Unit 5
BIG DATA APPLICATIONS:

Application Evolution, Big Data Analysis Fields - Structured Data Analysis, Text Data Analysis, Web Data Analysis, Multimedia Data Analysis, Network Data Analysis, Mobile Traffic Analysis, Key Applications - Application of Big Data in Enterprises, Application of IoT Based Big Data, Application of Online Social Network Oriented Big Data, Applications of Healthcare and Medical Big Data, Collective Intelligence, Smart Grid.

Application Evolution

Evolution of Commercial Applications:

The earliest business data was generally structured data, which was collected by companies from old systems and then stored in RDBMSs.

Analytical technologies used in such systems were prevailing in 1990s and was intuitive and simple, e.g., reports, instrument panels, special queries, search-based business intelligence, online transaction processing, interactive visualization, score cards, predictive modeling, and data mining.

Abundant products and customer information, including clickstream data logs and user behavior, etc., can be acquired from the websites.

Product layout optimization, customer trade analysis, product suggestions, and market structure analysis can be conducted by text analysis and website mining technologies.

Evolution of Network Applications:

The early Internet mainly provided email and web page services.

Text analysis, data mining, and webpage analysis technologies have been applied to the mining of email contents and building search engines.

Nowadays, most applications are web-based, regardless of their application field and design goals.

Network data accounts for a major percentage of the global data volume.

Web has become a common platform for interconnected pages, full of various kinds of data, such as text, images, videos, pictures, and interactive contents, etc.

Therefore, plentiful advanced technologies used for semi-structured or unstructured data emerged at the right moment.

For example, the image analysis technology may extract useful information from pictures, e.g., face recognition.

Multimedia analysis technologies can be applied to the automated video surveillance systems for business, law enforcement, and military applications.

Since 2004, online social media, such as Internet forums, online communities, blogs, social networking services, and social multimedia websites, etc., provide users with great opportunities to create, upload, and share contents generated by users.

Different user groups may search for daily news and celebrity news, publish their social and political opinions, and provide different applications with timely feedback.

Evolution of Scientific Applications:

Scientific research in many fields is acquiring massive data with high-throughput sensors and instruments, such as astrophysics, oceanology, genomics, and environmental research.

The U.S. National Science Foundation (NSF) has recently announced the BIGDATA Research Initiative to promote research efforts to extract knowledge and insights from large and complex collections of digital data.
Some scientific research disciplines have developed massive data platforms and obtained useful outcomes.
For example, in biology, iPlant [3] applies network infrastructure, physical computing resources, coordination environment, virtual machine resources, and inter-operative analysis software and data service to assist researchers, educators, and students in enriching all plant sciences.

Big Data Analysis Fields

Data analysis research can be divided into six key technical fields,
i.e., structured data analysis,
text data analysis,
website data analysis,
multimedia data analysis,
network data analysis, and mobile data analysis.
Such a classification aims to emphasize data characteristics, but some of the fields may utilize similar technologies.

Structured Data Analysis

Business applications and scientific research may generate massive structured data, of which the management and analysis rely on mature commercialized technologies, such as RDBMS, data warehouse, OLAP, and BPM (Business Process Management) .
Data analysis is mainly based on data mining and statistical analysis, both of which have been well studied over the past 30 years.
Data analysis is still a very active research field and new application demands drive the development of new methods.
Statistical machine learning based on exact mathematical models and powerful algorithms have been applied to anomaly detection and energy control .
Exploiting data characteristics, time and space mining may extract knowledge structures hidden in high-speed data flows and sensor data models and modes .
Driven by privacy protection in e-commerce, e-government, and health care applications, privacy protection data mining is an emerging research field .
Over the past decade, benefited by the substantial popularization of event data, new process discovery, and consistency check technologies, process mining is becoming a new research field especially in process analysis with event data

Text Data Analysis

The most common format of information storage is text, e.g., email communication, business documents, web pages, and social media.
Therefore, text analysis is deemed to feature more business-based potential than structured data mining.
Generally, tax analysis, also called text mining, is a process to extract useful information and knowledge from unstructured text.
Text mining is an interdisciplinary problem, involving information retrieval, machine learning, statistics, computing linguistics, and data mining in particular.

