基于文档主题结构的关键词抽取方法研究

答辩人：刘知远

导　师：孙茂松教授

2011年6月12日

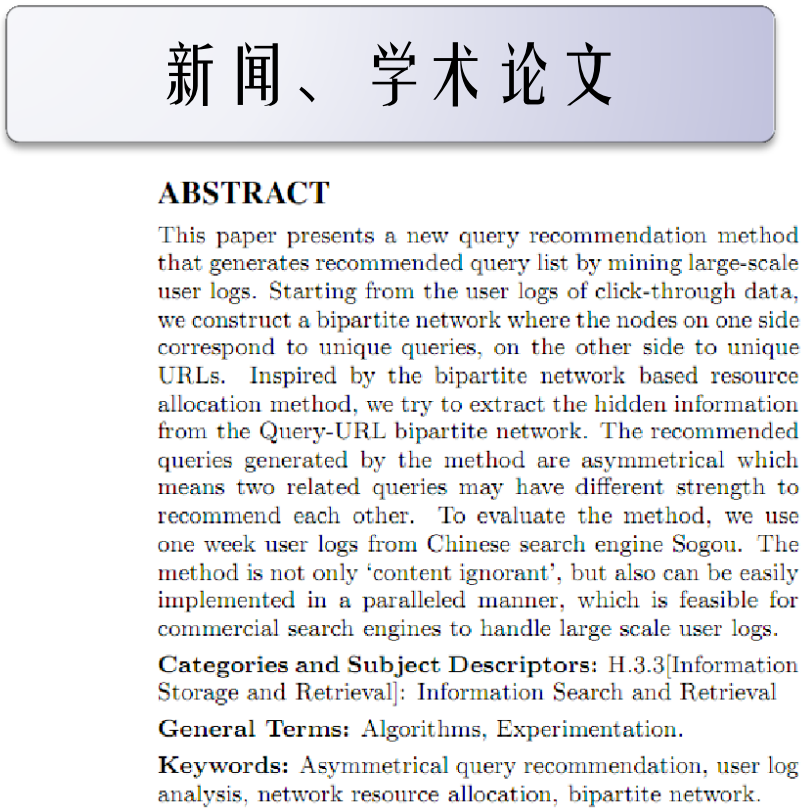
报告摘要

- 选题背景和意义
- 文献综述
- 研究内容
- 研究总结
- 未来工作与展望

问题描述-关键词自动标注

- 定义：选取若干关键词概括文档主题内容

新闻、学术论文 → 社会化标注



ABSTRACT

This paper presents a new query recommendation method that generates recommended query list by mining large-scale user logs. Starting from the user logs of click-through data, we construct a bipartite network where the nodes on one side correspond to unique queries, on the other side to unique URLs. Inspired by the bipartite network based resource allocation method, we try to extract the hidden information from the Query-URL bipartite network. The recommended queries generated by the method are asymmetrical which means two related queries may have different strength to recommend each other. To evaluate the method, we use one week user logs from Chinese search engine Sogou. The method is not only 'content ignorant', but also can be easily implemented in a paralleled manner, which is feasible for commercial search engines to handle large scale user logs.

Categories and Subject Descriptors: H.3.3[Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation.

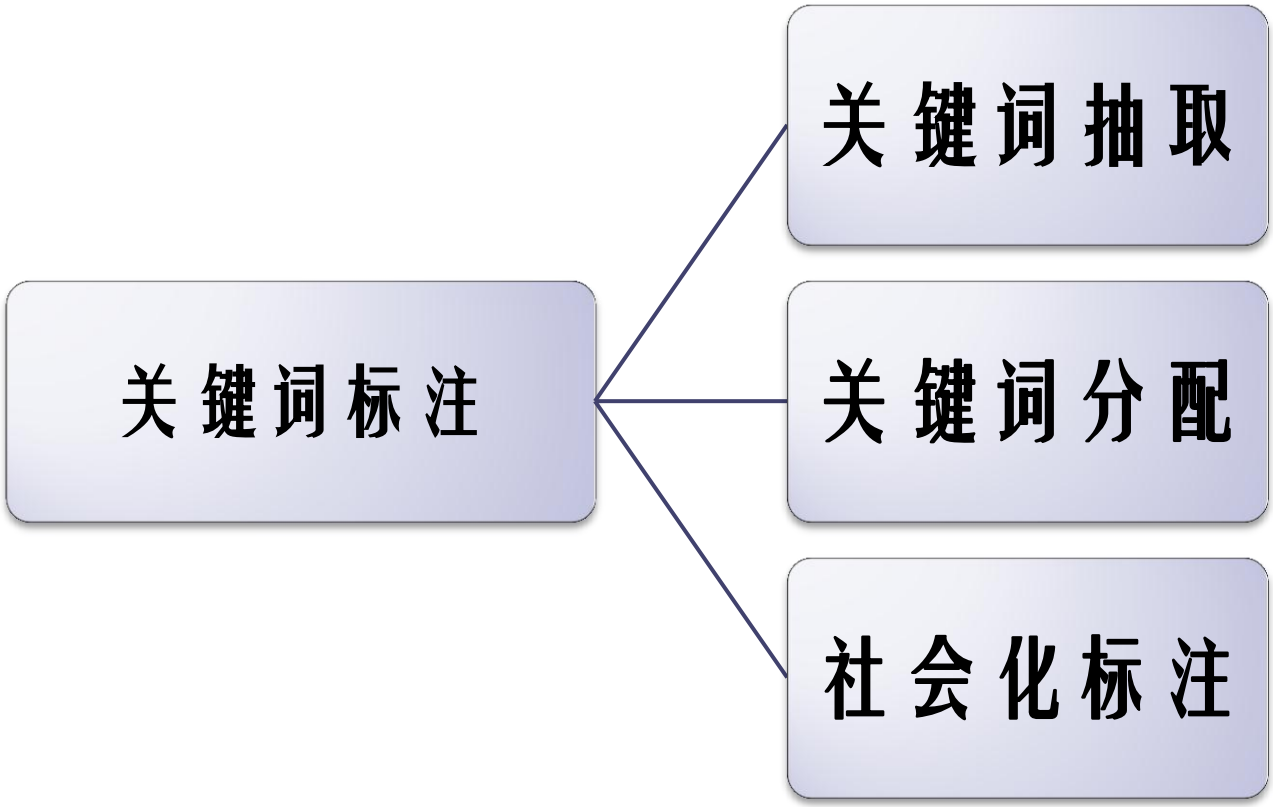
Keywords: Asymmetrical query recommendation, user log analysis, network resource allocation, bipartite network.

(a) 论文关键词

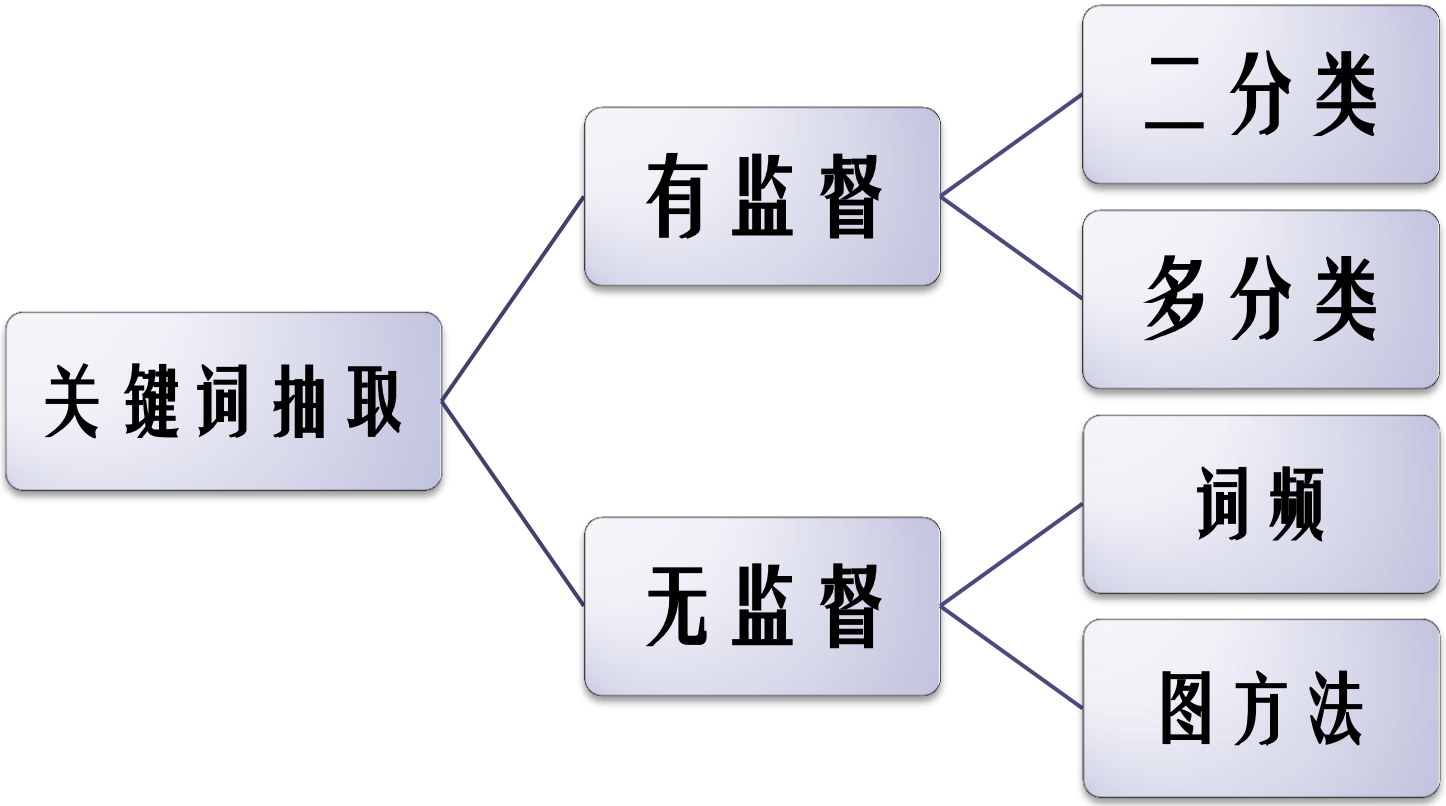
(b) 新闻关键词

(c) 社会标签

文献综述-关键词标注方式



文献综述-关键词抽取方法



文献综述-有监督方法

- 转化为二分类问题
  - 判断某个候选关键词是否为关键词
  - Frank 1999采用朴素贝叶斯分类器
  - Turney 2000采用C4.5决策树分类器
- 转化为多分类问题
  - 文本分类问题
  - 受控词表作为候选关键词集合（分类标签）

文献综述-无监督方法

- 词频
  - 基于TFIDF及其变形对候选关键词进行排序

$$TFIDF_w = tf_w \cdot log_2 \frac{D}{\{df_w\}}$$

- 图方法
  - Rada 2004: PageRank ➔ TextRank
  - Litvak and Last 2007: HITS

TFIDF:

仅考虑词自身频度

TextRank:

考虑文档内词间语义关系

文献综述- TextRank

构建词网 → PageRank → 选取排序最高的词为关键词

$$R(w_i) = \lambda \sum_{j:w_j \to w_i} \frac{e(w_j, w_i)}{O(w_j)} R(w_j) + (1-\lambda)\frac{1}{|V|}$$

- R(w): w的PageRank值
- O(w): w的出度
- e(w_j, w_i): $w_j \to w_i$ 边上的权重
- V: 节点集合
- λ: 平滑因子

研究问题

- 关键词应具备特点
  - 相关性，可读性，覆盖性
  - 关键词与文档主题保持一致性
- 在关键词抽取中考虑对文档主题的覆盖性
  - 一个文档往往有多个主题
  - 现有方法没有提供机制对主题进行较好覆盖
- 解决文档与关键词间的词汇差异问题
  - 许多关键词在文档中频度较低、甚至没有出现
    - “machine transliteration” vs “machine translation”
    - “iPad” vs “Apple”

研究思路

- 对文档主题结构进行建模，并用于提高关键词抽取的覆盖性
  - 利用文档内部信息构造文档主题
  - 利用文档外部信息构造文档主题
  - 结合文档内部、外部信息
- 利用无标注文档集中的文档与关键词的主题一致性，解决文档与关键词的词汇差异

