Module 5

Syllabus

Text Mining, Naïve-Bayes Analysis, Support Vector Machines, Web Mining, Social Network Analysis

5.1 Text Mining

- Text mining is the art and science of discovering knowledge, insights and patterns from an organized collection of textual databases.
- Textual mining can help with frequency analysis of important terms, and their semantic.

1) Explain Text Mining Applications.

Text mining is a useful tool in the hands of chief knowledge officers to extract knowledge relevant to an organization. Text mining can be used across industry sectors and application areas, including decision support, sentiment analysis, fraud detection, survey analysis, and many more.

- **Marketing:** The voice of the customer can be captured in its native and raw format and then analyzed for customer preferences and complaints.
- **Business operations:** Many aspects of business functioning can be accurately gauged from analyzing text. Social network analysis and text mining can be applied to emails, blogs, social media and other data to measure the emotional states and the mood of employee populations. Sentiment analysis can reveal early signs of employee dissatisfaction which can then can be proactively managed relationships.
- **Legal:** In legal applications, lawyers and paralegals can more easily search case histories and laws for relevant documents in a particular case to improve their chances of winning.
- **Governance and Politics:** Social network analysis and text mining of large-scale social media data can be used for measuring the emotional states and the mood of constituent populations. Micro-targeting constituents with specific messages gleaned from social media analysis can be a more efficient use of resources when fighting democratic elections.

2) Discuss Text mining process.

Text Mining is a rapidly evolving area of research. As the amount of social media and other text data grows, there is need for efficient abstraction and categorization of meaningful information from the text.

The first level of analysis is identifying frequent words. This creates a bag of important words. The next level is at the level of identifying meaningful phrases from words.

Text mining is a semi-automated process. Text data needs to be gathered, structured, and then mined, in a 3-step process

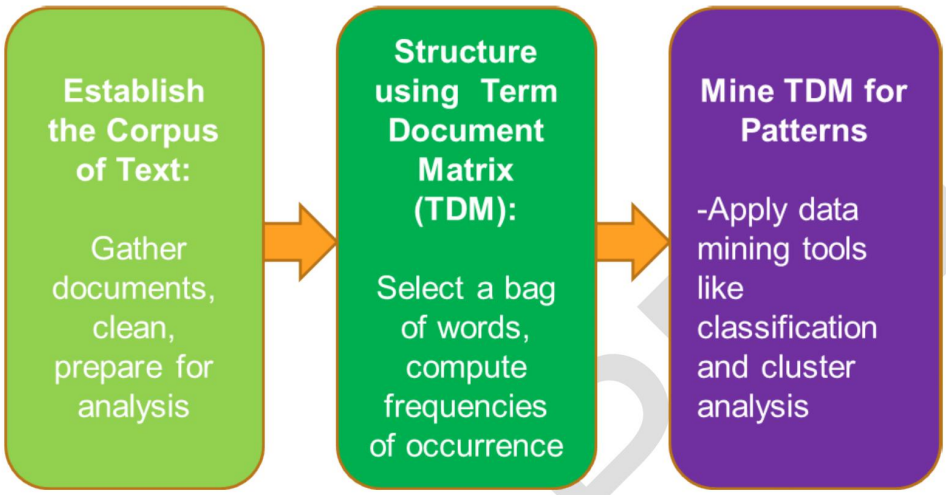

Figure 5.1 Text Mining Architecture

1. The text and documents are first gathered into a corpus, and organized.

2. The corpus is then analyzed for structure. The result is a matrix mapping important terms to source documents.

3. The structured data is then analyzed for word structures, sequences, and frequency.

3) Explain Text Document Practice (TDM).

This is the heart of the structuring process. Free flowing text can be transformed into numeric data in a TDM, which can then be mined using regular data mining techniques.