Most text mining systems are based on text expressions and natural language processing (NLP), with more focus on the latter.
Text summarization is to generate a reduced summary or extract from a single or several input text files.
Text summarization may be classified into concrete summarization and abstract summarization.
Concrete summarization selects important sentences and paragraphs from source documents and concentrates them into shorter forms.
Abstract summarization may interpret the source texts and, according to linguistic methods, use a few words and phrases to represent the source texts.
Text classification is to recognize probabilistic topics of documents by putting documents in scheduled topics.
Text classification based on the new graph representation and graph mining has recently attracted considerable interest .
Text clustering is used to group similar documents with scheduled topics, which is different from text classification that gathers documents together.
In text clustering, documents may appear in multiple subtopics.

Web Data Analysis

Web analysis has emerged as an active research field.
Web analysis aims to automatically retrieve, extract, and evaluate information from Web documents and services so as to discover useful knowledge.
Web analysis is related to several research fields, including database, information retrieval, NLP, and text mining.
According to the different parts of the Web to be mined, we classify Web analysis into three related fields:
Web content mining,
Web structure mining, and Web usage mining.
Web content mining is the process to discover useful knowledge in Web pages, which generally involve several types of data, such as text, image, audio, video, code, metadata, and hyperlink.
Supervised learning and classification play important roles in hyperlink mining, e.g., email, newsgroup management, and Web catalog maintenance
Web content mining can be conducted with two methods: the information retrieval method and the database method.
Information retrieval mainly assists in or improves information lookup, or filters user information according to deductions or configuration documents.
Web structure mining involves models for discovering Web link structures.
Models are built based on topological structures provided with hyperlinks with or without link description.
Such models reveal the similarities and correlations among different websites and are used to classify website pages.
Web usage mining aims to mine auxiliary data generated by Web dialogues or behaviors. Web content mining and Web structure mining use the master Web data.
Web usage data includes access logs at Web servers, logs at proxy servers, browsers' history records, user profiles, registration data, user sessions or trades, cache, user queries, bookmark

data, mouse click and scroll, and any other kind of data generated through interaction with the Web.
As Web services and Web2.0 are becoming mature and popular, Web usage data will have increasingly high variety.
Web usage mining plays key roles in personalized space, e-commerce, network privacy/security, and other emerging fields.
For example, collaborative recommender systems can personalize e-commerce by utilizing the different preferences of users.

Multimedia Data Analysis

Multimedia data (mainly including images, audios, and videos) have been growing at an amazing speed.
Multimedia content sharing is to extract related knowledge and understand semantics contained in multimedia data.
Because multimedia data is heterogeneous and most of such data contains richer information than simple structured data and text data, extracting information is confronted with the huge challenge of the semantic differences of multimedia data.
Research on multimedia analysis covers many disciplines.
Some recent research priorities include multimedia summarization, multimedia annotation, multimedia index and retrieval, multimedia suggestion, and multimedia event detection, etc.
Audio summarization can be accomplished by simply extracting the prominent words or phrases from metadata or synthesizing a new representation.
Video summarization is to interpret the most important or representative video content sequence, and it can be static or dynamic.
Static video summarization methods utilize a key frame sequence or context-sensitive keyframes to represent a video.
Such methods are very simple and have been applied to many business applications (e.g., Yahoo!, Alta Visa, and Google), but the playback performance is poor.
Dynamic summarization methods use a series of video clips to represent a video, configure low-level video functions, and take other smooth measures to make the final summarization look more natural.
Multimedia index and retrieval involve describing, storing, and organizing multimedia information and assisting users to conveniently and quickly look up multimedia resources.

Network Data Analysis

Network analysis evolved from the initial quantitative analysis and sociological network analysis into the emerging online social network analysis in the beginning of the twenty-first century.
Many prevailing online social networking services, include Twitter, Facebook, and LinkedIn, etc. have been increasingly popular over the years.
Such online social networking services generally include massive linked data and content data.
The linked data is mainly in the form of graphic structures, describing the communications between two entities.
The content data contains text, image, and other network multimedia data.
The rich contents of such networks bring about both unprecedented challenges and opportunities to data analysis.

In accordance with the data-centered perspective, the existing research on social networking service contexts can be classified into two categories:
link-based structural analysis and content-based analysis
The research on link-based structural analysis has always been committed on link prediction, community discovery, social network evolution, and social influence analysis, etc. SNS may be visualized as graphs, in which every vertex corresponds to a user and edges correspond to the correlations among users.
Link prediction is to predict the possibility of future connection between two vertices.
Many technologies can be used for link prediction, e.g., feature-based classification, probabilistic methods, and Linear Algebra.
Feature-based classification is to select a group of features for a vertex and utilize the existing link information to generate binary classifiers to predict the future link .
Probabilistic methods aim to build models for connection probabilities among vertexes in SNS .
Linear Algebra computes the similarity between two vertexes according to the singular similar matrix