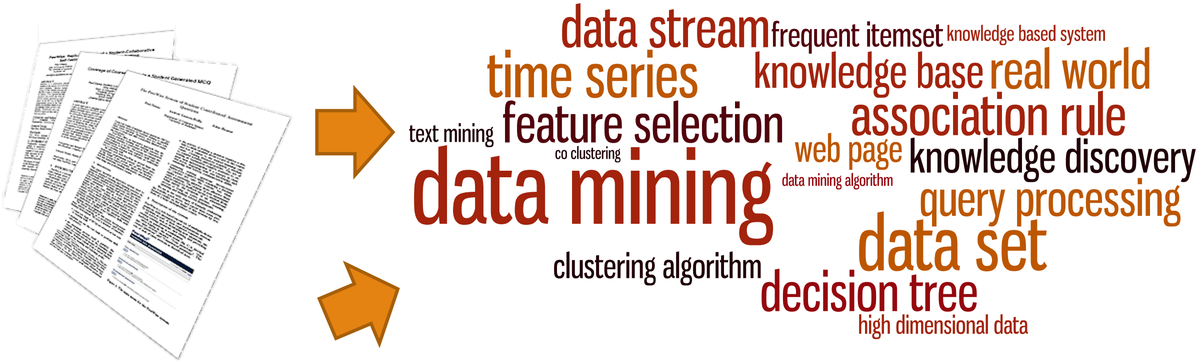
SegPhrase: Phrase Mining with Tiny Training Sets

SagPhrase: Phrase Mining with Tiny Training Sets

- A small set of training data may enhance the quality of phrase mining

J. Liu et al., Mining Quality Phrases from Massive Text Corpora. In *SIGMOD*’15

Raw Corpus **Quality Phrases** **Segmented Corpus**



+ A small set of labels by human or a general KB

Document 1
Citation recommendation is an interesting but challenging research problem in data mining area.

Document 2
In this study, we investigate the problem in the context of heterogeneous information networks using data mining technique.

Document 3
Principal Component Analysis is a linear dimensionality reduction technique commonly used in machine learning applications.

Input Raw Corpus → **Quality Phrases** ↔ **Segmented Corpus**

Phrase Mining **Phrasal Segmentation**

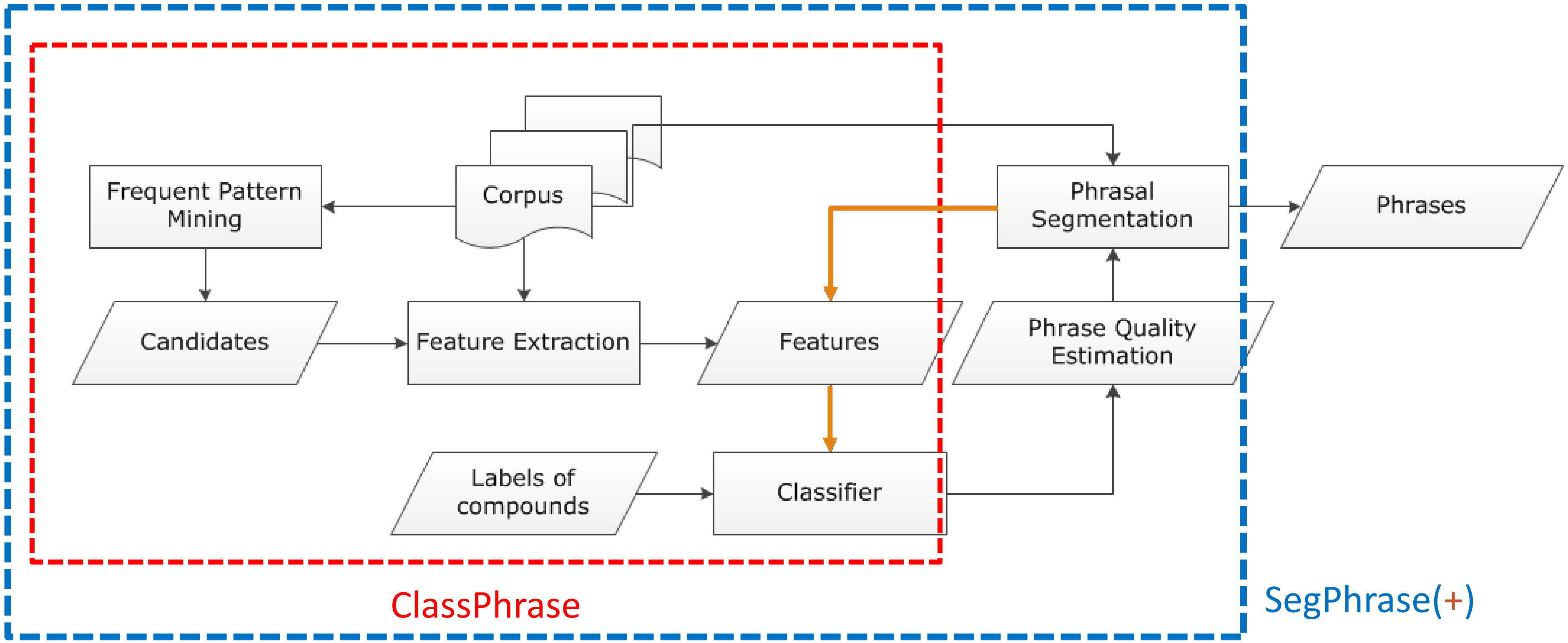
Integrating phrase mining with phrasal segmentation and classification

