



CHALLENGES THE WEB POSES FOR EFFECTIVE RESOURCE & KNOWLEDGE DISCOVERY

- The web seems to be too big for effective data warehousing & data mining
The size of the web is in the order of hundreds of terabytes & is rapidly growing. It is barely possible to replicate, store & integrate all the data on the web.
- The complexity of web pages is far greater than that of any traditional text document collection.
Web pages lack a unifying structure. The web is like a digital library, but the documents are not sorted in any order.
- The web page is a highly dynamic information source.
Not only does the web grow rapidly, but its info. is also constantly updated.
- The web serves a broad diversity of user communities. The web connects more than 100 million workstations & its user communities are rapidly expanding.
- Only a portion of the info. on the web is useful or relevant. It is said that 99.1% of the web info. is useless to 99.1% of the users.

- Compared to DOM-based methods, the segments obtained by VIPS are more semantically aggregated.
- Contents w/ different topics are distinguished as separate blocks.

MINING THE WEB PAGE LAYOUT STRUCTURE

- A web page has more structure than traditional plain text. Web pages are also regarded as semi-structured data.
- The basic structure of a webpage is its DOM (Document Object Model) structure.
- The DOM structure of a webpage is a tree structure, where every HTML tag corresponds to a node in the tree.
- Due to the flexibility of HTML syntax, many web pages don't obey the W3C specification, which may result in errors in the DOM tree structure.
- Initially, the DOM tree was introduced for presentation in the browser rather than description of the semantic structure of the webpage.
- In the sense of human perception, people always view a web page as different semantic objects rather than a single object.
- Research efforts show that users always expect that certain parts of a webpage appear at certain positions on the page.
- **VIPS** (VIision-based Page Segmentation) is an algo that aims to extract the semantic structure of a webpage based on its visual presentation.
- The algo 1st extracts all the suitable blocks from the HTML DOM tree, then finds separators b/w these blocks.

contd.

MINING THE WEB'S LINK STRUCTURES

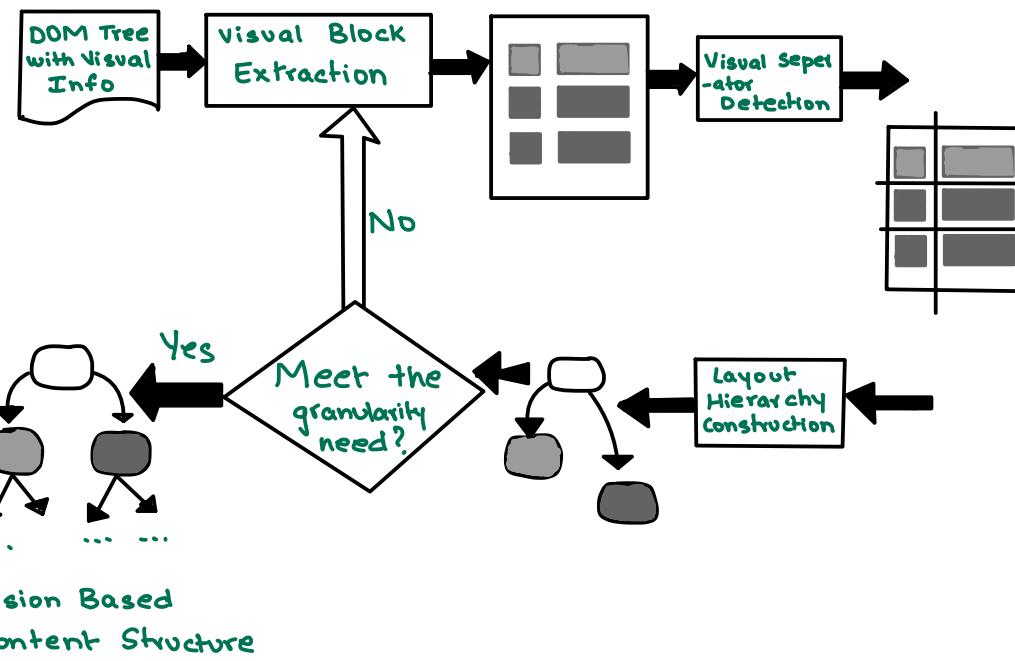
- In addition to retrieving pages that are relevant, you also hope that the pages retrieved (by a web search) will be of high quality, or authoritative on the topic.
 - The secrecy of authority is hiding in webpage linkages.
 - The web consists of not only pages, but also of hyperlinks pointing to another page
 - When an author of a web page includes a hyperlink on their webpage to another page, it is considered the author's endorsement of that page.
- The collective endorsement of a page by diff. authors may indicate the importance of the page & may naturally lead to the discovery of authoritative web pages.
- ∴ the tremendous amount of web linkage info provides rich info about the relevance, the quality, & the structure of the web's contents, & thus is a rich source for web mining.
 - A hub is 1 or a set of Web pages that provides collections of links to authorities.

In general, a good hub is a page that points to many good authorities; a good authority is a page pointed to by many good hubs.

- HITS** (Hyperlink-Induced Topic Search) - an algo. that uses hubs.

