An Introduction to Text Mining and Related Research Topics

Language modeling approach

Xin Zhao

Department of Computer Science

Outline

Google

Google

Google 北京大学 Google 搜索 高级

网页 打开百宝箱... 搜索 **北京大学** 获得约 **7,660,000** 条结果，以下是第 **1-10** 条。（用时 **0.13** 秒）

北京大学
www.pku.edu.cn

北京市海淀区颐和园路5号
010-62751032
获取行车路线 - 信息是否正确？

22 个评论
菜单及其他信息 »

欢迎访问北京大学主页
北京大学作为国内前茅的文理医工综合性大学，在培养高素质创新型人才、取得突破性科研进展，以及及为国民经济发展和社会进步提供智力支持等方面都发挥着极其重要的作用。
www.pku.edu.cn/ - 网页快照 - 类似结果

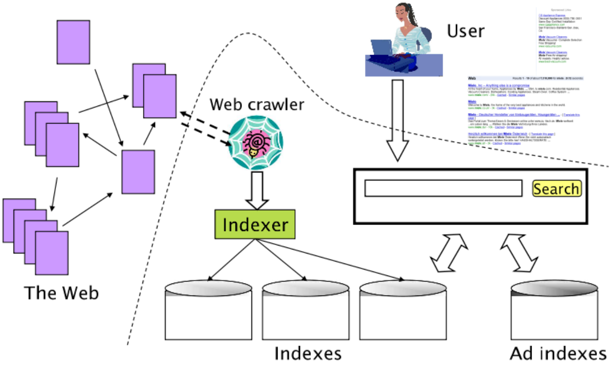
院系设置	本科生课程设置
招生信息	校内门户
北大研究生院	人才招聘
北大概况	北京大学英语系

pku.edu.cn站内的其它相关信息 »

北京大学招生网
由北京大学招生办公室主办，致力于全方位传递北大本科招生信息和与北大有关的各类资讯。
www.gotopku.com/ - 网页快照 - 类似结果

北京大学研究生院
北京大学研究生院站内搜索（请输入关键词）：... 地址：北京大学才斋(红二楼) 邮编：100871 维护：张林更新：2009年06月08日 ...
grs.pku.edu.cn/ - 网页快照 - 类似结果

Various components of a search engine



How to retrieve documents with an inverted index?(1)

- Given a document collection
  -
D1	Search engine is a quite powerful tool.
D2	Text mining techniques can be applied to search engines.
D3	Text mining aims to analyze documents better.

How to retrieve documents with an inverted index?(2)

- The respective inverted index:

search	D1:1	D2:1	
engine	D1:1	D2:1	
powerful	D1:1		
tool	D1:1		
text		D2:1	D3:1
mining		D2:1	D3:1
technique		D2:1	
apply		D2:1	
aim			D3:1
analyze			D3:1
well			D3:1

How to retrieve documents with an inverted index?(3)

search	*D1:1*	*D2:1*	
engine	*D1:1*	*D2:1*	
powerful	D1:1		
tool	D1:1		
text		D2:1	D3:1
mining		D2:1	D3:1
technique		D2:1	
apply		D2:1	
aim			D3:1
analyze			D3:1
well			D3:1

- query="search" and "engine"
  - Relevant docs={D1, D2}

How to retrieve documents with an inverted index?(4)

search	D1:1	D2:1	
engine	D1:1	D2:1	
powerful	D1:1		
tool	D1:1		
text		*D2:1*	*D3:1*
mining		*D2:1*	*D3:1*
technique		D2:1	
apply		D2:1	
aim			D3:1
analyze			D3:1
well			D3:1

- query="text" and "mining"
  - Relevant docs={D2, D3}

How to retrieve documents with an inverted index?(5)

search	*D1:1*	*D2:1*	
engine	*D1:1*	*D2:1*	
powerful	D1:1		
tool	D1:1		
text		*D2:1*	*D3:1*
mining		*D2:1*	*D3:1*
technique		D2:1	
apply		D2:1	
aim			D3:1
analyze			D3:1
well			D3:1

- query="search" and "engine" and "text" and "mining"
  - Relevant docs={D2}

Lots of Interesting Research Problems for Search Engines

- crawl
- compression
- term weighting
- phrase search
- user interface
- personalization
- query suggestion
- evaluation
- ...