- Measures the frequencies of select important terms occurring in each document. This creates a t x d Term–by–Document Matrix (TDM) where t is the number of terms and d is the number of documents (Table 5.1)
- Creating a TDM requires making choices of which terms to include. Then terms chosen should reflect the stated purpose of the text mining exercise. The list of terms should be as extensive as needed, but should not include unnecessary stuff that will serve to confuse the analysis, or slow the computation.

Table 5.1: Term-Document Matrix

	Term Document Matrix				
Document / Terms	**investment**	**Profit**	**happy**	**Success**	**…**
Doc 1	10	4	3	4	
Doc 2	7	2	2		
Doc 3			2	6	
Doc 4	1	5	3		
Doc 5		6		2	
Doc 6	4		2		
…					

Here are some considerations in creating a TDM.

1. A large collection of documents mapped to a large bag of words will likely lead to a very sparse matrix if they have few common words. Reducing dimensionality of data will help improve the speed of analysis and meaningfulness of the results.
2. Data should be cleaned for spelling errors.
3. When many variants of the same term are used, just the stem of the word would be used to reduce the number of terms. For instance, terms like customer order, ordering, order data, should be combined into a single token word, called ‘Order’.
4. On the other side, homonyms (terms with the same spelling but different meanings) should be counted separately.
5. Terms with very few occurrences in very few documents should be eliminated from the matrix.

4) Compare Text Mining and Data Mining

Dimension	Text Mining	Data Mining
Nature of data	Unstructured data: Words, phrases, sentences	Numbers; alphabetical and logical values
Language used	Many languages and dialects used in the world; many languages are extinct, new documents are discovered	Similar numerical systems across the world
Clarity and precision	Sentences can be ambiguous; sentiment may contradict the words	Numbers are precise.
Consistency	Different parts of the text can contradict each other	Different parts of data can be inconsistent, thus, requiring statistical significance analysis
Sentiment	Text may present a clear and consistent or mixed sentiment, across a continuum. Spoken words adds further sentiment	Not applicable
Quality	Spelling errors. Differing values of proper nouns such as names. Varying quality of language translation	Issues with missing values, outliers, etc
Nature of	Keyword based search; coexistence of themes; Sentiment mining,	A full wide range of statistical and machine learning analysis for relationships and differences

5.2 Naive Bayes Analysis

- Naïve Bayes (NB) technique is the supervised learning technique that uses probability theory based analysis.
- Naïve Bayes is used often in classifying text documents into one of multiple predefined categories.
- Basically, it's "naive" because it makes assumptions that may or may not turn out to be correct.

1. Define probability

Probability is defined as chance of something happening. Its values may range from 0 to 1.

Using the event records, the probability of something happening in the future can be reliably assessed.

2. Explain Naive Bayes Model.

Naïve Bayes is a conditional probability model for classification purpose. A function f: X→Y is to find a way to predict the class variable (Y) using a vector variable (X). In probability terms, the goal is to find P(Y|X) i.e probability of Y belonging to certain class of X.

Given an instance to be classified, represented by a vector $x=(x_1,x_2,\ldots x_n)$ represents n features, the Naïve- Bayes model assigns to an instance, probabilities of belonging to any of the K classes.

The posterior probability can be calculated using the following equation

$$P(C_k|x)= \frac{P(Ck)P(x|Ck)}{P(x)}$$

$P(C_k|x)$ is the posterior probability of class k, given predictor x.

$P(C_k)$ is the prior probability of class k.

P(x) is the prior probability of predictor

$P(x|C_k)$ is the current likelihood of predictor given class.

Note: **Text Classification Example**

The probability of the document 'd' being in class 'c' is computed as,

$$P(c|d) \propto P(c) \prod_{1<=k<=n_d} P(t_k|c)$$

Where $P(t_k|c)$ is the conditional probability of term t_k occuring in a document of class c.

