Representing Documents via Latent Keyphrase Inference

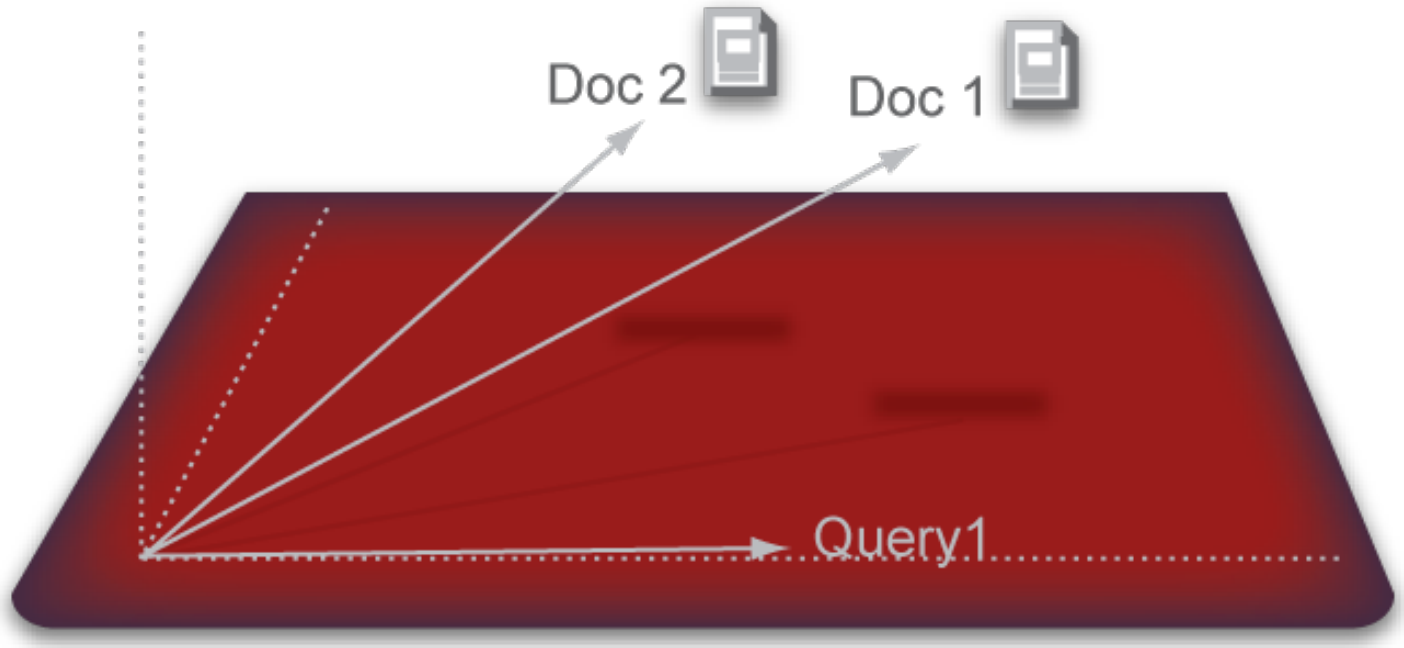
Jialu Liu† Xiang Ren† Jingbo Shang†
Taylor Cassidy‡ Clare R. Voss‡ Jiawei Han†

†Department of Computer Science, University of Illinois at Urbana-Champaign
‡Computational & Information Sciences Directorate, Army Research Laboratory