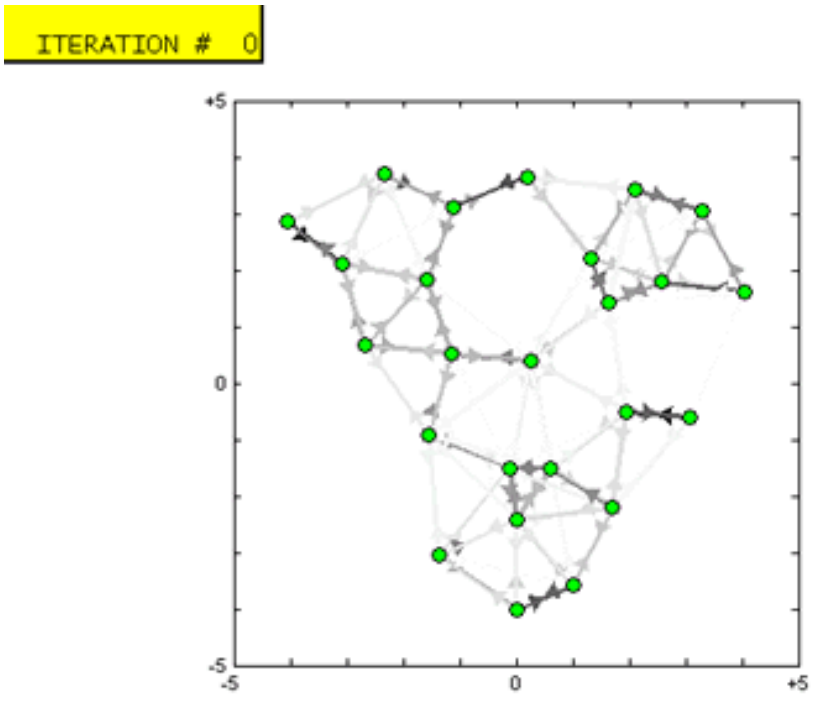
研究内容

1. 通过文档词聚类构建主题进行关键词抽取
2. 利用隐含主题构建主题进行关键词抽取
3. 综合利用隐含主题和文档结构进行关键词抽取
4. 利用机器翻译弥合词汇差异进行关键词抽取
5. 关键词抽取的典型应用

通过文档词聚类构建主题进行关键词抽取

研究动机与方法

- 动机：利用文档内部信息对文档主题进行建模
- 方法
  - 在文档中选取候选关键词
  - 计算候选关键词之间的语义相似度
  - 对文档中的词进行聚类
  - 在每个聚类中选取聚类中心（exemplar）扩展出关键词



算法细节

- 候选关键词相似度度量
  - 基于同现关系的相似度
  - 基于维基百科的相似度
    - Cosine，Euclid，PMI，NGD
- 聚类方法选取
  - 层次聚类（hierarchical clustering）
  - 谱聚类（spectral clustering）
  - 消息传递聚类（Affinity Propagation）

实验结果

- 数据集合：Hulth论文摘要
- 参数影响

Parameters	Precision	Recall	F1-measure
Cooccurrence-based Relatedness			
$w = 2$	0.331	0.626	0.433
$w = 4$	0.333	0.621	0.434
$w = 6$	0.331	0.630	0.434
$w = 8$	0.330	0.623	0.432
$w = 10$	0.333	0.632	0.436
Wikipedia-based Relatedness			
cos	0.348	0.655	0.455
euc	0.344	0.634	0.446
pmi_p	0.344	0.621	0.443
pmi_t	0.344	0.619	0.442
pmi_c	0.350	0.660	**0.457**
ngd	0.343	0.620	0.442

Parameters	Precision	Recall	F1-measure
Hierarchical Clustering			
$m = \frac{1}{4}n$	0.365	0.369	0.367
$m = \frac{1}{3}n$	0.365	0.369	0.367
$m = \frac{1}{2}n$	0.351	0.562	0.432
$m = \frac{3}{4}n$	0.346	0.629	0.446
$m = \frac{4}{5}n$	0.340	0.657	0.448
Spectral Clustering			
$m = \frac{1}{4}n$	0.385	0.409	0.397
$m = \frac{1}{3}n$	0.374	0.497	0.427
$m = \frac{1}{2}n$	0.374	0.497	0.427
$m = \frac{3}{4}n$	0.350	0.660	**0.457**
$m = \frac{4}{5}n$	0.340	0.679	0.453
Affinity Propagation			
$p = max$	0.331	0.688	0.447
$p = mean$	0.433	0.070	0.121
$p = median$	0.422	0.078	0.132
$p = min$	0.419	0.059	0.103

实验结果

- 与其他算法的比较
- 举例

Keyphrases when $m = \frac{1}{4}n, \frac{1}{3}n, \frac{1}{2}n$
unsupervis method; various **unsupervis** rank method; **exemplar term**; state-of-the-art **graph-bas** rank method; **keyphras**; **keyphras** extract

Keyphrases when $m = \frac{2}{3}n$
unsupervis method; manual assign; brief summari; various unsupervis rank method; exemplar term; document; state-of-the-art graph-bas rank method; experi; keyphras; import score; keyphras extract

	Assigned		Correct				
Method	Total	Mean	Total	Mean	Precision	Recall	F1-measure
Hulth's	7,815	15.6	1,973	3.9	0.252	0.517	0.339
TextRank	6,784	13.7	2,116	4.2	0.312	0.431	0.362
HC	7,303	14.6	2,494	5.0	0.342	0.657	0.449
SC	7,158	14.3	2,505	5.0	**0.350**	0.660	**0.457**
AP	8,013	16.0	2,648	5.3	0.330	**0.697**	0.448

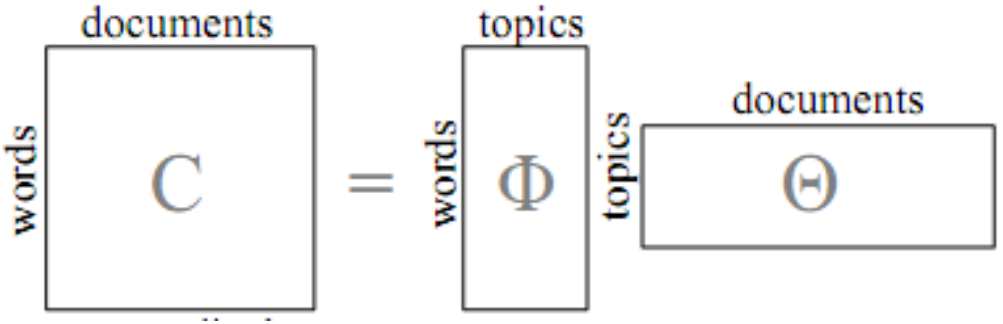
小结

- 提出了利用聚类对文档内部主题结构建模的关键词抽取算法
- 对比了不同的相似度度量算法、聚类算法
- 较好地实现推荐关键词的覆盖性
- 问题
  - 不同聚类个数较大地影响关键词抽取效果
  - 仅利用文档内部信息受到较大局限

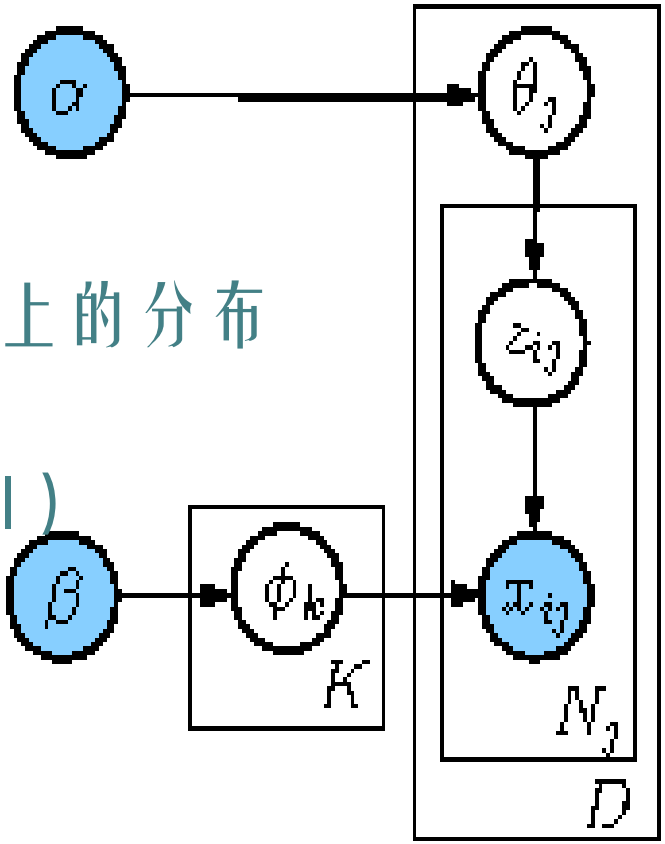
Zhiyuan Liu, Peng Li, Yabin Zheng, Maosong Sun. Clustering to Find Exemplar Terms for Keyphrase Extraction. The Conference on Empirical Methods in Natural Language Processing (EMNLP), 2009.

通过隐含主题模型构建主题进行关键词抽取

隐含主题模型



- 对文档主题进行建模的无监督学习模型
  - 由用户指定隐含主题个数
  - 根据大规模文档集合中学习
  - 每个主题是在词上的分布
  - 每个词和文档都可以表示为主题上的分布
- 常见隐含主题模型
  - Latent Semantic Analysis （LSA/LSI）
  - Probabilistic LSA （pLSA）
  - Latent Dirichlet allocation （LDA）



隐含主题模型示例

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

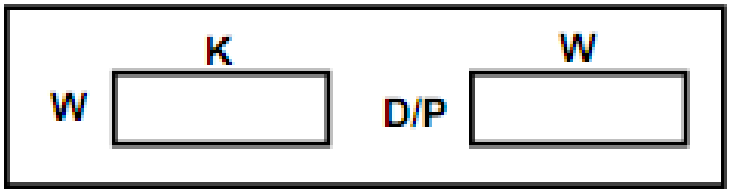
word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

利用隐含主题模型进行关键词抽取

- 文档的主题分布：$P(z|d)$
- 词的主题分布：$P(z|w)$
- 通过多种方式度量其语义关系
  - Cosine similarity
  - KL-divergence
  - $P(w|d) = \sum_z P(w|z)P(z|d)$
- 存在问题
  - LDA运算复杂度较高，在大规模数据集合上运行速度较慢
  - 解决方案：并行化

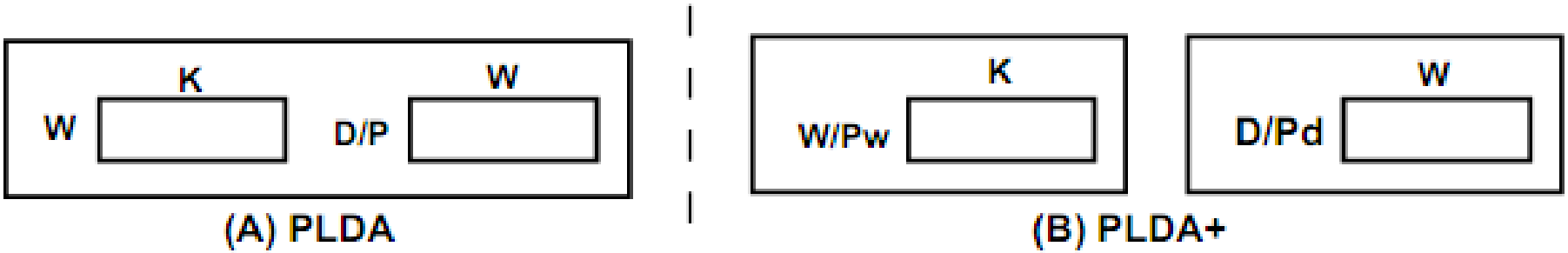
隐含主题模型的并行研究

- LDA的已有并行算法
  - Approximate Distributed LDA （AD-LDA）
  - Asynchronous LDA（AS-LDA）
- 主要问题
  - 内存瓶颈：要求主题模型（$W \times K$）保存于每台机器内存
  - 通信瓶颈：要求每次迭代机器间都要交互整个主题模型

