DHANALAKSHMI COLLEGE OF ENGINEERING, CHENNAI

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CS6010 - SOCIAL NETWORK ANALYSIS

UNIT I : INTRODUCTION

PART – A (2 Marks)

1. What is the main function of semantic web?

- Semantic web is a collaborative effort by the W3C and it is used to promote the common formats for data.
- It removes ambiguity from the form of data being represented on the World Wide Web.
- Semantic web allows the inclusion of the semantic content that describes the format in the web pages.
- It converts the unstructured content on the web and makes it more structured for the daily use.
- It consists of the web of data to represent it on the website and build it using the Resource Description Framework.

2. Why is Semantic Web used in current system?

- Semantic Web provides framework on which the applications can be made and developed using the tools.
- It allows the data to be shared and reused between many applications and other enterprise level applications.
- W3C also known as World Web Consortium uses the development libraries for the Semantic web standards.
- The semantic web describes the web of data that can be directly or indirectly gets executed on the client machine.
- This uses disambiguity principle that doesn't allow the ambiguous solution to be provided with the system.

3. What is the purpose of Semantic Web?

- Semantic web allowed the user to find, share and combine the information to transfer it from one place to another very easily.
- It allows the working of current web and makes it more secure and usable by showing the information.
- The users are allowed to use the Web for carrying out the tasks of finding the folder or categories and it just makes it easy.
- Semantic web provides the instructions for the machine to execute the tasks by providing the interpreter that can interpret it.
- Machines can perform the task provided by the Semantic Web and it involves finding, combining and acting on the information that is present on the web.

4. Why is Semantic Web so useful for the development of web?

- Semantic web provides the instructions for the machines that can be understood and the response should be produced from the system.
- Semantic web provides an interpreter that can interpret the instructions to the machine and translate them further to make it in human readable form with their meaning.
- Semantic web provides the information regarding the data format that requires understanding of the semantically structured data.
- Semantic web allows the user to use the tools to analyze the data on the web and it also have the content, links and other transactions between the people.
- It provides the applications in many areas like blogging, publishing, etc. This way the applications can be created and circulated around.

5. Why is Semantic web regarded as integrator?

- Integrator allows more than one data to be integrated using different content and information in the system.
- Semantic web provides the integrator that runs across different platforms for the applications that need to be published on the web.
- Semantic web provides the semantics or the metadata for the web that can be used to represent the status model reflecting the current technologies.
- It provides and support different fields to be integrated in one technology and can be worked upon.
- It provides tools that can be supported by applications and integrated in a platform that is used to create the applications.

6. What are the limitations of HTML?

- HTML is also known as HyperText Markup Language provides the creation of the web pages.
- The HTML pages are the documents that can be read by the server, and are not the best fit to be read by humans.
- HTML forms have the dependency on scripting languages and it results in complex document creation that consumes more time.
- HTML doesn't initialize the form data properly and doesn't make it easier for the users to enter the information once.
- HTML is having some limitations with the use of forms that doesn't allow encoding formats, urlencoded or multipart forms.

7. Why is HTML used in Semantic web?

- HTML is a standard language that communicates between the server and the client's system.
- The files that are given on the computer can be divided into human and machine readable form.
- Most of the documents are written in HTML form and it uses multimedia objects in a better way by using the images and forms.
- HTML is a standard output method for responding to the client's request and respond accordingly.
- HTML provides a way to generate the response of the web when the client request any data from the server.

8. What is the limitation of HTML forms?

- HTML forms are hard to initialize the data of the form and it provides no user experience as user needs to remember the form information.
- HTML form provides a unique control of defining the data that is initially being filled up.
- It uses the small bits of initialization data that is present in the overall document while defining the control.
- A new form needs to be constructed to fill the form again as it holds no data as the backup to fill the information with.
- A template replacement facility is not being provided on application servers that stores the data and doesn't allow the users to fill up again and again.

9. What are the design flaws involved in HTML forms?

- The design flaws are involved in HTML as it provides one step process i.e. from client to server.
- The processing finishes there and it doesn't provide further processes to be done on the forms.
- Forms involve the complicated path to traverse and HTML failed to make the traversing easier.
- Management of the HTML forms isn't easy as it requires reinterpreting the data format at every stage of the life cycle.
- HTML forms are not used due to its bad management and the provisions that are being provided for creations and modification.

10. What is being provided by Metadata tags?

- Metadata tags provide the keywords that are used for the search engine to make the website or the web page search engine friendly.
- It is a method to categorize the content of the web pages on the search engine so that it can be easily found by the browsers.
- Metadata tags are represented as:
  <meta name="keywords" content="computing, computer, comp" />
  <meta name="description" content="Hello world" />
  <meta name="author" content="Rohit Kumar" />
- Metadata tags provide good description in the tags and allow the content to be displayed for better performance of the web pages.

11. What are the activities performed using HTML?

- HTML is a tool that allows the rendering of the web pages and creation of it using the editor.
- The web page can be created with easy to use tags, browser compatible code and list of items.
- Simple documentation can be created using the tools that is being provided by HTML.
- Images can be displayed in variety of ways and text can be made floated using the special tags defined in HTML version.
- The pieces of information can be combined together to describe the items and other items on different web pages.

12. What is the function of semantic HTML?

- Semantic HTML provides the traditional methodologies to work and markup the code according to the guidelines.
- It doesn't specify the layout details in which the HTML needs to be presented or written.
- Semantic HTML uses the old tags like that denotes emphasis rather than <i> tag that used to denote italics.
- Layout details are web browser dependent and it is placed according to the combination of Cascading style sheets.
- The semantic of the objects are also not described by the use of items and by using their sales and price details.