3. Classify the test data into right class as h or ~h using Naïve Bayes classification

Table 5.2.1 Dataset

Training set	Document ID	Keywords in the document	Class=h(healthy)
	1	Love Happy Joy Joy Love	Yes
	2	Happy Love Kick Joy Happy	Yes
	3	Love Move Joy Good	Yes
	4	Love Happy Joy Pain Love	Yes
	5	Joy Love Pain Kick Pain	No
	6	Pain Pain Love Kick	No
Test Data	7	**Love Pain Joy Love Kick**	**?**

Step 1: The prior probabilities or probability of each class.

There are two class Yes and No: Yes=h No=~h

Number of documents= 6

Number of documents belongs to h class= 4

Number of documents belongs to ~h class= 2

P(h)=4/6=2/3=0.66

P(~h)=2/6=0.33

Step 2: calculate the conditional probability for each term in the document for h class and ~h class. Number of words in h class =19, Number of words in ~h class =9.

Table 5.2.2. Conditional Probability table

Class h	Class ~h
P(Love|h)=6/19	P(Love|~h)=2/9
P(Pain|h)=1/19	P(Pain|~h)=4/9
P(Joy|h)=5/19	P(Joy|~h)=1/9
P(Kick|h)=1/19	P(Kick|~h)=2/9

Step 3: Compute the test instance belonging to class h and ~h

For class h:

P(h|d7)= P(h) * P(Love|h) * P(Pain|h) * P(Joy|h) * P(Love|h) * P(Kick|h)

P(h|d7)=2/3 * 6/19 * 1/19 * 5/19 * 6/19 * 1/19

P(h|d7)=0.0000483

For class ~h:

P(~h|d7)= P(~h) * P(Love|~h) * P(Pain|~h) * P(Joy|~h) * P(Love|~h) * P(Kick|~h)

P(~h|d7)=1/3 * 2/9 * 4/9 * 1/9 * 2/9 * 2/9

P(~h|d7)=0.00018

The Naïve Bayes probabilty of test instance being ~h is higher than h. Therefore test document will be classified as **~h (Not Healthy).**

4. Classify the test data into right class as Good or Bad using Naïve Bayes classification

Table 5.2.3 Movie Data Set

ID	Keywords	Review class
1	I loved the movie	Good
2	I hated the movie	Bad
3	A great movie good movie	Good
4	Poor acting	Bad
5	Great acting. A good movie	Good
6	I hated the poor acting	?

Answer:

We need to find the probability of text document belonging to both class.

There are two classes good and bad.

Number of documents= 5

Number of documents belongs to good class= 3

Number of documents belongs to bad class= 2

P(good)=3/5

P(bad)=2/5

Total number of words in good=14

Total number of words in bad=6

Totall number of unique words=10

Table 5.2.4 Conditional Probability table

Good	Bad
P(I|good)	P(I|bad)
P(hated|good)	P(hated|bad)
P(the|good)	P(the|bad)
P(poor|good)	P(poor|bad)
P(acting|good)	P(acting|bad)

But for this table we can obtain a value P(hated|good)=0/14=0, which will nullify the entire probability. Hence we cannot accurately classify to any respective classes.

This problem can be solved using Laplace smoothing.

$$\Theta_i = \frac{x_i + \alpha}{n + \alpha d}$$

In simple term

$$\text{P(word)} = \frac{word\ count + 1}{Total\ words + number\ of\ unique\ words}$$

We are adding the constant value α as 1, but the probability value should be always less than 1. Hence we are adding αd in the denominator to make the probability less than 1.

By calculating the probability and substituting its value, we get

P(I hated the poor acting|good)=

P(good)*P(I|good)*P(hated|good)*P(the|good)*P(poor|good)*P(acting|good)

=3/5*2/24*1/24*2/24*1/24*2/24

=0.0000006028

P(I hated the poor acting|bad)=

P(bad)*P(I|bad)*P(hated|bad)*P(the|bad)*P(poor|bad)*P(acting|bad)

=2/5*2/16*2/16*2/16*2/16*2/16

=0.00001220

P(I hated the poor acting|good)=0.0000006028

P(I hated the poor acting|bad)=0.00001220

The Naïve Bayes probabilty of test instance being bad is higher than good. Therefore test document will be classified as **bad.**

5. Mention the Advantages and Disadvantages of Naive Bayes Analysis.

Advantages:

- NB logic is simple
- Conditional probabilities can be computed for discrete data and for probabilistic distributions.