April. 15th, 2016

Document Representation in Vector Space

Critical for document retrieval, categorization



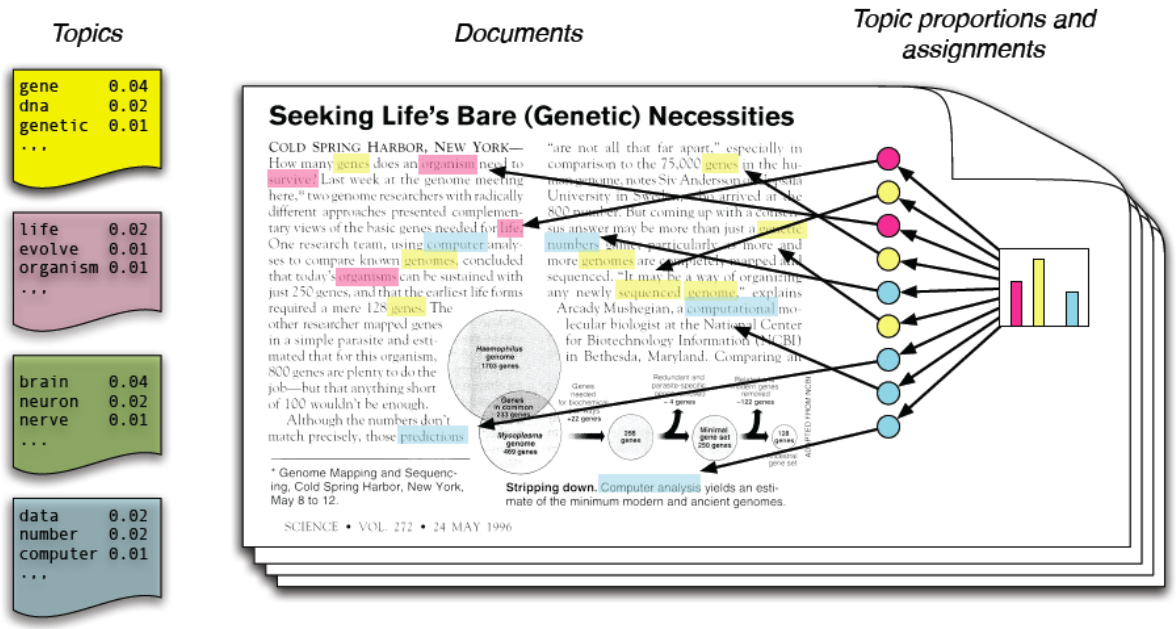
Traditional Methods

- Bag-of-Words or Phrases

	doc1	doc2	doc3
I	1	0	0
like	1	0	0
football	1	1	0
John	0	1	1
likes	0	1	1
football	1	1	0
basketball	0	0	1

- Cons: Sparse on short texts

- Topic models [LDA]

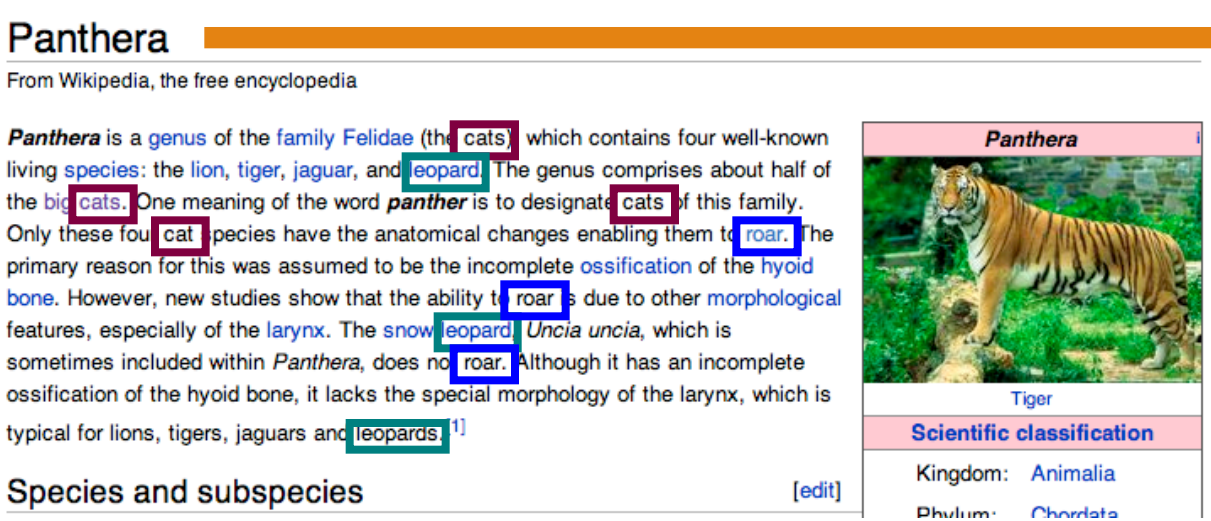


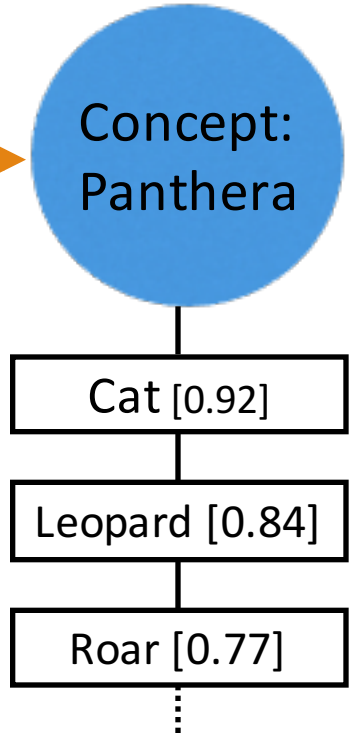
Each **topic** is a distribution over words, each **document** is a mixture of corpus-wide topics

- Cons: Difficult for human to infer topic semantics

- Concept-based models [ESA]