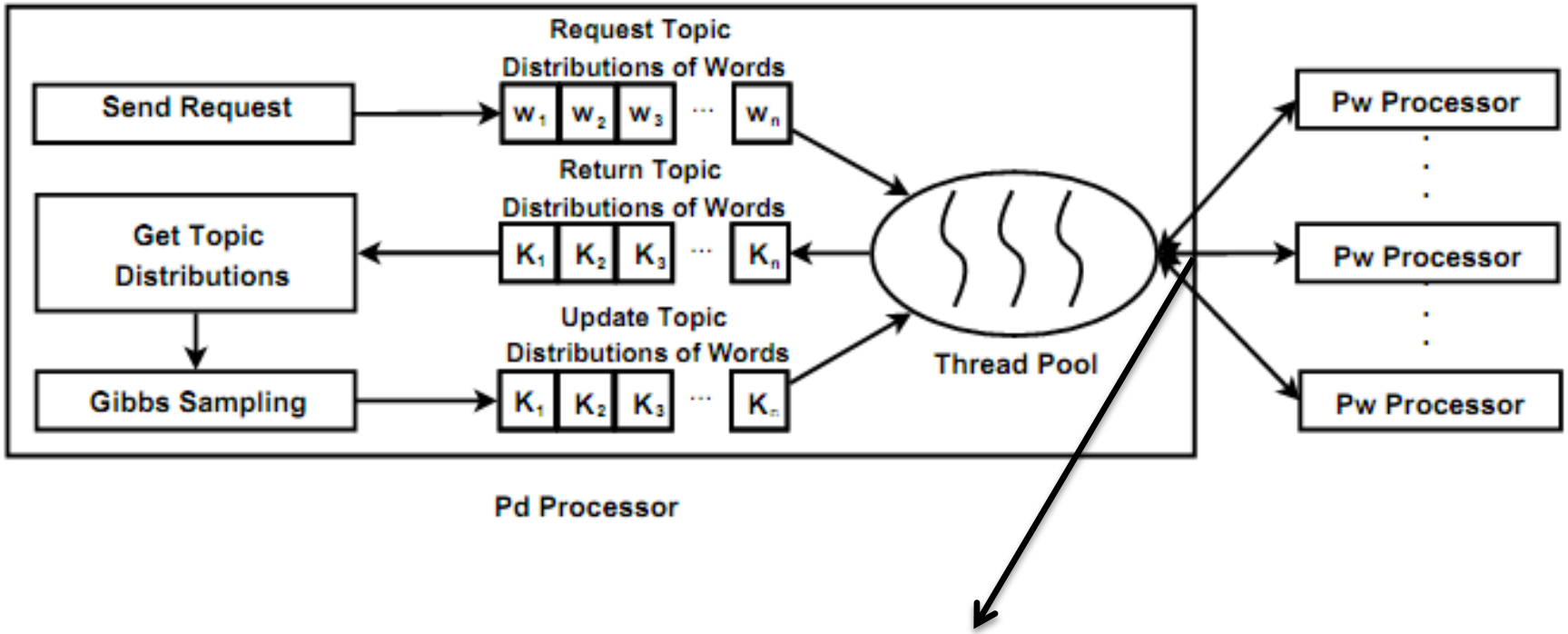


PLDA+算法

- 机器分为两种功能：
  - 一部分机器用于维护训练文档
  - 一部分机器用于维护主题模型

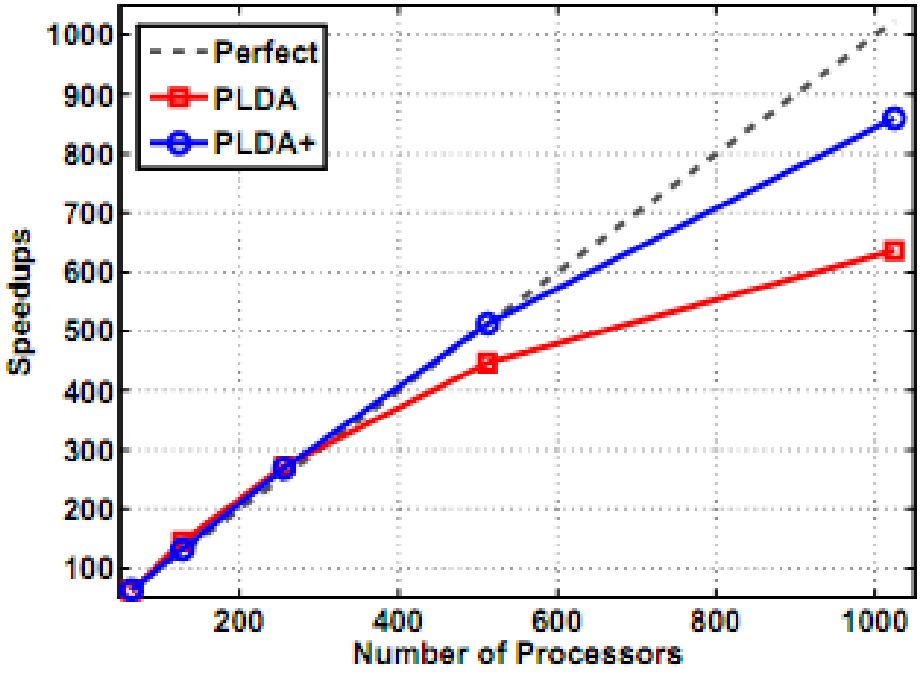


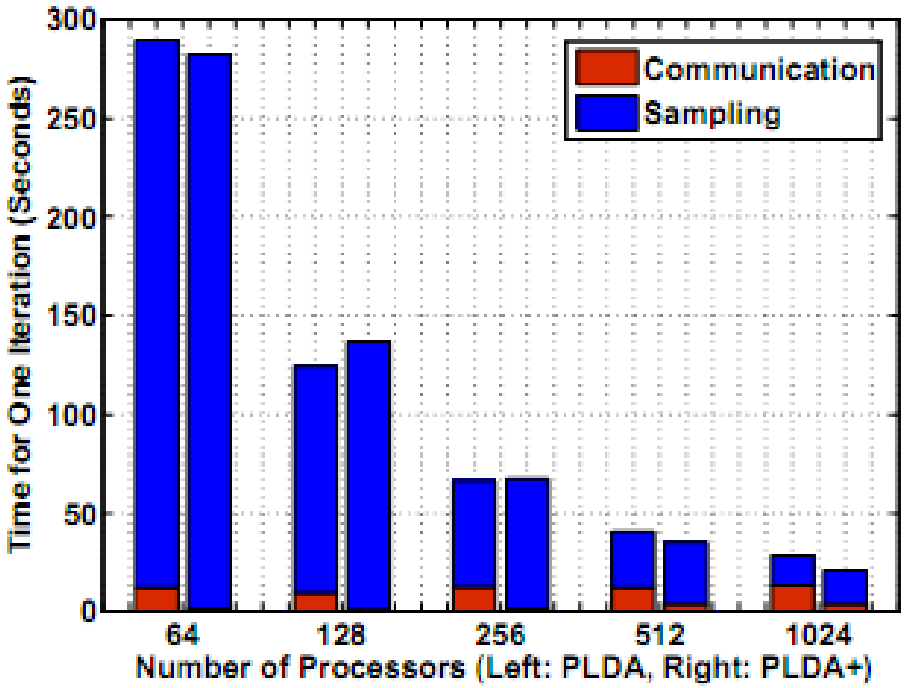
PLDA+算法



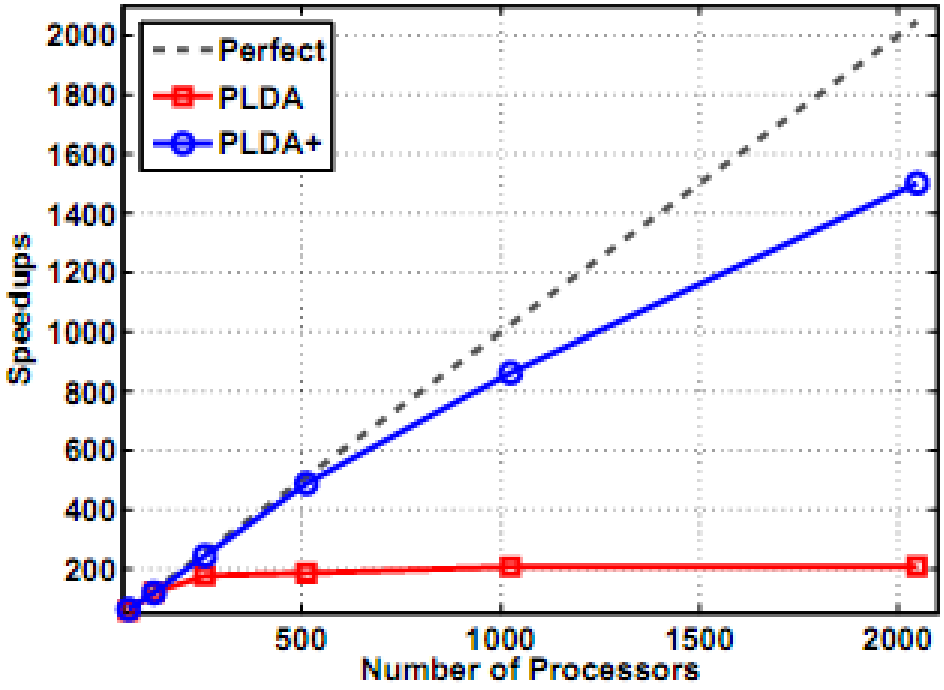
当网络不稳定时，可能会有部分请求不会被响应，超过一定时间后会被丢弃。我们称平均被丢弃的请求比例为missing ratio δ。

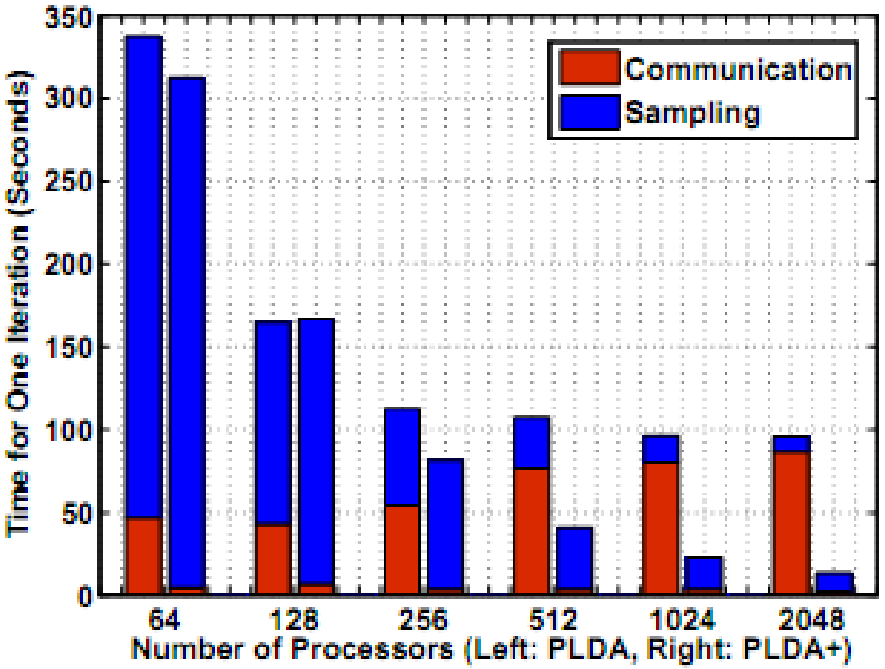
实验效果-维基百科（2万词汇）





实验效果-维基百科（20万词汇）





小结

- 通过PLDA+有效解决了通信瓶颈和内存瓶颈，使得LDA得到2000+以上的加速
- 下面展示利用隐含主题模型进行关键词抽取的效果

Zhiyuan Liu, Yuzhou Zhang, Edward Y. Chang, Maosong Sun. PLDA+: Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing. ACM Transactions on Intelligent Systems and Technology (ACM TIST), 2010.

LDA进行关键词抽取效果

- 在NEWS数据集合上推荐10个关键词的效果

方法	Precision	Recall	F_1-Measure
TFIDF	0.239	0.295	0.264
TextRank	0.242	0.299	0.267
NEWS,PL,$K = 50$	0.258	0.318	0.285
Wiki,PL,$K = 1500$	0.267	0.329	0.294

- 在RESEARCH数据集合上推荐5个关键词的效果

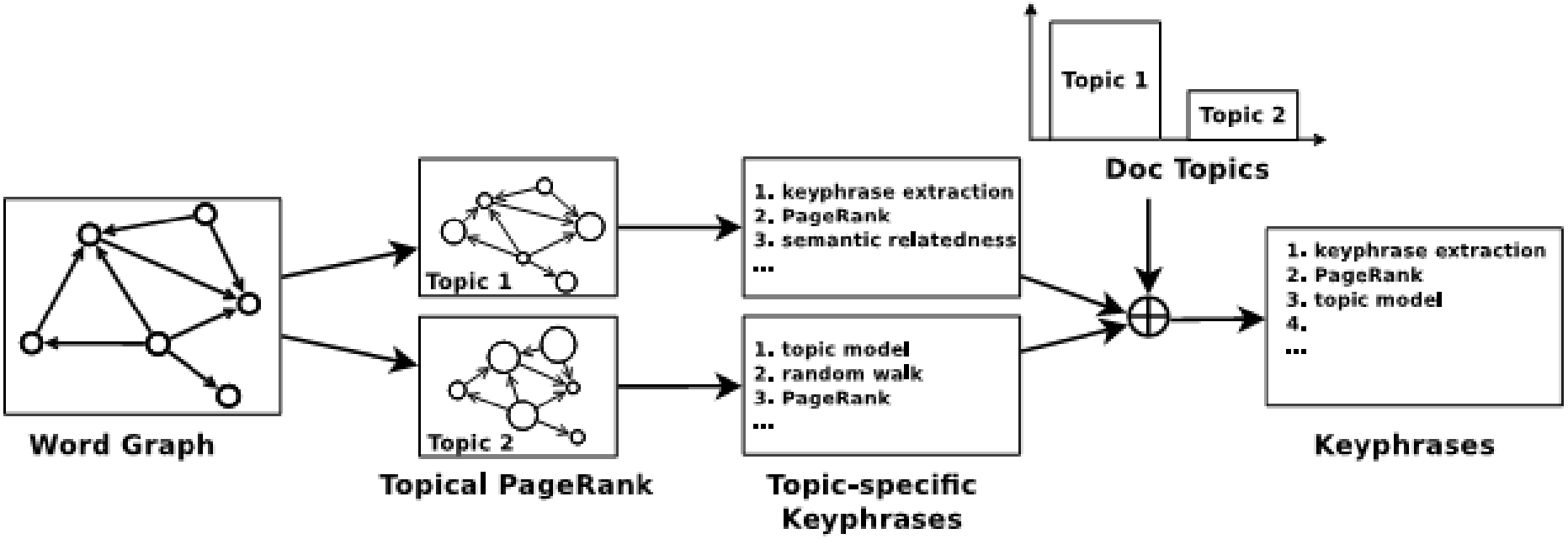
方法	Precision	Recall	F_1-Measure
TFIDF	0.333	0.173	0.227
TextRank	0.330	0.171	0.225
RESEARCH,cos,$K = 50$	0.343	0.178	0.234
Wiki,PL,$K = 1500$	0.349	0.181	0.238

