Unit 6
Web Analytics

- WEB MINING –
- Web mining (or Web data mining) is the process of discovering intrinsic relationships (i.e., interesting and useful information) from Web data, which are expressed in the form of textual, linkage, or usage information.
- Web analytics is primarily Web site usage data focused, Web mining is inclusive of all data generated via the Internet, including transaction, social, and usage data.
- While Web analytics aims to describe what has happened on the Web site (employing a predefined, metrics-driven descriptive analytics methodology), Web mining aims to discover previously unknown patterns and relationships.
- Web mining is divided into three main areas: Web content mining, Web structure mining, and Web usage mining.

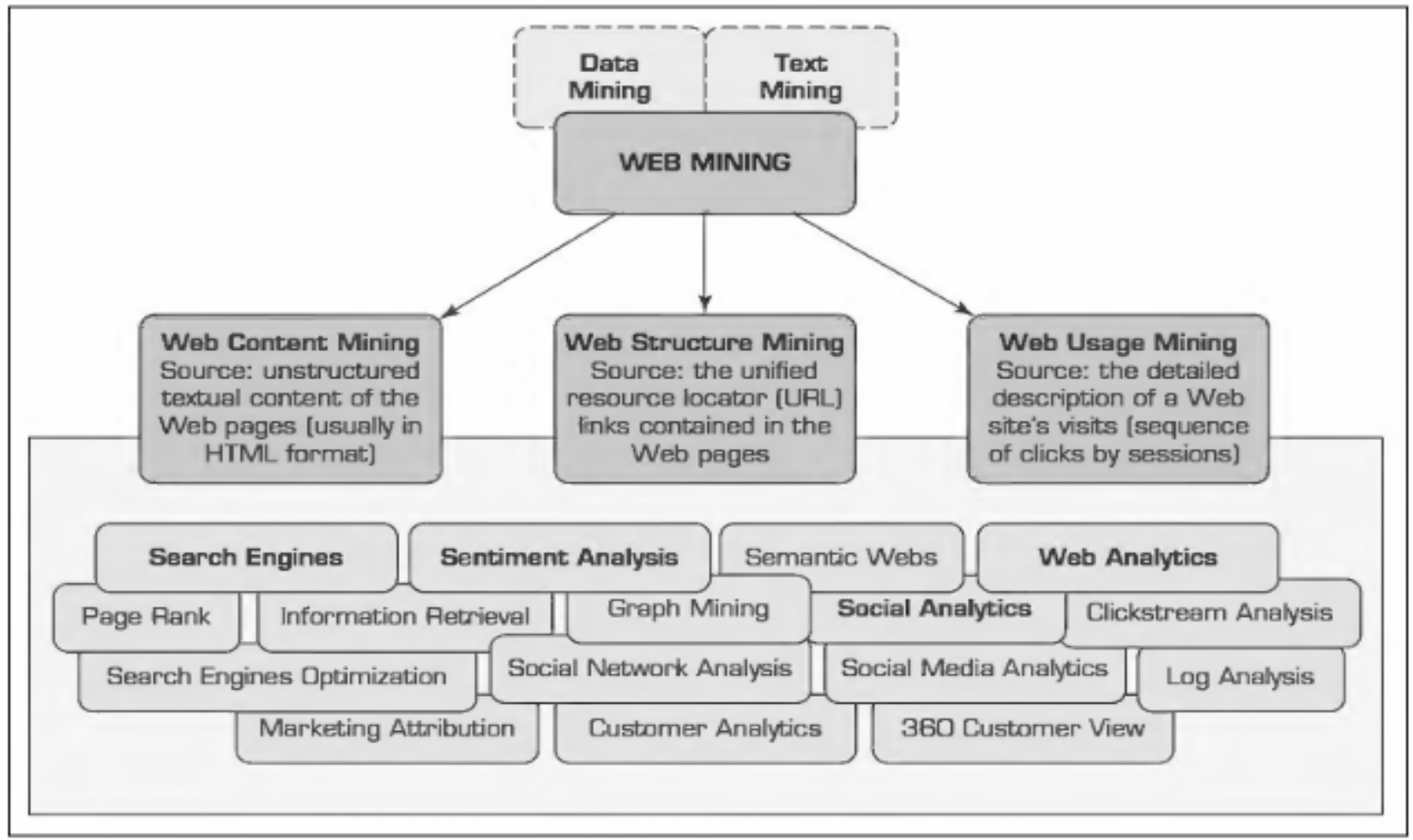

FIGURE 8.1 **A Simple Taxonomy of Web Mining.**

- **Web content and Web structure Mining**
- Web content mining-
- It refers to the extraction of useful information from Web pages
- The documents may be extracted in some machine-readable format so that automated techniques can extract some information from these Web pages.
- Web crawlers (also called spiders) are used to read through the content of a Web site automatically.
- The information gathered may include document characteristics similar to what are used in text mining, but it may also include additional concepts, such as the document hierarchy.
- Such an automated (or semi automated) process of collecting and mining Web content can be used for competitive intelligence (collecting intelligence about competitors' products, services, and customers).
