WEB MINING

What is Web Mining?

- ***Web*** is a collection of files/documents over the Internet which are inter-linked.
- The World Wide Web contains a large amount of data that provides a rich source to data mining.
- ***Web Mining*** is the application of Data Mining techniques to automatically discover and extract information from Web documents and services.
- The objective of Web mining is to look for patterns in Web data by collecting and examining data in order to gain insights.
- Web mining helps to improve the power of web search engine by identifying the web pages and classifying the web documents.
- Web mining is very useful to e-commerce websites and e-services.

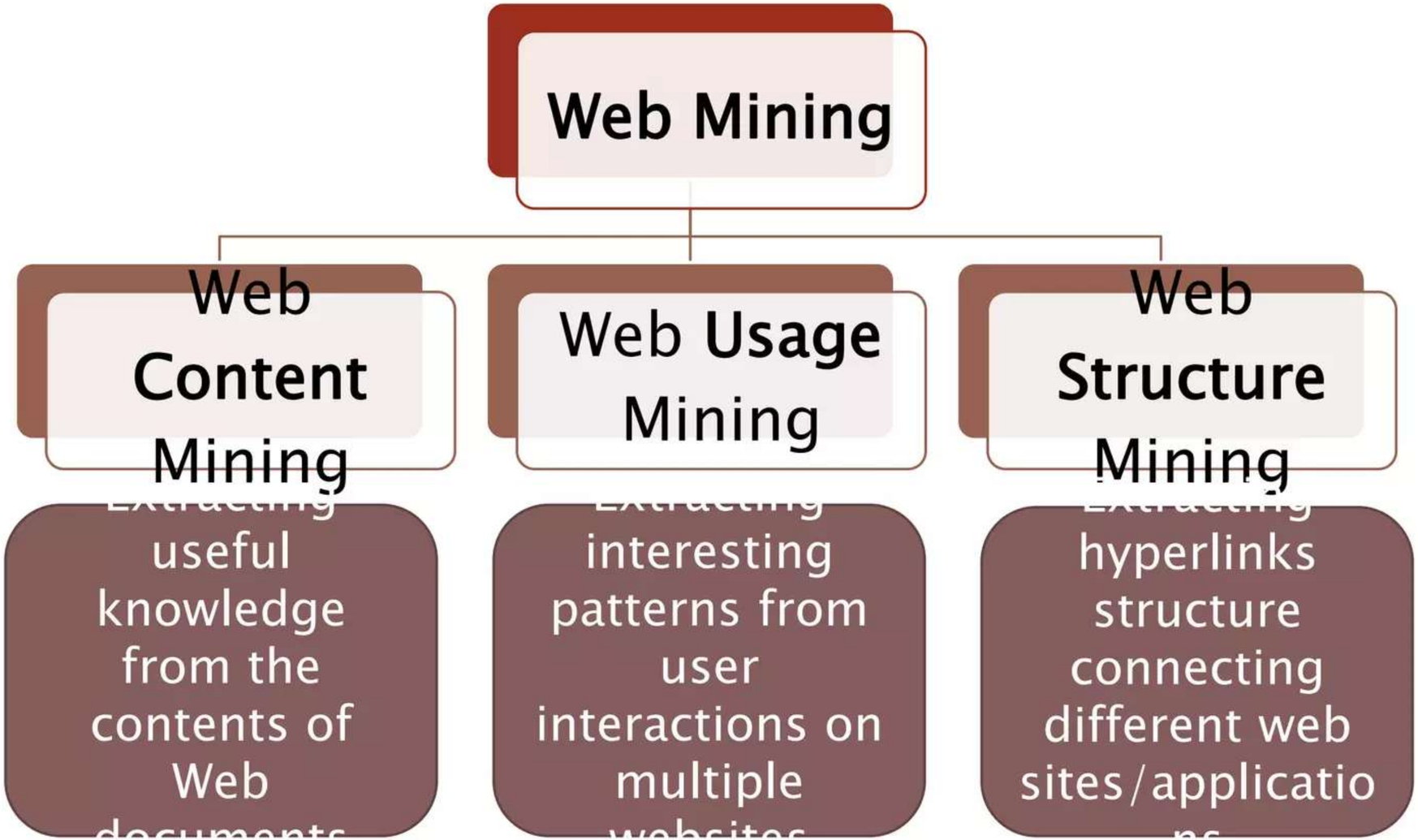
What is Web Mining?

- Web mining is used to predict user behavior.
- It is used for Web Searching e.g., Google, Yahoo etc. and in Vertical Searching.
- Web mining has a distinctive property to provide a set of various data types.
- The web has multiple aspects that yield different approaches for the mining process, such as
  - Web pages consist of text,
  - Web pages are linked via hyperlinks, and
  - User activity can be monitored via web server logs.

Challenges in Web Mining

- Amount of information on the Web is huge
- Coverage of Web information is very wide diverse
- Web information is semi-structured.
- Web information is linked.
- Web information is redundant
- Web consists of Surface web and Deep web
- Web information is dynamic
- Web information is noisy

Web Mining Taxonomy

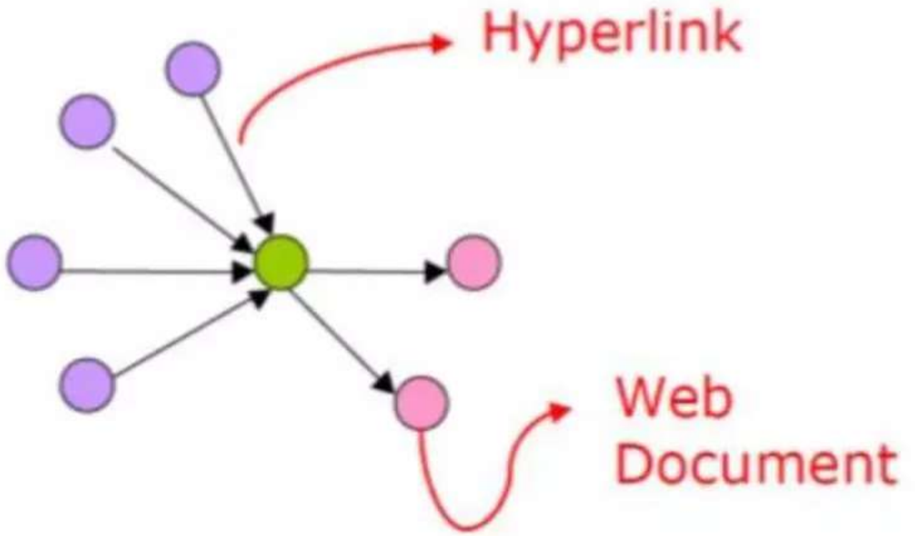


Web Structure Mining

- Web structure mining is the application of discovering structure information from the web.
- Structure mining basically shows the structured summary of a particular website.
- It identifies relationship between web pages linked by information or direct link connection.
- To determine the connection between two commercial websites, Web structure mining can be very useful.

