Chapter 12. Data Mining Trends and Research Frontiers

- Mining Rich Data Types 
- Data Mining Applications
- Data Mining Methodologies and Systems
- Data Mining, People and Society
- Summary

Mining Rich Data Types

- Mining Text Data
- Mining Spatial-Temporal Data
- Mining Graph and Networks

Mining Text Data

- Over 80% of the data is in an unstructured format, e.g., news, image, audio.
- Goal: deriving high-quality information, such as structures, patterns, and summaries, from **unstructured** text data.
  - Process: mining structures from text ⇒ deriving patterns within the structured data ⇒ evaluating and interpreting output
- Main challenge:
  - How to represent text primitives (e.g., words, sentences, and documents) and how to learn effective representation from unstructured text?

Mining Text Data: Major Tasks

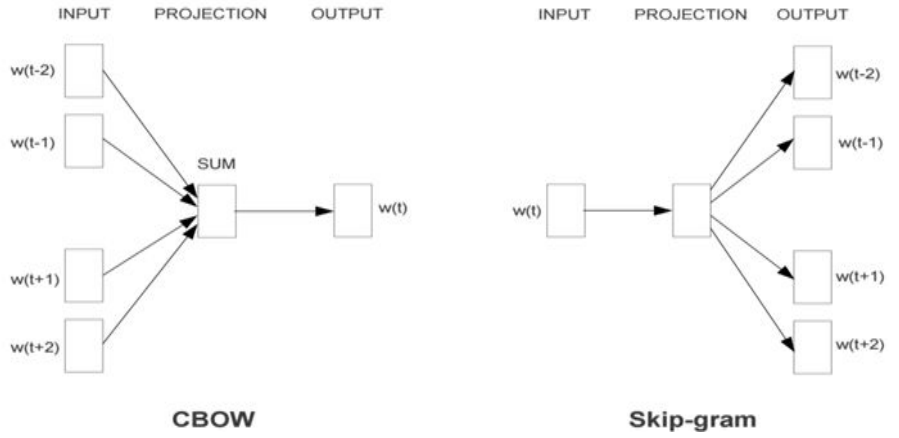
- Information extraction (IE): automatically extracting structured information from unstructured and semi-structured documents.
- Taxonomy construction: constructing different multi-faceted taxonomies based on different corpora or applications.
- Text clustering: grouping unlabeled texts with similar contextual patterns.
- Text classification: analyzing open-ended texts (e.g., webpages) and assigning them a set of predefined topics or categories
- Text summarization: creating a subset of text that represents the most important information from a single or multiple pieces of text.

Mining Text Data: Techniques

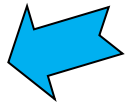
- Symbol-based representation: using one-hot vectors for words and bag-of-words model for documents.
  - Issues: high-dimensionality and data sparsity
- Distributed representation: low-dimensional real-valued dense vectors with large amount of unlabeled data using deep neural network.
  - Word embedding: words similar in meaning are expected to be closer in the embedded space.
  - Pre-trained language model: using large-scale data and more computing resources to generate contextualized embeddings.

Mining Text Data: Techniques

- Word2vec: words that share common contexts are embedded in close proximity, e.g., “Sweden" can be embedded close to Norway, Finland.
  - CBOW: predicts the current word from surrounding context words.
  - Skip-gram: uses the current word to predict the surrounding words.
- GloVe: combining global matrix factorization and local context window methods.
- JoSE: learning word embeddings in hyperbolic space to model hierarchical structures.
- Pre-trained language model: ELMo and BERT



Mining Rich Data Types

- Mining Text Data
- Mining Spatial-Temporal Data 
- Mining Graph and Networks

Mining Spatial-Temporal Data: Properties and Types

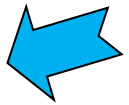
- Key Properties in Spatial and Temporal Data
  - Auto-correlation: dependencies and correlations in proximate locations and time, e.g., similar traffic status for adjacent intersections.
  - Heterogeneity: heterogeneous and non-stationary in distribution.
- Types of Spatial and Temporal Data

Data type	Definition	Operation	Examples
Event	at a spatial location and a time point	intersections and similarity/difference	intersection of forest fire and a rainfall returns the overlapping area and time
Trajectory	a path of an object over space and time	finding relationships between two trajectories	transportation (e.g., tracing vehicle), epidemiology, ecology
Point reference data	collected using discrete reference points	reconstructing fields and modeling non-stationary random process	using mobile sensors to collect temperature data in a certain region
Raster data	recording observation data at fixed locations and fixed time	converting a raster to a finer or coarser resolution	estimate traffic volume using data from nearby sensors/cameras

Mining Spatial-Temporal Data: Data Models

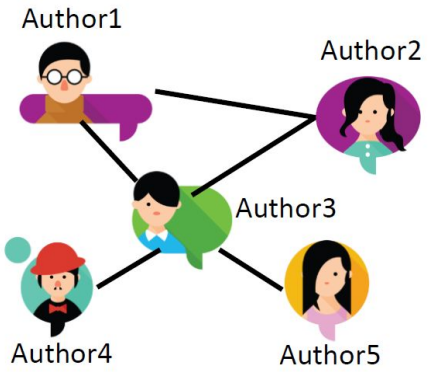
- Data Models for Spatial Data
  - Object model: points, lines, polygons (e.g., point ⇒ vehicle, line ⇒ road)
  - Field model: spatial information as a function (e.g., temperature in an area)
  - Spatial network model: graphs for relationship (e.g., road network)
- Data Models for Temporal Data
  - Temporal snapshot model: multiple spatial layers associated with time.
  - Temporal change model: a spatial theme using a start time and the incremental changes (e.g., moving of a vehicle with initial location, speed)
  - Event/process model: multiple events and processes over time