13. What is the use of Semantic Web solutions?

- Semantic web solutions provide publishing methodologies that is designed for the data.
- Resource description framework or RDF is also used and included in the semantic web solutions.
- The technologies used in are being combined to provide the descriptions and replacement of the web documents.
- Web ontology languages are used to describe the links between various texts and languages.
- It includes a manifest that consists of all the descriptive data stored in the web-accessible databases.
- The markup is used within the documents that are related to XML and the layout is being rendered using it.

14. What are the functions of machine readable descriptions?

- Machine readable descriptions allow the managers to manage the content by adding the meaning to the content used.
- It provides a structured knowledge of the system for which the content is being written.
- Machine processes the knowledge of changing the content using the processes by reasoning and inference.
- It provides meaningful resources and results that can be used to perform the information task automatically and more easily.
- Research gathering information is being provided in the semantic web solutions and provides the content to be written accordingly.

15. What are the examples of using the non-semantic web page?

- To make the web page more meaningful by adding the content or performing the automated tasks semantic web is used.
- Non-Semantic web page is used to provide the easy to use tags in there and get the functions performed to execute the tasks.
- The tags that are used:
  <item>cat</item>
- This tag provides an easy way to represent the information without following a pattern like semantic web pages.
  Semantic web pages are described like for the same web page content:
  <item rdf:about="http://hello.org/Cat">Cat</item>

16. What are the ways in which the web page can be accessed?

- The web page requires some functions that allow accessing of it in an easy and comfortable way.
- There are three ways in which the web page can be accessed and the data can be retrieved from it.
- The three ways are as follows:

  - The URL first should always point to the data that needs to be represented or accessed.
  - Accessing of the URL should provide the data back to the client that has requested for it.
  - The relationship between the data and the server is represented in such a way that it points in additional URLs as well.
  - The other URLs consist of the data residing on their server through which it can be accessed.

17. What are the challenges faced by the technology?

- The challenge that is being provided by semantic web includes the following:
- Vastness: this includes the large group of pages that is being accessed by the users using the existing technology.
- This consists of any automated system that is good in reasoning and deals with the very high inputs.
- Vagueness: it occurs due to the queries that are being provided by the content providers.
- If the query terms are matched then the knowledge can be combined together to find the knowledge.
- Uncertainty: includes uncertain value that can provide the correspondence using the different probability.
- Inconsitency: is the very big challenge that provides logical contradictions between the ontologies.
- It combines the resources to answer the questions that are being raised by the theories and sources of it.

18. What are the different components used in Semantic web?

- Semantic web uses different formats and technologies that enables it to provide great extent on the web.
- Semantic web provides the collection of data that are having relationship with each other.
- It also has the components that are enabled by technologies and provide the description of concepts, terms and relationships.
- The components that are used in semantic web follows:
- Resource Description Framework (RDF): this is used as a method to define the information and general queries of the system.
- RDF Schema (RDFS): this consists of the file data type format and helps in storing the data.
- Simple Knowledge Organization System (SKOS)
- SPARQL, an RDF query language

19. What is the function of Semantic Web Stack?

- Semantic web stack is used to provide architecture for the Semantic web and it deals in relationships related to the components.
- Semantic web stack provides the functions to be used in the components and provides the content structure.
- Syntax of the XML can be provided within the documents and it has the association with no semantics having the meaning of the content.
- XML is represented as the major component used with the technologies and it provides the process to be made standardized.
- Semantic web stack uses the programs and store it in the stack so the technologies are gathered at one place and used for the benefit to provide something easy and useful.

20. Explain the components of the semantic web in detail?

- The components used in semantic web are as follows:
- XML schema is used to store the data but it also provides and restricts the user to use the structure and content of the element being in use.
- RDF provides an expressing data models. It also deals in the relationships between them.
- The model based in RDF can be represented in much syntax that meets the standard web quality.
- RDF schema provided added functionality that describes the properties and classes using the based resources.
- The semantic web is used in a generalized-hierarchy way that can be used with clients and properties.

21. What are the Security Design Principles used in Web Security?

- The Security Design Principles used in Web Security are as follows:
- Least Privilege: this provides the security for the system and provides a way to limit the resources given to a process when it starts.
- Defence in Depth: the defence of the website is to provide the depth in the content such that it becomes hard for someone to break it.
- Secure Weakest Link: this way the security can't be breached as most of the attacks will be on the weak links only.
- Fail-safe Stance: provide a way to have the security such that if one security fails then it will have the model that will support it.
- Secure By Default: there are security that can be provided by default to secure the websites from being hacked.
- Simplicity: the design principles of the website should be simple to use and it should be easy customizable.
- Usability: the design of the website should be usable such that anyone can use the website.

PART – B (16 Marks)

1. What are the limitations of current Web? Explain the development of semantic Web and the emergence of Social Web.
2. Briefly explain the development of Social Network Analysis.
3. Enumerate the static properties of social networks.
4. Explain the dynamic properties of social networks.
5. Illustrate the Global structure of networks with an example.
6. Discuss in detail about the macro-structure of social networks.
7. Enumerate the different dimensions of social capital and their related concepts and measures.
8. Briefly explain the following:
    a) Electronic discussion networks
    b) Blogs and online communities
    c) Web-based Networks
    d) Personal Networks
9. Explain the statistical properties of social network analysis.
10. Discuss the business applications of Social Network Analysis.

