NLP&CC
2014

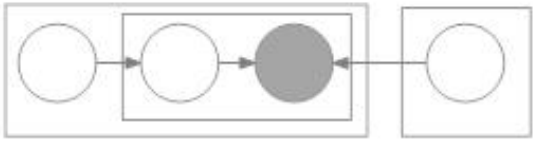
A Supervised Dynamic Topic Model

Zhuoren Jiang
Dalian Maritime University
December 2014

Motivation

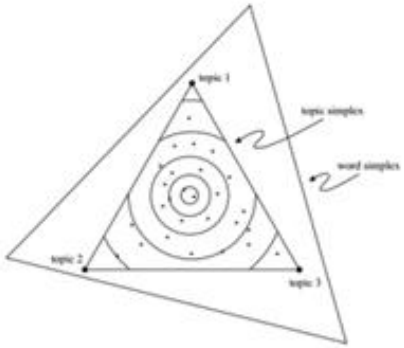
Probabilistic topic models

Assumptions



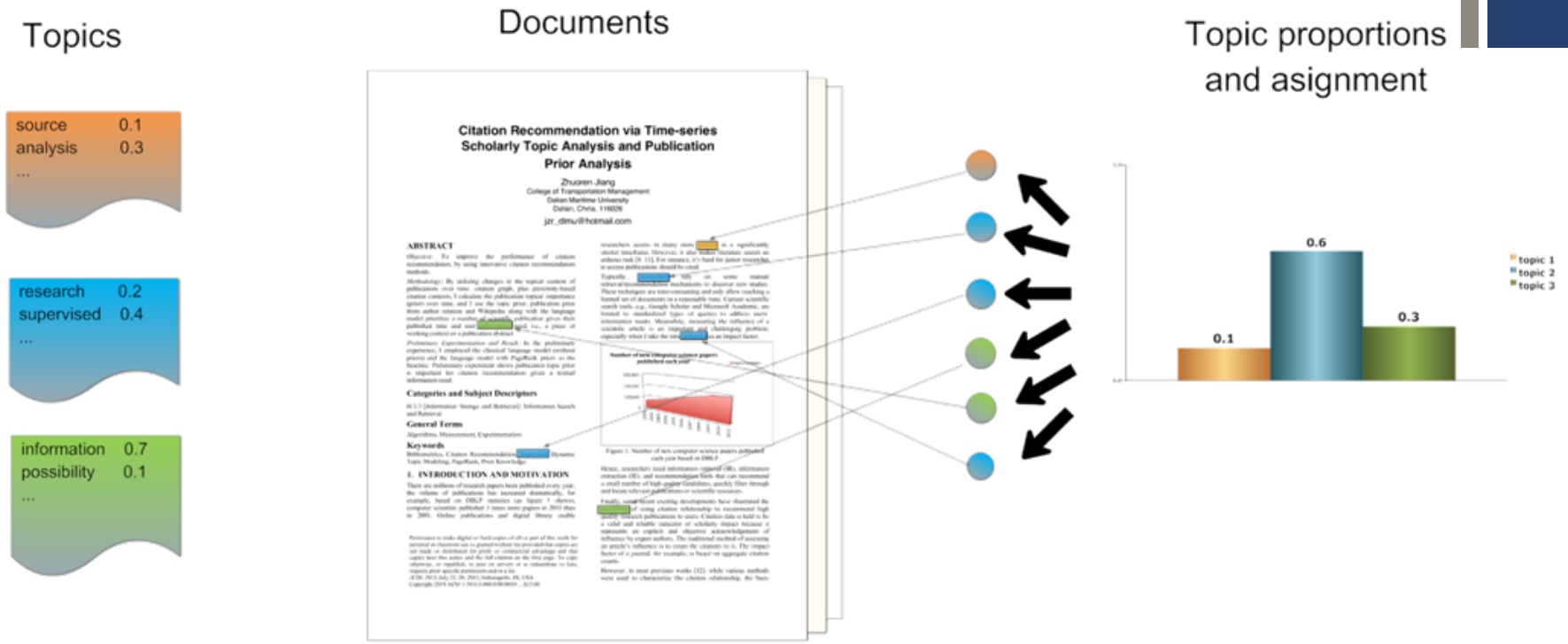
Data

Inference algorithm



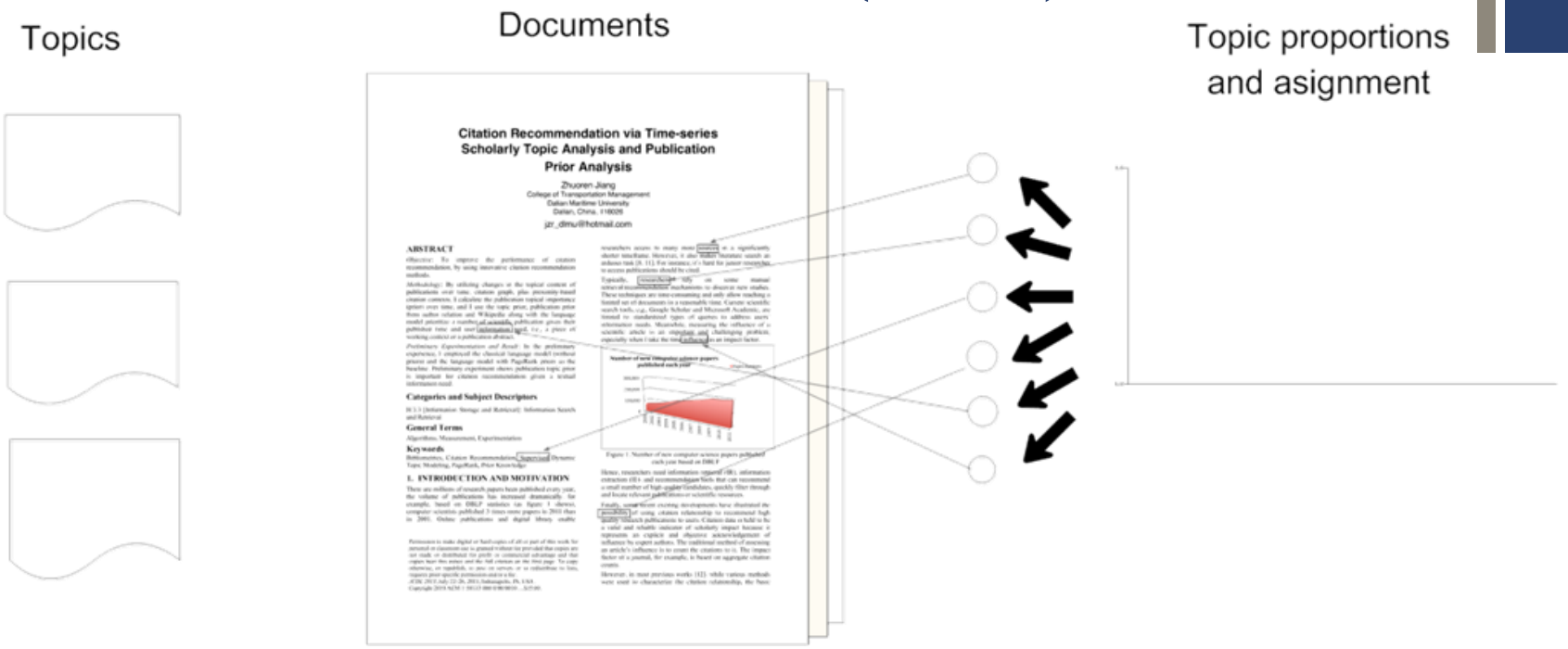
Discovered structrue

Topic Model-- Latent Dirichlet Allocation (LDA)



- Each topic is a distribution over words
- Each document is a mixture of corpus-wide topics
- Each word is drawn from one of those topics

Topic Model-- Latent Dirichlet Allocation (LDA)



- In reality, we only observe the documents
- The other structure are hidden variables
- Our goal is to infer the hidden variables

Limitation 1

- Assumption:
- The order of documents does not matter.

