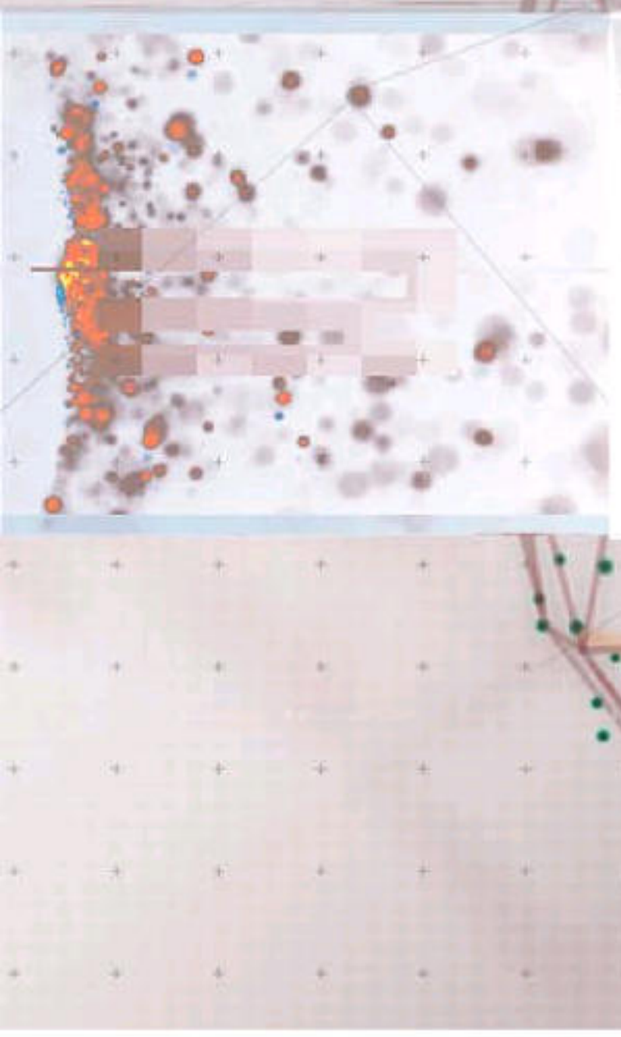
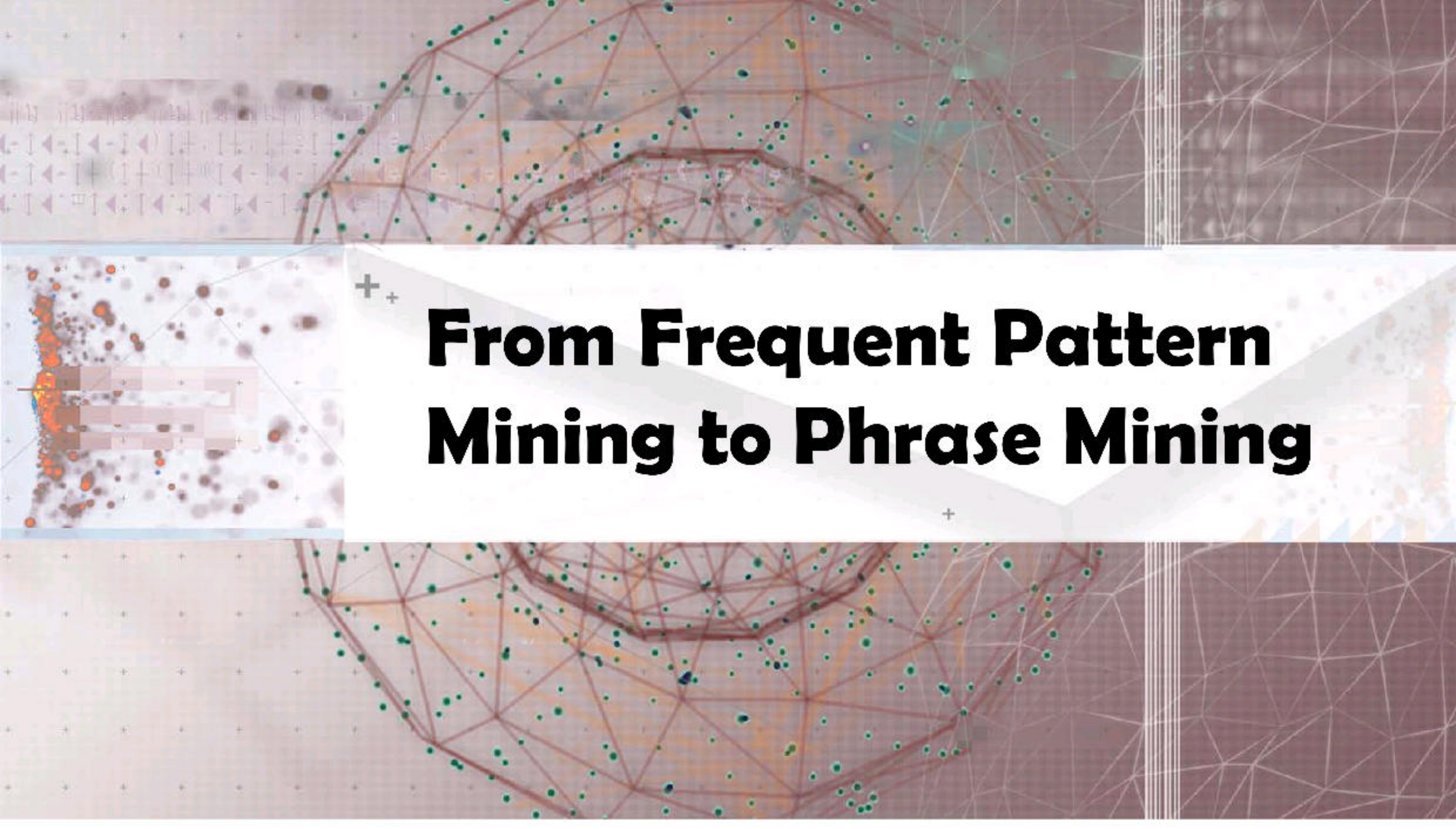