UNIT – II : MODELLING, AGGREGATING AND KNOWLEDGE REPRESENTATION

PART – A (2 Marks)

1. What are the uses of statistics in data mining?

   Statistics is used to

   - to estimate the complexity of a data mining problem;
   - suggest which data mining techniques are most likely to be successful; and
   - identify data fields that contain the most "surface information".

2. What are the factors to be considered while selecting the sample in statistics?

   The sample should be

   - Large enough to be representative of the population.
   - Small enough to be manageable.
   - Accessible to the sampler.
   - Free of bias.

3. Name some advanced database systems.

   Object-oriented databases,
   Object-relational databases.

4. Name some specific application oriented databases.

   - Spatial databases,
   - Time-series databases,
   - Text databases and multimedia databases.

5. What is meant by relational databases?

   A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.

6. What is meant by transactional databases?

   A transactional database consists of a file where each record represents a transaction. A

   Transaction typically includes a unique transaction identity number (trans_ID), and a list of the items making up the transaction.

7. What is Spatial Databases?

   Spatial databases contain spatial-related information. Such databases include geographic (map) databases, VLSI chip design databases, and medical and satellite image databases. Spatial data may be represented in raster format, consisting of n-dimensional bit maps or pixel maps.

8. What is Temporal Database?

Temporal database store time related data .It usually stores relational data that include time related attributes. These attributes may involve several time stamps, each having different semantics.

9. What are Time-Series databases?

A Time-Series database stores sequences of values that change with time, such as data Collected regarding the stock exchange.

10. Why machine learning is done?

- To understand and improve the efficiency of human learning.
- To discover new things or structure that is unknown to human beings.
- To fill in skeletal or computer specifications about a domain.

11. Give the components of a learning system.

- Critic
- Sensors
- Learning Element
- Performance Element
- Effectors
- Problem generators.

12. What are the steps in the data mining process?

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge representation

13. What is data cleaning?

Data cleaning means removing the inconsistent data or noise and collecting necessary information

14. What is data mining?

Data mining is a process of extracting or mining knowledge from huge amount of data.

15. What is meant by pattern evaluation?

Pattern evaluation is used to identify the truly interesting patterns representing knowledge based on some interesting measures.

16. What is knowledge representation?

Knowledge representation techniques are used to present the mined knowledge to the user.

17. What is Visualization?

Visualization is for depiction of data and to gain intuition about data being observed. ItAssists the analysts in selecting display formats, viewer perspectives and data representation schema

18. What is Spatial Visualization?

Spatial visualization depicts actual members of the population in their feature space

19. What is Descriptive and predictive data mining?

Descriptive data mining describes the data set in a concise and summertime manner and Presents interesting general properties of the data. Predictive data mining analyzes the data in order to construct one or set of models and attempts to predict the behavior of new data sets.

20. What is Data Generalization?

It is process that abstracts a large set of task-relevant data in a database from a relatively low conceptual to higher conceptual levels 2 approaches for Generalization

a. Data cube approach
b. Attribute-oriented induction approach

21. What is meant by Attribute Oriented Induction?

These method collets the task-relevant data using a relational database query and then perform generalization based on the examination in the relevant set of data.

22. What is bootstrap?

An interpretation of the jack knife is that the construction of pseudo value is based on Repeatedly and systematically sampling with out replacement from the data at hand. This lead to generalized concept to repeated sampling with replacement called bootstrap.

23. What is meant by the view of statistical approach?

Statistical method is interested in interpreting the model. It may sacrifice some performance to be able to extract meaning from the model structure. If accuracy is acceptable then the reason that a model can be decomposed in to revealing parts is often more useful than a 'black box' system, especially during early stages of investigation and design cycle.

24. What is deterministic models?

Deterministic models, which takes no account of random variables, but gives precise, fixed reproducible output.

25. What is menat by systems and models?

System is a collection of interrelated objects and Model is a description of a system. Models are abstract, and conceptually simple.

26. What are the ways the models are explored?

All things being equal, the smallest model that explains the observations and fits the objectives that should be accepted. In reality, the smallest means the model should optimizes a certain scoring function (e.g. Least nodes, most robust, least assumptions)

27. What is clustering?

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

28. What are the requirements of clustering?

- Scalability
- Ability to deal with different types of attributes
- Ability to deal with noisy data
- Minimal requirements for domain knowledge to determine input parameters
- Constraint based clustering
- Interpretability and usability

29. State the categories of clustering methods?

- Partitioning methods
- Hierarchical methods
- Density based methods
- Grid based methods
- Model based methods

30. What is linear regression?

In linear regression data are modeled using a straight line. Linear regression is the simplest form of regression. Bivariate linear regression models a random variable Y called response variable as a linear function of another random variable X, called a predictor variable.

$$Y = a + b X$$

31. State the types of linear model and state its use?

Generalized linear model represent the theoretical foundation on which linear regression can be applied to the modeling of categorical response variables. The types of generalized linear model are

Logistic regression

Poisson regression

32. Write the preprocessing steps that may be applied to the data for classification and prediction.
    - Data Cleaning
    - Relevance Analysis
    - Data Transformation

33. What is data classification?

    It is a two-step process. In the first step, a model is built describing a pre-determined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. In the second step the model is used for classification.

34. What is a “decision tree”?

    It is a flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. Decision tree is a predictive model. Each branch of the tree is a classification question and leaves of the tree are partition of the dataset with their classification.

35. Where are decision trees mainly used?

    Used for exploration of dataset and business problems Data preprocessing for other predictive analysis Statisticians use decision trees for exploratory analysis

36. What is Association rule?

    Association rule finds interesting association or correlation relationships among a large set of data items, which is used for decision-making processes. Association rules analyzes buying patterns that are frequently associated or purchased together.

