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

Text Mining

INTRODUCTION

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 relationships.
- Text mining can be applied to large scale social media data for gathering preferences, and measuring emotional sentiments.
- It can also be applied to societal, organizational and individual scales.

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.

1. **Marketing** The voice of the customer can be captured in its native and raw format and then analyzed for customer preferences and complaints.
   a. Social personas are a clustering technique to develop customer segments of interest. Consumer input from social media sources, such as reviews, blogs, and tweets, contain numerous leading indicators that can be used toward5 anticipating and predicting consumer behavior.
   b. A listening platform is a text mining application that in real time, gathers social media, blogs, and other textual feedback, and filters out the chatter to extract true consumer sentiments. The insights can lead to more effective product marketing and better customer service.
   c. The customer call center conversations and records can be analyzed for patterns of customer complaints. Decision trees can organize this data to create decision choices that could help with product management activities and to become proactive in avoiding those complaints.
2. **Business Operation** Many aspects of business functioning can be accurately gauged from analyzing text.

a. 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 be proactively managed.
b. Studying people as emotional investors and using text analysis of the social Internet to measure mass psychology can help in obtaining superior investment returns.

3. **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.
   a. Text mining is also embedded in e-discovery platforms that help in minimizing risk in the process of sharing legally mandated documents.
   b. Case histories, testimonies, and client meeting notes can reveal additional information, such as morbidities in healthcare situations that can help better predict high-cost injuries and prevent costs.
4. **Governance and Politics** Government can be overturned based on a tweet originating from a self-immolating fruit vendor in Tunisia.
   a. 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.
   b. In geopolitical security, internet chatter can be processed for real-time information and to connect the dots on any emerging threats.
   c. In academics, research streams could be meta-analyzed for underlying research trends.

TEXT MINING PROCESS

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

1. The first level of analysis is identifying frequent words. This creates a bag of important words. Texts documents or smaller messages can then be ranked on how they match to a particular bag-of-words. However, there are challenges with this approach. For example, the words may be spelled a little differently or there may be different words with similar meanings.
2. The next level is identifying meaningful phrases from words. Thus ‘ice’ and ‘cream’ will be two different key words that often come together. However, there is a more meaningful phrase by combining the two words into ‘ice cream’. There might be similarly meaningful phrases like ‘Apple Pie’.
3. The next higher level is that of Topics. Multiple phrases can be combined into Topic area. Thus the two phrases above can be put into a common basket, and this bucket is called ‘Desserts’.

Text mining is a semi-automated process. Text data needs to be gathered, structured, and then mined, in a 3 step process (Figure 11.1)

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.

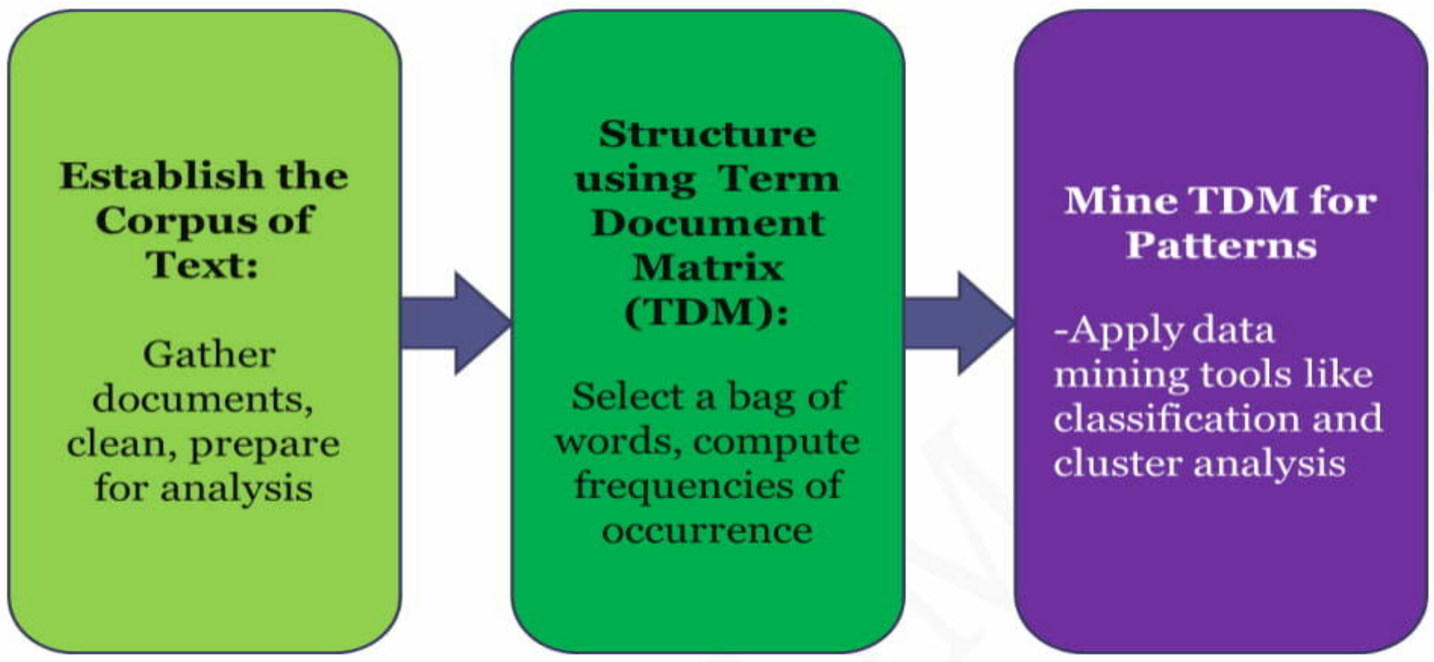


FIGURE 11.1 Text Mining Architecture

TERM DOCUMENT MATRIX

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.

	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		
…					

- There are several efficient techniques for identifying key terms from a text.
- There are less efficient techniques available for creating topics out of them.

- This approach 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 11.1).
- Creating a TDM requires making choices of which terms to include.
- The 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.

Here are some considerations in creating a TDM

- 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.
- Synonyms or terms with similar meanings should be combined and should be counted together as a common term. This would help reduce the number of distinct terms of words or 'tokens'.
- Data should be cleaned for spelling errors. Common spelling errors should be ignored and the terms should be combined. Uppercase-lowercase terms should also be combined.
- 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'.
- On the other side, homonyms (terms with the same spelling but different meanings) should be counted separately. This would enhance the quality of analysis.

For example, the term order can mean a customer order or the ranking of certain choices. These two should be treated separately "The boss ordered that the customer orders data analysis be presented in chronological order". This statement shows three different meanings for the word 'order thus, there will be a need for manual review of the TD matrix

- Terms with very few occurrences in the documents should be eliminated from the matrix. This would help increase the density of the matrix and the quality of analysis.
- The measures in each cell of the matrix could be one of the several possibilities. It could be a simple count of the number of occurrences of each term in a document. It could also be the log of that number. It could be the fraction number computed by dividing the frequency count by the total number of words in the document. Or there may be binary values in the matrix to represent whether a term is mentioned or not.
- The choice of value in the cells will depend upon the purpose of the text analysis.

- At the end of this analysis and cleansing, a well formed, densely populated, rectangular, TDM will be ready for analysis. The TDM can be mined using all the available data mining techniques.

MINING THE TDM

- The TDM can be mined to extract patterns/knowledge. A variety of techniques could be applied to the TDM to extract new knowledge.
- A simple application is to Visualize the highest frequency terms. This can be done very attractively and colorfully in the form of a 'word-cloud'.
- The word-cloud can be created after removing the common words like prepositions.
- It can be done for the top n words such as top 100 words, to focus on the key terms used in the document. The attached word-cloud represents the speech by US President Barack Obama on the topic of terrorism.