Outline

Basic Notations in Information Retrieval

Notations and Assumption

- Document collection: $C = \{D_1, D_2, ..., D_n\}$
- Vocabulary: $V = \{w_1, w_2, ..., w_n\}$
- Document: $D = \{w_1, w_2, ..., w_{N_d}\}, N_d = |D|$
- Query: $Q = \{q_1, ..., q_m\}$
- Assumption: *bag of words*

Outline

Statistical Language Models

Statistical Language Models

- A statistical language model (or just language model for short) is a probability distribution over word sequences.
- Can also be regarded as a probabilistic mechanism for "generating" text, thus also called a "generative" model.
- A piece of text can be regarded as a sample drawn according to this word distribution
- Examples:
  - $p(\text{“}\textit{Today is Wednesday}\text{”}) \approx 0.001$
  - $p(\text{“}\textit{Today Wednesday is}\text{”}) \approx 0.000000001$

Unigram Language Model

Unigram Language Models

- Unigram language model: $p(w_1, w_2, \ldots, w_n) = \prod p(w_i)$
  - Model parameters: $\theta = \{p(w|\theta)|w \in V\}$
  - Constraint: $1)\sum_{w \in V} p(w|\theta) = 1;\ 2)p(w|\theta) >= 0$
  - Essentially a multinomial distribution over words
  - Trade off between accuracy and complexity(data sparseness)
  - Examples:
    - $p(\text{“}\textit{Today is Wednesday}\text{”}) \approx 0.001$
    - $p(\text{“}\textit{Today Wednesday is}\text{”}) \approx 0.001$

Examples

D_1	Text	mining	is	an	interesting	research	field	.
D_2	A	good	diet	would	help	keep	healthy	.

- which model has higher probability to generate D_1?
- which model has higher probability to generate D_2?

$p(w|\theta_1)$

…
text 0.2
mining 0.1
assocation 0.01
clustering 0.02
…
food 0.00001
…

$p(w|\theta_2)$

…
food 0.25
nutrition 0.1
healthy 0.05
diet 0.02
…
text 0.00001
…

- Problem:Assume we have observed a document D, what's the underlying document language model? (with unigram language model assumption)

Estimating the document language model

- We view the observed document as a sample from the underlying document language model
- This is a standard problem in statistics:
  - Maximum likelihood (ML) : $p(w|\hat{\theta}_D) = \frac{C(w,D)}{|D|}$
  - Maximum A Posteriori (MAP) , such as *Dirichlet Prior*
- Smoothing: quite important for estimating document language model
  - Linear interpolation(Fixed Coefficient)
  - *Dirichlet Prior*
  - ...lots of research work

A simple example for ad-hoc retrieval

- Document collection

D_1	Text	mining	is	an	interesting	research	field	.
D_2	A	good	diet	would	helps	keep	healthy	.

- Query = "text mining"
- Which document is more relevant?

Query Likelihood Retrieval Model

- Assuming to use a multinomial language model, we would generate a sequence of words by generating each query word independently:
  - $$Score(D,Q) = P(Q|\theta_D) = \prod_{i=1}^{m} p(q_i|\theta_D) = \prod_{v=1}^{|V|} p(w|\theta_D)^{C(w,Q)}$$
  - It's "query generation" approach.
  - Other forms of language models can be also used:
    - Multiple Bernoulli
    - Multiple Possion

Kullback-Leibler Divergence Retrieval Model(1)

- We view documents and queries both as bag-of-words, since we use θ_D to model document language model, why not θ_Q?
- How to estimate θ_Q?
    - ML estimator
    - Model-Based feedback (Chengxiang Zhai et al. CIKM 2001)
        - $\theta_{Q'} = \lambda\theta_Q + (1-\lambda)\theta_F$
    - Relevance model (Victor Lavrenko et al. SIGIR 2001)
        - $\theta_R = \{p(w|Q, R = r)|w \in V\}$
    - ...