37. What is menat by support?

    Support is the ratio of the number of transactions that include all items in the antecedent and consequent parts of the rule to the total number of transactions. Support is an association rule interestingness measure.

38. What is confidence?

    Confidence is the ratio of the number of transactions that include all items in the consequent as well as antecedent to the number of transactions that include all items in antecedent. Confidence is an association rule interestingness measure.

39. How are the association rules mined from large databases?
    - Association rule mining is a two-step process.
    - Find all frequent itemsets.
    - Generate strong association rules from the frequent itemsets.

40. What are the advantages of Dimensional modeling?

- Ease of use.
- High performance
- Predictable, standard framework
- Understandable
- Extensible to accommodate unexpected new data elements and new design decisions

41. What is ment by dimensional modeling?

Dimensional modeling is a logical design technique that seeks to present the data in a Standard framework that intuitive and allows for high-performance access. It is inherently Dimensional and adheres to a discipline that uses the relational model with some important restrictions.

42. What comprises of a dimensional model?

Dimensional model is composed of one table with a multipart key called fact table and a set of smaller tables called dimension table. Each dimension table has a single part primary key that corresponds exactly to one of the components of multipart key in the fact table.

43. What is a data mart?

Data mart is a pragmatic collection of related facts, but does not have to be exhaustive or

Exclusive. A data mart is both a kind of subject area and an application. Data mart is a collection of numeric facts.

44. What are the advantages of a data-modeling tool?

- Integrates the data warehouse model with other corporate data models.
- Helps assure consistency in naming.
- Creates good documentation in a variety of useful formats.
- Provides a reasonably intuitive user interface for entering comments about objects.

45. What is data warehouse performance issue?

The performance of a data warehouse is largely a function of the quantity and type of data stored within a database and the query/data loading workload placed upon the system.

46. What are the types of performance issue?

- Capacity planning for the data warehouse
- data placement techniques within a data warehouse
- Application Performance Techniques.
- Monitoring the Data Warehouse.

47. Why do you need data warehouse life cycle process?

Data warehouse life cycle approach is essential because it ensures that the project pieces are brought together in the right order and at the right time.

48. What are the steps in the life cycle approach?
    - Project Planning
    - Business Requirements definition
    - Data track: Dimensional modeling, Physical Design, Data Staging Design & Development
    - Technology track: Technical Architecture design, Product Selection & Installation
    - Application track: End user Application Specification, End user Application Development
    - Deployment
    - Maintenance & Growth
    -

49. List the merits of Data Warehouse.
    - Ability to make effective decisions from database
    - Better analysis of data and decision support
    - Discover trends and correlations that benefits business
    - Handle huge amount of data.

50. What are the characteristics of data warehouse?
    - Separate
    - Available
    - Integrated
    - Subject Oriented
    - Not Dynamic
    - Consistency
    - Iterative Development
    - Aggregation Performance

51. List some of the Data Warehouse tools?
    - OLAP (Online Analytic Processing)
    - ROLAP (Relational OLAP)
    - End User Data Access tool
    - Ad Hoc Query tool
    - Data Transformation services
    - Replication

52. What is OLAP?

    The general activity of querying and presenting text and number data from Data Warehouses, as well as a specifically dimensional style of querying and presenting that is exemplified by a number of “OLAP Vendors” .The OLAP vendors technology is no relational and is almost always biased on an explicit multidimensional cube of data. LAP databases are also known as multidimensional cube of databases.

53. What is ROLAP?

    ROLAP is a set of user interfaces and applications that give a relational database a dimensional flavour. ROLAP stands for Relational Online Analytic Processing.

54. What is the need for End User Data Access tool?

End User Data Access tool is a client of the data warehouse. In a relational data warehouse, such a client maintains a session with the presentation server, sending a stream of separate SQL requests to the server. Evevtually the end user data access tool is done with the SQL session and turns around to present a screen of data or a report, a graph, or some other higher form of analysis to the user. An end user data access tool can be as simple as an Ad Hoc query tool or can be complex as a sophisticated data mining or modeling application.

55. What is meant by Ad Hoc query tool?

A specific kind of end user data access tool that invites the user to form their own queries by directly manipulating relational tables and their joins. Ad Hoc query tools, as powerful as they are, can only be effectively used and understood by about 10% of all the potential end users of a data warehouse.

56. Name some of the data mining applications?

- Data mining for Biomedical and DNA data analysis
- Data mining for Financial data analysis
- Data mining for the Retail industry
- Data mining for the Telecommunication industry

57. Name some of the data mining applications?

- Data mining for Biomedical and DNA data analysis
- Data mining for Financial data analysis
- Data mining for the Retail industry
- Data mining for the Telecommunication industry

58. Differentiate “supervised” from “unsupervised”.

In data mining during classification the class label of each training sample is provided, this type of training is called supervised learning (i.e.) the learning of the model is supervised in that it is told to which class each training sample belongs. Eg. Classification In unsupervised learning the class label of each training sample is not known and the member or set of classes to be learned may not be known in advance. Eg.Clustering.

59. Why is data quality so important in a data warehouse environment?

Data quality is important in a data warehouse environment to facilitate decision-making. In order to support decision-making, the stored data should provide information from a historical perspective and in a summarized manner.

60. How can data visualization help in decision-making?

Data visualization helps the analyst gain intuition about the data being observed. Visualization applications frequently assist the analyst in selecting display formats, viewer perspective and data representation schemas that faster deep intuitive understanding thus facilitating decision-making.

61. What do you mean by high performance data mining?

Data mining refers to extracting or mining knowledge. It involves an integration of techniques from multiple disciplines like database technology, statistics, machine learning, neural networks, etc. When it involves techniques from high performance computing it is referred as high performance data mining.