- Time attributes

- Language is changing over time

An example

1789
April 30

Inaugural addresses

Language change

2013
January 21

Solution

Topic should evolve over time

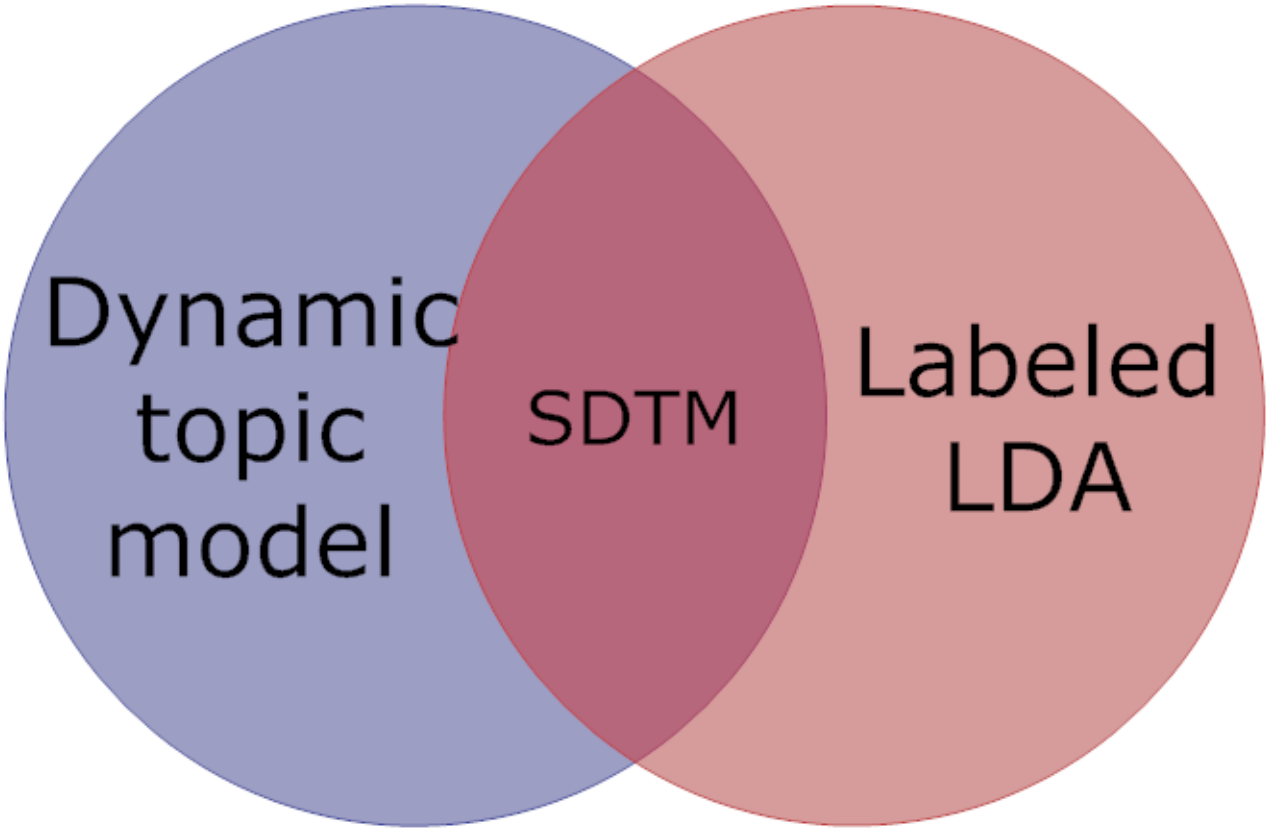
Limitation 2

- Topic is latent
- Hard to interpret
- Difficult for researchers to set topics number in advance