Kullback-Leibler Divergence Retrieval Model(2)

- It ranks the documents by computing the cross-entropy between θ_Q and θ_D
- $$Score(D,Q) = -D(\theta_Q||\theta_D) = -\sum_{v \in V} p(w|\theta_Q) log \frac{p(w|\theta_Q)}{p(w|\theta_D)}$$
- $$Score(D,Q) \stackrel{rank}{=} \sum_{v \in V} p(w|\theta_Q) logp(w|\theta_D)$$
- More flexiable than simple query likelihood retrieval model(QL)
  - In fact, QL retrieval model is a special case of KL divergence retrieval model

Outline

Introduction to Topic Models(1)

- $|V|$ can be quite a large number
- Retrieval is based on exact match
- So we seek:
  - a low-dimension semantic representation of text
    - it allows different words capturing the same semantic concept to match each other(automobile,car)
  - summarize search results through revealing the major topics in the results
    - each topic provides a coherent semantic dimension for representing text

Introduction to Topic Models(2)

Topic

- Each topic would be represented using a word distribution, or a unigram language model over the vocabulary

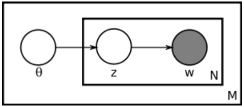
- Example[1]

$Topic_1$	$Topic_2$	$Topic_3$	$Topic_4$	$Topic_5$
coast	government	stars	laboratory	fig
sea	national	star	research	shown
fish	institution	observations	work	curve
island	scientific	observatory	room	curves

[1] from http://topics.cs.princeton.edu/Science/browser/

Probabilistic Latent Semantic Analysis(pLSA)(1)

Thomas Hofmann et al. 2001

- Plate notation representing the PLSA model
  - 

- Parameters of PLSA(multinomial distribution)
  - $p(w|z)$——topic˜word(coherent semantic topic)
  - $p(z|d)$——document˜topic(low dimension representation)
- The joint probability of (w,d)

$$p(w,d) = p(d)p(w|d)$$

$$p(w|d) = \sum_z p(w,z|d) = \sum_z p(w|z,d)p(z|d) = \sum_z p(w|z)p(z|d)$$

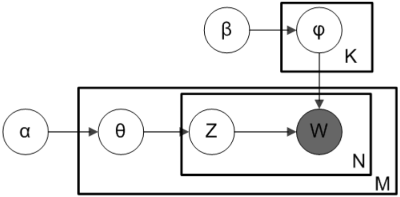
Probabilistic Latent Semantic Analysis(pLSA)(2)

Thomas Hofmann et al. 2001

- Parameters estimation(Quite insteresting parts for application)
  - ML estimator (EM)
  - MAP (If we know some topic should be ...) (Yue Lu et al. WWW 2008)
  - Estimation with regularization (in next pages...)
- Relationship with mixture of Gaussian model
- Relationship with mixture of unigram model

Latent Dirichlet Allocation (LDA)(1)

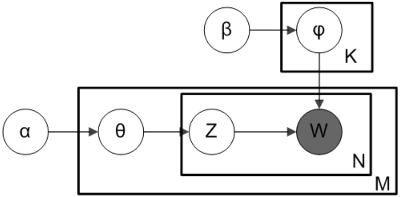
D. Blei et al. 2004

- Plate notation representing the LDA model
    - 

- Hierarchy Bayesian Model
    - $\varphi \approx p(w|z)$, $\theta \approx p(z|d)$
    - $w|z, \varphi_z \sim Multinomial(\varphi_z)$, $\varphi \sim Dirichlet(\beta)$
    - $z|\theta_d \sim Multinomial(\phi_d)$, $\theta \sim Dirichlet(\alpha)$