62. What are the various data mining issues?

- Knowledge Mining
- User interaction
- Performance
- Diversity in data types

63. What are the various data mining functionalities?

The data mining functionalities are:

- Concept class description
- Association analysis
- Classification and prediction
- Cluster Analysis
- Outlier Analysis

64. Explain the different types of data repositories on which mining can be performed?

The different types of data repositories on which mining can be performed are:

- Relational Databases
- Data Warehouses
- Transactional Databases
- Advanced Databases
- Flat files
- World Wide Web

PART – B (16 Marks)

1. Explain the architecture of data warehouse.
2. What is Data Mining? Explain the steps in Knowledge Discovery?
3. Explain the data pre-processing techniques in detail? Explain the smoothing Techniques?
4. Explain Data transformation in detail?
5. Explain Normalization in detail?
6. Explain data reduction?
7. Explain Data Discrimination and Concept Hierarchy Generation?
8. Explain Statistical measures in databases?
9. Explain multilevel association rule?
10. Explain Multidimensional Database briefly?
11. Explain star, snowflake, fact constellation schema and Diagrams.
12. Explain Indexing with suitable examples?
13. Explain the Back Propagation technique?
14. Explain Partition Methods?
15. Explain Hierarchical method of classifications?
16. Explain classification by Decision tree induction?
17. Explain the types of data in cluster analysis.
18. Explain Outlier analysis?
19. Explain Mining complex types of data?
20. Briefly explain about Data Mining Application?
21. Explain social impacts of data mining?
22. Explain Additional themes in data mining?

UNIT – III : EXTRACTION AND MINING COMMUNITIES IN WEB SOCIAL NETWORKS

PART – A : (2 Marks)

1. What is a Web Community?

A web community is a web site (or group of web sites) where specific content or links are only available to its members. A web community may take the form of a social network service, an Internet forum, a group of blogs, or another kind of social software web application.

2. How a Web Community does differs from a community of people?

An online community is a virtual community whose members interact with each other primarily via the Internet. For many, online communities may feel like home, consisting of a "family of invisible friends. An online community can act as an information system where members can post, comment on discussions, give advice or collaborate. Commonly, people communicate through social networking sites, chat rooms, forums, e-mail lists and discussion boards. People may also join online communities through video games, blogs and virtual worlds.

3. How is Web community extracted?

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

4. What is meant by virtual community?

A virtual community is a social network of individuals who interact through specific social media, potentially crossing geographical and political boundaries in order to pursue mutual interests or goals. Some of the most pervasive virtual communities are online communities operating under social networking services.

5. What is the purpose of evolution metrics?

A system of measurement is a collection of units of measurement and rules relating them to each other. Systems of measurement have historically been important, regulated and defined for the purposes of science and commerce. Systems of measurement in modern use include the metric system, the imperial system, and United States customary units.

6. What attributes are used to represent how many URLs the focused community obtains or loses?

HTML5 defines a <nav> menu, which is to be used to contain the primary navigation of a web site, be it a list of links or a form element such as a search box. This is a good idea, as previous to this we would contain the navigation block inside something like .

7. Justify the statement "The Web is extremely dynamic".

To facilitate this task we would appreciate that the largest amount of meta-data would be supplied along with the contents, specially.

- the web site address(es). If there are several web sites, please group the contents belonging to each one of them on a separate directory;
- the content addresses (URL). If you are providing a local copy of a site please maintain the original file names. If you are supplying contents that you gathered from the web please provide their original URLs;
- the content dates. Supply the date when each content was published or saved. If you do not know the exact dates, please supply approximate dates;
- the content media type (MIME). Please maintain the original file name extensions of the contents (e.g. .gif, .html, .jpg). If possible, provide the full HTTP header for each content. It is particularly important to provide the media type for contents dynamically generated that do not contain file name extensions.

8. Write notes on Web Community Charts.

Gantt chart is a type of bar chart, devised by Henry Gantt in the 1910s, that illustrates a project schedule. Gantt charts illustrate the start and finish dates of the terminal elements and summary elements of a project. Terminal elements and summary elements comprise the work breakdown structure of the project.

9. What is the size distribution of communities?

Rank-size distribution is the distribution of size by rank, in decreasing order of size. For example, if a data set consists of items of sizes 5, 100, 5, and 8, the rank-size distribution is 100, 8, 5, 5 (ranks 1 through 4). This is also known as the rank-frequency distribution

10. What is meant by community structure?

Complex networks, a network is said to have community structure if the nodes of the network can be easily grouped into (potentially overlapping) sets of nodes such that each set of nodes is densely connected internally. In the particular case of *non-overlapping* community finding, this implies that the network divides naturally into groups of nodes with dense connections internally and sparser connections between groups.

11. Give the significance of communitydiscovery in social network analysis.

The community detection in complex networks has attracted a growing interest and is the subject of several researches that have been proposed to understand the network structure and analyze the network properties.

12. What are the uses of community discovery?

Discovering communities in a social network environment is graph partitioning problem, which subdivides the entire graph into smaller partitions. Graph partitioning is believed as NP – hard problem, due to its complexity to split the number of vertices. We introduced the method of mutual accessibility to find communities in social networking environments. Existing work presents community discovery from blog posts. In this research, we discovered community structures from blogs which are posted by mobile devices such as mobile phones, specialized devices like personal digital assistants (PDA).

13. Mention the advantages of hierarchical algorithms.

- No apriori information about the number of clusters required.
- Easy to implement and gives best result in some cases.