- It can also be used for information/ news/ opinion collection and summarization, sentiment analysis, automated data collection, and structuring for predictive modeling.
- Web content mining can also be used to enhance the results produced by search engines.

- **Web structure mining and its terms-**
- **Web Structure Mining Definition**: It's about finding useful info from the links in web pages.
- **Identifying Authority**: Helps find important pages (authoritative pages) and hubs in a network.
- **Cornerstones of Page-Rank Algorithms**: These algorithms (like Google's) rely on web structure mining to rank pages.
- **Understanding Popularity**: Links to a page show its popularity.
- **Depth of Coverage**: Links within a page or site indicate how much it covers a specific topic.
- **Understanding Relationships**: Analyzing links helps see how web pages are connected, revealing web communities or groups.

- **SEARCH ENGINES-**
- A search engine is a software program that searches for documents (Internet sites or files) based on the keywords (individual words, multi-word te1ms, or a complete sentence) that users have provided that have to do with the subject of their inquiry.
- **Search Engine Goal**: Find and return documents/pages that best match user queries.
- **Metrics**: Effectiveness (finding right documents) and efficiency (speed of response) are key.
- **Trade-off**: Improving one metric often worsens the other; they work in reverse.
- **User Expectations**: Search engines may prioritize one metric over the other based on user expectations.
- **Ideal**: The best search engines excel in both effectiveness and efficiency simultaneously.