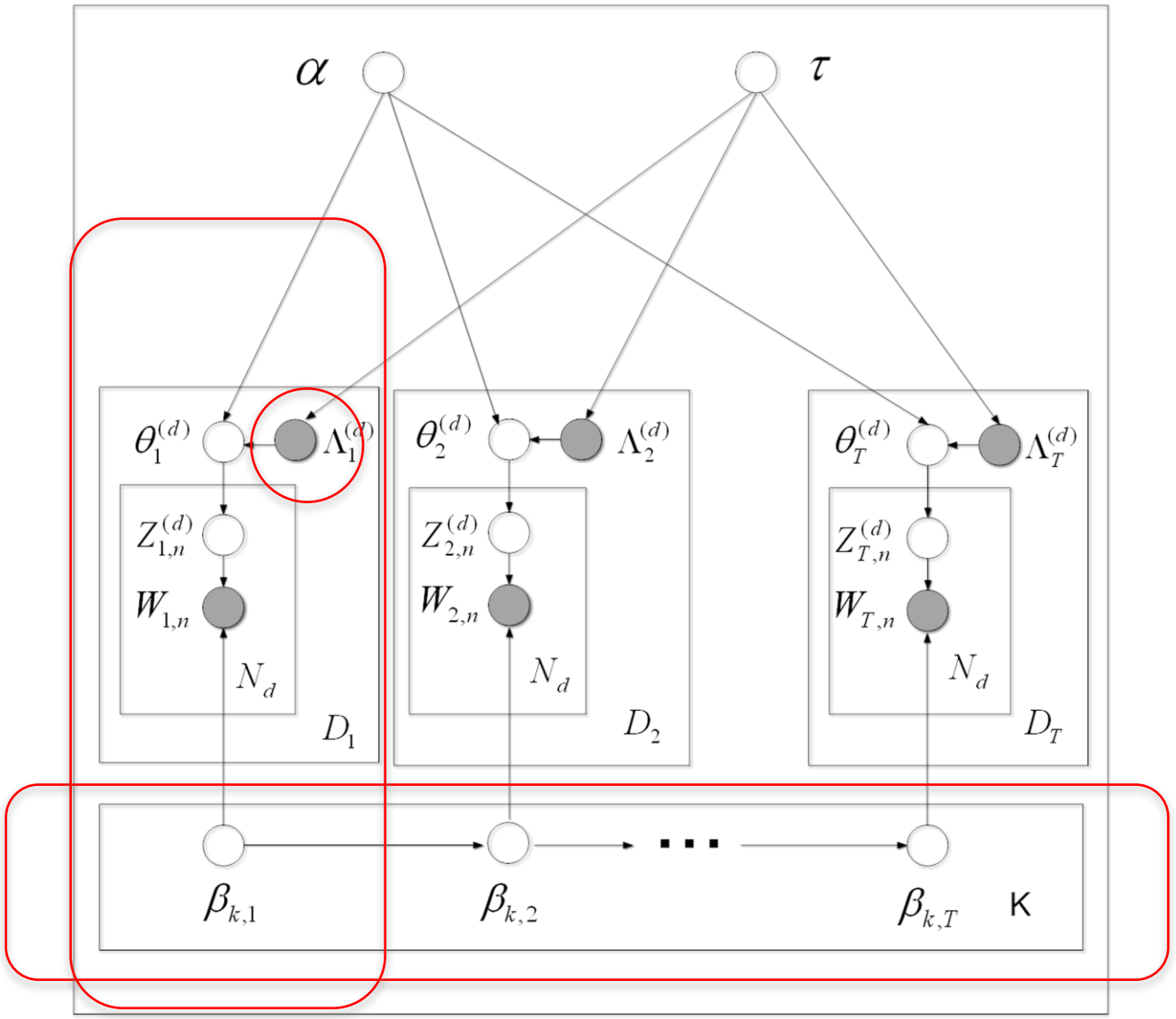
Solution

The problem of topic interpretation and topic number setting

Supervised dynamic topic model



Graphical representation



Variational Inference

- The idea behind variational methods is to optimize the free parameters of a distribution over the latent variables so that the distribution is close in Kullback-Liebler (KL) divergence to the true posterior; this distribution can then be used as a substitute for the true posterior.