14. Write notes on spectral methods.

Spectral methods are a class of techniques used in applied mathematics and scientific computing to numerically solve certain differential equations, often involving the use of the Fast Fourier Transform. The idea is to write the solution of the differential equation as a sum of certain "basis functions" (for example, as a Fourier series which is a sum of sinusoids) and then to choose the coefficients in the sum in order to satisfy the differential equation as well as possible.

15. What is Markov Clustering?

The MCL algorithm is short for the Markov Cluster Algorithm, a fast and scalable unsupervised cluster algorithm for graphs (also known as networks) based on simulation of (stochastic) flow in graphs.

16. What is the objective of *Kernighan-Lin (KL)* algorithm*?*

Kernighan–Lin algorithm. This article is about the heuristic algorithm for the graph partitioning problem. For a heuristic for the traveling salesperson problem, see Lin–Kernighan heuristic. The Kernighan–Lin algorithm is a heuristic algorithm for finding partitions of graphs.

17. What is meant by modularity?

Modular programming is the process of subdividing a computer program into separate sub-programs. A module is a separate software component. It can often be used in a variety of applications and functions with other components of the system.

18. Differentiate between agglomerative and divisive clustering?

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC . Top-down clustering requires a method for splitting a cluster.

19. What is a Dendrogram?

A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.Dendrograms are often used in computational biology to illustrate the clustering of genes or samples, sometimes on top of heatmaps.

20. What is Girvan and Newman's divisive algorithm.

The Girvan–Newman algorithm detects communities by progressively removing edges from the original network. The connected components of the remaining network are the communities. Instead of trying to construct a measure that tells us which edges are the most central to communities, the Girvan–Newman algorithm focuses on edges that are most likely "between" communities.

21. Write short notes on multi-level graph partitioning.

The graph partition problem is defined on data represented in the form of a graph $G = (V,E)$, with V vertices and E edges, such that it is possible to partition G into smaller components with specific properties. For instance, a k-way partition divides the vertex set into k smaller components. A good partition is defined as one in which the number of edges running between separated components is small.

22. What is stochastic flow?

It is a known fact that solutions to a certain second order parabolic par- tial differential equation are represented by means of a diffusion process or a stochastic flow.

23. Mention the limitations of Markov Clustering.

The Markov Cluster Algorithm (MCL) (Van Don-gen, 2000) is well-recognized as an effective method of graph clustering. It involves changing the values of a transition matrix toward either 0 or 1 at each step in a random walk until the stochastic condition is satisfied. When the hadamard power for each transition probability value is divided by the sum of each column, the rescaling process yields a transition matrix for the next stage.

24. What is the purpose of Regularized MCL?

Markov clustering (MCL) has emerged as an effective algorithm for clustering biological networks-for instance clustering protein-protein interaction (PPI) networks to identify functional modules. However, a limitation of MCL and its variants (e.g. regularized MCL) is that it only supports hard clustering often leading to an impedance mismatch given that there is often a significant overlap of proteins across functional modules.

25. What are heterogeneous social networks?

Community mining is one of the major directions in social network analysis. ... However, in reality, there exist multiple, heterogeneous social networks, each representing a particular kind of relationship, and each kind of relationship may play a distinct role in a particular task.

26. What is ensemble clustering?

The cluster ensemble problem is formulated as partitioning the hypergraph by cutting a minimal number of hyperedges. They make use of hMETIS which is a hypergraph partitioning package system.

27. What is co-citation regularity?

Co-citation, like Bibliographic Coupling, is a semantic similarity measure for documents that makes use of citation relationships. Co-citation is defined as the frequency with which two documents *are cited* together by other documents

28. What are the methods of inducing the graph?

In graph theory, an induced subgraph of a graph is another graph, formed from a subset of the vertices of the graph and all of the edges connecting pairs of vertices in that subset.

The induced subgraph isomorphism problem is a form of the subgraph isomorphism problem in which the goal is to test whether one graph can be found as an induced subgraph of another. Because it includes the clique problem as a special case, it is NP-complete.

PART – B (16 Marks)

1. What is a Web Community? How will you extract the evolution of Web Community from a series of Web Archives?
2. a. Discuss the various evolution metrics.
   b. Describe the various definitions of community.
3. Describe the core methods of community discovery in social networks.
4. Write notes on :
   a. Local graph clustering
   b. Flow-Based Post-Processing for Improving Community Detection
   c. Community Discovery via Shingling
   d. Explain the quality function to evaluate the community structure.
5. Explain the Node Classification problem.
6. Discuss the various local classifiers to solve node classification problem.
7. Describe the random walk-based methods of node classification.
8. Explain adsorption method of node classification.
9. Explain how to apply node classification to large social networks.
10. Discuss the applications of community mining algorithms.

UNIT – IV : PREDICTING HUMAN BEHAVIOUR AND PRIVACY ISSUES

PART – A : (2 Marks)

1. What is meant by evolution in Social Networks?

Visual representation of social networks is important to understand the network data and convey the result of the analysis.Signed graphs can be used to illustrate good and bad relationships between humans location-based interaction analysis, social sharing and filtering, recommender systems development, and link prediction and entity resolution.

2. What is stream paradigm of computation?

Stream processing is a computer programming paradigm, equivalent to dataflow programming, event stream processing, and reactive programming, that allows some applications to more easily exploit a limited form of parallel processing. Such applications can use multiple computational units, such as the FPUs on a GPU or field programmable gate arrays (FPGAs), without explicitly managing allocation, synchronization, or communication among those units.