综合利用隐含主题模型和文档结构进行关键词抽取

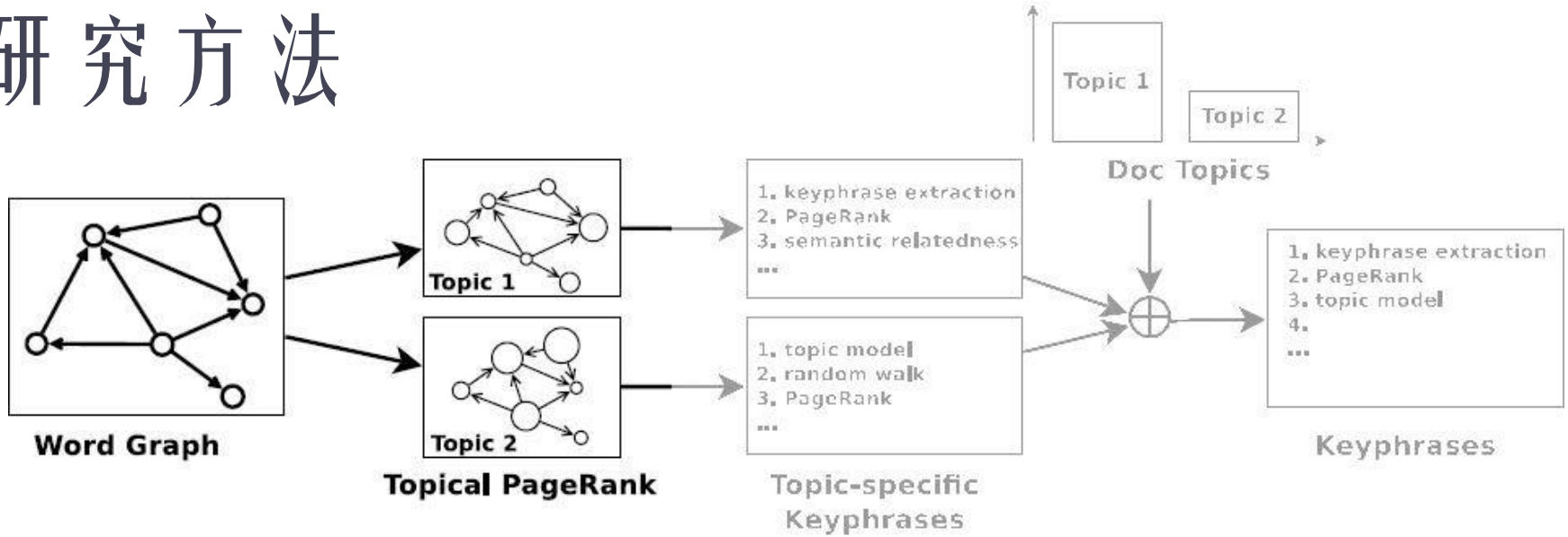
研究思路

- 前述工作
  - LDA：利用隐含主题模型发现文档主题
  - TextRank：利用文档内部结构信息
- 综合考虑文档主题和内部结构进行关键词抽取
  - Topical-PageRank（TPR）

研究方法


Topic 1
Topic 2
Doc Topics
1. keyphrase extraction
2. PageRank
3. semantic relatedness
...
1. keyphrase extraction
2. PageRank
3. topic model
4.
...
Topic 1
Topic 2
1. topic model
2. random walk
3. PageRank
...
Word Graph
Topical PageRank
Topic-specific Keyphrases
Keyphrases

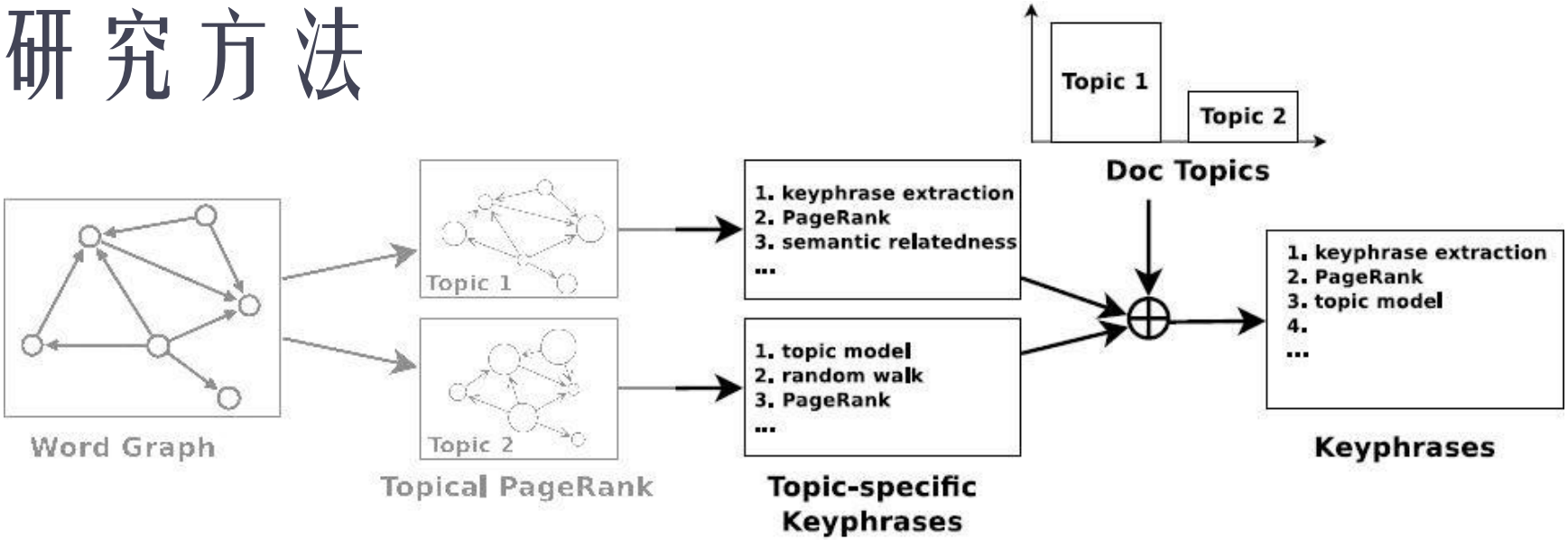
研究方法



$$R_z(w_i) = \lambda \sum_{j:w_j \to w_i} \frac{e(w_j, w_i)}{O(w_j)} R_z(w_j) + (1-\lambda) p_z(w_i)$$

- $p_z(w_i) = P(w|z)$, probability of word w given topic z.
- $p_z(w_i) = P(z|w)$, probability of word z given topic w.
- $p_z(w_i) = P(w|z) \times P(z|w)$, product of hub and authority.

研究方法



Candidate Phrases	noun phrases (Hulth, 2003) (adjective)*(noun)+
Doc topic distribution	$P(z\|d)$ for each topic z
Phrase Score	$R(p) = \sum_{z=1}^{K} R_z(p) \times P(z\|d)$

示例

(a) Topic on “Terrorism”

(b) Topic on “Israel”

(c) Topic on “U.S.”

(d) TPR Result

实验

- 实验数据
  - 新闻数据: 308 篇，来自DUC2001
  - 论文摘要: 2,000 篇，来自(Hulth, 2003)
- 评价指标
  - precision, recall, F–measure

$$p = \frac{C_{correct}}{C_{extract}}, r = \frac{C_{correct}}{C_{standard}}, f = \frac{2pr}{p + r}$$

  - binary preference measure (Bpref)

$$Bpref = \frac{1}{R}\sum_{r \in R} 1 - \frac{|n \text{ ranked higher than } r|}{M}$$

  - mean reciprocal rank (MRR)

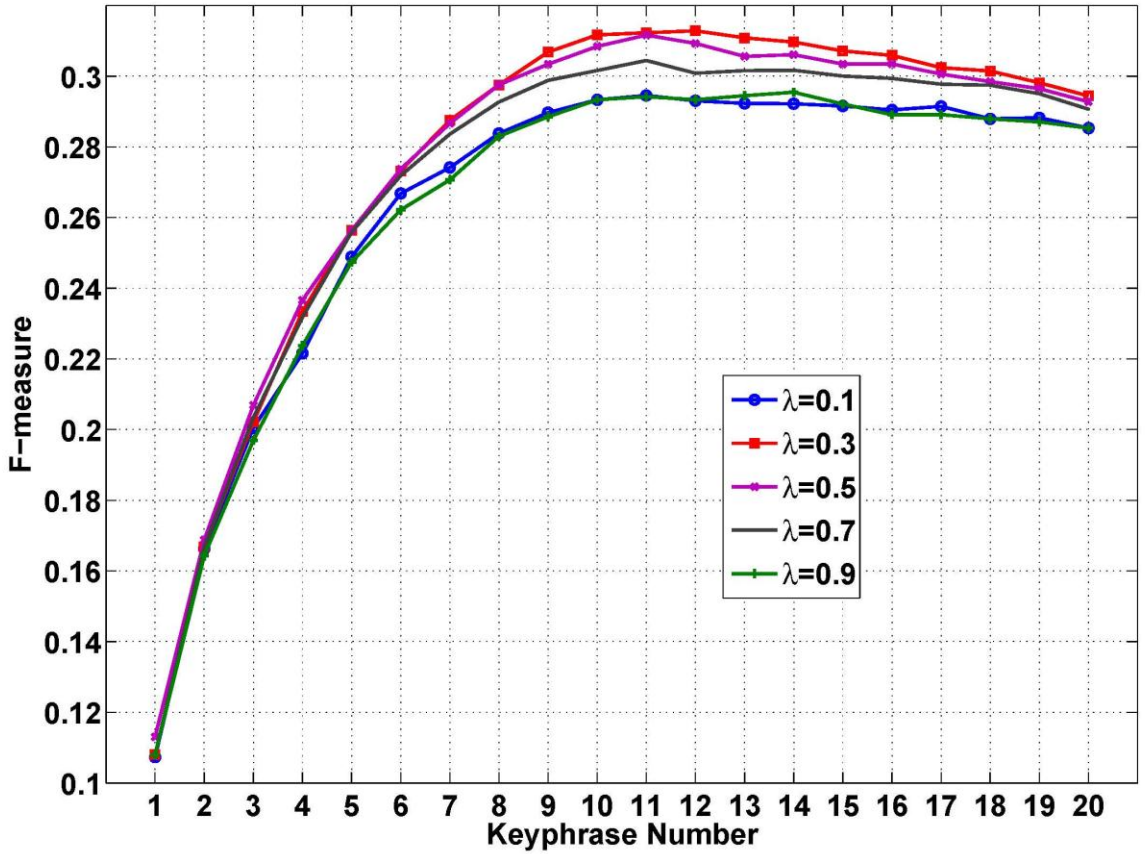
$$MRR = \frac{1}{|D|}\sum_{d \in D} \frac{1}{rank_d}$$

参数影响

K	Pre.	Rec.	F.	Bpref	MRR
50	0.268	0.330	0.296	0.204	0.632
100	0.276	0.340	0.304	0.208	0.632
500	0.284	0.350	0.313	0.215	0.648
1000	0.282	0.348	0.312	0.214	0.638
1500	0.282	0.348	0.311	0.214	0.631

新闻数据上LDA主题个数K影响（推荐$M = 10$个关键词）

参数影响



新闻数据上$\lambda = 0.1, 0.3, 0.5, 0.7$ and 0.9 的影响

不同偏好参数设置的影响

Pref	Pre.	Rec.	F.	Bpref	MRR
$pr(w\|z)$	0.256	0.316	0.283	0.192	0.584
$pr(z\|w)$	0.282	0.348	0.312	0.214	0.638
prod	0.259	0.320	0.286	0.193	0.587

新闻数据上不同偏好设置的影响（推荐$M = 10$个关键词）

与其他方法比较

Method	Pre.	Rec.	F.	Bpref	MRR
TFIDF	0.239	0.295	0.264	0.179	0.576
PageRank	0.242	0.299	0.267	0.184	0.564
LDA	0.259	0.320	0.286	0.194	0.518
TPR	**0.282**	**0.348**	**0.312**	**0.214**	**0.638**

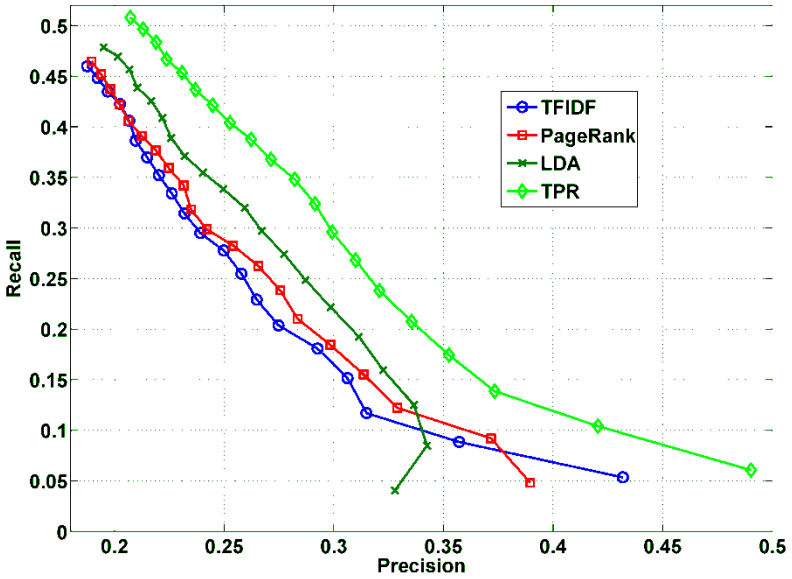
在论文摘要数据上的比较($M = 10$)