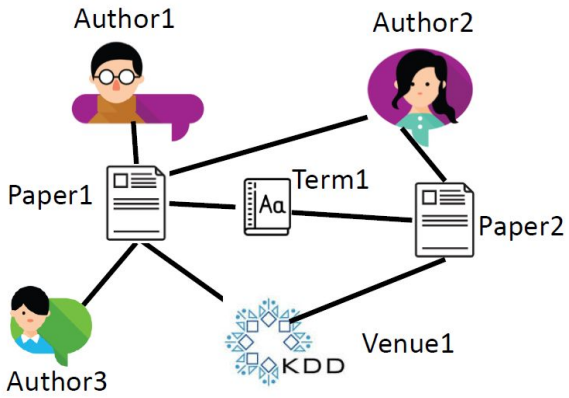
Mining Rich Data Types

- Mining Text Data
- Mining Spatial-Temporal Data
- Mining Graph and Networks 

Mining Graph and Networks

- Graph/Networks: consist of (1) a set of nodes and (2) a collection of edges, to model complex systems, e.g., online social platform
- Categorization
  - attributed vs. plain graphs: node/edge attributes exist or not.
  - dynamic vs. static graphs: graph topology/attribute change or not.
  - homogeneous vs. heterogeneous graphs: nodes/edges are of the same type or not.





Homogeneous vs. heterogeneous networks

Mining Graph and Networks: Problems

- Graph Modeling: to study and simulate the generation mechanisms.
  - Motivation: (1) understanding how graphs are created and formed, (2) helping protect user privacy by anonymizing social networks.
  - Graph models:
    - Erdős–Rényi model: assumes certain probability distributions.
    - realistic graph generators: mimic one or more properties of real-world graphs, e.g., preferential attachment based (Barabási–Albert model).
    - deep generative models: mimic the real graphs with neural networks.
  - Challenge: heterogeneous networks

Mining Graph and Networks: Problems

- Heterogeneous Network Mining: capturing rich semantics
  - Core concept: **meta-path** represents a sequence of relations, e.g., author-paper-author $\Rightarrow$ co-authorship relation.
  - Applications of meta-path
    - Similarity search: improving ranking accuracy
    - Node classification: similarity matrices can be used to infer node labels
    - Clustering: representing different semantics
    - heterogeneous network embedding: constructing neighborhood
  - Challenge: automatically inferring structure and semantics from input data

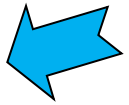
Mining Graph and Networks: Problems

- Knowledge Graph Mining
  - Knowledge graph (KG): a directed heterogeneous graph to link concepts and entities through human-interpretable semantics, e.g., DBPedia, Google KG.
  - KG construction methods: labor intensive, relying on pre-specified ontology
    - Extractive construction: information extraction, e.g., semantic labeling.
    - Language model-based
    - Representation learning-based: link prediction, entity resolution
  - Challenges: (1) automatic knowledge graph construction and dynamic update, (2) medical AI, conversational AI, and academic search engine

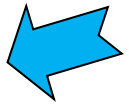
Summary: Mining Rich Data Types

- Mining Text Data
  - Major tasks: information extraction, taxonomy construction, text clustering/classification/summarization
  - Techniques: symbol-based and distributed representation (e.g., word2vec)
- Mining Spatial-Temporal Data
  - Types: event, trajectory, point reference data, raster data
  - Models: object/field/spatial network and snapshot/change/event-process
- Mining Graph and Networks
  - Problems: graph modeling, heterogeneous graph mining, knowledge graph

Chapter 12. Data Mining Trends and Research Frontiers

- Mining Rich Data Types
- Data Mining Applications 
- Data Mining Methodologies and Systems
- Data Mining, People and Society
- Summary

Data Mining Applications

- Data Mining for Sentiment and Opinion 
- Truth Discovery and Misinformation Identification
- Information and Disease Propagation
- Productivity and Team Science

Sentiment and Opinion

- **Sentiment**: a view of attitude/emotion toward a situation or event.
- **Opinion**: a view or judgement formed about something.
- Sentiment and opinion are heavily correlated.
  - Note: many opinions may express sentiment, but some do not. (e.g., “I think she will move to Canada after graduation.”)
- Subtle difference between sentiment and opinion:
  - Sentiment sentence example: “I find data mining highly interesting.”
  - Opinion sentence example: “I believe data mining is promising and useful.”
  - The subtle difference is that one may share the same sentiment expressed in the example of sentiment sentence and agree or disagree with the opinion in the second sentence.

Sentiment Analysis and Opinion Mining Techniques

- Goal: understand sentiment expressed in text.
- Techniques being used:
  - Natural language processing methods.
  - Machine learning and data mining methods.
  - Psychology and social sciences.
- Technical directions being explored:
  - Analysis at different levels: Document-level, sentence-level, aspect-level.
  - Sentiment classification based on lexicons.
  - Two types of text content: *stand-alone* and *online conversations*.
  - Mining intent: intent may imply sentiments and express opinions.
  - Opinions spam detection: acquire data for social goodness.