Latent Dirichlet Allocation (LDA)(2)

D. Blei et al. 2004

- Plate notation representing the LDA model
  - 

- Parameters of LDA
  - α,β
- Inference(from known to unknown)
  - Variational EM(D. Blei et al. 2004)
  - Gibbs sampling (T. L. Griffiths et al. PNAS 2004)
  - Expectation propagation

A summary of topic model

- A topic is just a distribution over vocabulary; LDA and pLSA model documents as mixture of topics
- Input of LDA and pLSA
  - Collection of documents{online(Loulwah AlSumait et al. ICDM 2008) and scalability(Yi Wang et al. AAIM 2009)}
  - Topic number(how to find the right number)
- Output of LDA and pLSA
  - $p(w|z)$——topic~word(coherent semantic topic)
  - $p(z|d)$——document~topic(low dimension representation)
- Evaluation(for recent work, see J. Chang et al. NIPS 2009)

$$perplexity(C_{test}) = exp\{-\frac{\sum_d logp(\vec{w}_d)}{\sum_d N_d}\}$$

An Example for Toipc Model

- Assume the text is : National scientific institution starts star observatory research work as the fig shown.
- After removing the stopwords:
  - national scientific institution star observatory research work fig shown
- Let's show the process of topic modeling on text

An Example for Toipc Model

- Word by word

nat.	sci.	ins.	star	obs.	res.	work	fig	shown
?	?	?	?	?	?	?	?	?

- Topic distribution for this document
0.05	0.3	0.25	0.2	0.2

- Reuse the topics from blei's website
$Topic_1$	$Topic_2$	$Topic_3$	$Topic_4$	$Topic_5$
coast	government	stars	laboratory	fig
sea	national	star	research	shown
fish	institution	observations	work	curve
island	scientific	observatory	room	curves

An Example for Toipc Model

- Word by word

nat.	sci.	ins.	star	obs.	res.	work	fig	shown
2	?	?	?	?	?	?	?	?

- Topic distribution for this document
0.05	**0.3**	0.25	0.2	0.2
- Reuse the topics from blei's website
$Topic_1$	$Topic_2$	$Topic_3$	$Topic_4$	$Topic_5$
coast	government	stars	laboratory	fig
sea	**national**	star	research	shown
fish	institution	observations	work	curve
island	scientific	observatory	room	curves

An Example for Toipc Model

- Word by word

nat.	sci.	ins.	star	obs.	res.	work	fig	shown
2	2	2	**3**	?	?	?	?	?

- Topic distribution for this document
0.05	0.3	**0.25**	0.2	0.2

- Reuse the topics from blei's website
$Topic_1$	$Topic_2$	$Topic_3$	$Topic_4$	$Topic_5$
coast	government	stars	laboratory	fig
sea	national	**star**	research	shown
fish	institution	observations	work	curve
island	scientific	observatory	room	curves

An Example for Toipc Model

- Word by word

nat.	sci.	ins.	star	obs.	res.	work	fig	shown
2	2	2	3	3	**4**	?	?	?

- Topic distribution for this document
0.05	0.3	0.25	**0.2**	0.2

- Reuse the topics from blei's website
$Topic_1$	$Topic_2$	$Topic_3$	$Topic_4$	$Topic_5$
coast	government	stars	laboratory	fig
sea	national	star	**research**	shown
fish	institution	observations	work	curve
island	scientific	observatory	room	curves

An Example for Toipc Model

- Word by word

nat.	sci.	ins.	star	obs.	res.	work	fig	shown
2	2	2	3	3	4	4	**5**	?

