the according equation system to determine PageRank values. In practice, the **web consists of billions of documents** and it is not possible to find a solution by inspection.

- Because of the size of the actual web, the Google search engine **uses an approximate**, iterative computation of PageRank values. This means that **each page is assigned an initial starting value** and the PageRanks of all pages are then calculated in several computation circles based on the equations determined by the PageRank algorithm.

- The iterative calculation shall again be illustrated by the three-page example, whereby each page is assigned a starting **PageRank value of 1**.

The Iterative Computation of PageRank (example)

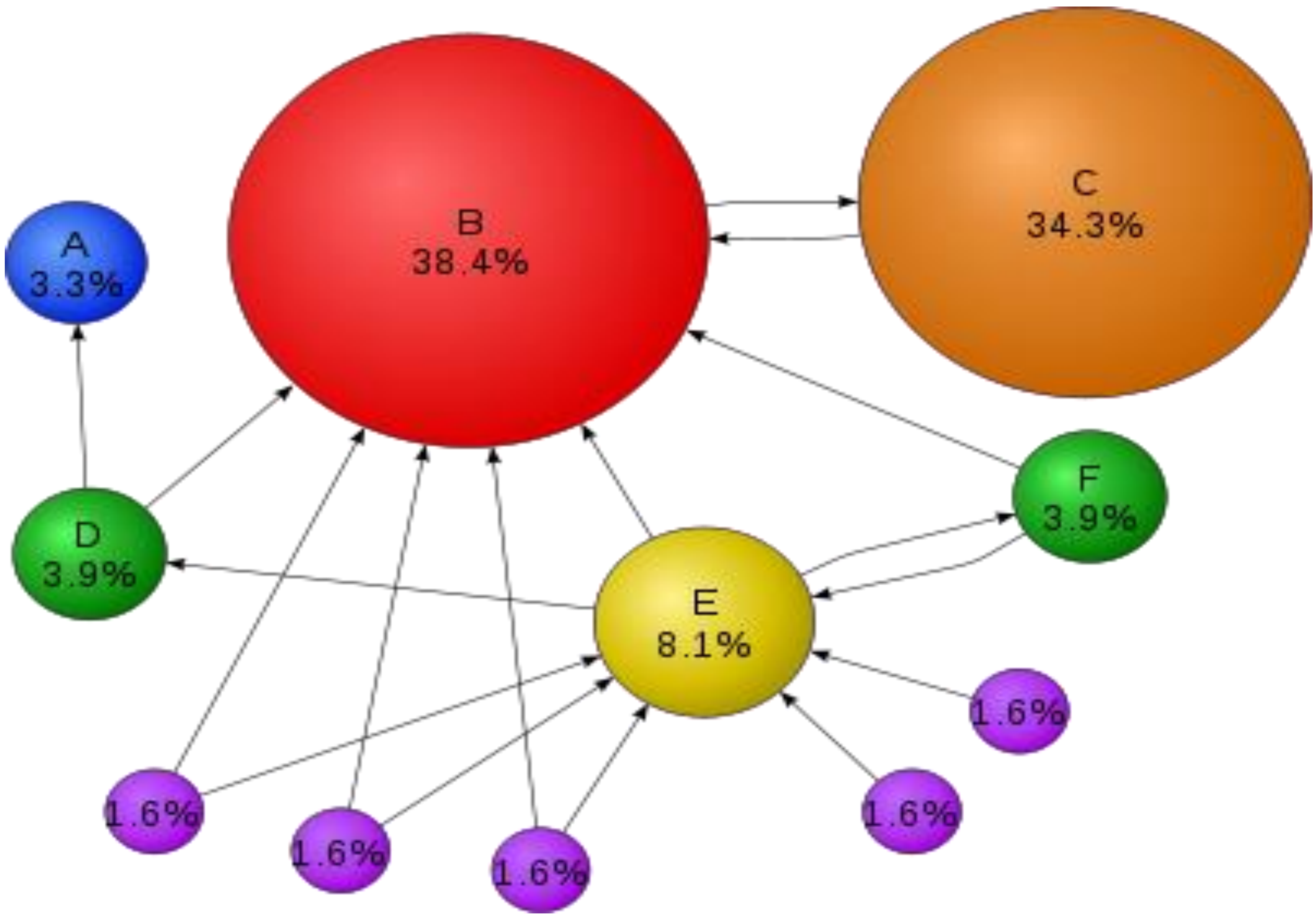
Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592
9	1.07692296	0.76923074	1.15384611
10	1.07692305	0.76923076	1.15384615
11	1.07692307	0.76923077	1.15384615
12	1.07692308	0.76923077	1.15384615