Sentiment Analysis and Opinion Mining Applications

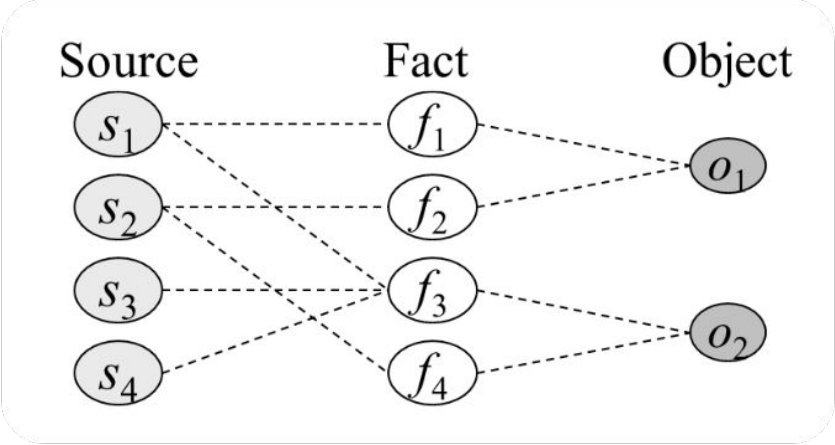
- Understand advantages/disadvantages of targets:
  - Sentiment analysis/opinion mining over product reviews.
  - Help customer make decisions; Help platform optimize supply chain; Help producers and manufacturers improve products, etc.
- Outcome can be consumed by other AI agents:
  - Example: sentiment from news ➡ trading signals for stock markets.
  - Challenges: highly efficient/accurate/robust system.
- Monitoring and government administration:
  - Fight against cyber-violence, terrorism and racism.
  - Challenges: fairness and privacy issues.

Data Mining Applications

- Data Mining for Sentiment and Opinion
- Truth Discovery and Misinformation Identification 
- Information and Disease Propagation
- Productivity and Team Science

Truth Discovery

- Goal: assess and choose the actual true value for a data item when different sources provide conflicting information on it.
- Single-truth discovery:
  - Only one true value is allowed for data item.
- Multi-truth discovery:
  - Multiple true values are allowed for data items.
- Challenges:
  - Source trustworthiness and claim reliability.
  - Limited amount of labeled data source: semi-supervised/unsupervised.
  - Detect copy behaviors of false values.
  - Evaluation: based on data distribution and treat those deviated substantially from norm as wrong answers.



Identification of Misinformation

- Goal: detect false, inaccurate, or misleading information from data source.
- Potential solutions:
  - Explore mutual enhancement between source trustworthiness and claim credibility (similar idea from truth discovery).
  - Methods should distinguish misinformation with additional measures:
    - Building up information literacy and media literacy through education.
    - Using commonsense knowledge and open-mindedness.
    - Developing critical thinking.
  - Determine misinformation in the final stage based on:
    - Individual's mental model and worldview beliefs.
    - Consideration of repetition of misinformation.
    - Time-lag/relative coherency between misinformation and corrective information.

Data Mining Applications

- Data Mining for Sentiment and Opinion
- Truth Discovery and Misinformation Identification
- Information and Disease Propagation 
- Productivity and Team Science

Information and Disease Propagation

- Data mining research problems:
  - Prediction problems of propagation.
  - Optimization problems of propagation.
- Applications:
  - Social media: predict which piece of information is likely to go viral; detect the rumor source who started a misinformation campaign; neutralize the propagation of misinformation.
  - Computational epidemiology: determine critical network condition (e.g., epidemic threshold) under which an epidemic is likely to happen.

Information and Disease Propagation

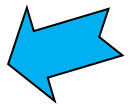
- Prediction problems of propagation:
  - Classification and regression: whether information being popular/popular scores.
  - Prediction around publication: before/after information is published.
  - Prediction at different granularities: individual/groups of information.
- Optimization problems of propagation (in a social network perspective):
  - Influence maximization: choose a set of initial nodes to maximize the number of infected nodes.
  - Source localization: identify source of a cascade over network with partial observations.
  - Activity shaping: steer user's activity.
  - Graph connectivity optimization: manipulate the graph topology to affect the information propagation results.

Information and Disease Propagation

- Models for information diffusion problems:
  - Feature based model:
    - Idea: utilize supervised models for information diffusion prediction.
    - Features used: temporal, local and global structure, user and item, etc.
  - Generative model:
    - Idea: take information diffusion as event sequences in the continuous temporal domain, and formulate as statistical generative approaches.
    - Methods: Poisson Process, Survival Analysis, Hawkes Process, Epidemic Model.
  - Deep learning based model:
    - Idea: no assumptions on the information diffusion process.
    - Methods: multi-modal data with various neural modules.
- Computational epidemiology models: SIS, SIR and SEIR and their variants.

Data Mining Applications

- Data Mining for Sentiment and Opinion
- Truth Discovery and Misinformation Identification
- Information and Disease Propagation
- Productivity and Team Science



Productivity and Team Science

- Team components: (1) team leader and members, (2) the environment in which team members collaborate, (3) the task on which the team works.
- Data mining research questions:
  - Team performance characterization: how to reveal key characteristic patterns that differentiate a high-performing team from a struggling one?
  - Team performance prediction: how to forecast the performance of the team before or soon after the start of the task?
  - Team performance optimization: how to further enhance the team performance by adjusting the team composition?
  - Team performance explanation: how to interpret the team performance prediction and optimization results intuitively?

Productivity and Team Science

- Team performance characterization:
  - Challenges:
    - Team members possess different type of skills and social connectivity.
    - The nature of tasks varies depends on the application scenarios.
    - Environments are large, noisy, incomplete and volatile to dynamics.
  - Treat as outcome of collective intelligence from both virtual teams and teams with face-to-face interactions.
- Team performance prediction:
  - Challenges :
    - Identify crucial features.
    - Model the correlation between features and the team performance.
    - Encode the dynamics of team evolution.
  - Predict the impact of contents that team produces at a finer granularity.