- **Anatomy of a Search Engine -**

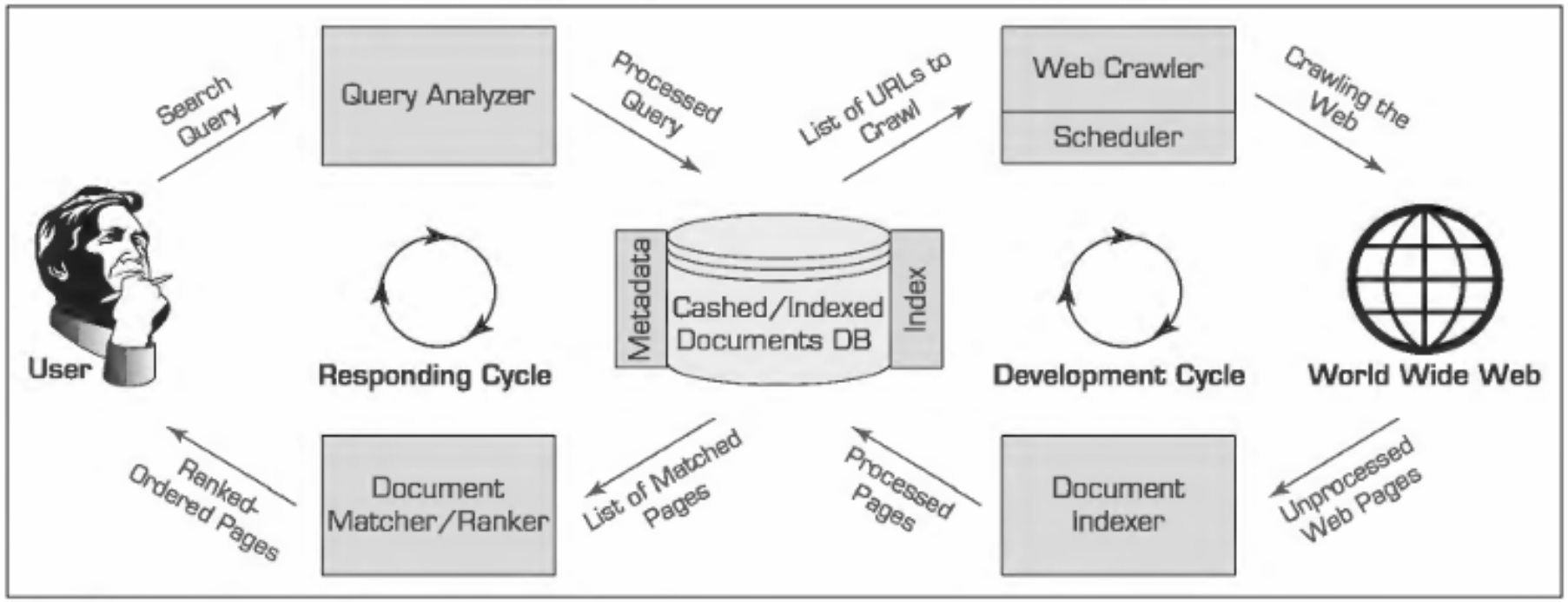

FIGURE 8.2 Structure of a Typical Internet Search Engine.

- **Anatomy of a Search Engine** -
- A search engine system is composed of two main cycles: a development cycle and a responding cycle.
- Development Cycle-
- The two main components of the development cycle are the Web crawler and document indexer.
- The purpose of this cycle is to create a huge database of documents/pages organized and indexed based on their content and information value.
- Search engines "cashes the Web" into their database, and uses the cashed version of the Web for searching and finding.
- Once created, this database allows search engines to rapidly and accurately respond to user queries.
- **Web Crawler Function**: It's software that systematically browses the web to find and fetch web pages.
- Copying Pages: Crawlers often copy visited pages for later processing by search engine functions.
- Starting Point: It begins with a list of URLs (seeds) to visit, sourced from webmaster submissions or internal hyperlinks.
- Identifying Links: As it visits URLs, it finds and adds hyperlinks to its list of URLs to visit (scheduler).
- Recursion and Policies: URLs in the scheduler are recursively visited based on policies set by the search engine.
- Download Limitations: Due to the vast number of web pages, crawlers can only download a limited number within a given time, so they prioritize downloads.

Document Indexer

- As the documents are found and fetched by the crawler, they are stored in a temporary staging area for the document indexer to grab and process.
- The document indexer is responsible for processing the documents (Web pages or document files) and placing them into the document database.

- **Step 1: Preprocessing Documents**:
  - Convert fetched documents to a standard format.
  - Separate different content types (text, hyperlink, image).
  - Format and store for further processing.
- **Step 2: Parsing Documents**:
  - Use text mining tools to analyze documents.
  - Parse documents to identify index-worthy words/terms.
  - Apply tokenization rules to extract words/terms.
  - Correct spelling errors and anomalies.
  - Eliminate non-discriminating words (stop words).
  - Apply stemming to reduce words to root forms.
  - Identify synonyms and homonyms.
  - Prepare documents for indexing.
- **Step 3: Creating Term-by-Document Matrix**:
  - Establish relationships between words/terms and documents/pages.
  - Assign weights to words/terms based on occurrence.
  - Weighting schemas can be binary or frequency-based.
  - Common schema: TF/IDF algorithm (Term Frequency/Inverse Document Frequency).
  - Calculate weights based on word/term frequency in documents.
  - Consider document discrimination and domain relevance.
  - Create term-by-document index file with calculated weights.

- **Response Cycle Components**:
  Query Analyzer and Document Matcher/Ranker are main components.
- **Query Analyzer**:
  - Receives search request and standardizes it.
  - Parses query into words/terms.
  - Tasks include tokenization, stop word removal, stemming, and disambiguation.
  - Similar to document indexer in processing.
- **Document Matcher/Ranker**:
  - Matches structured query data against document database.
  - Finds and ranks most relevant documents/pages.
  - Proficiency crucial for search engine comparison.
- **Historical Evolution**:
  - Early search engines used simple keyword matching.
  - Returned ordered documents based on matched words/terms and weights.
  - Quality of results not optimal.
- **Introduction of PageRank**:
  - Developed by Google creators in 1997.
  - Algorithmic approach to rank documents based on relevance and importance.
  - Improved quality and usefulness of search results.