Disadvantages:

- If there are no joint occurrences at all of a class label with a certain attribute, then the frequency based conditional probability will be zero.
- Posterior probability computations are good only for comparison and classification of the instances

5.3 Support Vector Machine

SVM is a mathematically rigorous, machine learning technique to build a linear binary classifier. It creates a hyper plane in a high dimensional space that can accurately slice data set into two segments according to the desired objectives.

1. Explain in detail SVM model and its classifier

- SVM is a classifier function that defines a decision boundary between two classes.
- The support vectors are the data points that define the 'gutter' or the boundary condition on either side of the hyper plane for each of the two classes.
- There is a labelled set of points classified into two classes as shown in figure 1.1 and the goal is to find the best classifier between the points of the two types.

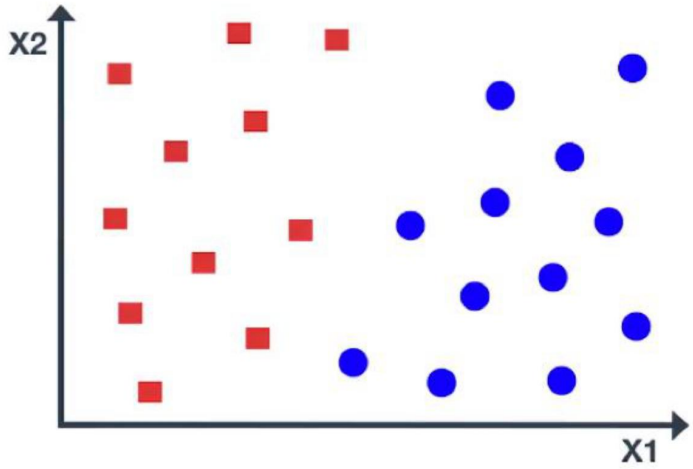

Figure 5.3.1

SVM takes the widest street approach to demarcate the two classes and thus finds the hyper plane that has the widest margin i.e. largest distance to the nearest training data points of either class.

- In figure 5.3.2 the hard line is the optimal hyper plane.
- The dotted lines are the gutters on the side of the two classes.
- The gap between the gutters is the maximum or widest margin.

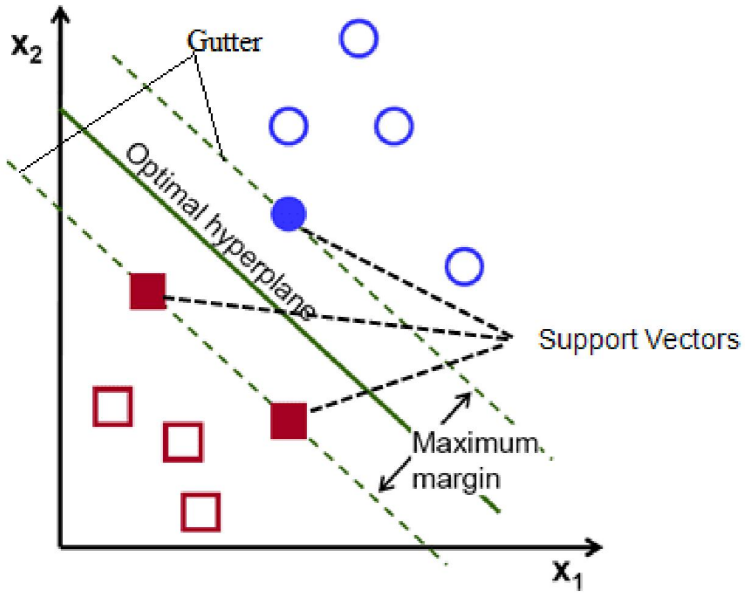

Figure 5.3.2

- The classifier is defined by only those points that fall on the gutters on both the sides. These points are called support vectors.
- The remaining data points are irrelevant for defining the classifier.