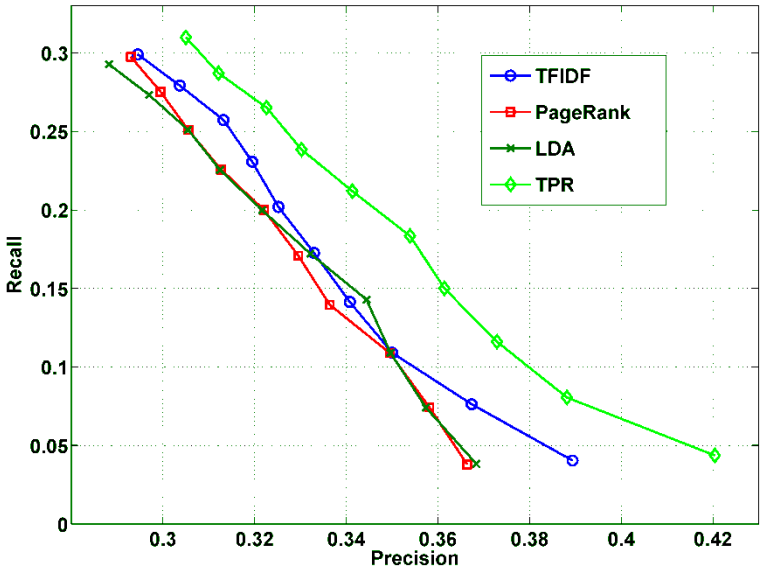
与其他方法比较

Method	Pre.	Rec.	F.	Bpref	MRR
TFIDF	0.333	0.173	0.227	0.255	0.565
PageRank	0.330	0.171	0.225	0.263	0.575
LDA	0.332	0.172	0.227	0.254	0.548
TPR	**0.354**	**0.183**	**0.242**	**0.274**	**0.583**

在论文摘要数据上的比较($M = 5$)

与其他方法比较





在新闻数据上，M 从1 到 20变化

在论文摘要数据上，M 从1 到 10变化

小结

- LDA通过文档主题进行关键词抽取，因此取得较TFIDF、TextRank较优的结果
- TPR综合了TextRank和LDA的优点，在两个数据集合上都表现出了它的优势
- 由于TPR可以按照主题推荐关键词，因此可以用于文档可视化，也可以用来进行查询导向（query focused）的关键词抽取

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, Maosong Sun. Automatic Keyphrase Extraction via Topic Decomposition. The Conference on Empirical Methods in Natural Language Processing (EMNLP), 2010.

利用机器翻译模型进行关键词抽取

研究问题

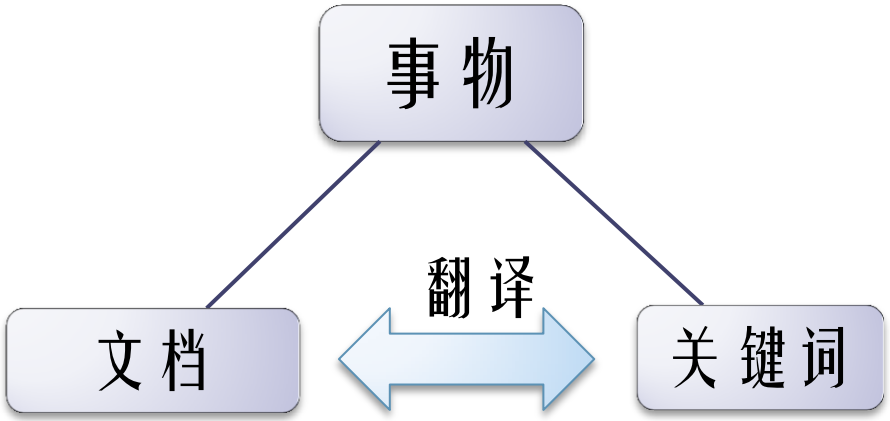
- 文档和关键词都是对同一事物的描述
  - 主题一致，词汇差异
- 词汇差异的表现
  - 很多关键词在文档中出现次数不高
  - 有的关键词在文档中根本没有出现（尤其是短文本）
- 问题
  - TFIDF、TextRank及其扩展、LDA等方法均没有很好解决词汇差异问题

相关工作

- TextRank的扩展ExpandRank
  - 在构建词网时，同时考虑文档的近邻文档
  - 从“文档层次（document level）”利用外部信息
  - 容易引入噪音
- LDA
  - 通过主题分布的相似度来对候选关键词排序
  - 从“主题层次（topic level）”利用外部信息
  - 由于主题一般是粗粒度的
    - 倾向于推荐普通词
    - 容易发生主题漂移

研究思路

- 在“词汇层次（word level）”利用外部信息
- 文档和关键词是对同一事物的描述
- 关键词抽取问题➔翻译问题



研究方法

- 构建翻译对（translation pairs）
- 学习两种语言间词汇的翻译概率（translation probabilities）$P(w_k|w_d)$
  - 利用SMT中的词对齐（word alignment）算法
- 给一个新的文档d
  - 计算每个候选关键词p的似然概率

$$P(p|d) = \sum_{i \in p} \sum_{j \in d} P(w_i|w_j) P(w_j|d)$$

  - 按照候选关键词的值进行排序

研究方法-构建翻译对集合

- 将文档标题或摘要看作近似用关键词语言写成
  - 大部分文档有标题或摘要信息
  - 将标题/摘要与文档正文形成翻译对
- 问题
  - 摘要、文档往往较长
    - 直接使用词对齐算法效率较低、效果较差
  - 没有标题/摘要的时候怎么办

研究方法-构建翻译对集合

- 给定标题和文档，提出两种构建翻译对的办法
  - 采样法（sampling）：将较长的文档进行抽样，直到与标题长度一致
    - 基于词在文档中的重要性（TFIDF）进行采样
  - 分割法（split）：将较长文档划分为句子，用每句话与标题构成一个翻译对
    - 只有句子与标题之间相似度大于某个阈值δ才放入训练集

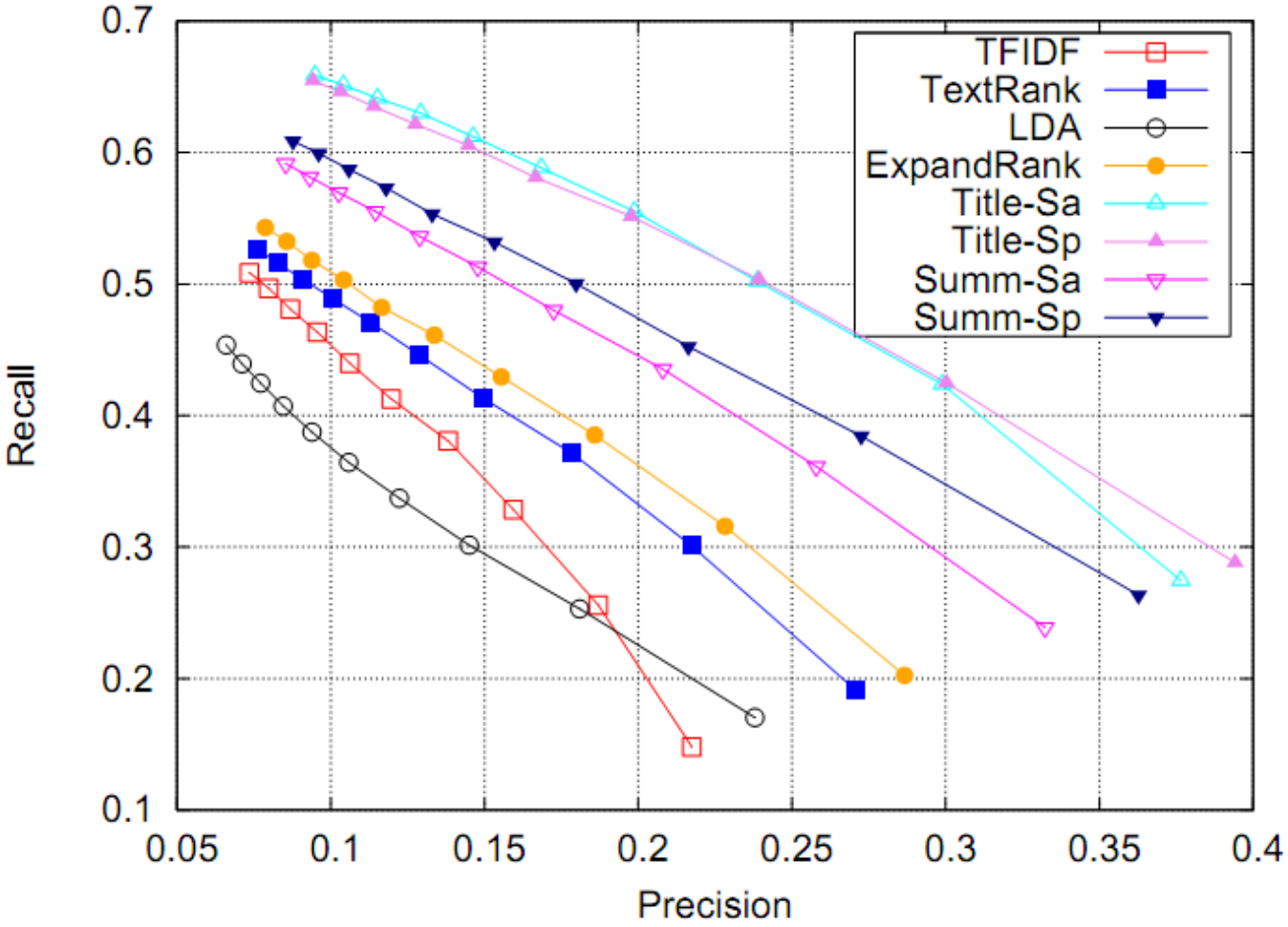
研究方法-构建翻译对集合

- 当没有标题或摘要，从文档正文中选择重要的句子来与正文构成翻译对
  - 选择文档第一句话
  - 选择与文档最相关的一句话

实验设置

- 词对齐算法采用IBM Model-1的工具GIZA++
- 在13,702篇中文新闻上进行试验

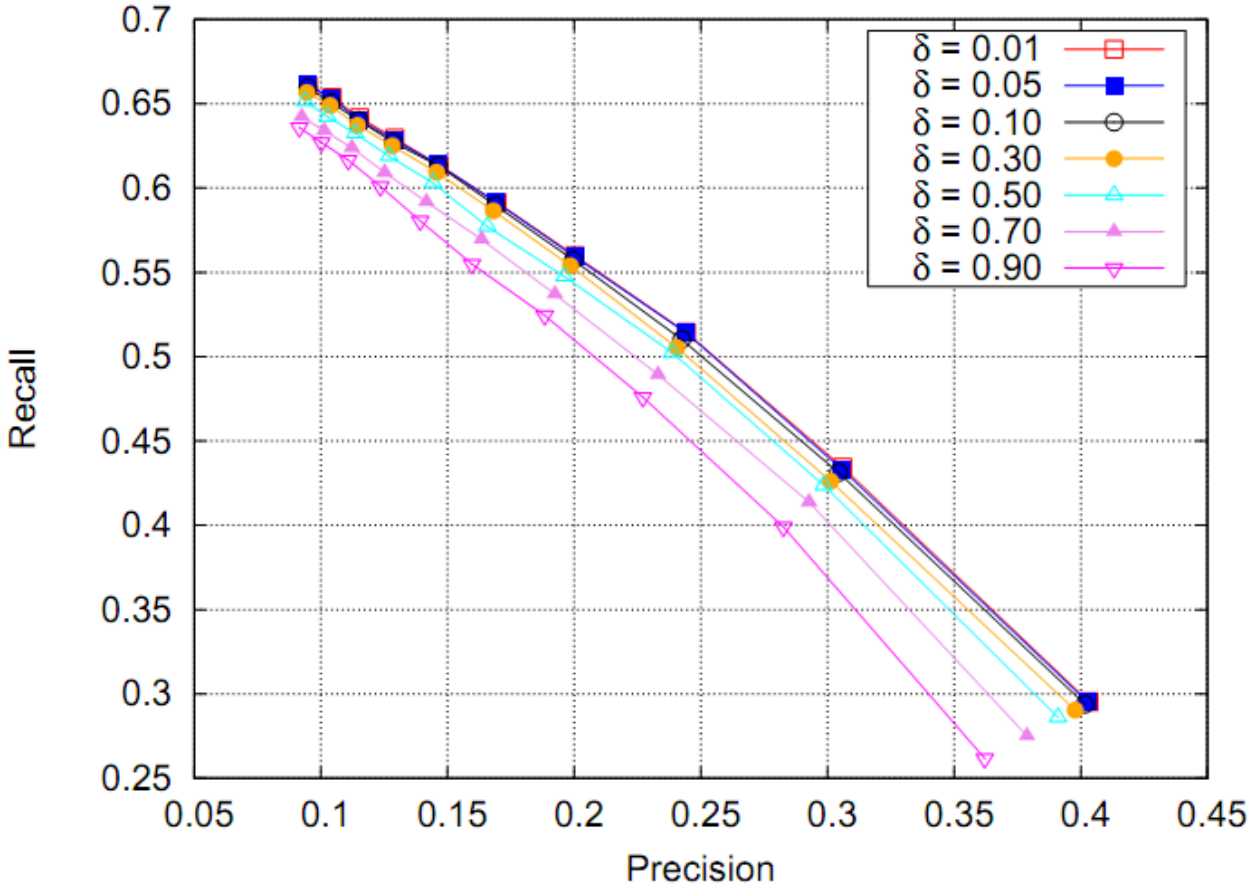
实验结果


0.7
0.6
0.5
0.4
0.3
0.2
0.1
Recall
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Precision
TFIDF
TextRank
LDA
ExpandRank
Title-Sa
Title-Sp
Summ-Sa
Summ-Sp