Web Structure Mining

- The structure of the web graph consists of ***web pages as nodes***, and ***hyperlinks as edges*** connecting related pages.
- Type of mining
  - Document level mining (Intra-page)
  - Hyperlink level mining (Inter-page)
- The research at hyperlink level is known as ***Hyperlink Analysis***
- Web pages can be accessed using URL, hyperlinks or using navigational tools.



Web Structure Mining

- Web page ranking algorithms PageRank (PR)
  - Google Search engine uses PageRank algorithm to rank websites in their search engine results.
  - It is a way of measuring the importance of website pages.

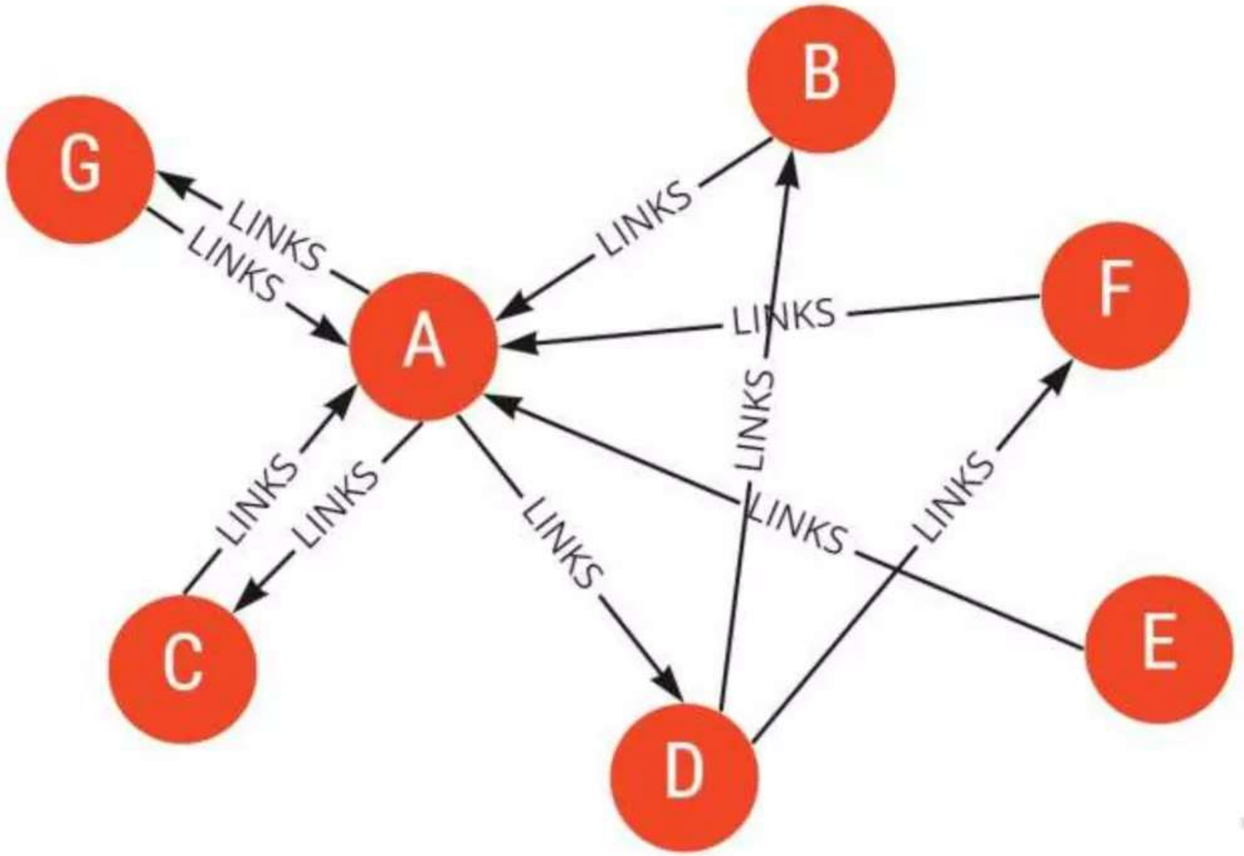
> *"It counts the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites."*

  - The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page.
  - PageRank can be calculated for collections of documents of any size.