SegPhrase+: The Overall Framework

- ClassPhrase: Frequent pattern mining, feature extraction, classification
- SegPhrase: Phrasal segmentation and phrase quality estimation
- SegPhrase+: One more round to enhance mined phrase quality

SegPhrase (a classifier is used)

Small labeled dataset provided by experts
or
a distant supervised KB (e.g., Wikipedia / DBPedia)



SegPhrase: Pattern Mining and Feature Extraction

- **Pattern Mining for Candidate Set**
  - Build a candidate phrases set by frequent pattern mining
    - Mining frequent *k*-grams (*k* is typically small, e.g., 6 in the experiments)
    - **Popularity** measured by *raw* frequent words and phrases mined from the corpus
- **Feature Extraction: Concordance**
  - Partition a phrase into two parts to check whether the co-occurrence is significantly higher than pure random
- **Feature Extraction: Informativeness**
  - Quality phrases typically start and end with a non-stopword
    - “machine learning is” vs. “machine learning”
  - Use average IDF over words in the phrase to measure the semantics
  - Usually, the probabilities of a quality phrase in quotes, brackets, or connected by hyphen should be higher (punctuations information)
      - e.g., “state-of-the-art”

SegPhrase: Classification Using Tiny Training Sets

- Use tiny training sets (300 labels for 1GB corpus; can also use phrases extracted from KBs)
  - Label: indicating whether a phrase is a high quality one
    - E.g., “support vector machine”: 1; “the experiment shows”: 0
- Classification: Construct models to distinguish quality phrases from poor ones
  - Use *Random Forest* algorithm to bootstrap different datasets with limited labels
- Phrasal segmentation can tell which phrase is more appropriate
  - Ex: “A standard [feature vector] [machine learning] setup is used to describe .....”

    Not counted towards the rectified frequency
  - Partition a sequence of words by maximizing the likelihood
  - Consider length penalty and filter out phrases with low rectified frequency
- Process: Classification → Phrasal segmentation // SegPhrase

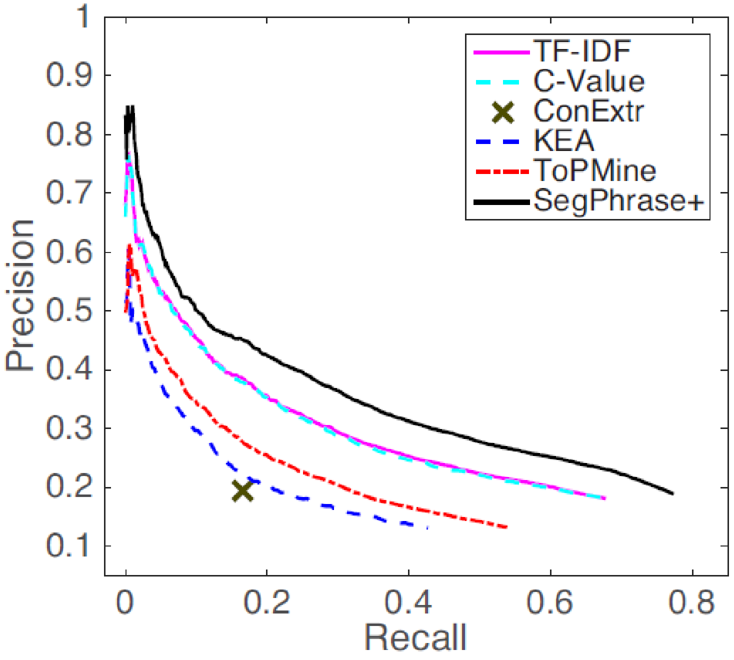
  → Classification → Phrasal segmentation // SegPhrase+