The Iterative Computation of PageRank

- We **get a good approximation of the real PageRank values after only a few iterations**. According to publications of Lawrence Page and Sergey Brin, about 100 iterations are necessary to get a good approximation of the PageRank values of the whole web.

- The **sum of all pages' PageRanks** still converges to the **total number of web pages.** So the **average PageRank of a web page is 1**.

Example Webstructure



The damping factor d

- The **probability for the random surfer not stopping to click on links is given by the damping factor d**, which depends on probability therefore, is set **between 0 and 1**

- The **higher d is**, the more likely will the **random surfer keep clicking links**. Since the **surfer jumps to another page** at random after he stopped clicking links, the probability therefore is implemented as a **constant (1-d)** into the algorithm.

- Regardless of inbound links, the **probability** for the random surfer **jumping to a page is always (1-d),** so a **page has always a minimum PageRank**.

The Effect of Inbound Links

- **Each additional inbound link for a web page always increases that page's PageRank**. Taking a look at the PageRank algorithm, which is given by

$$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

- One may assume that an additional inbound link from page X increases the PageRank of page A by

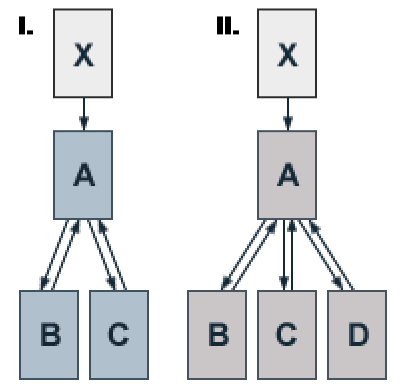
$$\mathbf{d \times PR(X) / C(X)}$$

where PR(X) is the PageRank of page X and C(X) is the total number of its outbound links.

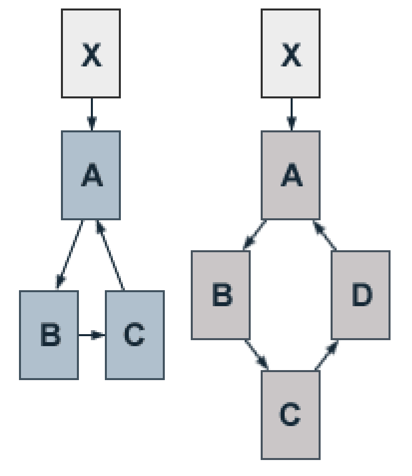
Tips for raising your website's PageRank value

- **Add new pages to your website** (as many as you can)
- **Swap links with websites which have high PageRank value**
- **Raise the number of inbound links (**Advertise your website on other sites)
- When you **add a new page** to your site, be sure to **link it to your front page** and vice versa as it is shown on the picture

Right



Wrong



The effect of additional pages

Sub-pages	PageRank of the front page
1	1.000000
2	1.428673
3	1.857347
4	2.286020
5	2.714694
10	4.858060
20	9.144795
50	22.005003
100	43.438648
250	107.739838
500	214.907135
700	300.642426
1000	429.246613