Suppose training data of n point is

$(X_1,y_1),(X_2,y_2)\ldots\ldots(X_i,y_i)$

Where X_i represents p-value vector for point I and y_i is its binary class value of 1 or -1. Thus there are two classes 1 and -1.

Assuming that the data is indeed linearly separable, the classifier hyper plane is defined as set of points that satisfy the equation

$$W \cdot X + b=0$$

Where W is the normal vector for the hyper plane.

The hard margins can be defined as

$$W \cdot X + b=1 \text{ and } W \cdot X + b= -1$$

The width of the hard margin is (2/|W|)

The y value will have to be either greater than 1 or less than -1

The SVM algorithm finds the weight vector (W) for the features such that there is a widest margin between the two categories.

2. Explain Kernel Method of SVM.

- The heart of an SVM algorithm is the Kernel Method. This method operate using 'kernel trick'.
- This trick involves computing and working with the inner products of only the relevant pairs of data in the feature space.
- The kernel tricks make the algorithm much less demanding in computational and memory resources.
- Kernel method achieves this by learning from instances called as instance-based learning.
- There are several types of support vector models including linear, polynomial and sigmoid.

3. Mention the advantages and disadvantages of SVM.

Advantages:

- They work well even when the number of features is much larger than number of instances.
- SVMs transform the variables to create to create new dimensions even when the optimal decision is non-linear curve.
- Easy to understand
- With only a subset of relevant data, they are computationally efficient.

Disadvantages:

- It works only well with the real numbers i.e all the data points in all the dimensions must be defined by numeric values.
- It works only with the binary classification problems.
- Training the SVMs is an inefficient and time consuming process when the data is large.
- It doesn't work well when there is a noise in the data.
- It will not provide confidence level for classifying an instance.

5.4 Web Mining

Web mining is the art and science of discovering patterns and insights from the World Wide Web so as to improve it. Web mining analyzes data from the Web and helps find insights that could optimize the web content and improve the user experience.

1. What are the characteristics of optimized websites?

1. ***Appearance:*** Aesthetic design; well-formatted content, easy to scan and navigate; and good color contrasts.

2. ***Content:*** Well-planned information architecture with useful content; fresh content; search-engine optimized; and links to other good sites.

3. ***Functionality:*** Accessible to all authorized users; fast loading times; usable forms; and mobile enabled.

2. Explain the three types of web mining. (Explain web mining structure)

Web mining can be divided into three different types:

a. web usage mining
b. web content mining
c. web structure mining

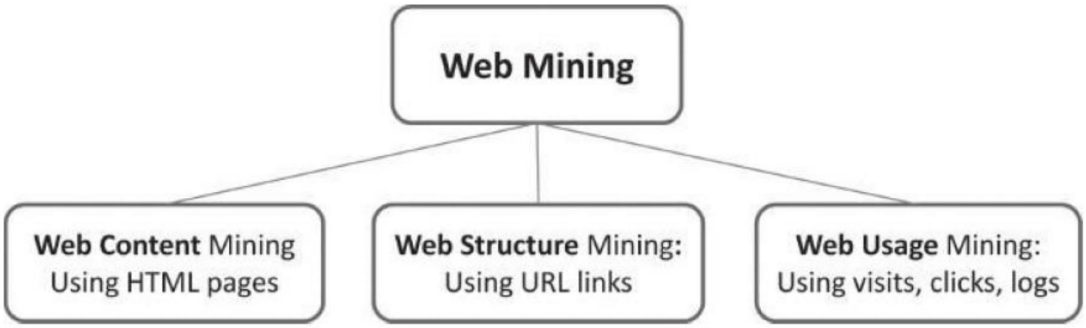

Figure 5.4.1

Web Content Mining

- Large website may contain thousands of pages. Those pages and their content are managed using content management systems.
- Every page can have text, graphics, audio, video, forms, applications, and more kinds of content, including user-generated content.
- The websites make a record of all requests received for its page/URLs.
- The log of these requests could be analyzed to gauge the popularity of those pages.
- The textual and application content could be analyzed for its usage by visits to the website.

