UNIT 5

DATA MINING

- It is also known as knowledge Discovery (KDD) from data.
- Extracting hidden info or hidden patterns
- Data mining is also known as KDD or knowledge discovery in database.
- Data processing using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large pre-existing databases. A way to discover new meaning in data.

- Supervised and Unsupervised algorithms are used to identify the hidden patterns in data

Datasets can be generalized into 3 types

1) Data with labels → Supervised
2) Data without labels → Unsupervised
3) Data with only a small portion of labels

Supervised algorithm example:

→ Classification

Supervised example:

→ Classification

* Supervised approaches depend on some a priori knowledge of the data (class labels)

* Unsupervised algorithms are used to classify data without any prior knowledge. All kinds of patterns will be discovered by the algorithm.

* Classification is a common supervised approach as it is appropriate when the dataset has labels of a store portion of the data has labels

* Clustering is a common unsupervised data mining technique that is used when dealing with datasets without labels

* For classification - The algorithm learns from the training data and builds a model that will automatically categorize new data elements into one of the predefined classes providing with the training data.

- for clustering - clustering algo determine which elements in the dataset are similar to each other based on the similarity of the data elements.
  It makes use of euclidean distance.

- Clustering techniques use keywords that are represented as a vector (To represent a document) and the cosine similarity measure is used to distinguish how similar one vector (data element) is to another

Motivations for Data mining in social media

→ The data available via social media can give us insights into social networks in ways that were not previously possible in both scale and extend

→ Data mining techniques can help effectively deal with the 3 main challenges with social media data

(2) Social media datasets are large. Eg: Fb
(1) SM datasets can be noisy & Missing data.
(3) Data from online SM is dynamic
~~dynamic~~

* Applying data mining techniques to large social media datasets has the potential to continue to improve search results for everyday search engines (eg google search engine), realize ~~sep~~ specialized target marketing for business, help psychologist to study behaviors. Provide new insight into social sciences for sociologists, personalized web services for consumers and even help detect & prevent spam for all of us.

Data mining methods for social media

Applications that apply data mining techniques developed by industry and academia are already being used commercially

Eg: Samepoint, an social media analytics company, provides services to mining and monitor social media to provide business information about how

Relationships are annotation the individuals are represented as link in the graph

A graph representation enables the application of classical representation of mathematical graph theory, traditional social network analysis method, and work on mining graph data.

Representing online social media data as graphs enables leveraging work done with graphs related to influence propagation, community detection, and link prediction.

Data mining — A process

Different datasets and data questions require diff types of tools. If it is known how data can should be organized & classification tool will be appropriate

If you understand what the data is about & correct answers

trends and patterns in data, a clustering tool may be used,

Before data mining can take place

With Data represented as a graph the work begins with the selected number of nodes, known as seeds. Graphs are traversed beginning with a set of seeds and as the link structure from the seed nodes is exploited, data is collected and the structure itself is also analyzed. Using the link structure to extent from the seed set & gather new information is known as Crawling the network.

As the crawler discovers new information, it stores the new information in a repository for further analysis of data.

* 2 specific types of SM data :-

1) Social Networking Site → FB

2) Blogosphere.

A social networking site like FB / LinkedIn consist of connected users & user unique profile

Eg: Hypothetical graph structure diagram for a typical social networking site

Social networking site provide excellent source of data for studying collaboration relationships, group structure and who talks to whom

Thus a graph structure should be utilized

The driving factors for mining social networking sites is the "unique opportunity to understand the impact of a person position in the network on everything from their tastes to, moods to their health

The most common data mining application related to social networking site include

1) Group Detection

Finding and identifying the group

Finding individuals that associate more with each other than other users.

A clustering approach such as modularity maximization can be used to detect the subgroups

(2) Group Profiling

"What is this group about?" Advanced data mining techniques such as Topic Taxonomy, a tree structure can be utilized.

(3) Recommendation Systems

A recommendation system analyzes social networking data & recommends new friends or new groups to a user. Eg:- Facebook.

- friend suggestion. The ability to recommend group membership to an individual is advantageous for the group that would like to have additional members & can be helpful to an individual who is looking to find other individuals with similar interests or goals.

Recommendations are used on user profile data & users associations both structural which can be used to provide suggestions to users. There are 3 steps for this

(1) Identify features of the profile.

2) Group Members are clustered.

(3) A decision tree is created.

Blogosphere

Need not be blogs

Web logs of blogs a user published journals available on the web.

Blogs are typically open to the public & provide a mechanism for readers to comment on specific post.

The set of all blogs and blog posts is referred to as blogosphere.

Clustering, matrix factorization and ranking are the most commonly used techniques for data mining in blogosphere.

4 measures / factors to identify influential nodes

(1) Recognition

(2) Activity Generation

(3) Novelty

(4) Eloquence

B) Topic detection and change

(4) Sentiment Analysis

The most common data mining application related to the blogosphere include.

(1) Blog classification

- Organize blogs by topic

  Using SVM (Support Vector Machine) automatically blogs can be classified using descriptive words (tags).

- Tags are assigned to the blogpost using SVM classifier.

(2) Identifying influential nodes.

A blogger is influential if he/she has the capacity to effect the behaviour of fellow bloggers

The benefit of being able to identify influential bloggers, blogs, blogpost is that marketing efforts for sales & service could be focused on points of influence that are most likely to gain support for a topic, product or essay

[illegible] and services perceive and discussed [illegible] social media. Researches in other [illegible] have applied text mining techniques and propagation models to [illegible] to develop approaches for better understanding how information moves [illegible] the ~~global~~ blogosphere.

Data mining technique can be applied in social media to understand behavior and to make use of data for research and business purpose. Representative areas include community and group detection, information diffusion, influence propagation, topic detection and monitoring, individual behaviour analysis, groups behaviour analysis and marketing researches for business.

Data Representation

It is common to use a graph representation to study social media datasets. A graph consists of a set containing vertexes, nodes and edges (links). Individuals are usually represented as nodes of the graph.