First, HITS uses the query term to collect a starting set of, say, 200 pages from an index-based search engine. These pages form the root set.

The process flow of vision-based page segmentation algo.



Second, a weight-propagation phase is initiated. This iterative process determines numerical estimates of hub & authority weights.

Finally, the algo outputs a short list of the pages with large hub weights, & the pages with large authority weights for the given search topic.

WEB USAGE MINING

- It mines weblog records to discover user access patterns of web pages.
- Analyzing & exploring regularities in weblog records can identify potential customers for electronic commerce, enhance the quality & delivery of Internet information services to the end user, & improve Web server system performance.
- A web server usually registers a log entry for every access of a web page. It includes the URL requested, IP address from which the request originated, & the timestamp.
- In developing techniques for web usage mining, we consider the following:
 - Although it is encouraging to imagine the various potential applications of weblog file analysis, note that the success of such applications depends on how much valid & reliable knowledge can be discovered from the large raw log data.
 - With the available URL, time, IP address, & web page content info, a multidimensional view can be constructed on the weblog database, & multidimensional OLAP analysis can be performed to find the top N users, top N accessed webpages, & so on.
 - Data mining can be performed on weblog records to find association patterns, sequential patterns, & trends of web accessing.

- Because weblog data provides info about what kind of users that will access what kind of web pages, weblog info can be integrated w/ web content & Web linkage structure mining to help Web page ranking, web document classification, & the construction of a multilayered Web Info base.

DM FOR FINANCIAL DATA ANALYSIS

Financial data collected in the banking & financial industry are often relatively complete, reliable, & of high quality, which facilitates systematic data analysis & data mining. A few cases:

- Design & construction of data warehouses for multidimensional data analysis & data mining - MultiD data analysis methods should be used to analyze the general properties of such data.
Eg: To view the debt & revenue changes by month, region or sector along w/ statistical info.
- Loan payment prediction & customer credit policy analysis - " " " are critical to the business of a bank. Many factors can strongly or weakly influence loan payment performance & customer credit rating. DM methods like attribute selection & attribute relevance ranking can help identify important factors & eliminate irrelevant ones.

DM FOR THE RETAIL INDUSTRY

The retail industry is a major application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption, & service. The quantity of data continues to expand rapidly, especially due to the increasing ease & availability of e-commerce.

Retail DM can help identify customer buying behaviors, discover customer shopping patterns & trends, reduce the cost of business, etc.

Some examples:

- Multidimensional analysis of sales, customers, products, time & region: The retail industry requires timely data regarding customer needs, product sales, trends, & fashions, as well as quality, cost & profit. It is :: imp. to provide powerful multiD analysis & visualization tools.
- Analysis of the effectiveness of sales campaigns: careful analysis " " can help improve company profits. Association analysis may disclose which items are likely to be purchased together w/ the items on sale, especially in comparison w/ the sales before the campaign.

- Classification & clustering of customers for targeted marketing - " " can be used for customer group identification & targeted marketing.
Eg: customers w/ similar behaviors regarding loan payments may be identified by multiD clustering.
- Detection of money laundering & other financial crimes - For this, it is important to integrate info from multiple databases, as long as they are potentially related to the study. Multiple data analysis tools can then be used to detect unusual patterns, such as large amounts of cash flow.

UNIT 4

REQUIREMENTS OF CLUSTERING

SCALABILITY - Many clustering algs work well on small datasets; however, a large database may contain millions of objects. Clustering on a sample of a given large dataset may lead to biased results.
∴ Highly scalable algos are needed

ABILITY TO DEAL WITH DIFF TYPES OF ATTRIBUTES

Many algos are designed to cluster numerical data. However, app's may require clustering other types of data, like binary, categorical & ordinal data or a mixture of these types.

DISCOVERY OF CLUSTERS WITH ARBITRARY SHAPE

Many algos determine clusters based on Euclidean or Manhattan distance measures. Algos based on such measures tend to find spherical clusters w/ similar size & density. However, a cluster could be of any shape. Algos have to be able to detect clusters of arbitrary shape.

ABILITY TO DEAL WITH NOISY DATA

Most databases contain outliers or missing, or erroneous data. Some clustering algos are sensitive to such data & may lead to clusters of poor quality

MINIMAL REQUIREMENTS FOR DOMAIN KNOWLEDGE TO DETERMINE INPUT

PARAMETERS - Many clustering algos require users to i/p certain parameters in cluster analysis. The clustering results can be quite sensitive to i/p parameters. Parameters are often difficult to determine.

INCREMENTAL CLUSTERING & INSENSITIVITY TO THE ORDER OF I/P RECORDS

Some algos cannot incorporate newly inserted data into clustering structures & must determine a new clustering from scratch. Some algos are sensitive to the order of i/p data, such that, it may return dramatically different clusterings depending on the order of the i/p objects.

HIGH DIMENSIONALITY - A database can contain several dimensions of attributes. Finding clusters in high-dimensional space is challenging.

CONSTRAINT-BASED CLUSTERING - Real world app's may need to perform clustering under various kinds of constraints.