- The pages on a website themselves could be analyzed for quality of content.
- The unwanted pages could be transformed with different content and style, or they may be deleted altogether.

Web Structure Mining

The structure of web pages could also be analyzed to examine the structure of hyperlinks among pages. There are two basic strategic models for successful websites: hubs and authorities.

1. *Hubs:* The pages with a large number of interesting links would serve as a hub, or a gathering point, where people access a variety of information. Media sites like Yahoo.com or government sites would serve that purpose.

2. *Authorities:* Ultimately, people would gravitate toward pages that provide the most complete and authoritative information on a particular subject, including user reviews. These websites would have the most number of inbound links.

Web Usage Mining

- As a user clicks anywhere on a web page or application, the action is recorded by many entities in many locations.
- The browser at the client machine will record the click, and the web server providing the content would also log onto the pages-served activity.
- The entities between the client and the server, such as the router, proxy server, or ad server, too, would record that click.
- The activity data comes from data stored in server access logs, referrer logs, agent logs, and client-side cookies.
- The user characteristics and usage profiles are also gathered directly, or indirectly, through syndicated data.
- Further, metadata, such as page attributes, content attributes, and usage data, are also gathered.

The web content could be analysed at multiple levels.

1. The server side analysis would show the relative popularity of the web pages accessed. Those websites could be hubs and authorities.

2. The client-side analysis could focus on the usage pattern or the actual content consumed and created by users.

a. Usage pattern could be analyzed using "clickstream" analysis, that is, analyzing web activity for patterns of sequence of clicks, and the location and duration of visits on websites.

Clickstream analysis is useful for web activity analysis, software testing, market research, and analyzing employee productivity.

b. Textual information accessed on the pages retrieved by users could be analyzed using text mining techniques. The text would be gathered and structured using the bag-of-words technique to build a term-document matrix. This matrix could then be mined using cluster analysis and association rules for patterns, such as popular topics, user segmentation, and sentiment analysis (Refer figure 5.4.2).

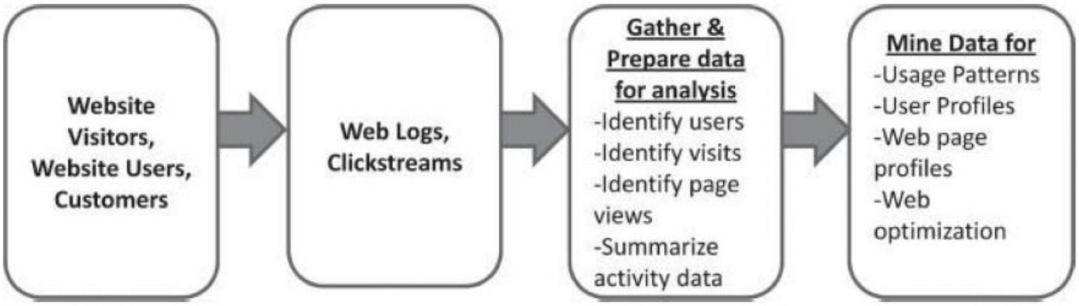

Figure 5.4.2

5.5 Social Network Analysis

- Social networks are graphical representation of relationship among people and entities.
- SNA is the art and science of discovering patterns of interaction and influence within the participants within the network.
- Participants could be people, organization, machines or any other kind of entities.

1. What are the Applications of SNA?

a. **Self-awareness:**

Visualizing social network can help a person organize their relationships and support network.

b. **Communities:**

This help in identification, construction and strengthening of networks within communities to build wellness and comfort.

c. **Marketing:**

There is a popular network insight that any two people are related to each other through at most seven degrees of links. It is also used for an organization to understand their customer need.

d. **Public health:**

Awareness of network can help identify the paths that certain diseases take to spread. Public health professionals can isolate and contain diseases before they expand to other networks.