Productivity and Team Science

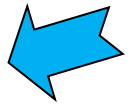
- Team performance optimization:
  - Optimize the team composition to maximize the performance.
  - Aspects considered by the replacement algorithms:
    - Skills of the team members.
    - Network connectivity (the environment).
  - Generalizations: team shrinkage, team expansion, team conflict resolution, real-time team optimization, etc.
- Team performance explanation:
  - Limited literature exists for intuitive explanations of team performance/algorithms.
  - Use influence function to identify key elements in teams for interpretation.
  - Interpret from multiple levels.

Summary: Data Mining Applications

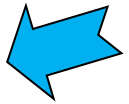
- Data Mining for Sentiment and Opinion
  - detect sentiment in texts with techniques from NLP, ML/DM, psychology and social science.
- Truth Discovery and Misinformation Identification
  - determine true values for data items.
  - determine false/inaccurate and misleading values for data items.
- Information and Disease Propagation
  - prediction problem and optimization problem.
- Productivity and Team Science
  - team performance characterization, prediction, optimizaion and explanation.

Chapter 12. Data Mining Trends and Research Frontiers

- Mining Rich Data Types
- Data Mining Applications
- Data Mining Methodologies and Systems
- Data Mining, People and Society
- Summary



Data Mining Methodologies and Systems

- Structuring Unstructured Data for Knowledge Mining: A Data-Driven Approach 
- Data Augmentation
- From Correlation to Causality
- Network as a Context
- Auto-ML: Methods and Systems

Taxonomy construction and refinement

- Taxonomy
  - Organizing important concepts into semantically rich structures
  - Playing an essential role at organizing massive unstructured text data into relatively organized structures
- Suitable taxonomy properties
  - multi-faceted
  - corpus- and application-dependent
- Taxonomy generation and expansion
  - Weakly or distantly supervised methods
    - Set expansion
    - Embedding-based hierarchical clustering
    - Adding new emerging terms or deleting old, obsolete ones

Weakly supervised text classification

- Text: massive, diverse and in multiple granularities
- Weakly supervised text classification
  - Idea #1:
    - Generate pseudo documents by Embedding methods
    - Train neural networks based on the pseudo documents and unlabeled text data
  - Idea #2:
    - Generate class-distinctive keywords or phrases by category-guided embedding
  - Idea #3:
    - Pre-trained language model to generate class-distinctive keywords

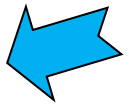
Fine-grained information extraction

- Fine-grained entity recognition
  - Extract the concrete role that an entity plays in a particular context
- Identification of the context of an entity
  - Generate entity mention candidates and relation phrases
  - Find the most appropriate fine-grained types for the entities to be examined
- Methods
  - Distant supervision based methods
  - Embedding based methods
  - Taxonomy-guided supervision and pre-trained language model

Knowledge graph/information network construction

- Knowledge graphs
  - A set of entities
  - A set of (possibly labeled) edges linking among entities
- Construction of knowledge graph
  - Use distant or weak supervision
  - Taxonomy-guided text classification
  - Problems of building knowledge graph
    - An entity can be associated with different attributes/value
    - Easy to cause confusion if different associations are merged into a single knowledge graph
  - Solutions
    - Construct local knowledge graphs
    - Use such local knowledge graphs under similar conditions

Data Mining Methodologies and Systems

- Structuring Unstructured Data for Knowledge Mining: A Data-Driven Approach
- Data Augmentation 
- From Correlation to Causality
- Network as a Context
- Auto-ML: Methods and Systems

Data Augmentation Details

- Why do we need data augmentation?
  - A high-performing data mining model often requires a great amount of labelled data samples for training
  - However, for many domains, we might be only provided by a limited number of labelled data
  - Training a data mining model with scarce training data, which could lead to the overfitting issue
- How
  - Basic Methods: Flipping, rotation, equalization and so on
  - Augmentation Policy Learning: Combine multiple basic augmentations
  - GANs-Based Methods: Training a GAN model to generate new data
  - Adversarial Training: Mixing the trickily-perturbed samples into the training samples

