DATA MINING - MINING WORLD WIDE WEB

The World Wide Web contains huge amounts of information that provides a rich source for data mining.

Challenges in Web Mining

The web poses great challenges for resource and knowledge discovery based on the following observations −

- **The web is too huge** − The size of the web is very huge and rapidly increasing. This seems that the web is too huge for data warehousing and data mining.
- **Complexity of Web pages** − The web pages do not have unifying structure. They are very complex as compared to traditional text document. There are huge amount of documents in digital library of web. These libraries are not arranged according to any particular sorted order.
- **Web is dynamic information source** − The information on the web is rapidly updated. The data such as news, stock markets, weather, sports, shopping, etc., are regularly updated.
- **Diversity of user communities** − The user community on the web is rapidly expanding. These users have different backgrounds, interests, and usage purposes. There are more than 100 million workstations that are connected to the Internet and still rapidly increasing.
- **Relevancy of Information** − It is considered that a particular person is generally interested in only small portion of the web, while the rest of the portion of the web contains the information that is not relevant to the user and may swamp desired results.

Mining Web page layout structure

The basic structure of the web page is based on the Document Object Model *DOM*. The DOM structure refers to a tree like structure where the HTML tag in the page corresponds to a node in the DOM tree. We can segment the web page by using predefined tags in HTML. The HTML syntax is flexible therefore, the web pages does not follow the W3C specifications. Not following the specifications of W3C may cause error in DOM tree structure.

The DOM structure was initially introduced for presentation in the browser and not for description of semantic structure of the web page. The DOM structure cannot correctly identify the semantic relationship between the different parts of a web page.

Vision-based page segmentation *VIPS*

- The purpose of VIPS is to extract the semantic structure of a web page based on its visual presentation.
- Such a semantic structure corresponds to a tree structure. In this tree each node corresponds to a block.
- A value is assigned to each node. This value is called the Degree of Coherence. This value is assigned to indicate the coherent content in the block based on visual perception.
- The VIPS algorithm first extracts all the suitable blocks from the HTML DOM tree. After that it finds the separators between these blocks.
- The separators refer to the horizontal or vertical lines in a web page that visually cross with no blocks.
- The semantics of the web page is constructed on the basis of these blocks.

The following figure shows the procedure of VIPS algorithm −