# Employing Topic Models for Pattern-based Semantic Class Discovery

Huibin Zhang[2], Mingjie Zhu[3], Shuming Shi[1], Ji-Rong Wen[1]

[1] **Microsoft Research Asia**
[2] Nankai University
[3] University of Science and Technology of China

Send feedback to: shumings@microsoft.com

# Outline

- Problem statement
- Approach
  - *Topic models + Postprocessing*
- Experiments
- Related work
- Conclusion

Microsoft Research

# Semantic Class

- A set of terms or phrases with the peer or sibling relationship among them
  - {white, black, red, blue, green, orange, brown…}
  - {first, second, third, fourth, fifth…}
- Extract **Raw** Semantic Classes (RASCs)
  - Data sources
    - Document collection
    - The search results of search engines
  - Extraction techniques
    - Parsing
    - Pattern matching

# Pattern-based Semantic Class Extraction

- Sample patterns:

| Type | Pattern |
|---|---|
| SENT | NP {, NP}* {,} (and\|or) {other} NP |
| TAG | <UL> <LI>item</LI> … <LI>item</LI> </UL> |
| TAG | <SELECT> <OPTION>item…<OPTION>item </SELECT> |

- Example RASCs

| |
|---|
| $R_1$: {gold, silver, copper, coal, iron, uranium} |
| $R_2$: {red, yellow, *color*, gold, silver, copper} |
| $R_3$: {red, green, blue, yellow} |
| $R_4$: {HTML, Text, PDF, MS Word, *Any file type*} |
| $R_5$: {*Today*, *Tomorrow*, Wednesday, Thursday, Friday, Saturday, Sunday} |
| $R_6$: {*Bush*, Iraq, *Photos*, USA, *War*} |

RASCs should not be the final semantic classes

1. Noisy
2. Duplication

Microsoft Research

# Our Problem

- Source data: A collection of RASCs
- Input: A term or a phrase as the query
- Required output: Semantic class**es** the query belongs to
  - **Multi-membership**: A term/phrase belongs to multiple semantic classes
    - "Singapore" is both a **county** and a **city**.
  - Multi-membership is popular: Lots of English words are borrowed as company names, places, product names…

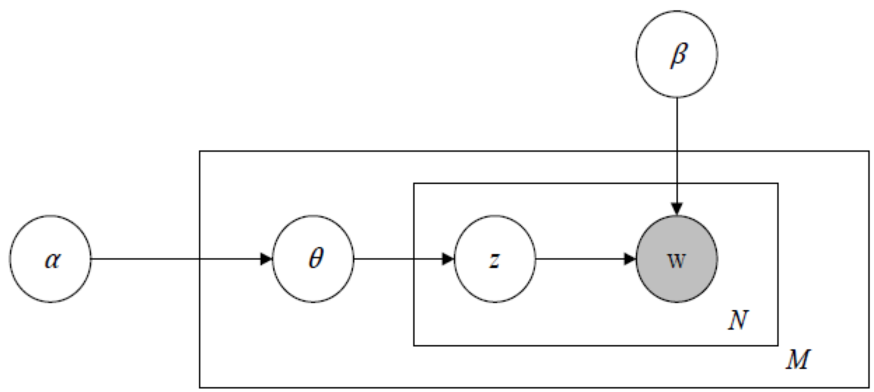
- Online research prototype:

  http://needleseek.msra.cn

Microsoft Research

# Our Approach: Main Ideas

- Main Idea: Employing topic models
- Topic modeling: Every "document" is modeled as a mixture of hidden "topics"
  - pLSI (Hofmann, 1999) ; LDA (Blei et al., 2003)…




Graphical model representation of LDA, from Blei et al. (2003)

# Our Approach: Main Ideas (cont.)

- Why topic modeling? Observations:
  - In our problem
    1) One item may belong to multiple semantic classes
    2) Some RASCs are comprised of items in multiple semantic classes
  - In (the typical application of) topic modeling
    1) A word can appear in multiple topics
    2) One document could be related to multiple topics