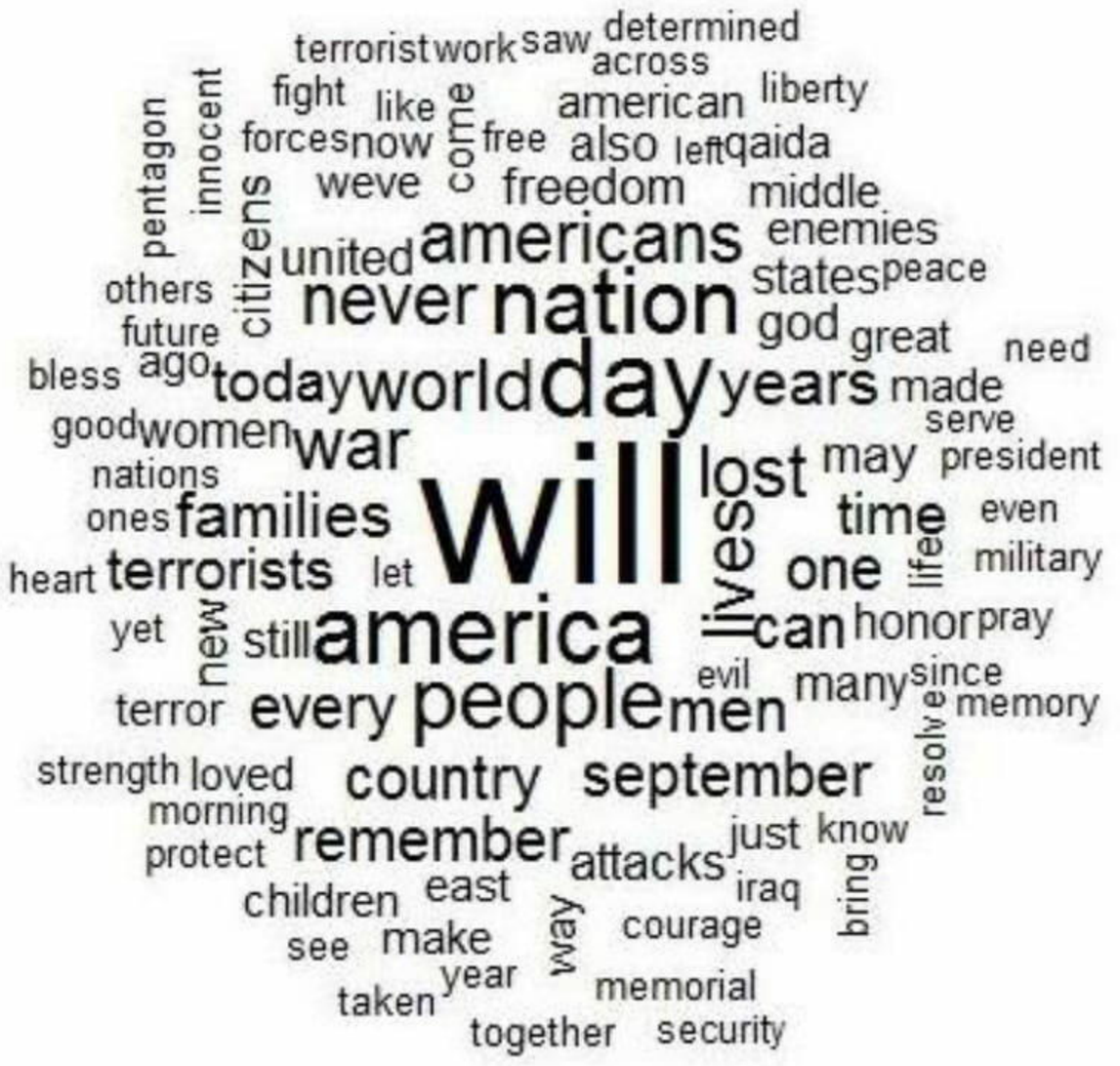


FIGURE 11.2 Word-cloud of Top 100 Words from US President's Speech

COMPARING TEXT MINING AND DATA MINING

- Text Mining is a form of data mining. There are many common elements between Text and Data Mining.
- However, there are some differences (Table 11.2). The key difference is that text mining requires conversion of text data into frequency data, before data mining techniques can be applied.

Table 11.2 Comparing 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 add 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 Analysis	Keyword based search; co-existence of themes; sentiment mining;	A full wide range of statistical and machine learning analysis for relationships and differences

TEXT MINING BEST PRACTICES

Many of the best practices that apply to the use of data mining techniques will also apply to text mining.

- The first and most important practice is to ask the right question. A good question is the one which gives an answer and would lead to large payoffs for the organization. The purpose and the key question will define how and at what levels of granularity the TDM would be made.
  - **For example**, TDM defined for simpler searches would be different from those used for complex semantic analysis or network analysis.
- A second important practice is to be creative and open in proposing imaginative hypotheses for the solution. Thinking outside the box is important, both in the quality of the proposed solution as well as in finding the high quality datasets required to test the hypothesized solution.
  - **For example**, a TDM of consumer sentiment data should be combined with customer order data in order to develop a comprehensive view of customer behavior. It's important to assemble a team that has a healthy mix of technical and business skills.

- Another important element is to pursue the problem iteratively. Too much data can overwhelm the infrastructure and also befuddle the mind. It is better to divide and conquer the problem with a simpler TDM, with fewer terms and fewer documents and data sources. Expand as needed, in an iterative sequence of steps. In the future, add new terms to help improve predictive accuracy.
- A variety of data mining tools should be used to test the relationships in the TDM. Different decision tree algorithms could be run alongside cluster analysis and other techniques. Triangulating the findings with multiple techniques, and many what-if scenarios helps build confidence in the solution. Test the solution in many ways before committing to deploy it.

NAIVE-BAYES ANALYSIS

INTRODUCTION

- Naive-Bayes (NB) technique is a supervised learning technique that uses probability-theory-based analysis.
- It is a machine-learning technique that computes the probabilities of an instance belonging to each one of many target classes, given the prior probabilities of classification using individual factors.
- Naive-Bayes technique is used often in classifying text documents into one of multiple predefined categories.

PROBABILITY

- Probability is defined as the chance of something happening.
- The probability Values thus range from zero to one; with a value of zero representing no chance, and to one representing total certainty.
- Using past event records, the probability of something happening in the future can be reliably assessed.

For example, one can assess the probability of dying from an airline accident, by dividing the total number of airline accident related deaths in a time period by the total number of People flying during that period. These probabilities can then be compared to come to the conclusions, such as the safety levels of various event types.

For example, past data may show that the probability of dying from airline accident is less than that of dying from being hit by lightning.

- The Naive-Bayes algorithm is special in that it takes into consideration the prior probability of an instance belonging to a class, in addition to the recent track record of the instance belonging to that class.

- The word Bayes refers to Bayesian analysis (based on the work of mathematician Thomas Bayes) which computes the probability of a new occurrence not only on the recent record, but also on the basis of prior experience.
- The word Naive represents the strong assumption that all parameters/features of an instance are independent variables with little or no correlation Thus if people are identified by their height, weight, age, and gender; all these variables are assumed to be independent of each other.
- NB algorithm is easy to understand and works fast. It also performs well in multiclass prediction, such as when the target class has multiple options beyond binary yes/no classification.
- NB can perform well even in case of categorical input variables compared to numerical variable(s).