Variational Inference

- For document- level: $q \quad Z$
- EM algorithm:
- linear-scaling Newton-Rhapson algorithm & Coordinate ascent algorithm

- For topic-level: b
- Standard Variational Kalman Filtering algorithm

Experiment

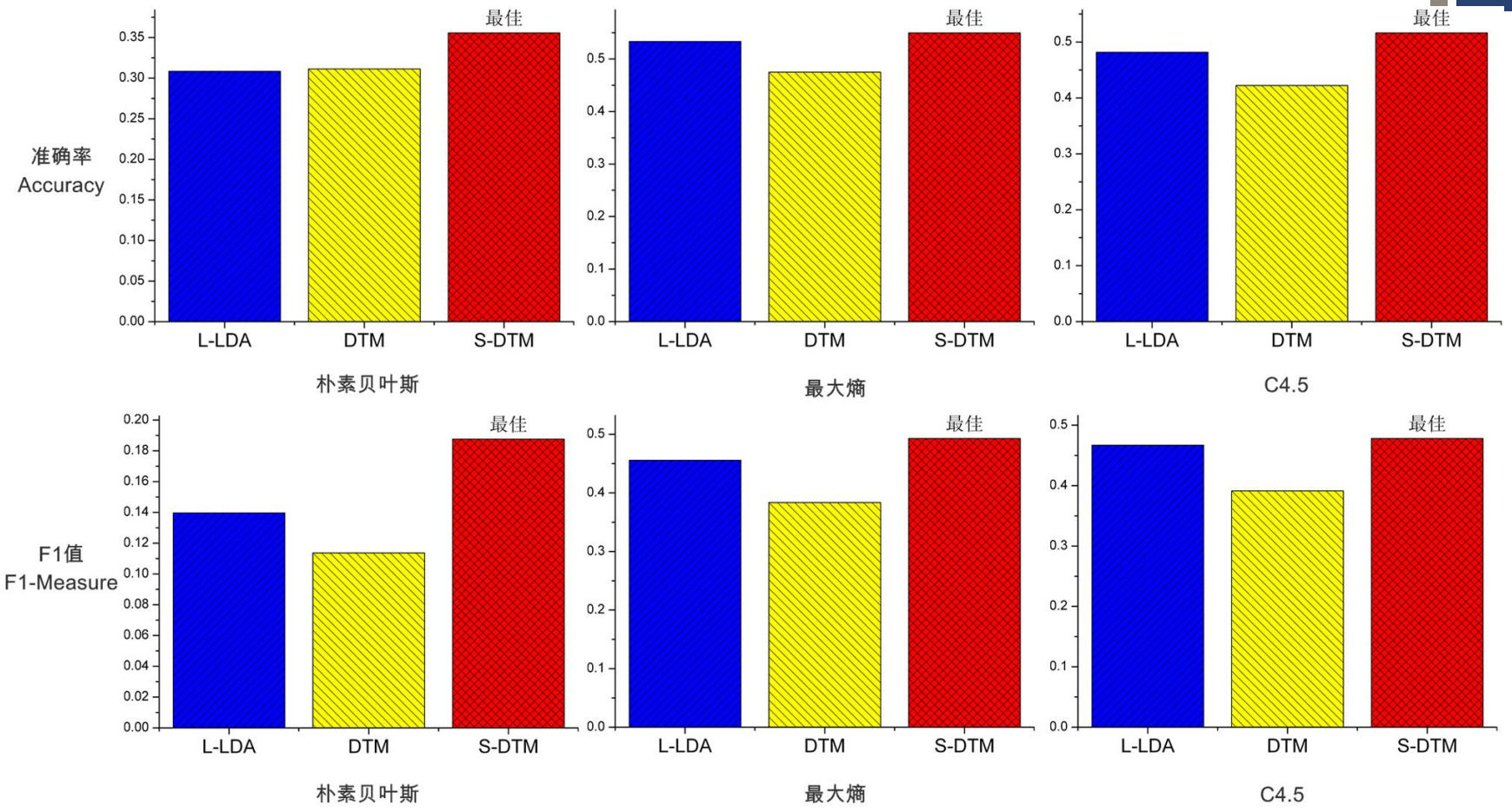
- Corpus: A twenty-five-year-spanning (1985-2009) Chinese journal paper corpus that is mainly focusing on natural language processing.
- Author provided keywords as the label of topic: 65 keywords (topics)
- Time slices (10)