2. What are the two major level of social network analysis? Explain.

The two levels of SNA are:

- Finding sub networks
- Computing Importance of nodes

Finding sub networks:

- Finding sub networks are like doing a cluster analysis of nodes.
- Nodes with strong ties between them would belong to the same sub network, while those with weak or no ties would belong to separate sub networks.

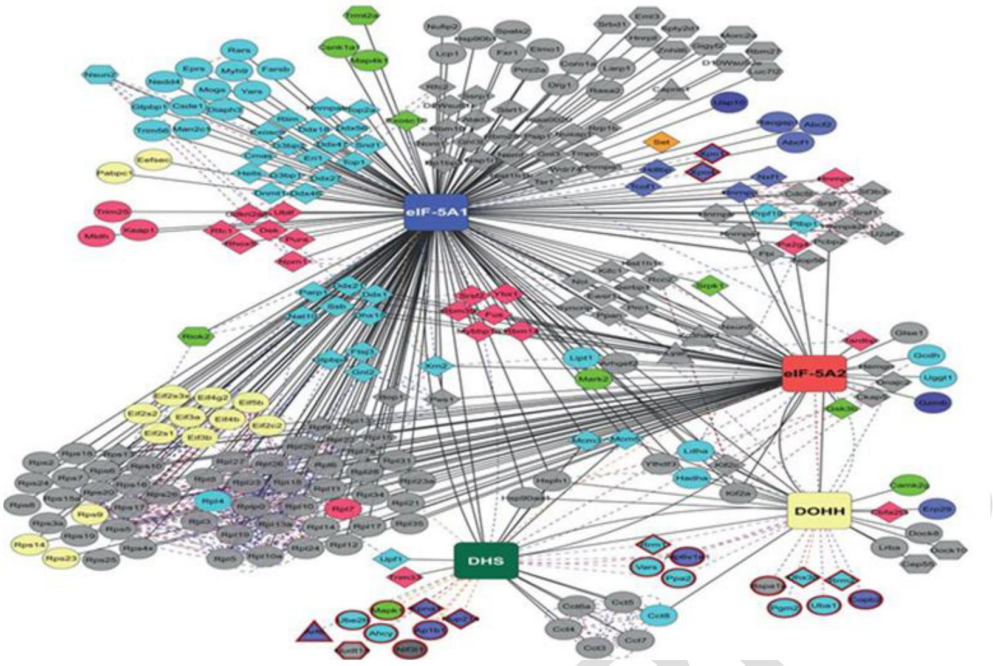

Figure 5.5.1

- Visual representation of networks can help identify networks. Use of colors can help differentiate the types of nodes.

Computing Importance of nodes

- When the connections between nodes in the network have a direction to them, then the nodes can be compared for their relative influence or rank.
- This is done using Influence Flow Model. Every outbound link from a node can be considered an outflow of an influence.
- Every incoming link is similarly an inflow of influence. More in links to a node means greater importance.
- Computing a relative influence of each node is done on the basis of an input output matrix of flows of influence among the nodes.
- It is an iterative task, begin with some initial value and continue until values get stabilize.

3. Compute the rank values for the nodes for the following network. Which is the highest ranked node?

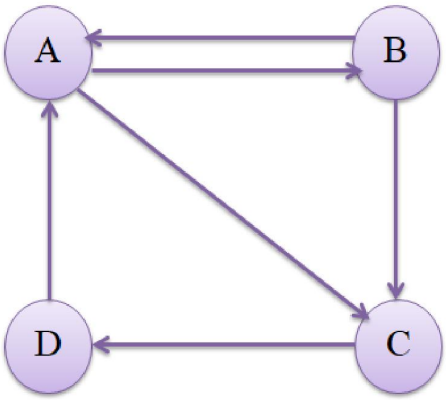

Figure 5.5.2

Step 1

The network with 4 nodes A,B,C,D and 6 directed link between them and also There is a bi directional link.