https://www.youtube.com/watch?v=meonLcN7LD4

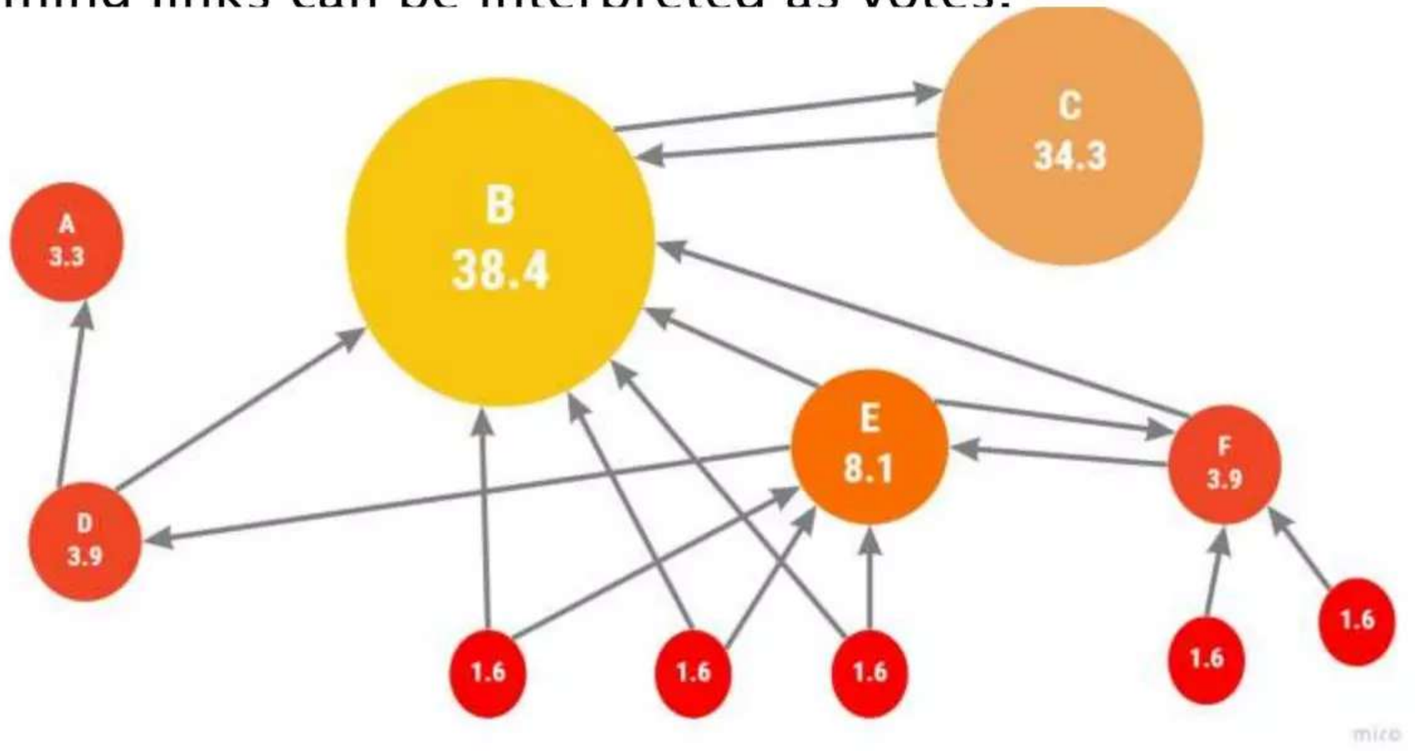
Web Structure Mining

- Web page ranking algorithms PageRank (PR)



Web Structure Mining

- Web page ranking algorithms PageRank (PR)
  - Google interprets a link from page A to page B as a vote from page A to page B.
  - All incoming links can be interpreted as votes.



Web Structure Mining

- Web page ranking algorithms PageRank (PR)
  - It also takes into consideration the “importance” of the page that is “giving” out the vote.
  - If the page that’s casting a vote is more important, then the links are worth more and it will help rank up the other pages.
  - Page’s importance is equal to the sum of the votes of its incoming links.
  - A web page is important if it is pointed by other important web pages.
  - Types of Links
    - Inbound links (In-links)
    - Outbound links (Out-links)

Web Structure Mining

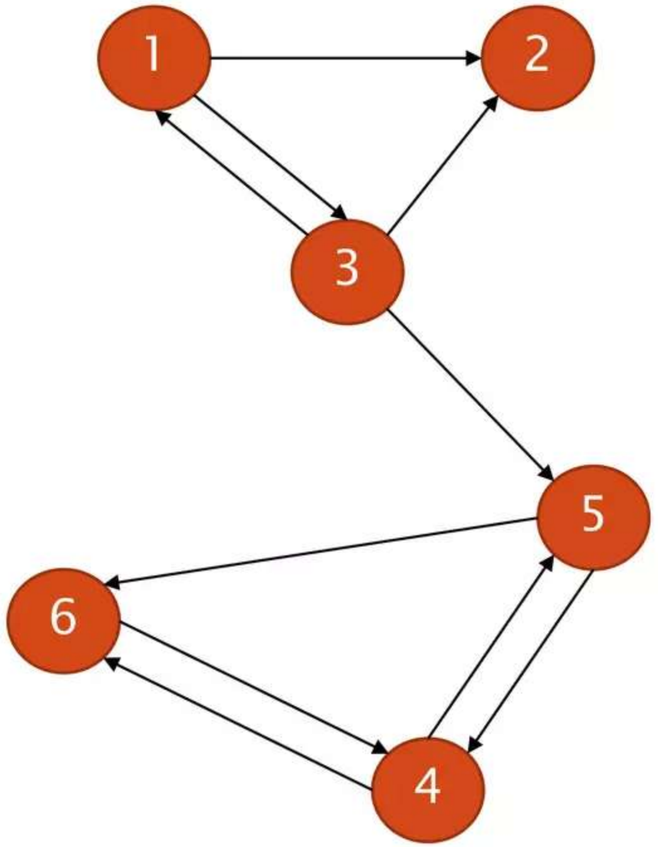
- Web page ranking algorithms PageRank (PR)
  - Page rank of a page P_i denoted by $PR(P_i)$ is the sum of page ranks of all pages pointing into P_i.

$$PR(P_i) = \sum_{P_j \to P_i} \frac{PR(P_j)}{|P_j|}$$

  - $P_j \to P_i$ set of pages pointing to P_i
  - $|P_j| =$ number of outlinks from P_j
  - PR is page rank initially unknown
    - All pages are given equal page rank initially $1/n$
    - n is number of pages in Google index of web

Web Structure Mining

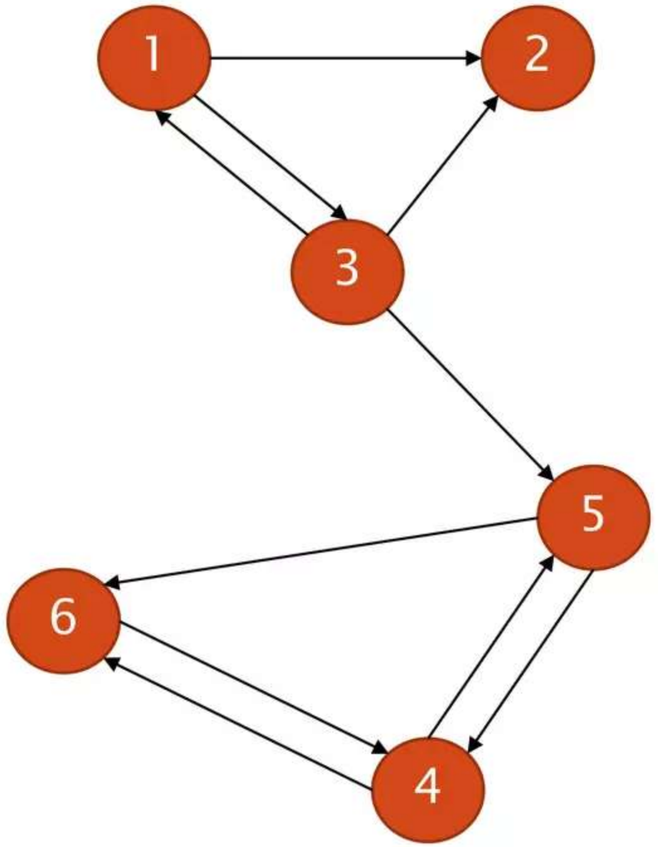
- Web page ranking algorithms PageRank (PR)