Every Wikipedia article represents a **concept**

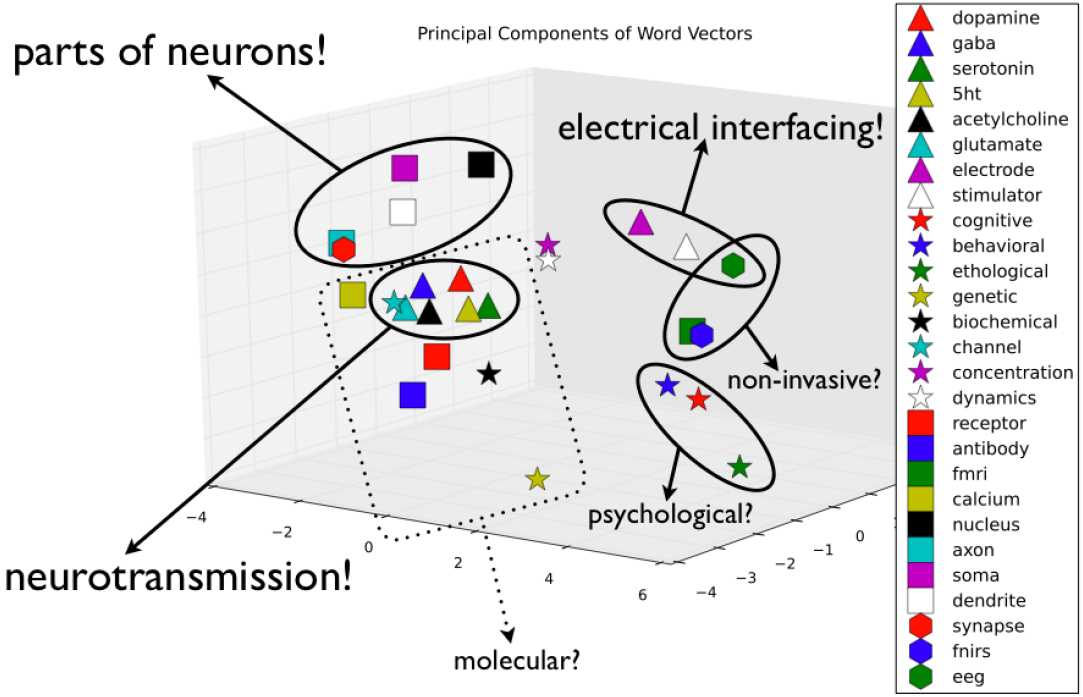




Article words are associated with the concept (TF.IDF), which help infer concepts from document

- Cons: Low coverage of concepts in human-curated knowledge base

- Word/Document embedding models [word2vec paragraph2vec]

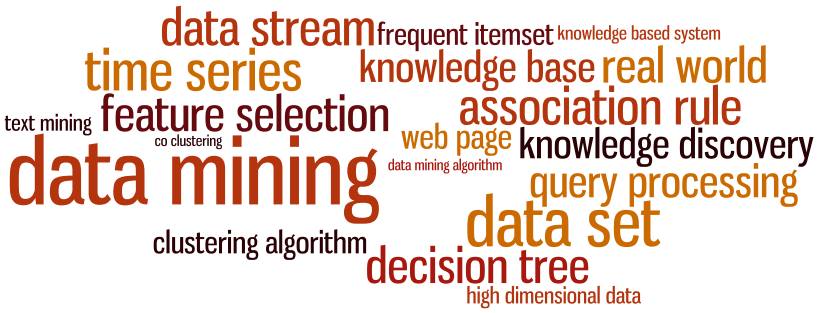


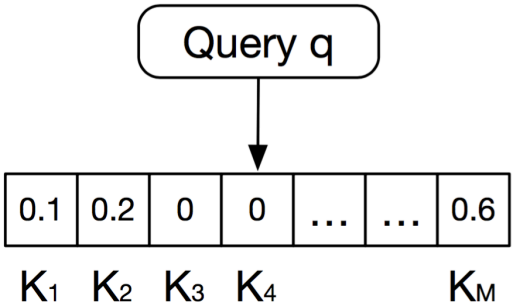
- Cons: Difficult to explain what each dimension means

Document Representation Using Keyphrases

Corpus

Domain Keyphrases <K_1, K_2, …, K_M>





- Use **domain keyphrases** as the entries in the vector and
- Identify **document keyphrases** (subset of domain keyphrases) by evaluating relatedness between (doc, domain keyphrase)
- Unsupervised model

