

Some Challenging Data Types for Data Mining

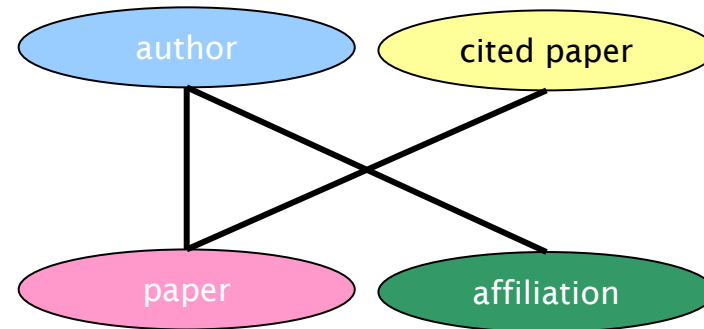
Qing Li*

Dept of Computer Science
City University of Hong Kong

* With input from Ms. Jing Chen and Mr. Yipu Wu

Data as Relations

- Motivation
 - Non-relational learning methods often require homogeneous data
 - Relations in heterogeneous data are often unexplored
 - Attributes for individual objects: author-affiliation
 - Intra-Relations between homogeneous data: paper-cited paper
 - Inter-Relations between heterogeneous data: author-paper
- Objective
 - Clustering relational data



- Challenges
 - Relational clustering: design similarity measures for different objects
 - Important relation selection: not all relations are helpful for clustering, how to select interesting relations

References

- Bo Long, Zhongfei Zhang, and Philip S. Yu: A Probabilistic Framework for Relational Clustering. KDD'07: 470-479
- Glen Jeh, and Jennifer Widom: Simrank: A Measure of Structural-Context Similarity. KDD'02: 538-543
- Adam Anthony, and Marie desJardins: Open Problems in Relational Data Clustering. ICML Workshop on Open Problems in Statistical Relational Learning, 2006

Computer-Generated Data

- Objective
 - Distinguishing whether a piece of digital data is a depiction of real-life occurrences or a synthetically generated one
- Characteristics
 - Data acquisition with sensors is fundamentally different from the generative algorithms deployed by computer-generated data
 - Data from a given sensor exhibit a unique stochastic characteristic due to the pattern noise introduced
- Methods
 - Verifying and evaluating the data statistics that are inherent to real-life sceneries



- Identifying signatures to detect traces of certain types of operations used in computer data generation processes
- Learning classifiers from features extracted from data samples and using them to differentiate between the two types of data
- Applications
 - Verifying the integrity and authenticity of multimedia data and data collected from sensors
 - Data forensics

References

- Sintayehu Dehnie, Taha H. Sencar, Nasir D. Memon: Digital Image Forensics for Identifying Computer Generated and Digital Camera Images. ICIP 2006: 2313-2316.
- AE Dirik, S Bayram, HT Sencar, N Memon: New Features to Identify Computer Generated Images. ICIP 2007: 120-127 .

Uncertain data

- An uncertain datum is one that represents multiple possible instances, each one corresponds to a single possible state.
- Causes
 - The semantic mappings between the data sources and the mediated schema may be inappropriate.
 - Transformation between keyword queries and a set of candidate structured queries is a second source of uncertainty.
 - Data are often extracted from unstructured sources using information extraction techniques, and these techniques are inappropriate.
- Methods
 - Modeling tuples and semantic mappings with probabilities associated with them.

• Table: a source instance S, a mediated schema T and a probabilistic schema mapping between S and T

pname	permanent-addr	Current-addr
Alice	Mountain View	Sunnyvale
Bob	Sunnyvale	Sunnyvale

name	Home-address

Possible Mapping	Current-addr
$M_1 = \{(pname=name), (permanent-addr=home-address)\}$	0.3
$M_2 = \{(pname=name), (Currentt-addr=home-address)\}$	0.7

- Querying with lineage and uncertainty together presents computational benefits than treating them separately.
- Using several kinds of probabilistic information to rank the multiple information sources.
- Applications
 - Integrating inconsistent data sources
 - Query answering in probabilistic databases

References

- Dong, X. and Halevy, A.Y. and Yu, C. :
Data integration with uncertainty. VLDB
2007: 687-698.
- Benjelloun, O. and Sarma, A.D. and
Halevy, A. and Widom, J. :
ULDBs: databases with uncertainty and
lineage. VLDB 2006: 953-964.