NAIVE-BAYES MODEL

- In the abstract, Naive-Bayes is a conditional probability model for classification purposes.
- The goal is to find a way to predict the class variable (Y) using a vector of independent variables (X), i.e., finding the function

$$f: \mathbf{X} \rightarrow \mathbf{Y}$$

- In probability terms, the goal is to find $P(Y \mid X)$, i.e., the probability of Y belonging to a certain class X. Y is generally assumed to be a categorical variable with two or more discrete values.
- Given an instance to be classified, represented by a vector $x = (x_1, ..., x_n)$ representing 'n' features (independent variables), the Naive-Bayes model assigns, to an instance, probabilities of belonging to any of the K classes. The class K with the highest posterior probability is the label assigned to the instance.
- The posterior probability (of belonging to a Class K) is calculated as a function of prior probabilities and current likelihood value, as shown in the equation below

$$p(C_k \mid x) = \frac{p(C_k)p(x \mid C_k)}{p(x)}$$

- $P(C_k \mid x)$ 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.
- $P(x \mid C_k)$ is the current likelihood of predictor given class.

SIMPLE CLASSIFICATION EXAMPLE

Suppose a salon needs to predict the service required by the incoming customer If there are only two services offered -Hair cut (R) and Manicure-Pedicure (M) then the value to be predicted is whether the next customer will be for R or M. The number of classes (K) is 2.

- The first step is to compute prior probability. Suppose the data gathered for the last one year showed that during that period there were 2500 customers for R and 1500 customers for M.
- Thus, the default (or prior) probability for the next customer to be for R is 2500/4000 or 5/8.
- Similarly, the default probability for the next customer to be for M is 1500/4000 or 3/8.
- Based on this information alone, the next customer would likely be for R.

Another way to predict the service requirement by the next customer is to look at the most recent data. One can look at the last few (choose a number) customers, to predict the next customer. Suppose the last five customers were for the services R, M, R, M, M order.

- Thus, the data shows the recent probability of R is 2/5 and that of M is 3/5.
- Based on just this information, the next customer will likely to be for M.
- Thus in this example, the NB posterior probability P(R) is (5/8 x 2/5) = 10/40.
- Similarly, the NB probability P (M) is (3/8 x 3/5) = 9/40.
- Since P(R) is greater than P (M), it follows that there is a greater probability of the next customer to be for R. Thus the expected class label assigned to the next customer would be R.
- Suppose, however the next customer coming in was for M service. The last five customer sequence now becomes M, R, M, M, M.
- Thus, the recent data shows the probability for R to be 1/5 and that of M to be 4/5.
- Now the N B probability for R is (5/8 x 1/5) = 5/40.
- Similarly, the NB probability for M is (3/8 x 4/5) = 12/40.

Since P (M) is greater than P(R), it follows that there is a greater probability of the next customer to be for M. Thus the expected class label assigned to the next customer is M.

The NB predictor thus dynamically changes its prediction value based on the recent data.

TEXT CLASSIFICATION EXAMPLE

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

$$P(c \mid d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k \mid c)$$

Where, P (t_k | c) is the conditional probability of term tk occurring in a document of class c. Dataset 12.1 shows the text classification training and test data. The goal is to classify the test data into the right class as h or ~h (read as not h).

Dataset 12.1

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	?

- The prior probabilities of a document being classified using the six documents are

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

$$P(\sim h) = 2/6 = 1/3$$

- i.e., there is 2/3 prior probability that a document will be classified as h and 1/3 probability of not h.
- The conditional probability for each term is the relative frequency of the term occurring in each class of the documents 'h class' and 'not h class'.

Conditional Probabilities	
Class *h*	Class ~*h*
$P(\text{Love} \mid h) = 5/19$	$P(\text{Love} \mid \sim h) = 2/9$
$P(\text{Pain} \mid h) = 1/19$	$P(\text{Pain} \mid \sim h) = 4/9$
$P(\text{Joy} \mid h) = 5/19$	$P(\text{Joy} \mid \sim h) = 1/9$
$P(\text{Kick} \mid h) = 1/19$	$P(\text{Kick} \mid \sim h) = 2/9$

The probability of the test instance belonging to class h can be computed by multiplying the prior probability of the instance belonging to class h, with the conditional probabilities for each of the terms in the document for class h. Thus,

$$P(h \mid d7) = P(h) * (P(\text{Love} \mid h))\text{^}2 * P(\text{Pain} \mid h) * P(\text{Joy} \mid h) * P(\text{Kick} \mid h)$$
$$= (2/3) * (5/19) * (5/19) * (1/19) * (5/19) * (1/19) = \sim 0.0000067$$

The NB probability of the test instance being 'not h' is much higher than its being h. Thus the test document will be classified as 'not h'.

$$P(\sim h \mid d7) = P(\sim h) * P(\text{Love} \mid \sim h) * P(\text{Love} \mid \sim h) * P(\text{Pain} \mid \sim h) * P(\text{Joy} \mid \sim h) * P(\text{Kick} \mid \sim h)$$
$$= (1/3) * (2/9) * (2/9) * (4/9) * (1/9) * (2/9) = 0.00018$$

ADVANTAGES AND DISADVANTAGES OF NAIVE-BAYES

1. The NB logic is simple and so is the NB posterior probability computation.
2. Conditional probabilities can be computed for discrete data and for probabilistic distributions. When there are number of variables in the vector X. then the problem can be modeled using probability functions to simulate the incoming values. A variety of methods exist for modeling the conditional distributions of the X variables, including normal, lognormal, gamma, and Poisson.
3. Naive-Bayes assumes that all the features are independent for most instances that work fine. However, it can be a limitation. 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. When all the probabilities are multiplied, it will make the entire posterior probability estimate to be zero. This can be rectified by adding 1 to all the numerators and adding n, the number of variables in X, to all the denominators. This will make those probabilities very small but not zero.
4. A limitation of NB is that the posterior probability computations are good for comparison and classification of the instances. However, the probability values themselves are not good estimates of the event happening.

Support Vector Machines

INTRODUCTION

- Support Vector Machine (SVM) is a mathematically rigorous, machine learning technique to build a linear binary classifier.
- It creates a hyperplane in a high-dimensional space that can accurately slice a dataset into two segments according to the desired objective.
- The algorithms for developing the classifier can be mathematically challenging though.
- SVMs are popular since they are state-of-the-art for many practical problems, such as identifying spam emails and other text mining applications.

SVM MODEL

- An SVM is a classifier function in a high dimensional space that defines the decision boundary between two classes.
- The support vectors are the data points that define the 'gutters' or the boundary condition on either side of the hyperplane, for each of the two classes.

- The SVM model is thus conceptually easy to understand.
- Suppose there is a labeled set of points classified into two classes. The goal is to find the best classifier between the points of the two types.