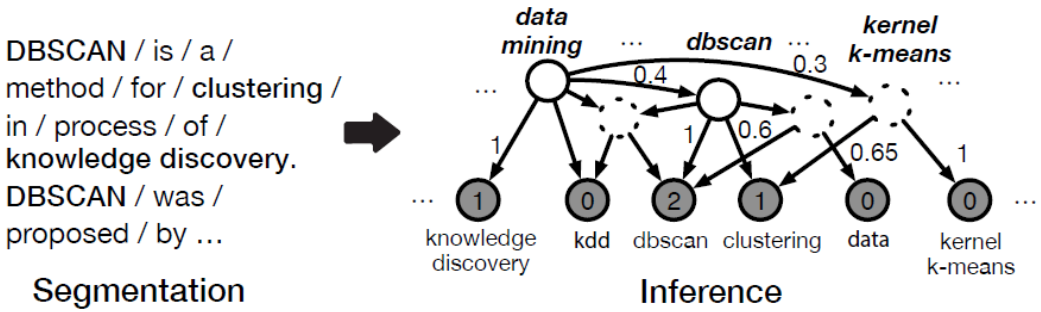
Challenges

- Where to get domain keyphrases from a given corpus?
  - **Mining Quality Phrases from Massive Text Corpora [SIGMOD15]**
- How to identify document keyphrases?
  - Can be latent mentions (short text)
  - Relatedness scores

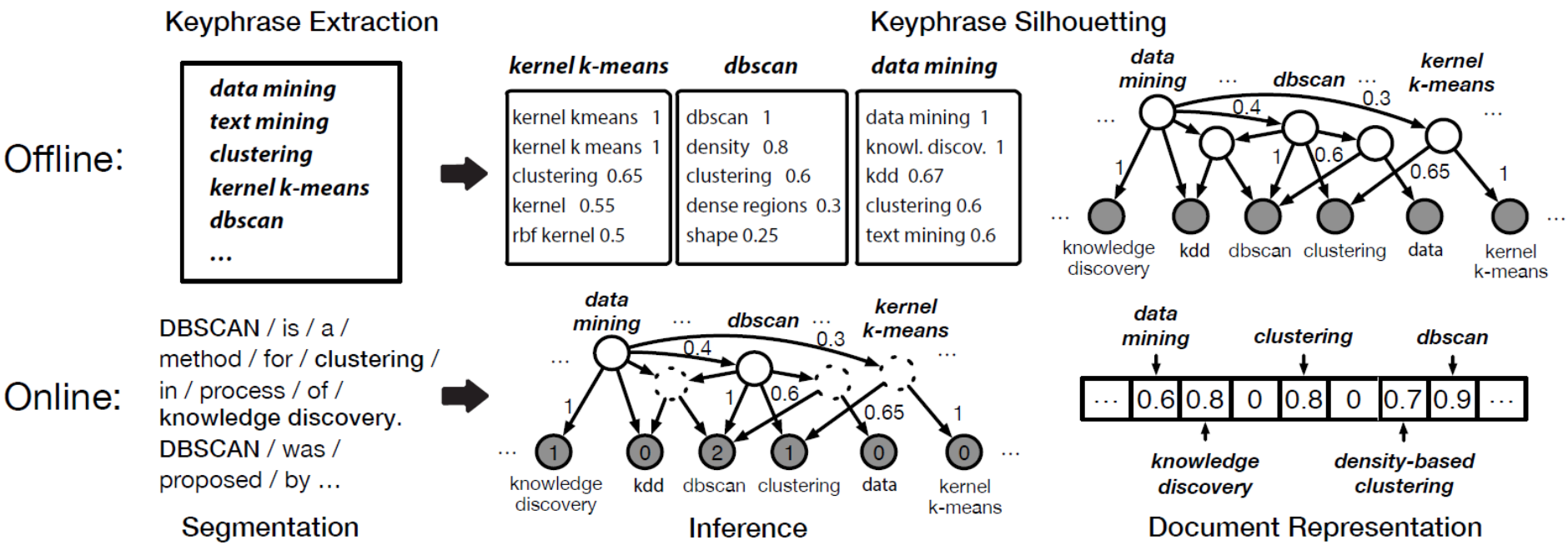
How to identify document keyphrases?

- Powered by Bayesian Inference on "Domain Keyphrase Silhouette"
  - Domain Keyphrase Silhouette: Topic centered on domain keyphrase
    - "Reverse" topic models
    - Learned from corpus

kernel k-means	*dbscan*	*data mining*
kernel kmeans 1	dbscan 1	data mining 1
kernel k means 1	density 0.8	knowl. discov. 1
clustering 0.65	clustering 0.6	kdd 0.67
kernel 0.55	dense regions 0.3	clustering 0.6
rbf kernel 0.5	shape 0.25	text mining 0.6



Framework for Latent Keyphrase Inference (LAKI)



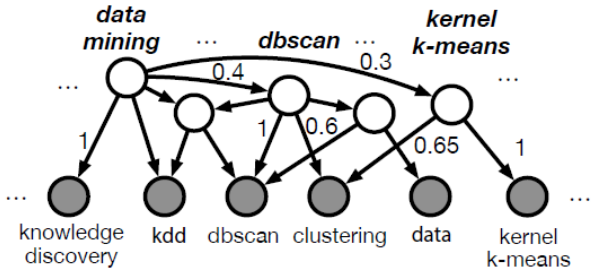
Offline:

Keyphrase Extraction

data mining
text mining
clustering
kernel k-means
dbscan
...