Pattern Mining Applications: Mining Quality Phrases from Text Data

Pattern Mining Applications: Mining Quality Phrases from Text Data

- From Frequent Pattern Mining to Phrase Mining
- Previous Phrase Mining Methods
- ToPMine: Phrase Mining without Training Data
- SegPhrase: Phrase Mining with Tiny Training Sets

Thanks to Ahmed El-Kishky@UIUC, Jialu Liu@UIUC, Jingbo Shang@UIUC, Xiang Ren@UIUC, Chi Wang@MSR and Marina Danilevsky@IBM for their contributions



From Frequent Pattern Mining to Phrase Mining

Why Phrase Mining?

- Unigrams vs. phrases
  - **Unigrams** (single words) are often *ambiguous*
    - Example: "United": United States? United Airline? United Parcel Service?
  - **Phrase**: A natural, meaningful, *unambiguous* semantic unit
    - Example: "United States" vs. "United Airline"
- Mining semantically meaningful phrases
  - Transform text data from *word granularity* to *phrase granularity*
  - Enhance the power and efficiency at manipulating unstructured data

From Frequent Pattern Mining to Phrase Mining

- General principle
  - Exploit information redundancy and data-driven criteria to determine phrase boundaries and salience
- Methodology: Exploring three ideas
  - Frequent pattern mining and colocation analysis
  - Phrasal segmentation
  - Quality phrase assessment
- Recent developments of phrase mining methods
  - ToPMine: Mining quality phrase without training (A. El-Kishky, et al., 2015)
  - SegPhrase: Mining quality phrase with tiny training sets (J. Liu, et al., 2015)