Node A links to B

Node B links to C

Node C links to D

Node D links to A

Node A links to C

Node B links to A

- Begin by assigning the variables for influence value for each node as Ra,Rb,Rc and Rd, the goal is to find the relative values of these variable.
- There are two outbound links from node A to node B and C. Thus both B and C receives half of node A's influence.
- B to C and A: C and A receives half influence from node B
- D to A: A receives full influence from node D
- C to D: D receives full influence from node C

Write the equation for all nodes:

Ra=0.5*Rb+Rd

Rb=0.5*Ra

Rc=0.5*Ra+0.5*Rb

Rd=Rc

Table 5.5.1 influence matrix

	Ra	Rb	Rc	Rd
Ra	**0**	**0.50**	**0**	**1**
Rb	**0.50**	**0**	**0**	**0**
Rc	**0.50**	**0.50**	**0**	**0**
Rd	**0**	**0**	**1.0**	**0**

The table 5.5.1 is called as influence matrix. The zero value represent that the term is not present in the equation.

Step 2:

Let us start with the initial set of rank values and then perform the iteration. Start with the initial rank value as 1/n = 1/4, and initial value of all 4 variables are assigned with 0.250 as shown in table 5.5.2.

Table 5.5.2 Initial values

Variable	Initial value
Ra	0.250
Rb	0.250
Rc	0.250
Rd	0.250

Step 3:

Based on the initial value and the equation, compute the revised set of values that gives iteration 1 as shown in table 5.5.3.

Table 5.5.3 Iteration 1

Variable	Initial value	Iteration 1
Ra	0.250	0.375
Rb	0.250	0.125
Rc	0.250	0.250
Rd	0.250	0.250

Step 4:

Based on the rank values from iteration 1, compute the new values that give iteration 2 as shown in table 3.4.

Table 5.5.4

Variable	Initial value	Iteration 1	Iteration 2
Ra	0.250	0.375	0.3125
Rb	0.250	0.125	0.1875
Rc	0.250	0.250	0.250
Rd	0.250	0.250	0.250

- We have to perform the iteration until all the values get stabilize.
- For this problem it takes eight numbers of iterations to stabilize the values. And all the values are shown in following table.

Table 5.5.5

Variable	Initial value	Iter 1	Iter 2		Iter 7	Iter 8
Ra	0.250	0.375	0.3125	……	0.3339	0.333
Rb	0.250	0.125	0.1875	……	0.166	0.1669
Rc	0.250	0.250	0.250	……	0.2499	0.2499
Rd	0.250	0.250	0.250	……	0.250	0.250

The final rank shows that rank of node A has highest value i.e **0.333**

Therefore most important node is: **node A**

The lowest rank is for node B I,e **0.166**

Hence least important node is: **node B**

Node C and D are the middle nodes

4. Explain the practical consideration of SNA.

1. Network size:

Collecting data about large networks can be very challenging.

2. Gathering data:

E-mails, chats can gather social network data more easily. Capturing, cleansing and organizing the data can take lot of time and effort.

3. Computation and visualization:

Modelling large networks can be computationally challenging and visualizing them also would require special skills.

4. Dynamic Networks:

The relationship between the nodes can change in strength and functional nature.

5. Differentiate Social Network Analysis and Traditional Data Mining.

Social Network Analysis	Traditional Data Mining
Unsupervised learning in nature	Supervised and Unsupervised learning in nature
Goal is to identify hub nodes, important nodes and sub networks	Goal is to key decision rules and cluster centroids
It uses graph of nodes and links	It uses data of variables and instances
It uses visualization and statistics technique	It uses machine learning and statistics technique.
Usefulness is key criterion	Predictive accuracy for classification techniques

~~~~~~~~~~~~~~~~~~~~ *** ~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~