Iteration 1
$PR(1) = \frac{1}{6}$
$PR(2) = \frac{1}{6}$
$PR(3) = \frac{1}{6}$
$PR(4) = \frac{1}{6}$
$PR(5) = \frac{1}{6}$
$PR(6) = \frac{1}{6}$

Web Structure Mining

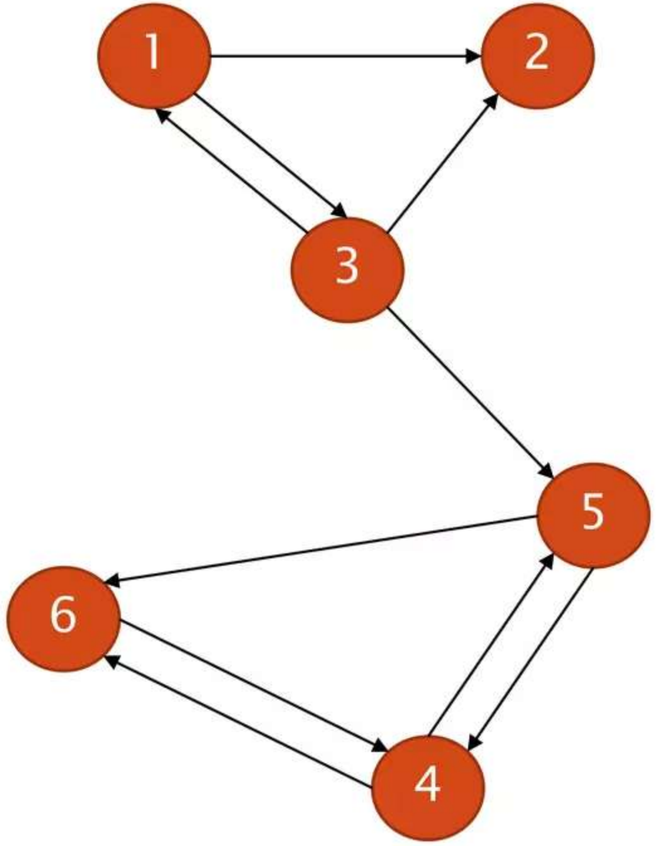
- Web page ranking algorithms PageRank (PR)



Iteration 1	Iteration 2
$PR(1) = \frac{1}{6}$	$PR(1) = \frac{1}{18}$
$PR(2) = \frac{1}{6}$	$PR(2) = \frac{5}{36}$
$PR(3) = \frac{1}{6}$	$PR(3) = \frac{1}{12}$
$PR(4) = \frac{1}{6}$	$PR(4) = \frac{1}{4}$
$PR(5) = \frac{1}{6}$	$PR(5) = \frac{5}{36}$
$PR(6) = \frac{1}{6}$	$PR(6) = \frac{1}{6}$

Web Structure Mining

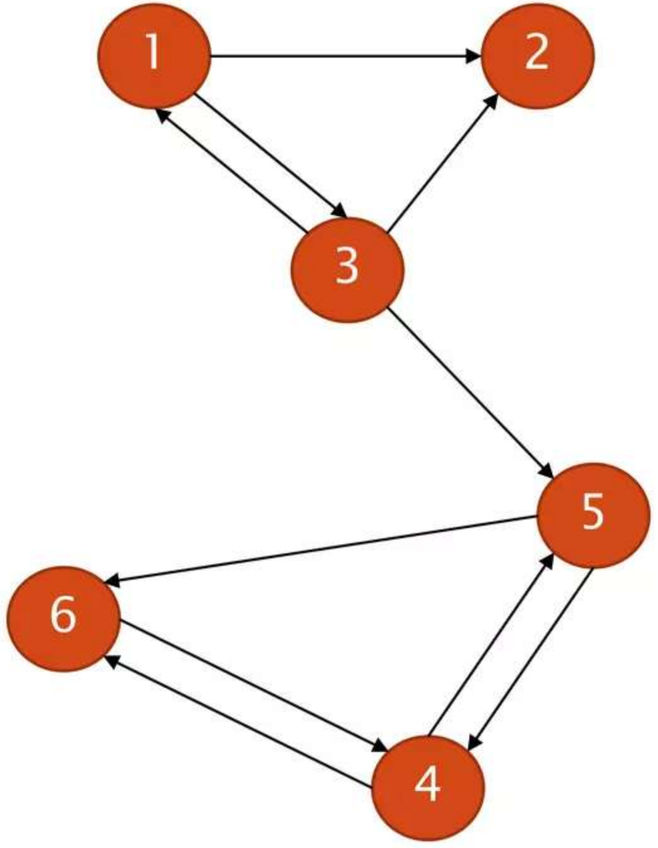
- Web page ranking algorithms PageRank (PR)



Iteration 1	Iteration 2	Iteration 2
$PR(1) = \frac{1}{6}$	$PR(1) = \frac{1}{18}$	$PR(1) = \frac{1}{36}$
$PR(2) = \frac{1}{6}$	$PR(2) = \frac{5}{36}$	$PR(2) = \frac{1}{18}$
$PR(3) = \frac{1}{6}$	$PR(3) = \frac{1}{12}$	$PR(3) = \frac{1}{36}$
$PR(4) = \frac{1}{6}$	$PR(4) = \frac{1}{4}$	$PR(4) = \frac{17}{72}$
$PR(5) = \frac{1}{6}$	$PR(5) = \frac{5}{36}$	$PR(5) = \frac{11}{72}$
$PR(6) = \frac{1}{6}$	$PR(6) = \frac{1}{6}$	$PR(6) = \frac{14}{72}$

Web Structure Mining

- Web page ranking algorithms PageRank (PR)