3. Give the purpose of stream mining algorithm.

A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities. Examples of data streams include computer network traffic, phone conversations, ATM transactions, web searches, and sensor data. Data stream mining can be considered a subfield of data mining, machine learning, and knowledge discovery.

4. What is the use of sliding window in stream mining?

Finding frequent patterns in a continuous stream of transactions is critical for many applications such as retail market data analysis, network monitoring, web usage mining, and stock market prediction.

5. What are the two different threads of research on the analysis of dynamic social networks?

Social and temporal analysis methods.

6. List the characteristics of *perennial objects?*

An object is made of tangible material (the pen is made of plastic, metal, ink).

An object holds together as a single whole (the whole pen, not a fog).

An object has properties (the color of the pen, where it is, how thick it writes...).

An object can do things and can have things done to it.

7. How will you compute the entity similarity matrix?

The term "cosine similarity" is sometimes used to refer to different definition of similarity provided below. However the most common use of "cosine similarity" is as defined above and the similarity and distance metrics defined below are referred to as "angular similarity" and "angular distance" respectively. The normalized angle between the vectors is a formal distance metric and can be calculated from the similarity score defined above. This angular distance metric can then be used to compute a similarity function bounded between 0 and 1, inclusive.

8. What is an Evolution Net?

A social network is a social structure made up of a set of social actors sets of dyadic ties. The study of these structures uses *social network analysis* to identify local and global .The Barabási model of network *evolution* shown above is an example of a scale-free network and criminology.

9. What are the challenging issues in (dynamic) probabilistic modeling?

Bayesian Networks assume a static model of the system which does not account for failure/repair dynamics (i.e., the system state is assumed to be static during diagnosis process). In highly dynamic systems, this is not the case. There is a need to expand a static Bayesian Network model into a dynamic Bayesian Network model, in order to model situations where the node states change over time.

10. What are the two risk functions of non-parametric method?

Modelling the risk function non-parametrically, estimating it, for example, by a smoothing (thin plate) spline is attractive as a more explorative approach. For prospective studies this amounts to smoothing within the framework and distributional assumptions of generalized regression models (for binary observations). Case-control studies as retrospective studies with exposure to risk factors being observed do not immediately fit into this setting.

11. What is meant by social influence?

Social influence occurs when one's emotions, opinions, or behaviors are affected by others. Social influence takes many forms and can be seen in conformity, socialization, peer pressure, obedience, leadership, persuasion, sales and marketing.

12. What is meant by social correlation?

The correlation is one of the most common and most useful statistics. A correlation is a single number that describes the degree of relationship between two variables. Let's work through an example to show you how this statistic is computed.

13. What is meant by triadic closure?

Triadic closure is the property among three nodes A, B, and C, such that if a strong tie exists between A-B and A-C, there is a weak or strong tie between B-C.

14. What is node-based centrality?

A star network with 5 nodes and 4 edges. ... Based on these three features, Freeman (1978) formalized three different measures of node centrality: degree, closeness, and betweenness. Degree is the number of nodes that a focal node is connected to, and measures the involvement of the node in the network.

15. What is meant by katz centrality?

Katz centrality of a node is a measure of centrality in a network. It was introduced by Leo Katz in 1953 and is used to measure the relative degree of influence of an actor (or node) within a social network. Unlike typical centrality measures which consider only the shortest path (the geodesic) between a pair of actors, Katz centrality measures influence by taking into account the total number of walks between a pair of actors.

16. What is social action tracking?

Event tracking measures general user-interactions very well, Social Analytics provides a consistent framework for recording social interactions. This in turn provides a consistent set of reports to compare social network interactions across multiple networks.

17. What is meant byLatent action state?

Latent functions are the unintended, unpredicted or unseen consequences that might arise as a result of certain manifest functions that have taken place.

18. What is meant bygrouping behavior?

A group can be defined as two or more interacting and interdependent individuals who come together to achieve particular objectives. A group behavior can be stated as a course of action a group takes as a family. For example: Strike.

19. What is meant by diffusion influence model?

A diffusion model attempts to replicate the temporal adoption of a new product as word of mouth travels through the target population and external communications attempt to influence demand. A sample diffusion model worksheet with a graph of projected adoption appears below. Click on sections of the image to link to explanations of its contents.

20. State Expert location problem.

Human expertise is more valuable than capital, means of production or intellectual property. Contrary to expertise, all other aspects of capitalism are now relatively generic: access to capital is global, as is access to means of production for many areas of manufacturing. Intellectual property can be similarly licensed. Furthermore, expertise finding is also a key aspect of institutional memory, as without its experts an institution is effectively decapitated. However, finding and "licensing" expertise, the key to the effective use of these resources, remain much harder, starting with the very first step: finding expertise that you can trust.

PART – B (16 Marks)

1. a. Discuss the four dimensions that are associated to knowledge discovery in social networks and elaborate on their interplay in the context of evolution.

    b. Discuss the challenges of social network streams.

2. Explain how communities evolve into the learning process as smoothly evolving constellations of interacting entities.
3. Discuss the various influence related statistics.
4. Explain briefly social similarity and influence.
5. Describe influence maximization in viral marketing.
6. Describe expert location without graph constraints.
7. Explain expert location with score propagation.
8. a. Describe in detail expert score propagation.

    b. Explain probabilistic relational models

9. Explain in detail Bayesian probabilistic models.
10. Describe feature based link prediction.

UNIT – V : VISUALIZATION AND APPLICATIONS OF SOCIAL NETWORKS

PART – A (2 Marks)

1. What is visualization of online social networks?

Visualization system for playful end-user exploration and navigation of large scale online social networks. Our design builds upon familiar node link network layouts to contribute customized techniques for exploring connectivity in large graph structures, supporting visual search and analysis, and automatically identifying and visualizing community structures.