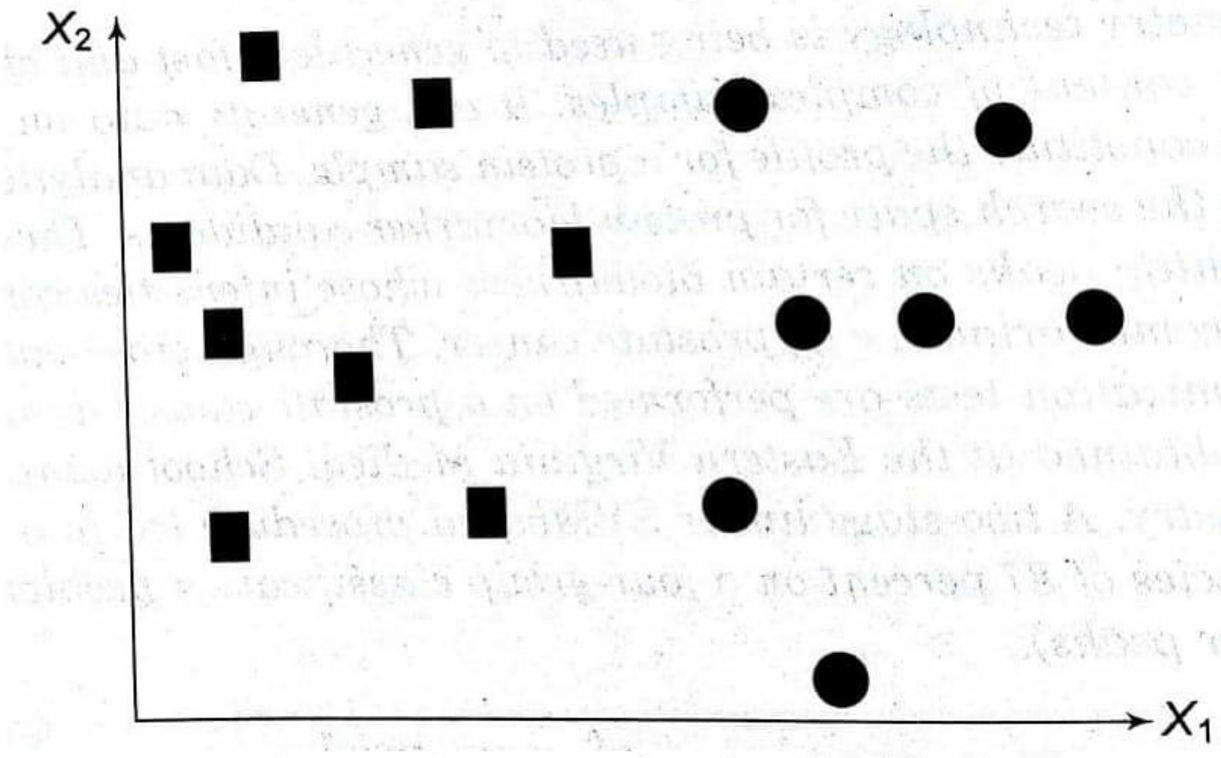


FIGURE 13.1 Data Points for Classification

SVM takes the widest street (a vector) approach to demarcate the two classes and thus finds the hyperplane that has the widest margin, i.e., largest distance to the nearest training data points of either class (Figure 13.2).

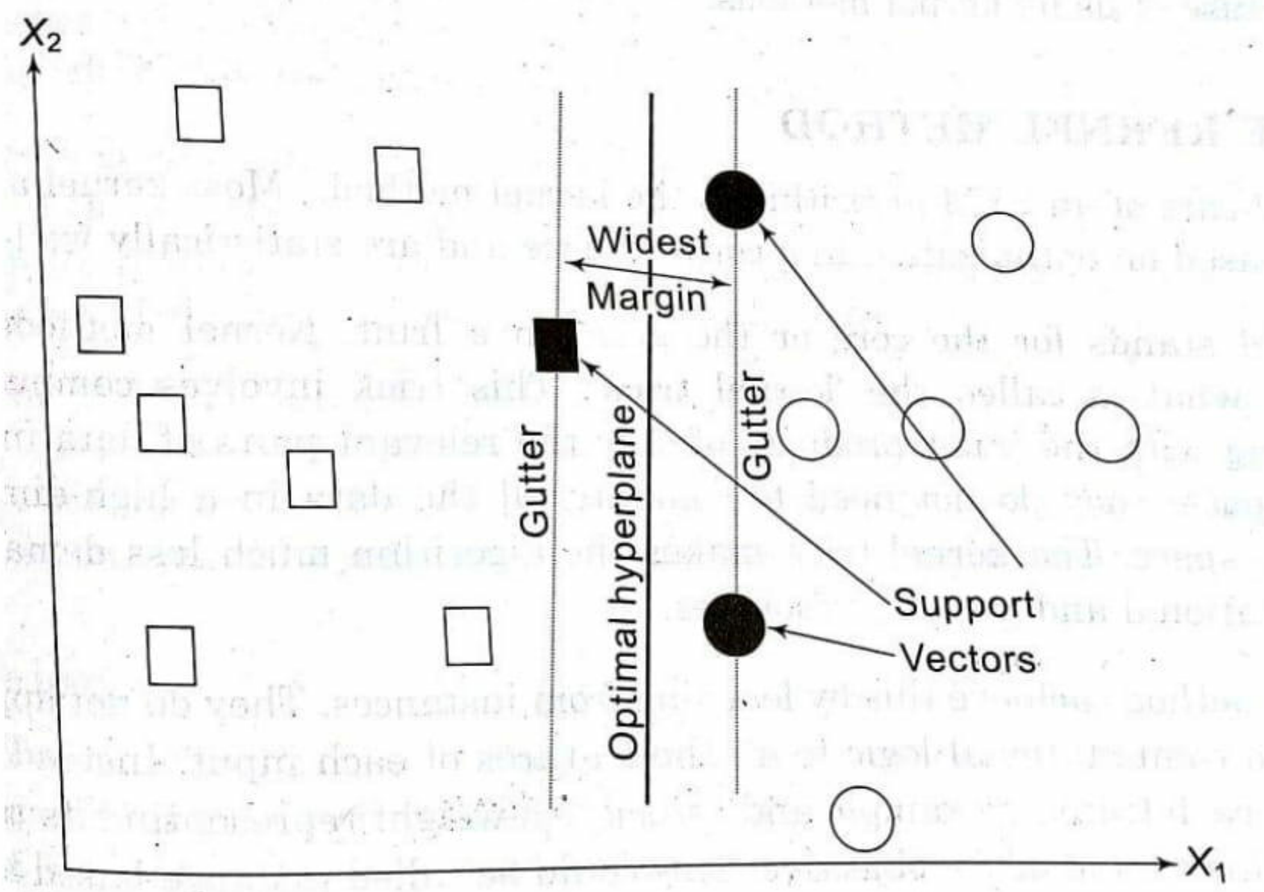


FIGURE 13.2 Support Vector Machine. Classifier

- In Figure 13.2, the hard line is the optimal hyperplane. The dotted lines are the gutters on the sides of the two classes.
- The gap between the gutters is the maximum or widest margin. The classifier (hyperplane) is defined by only those points that fall on the gutters on both sides.

- These points are called the support vectors (shown in their bold shape). The rest of the data points in their class are irrelevant for defining the classifier (shown unfilled).
- Abstractly, suppose that the training data of n points is

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

Where $\mathbf{X_i}$ represents the p-value vector for point i and $\mathbf{y_i}$ is its binary class value of 1 or -1. Thus there are two classes represented as 1 and -1.

Assuming that the data is indeed linearly separable, the classifier hyperplane is defined as a set of points (which is a subset of the training data) that satisfy the equation

$$\mathbf{W \cdot X + b = 0}$$

Where W is the normal vector to the hyperplane.

The hard margins can be defined by the following hyperplanes

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

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

- For all points not on the hyperplane, they will be safely in their own class.
- Thus, the y values will have to be either greater than 1 (for point in class 1) or less than -1 (for points in class -1).
- The SVM algorithm finds the weights vector (W) for the features, such that there is a widest margin between the two categories.
- Computing an SVM using these equations is a hill-climbing process problem in a convex space.
- However, by working with points nearest to the classifying boundary only, it reduces sharply the number of data instances to work with this approach reduces its memory requirements for computation.
- This is possible because of using kernel methods.