Iteration 1	Iteration 2	Iteration 2	Rank
$PR(1) = \frac{1}{6}$	$PR(1) = \frac{1}{18}$	$PR(1) = \frac{1}{36}$	0.0277 **5**
$PR(2) = \frac{1}{6}$	$PR(2) = \frac{5}{36}$	$PR(2) = \frac{1}{18}$	0.0555 **4**
$PR(3) = \frac{1}{6}$	$PR(3) = \frac{1}{12}$	$PR(3) = \frac{1}{36}$	0.0277 **5**
$PR(4) = \frac{1}{6}$	$PR(4) = \frac{1}{4}$	$PR(4) = \frac{17}{72}$	0.2361 **1**
$PR(5) = \frac{1}{6}$	$PR(5) = \frac{5}{36}$	$PR(5) = \frac{11}{72}$	0.1527 **3**
$PR(6) = \frac{1}{6}$	$PR(6) = \frac{1}{6}$	$PR(6) = \frac{14}{72}$	0.1944 **2**

Web Structure Mining

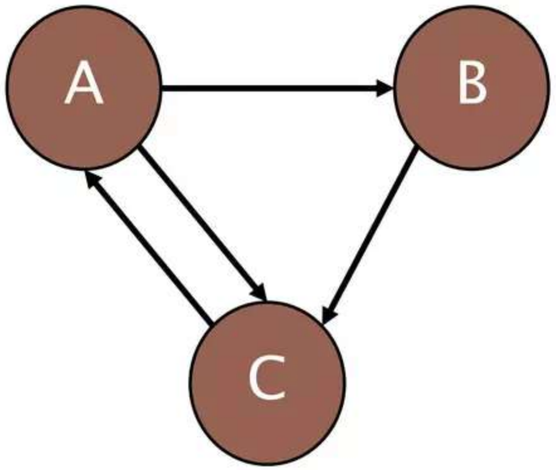
- Web page ranking algorithms PageRank (PR)
  - Simple PageRank Algorithm

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right)$$

  - where,
    - $PR(A)$ –Page Rank of page A
    - $PR(T_i)$– Page Rank of pages T_i which link to page A
    - $C(T_i)$ –number of outbound links on page T_i
    - d –**damping factor** which can be set between 0 and 1 (Usually set to 0.85)
      - The probability of a random user continuing to click on links as they browse the web.

Web Structure Mining

- Web page ranking algorithms PageRank (PR)
  - Simple Page Rank Algorithm



$$PR(A) = (1-d) + d(PR(C)/C(C))$$
$$= (1-0.85) + 0.85(1/1)$$
$$= 0.15 + 0.85 = 1$$

$$PR(B) = (1-d) + d(PR(A)/C(A))$$
$$= (1-0.85) + 0.85(1/2)$$
$$= 0.15 + 0.425 = 0.575$$

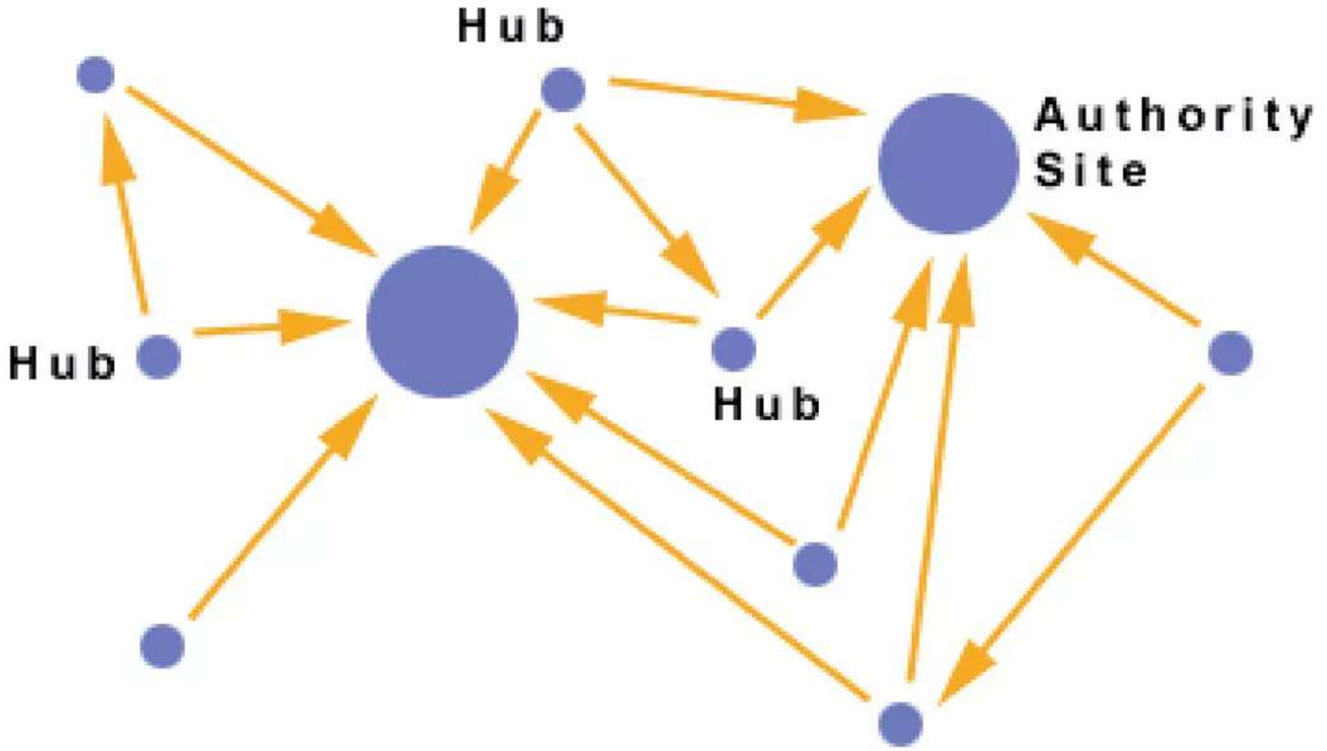
$$PR(C) = (1-d) + d\left(\frac{PR(A)}{C(A)} + \frac{PR(B)}{C(B)}\right)$$
$$= (1-0.85) + 0.85(1/2 + 0.575)$$
$$= 0.15 + 0.85(0.5 + 0.575)$$
$$= 0.15 + 0.91375 = 1.06$$

Web Structure Mining

- Hyperlink Induced Topic Search (HITS) Algorithm
  - A Link Analysis Algorithm that rates webpages, developed by Jon Kleinberg.
  - This algorithm is used to discover and rank the webpages relevant for a particular search.
  - HITS uses ***hubs*** and ***authorities*** to define a recursive relationship between webpages.
    - Given a query to a Search Engine, the set of highly relevant web pages are called **Roots**. They are potential **Authorities**.
    - Pages that are not very relevant but point to pages in the Root are called **Hubs**.
    - An Authority is a page that many hubs link to, whereas a Hub is a page that links to many authorities.