Performance: Precision Recall Curves on DBLP

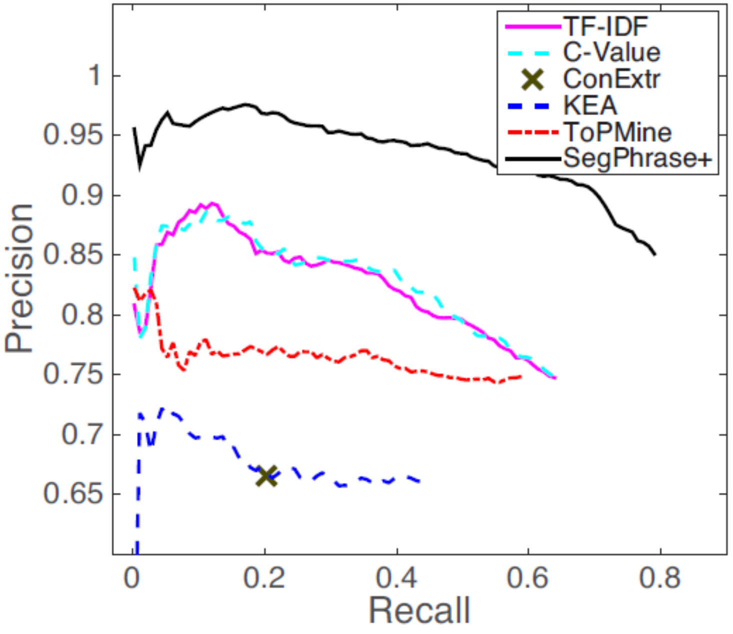
- Datasets:
- Evaluation
  - Wiki Phrases (based on internal links, ~7K high quality phrases)
  - Sampled 500*7 Wiki-uncovered phrases: Results evaluated by 3 reviewers
- Compared with other phrase-mining methods
  - TF-IDF, C-Value, ConExtr, KEA, and ToPMine
- Also, Segphrase+ is efficient, linearly scalable

Dataset	#docs	#words	#labels
DBLP	2.77M	91.6M	300
Yelp	4.75M	145.1M	300

Use only 300 human labeled phrases for training



Precision-Recall Curves on DBLP Data (Wiki Phrases)



Precision-Recall Curves on DBLP Data (Non Wiki-phrases)

Experimental Results: Interesting Phrases Generated (From Titles & Abstracts of SIGKDD)

Query	SIGKDD	
Method	SegPhrase+	Chunking (TF-IDF & C-Value)
1	data mining	data mining
2	data set	association rule
3	association rule	knowledge discovery
4	knowledge discovery	frequent itemset
5	**time series**	decision tree
…	…	…
51	association rule mining	search space
52	rule set	domain knowledge
53	concept drift	**important problem**
54	knowledge acquisition	concurrency control
55	**gene expression data**	conceptual graph
…	…	…
201	web content	optimal solution
202	**frequent subgraph**	semantic relationship
203	intrusion detection	**effective way**
204	**categorical attribute**	space complexity
205	user preference	**small set**
…	…	…

Only in SegPhrase+

Only in Chunking

Mining Quality Phrases in Multiple Languages

- ❑ Both ToPMine and SegPhrase+ are extensible to mining quality phrases in multiple languages
- ❑ SegPhrase+ on Chinese (From Chinese Wikipedia)
- ❑ ToPMine on Arabic (From Quran (Fus7a Arabic)(no preprocessing)
  - ❑ Experimental results of Arabic phrases:

كفروا→ Those who disbelieve

بسم الله الرحمن الرحيم→ In the name of God the Gracious and Merciful

Rank	Phrase	In English
…	…	…
62	首席_执行官	CEO
63	中间_偏右	Middle-right
…	…	…
84	百度_百科	Baidu Pedia
85	热带_气旋	Tropical cyclone
86	中国科学院_院士	Fellow of Chinese Academy of Sciences
…	…	…
1001	十大_中文_金曲	Top-10 Chinese Songs
1002	全球_资讯网	Global News Website
1003	天一阁_藏_明代_科举_录_选刊	A Chinese book name
…	…	…
9934	国家_戏剧_院	National Theater
9935	谢谢_你	Thank you
…	…	…

Summary

Summary: Pattern Mining Applications: Mining Quality Phrases from Text Data

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

Recommended Readings

- S. Bergsma, E. Pitler, D. Lin, Creating robust supervised classifiers via web-scale n-gram data, ACL'2010
- D. M. Blei and J. D. Lafferty. Visualizing Topics with Multi-Word Expressions. arXiv:0907.1013, 2009
- D.M. Blei, A. Y. Ng, M. I. Jordan, J. D. Lafferty, Latent Dirichlet allocation. JMLR 2003
- K. Church, W. Gale, P. Hanks, D. Hindle. Using Statistics in Lexical Analysis. In U. Zernik (ed.), Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Lawrence Erlbaum, 1991
- M. Danilevsky, C. Wang, N. Desai, X. Ren, J. Guo, J. Han. Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents. SDM'14
- A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable Topical Phrase Mining from Text Corpora. VLDB'15
- R. V. Lindsey, W. P. Headden, III, M. J. Stipicevic. A Phrase-Discovering Topic Model Using Hierarchical Pitman-Yor Processes. EMNLP-CoNLL'12.
- J. Liu, J. Shang, C. Wang, X. Ren, J. Han, Mining Quality Phrases from Massive Text Corpora. SIGMOD'15
- A. Parameswaran, H. Garcia-Molina, and A. Rajaraman. Towards the Web of Concepts: Extracting Concepts from Large Datasets. VLDB'10
- X. Wang, A. McCallum, X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. ICDM'07