Challenges in Web Mining

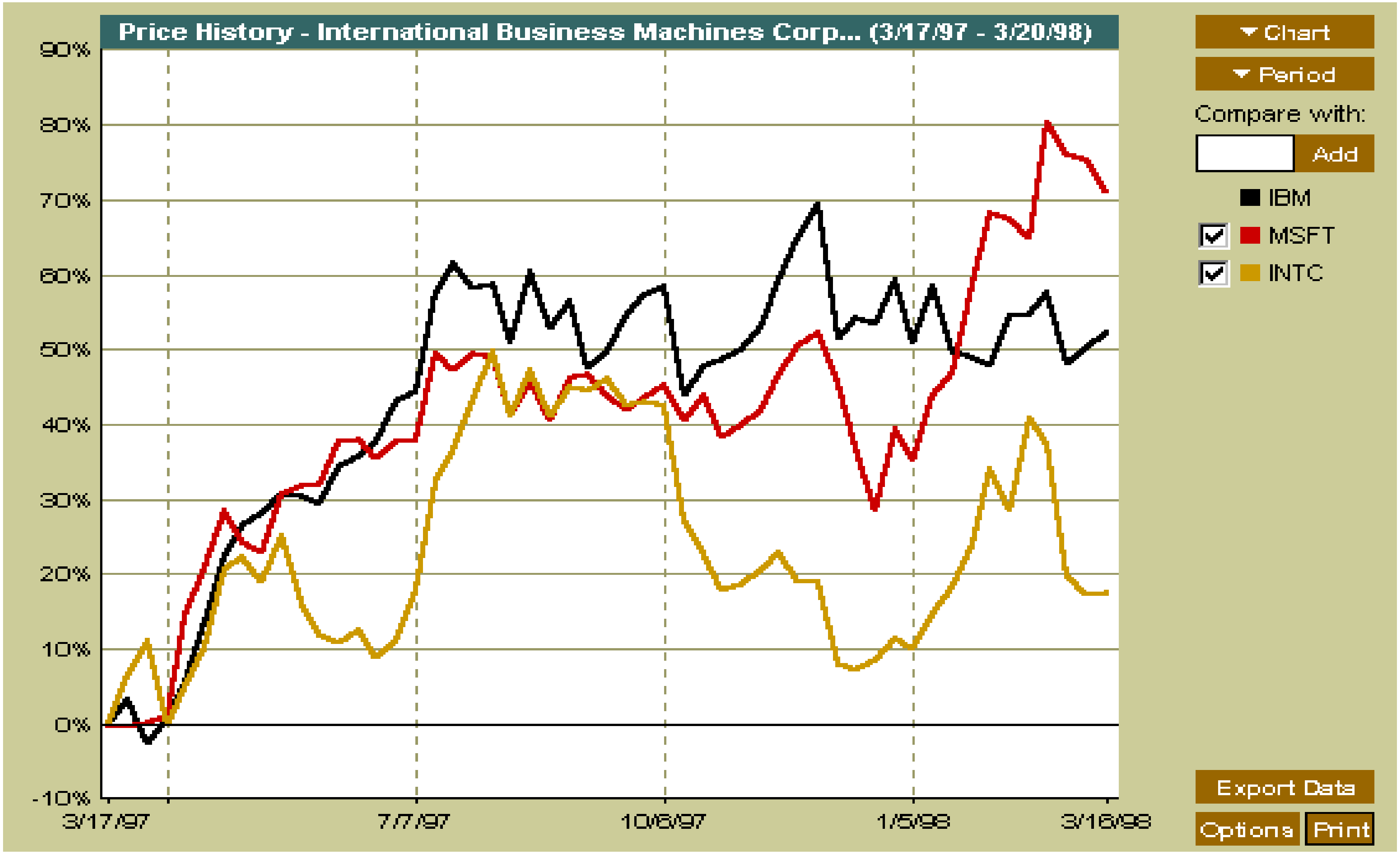
- Too huge for effective data warehousing and data mining.
- Too complex and heterogeneous web structure.
- Growing and changing rapidly
- Broad diversity of user communities.
- Only small portion of the information on the web is truly relevant or useful.