THE KERNEL METHOD

- The heart of an SVM algorithm is the kernel method. Most kernel algorithms are based on optimization in a convex space and are statistically well-founded.
- Kernel stands for the core or the germ in a fruit. Kernel methods operate using what is called the *'kernel trick'*.
- This trick involves computing and working with the inner products of only the relevant pairs of data in the feature space; they do not need to compute all the data in a high-dimensional feature space.
- The kernel trick makes the algorithm much less demanding in computational and memory resources.
- Kernel methods achieve this by learning from instances. So it is called instance-based learning.

- There are several types of support vector models including linear, polynomial, RBF, and sigmoid.
- SVMs have evolved to be more flexible and be able to tolerate some amount of misclassification the margin of separation between the categories is thus a *'soft margin'* as against a *hard margin*.

ADVANTAGES AND DISADVANTAGES OF SVMs

1. The main strength of SVMs is that they work well even when the number of features is much larger than the number of instances. It can work on datasets with huge feature space; such is the case in spam filtering, where a large number of words are the potential signifiers of a message being spam.
2. Another advantage of SVMs is that even when the optimal decision boundary is a nonlinear curve, the SVM transforms the variables to create new dimensions such that the representation of the classifier is a linear function of those transformed dimensions of the data. ‘
3. SVMs are conceptually easy to understand. They create an easy-to-understand linear classifier. By working on only a subset of relevant data, they are computationally efficient. SVMs are now available with almost all data analytics toolsets.
4. The SVM technique has two major constraints
   a. It works well only with real numbers, i.e, all the data points in all the dimensions must be defined by numeric values only
   b. It works only with binary classification problems. One can make a series of cascaded SVMs to get around this constraint.
5. Training the SVMs is an inefficient and time consuming process, when the data is large. It does not work well when there is much noise in the data, and thus has to compute soft margins. The SVMs will also not provide a probability estimate of classification, i.e., the confidence level for classifying an instance.

WEB MINING

INTRODUCTION

Web Mining is the art and science of discovering patterns and insights from the World Wide Web so as to improve it.

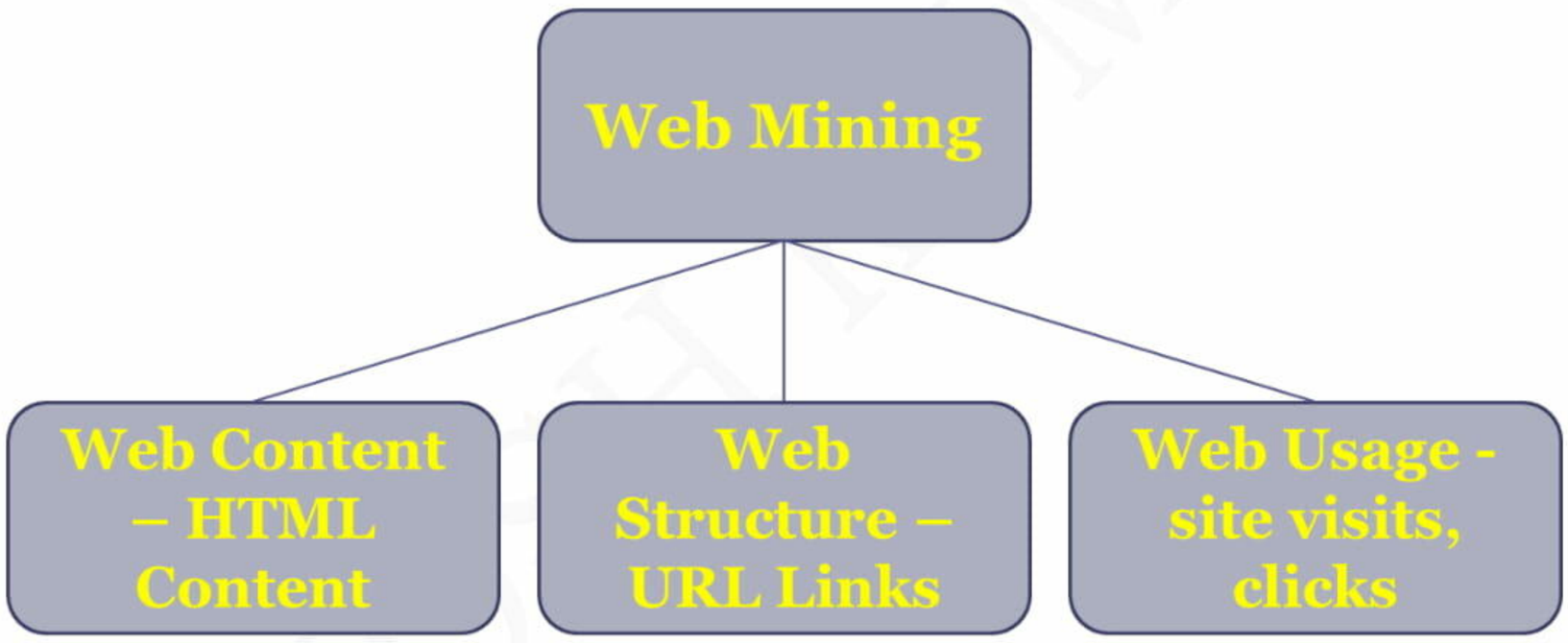
- Web mining analyzes data from the web and helps find insights that could optimize the web content and improve the user experience.
- Data for web mining is collected via web crawlers, web logs, and other means.

Here are some characteristics of optimized websites

1. **Appearance** Aesthetic design, well-formatted content, easy to scan and navigate, good color contrasts
2. **Content** well-planned information architecture with useful content, Fresh content, search engine optimized, links to other good sites
3. **Functionality** accessible to all authorized users, fast loading times, usable forms, mobile enabled.

- The analysis of web usage provides feedback on the web content and also the consumer's browsing habits.
- This data can be of immense use for commercial advertising and even for social engineering.
- The web could be analyzed for its structure as well as content.
- The usage pattern of web pages could also be analyzed.
- Depending upon objectives, web mining can be divided into three different types-web usage mining.

Web content mining and web structure mining (Figure 14.1)



WEB CONTENT MINING

- A website is designed in the form of pages with a distinct URL (Universal Resource Locator).
- A large website may contain thousands of pages.
- These pages (and their content) are managed using specialized software systems called Content Management Systems.
- Every page can have text, graphics, audio, video, forms, applications, and more kinds of content including user generated content.
- The websites keep a record of all requests received for its page/URLs, including the requester information using 'cookies'.
- The log of these requests could be analyzed to gauge the popularity of those pages among different segments of the population.

- The text and application content on the pages could be analyzed for its usage by visit counts.
- The pages on a website themselves could be analyzed for quality of content that attracts most users.
- Thus, the unwanted or unpopular pages could be weeded out or they can be transformed with different content and style. Similarly, more resources could be assigned to keep the more popular pages fresh and inviting.

WEB STRUCTURE MINING

- The web works through a system of hyperlinks using the hypertext protocol (http).
- Any page can create a hyperlink to any other page, i.e., it can be linked to by another page.
- The intertwined or self-referral nature of web lends itself to some unique network analytical algorithms.
- The structure of web pages could also be analyzed to examine the pattern of hyperlinks among pages.

There are two basic strategic models for successful websites -Hubs and Authorities.