Incomplete Data

- An incomplete datum indicates that fields of some attributes are empty.
- causes
 - Incomplete Entry: some data are often populated by lazy individuals without any central curation.
 - Inaccurate Extraction: some data are being populated using automated information extraction techniques. As a result of the inherent imperfection of these extractions, many data may contain missing values.
 - User-defined data: Another type of incompleteness occurs in the context of some applications which allow users significant freedom to define and list their own attributes.
- Methods
 - Modifying the data directly by replacing null values with likely values.
 - Returning all the certain answers, and all the tuples with missing values on the constrained attribute(s).
 - First retrieving all the tuples with null values on the constrained attributes, predicting their missing values, and then deciding the set of relevant query answers to show to the users.
 - Rewritten queries according to a set of mined attribute correlation rules, these methods are able to retrieve possible answers without binding null values or modifying data.
- Applications
 - Querying over Incomplete Databases
 - Supervised learning from incomplete data

References

- Khatri, H. and Fan, J. and Chen, Y. and Kambhampati, S. : QPIAD: Query Processing over Incomplete Autonomous Databases . ICDE 2007: 1430-1432.
- Ghahramani, Z. and Jordan, M.I. and Cowan, J.D. and Tesauro, G. and Alspector, J. : Supervised learning from incomplete data via an EM approach. Advances in Neural Information Processing Systems. V.6: 120-127, 1994.

Untraditional Data

- Data Types
 - blogs, forums, emails, message boards
- Characteristics
 - Multiple authorship
 - Noisy
 - Semi-structured
- Challenges
 - Data quality: how to effectively preprocess data to reduce noise; how to extract the real content from semi-structured data automatically
 - Large scale: much of these data are not completed indexed by the search engine, some of them hidden. How to handle the large amount of data efficiently
 - Modeling: is the traditional vector space model appropriate?
- Tasks involved
 - Data cleaning
 - Data clustering
 - Social community mining

References

- Yipu Wu, Jing Chen, Qing Li: Extracting Loosely Structured Data Records Through Mining Strict Patterns. ICDE 2008:1322-1324
- Nishith Pathak, Sandeep Mane, Jaideep Srivastava: Who Thinks Who Knows Who? Socio-cognitive Analysis of Email Networks. ICDM 2006: 466-477
- Dou Shen, Jian-Tao Sun, Qiang Yang, Zheng Chen: Latent Friend Mining from Blog Data. ICDM 2006: 552-561
- Dou Shen, Qiang Yang, Jian-Tao Sun, Zheng Chen: Thread detection in dynamic text message streams. SIGIR 2006:35-42