Keyphrase Silhouetting

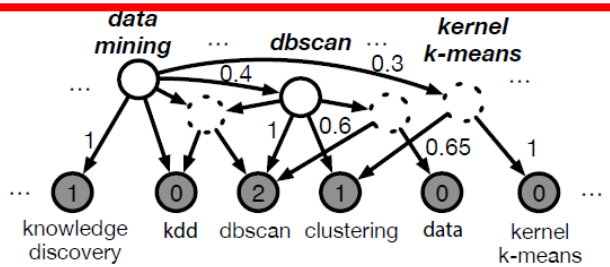
kernel k-means	***dbscan***	***data mining***
kernel kmeans 1	dbscan 1	data mining 1
kernel k means 1	density 0.8	knowl. discov. 1
clustering 0.65	clustering 0.6	kdd 0.67
kernel 0.55	dense regions 0.3	clustering 0.6
rbf kernel 0.5	shape 0.25	text mining 0.6



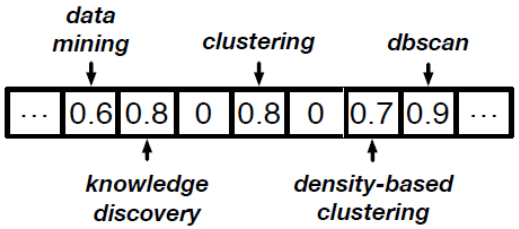
Online:

DBSCAN / is / a / method / for / clustering / in / process / of / knowledge discovery. DBSCAN / was / proposed / by …

Segmentation



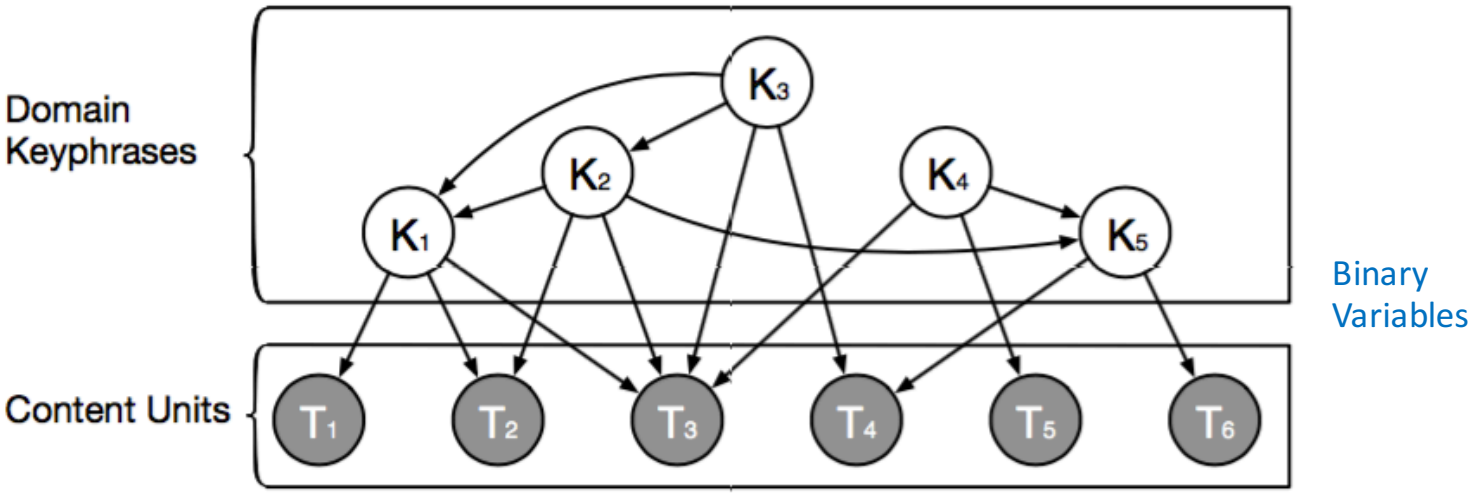
Inference



Document Representation

Domain Keyphrase Silhouette

- Learning Hierarchical Bayesian Network (DAG)