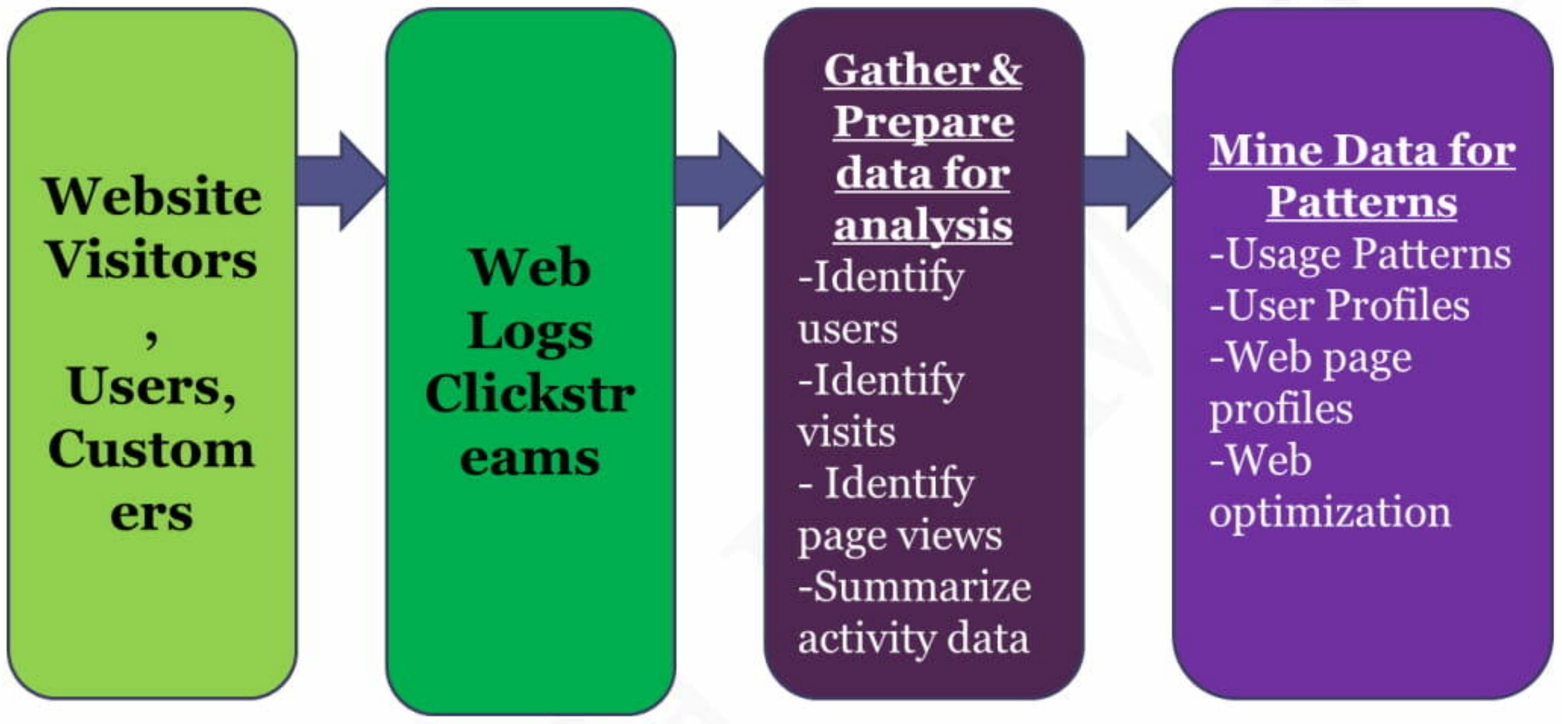
1. **Hubs**
   - These are pages with a large number of interesting links.
   - They serve as a hub or a gathering point, where people visit to access a variety of information.
   - Media sites like yahoo.com or government sites could serve that purpose.
   - More focused sites like traveladvisor.com and yelp.com could aspire to becoming hubs for new emerging areas.
2. **Authorities**
   - Ultimately, people would gravitate towards pages that provide the most complete and authoritative information on a particular subject. This could be factual information, news, advice, user reviews etc.
   - These websites have the most number of inbound links from other websites.
   - Thus, mayoclinic.com could serve as an authoritative page for expert medical opinion; NYtimes.com could serve as an authoritative page for daily news.

WEB USAGE MINING

- As a user clicks anywhere on a webpage 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 make a record of the pages served and the user activity on those pages.

- The entities between the client and the server, such as the router, proxy server, or ad server too would record that click.
- The goal of web usage mining is to extract useful information and patterns from data generated through web page visits and transactions.
- 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 analyzed at multiple levels (Figure 14.2).



- The server side analysis would show the relative popularity of the web pages accessed. Those websites could be hubs and authorities.
- The client side analysis would focus on the usage pattern or the actual content consumed and created by users.
    a) Usage pattern could be analyzed using 'clickstream' analysis, i.e., analyzing web activity for patterns of sequence of clicks, and the location duration of visits on websites. Clickstream analysis can be useful forweb 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 t0pics, user segmentation, and sentiment analysis.
- Web usage mining has many business applications. It can help predict user behavior based on previously learned rules and users' profiles, and can help determine lifetime value of clients. It can also help design cross-marketing strategies across products by observing association rules among the pages on the website.

- Web usage can help evaluate promotional campaigns and see if the users were attracted to the website and used the pages relevant to the campaign.
- Web usage mining could be used to present dynamic information to users based on their interests and profiles. This includes targeted online ads and coupons at user groups based on user access patterns.

WEB MINING ALGORITHMS

- Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that rates web pages as being hubs or authorities.
- The most famous and powerful of these algorithms is the *Page Rank algorithm*. Invented by Google cofounder Larry Page, this algorithm i9 used by Google to organize the results of its search function.
- This algorithm helps determine the relative importance of any particular web page by counting the number and quality of links to a page.
- The websites with more number of links, and/or more links from higher quality websites, will be ranked higher.
- It works in a similar way as determining the status of a person in a society of people.
- Those with relations to more people and/or relations to people of higher status will be accorded a higher status.
- Page Rank is the algorithm that helps determine the order of pages listed upon a Google search query.
- The original Page Rank algorithm formulation has been updated in many ways and the latest algorithm is kept a secret so other websites cannot take advantage of the algorithm and manipulate their website according to it.
- However, there are many standard elements that remain unchanged. These elements lead to the principles for a good website. This process is also called Search Engine Optimization (SEO).

Social Network Analysis

INTRODUCTION

Social Networks are a graphical representation of relationships among people and or entities.

Social Network Analysis (SNA) is the art and science of discovering patterns of interaction and influence within the participants in' a network.

- These participants could be people, organizations, machines, concepts, or any other kinds of entities.
- An ideal application of social network analysis will discover essential characteristics of a network including its Central nodes and its subnetwork structure.

- Subnetworks are clusters of nodes, Where the within subnetwork connections are stronger than the connections with nodes outside the subnetwork.
- SNA is accomplished through graphically representing social relationships into a network of nodes and links and applying iterative computational techniques to measure the strengths of relationships.
- The social network analysis ultimately helps relate the totality of network to the Unified Field which is the ultimate entity with infinite relationships among everything.

Applications of SNA

1. **Self-awareness** Visualizing his/her social network can help a person organiz their relationships and support network.
2. **Communities** Social Network Analysis can help identification, construction and strengthening of networks within communities to build wellness, comfort and resilience. Analysis of joint authoring relationships and citations help identify subnetworks of specializations of knowledge in an academic field. Researchers at Northwestern University found that the most determinant factor in the success of a Broadway play was the strength of relationships amongst the crew and cast.
3. **Marketing** There is a popular network insight that any two people are related to each other through at most seven degrees of links. Organizations can use this insight to reach out with their message to large number of people and also to listen actively to opinion leaders as ways to understand their customers' needs and behaviors. Politicians can reach out to opinion leaders to get their message out.
4. **Public Health** Awareness of networks 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.