Workflow-like Data

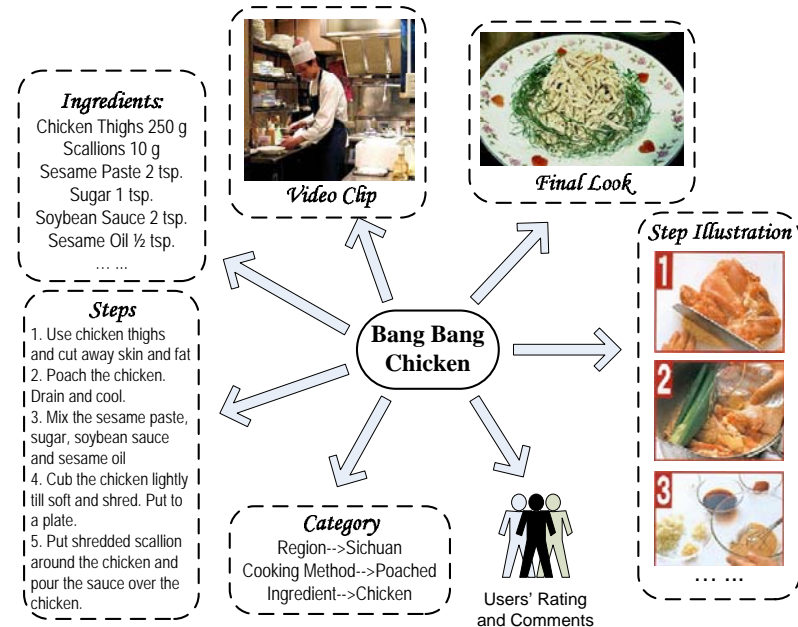
- Motivation

Unlike traditional data types, workflow-like data such as recipes have their distinct characteristics which make most (if not all) of the conventional data models unsuitable/inapplicable for such data:

- **Loosely structured:** they are usually semi-/loosely- structured data, with various levels of details and granularities. Some are more detailed/ wordy than others.
- **Behavior oriented:** they are not only data-intensive, but also behavior oriented, (eg, the main part of a recipe is about the procedure to follow in cooking a dish).
- **Constraints bound:** they are usually also bound by various constraints which are applicable to either individual actions or a sequence of actions.

- Objectives

- Mining the cooking skills (patterns) of different recipes from different regions



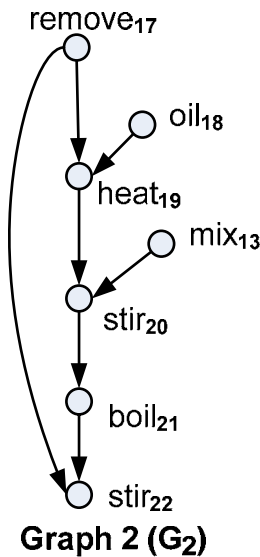
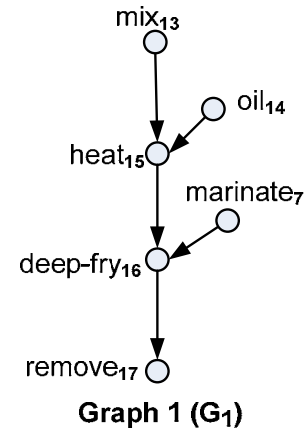
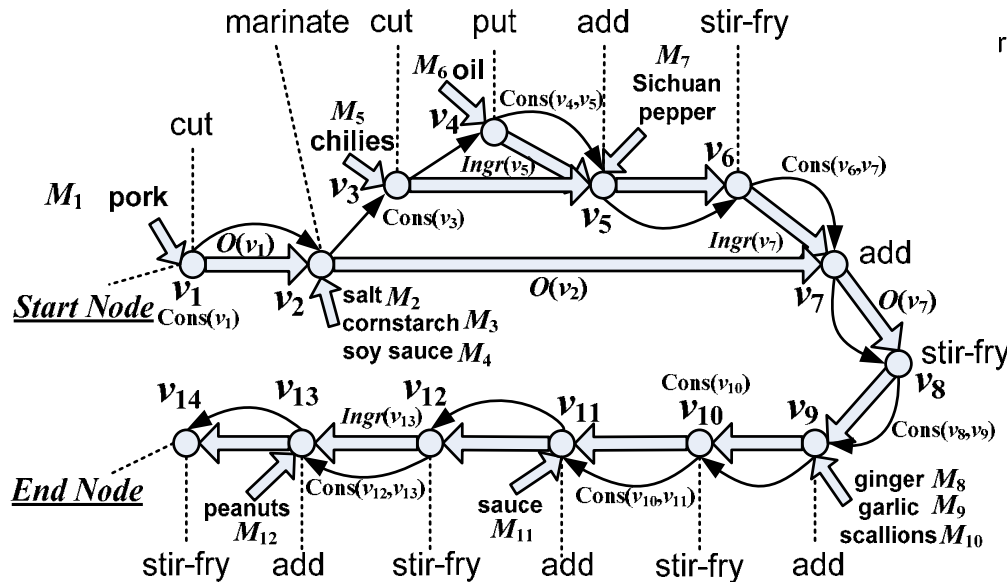
- Challenges