Data Mining Methodologies and Systems

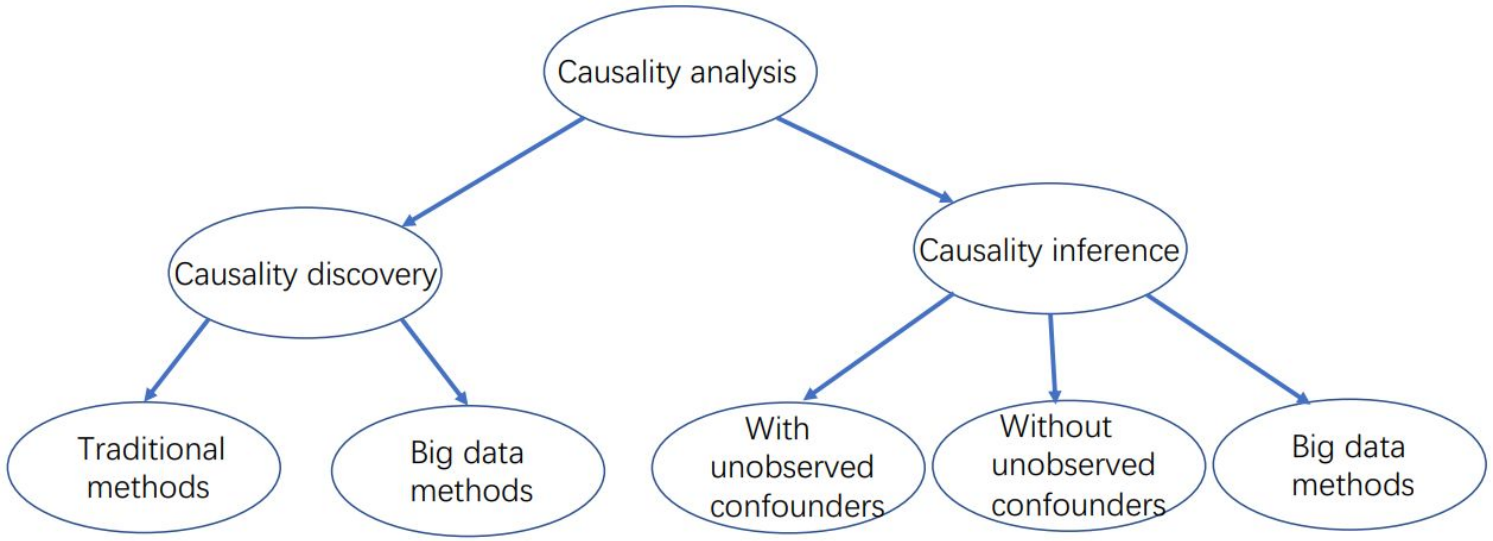
- Structuring Unstructured Data for Knowledge Mining: A Data-Driven Approach
- Data Augmentation
- From Correlation to Causality
- Network as a Context
- Auto-ML: Methods and Systems

From Correlation to Causality

- Correlation analysis
  - Study the statistical associations between different observed variables
- Causal analysis
  - Causality relation between observed variables
- Causality analysis tasks
  - Causality discovery
    - constraint-based algorithms, score-based algorithms and functional causal models
  - Causality inference
    - Regression adjustment, propensity score methods and covariate balancing methods

From Correlation to Causality

- Causality analysis benefits
  - Enhance the interpretability of black box deep learning models
  - Help avoid unfair results
  - Improve the robustness of models
- The taxonomy of causality analysis



Data Mining Methodologies and Systems

- Structuring Unstructured Data for Knowledge Mining: A Data-Driven Approach
- Data Augmentation
- From Correlation to Causality
- Network as a Context
- Auto-ML: Methods and Systems

Network as a Context

- Network of X
  - Each node represents an entity, a data set or a data mining model
- Examples: Network of time series
  - The key idea for mining network of time series data
    - Model each time series by either traditional signal processing approaches or deep neural networks
    - Leverage the contextual network to regularize different time series models

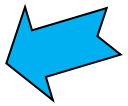
Network as a Context

- Network of X
  - Each node represents an entity, a data set or a data mining model
- Examples: Network of networks
  - Each X (i.e., node) itself of the contextual network represents another domain-specific network
  - The main advantages of the network of networks model
    - Explicitly encodes the hierarchical structure
    - The contextual network provides additional regularization

Network as a Context

- Network of X
  - Each node represents an entity, a data set or a data mining model
- Examples: Network of data mining models
  - Each X itself is a data mining model
  - Different performance prediction models 'borrow' data from each other
  - Mutually boost each other's prediction performance
- Future directions
  - The contextual network construction
  - Can be applied to more applications, e.g., knowledge graphs, heterogeneous networks
  - Integrate other data mining problems, e.g., adversarial learning, fairness, explainable learning.

Data Mining Methodologies and Systems

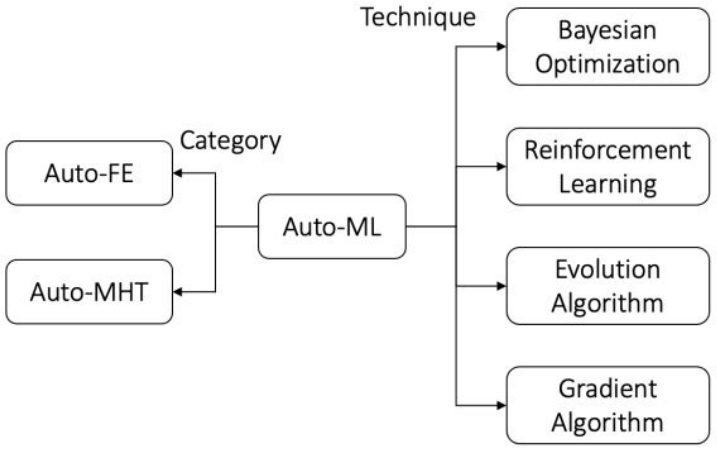
- Structuring Unstructured Data for Knowledge Mining: A Data-Driven Approach
- Data Augmentation
- From Correlation to Causality
- Network as a Context
- Auto-ML: Methods and Systems 

Auto-ML: Methods and Systems

- Why do we need Auto-ML
  - Different data mining and machine learning models have been widely used
  - Different task has different settings
  - Building and training appropriates mining models for each task could be time consuming
  - Domain experts might not be familiar with the mining models
- Auto-ML classification
  - Automated Feature Engineering (Auto-FE)
    - Automatically detects the most representative and informative features
  - Automated Model and Hyperparameter Tuning (Auto-MHT)
    - Automatically builds machine learning models and tunes the hyper-parameters