- Mapping of concepts

| Topic modeling | Semantic class construction |
|---|---|
| word | item (word or phrase) |
| document | RASC |
| topic | semantic class |

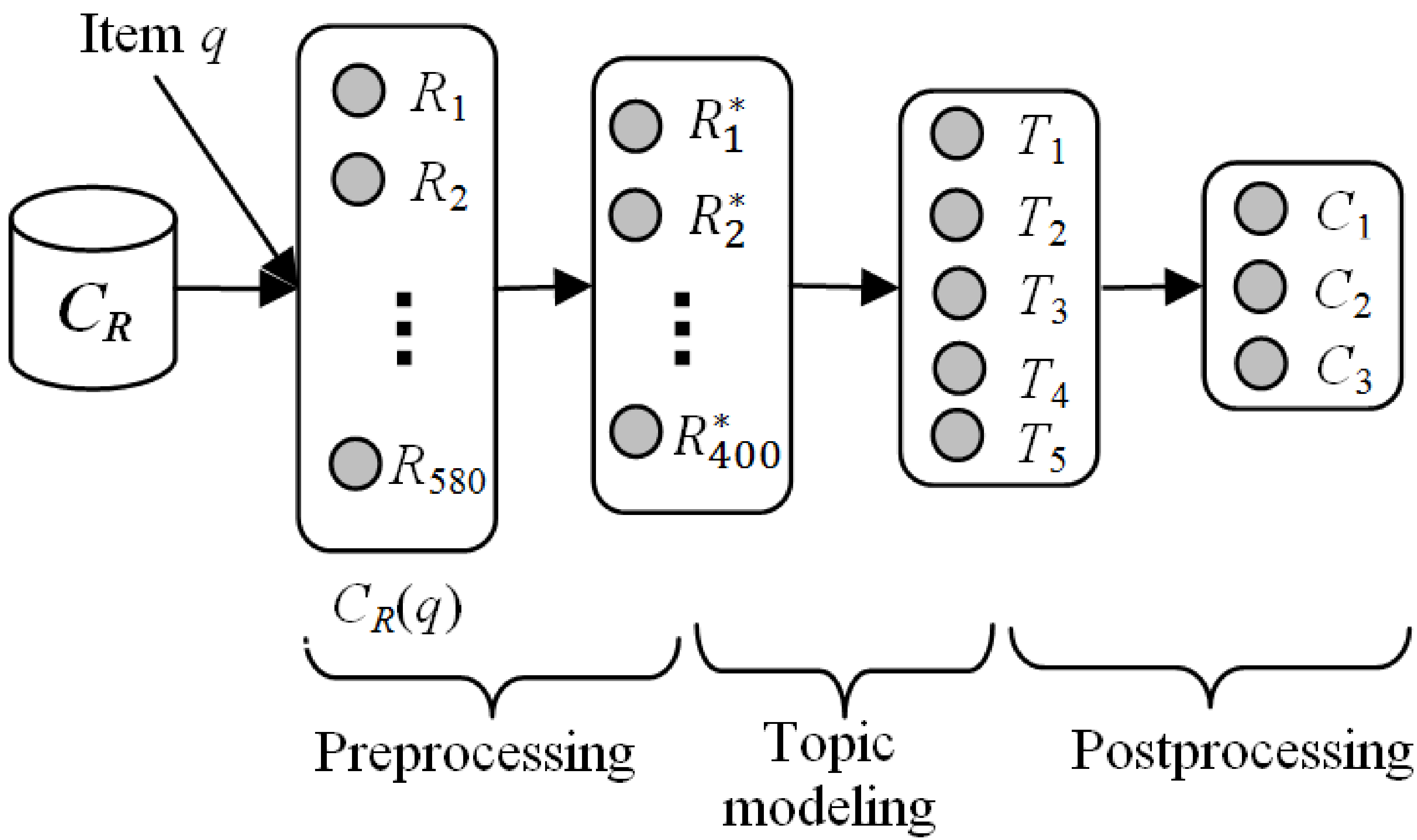
Microsoft Research

# Our Approach: Main Ideas (cont.)

- Challenges of adopting topic models here
  - Computation is intractable
    - 2.7 million unique RASCs extracted from 40 million web pages
  - Typical topic models require the number of topics (*k*) to be given
- Our solutions
  - Making computation feasible
    - Apply topic models to $C_R(q)$ rather than $C_R$
    - Preprocessing: Remove low frequency items
  - Set *k*: the number of topics
    - Set (for all items *q*) the topic number to be a fixed value ($k$=5).
    - A post-processing step to merge "topics" (very important!)