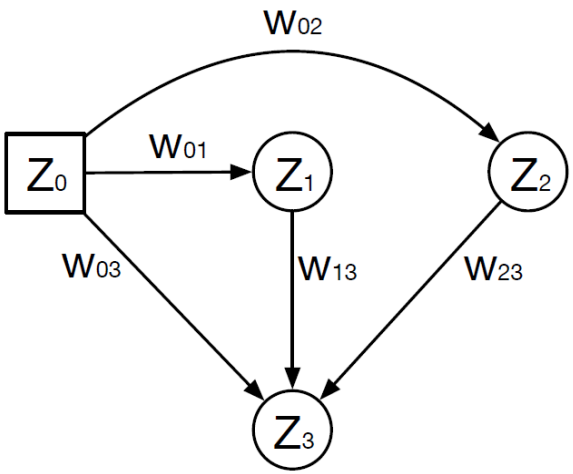


Task 1: Model Learning: learning link weights

Task 2: Structure Learning: learning network structure

Task 1: Model Learning given Structure

- Use Z to represent K (domain keyphrases) and T (content units)
- Noisy-OR
  - A parent node is easier to activate its children when the link weight is larger
  - A child node is influenced by all its parents



Toy example

$$p(Z_j = 1 | Pa(Z_j)) = 1 - exp\left(-W_{0j} - \sum_i W_{ij} \mathbb{1}_{Pa_j^i}\right)$$

Noise / Prior

Aggregated over all other links connected with Z_j

Maximum Likelihood Estimation

- Training data: Documents
- Expectation-step:
  - For each document, collect sufficient statistics
    - Link firing (Parent, child both being activated) probability
    - Node activation probability
- Maximization-step:
  - Update link weight

Partially observed document keyphrases

$$L(D) = \sum_{d=1}^{N} \log \sum_{k \in \Omega^{(d)}} p(K = k, T = t^{(d)})$$

Fully observed content units

Task 2: Structure Learning

- Domain keyphrases are connected to content units
  - Help infer document keyphrases from content units
- Domain keyphrases are interconnected
  - Help infer document keyphrases from other keyphrases

A Heuristic Approach

- Data-Driven, DAG, similar to ontology
- Heuristic:
  - Two nodes are connected only
    - Closely Related: word2vec
    - Co-occur frequently
  - Links are always point to less frequent nodes
- Work well in practice

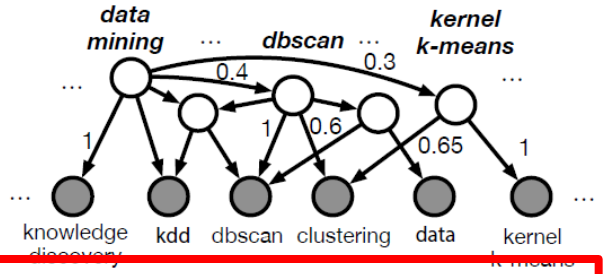
Keyphrase Extraction

Keyphrase Silhouetting

Offline:

data mining
text mining
clustering
kernel k-means
dbscan
...

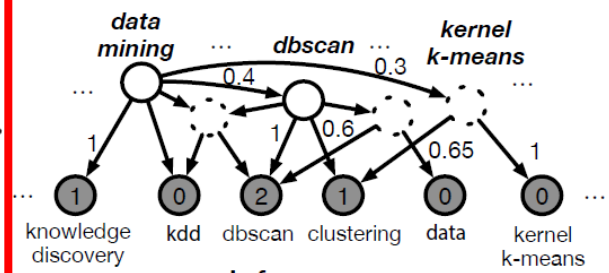
kernel k-means	***dbscan***	***data mining***
kernel kmeans 1	dbscan 1	data mining 1
kernel k means 1	density 0.8	knowl. discov. 1
clustering 0.65	clustering 0.6	kdd 0.67
kernel 0.55	dense regions 0.3	clustering 0.6
rbf kernel 0.5	shape 0.25	text mining 0.6



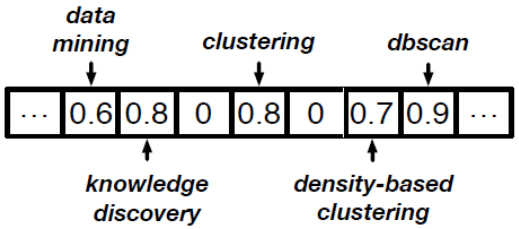
Online:

DBSCAN / is / a / method / for / clustering / in / process / of / knowledge discovery. DBSCAN / was / proposed / by …

Segmentation



Inference



Document Representation

Inference

- Exact inference is slow!
  - NP hard to compute posterior probability for Noisy-Or networks
  - Approximate inference instead
    - Pruning irrelevant nodes using an efficient scoring function
    - Gibbs sampling

Experiments

- Two text-related tasks to evaluate document representation quality
  - Phrase relatedness
  - Document classification

- Two datasets

Dataset	#Docs	#Words	Content type	#Domain Keyphr. (in Wiki)
Academia	0.43M	28M	title & abstract	33K (6,367)
Yelp	0.47M	98M	review	25K (4,996)