实验结果

Method	Precision	Recall	F-measure
TFIDF	0.187	0.256	0.208±0.005
TextRank	0.217	0.301	0.243±0.008
LDA	0.181	0.253	0.203±0.002
ExpandRank	0.228	0.316	0.255±0.007
Title-Sa	0.299	0.424	0.337±0.008
Title-Sp	**0.300**	**0.425**	**0.339±0.010**
Summ-Sa	0.258	0.361	0.289±0.009
Summ-Sp	0.273	0.384	0.307±0.008

实验结果-阈值δ的影响


δ = 0.01
δ = 0.05
δ = 0.10
δ = 0.30
δ = 0.50
δ = 0.70
δ = 0.90
Recall
Precision
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45

实验结果-抽取重要句子构建翻译对

Method	Precision	Recall	F-measure
First	0.290	0.410	0.327±0.013
Importance	0.260	0.367	0.293±0.010

实验结果-关键词生成（keyword Generation）

- 在测试时，只能够根据新闻标题产生关键词

Method	Precision	Recall	F-measure
TFIDF	0.105	0.141	0.115±0.004
TextRank	0.107	0.144	0.118±0.005
LDA	0.180	0.256	0.204±0.008
ExpandRank	0.194	0.268	0.216±0.012
WAM	**0.296**	**0.420**	**0.334±0.009**

实验结果-关键词生成举例

- 文档题目：“以军方称伊朗能造核弹 可能据此对伊朗动武”

方法	推荐关键词
标准答案	"核武器","以色列","伊朗"
SMT	"伊朗","动武","以军","以色列","军事","核武器"
TFIDF	"伊朗","动武","核弹","以军","据此 “
TextRank	"伊朗","可能","据此","核弹","动武"
LDA	"伊朗","美国","谈判","以色列","制裁"
ExpandRank	"伊朗","以色列","黎巴嫩","美国","以军"

小结

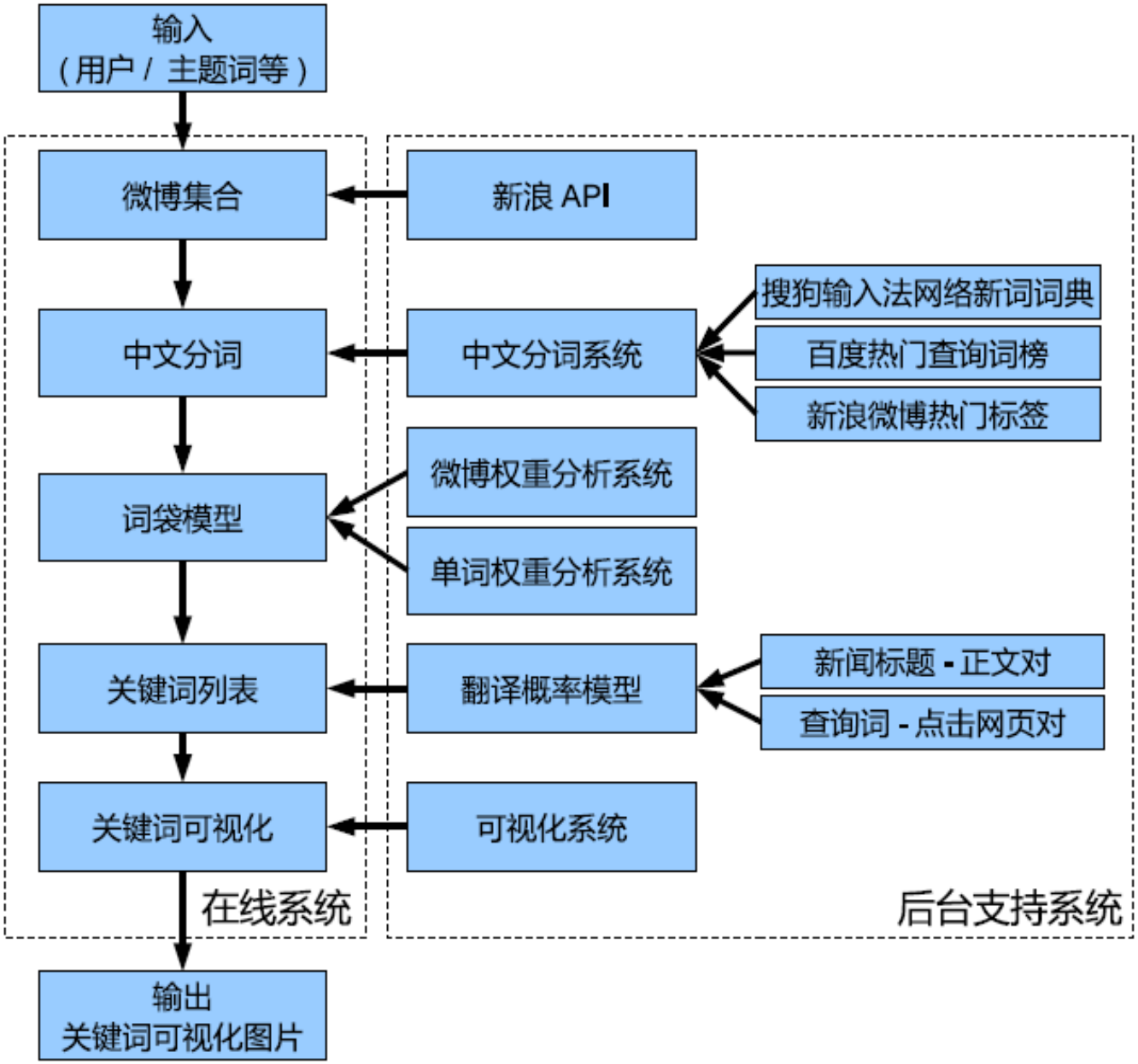
- 机器翻译技术可以有效解决词汇差异问题
  - 推荐更符合文档主题的关键词
  - 甚至能够胜任关键词生成任务
- 标题/摘要与文档能够构建高质量的翻译对
  - 对于新闻文档而言，正文第一句也可以用来构建高质量翻译对

Zhiyuan Liu, Xinxiong Chen, Yabin Zheng, Maosong Sun. Automatic Keyphrase Extraction by Bridging Vocabulary Gap. The 15th Conference on Computational Natural Language Learning (CoNLL'11).

典型应用：微博关键词抽取

应用简介

- 以新浪微博为平台
- 利用关键词抽取技术获取用户发表微博的关键词
- 应用前景
  - 发现和建模用户兴趣
  - 为用户之间链接赋予更丰富信息
  - 推荐用户感兴趣的产品、信息和好友等
  - 具有广阔的商业前景


输入
（用户 / 主题词等）
微博集合
新浪 API
中文分词
中文分词系统
搜狗输入法网络新词词典
百度热门查询词榜
新浪微博热门标签
词袋模型
微博权重分析系统
单词权重分析系统
关键词列表
翻译概率模型
新闻标题 - 正文对
查询词 - 点击网页对
关键词可视化
可视化系统
在线系统
后台支持系统
输出
关键词可视化图片

应用界面

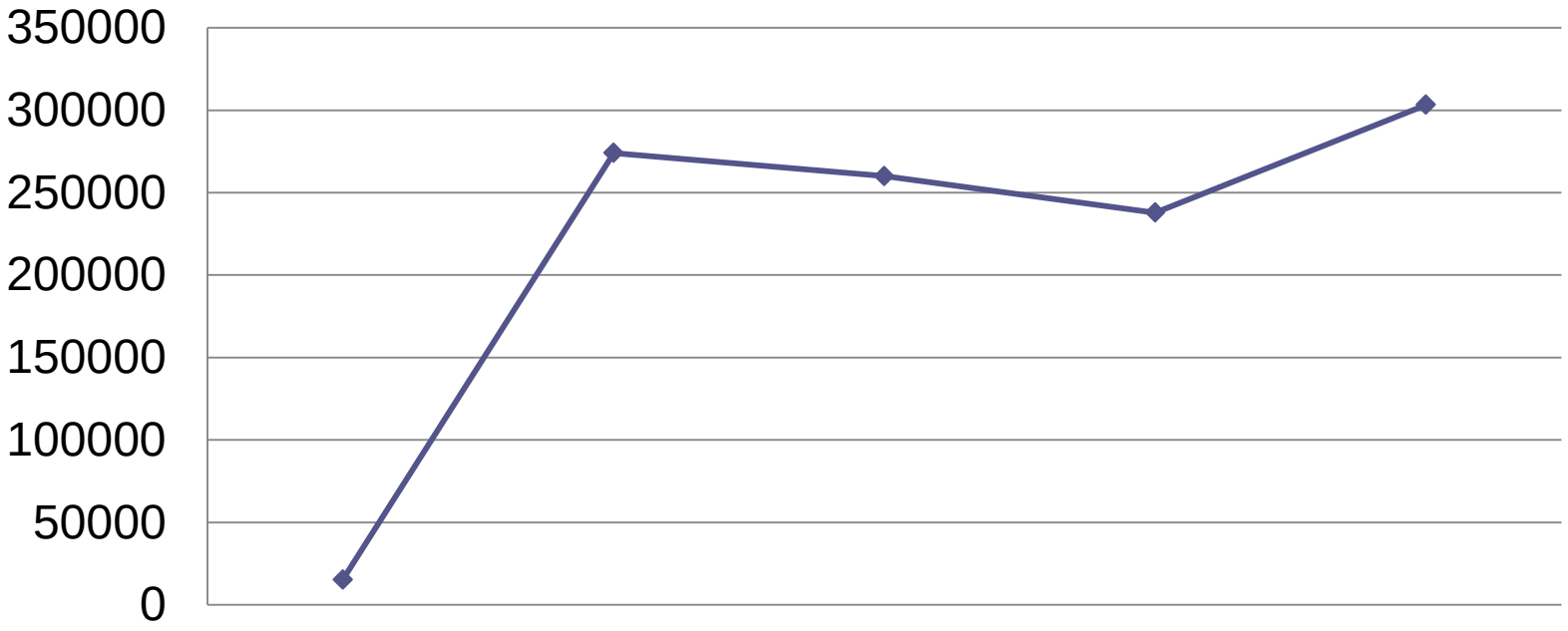

我的首页 新浪微博-随时...
Zhiyuan Liu's Homepage
166.111.138.15:8080/w...
166.111.138.15:8080/weibo/genImage.jsp
cfp
persons
research
jcst
VW
Lymbix
Pingar
emnlp
chromext
其他书签
我的微博关键词
Ta的微博关键词
主题的微博关键词
好友关键词
新闻关键词
我的微博关键词
以最近200条微博产生
词汇
批评
简单
翻译
搜索引擎
蔡康永
三国
喜欢
学习
2008
杂志
交互
记者
微软
讲座
凤凰卫视
真相
产品
舞蹈
人文
分析
思考
设计
字体
学习
英语
计算机
机器
企业家
商业
人物
生命
广告
生活
真实
技术
清华
人生
apple
创业
科学
时间
政治
世界
艺术
成功
google
丰子恺
南方周末
母亲
基金
李敖
人工智能
统计
电影
未来
女人
社会
复杂
博物馆
搜索
方法
工具
图书馆
组织