Auto-ML: Methods and Systems

- Key challenges for Auto-ML
  - The lack of authoritative benchmarks for Auto-FE and Auto-MHT
  - The efficiency of Auto-ML
  - How to incorporate human knowledge or experiences into the AutoML to find suitable sub-space of the entire search space
  - The interpretability of Auto-ML
  - The scope of applications
- Taxonomy of Auto-ML



Summary: Data Mining Methodologies and Systems

- Structuring Unstructured Data for Knowledge Mining: A Data-Driven Approach
  - Transform unstructured text-rich data into organized, relatively structured data
- Data Augmentation
  - We might be only provided by a limited number of labelled data
  - Data augmentation can prevent overfitting
- From Correlation to Causality
  - Focus on the causality relation between observed variables

Summary: Data Mining Methodologies and Systems

- Network as a Context
  - Provided a powerful context that links different types of data from different sources with different data mining algorithms
- Auto-ML: Methods and Systems
  - Allow non-experts to make use of machine learning models and techniques without requiring them to become experts in machine learning.

Chapter 12. Data Mining Trends and Research Frontiers

- Mining Rich Data Types
- Data Mining Applications
- Data Mining Methodologies and Systems
- Data Mining, People and Society
- Summary

Data Mining, People and Society

- Privacy-Preserving Data Mining 
- Human-Algorithm Interaction
- Mining beyond Maximizing Accuracy: Fairness, Interpretability, and Robustness
- Data Mining for Social Good
- Summary

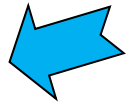
Privacy-Preserving Data Mining

- Privacy leakage between users and data owners
  - Example: re-identify users based on anonymous reviews on products.
- Anonymize data for privacy protection
  - k-anonymity (P. Samarat et al., 1998): generalize the data so that it is identical to at least k-1 other records on the identifying attribute.
- Differential privacy (A. Machanavajjhala et al. 2006)
  - Describe patterns of groups without sharing individual information.
  - Global sensitivity: for a function f and a pair of neighboring datasets D_1 and D_2, GS_f is the maximum difference between function value $f(D_1)$ and $f(D_2)$.

Privacy-Preserving Data Mining

- Privacy leakage between data owners
  - Example: multiple companies collaboratively conduct data mining but refuse to share user information.
  - Solution: federated learning and analytics.
- Federated learning
  - Run local computation over each data owners' devices and aggregate results without leaking detailed data from individual data owners.
  - Example: federated K-Means.
  - Horizontal and vertical federated learning
    - Horizontal: data owned by each data owners follows the same schema.
    - Vertical: different data owners have the data on only some attributes/objects.

Data Mining, People and Society

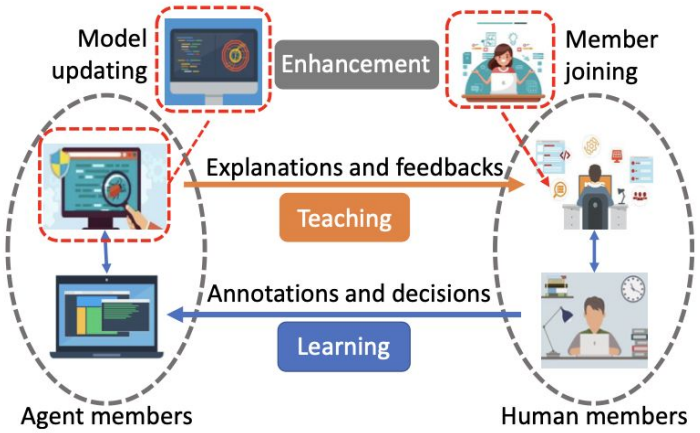
- Privacy-Preserving Data Mining
- Human-Algorithm Interaction 
- Mining beyond Maximizing Accuracy: Fairness, Interpretability, and Robustness
- Data Mining for Social Good

Human-Algorithm Interaction

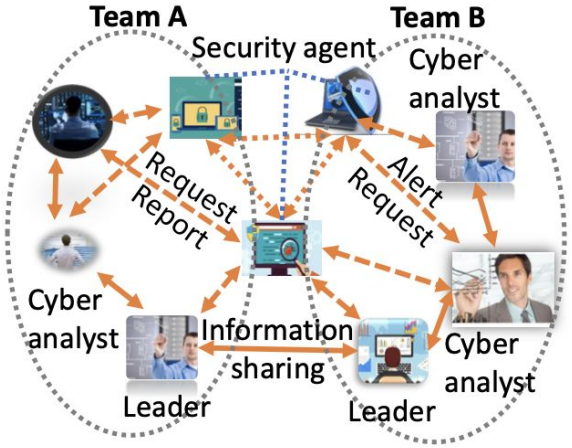
- Human-algorithm interaction
  - Goal: optimize the collaboration between human intelligence and algorithm, and consequently enhance the system performance and user experience.
- Crowdsourcing
  - Goal: harness the human intelligence to resolve difficult problems.
  - Applications: label generation, model evaluation, evaluating prediction interpretability, debug AI systems, etc.
- Human-in-the-loop
  - Goal: leveraging human intervention for improving algorithms.
  - Applications: high-quality label generation for supervised learning, enhancing model explainability with background knowledge, model evaluation, etc.