<1996	1996-1999	2000-2001	2002-2003	2004	2005	2006	2007	2008	2009
392	402	413	548	438	426	511	506	531	448

Experiment

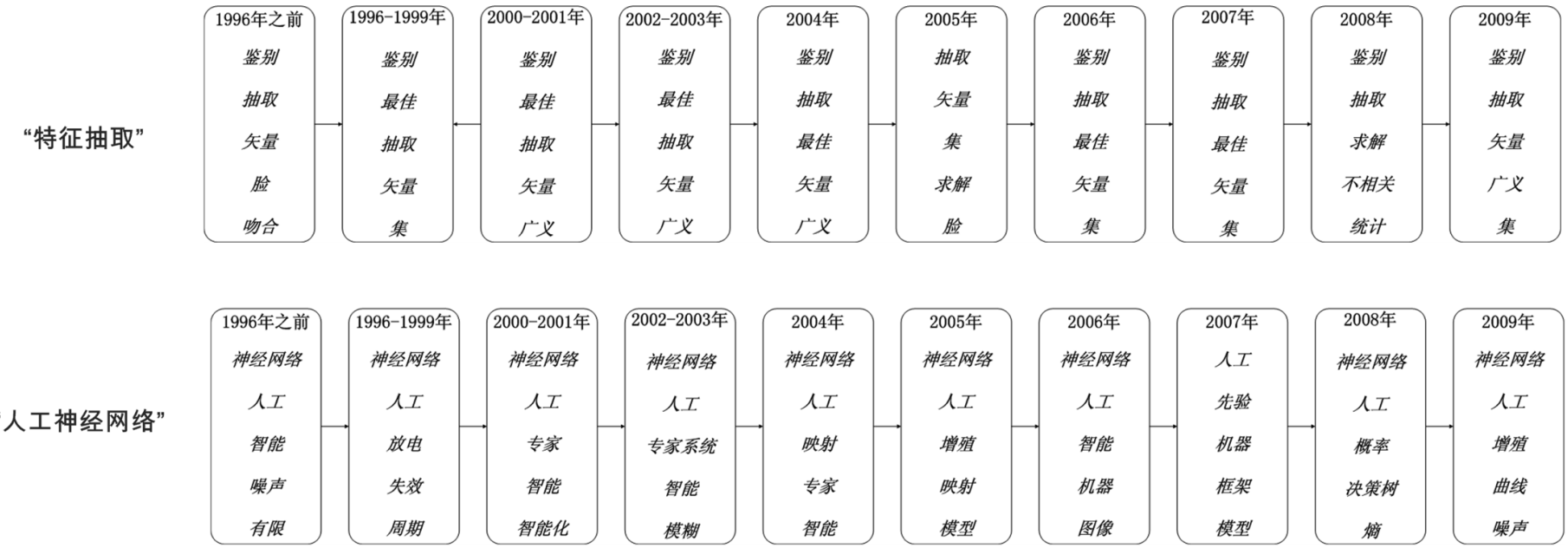
- Text classification
- 2,072 documents that has only one topic(label)
- 13 classes (topics)
- 1. Representing text with these topic distribution trained by these approaches
- 2. Perform text classification based on the representation results
- Naïve Bayes, Maximum Entropy, C4.5