- Topic distribution for this document
0.05	0.3	0.25	0.2	**0.2**

- Reuse the topics from blei's website
Topic$_1$	*Topic*$_2$	*Topic*$_3$	*Topic*$_4$	*Topic*$_5$
coast	government	stars	laboratory	**fig**
sea	national	star	research	shown
fish	institution	observations	work	curve
island	scientific	observatory	room	curves

An Example for Toipc Model

- Word by word

nat.	sci.	ins.	star	obs.	res.	work	fig	shown
2	2	2	3	3	4	4	5	5

- Reuse the topics from blei's website

-

$Topic_1$	$Topic_2$	$Topic_3$	$Topic_4$	$Topic_5$
coast	government	stars	laboratory	fig
sea	national	star	research	shown
fish	institution	observations	work	curve
island	scientific	observatory	room	curves

Outline

Author-Topic Analysis

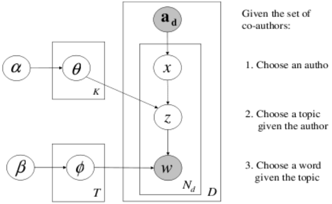
- 

online learning	computer vision	system and control
dynamic,theory,learn	fingerprint,classification	robust,design,system

Author-Topic Model
Mark Steyvers et al. WWW'04

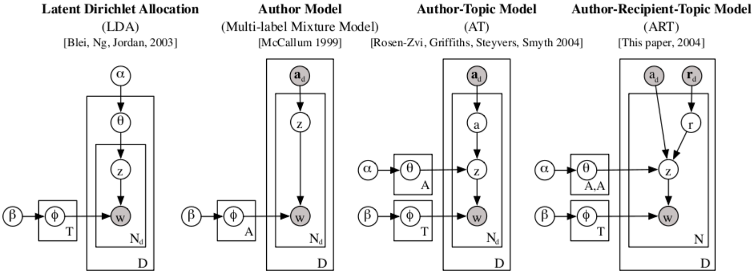
TOPIC 276	
WORD	**PROB.**
DATA	0.1468
MINING	0.0631
DISCOVERY	0.0396
ATTRIBUTES	0.0392
ASSOCIATION	0.0316
RULES	0.0252
PATTERNS	0.0210
LARGE	0.0207
ATTRIBUTE	0.0183
DATABASES	0.0179
AUTHOR	**PROB.**
Han_J	0.0157
Zaki_M	0.0104
Liu_B	0.0080
Cheung_D	0.0075
Hamilton_H	0.0058
Mannila_H	0.0056
Brin_S	0.0055
Ganti_V	0.0050
Liu_H	0.0050
Toivonen_H	0.0049

....



Author-Recipient-Topic Models

Andrew McCallum et al. 2005

- 

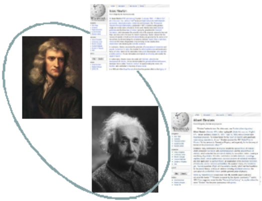
Outline

1. What's Behind a Search Engine?
2. Statistical Language Modeling
   - Basic Notations in Information Retrieval
   - Introduction to Statistical Language Models
   - Language Models for Latent Topic Analysis
3. Interesting Text Mining Problems
   - Author Topic Analysis
   - Text Mining with Network
   - Opinion Mining
   - Other Interesting Models on Text Mining
4. The Research of SEWM Group
   - Research Topics
   - Research Projects

Topic Modeling with Network Regularization(1)

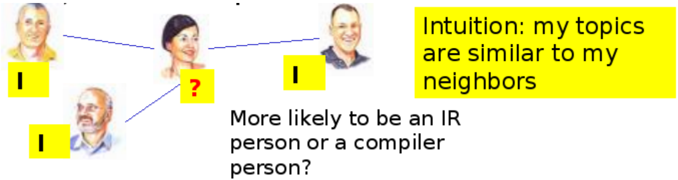
Qiaozhu Mei et al. WWW'08

- Can Topic Modeling help community extraction?