关键词抽取举例-我的微博关键词

关键词抽取举例-马少平老师的微博关键词

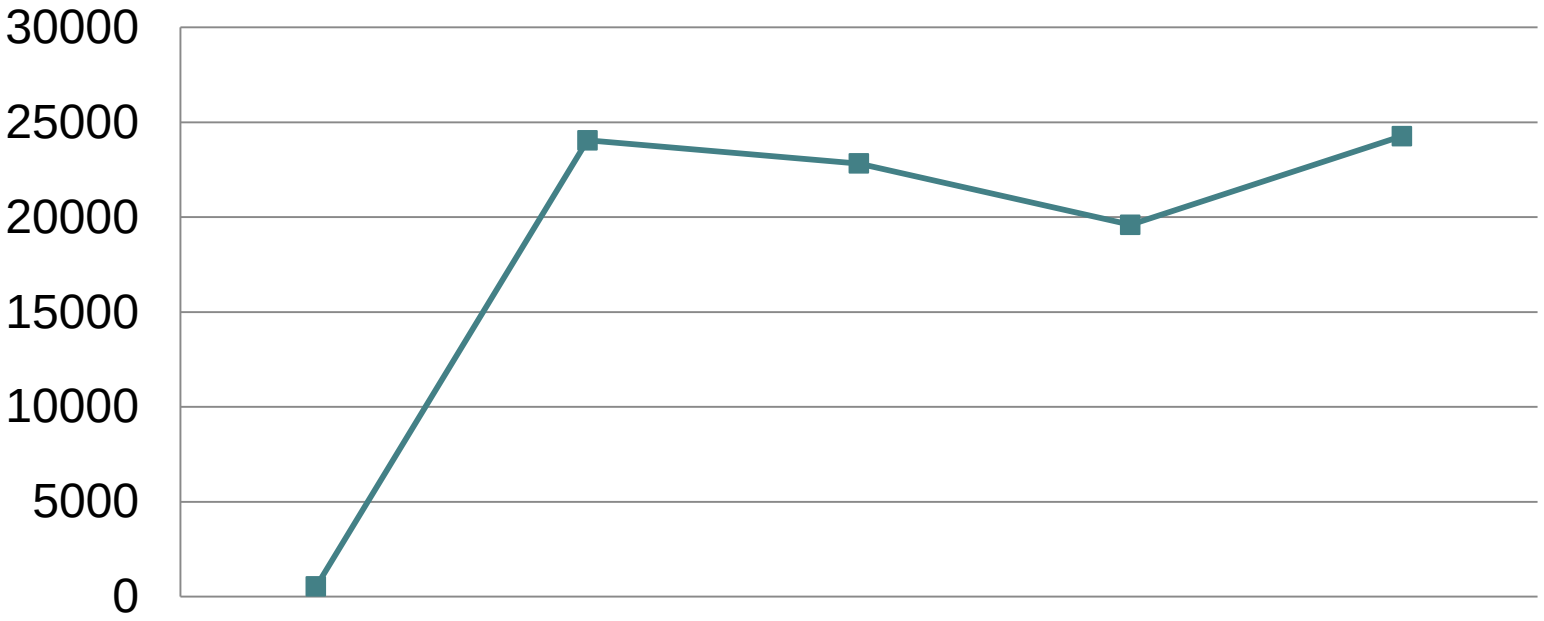
关键词抽取举例-MSRA的微博关键词

应用使用情况-接口调用数



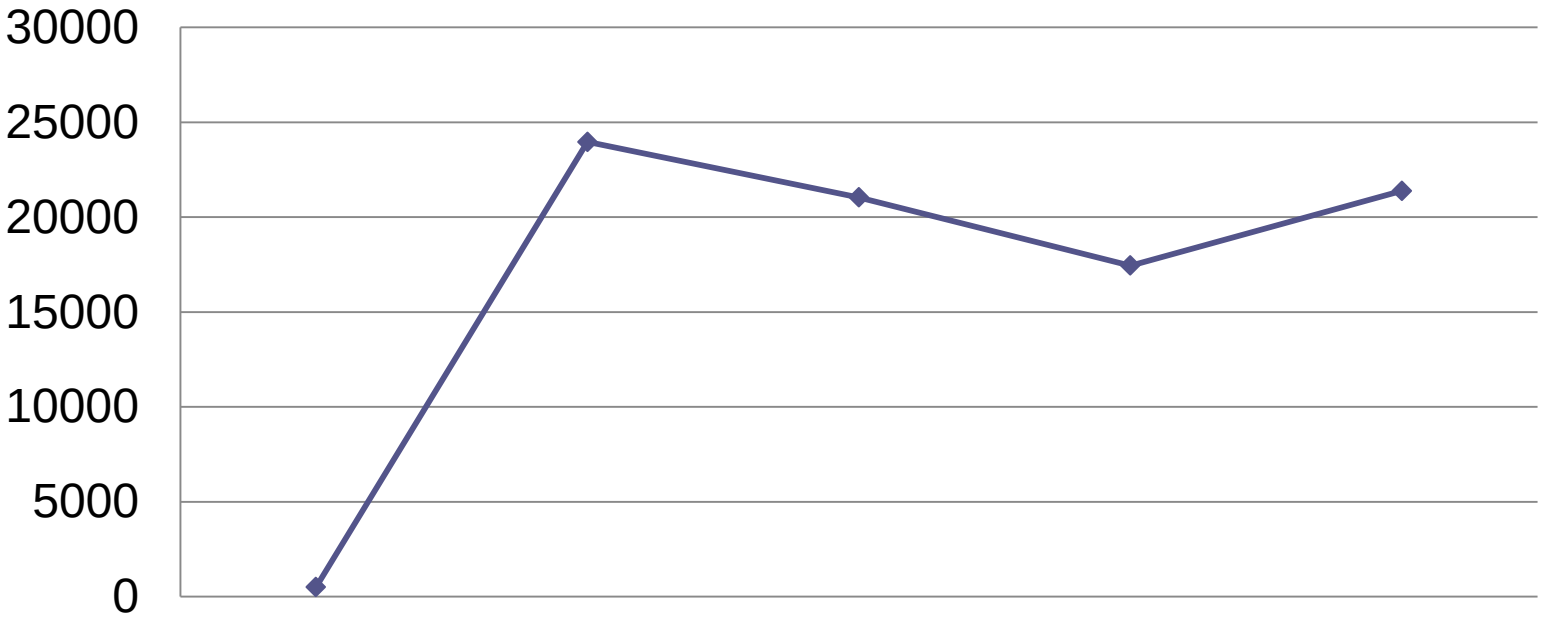
	5月25日	5月26日	5月27日	5月28日	5月29日
接口调用数	15201	274089	260023	237878	303315

应用使用情况-使用用户量



	5月25日	5月26日	5月27日	5月28日	5月29日
使用用户量	526	24047	22826	19599	24273

应用使用情况-新增用户数



	5月25日	5月26日	5月27日	5月28日	5月29日
新增用户数	501	23964	21043	17448	21381

应用使用情况-统计概览(5.25-5.29)

接口调用总次数	最近一周总用户量	最近一个月总用户量	累计总用户量
1099,979	84,427	84,626	84,626

小结

- 系统受到了微博用户的普遍认可
- 微博关键词抽取系统验证了本文对于基于文档主题结构关键词抽取研究的有效性
- 不足：交互机制

研究总结

- 利用文档主题结构对关键词抽取覆盖度的作用进行了深入研究
  - 通过文档内词聚类构建文档主题
  - 通过隐含主题模型构建文档主题
    - 提出隐含主题模型的高效并行算法
  - 综合考虑隐含主题和文档结构
- 以文档-关键词主题一致性为基础，提出基于机器翻译模型的算法，解决关键词抽取的词汇差异问题
- 以该研究为基础的微博关键词抽取系统在新浪微博上取得成功

未来工作与展望

- 实现一个高效实用的（中文）关键词抽取系统
- 关键词抽取在社会标签自动推荐中的应用
  - 解决冷启动问题：新标签、新对象、新用户
- 关键词抽取在Web数据中的应用
  - 用户兴趣建模和基于内容的推荐系统
  - 趋势检测和分析

主要发表论文

1. **Zhiyuan Liu**, Xinxiong Chen, Maosong Sun. A Simple Word Trigger Method for Social Tag Suggestion. The Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011.
2. **Zhiyuan Liu**, Xinxiong Chen, Yabin Zheng, Maosong Sun. Automatic Keyphrase Extraction by Bridging Vocabulary Gap. The 15th Conference on Computational Natural Language Learning (CoNLL), 2011.
3. **Zhiyuan Liu**, Yabin Zheng, Lixing Xie, Maosong Sun, Liyun Ru. User Behaviors in Related Word Retrieval and New Word Detection: A Collaborative Perspective. ACM Transactions on Asian Language Information Processing (ACM TALIP) (Special Issue on Chinese Language Processing), 2011.
4. **Zhiyuan Liu**, Yuzhou Zhang, Edward Y. Chang, Maosong Sun. PLDA+: Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing. ACM Transactions on Intelligent Systems and Technology (ACM TIST), 2010.
5. **Zhiyuan Liu**, Wenyi Huang, Yabin Zheng, Maosong Sun. Automatic Keyphrase Extraction via Topic Decomposition. The Conference on Empirical Methods in Natural Language Processing (EMNLP), 2010.
6. **Zhiyuan Liu**, Peng Li, Yabin Zheng, Maosong Sun. Clustering to Find Exemplar Terms for Keyphrase Extraction. The Conference on Empirical Methods in Natural Language Processing (EMNLP), 2009.

主要发表论文

7. **Zhiyuan Liu**, Maosong Sun. Domain–Specific Term Rankings Using Topic Models. The Sixth Asia Information Retrieval Society Conference (AIRS), 2010.
8. **Zhiyuan Liu**, Chuan Shi, Maosong Sun. FolkDiffusion: A Graph–based Tag Suggestion Method for Folksonomies. The Sixth Asia Information Retrieval Society Conference (AIRS), 2010.
8. **Zhiyuan Liu**, Yabin Zheng, Maosong Sun. Quantifying Asymmetric Semantic Relations from Query Logs by Resource Allocation. The 13th Pacific–Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2009.
9. **Zhiyuan Liu**, Maosong Sun. Asymmetrical Query Recommendation Method Based on Network–resource–allocation Dynamics. The 17th International World Wide Web Conference (WWW), 2008. **刘知远**, 郑亚斌, 孙茂松. 汉语依存句法网络的复杂网络性质. 复杂系统与复杂性科学, Vol. 5, No. 2, pp. 37–45, 2008.
10. **刘知远**, 孙茂松. 汉语词同现网络的小世界效应和无标度特性. 中文信息学报, Vol. 21, No. 6, pp. 52–57, 2007.
11. **刘知远**, 司宪策, 郑亚斌, 孙茂松. 中文博客标签的若干统计性质. 第七届中文处理国际会议 (ICCC), 2007.
12. **刘知远**, 孙茂松. 基于WEB的计算机领域新术语的自动检测. 第九届全国计算语言学学术会议 (CNCCL), 2007.

申请专利

1. 国内专利. 第二发明人.获取新词的方法和装置. 申请号: 200910083143.2. 公开号: CN101539940.
2. 国际专利. 第二发明人. Category-Sensitive Ranking for Text. 申请号: PCT/CN2009/001584.
3. 国际专利. 第一发明人. Parallel Generation of Topics from Documents. 申请中.

谢谢各位老师！请提出宝贵意见!

LDA学习算法

- Gibbs Sampling

其他位置上的词w的主题分布

该文档其他位置上词的主题分布

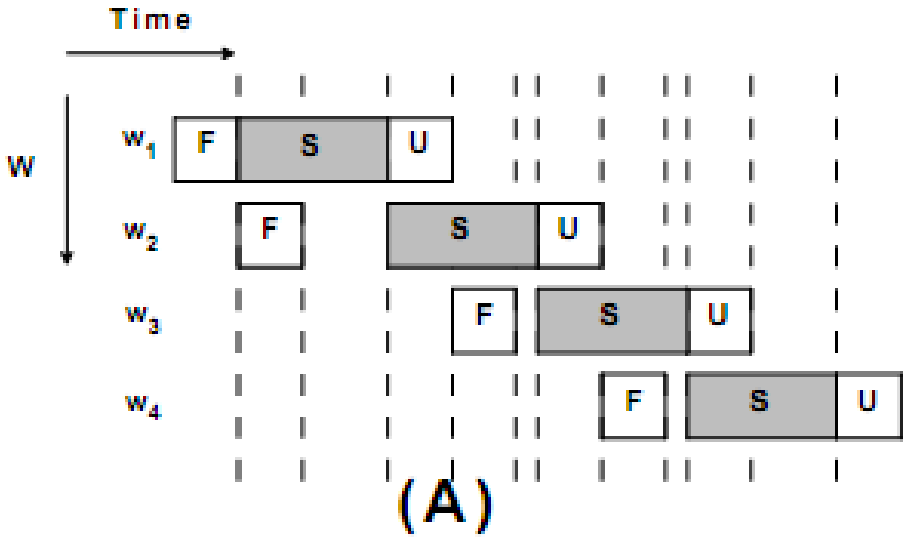
$$P\left(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot\right) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W\beta} \cdot \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^{T} C_{d_i t}^{DT} + T\alpha}$$