- # SEARCH ENGINE OPTIMIZATION
- **Definition of SEO**:
  - SEO is the intentional effort to improve a website's visibility in unpaid search engine results.
- **Impact of Rankings**:
  - Higher ranking and more frequent appearance in search results lead to increased website traffic.
- **Considerations in SEO**:
  - Understanding search engine operations, user search behavior, and preferences.
  - Optimization involves content editing, HTML, and coding to enhance relevance and ease of indexing.
- **SEO Types:**
- On Page- It Refers to optimizing the parts of your website you control like content ,HTML,Title of page,URL etc.
- Off page- It focuses on the ranking factors that occur outside of your website like brand mentions and backlinks.
- **Importance of Ranking**:
  - Being indexed by search engines isn't sufficient; ranking higher than competitors is crucial for business success.
- **Methods to Improve Ranking**:
  - Cross-linking between website pages to provide more links.
  - Writing content with relevant keywords to match search queries.
  - Regularly updating content to attract search engine crawlers.
  - Adding relevant keywords to metadata like title tags and meta descriptions.
  - URL normalization and canonical link elements ensure proper indexing and link popularity.
- # Methods for Search Engine Optimization-
- **White Hat vs. Black Hat SEO**:
  - White hat techniques follow search engine guidelines and focus on user-centric content.
  - Black hat techniques attempt to manipulate rankings and involve deception.

- **White Hat SEO**:
- It includes optimizing your websites by following the restrictions imposed by Search Engines.
- Methods-
- Content Optimize
- Quality Content
- Relevant internal links
- Relevant strong Backlinks
- Better UX
- **Black Hat SEO**:
- It includes unethical techniques which are disapproved by Search Engines.
- Methods-
- Keyword stuffing
- Duplicate Content
- Clocking
- Hidden text links
- Link farming
- **SEO Challenges**:
  - Search engine algorithms constantly change.
  - No guarantee of continued traffic or referrals.
- **Approaches to SEO**:
  - Hire SEO specialists to adapt to search engine practices.
  - Pay for sponsored listings on search engines.
  - Diversify traffic sources to reduce dependence on search engines.
- **E-commerce Focus**:
  - Prioritize maximizing customer transactions over mere visitor numbers.
  - Case study emphasizes using data analysis to enhance conversion rates.

- **WEB ANALYTICS MATURITY MODEL AND WEB ANALYTICS TOOLS –**
- The concept of maturity in business refers to the progression from ad hoc practices to structured, optimized processes.
- Maturity models provide a framework for assessing and improving organizational proficiency in specific areas. Examples include the TDWI BI Maturity Model for data warehousing and the business analytics maturity model for advancing from descriptive to prescriptive analytics.
- For web analytics, Stephane Hamel proposed a comprehensive model with six dimensions and proficiency levels ranging from Analytically Impaired to Analytical Competitor.
- These levels depict the organization's analytical maturity and capabilities:
  1. Analytically Impaired: Basic use of out-of-the-box tools with limited resources and scope.
  2. Initiated: Metrics used to optimize specific areas, but processes may be outdated.
  3. Operational: Key performance indicators aligned with strategic objectives; multidisciplinary teams and segmentation used.
  4. Integrated: Correlation of online and offline data for comprehensive optimization; insight reaches executive level.
  5. Competitor: Strong analytical culture with predictive modeling and enterprise-wide use.
  6. Addicted: Deep strategic insight and continuous improvement beyond online channels.
- These levels can be mapped across six dimensions to create an organization's maturity model, often represented as a spider diagram.