Web Structure Mining

- Hyperlink Induced Topic Search (HITS) Algorithm



Web Structure Mining

- Hyperlink Induced Topic Search (HITS) Algorithm
  - Hub and authority scores:
    - Compute for each page d in the base set a hub score $h(d)$ and an authority score $a(d)$.
    - Initialize for all d : $h(d) = 1$; $a(d) = 1$.
    - Update h(d) and $a(d)$ for each iteration.
  - Output pages with highest h scores as top hubs.
  - Output pages with highest a scores as top authorities

Web Structure Mining

- Hyperlink Induced Topic Search (HITS) Algorithm

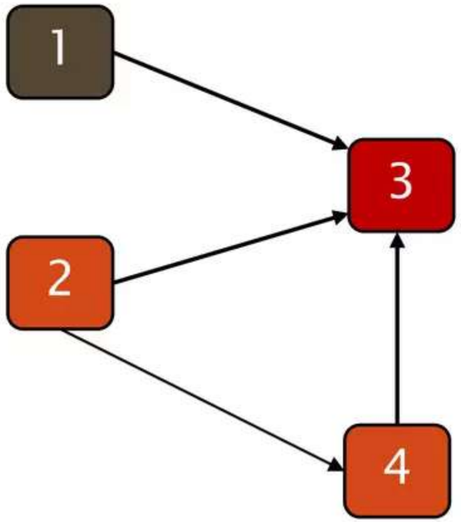
$$a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ . \\ . \\ a_n \end{bmatrix} \qquad h = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ . \\ . \\ h_n \end{bmatrix} \qquad \begin{cases} a = A^T \cdot h \\ h = A \cdot a \end{cases}$$

$$a_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad h_0 = A^T \cdot \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \begin{cases} a_k = (A^T \cdot A) \cdot a_{k-1} \\ h_k = (A \cdot A^T) \cdot h_{k-1} \end{cases}$$

a = Authority Matrix
h = Hub matrix
A = Adjacency Matrix of web

Web Structure Mining

- Hyperlink Induced Topic Search (HITS) Algorithm



$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad A^T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad h = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$a = A^T \cdot h = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 3 \\ 1 \end{bmatrix}$$

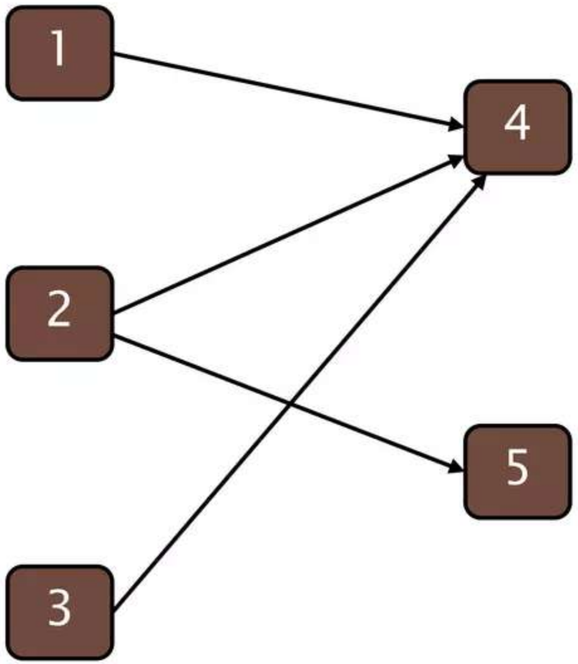
$$h = A \cdot a = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 0 \\ 3 \end{bmatrix}$$

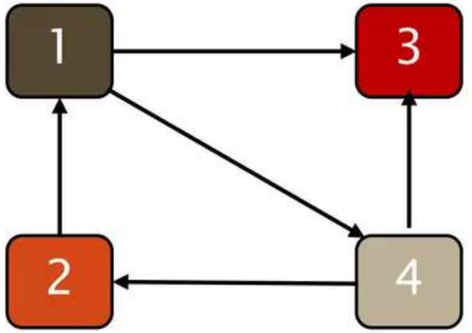
Top hub = Node 2

Top authority = Node 3

Web Structure Mining

- Hyperlink Induced Topic Search (HITS) Algorithm





Web Structure Mining

- PageRank vs HITS
  - PageRank:
    - Used for ranking all the nodes of the complete graph and then applying a search.
    - Based on the 'random surfer' idea
    - Example : Google
  - HITS
    - Applied when a search is done on the complete graph.
    - Applied on a subgraph based on query.
    - Defines hubs and authorities recursively.
    - Example :ASK, Yahoo, Clever

Web Content Mining

- Web content mining is the application of extracting useful information from the content of the web documents.
- Web content consist of several types of data – text, image, audio, video etc.
- Content data is the group of facts that a web page is designed.
- It can provide effective and interesting patterns about user needs.
- Text documents are related to text mining, machine learning and natural language processing.
- This mining is also known as text mining.
- This type of mining performs scanning and mining of the text, images and groups of web pages according to the content of the input.

Web Content Mining

- With web mining, one can
  - Cluster web documents
  - Classify web pages
  - Extract patterns
- Research activities in this field involve using other techniques from other discipline such as Information Retrieval (IR) and Natural Language Processing (NLP).
- Applications
  - Document clustering / categorization
  - Topic identification / tracking
  - Concept discovery

Web Content Mining

- **Web Scraping**
  - Web Scraping is an automatic method to obtain large amounts of data from websites.
  - Unstructured data in an HTML format is converted into structured data in a spreadsheet or a database so that it can be used in various applications.
  - There are many different ways to perform web scraping to obtain data from websites.
    - Using online services,
    - Using particular API's or
    - creating your code for web scraping from scratch.
  - Websites like Google, Twitter, Facebook, StackOverflow, etc. have API's that allow you to access their data in a structured format.
  - Some websites that don't allow users to access large amounts of data in a structured form, use Web Scraping to scrape the website for data.