Network Topologies

There are two primary types of network topologies *ring-type* and *hub-spoke* topologies. Each of the topologies has different characteristics and benefits.

- In the *ring network*, nodes typically connect to their adjacent nodes in the network and all nodes can be connected to each other.
- A ring network could be dense where every node has a direct connection with practically every node or it could be sparse where every node connects to a small subset of the nodes.
- A dense network, with more connections, will allow many direct connections between pairs of nodes.
- In a sparse network, one may need to traverse many connections to reach the desired destination. A peer-to-peer email (or messaging) network is an example on the ring model, as anyone can potentially directly connect with anyone else.

- In the hub-spoke model, there is one central hub node to which all the other node are connected and no direct relationships between the nodes.
- Nodes connect with each other through the hub node.
- This is a hierarchical network structure since the hub node is central to the network.
- The hub node is structurally more important as it is central to all communications between other peripheral nodes,
- The hub-spoke network is that one could predictably reach from any node to any other node through traversing exactly just two connections.
- As an example, modern airlines operate on this model to maintain hub networks from which they operate flights to a large number of airports.
- The density of a network can be defined as the average number of connections per node.
- The cohesiveness of the network is a related concept, which is the average number of connections needed to reach from one node to the other.
- Another way of analyzing networks is to define the centrality (or importance) of wary node. The number of links associated with a node is a sign of centrality of the node.
- In the ring network in the figure below, each node has exactly 2 links. Thus, there is no central node. However, in the hub-spoke network, the hub-node N has 8 links while all other nodes have only 1 link each. Thus, node N has a high centrality.

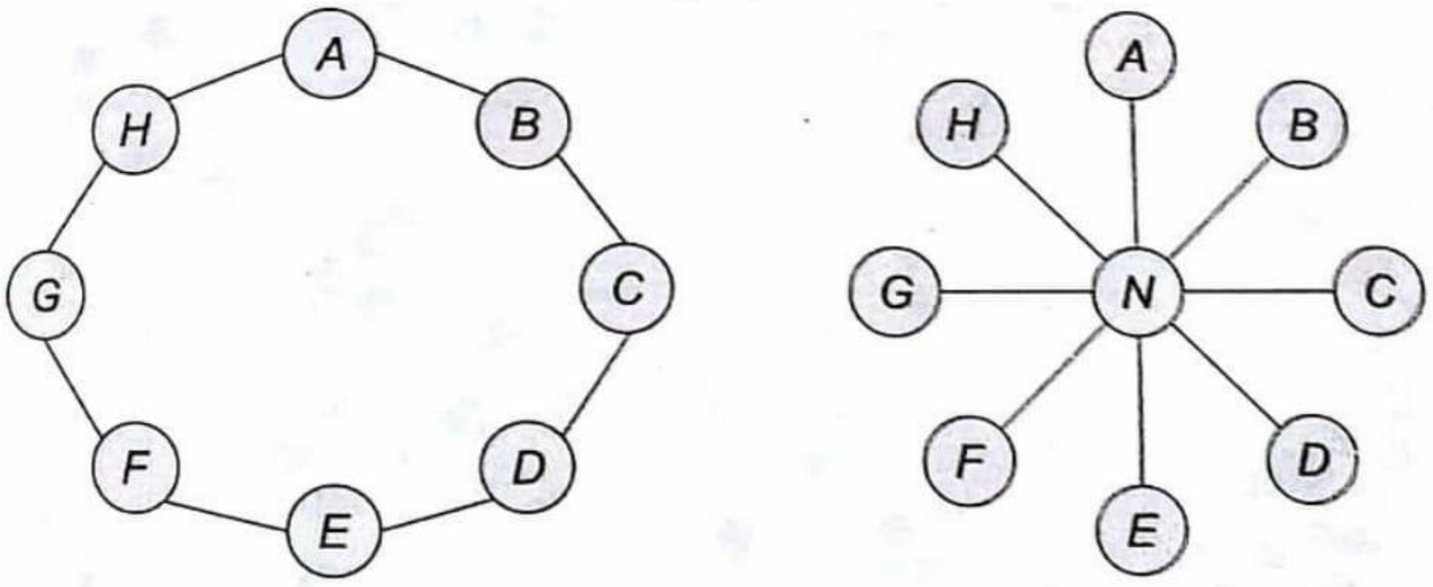


FIGURE 15.1 Network Topologies: Ring (left) and Hub-spoke (right)

A variant that combines both the above types is the network of networks in which each participating network will connect with other networks at selected points of contact.

For example, the internet is a network of networks. Here, all the commercial, university, government and similar computer networks connect to one other at certain designated nodes (called gateways) to exchange information and conduct business.

TECHNIQUES AND ALGORITHMS

There are two major levels of social network analysis *discovering subnetworks* within the network *and ranking the nodes* to find more important nodes or hubs

Finding Subnetworks

- A large network could be better analyzed and managed if it can be seen as an interconnected set of distinct subnetworks each with its own distinct identity and unique characteristics.
- This is like doing a cluster analysis of nodes. Nodes with strong ties between them would belong to the same subnetwork, while those with weak or no ties would belong to separate subnetworks.
- This is unsupervised learning technique, as in Apriori there is no correct number of subnetworks in a network.
- The usefulness of the subnetwork structure for decision-making is the main criterion for adopting a particular structure.

FIGURE 15.2 A Network with Distinct Subnetworks

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 influence.
- Every incoming link is similarly an inflow of influence. More in-links to a node means greater importance.
- Thus there will be many direct and indirect flows of influence between any two nodes in the network.
- Computing the relative influence of each node is done on the basis of an input output matrix of flows of influence among the nodes.

- Assume each node has an influence value. The computational task is to identify a set of rank values that satisfies the set of links between the nodes.
- It is an iterative task where we begin with some initial values and continue to iterate till the rank values stabilize.

Consider the following simple network with 4 A ' B nodes (A, B, C, D) and 6 directed links between them as shown in Figure 15.3. Note that there is a bidirectional link. Here are the links

Node A links into B
Node B links into C
Node C links into D
Node D links into A
Node A links into C
Node B links into A.