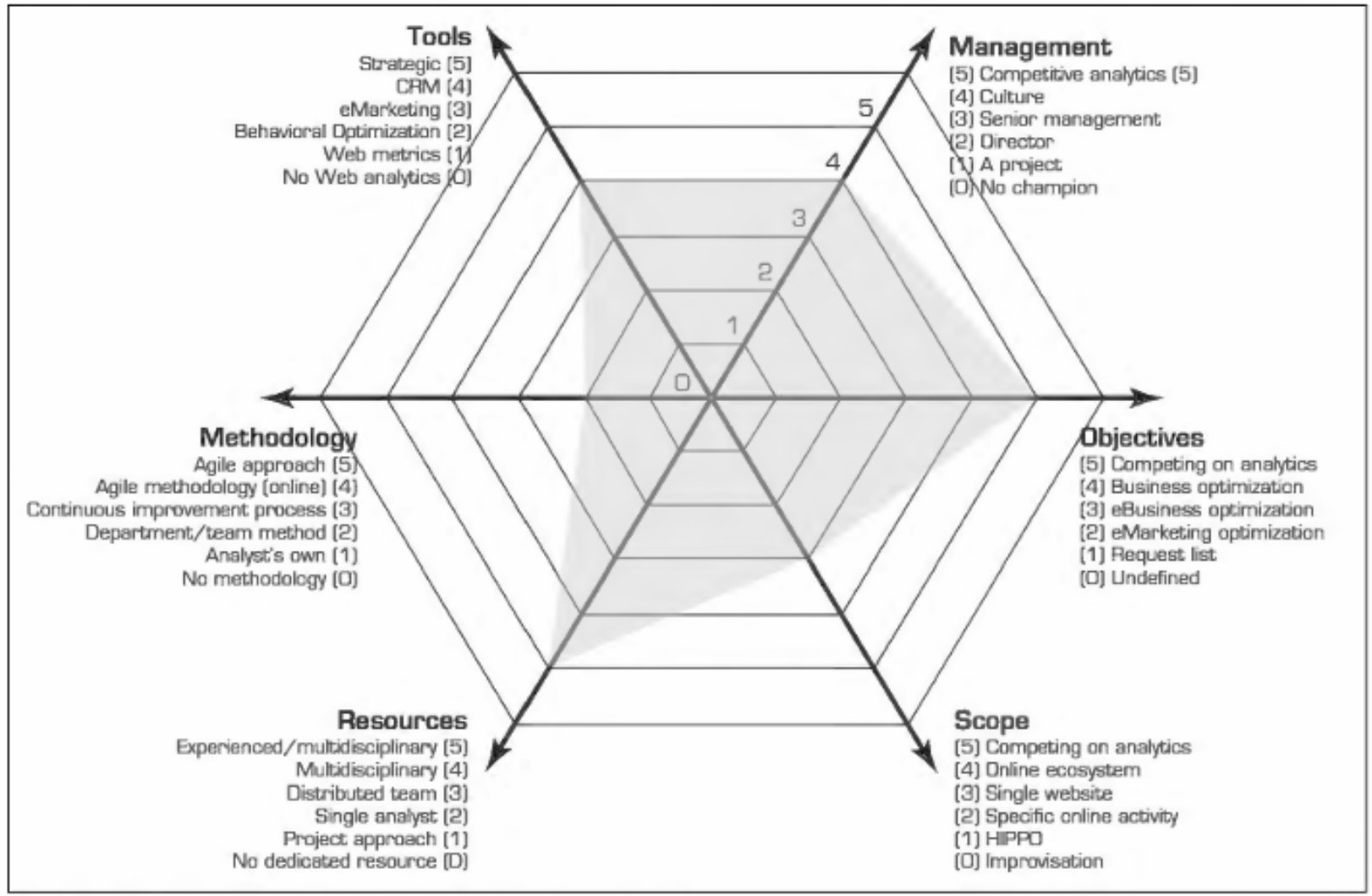

FIGURE 8.7 A Framework for Web Analytics Maturity Model.

- **SOCIAL ANALYTICS AND SOCIAL NETWORK ANALYSIS –**

- Social network analysis (SNA) studies the structure and dynamics of social entities, identifying patterns and influential entities.
- It emerged from various disciplines like social psychology, sociology, statistics, and graph theory, with formalization dating back to the 1950s.
- SNA is now a major paradigm in business analytics, consumer intelligence, and contemporary sociology.
- Social networks are theoretical constructs used to study relationships between individuals, groups, organizations, or entire societies.
- They are self-organizing, emergent, and complex systems where globally coherent patterns arise from local interactions.
- Types of social networks relevant to business activities include communication networks, community networks, criminal networks, and innovation networks.
- Communication networks involve the transfer of information and are utilized by telecommunication companies for optimizing business practices.
- Community networks, once based on geographic location, now include online communities generated through social networking tools.
- Criminal networks are studied in criminology to understand and prevent illegal activities, with modern technology aiding law enforcement efforts.
- Innovation networks in business focus on the spread and use of ideas among members, examining how network structure influences innovation diffusion and behavior.

- **Social Network Analysis Metrics-**
- Social Network Analysis (SNA) systematically examines social networks, viewing relationships through network theory.
- Networks consist of nodes (individuals or organizations) and ties/connections representing relationships like friendship or organizational position.
- Social network diagrams visually represent networks with nodes as points and ties as lines.

- Tutorial 6

1. Explain web Mining and Taxonomy of web mining in detail.
2. Explain Search Engine and its structure in detail.
3. Explain Search Engine Optimization and its Methods.
4. Explain Web Analytics maturity of model in detail with diagram.