TYPES OF DATA IN CLUSTER ANALYSIS

① **DATA MATRIX** - represents n objects, such as people, with p variables (attributes), such as age, height & so on.

The structure is in the form of a relational table, or n-by-p matrix

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

② **DISSIMILARITY MATRIX** - This stores a collection of proximities that are available for all pairs of n objects. It is represented by an n-by-n table

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

* DISTANCE MEASURES *

Euclidean distance

$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + \dots + (x_{in} - x_{jn})^2}$
 $i = (x_{i1}, x_{i2}, \dots x_{in})$ &
 $j = (x_{j1}, x_{j2}, \dots x_{jn})$ are 2
 n -dimensional data objects.

MANHATTAN (CITY BLOCK) DISTANCE

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{in} - x_{jn}|$$

JACCARD COEFFICIENT

$$d(i,j) = \frac{r+s}{q+r+s}$$

$$\text{sim}(i,j) = \frac{q}{q+r+s}$$

$q \rightarrow$ no. of vars that equal 1 for both objects ($i \& j$).

$r \rightarrow$ " " " 1 for i but are 0 for j .

$s \rightarrow$ " " " 0 for i but 1 for j .

$t \rightarrow$ " " " 0 for both ($i \& j$)

Dissimilarity between binary variables. Suppose that a patient record table (Table 7.2) contains the attributes *name*, *gender*, *fever*, *cough*, *test-1*, *test-2*, *test-3*, and *test-4*, where *name* is an object identifier, *gender* is a symmetric attribute, and the remaining attributes are asymmetric binary.

For asymmetric attribute values, let the values *Y* (yes) and *P* (positive) be set to 1, and the value *N* (no or negative) be set to 0. Suppose that the distance between objects (patients) is computed based only on the asymmetric variables. According to Equation (7.10), the distance between each pair of the three patients, Jack, Mary, and Jim, is

$$d(i,j) = \frac{r+s}{q+r+s}$$

$$\begin{aligned} d(Jack, Mary) &= \frac{0+1}{2+0+1} = 0.33 \\ d(Jack, Jim) &= \frac{1+1}{1+1+1} = 0.67 \\ d(Mary, Jim) &= \frac{1+2}{1+1+2} = 0.75 \end{aligned}$$

A relational table where patients are described by binary attributes.

name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
:	:	:	:	:	:	:	:

These measurements suggest that Mary and Jim are unlikely to have a similar disease because they have the highest dissimilarity value among the three pairs. Of the three patients, Jack and Mary are the most likely to have a similar disease.

until eventually each object is in 1 cluster, or until a termination condition holds.

contd.

CLUSTERING METHODS

* PARTITIONING :

Given a database of n objects, a partitioning method constructs K partitions of the data, where each partition represents a cluster, & $K \leq n$. That is, it classifies the data into K groups, which satisfy:

- (1) each group must contain at least 1 object.
- (2) each object must belong to exactly

1 group

The 2nd requirement can be relaxed in some fuzzy partitioning techniques.

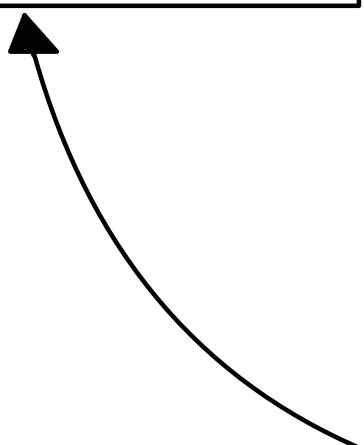
The general criterion of a good partitioning is that objects in the same cluster are "close" or related, whereas objects of different clusters are "far apart" or very different.

* HIERARCHICAL :

This creates a hierarchical decomposition of the given data objects. It can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

Agglomerative/bottom-up approach - starts w/ each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into 1, or until a termination condition holds.

Divisive/top-down" - Starts w/ all objects in the same cluster. In each iteration, a cluster is split up into smaller clusters,



contd.

Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone. This technique doesn't allow for correction of erroneous decisions.

* DENSITY-BASED:

Their general idea is to continue growing the given cluster as long as the density in the "neighborhood" exceeds some threshold. Such a method can be used to filter out noise (outliers) & discover clusters of arbitrary shape.

DBSCAN & DENCLUE.

* GRID-BASED:

These methods quantize the object space into a finite no. of cells that form a grid structure. All of the clustering operations are performed on the grid structure. The main advantage of this is its fast processing time.

* MODEL-BASED:

These methods hypothesize a model for each of the clusters & find the best fit of the data to the given model. This algo may locate clusters by constructing a density function that reflects the spatial distribution of the data points

* CLUSTERING HIGH-D DATA:

As the no. of dimensions increases, the data becomes increasingly sparse so that the distance measurement b/w pairs of points become meaningless & the avg density of the points anywhere in the data is likely to be low.

CLIQUE & PROCLUS are 2 influential subspace clustering methods, which search for clusters in subspaces of the data, rather than over the entire data space.

* CONSTRAINT-BASED CLUSTERING:

Performs clustering by incorporating user-specified or application-oriented constraints. Semi-supervised clustering employs pairwise constraints to improve the quality of the resulting clustering.