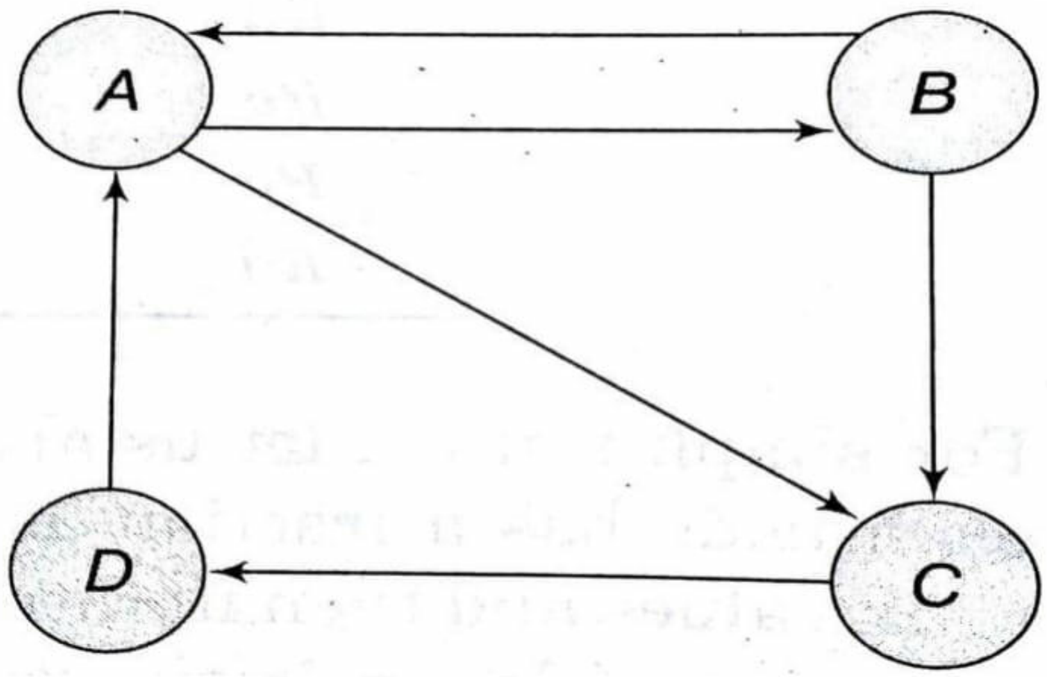


FIGURE 15-3

- The goal is to find the relative importance, or rank, or every node in the network. This will help identify the most important node(s) in the network.
- We begin by assigning the variables for influence (or rank) value for each node s **R_a, R_b, R_c, and R_d**. The goal is to find the relative values of these variables
- There are two outbound links from node A to nodes B and C. Thus, both B and C receive half of node A's influence. Similarly, there are two outbound links from node B to nodes C and A, so both C and A receive half of node B's influence.
- There is only outbound link from node D into node A. Thus, node A gets all the influence of node D. There is only outbound link from node C into node D and hence, node D gets all the influence of node C.
- Node A gets all of the influence of node D and half the influence of node B,

Thus, $$R_a = 0.5 \times R_b + R_d$$

Node B gets half the influence of node A.

Thus, $$R_b = 0.5 \times R_a$$

Node C gets half the influence of node A and half the influence of node B.

Thus, $R_c = 0.5 \times R_a + 0.5 \times R_b$

Node D gets all of the influence of node C and half the influence of node B.

Thus, $R_d = R_c$

We have 4 equations using 4 variables. These can be solved mathematically.

We can represent the coefficients of these 4 equations in a matrix form as shown in Dataset 15.1 given below. This is the Influence Matrix. The zero values represent that the term is not represented in an equation.

Dataset 15.1

	Ra	*Rb*	*Rc*	*Rd*
Ra	0	0.50	0	1.00
Rb	0.50	0	0	0
Rc	0.50	0.50	0	0
Rd	0	0	1.00	0

For simplification, let us also state that all the rank values add up to 1. Thus, each node has a fraction as the rank value. Let us start with an initial set of rank values and then iteratively compute new rank values till they stabilize. One can start with any initial rank values, such as 1/n or 1/4 for each of the nodes.

Variable	Initial Value
Ra	0.250
Rb	0.250
Rc	0.250
Rd	0.250

Computing the revised values using the equations stated earlier, we get a revised set of values shown as Iterationl. (This can be computed easily by creating formulae using the influence matrix in a spreadsheet such as Excel.)

Variable	Initial Value	Iteration1
Ra	0.250	0.375
Rb	0.250	0.125
Rc	0.250	0.250
Rd	0.250	0.250

Using the rank values from Iterationl as the new starting values, we can compute new values for these variables, shown as Iteration2 Rank values will continue to change.

Variable	Initial Value	Iteration1	Iteration2
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

Working from values of Iteration2 and so, we can do a few more iterations till the values stabilize. Dataset 15.2 shows the final values after the 8th iteration.

Dataset 15.2

Variable	Initial Value	Iteration1	Iteration2	...	Iteration8
Ra	0.250	0.375	0.313	...	0.333
Rb	0.250	0.125	0.188	...	0.167
Rc	0.250	0.250	0.250	...	0.250
Rd	0.250	0.250	0.250	...	0.250

- The final rank shows that rank of node A is the highest at 0.333. Thus, the most important node is A. The lowest rank is 0.167 of Rb.
- Thus, B is the least important node. Nodes C and D are in the middle. In this case, their ranks did not change at all.
- The relative scores of the nodes in this network would have been the same irrespective of the initial values chosen for the computations.
- It may take longer or shorter number of iterations for the results to stabilize for different sets of Initial values.

PAGERANK

- PageRank is a particular application of the social network analysis techniques above to compute the relative importance of websites in the overall World Wide Web.
- The data on websites and their links is gathered through web crawler bots that traverse through the webpages at frequent intervals.
- Every webpage is a node in a social network and all the hyperlinks from that page become directed links to other webpages.
- Every outbound link from a webpage is considered an outflow of influence of that webpage.

- An iterative computational technique is applied to compute a relative importance to each page.
- That score is called PageRank according to an eponymous algorithm invented by the founders of Google, the web search company.
- PageRank is used by Google for ordering the display of websites in response to search queries.
- To be shown higher in the search results, many website owners try to artificially boost their PageRank by creating many dummy websites whose ranks can be made to flow into their desired website.
- Also, many websites can be designed to cyclical sets of links from where the web crawler may not be able to break out. These are called spider traps.
- To overcome these and other challenges, Google includes a Teleporting factor into computing the PageRank.
- Teleporting assumed that there is a potential link from any node to any other node, irrespective of whether it actually exists.
- Thus, the influence matrix is multiplied by a weighting factor called Beta with a typical value of 0.85 or 85 percent.
- The remaining weight of 0.15 or 15 percent is given to teleportation.
- In teleportation matrix, each cell is given a rank of 1/n. where n is the number of nodes the web.
- The two matrices are added to compute the final influence matrix. This matrix can be used to iteratively compute the PageRank of all the nodes, just as shown in the example earlier.

PRACTICAL CONSIDERATIONS

1. **Network Size** Most SNA research is done using small networks. Collecting data about large networks can be very challenging. This is because the number of links is the order of the square of the number of nodes. Thus, in a network of 1000 nodes there are potentially 1 million possible pairs of links.
2. **Gathering Data** Electronic communication records (emails, chats, etc.) can be harnessed to gather social network data more easily. Data on the nature and quality of relationships need to be collected using survey documents. Capturing and cleansing and organizing the data can take a lot of time and effort, just like in a typical data analytics project.
3. **Computation and Visualization** Modeling large networks can be computationally challenging and visualizing them also would require special skills. Big data analytical tools may be needed to compute large networks.
4. **Dynamic Networks** Relationships between nodes in a social network can be fluid. They can change in strength and functional nature. For example, there could be multiple relationships between two people they could simultaneously be coworkers, coauthors,

and spouses. The network should be modeled frequently to see the dynamics of the network.

Table 15.1 Social Network Analysis vs Traditional Data Analytics

Dimension	Social Network Analysis	Traditional Data Mining
Nature of learning	Unsupervised learning	Supervised and unsupervised learning
Analysis of goals	Hub nodes, important nodes, and subnetworks	Key decision rules, cluster centroids
Dataset structure	A graph of nodes and (directed) links	Rectangular data of variables and instances
Analysis techniques	Visualization with statistics; iterative graphical computation	Machine learning, statistics
Quality measurement	Usefulness is key criterion	Predictive accuracy for classification techniques

Reference text book:

Anil Maheshwari, "Data Analytics", 1st Edition, McGraw Hill Education, 2017.
ISBN-13: 978-9352604180

For Study material Visit:

sites.google.com/view/dksbin