2. What is meant by taxonomy of visualization?

A new, comprehensive taxonomy of visualization techniques, drawing from the theories of Edward Tufte and citing examples from academia, government, and the excellent NYT visualization team. This list contains 12 steps for turning data into a compelling visualization: Visualize, Filter, Sort, Derive, Select, Navigate, Coordinate, Organize, Record, Annotate, Share, & Guide. 'For developers, the taxonomy can function as a checklist of elements to consider when creating new analysis tools.' The citations alone make this an article worth bookmarking."

3. Mention the different types of visualization.

There are two basic types of visualization techniques:

- Internalizing - visualization pictures in our mind's eye.
- Externalizing - visualization pictures outside of us with our eye's open.

4. What are the two approaches to structural visualization?

A very common approach to structural visualization is to guide the visualization process by underlying programming styles or computational models.

5. State the purpose of visualization.

Based on (non-visual) data. A visualization's purpose is the communication of data. That means that the data must come from something that is abstract or at least not immediately visible (like the inside of the human body).

6. What is meant by proximity of nodes?

Detailed level (i.e., node level) of link analysis, we want to figure out the relationship between two nodes on the graph, such as proximity, association, correlation and causality. For proximity, the goal is to measure the closeness (a.k.a, relevance, or similarity) between two nodes.

7. What are the various layout algorithms?

- force-based layout systems
- Spectral layout
- Orthogonal layout

8. Give the significance of graph layout algorithm?

In mathematics graph theory is the study of graphs, which are mathematical structures used to model pairwise relations between objects. A graph in this context is made up of vertices, nodes, or points which are connected by edges, arcs, or lines.

9. Write short notes on node-edge diagrams.

Node-edge diagrams based techniques as well as space-filling approaches incorporate the focus and context concept.

10. Write notes on matrix-oriented techniques.

In geometry the orientation, angular position, or attitude of an object such as a line, plane or rigid body is part of the description of how it is placed in the space it is in.[1] Namely, it is the imaginary rotation that is needed to move the object from a reference placement to its current placement. A rotation may not be enough to reach the current placement. It may be necessary to add an imaginary translation, called the object's location (or position, or linear position). The location and orientation together fully describe how the object is placed in space.

11. Write short notes on Web Communities.

A web community is a web site (or group of web sites) where specific content or links are only available to its members. A web community may take the form of a social network service, an Internet forum, a group of blogs, or another kind of social software web application.

12. What are digital libraries?

A digital library is a special library with a focused collection of digital objects that can include text, visual material, audio material, video material, stored as electronic media formats (as opposed to print, microform, or other media), along with means for organizing, storing, and retrieving the files and media

13. What do you mean by Content-centric visualization?

In contrast to IP-based, host-oriented, Internet architecture, content centric networking (CCN) emphasizes content by making it directly addressable and routable. Endpoints communicate based on named data instead of IP addresses.

14. What is the purpose of User-centric visualization?

The need to understand and track files (and inherently, data) in cloud computing systems is in high demand. Over the past years, the use of logs and data representation using graphs have become the main method for tracking and relating information to the cloud users.

15. Define semantic visualization.

Visualization can support effective and efficient interaction with a range of information for a variety of tasks.

16. What is meant by ontology engineering?

Ontology engineering in computer science and information science is a field which studies the methods and methodologies for building ontologies: formal representations of a set of concepts within a domain and the relationships between those concepts. A large-scale representation of abstract concepts such as actions, time, physical objects and beliefs would be an example of ontological engineering.

17. What is a semantic substrate?

A semantic wiki is a wiki that has an underlying model of the knowledge described in its pages. Regular, or syntactic, wikis have structured text and untyped hyperlinks. Semantic wikis, on the other hand, provide the ability to capture or identify information about the data within pages, and the relationships between pages, in ways that can be queried or exported like a database through semantic queries.

18. What is meant by data visualization?

Data visualization or data visualisation is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information".

19. What is the purpose of ontology mapping?

Ontology mapping may refer to:

- Semantic integration
- Ontology alignment

20. What is meant by semantic integration?

Semantic integration is the process of interrelating information from diverse sources, for example calendars and to do lists, email archives, presence information (physical, psychological, and social), documents of all sorts, contacts (including social graphs), search results, and advertising and marketing relevance derived from them. In this regard, semantics focuses on the organization of and action upon information by acting as an intermediary between heterogeneous data sources, which may conflict not only by structure but also context or value.

21. What is meant by ontology alignment?

Ontology alignment, or ontology matching, is the process of determining correspondences between concepts. A set of correspondences is also called an alignment. The phrase takes on a slightly different meaning, in computer science, cognitive science or philosophy.

PART – B (16 Marks)

1. What is visualization? Explain Social Network visualization on the Web.
2. Discuss the taxonomy of visualizations of social networks.
3. Explain the following:
    a. Clustering
    b. Centrality
    c. Node-link diagrams
4. Explain the Node-edge diagrams to visualize social networks.
5. Explain how to visualize social networks with matrix-based representation. Also discuss the pros and cons of matrix-based representation.
6. Discuss the various approaches to scale node-link diagrams to large networks with several thousand or millions of nodes.
7. Briefly explain the hybrid representation of visualization.
8. Briefly explain the concept of modeling and aggregating social network data.
9. Explain how clustering is performed with random walk based measures. Also discuss the algorithms for computing proximity measures.
10. a) Discuss the applications of random walks approach.

b) Briefly explain the use of Hadoop and Map Reduce002E