Microsoft Research

# Main Phases of our Approach



# Preprocessing

- Discard from all RASCs the items with frequency less than a threshold *h*
- Objective
  - Reduce the topic model training time without sacrificing results quality too much
- Effects
  - For some small ***h*** values, the results quality becomes *higher* after preprocessing is performed

Microsoft Research

# Adopting Topic Models

- For a query $q$, process the RASCs in $C_R(q)$
- Parameters
  - Topic number $\boldsymbol{k}$ is fixed (=5) for all queries
- Results
  - $\boldsymbol{k}$ topics (semantic classes) for query $q$
- Remarks
  - Inference is not needed
  - Why are the words/phrases within a resultant topic in peer relationship?

Microsoft Research

# Post-Processing

- Operations
  - Opr-1: Merge the “topics” yielded at the previous phase
  - Opr-2: Sort the items in each semantic class
- Operation-1: Merge topics (or semantic classes)
  - Repeatedly merge the two topics with the highest similarity until the similarity is under a threshold

$$sim(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$$

$$sim(C_1, C_2) = \frac{\sum_{a \in C_1} \sum_{b \in C_2} sim(a, b)}{|C_1| \cdot |C_2|} \quad ?$$

Microsoft Research

# Post-Processing (cont.)

- Operation-2: Sort items in a semantic class
  - Two factors
    - Average similarity between the item and the other items in the semantic class
    - Similarity between the item and the query item *q*
  - Define the importance of item *a* in semantic class *C*

$$g(a|C) = \lambda \cdot sim(a,C) + (1-\lambda) \cdot sim(a,q)$$

?

where

$$sim(a,C) = \frac{\sum_{b \in C} sim(a,b)}{|C|}$$

?

Microsoft Research

# Post-Processing (cont.)

- Item similarity calculation

$$sim(a,b) = \sum_{i=1}^{m} \log(1 + \sum_{j=1}^{k_i} w(P(C_{i,j})))$$

  - $C_{i,j}$: RASC $j$ in domain $i$
  - $P(C)$: The pattern by which RASC $C$ is extracted
  - $w(P)$: The weight of pattern P
  - $m$: The RASCs belong to $m$ domains

# Experimental Setup

- Datasets
  - Crawled 40 million English web pages
  - 2.7 million unique RASCs extracted (1 million distinct items)
- Query set
  - 55 queries provided by volunteers

Microsoft Research

# Experimental Setup

- Labeling
  - Manually determine the *standard semantic classes* (SSCs) for the query
    - Query: “Georgia”
    - The ideal/standard semantic classes may include Countries, and U.S. states
  - Each item is assigned a label of “Good”, “Fair”, or “Bad”, w.r.t. each SSC
    - “silver” is labeled “Good” with respect to “colors” and “chemical elements”
    - Term “color” is “Bad” w.r.t. “colors”