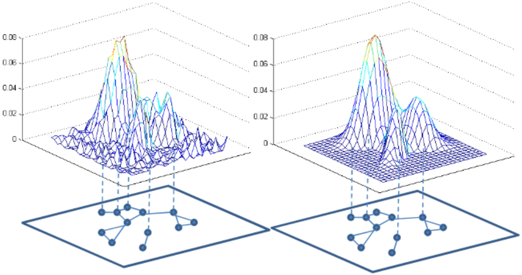


-

Topic Modeling with Network Regularization(2)

- Can Network help topic modeling?
    - 

Topic Modeling with Network Regularization(3)



Topic Modeling with Network Regularization(4)

- $$O(C,G) = -(1-\lambda) * \sum_{d}\sum_{w} c(w,d) log \sum_{z} p(w|z)p(z|d)$$
  $$+\frac{\lambda}{2}\sum_{(u,v)\in E} w(u,v)\sum_{z}(p(z|d_u) - p(z|d_v))^2$$

Outline

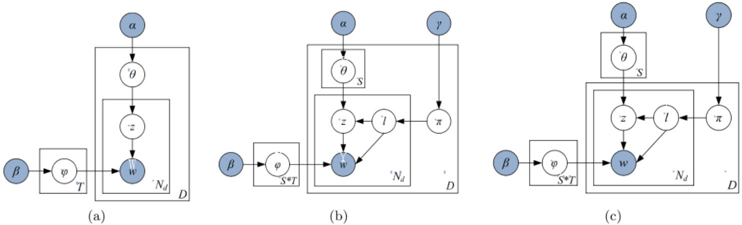
Sentiment Analysis

- **真惊了！百公里油耗14.8！？**

  14号刚买了09款1.8mt两厢小福，蓝色滴！
  昨刚提了车，没料到啊，行车电脑显示的百公里油耗竟然是14.8！！
  我狂晕！
  我可是为了省油才买的手动档的，竟然比自动档的还恐怖！
  到底是转速在2000左右换挡省油呢还是2200—2500之间嫩?
  它是高转速发动机吗？是不是油耗在8个左右才正常?

Joint Sentiment/Topic Model for Sentiment Analysis(1)

Chenghua Lin et al. CIKM'09

- 

Joint Sentiment/Topic Model for Sentiment Analysis(2)

Table 3: Example of topics extracted by JST under different sentiment labels.

Positive sentiment label						Negative sentiment label					
Topic 1		Topic 2		Topic 3		Topic 1		Topic 2		Topic 3	
w	$P(w\|z,l)$	w	$P(w\|z,l)$	w	$P(w\|z,l)$	w	$P(w\|z,l)$	w	$P(w\|z,l)$	w	$P(w\|z,l)$
good	0.084708	tom	0.035175	ship	0.059020	bad	0.079132	sex	0.065904	prison	0.073208
realli	0.046559	ryan	0.030281	titan	0.031586	worst	0.035402	scene	0.053660	evil	0.032196
plai	0.044174	hank	0.025388	crew	0.024439	plot	0.033687	sexual	0.031693	guard	0.031755
great	0.036645	comedi	0.021718	cameron	0.024439	stupid	0.029767	women	0.026291	green	0.029109
just	0.028990	star	0.020800	alien	0.022826	act	0.025602	rate	0.023770	hank	0.028227
perform	0.028362	drama	0.016519	jack	0.020751	suppos	0.025480	act	0.023230	wonder	0.027345
nice	0.026354	meg	0.015601	water	0.019137	script	0.024500	offens	0.018728	excute	0.026904
fun	0.025978	joe	0.014378	stori	0.017984	wast	0.024500	credict	0.016027	secret	0.025581
lot	0.025853	relationship	0.014072	rise	0.016601	dialogu	0.023643	porn	0.014587	mile	0.022936
act	0.022715	mail	0.013766	rose	0.013835	bore	0.022908	rape	0.013867	death	0.022495
direct	0.021586	blond	0.013460	boat	0.013374	poor	0.022908	femal	0.013686	base	0.022054
best	0.020331	run	0.012543	deep	0.013143	complet	0.020825	cut	0.013686	tom	0.019849
get	0.020331	phone	0.012237	ocean	0.012451	line	0.019968	gril	0.013506	convict	0.018967
entertain	0.018198	date	0.011931	board	0.011990	terribl	0.018988	parti	0.012426	return	0.018526
better	0.017445	got	0.011625	sink	0.011299	mess	0.015313	male	0.011886	franklin	0.016762
job	0.016692	busi	0.011319	sea	0.010838	wors	0.014333	bad	0.011346	happen	0.016321
talent	0.016064	cute	0.011013	rain	0.010838	dull	0.013598	nuditi	0.011166	power	0.014116
pretti	0.016064	sister	0.010708	dicaprio	0.010607	actor	0.012986	woman	0.010986	known	0.012352
try	0.015688	children	0.010096	storm	0.010377	total	0.012986	peopl	0.010986	instinct	0.011470
want	0.015186	dog	0.009790	disast	0.010146	isn	0.012863	nake	0.010625	inmat	0.011470