Result



Visualization

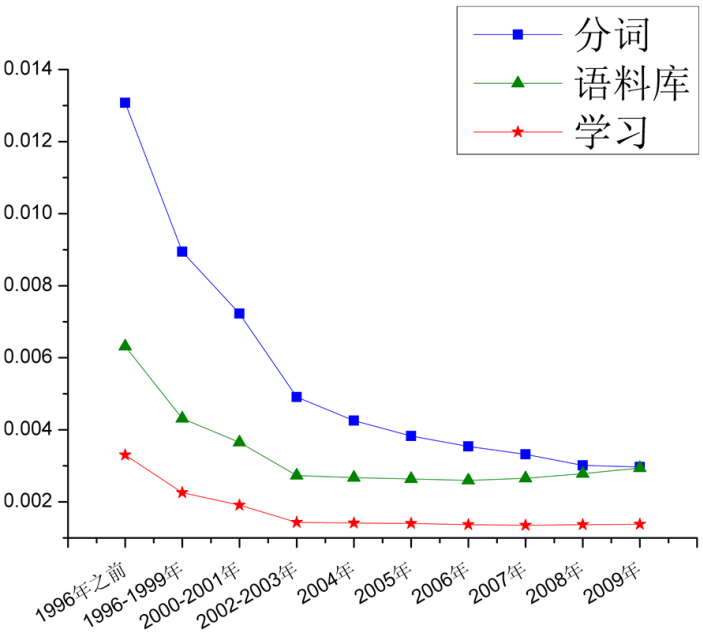
- The change of top words appear in two topics over time



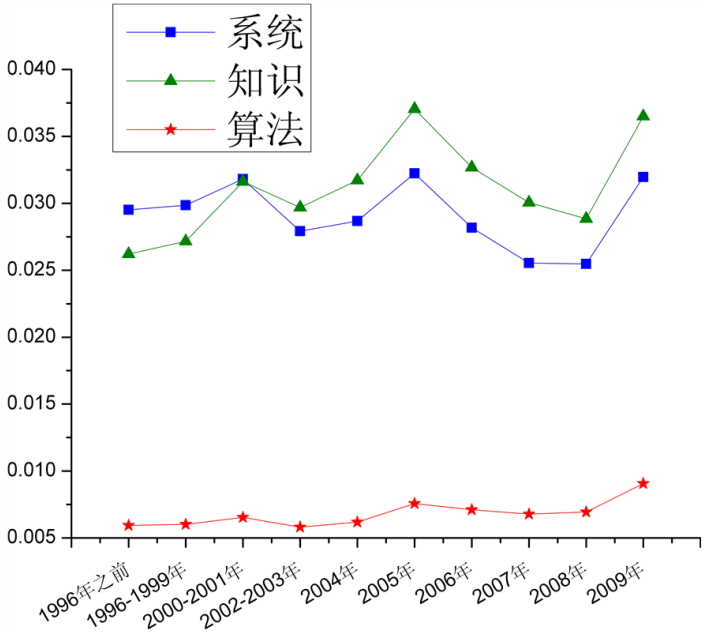
Visualization

- The probability change of specific words in two topics over time

“中文信息处理”



“语料库”



Conclusion

- For overcoming the the limitations of traditional topic models
- Model the time-varying language dynamics and is combined with supervised learning technology
- Comparing with static supervised topic model and unsupervised dynamic topic model, S-DTM has a better semantic interpretation performance

Our team

Yan Chen

Zhi Tang

Liangcai Gao

Xiaozhong Liu

Zhuoren Jiang

- Thank you!!
- Questions