Web Content Mining

- **Web Scraping**
  - Web scraping requires two parts namely the ***crawler*** and the ***scraper***.
  - The ***crawler*** is an artificial intelligence algorithm that browses the web to search the particular data required by following the links across the internet.
  - The ***scraper*** is a specific tool created to extract the data from the website.

Web Content Mining

- **Web Crawler**
  - A ***web crawler*** as known as ***web spider*** is a bot that searches and indexes content on the internet.
  - Web crawlers are responsible for understanding the content on a web page so they can retrieve it when an inquiry is made.
  - Web crawlers are operated by search engines with their own algorithms.
  - A web spider will search (crawl) and categorize all web pages on the internet that it can find and index them.
  - The most well known web crawler is the Googlebot.

Web Content Mining

- **How does a Web Crawler work?**
  - *Finds information by crawling*
    - Crawlers look at webpages and follow links on those pages, much like you would if you were browsing content on the web.
    - They go from link to link and bring data about those webpages back to the servers.
  - *Organizes information by indexing*
    - When crawlers find a webpage, systems render the content of the page, just as a browser does.
    - It takes note of key signals — from keywords to website freshness — and keeps track of it all in the Search index.

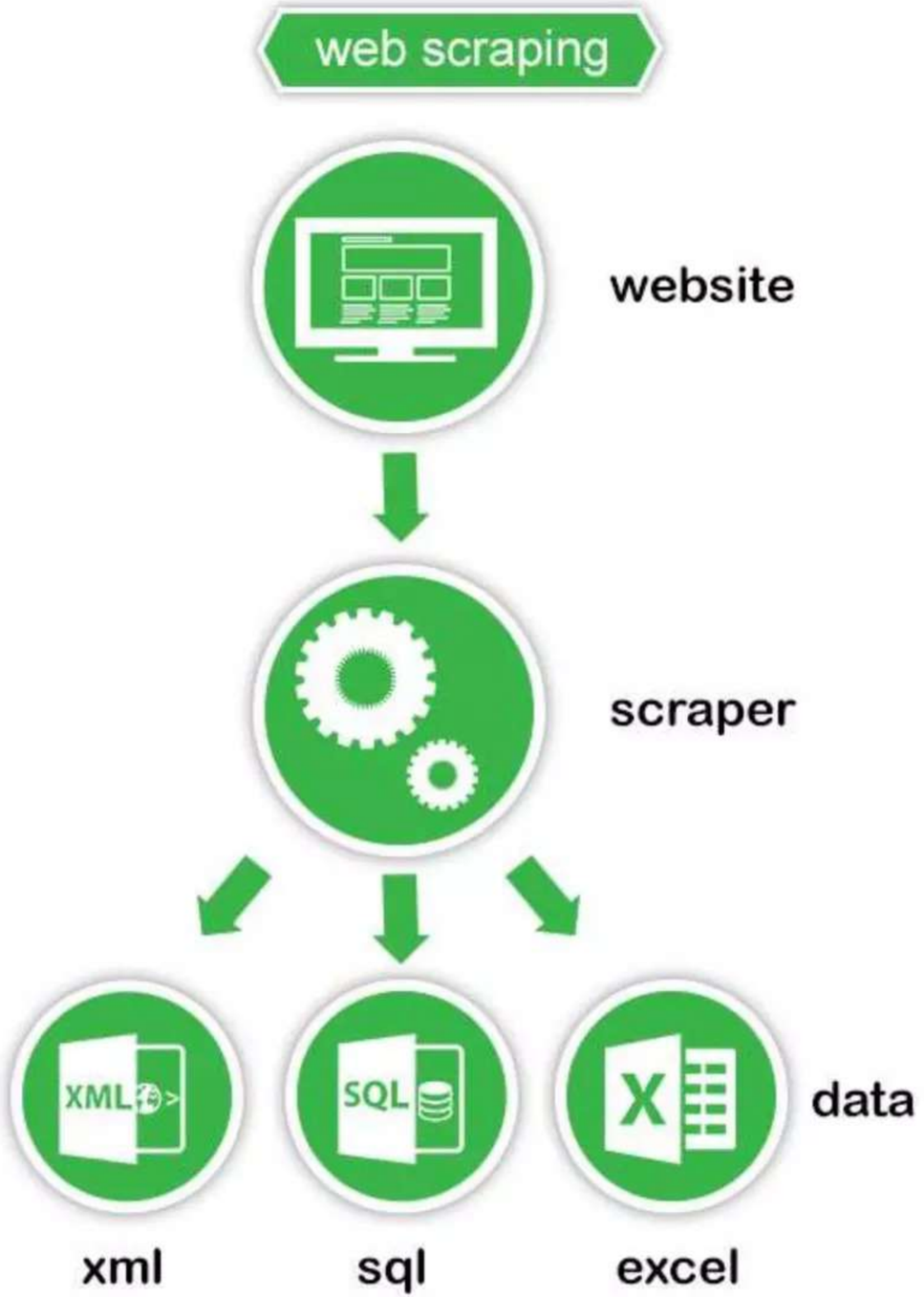
https://www.youtube.com/watch?v=jkuRWcH1-kk

Web Content Mining

- **Applications of Web Crawler**
  - The crawlers are the basis for the work of search engines.
  - Web crawlers are also used for other purposes:
    - **Price comparison portals** search for information on specific products on the Web, so that prices or data can be compared accurately.
    - In the area of **data mining**, a crawler may collect publicly available e-mail or postal addresses of companies.
    - **Web analysis** tools use crawlers or spiders to collect data for page views, or incoming or outbound links.
    - Crawlers serve to provide **information hubs** with data, for example, news sites.

Web Content Mining

- **Web Scraping**
  - First the web scraper is provided the URL's of the required sites.
  - Then it loads all the HTML code for those sites
  - Then scraper obtains the required data from this HTML code and outputs this data in the format specified by the user.
  - Output is in the form of an Excel spreadsheet or a CSV file but the data can also be saved in other formats such as a JSON file.



https://www.youtube.com/watch?v=US-HP1hiRHY

Web Content Mining

- **Web Scraping Applications**
  - Price Intelligence
  - Market Research
  - Alternative data for finance
  - Real estate
  - News and content monitoring etc.