Methods

Method	Semantic Space	Input Source
ESA	KB concepts	KB
KBLink	KB concepts	KB
BoW	Words	-
ESA-C	Documents	Corpus
LSA	Topics	Corpus
LDA	Topics	Corpus
Word2Vec	-	Corpus
EKM	Explicit Keyphrases	Corpus
LAKI	Latent Keyphrases	Corpus

- **ESA** (Explicit Semantic Analysis)
- **KBLink** uses link structure in Wikipedia
- **BoW** (bag-of-words)
- **ESA-C** extends ESA by replacing Wiki with domain corpus
- **LSA** (Latent Semantic Analysis)
- **LDA** (Latent Dirichlet Allocation)
- **Word2Vec** is a neural network computing word embeddings
- **EKM** uses explicit keyphrase detection

Phrase Relatedness Correlation

Method	Academia (w/ phrase)	Yelp (w/ phrase)
ESA	0.4320 (-)	0.4567 (-)
KBLink	0.1878 (-)	0.4179 (-)
ESA-C	0.4905 (0.5243)	0.4655 (0.5029)
LSA	0.5877 (0.6383)	0.6700 (0.7229)
LDA	0.3610 (0.5391)	0.3928 (0.5405)
Word2Vec	0.6674 (0.7281)	0.7143 (0.7419)
LAKI	**0.7504**	**0.7609**

Document Classification

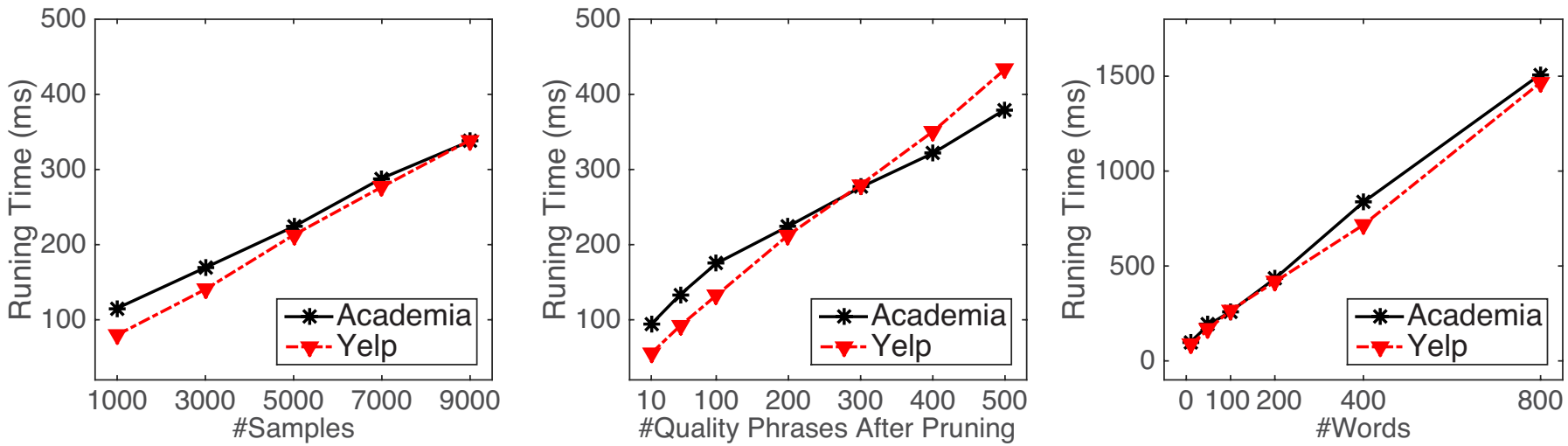
Method	Academia (w/ phrase)	Yelp (w/ phrase)
ESA	37.61 (-)	46.56 (-)
KBLink	36.37 (-)	35.94 (-)
BoW	48.05 (45.60)	51.26 (45.97)
ESA-C	39.75 (42.20)	49.13 (54.51)
LSA	72.50 (79.22)	66.55 (78.57)
LDA	77.27 (80.52)	75.55 (82.65)
EKM	45.46	40.57
LAKI	**84.42**	**90.58**