Microsoft Research

# Experimental Setup

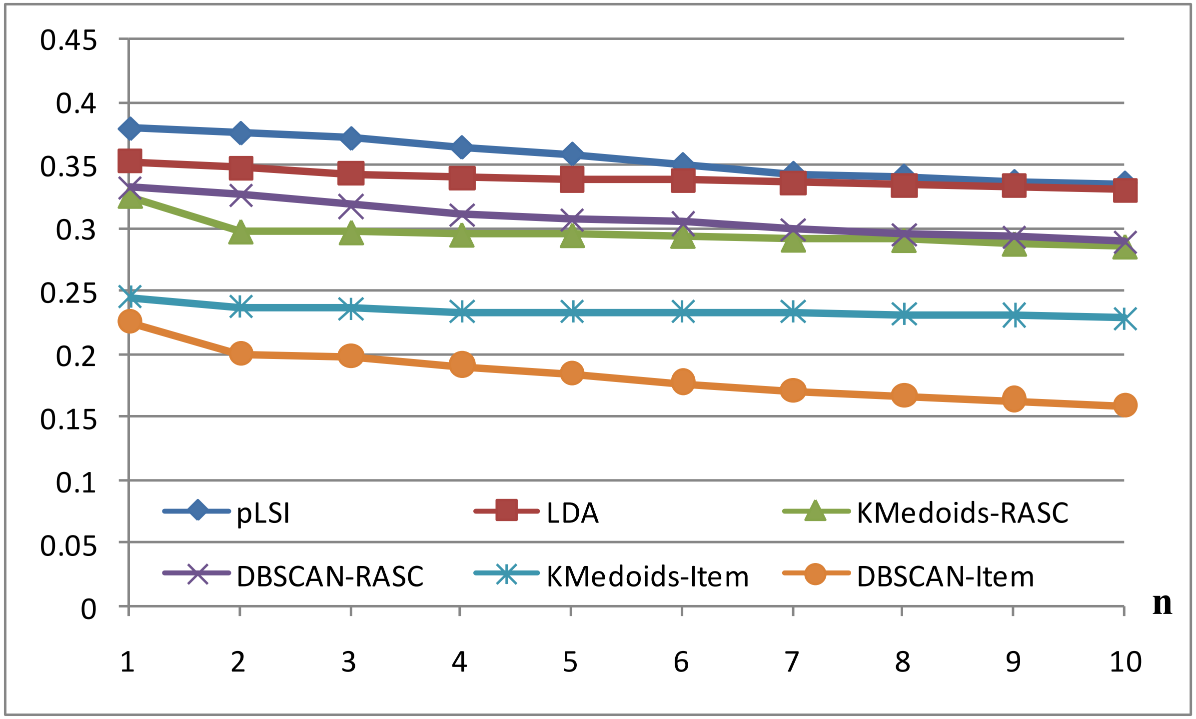
- Evaluation
  - Each resultant semantic class is an **ordered** item list
  - Adopting information retrieval (IR) metrics to evaluate it
    - Mean Average Precision (**MAP**)
    - Normalized Discounted Cumulative Gain (**nDCG**)
    - …
  - Evaluate multiple semantic classes
    - Extend existing IR metrics to support multiple ordered lists
    - nDCG → MnDCG

Microsoft Research

# Experimental Setup

- Approaches for comparison
  - LDA: Our approach with LDA as the topic model
  - pLSI: Our approach with pLSI as the topic model
  - KMedoids-RASC: RASC clustering using K-Medoids
  - DBSCAN-RASC: RASC clustering using DBSCAN
  - KMedoids-Item: Item clustering using K-Medoids
  - DBSCAN-Item: Item clustering using DBSCAN

# Quality comparison




- Frequency threshold $h = 4$ in preprocessing
- $k = 5$ in topic models
- Metrics: MnDCG@$k$