- data representation: directed cooking graph
- data clustering: need new similarity measures for different interests/emphasis

Workflow-like Data

- Data representation

- Data modeling

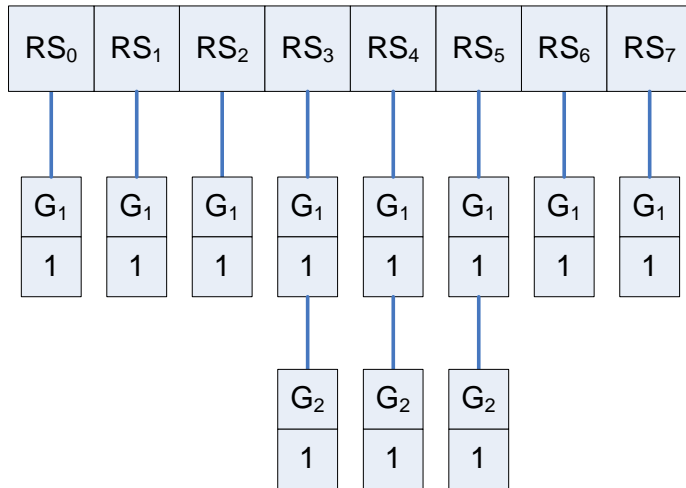


- RS0: <P, 'marinate', 'heat'>
- RS1: <P, 'mix', 'oil'>
- RS2: <F, null, 'marinate'>
- RS3: <F, null, 'mix'>
- RS4: <F, null, 'oil'>
- RS5: <F, 'oil', 'heat'>
- RS6: <F, 'heat', 'deep-fry'>
- RS7: <F, 'deep-fry', 'remove'>

- RS8: <P, 'mix', 'heat'>
- RS9: <P, 'remove', 'oil'>
- RS10: <P, 'remove', 'boil'>
- RS11: <S, 'heat', 'stir'>
- RS3: <F, null, 'mix'>
- RS12: <F, null, 'remove'>
- RS4: <F, null, 'oil'>
- RS5: <F, 'oil', 'heat'>
- RS13: <F, 'heat', 'stir'>
- RS14: <F, 'stir', 'boil'>
- RS15: <F, 'boil', 'stir'>

Workflow-like Data

- Cooking Graph Indexing



- Sample prototype

RecipeView screenshot: the target recipe (left window) with its most similar recipe (right window), and other top 10 similar recipes (shown on the right column)

- Data Clustering
 - new similarity measures

$$sim(SP_1, SP_2) =$$

$$\left[\left(\sum_{i=1}^m |E_{Si}| (\mu |E_{SAi}| + \gamma |E_{Sli}|) \cdot \log_2 \frac{N}{d_{Si}} \right) \cdot Per(SP_1, SP_2) \right]^{1/2}$$

RecipeView
Sharing recipe information

Welcome to RecipeView!
You can choose your favorite recipes and similar recipes as well!

Recipe Home | About Us | Contact | Help

Recipe Similarity Search: Fried Spareribs in Orange Juice

Quick Browsing: Guangdong cai

Sweet and Sour Pork

Fried Spareribs in Orange Juice

Cheng Du Young Chicken

Top10 Similar Recipes

- Boiled Chicken on Spinach
- Chicken with Chili
- Kung Pao Pork
- Crab with Scallion and Ginger
- Stir-fried Chicken Liver with Squid
- Dry Tofu Lamb Clay Pot
- Fried Chicken Sivers
- Ear Squid
- Fried Shrimp with Scrambled Eggs
- Braised chinese Cabbage with Black Mushroom

Want to see more? >>

References

- L. Wang, Q. Li, N. Li, G. Dong and Y. Yang: Substructure Similarity Measurement in Chinese Recipes. *Proc. WWW'08*: 979-988
- L. Wang, Q. Li: A Personalized Recipe Database System with User-Centered Adaptation and Tutoring Support. *SIGMOD2007 Workshop on Innovative Database Research (IDAR2007)*