Time Series Data Mining

- A Time Series is an **ordered sequence of data points at uniform time intervals.**
- Examples of time series are the **daily value of a stock, annual/monthly sales figures, weather data for long period of time** etc.
- Time Series Analysis comprises methods for **analyzing time series data in order to extract meaningful statistics, rules and patterns.**
- Later on **these rules and patterns might be used to build forecasting models that are able to predict future developments**.
- In case we want to **predict future trend directions (e.g. up/down)** we have to solve a **Classification** problem. If we try to **forecast future time series data points** (e.g. the nepal stock will be at 1800 point at end of the year) the relevant data mining technique is called **Regression**.

Time-Series Data Mining

- Time-series database
  - Consists of sequences of values or events changing with time
  - Data is recorded at regular intervals
  - Characteristic time-series components
    - Trend, cycle, seasonal, irregular
- Applications
  - Financial: stock price, inflation
  - Industry: power consumption
  - Scientific: experiment results
  - Meteorological: precipitation



- A time series can be illustrated as a time-series graph which describes a point moving with the passage of time

Categories of Time-Series Movements

- Categories of Time-Series Movements
  - **Long-term or trend movements (trend curve):** general direction in which a time series is moving over a long interval of time
  - **Cyclic movements or cycle variations**: long term oscillations about a trend line or curve
    - e.g., business cycles, may or may not be periodic
  - **Seasonal movements or seasonal variations**
    - i.e, almost identical patterns that a time series appears to follow during corresponding months of successive years.
  - **Irregular or random movements**
- **Time series analysis: decomposition of a time series into these four basic movements**
  - Additive Modal: $TS = T + C + S + I$
  - Multiplicative Modal: $TS = T \times C \times S \times I$

Estimation of Trend Curve

- **The freehand method**
  - Fit the curve by looking at the graph
  - Costly and barely reliable for large-scaled data mining
- **The least-square method**
  - Find the curve minimizing the sum of the squares of the deviation of points on the curve from the corresponding data points
- **The moving-average method**

Moving Average

- Moving average of order n

$$\frac{y_1 + y_2 + \cdots + y_n}{n}, \; \frac{y_2 + y_3 + \cdots + y_{n+1}}{n}, \; \frac{y_3 + y_4 + \cdots + y_{n+2}}{n}, \cdots$$

- Eg: Original Data: 3 7 2 0 4 5 9 7 2
- Moving average of order3: (3 + 7 + 2)/3 = 4, 3 2 3 6 7 6
- Weighted (1, 4, 1) average: ((1*3 +4*7 +1*2)/(1+4 +1))= 5.5, 2.5 1 3.5 5.5 8 6.5

Trend Discovery in Time-Series : Estimation of Seasonal Variations

- Seasonal index
  - Set of **numbers showing the relative values of a variable during the months of the year**
  - E.g., if the sales during **October, November, and December are 80%, 120%, and 140%** of the average monthly sales for the whole year, respectively, then **80, 120, and 140 are seasonal index** numbers for these months

Trend Discovery in Time-Series

- **Estimation of cyclic variations**
  - If (approximate) **periodicity of cycles occurs**, **cyclic index** can be constructed in much the same manner as **seasonal indexes**

- **Estimation of irregular variations**
  - By adjusting the data for trend, seasonal and cyclic variations

- With the **systematic analysis of the trend, cyclic, seasonal, and irregular components**, it is possible to make **long- or short-term predictions with reasonable quality**

Multimedia Mining

- Multimedia database system stores and manages a large collection of multimedia data such as audio, video, images, graphics, speech, text etc.

- Image/multimedia mining deals with extraction of implicit knowledge, data relationship or other patterns not explicitly stored in images/multimedia

- The challenges in images mining is to determine the low-level pixel representation contained in an image or image sequence and can be effectively and efficiently processed to identify high level spatial objects and relationships.

- Typical image/multimedia processing involves preprocessing, transformations and feature extraction mining, evaluation and interpretation of the knowledge.

- Different data mining techniques can be used such as association rules, clustering.