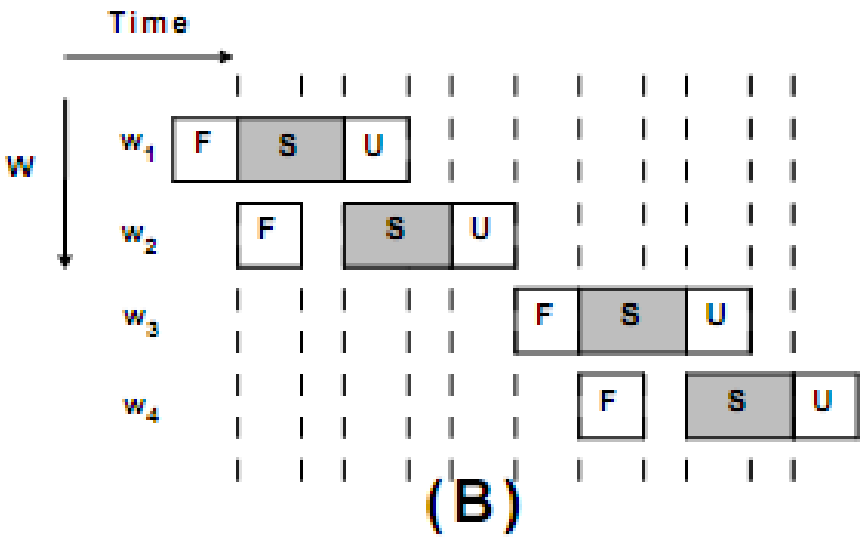
$$\phi_i'^{(j)} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\beta}$$

$$\theta_j'^{(d)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha}$$

PLDA+算法

- 基于新结构的Pipeline-based Gibbs Sampling





复杂度分析

Method	Time Complexity		Space Complexity
	Preprocessing	Gibbs sampling	
LDA	-	INK	$K(D+W)+N$
PLDA	$\frac{D}{\lvert P\rvert}$	$I\left(\frac{NK}{P}+cKW\log P\right)$	$\frac{(N+KD)}{P}+KW$
PLDA+,P_d	$\frac{D}{\lvert P_d\rvert}+cW\log W+\frac{WK}{\lvert P_w\rvert}$	$\frac{INK}{\lvert P_d\rvert}$	$\frac{(N+KD)}{\lvert P_d\rvert}$
PLDA+,P_w	-	-	$\frac{KW}{\lvert P_w\rvert}$

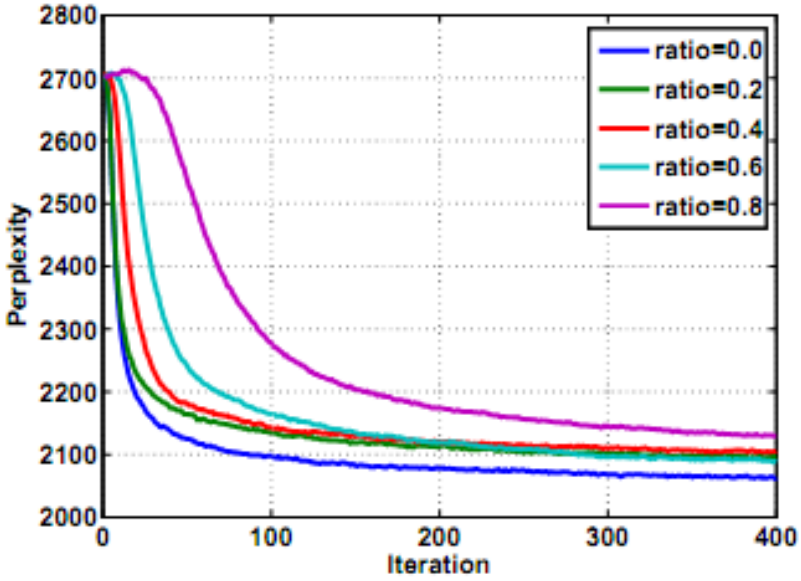
实验设置

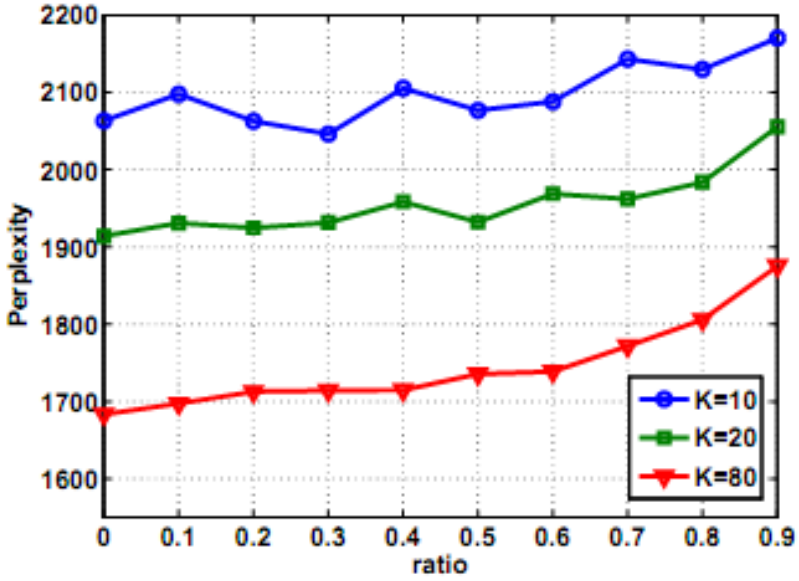
- 数据集合

	NIPS	Wiki-20T	Wiki-200T
D_{train}	1,540	2,122,618	2,122,618
W	11,909	20,000	200,000
N	1,260,732	447,004,756	486,904,674
D_{test}	200	-	-

实验效果-Missing ratio

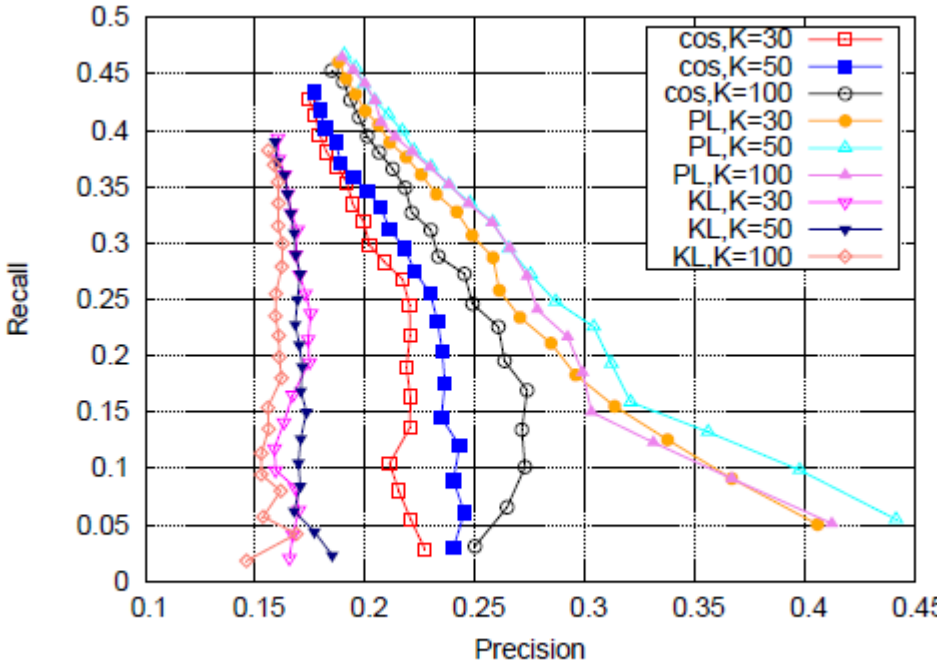
- Missing ratio与迭代次数和主题个数之间的关系



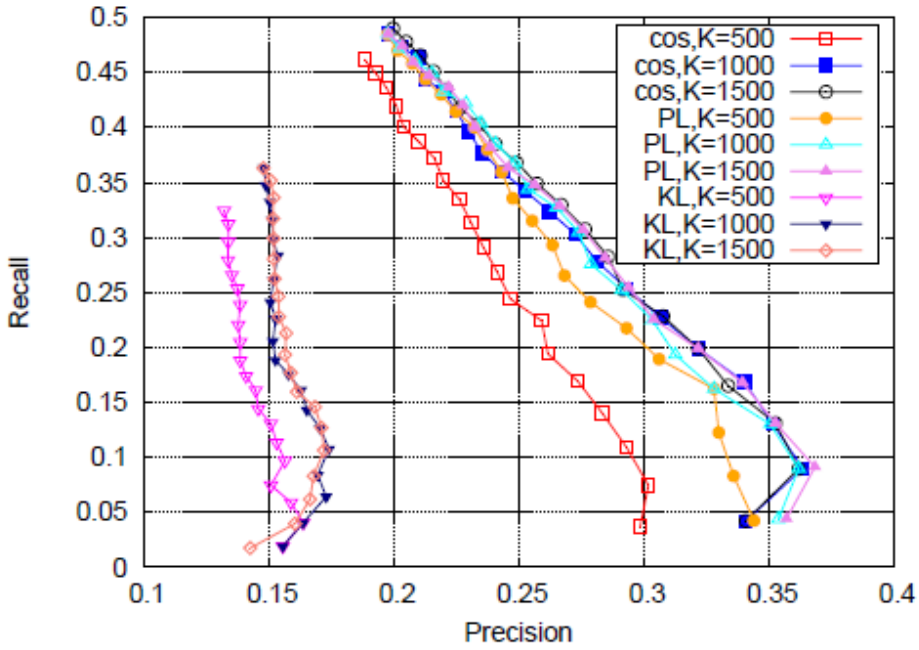


NEWS数据

- LDA分别在NEWS训练和在Wikipedia上训练



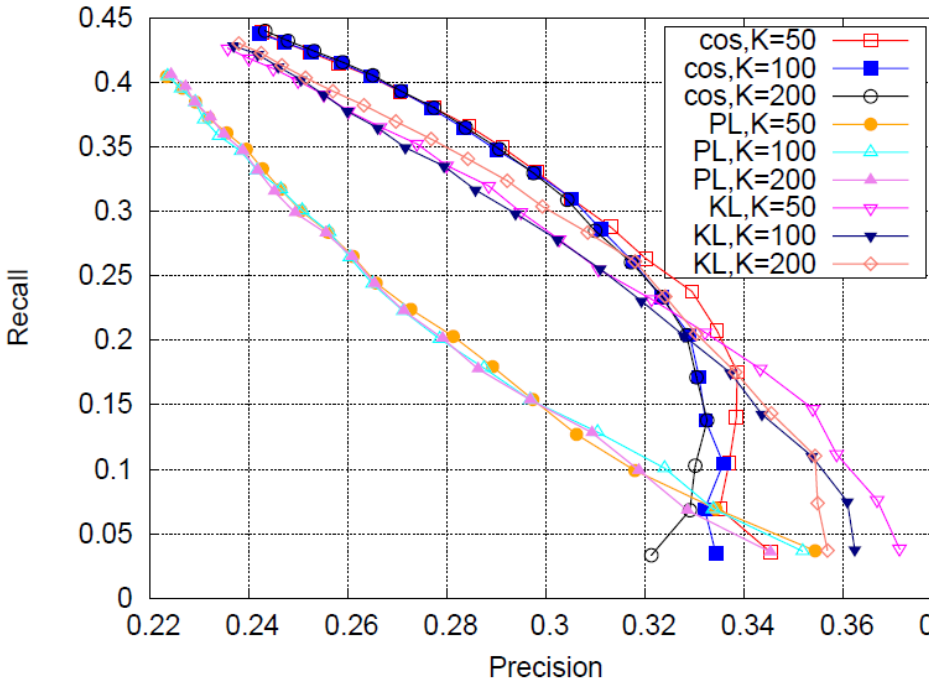
(a) 在NEWS数据集合学习LDA模型



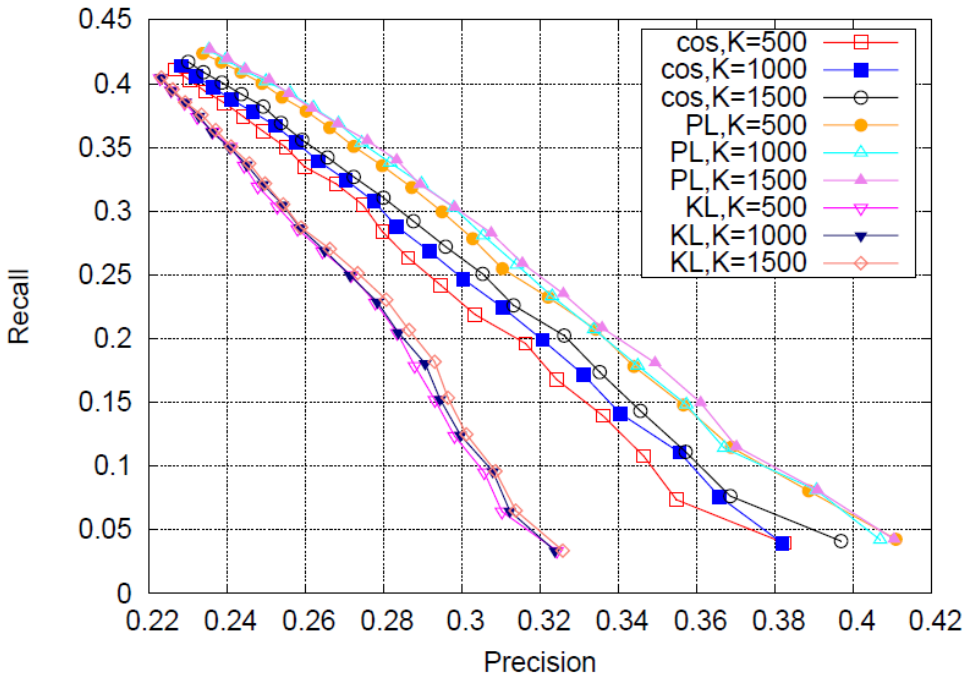
(b) 在Wikipedia数据集合学习LDA模型

RESEARCH数据

- LDA分别在RESEARCH训练和在Wikipedia上训练



(a) 在RESEARCH数据集合学习LDA模型



(b) 在Wikipedia数据集合学习LDA模型