Human-Algorithm Interaction

- Machine-in-the-loop
  - Goal: improving humans' query strategy by algorithms.
  - Applications: creative writing and drawing, achieving comparable model performance which reducing the demand for labelled data
- Human-machine-teaming
  - Goal: fostering the effective teamwork between humans and algorithms.
  - Applications:



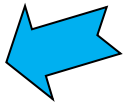
(a) HMT Overview



(b) HMT for Cyber Defense

Data Mining, People and Society

- Privacy-Preserving Data Mining
- Human-Algorithm Interaction
- Mining beyond Maximizing Accuracy: Fairness, Interpretability, and Robustness



- Data Mining for Social Good

Mining beyond Maximizing Accuracy: Fairness, Interpretability, and Robustness

- Algorithm fairness: definitions
  - Group fairness
    - Demographic parity: predictions are independent of sensitive attributes
    - Equalized odds: conditional probability of the same prediction results given the label should be equal for different sensitive attributes.
    - Predictive rate parity: conditional probability of label given the prediction results should be equal for different sensitive attributes.
  - Individual fairness: preserve individual similarities, i.e., similar individuals should be treated similarly.
  - Counterfactual fairness: correct predictions of a label variable that are unfairly altered by an individual's sensitive attribute.

Mining beyond Maximizing Accuracy: Fairness, Interpretability, and Robustness

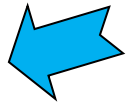
- Algorithm fairness: strategies
  - Pre-processing
    - Learn representations that removes the impact of sensitive attributes.
  - Optimization at training
    - Use regularizers or constraints to achieve trade-off between mitigating the bias and retaining the original mining accuracy.
  - Post-processing
    - Find a proper threshold for each group to achieve certain types of fairness

Mining beyond Maximizing Accuracy: Fairness, Interpretability, and Robustness

- Interpretability
  - Goal: explain results given by ‘black boxes’.
  - Visualize summary statistics for each feature as interpretation.
  - Explore self-interpretable components of the models (e.g., weights in the linear model, structure of decision trees)
  - Identifying key data points for interpretation.
- Robustness
  - Goal: ensuring consistency between test results and training results with unintentionally added noises or intentional adversarial attacks.
  - Reactive methods: detect adversarial examples after the model is built.
  - Proactive methods: make mining models more robust to attacks

Data Mining, People and Society

- Privacy-Preserving Data Mining
- Human-Algorithm Interaction
- Mining beyond Maximizing Accuracy: Fairness, Interpretability, and Robustness
- Data Mining for Social Good



Data Mining for Social Good

- Data mining applications

Applications	Key Data Mining Techniques
Education	Random Forests Logistic Regressions
Public Health	Long-Short Term Memory (LSTM) Auto-Encoder Deep Clustering
Combating Information Manipulation	SVM with RBF Kernel Function Markov Random Field (MRF) Graph Attention Network (GAT)
Social Care and Urban Planning	Gradient Boosting Decision Trees AdaBoost Transfer Learning
Public Safety	Generalized Linear Model (GLM) Naive Bayes (NB)
Transportation	Demand and Supplier Modeling Multi-agent Reinforcement Learning

- Challenges
  - Data scarcity: lack of large-scale data
  - Evaluation: standard evaluation metrics may be insufficient in real-world scenarios
  - Human-in-the-loop: insufficient use of domain knowledge from experts
  - Sustainable deployment

Summary: Data Mining, People and Society

- Privacy-Preserving Data Mining
  - Privacy between user and data owner: differential privacy
  - Privacy between data owners: federated learning
- Human-Algorithm Interaction
  - Crowdsourcing
  - Human-in-the-Loop
  - Machine-in-the-Loop
  - Human-Machine-teaming
- Beyond Accuracy
  - Fairness: group fairness, individual fairness, and counterfactual fairness
  - Interpretability
  - Robustness
- Data Mining for Social Good

References

- Pierangela Samarati and Latanya Sweeney. **Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression.** Technical report, SRI International, 1998.
- A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. **L-diversity: privacy beyond k-anonymity.** In Proc. 2006 Int. Conf. Data Engineering (ICDE'06), pages 24–24, Atlanta, GA, USA, Apr. 2006.
- Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. **Learning from crowds. Journal of Machine Learning Research**, 11(43):1297– 1322, 2010.

Chapter 12. Data Mining Trends and Research Frontiers

- Mining Rich Data Types
- Data Mining Applications
- Data Mining Methodologies and Systems
- Data Mining, People and Society
- Summary 

Summary: Data Mining Trends and Research Frontiers

- Mining Rich Data Types
  - Text data
  - Spatial-Temporal data
  - Graph and Networks
- Data Mining Methodologies and Systems
  - Structure Unstructured Data for Knowledge Mining
  - Data Augmentation
  - From Correlation to Causality
  - Network as a Context
  - Auto-ML

Summary: Data Mining Trends and Research Frontiers

- Data Mining Applications
  - Data Mining for Sentiment and Opinion
  - Truth Discovery and Misinformation Identification
  - Information and Disease Propagation
  - Productivity and Team Science
- Data Mining, People and Society
  - Privacy-Preserving Data Mining
  - Human-Algorithm Interaction
  - Mining beyond Maximizing Accuracy: Fairness, Interpretability, and Robustness
  - Data Mining for Social Good

References (Mining Rich Data Types)

- T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. **Distributed representations of words and phrases and their compositionality**. NIPS'13
- J. Pennington, R. Socher, and C. D. Manning. **GloVe: Global vectors for word representation.** EMNLP'14
- Y. Meng, J. Huang, G. Wang, C. Zhang, H. Zhuang, L. M. Kaplan, and J. Han. **Spherical text embedding.** NeurIPS'19
- M. Nickel and D. Kiela. **Poincaré embeddings for learning hierarchical representations.** NIPS'17
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. **BERT: Pre-training of deep bidirectional transformers for language understanding.** NAACL-HLT'19