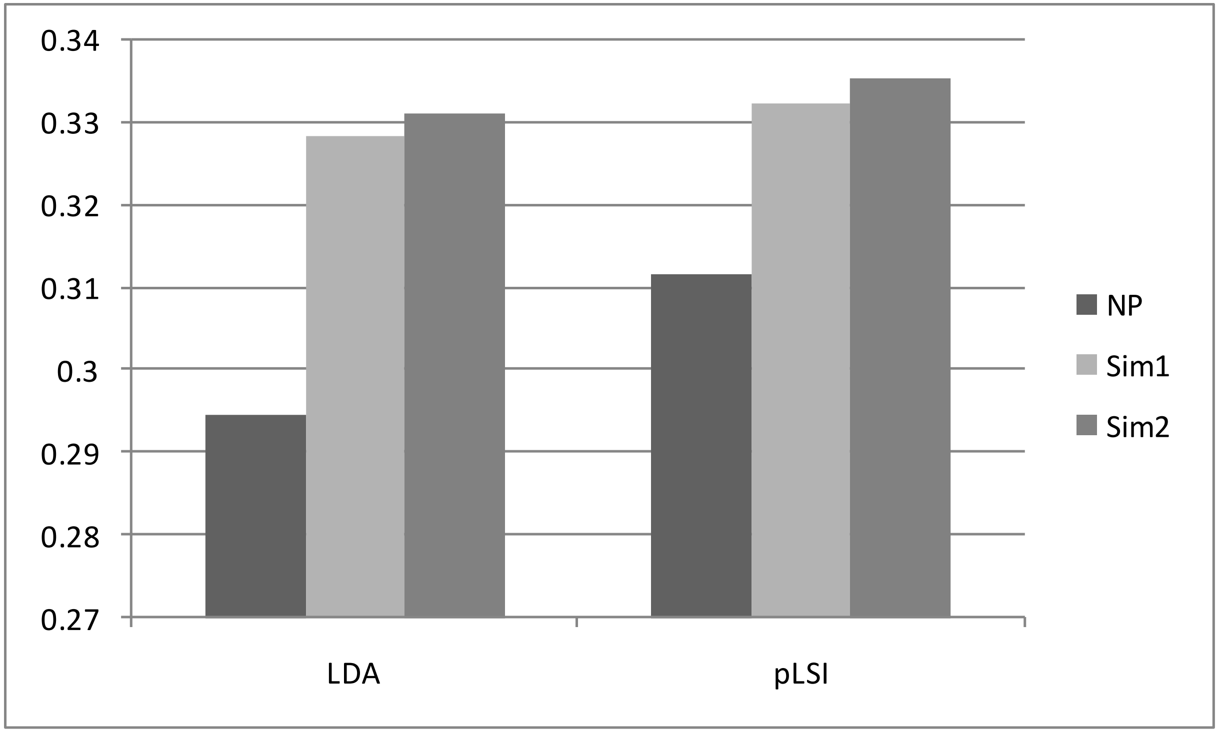
# Preprocessing

- LDA approaches
- Different preprocessing thresholds ( $0 < h < 10$)

| $h$ | Avg. Query Proc. Time (seconds) | Quality (MnDCG@10) |
|---|---|---|
| 1 | 0.414 | 0.281 |
| 2 | 0.375 | 0.294 |
| 3 | 0.320 | 0.322 |
| 4 | 0.268 | **0.331** |
| 5 | 0.232 | 0.328 |
| 6 | 0.210 | 0.315 |
| 7 | 0.197 | 0.315 |
| 8 | 0.184 | 0.313 |
| 9 | 0.173 | 0.288 |

# Post-Processing Results

- Topic modeling approaches with and without post-processing
- Metric: MnDCG@10



# Related work

- Semantic class discovery
  - Set expansion
    - Hindle  (1990; Ruge (1992); Lin (1998); Google sets; Ghahramani and Heller (2005); Wang and Cohen (2007); Kozareva (2008)
  - Pattern-based approaches
    - Shinzato and Torisawa (2004); Shinzato and Torisawa (2005)
    - Pasca (2004); Shi et al. (2008)
  - Distributional similarity approaches
    - Harris (1985); Lin  and Pantel (2001); Pantel and Lin (2002)
- Topic modeling applications
  - Lots of document clustering applications
  - Word Sense Disambiguation (WSD)
    - Cai et al (2007); Boyd-Graber et al. (2007)

Microsoft Research

# Summary

- Employ topic models to construct semantic classes
  - Query $q$ → Retrieve $C_R(q)$ → Preprocessing
    → Topic modeling → Post-processing (merge topics)
- Propose an evaluation methodology
- Contributions:
  - Find an effective way of constructing high-quality semantic classes with **multi-membership** support, in the pattern-based category
  - For the first time, demonstrate the effectiveness of topic modeling in semantic class construction

Microsoft Research

# Thank you

# Questions?

- Welcome to try our online research prototype:

  http://needleseek.msra.cn/

- Visit http://needleseek.msra.cn/rascsearch/ to search and download raw semantic classes (RASCs) for your research work.

Microsoft Research