Case Study

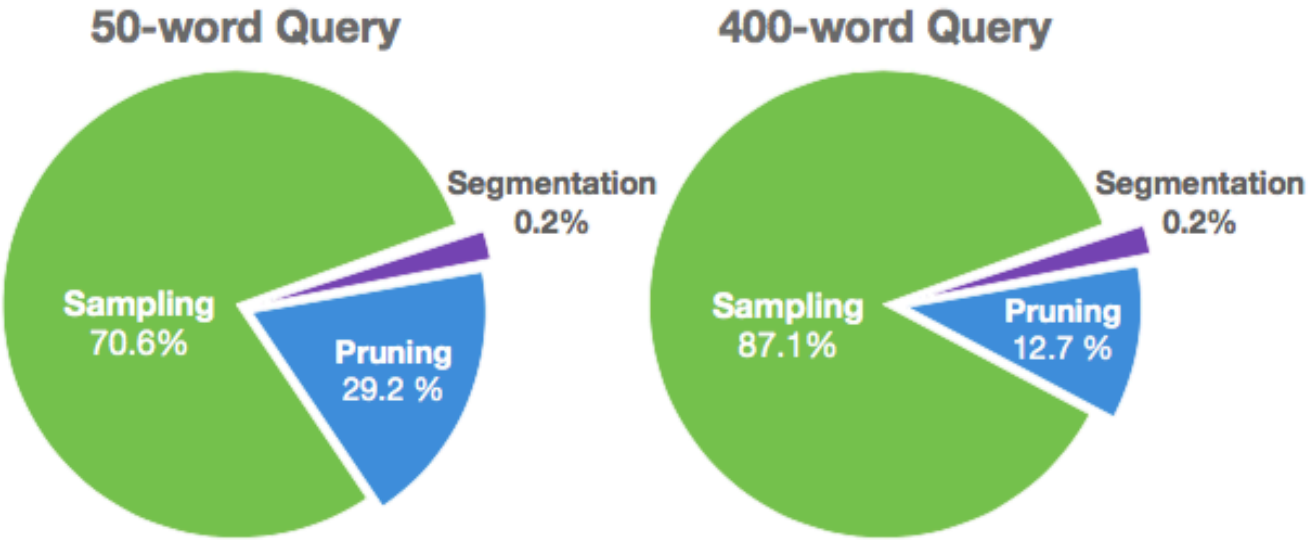
Query	*LDA*	*BOA*
Keyphrases	linear discriminant analysis, latent dirichlet allocation, topic models, topic modeling, face recognition, sda, latent dirichlet, generative model, topic, subspace models, ...	boa steakhouse, bank of america, stripsteak, agnolotti, credit card, santa monica, restaurants, wells fargo, steakhouse, prime rib, bank, vegas, las vegas, cash, cut, dinner, bank, money, ...
Query	*LDA topic*	*BOA steak*
Keyphrases	latent dirichlet allocation, topic, topic models, topic modeling, probabilistic topic models, latent topics, topic discovery, generative model, mixture, text mining, topic distribution, plsi, ...	steak, stripsteak, boa steakhouse, steakhouse, ribeye, craftsteak, santa monica, medium rare, prime, vegas, entrees, potatoes, french fries, filet mignon, mashed potatoes, texas roadhouse, ...
Query	*SVM*	*deep dish pizza*
Keyphrases	support vector machines, svm classifier, multi class, training set, margin, knn, classification problems, kernel function, multi class svm, multi class support vector machine, support vector, ...	deep dish pizza, chicago, deep dish, amore taste of chicago, amore, pizza, oregano, chicago style, chicago style deep dish pizza, thin crust, windy city, slice, pan, oven, pepperoni, hot dog, ...

	Academia	Yelp
Query	*Mining Frequent Patterns without Candidate Generation*	*I am a huge fan of the All You Can Eat Chinese food buffet.*
Keyphrases	mining frequent patterns, candidate generation, frequent pattern mining, candidate, prune, fp growth, frequent pattern tree, apriori, subtrees, frequent patterns, candidate sets, ...	all you can eat, chinese food, buffet, chinese buffet, dim sum, orange chicken, chinese restaurant, asian food, asian buffet, crab legs, lunch buffet, fan, salad bar, all you can drink, ...
Query	*Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through means such as statistical pattern learning.*	*It's the perfect steakhouse for both meat and fish lovers. My table guest was completely delirious about his Kobe Beef and my lobster was perfectly cooked. Good wine list, they have a lovely Sancerre! Professional staff, quick and smooth.*
Keyphrases	text analytics, text mining, patterns, text, textual data, topic, information, text documents, information extraction, machine learning, data mining, knowledge discovery, ...	kobe beef, fish lovers, steakhouse, sancerre, wine list, guests, perfectly cooked, lobster, staff, meat, fillet, fish, lover, seafood, ribeye, filet, sea bass, risotto, starter, scallops, steak, beef, ...

Time Complexity



Breakdown of Processing Time



Conclusion

- We have introduced a novel document representation method using **latent keyphrases**
  - Each dimension is explainable
  - Works for short text
  - Works for closed-domain text
- We have developed an efficient inference method to do **real time keyphrase identification**
- Future work
  - Better structure learning approach
  - Combined with knowledge base
  - Try other inference method other than Gibbs sampling
- Code available at http://jialu.info

26