References (Mining Rich Data Types)

- G. Atluri, A. Karpatne, and V. Kumar. **Spatiotemporal data mining: A survey of problems and methods.** ACM Comput. Surv., August 2018
- S. Shekhar, R. Vatsavai, and M. Celik. **Spatial and spatiotemporal data mining: Recent advances.** Next Generation of Data Mining, 2008.
- P. W. Holland, K. B. Laskey, and S. Leinhardt. **Stochastic blockmodels: First steps.** Social networks, 5, 1983.
- J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. **Kronecker graphs: an approach to modeling networks.** JMLR, 11, 2010.
- Y. Dong, N. V. Chawla, and A. Swami. **metapath2vec: Scalable representation learning for heterogeneous networks.** KDD'17

References (Data Mining Methodologies and Systems)

- J. G. Shanahan, Y. Qu, and J. Wiebe, editors. **Computing Attitude and Affect in Text: Theory and Applications**. Springer, 2006.
- B. Liu. **Sentiment Analysis: Mining Opinions, Sentiments, and Emotions**. Cambridge University Press, 2020.
- Y. Li, J. Gao, C.i Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. **A survey on truth discovery**. SIGKDD, 2015.
- L. Wu, F. Morstatter, K. M. Carley, and H. Liu. **Misinformation in social media: Definition, manipulation, and detection**. SIGKDD'19.
- Z. Chen, H. Tong, and L. Ying. **Inferring full diffusion history from partial timestamps**. TKDE'19.

References References (Data Mining Methodologies and Systems)

- Y. Yang and J. Pei. **Influence analysis in evolving networks: A survey**. TKDE'19.
- S. F Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R Varkonyi-Koczy, U. Reuter, T. Rabczuk, and P. M Atkinson. **Covid-19 outbreak prediction with machine learning**. Algorithms, 2020.
- J. Hackman and N. Katz. **Group Behavior and Performance,** 2010.
- L. Li and H. Tong. **Computational Approaches to the Network Science of Teams**. Cambridge University Press, 2020.

References (Data Mining Applications)

- Jingbo Shang, Jialu Liu, Meng Jiang, X. Ren, Clare R. Voss, and Jiawei Han. **Automated phrase mining from massive text corpora**. IEEE Transactions on Knowledge and Data Engineering, 2018.
- Chi Wang and Jiawei Han. **Mining Latent Entity Structures**. Morgan & Claypool, 2015.
- Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R. Voss, Heng Ji, and Jiawei Han. **ClusType: Effective entity recognition and typing by relation phrase-based clustering**. KDD, 2015.
- Jiaxin Huang, Yiqing Xie, Yu Meng, Jiaming Shen, Yunyi Zhang, and Jiawei Han. **Guiding corpus-based set expansion by auxiliary sets generation and co-expansion**. The Web Conf. (WWW'20), 2020.

References (Data Mining Applications)

- Ekin Dogus Cubuk, Barret Zoph, Dandelion Man´e, Vijay Vasudevan, and Quoc V. Le. **Autoaugment: Learning augmentation policies from data**. CoRR, 2018.
- Siyu Shao, Pu Wang, and Ruqiang Yan. **Generative adversarial networks for data augmentation in machine fault diagnosis**. Comput, 2019.
- Antreas Antoniou, Amos J. Storkey, and Harrison Edwards. **Data augmentation generative adversarial networks**. CoRR, 2017.
- Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. **Causation, prediction, and search**. MIT press, 2000.
- David Maxwell Chickering. **Optimal structure identification with greedy search**. Journal of machine learning research, 2002.

References (Data Mining Applications)

- Keisuke Hirano, Guido W Imbens, and Geert Ridder. **Efficient estimation of average treatment effects using the estimated propensity score**. Econometrica, 2003.
- Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. **Causal effect inference with deep latent variable models**. arXiv, 2017.
- Yongjie Cai, Hanghang Tong, Wei Fan, and Ping Ji. **Fast mining of a network of coevolving time series**. In Proc. 2015 SIAM Int. Conf. Data Mining (SDM'15), 2015.
- Yongjie Cai, Hanghang Tong, Wei Fan, Ping Ji, and Qing He. **Facets: Fast comprehensive mining of coevolving high-order time series**. KDD, 2015.

References (Data Mining, People and Society)

- Hannes Heikinheimo and Antti Ukkonen. **The crowd-median algorithm.** In Proc. 2013 Conf. Human Computation and Crowdsourcing (HCOMP'13), Palm Springs, CA, USA, Nov. 2013.
- Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. **Optimized pre-processing for discrimination prevention.** In Proc. 2017 Conf. Neural Information Processing Systems (NIP'17), pages 3995–4004, Long Beach, CA, USA, Dec. 2017.
- Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. **Adversarial attacks and defenses in deep learning. Engineering**, 6(3):346–360, 2020.
- Christoph Molnar. **Interpretable machine learning.** Lulu. com, 2020.

References (Data Mining, People and Society)

- Wojciech Samek, Thomas Wiegand, and Klaus-Robert Mu¨ller. **Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.** arXiv preprint arXiv:1708.08296, 2017.
- Jacob Abernethy, Alex Chojnacki, Arya Farahi, Eric Schwartz, and Jared Webb. **Activeremediation: The search for lead pipes in flint, michigan.** In Proc. 2018 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'18), pages 5–14, London, UK, Aug. 2018.
- Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. **Efficient large-scale fleet management via multi-agent deep reinforcement learning.** In Proc. 2018 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'18), pages 1774–1783, London, UK, Aug. 2018.