Web Content Mining

- **Web Scraping**
  - Python is popular programming language for web scraping.
  - **Scrapy**:
    - Open-source web crawling framework
    - Ideal for web scraping as well as extracting data using APIs.
  - **Beautiful soap**:
    - Highly suitable for Web Scraping.
    - It creates a parse tree that can be used to extract data from HTML on a website.

Web Usage Mining

- Focuses on predicting the behavior of users while they are interacting with the WWW.
- ***Focus:***
  - How web sites are used by users?
  - Understanding behavior of different user segment.
  - Predict user behavior in future
  - Extracting meaningful information to individual user or group.
  - What pages are being accessed frequently?
  - From what search engine are visitors coming?
  - Which browser and operating systems are most commonly used by visitors?
  - What are the most recent visits per page?
  - Who is visiting which page?

Web Usage Mining

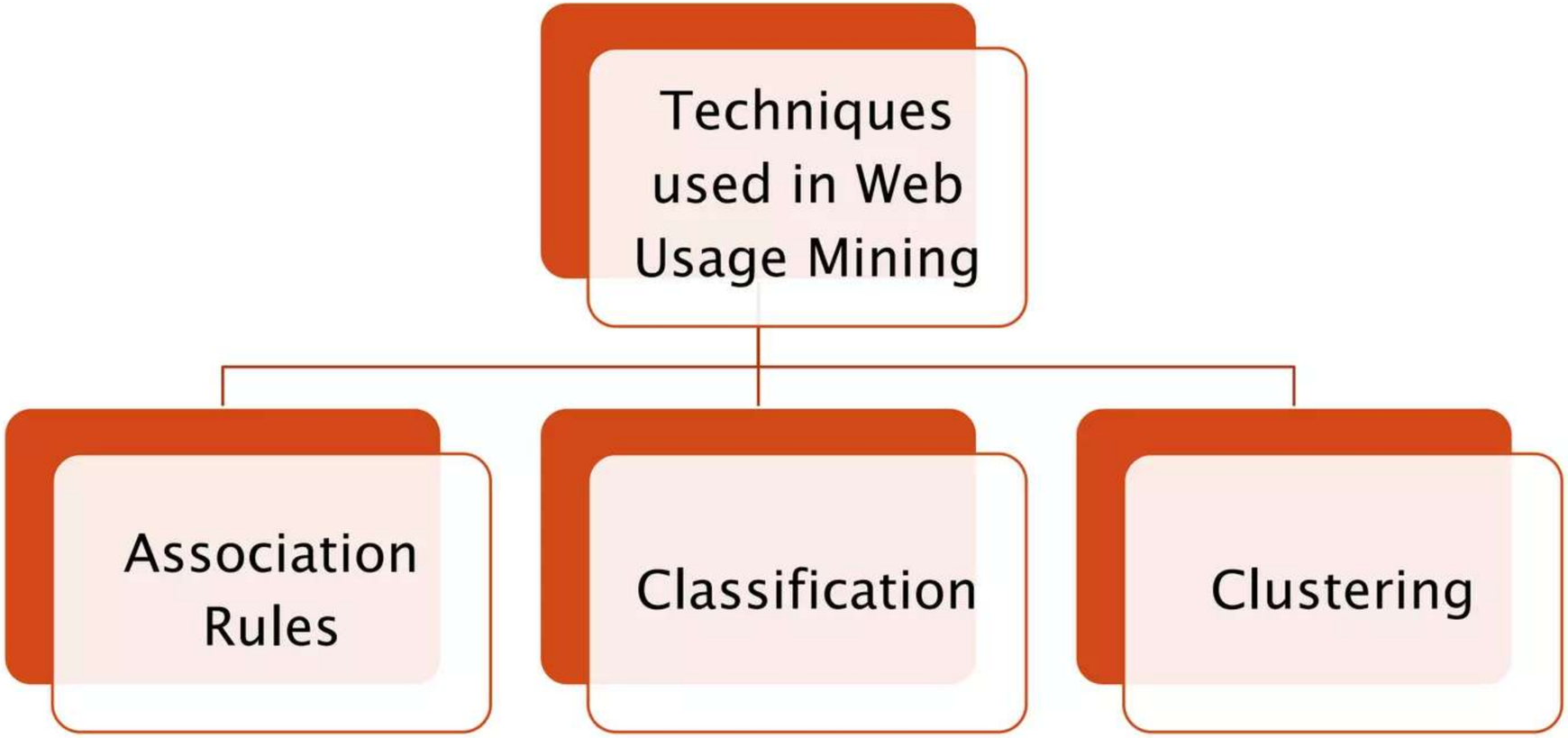
- Analyzes user activities on different web pages and track them over a period of time.
- Three types of Web Data
  - Web content data
  - Web structure data
  - Web usage data
    - The main source of data here is-Web Server and Application Server.
    - It involves log data which is collected by sources.
    - The data in this type can be mainly categorized into three types based on the source it comes from:
      - Server side.
      - Client side
      - Proxy side

Web Usage Mining

- Types of Web Usage mining based on the usage data
  - Web server data
    - includes the IP address, browser logs, proxy server logs, user profiles, etc.
  - Application server data
    - Tracking various business events and logging them into application server logs is mainly what application server data consists of.
  - Application level data
    - There are various new kinds of events that can be there in an application.

Web Usage Mining

- Techniques used in Web Usage Mining



Web Usage Mining

- Techniques used in Web Usage Mining
  - Association Rules
    - This technique focuses on relations among the web pages that frequently appear together in users' sessions.
    - The pages accessed together are always put together into a single server session. Association Rules help in the reconstruction of websites using the access logs.
    - So many sets of rules produced together may result in some of the rules being completely inconsequential.

Web Usage Mining

- Techniques used in Web Usage Mining
  - Classification
    - Develops that kind of profile of users/customers that are associated with a particular class/category.
    - One requires to extract the best features that will be best suitable for the associated class.
    - Classification can be implemented by various algorithms – some of them include- Support vector machines, K-Nearest Neighbors, Logistic Regression, Decision Trees, etc.

Web Usage Mining

- Techniques used in Web Usage Mining
  - Clustering
    - There are mainly 2 types of clusters- the first one is the usage cluster and the second one is the page cluster.
    - The clustering of pages can be readily performed based on the usage data.
    - In usage-based clustering, items that are commonly accessed /purchased together can be automatically organized into groups.

Virtual Web View

- A view on top of the World-Wide-Web
- An approach to handle an unstructured data
- Uses Multiple Layered Database (MLDB) built on top of web