Personal view on recent text mining research

- Besides text content, more and more work pay attention to
  - Web page semi-structured information (Deng Cai et al. SIGIR 2004)
  - User generating content (Lei Guo et al. KDD 2009)
  - Temporal factor(D. Blei et al. 2006)
  - Entity information(Andrew McCallum et al. 2005)
  - Social network(Jure Leskovec's ICML 2009 tutorial)

Outline

Hidden Markov Model

- Topic segmentation(D. Blei et al. SIGIR 2001)
- NLP
  - Part of Speech
- Opinion mining(Wei Jin et al. ICML 2009)
- Context-aware search(Huanhuan Cao et al. WWW 2009)
  - vlHMM
- HMM-LDA(T. Griffiths et al. NIPS 2005)
- How authors effect research topics(Ding Zhou et al. CIKM 2006)

PageRank

- Summarization
  - LexRank (Günes Erkan et al. JAIR 2004)
  - Manifold rank(Xiaojun Wan et al. IJCAI 2007)
- Web spam
  - Trust rank(Zoltán Gyöngyi et al. VLDB 2004)
- Language model prior

Clustering

- Event detection and tracking
  - probabilistic online clustering(Dirichlet process) (Jian Zhang et al. NIPS 2004)
- Person search disambiguation
  - http://nlp.uned.es/weps/
- Language model smoothing (Xiaoyong Liu et al. SIGIR 2004)
- Summarization(+PageRank || +HITS) (Xiaojun Wan et al. SIGIR 2008)

Top Researchers on Different Fields

- Information retrieval
  - W. Bruce Croft
  - Chengxiang Zhai
- Learning to rank
  - Tieyan Liu
- Text mining
  - Bing Liu
  - Qiaozhu Mei
- Topic modeling
  - David M. Blei
- Social networking
  - Jure Leskovec

Outline

Research Topics

- Search Engine
  - Towards a scalable web search engine
- Web Mining
  - Event detection and tracking
  - Entity mining
  - Information extraction
  - Other related research topics

Outline

1. What's Behind a Search Engine?
2. Statistical Language Modeling
   - Basic Notations in Information Retrieval
   - Introduction to Statistical Language Models
   - Language Models for Latent Topic Analysis
3. Interesting Text Mining Problems
   - Author Topic Analysis
   - Text Mining with Network
   - Opinion Mining
   - Other Interesting Models on Text Mining
4. The Research of SEWM Group
   - Research Topics
   - Research Projects

Research Projects

- Platform for Applying, Researching And Developing Intelligent Search Engine (PARADISE)
- Extraction and Analysis of Entities and Relations (EAER)
- New theory and method on Search Engine and Web Mining (NSEWM)

Systems(1)

- 

- www.pku.edu.cn

Systems(2)

- e.pku.edu.cn

Systems(3)

- www.infomall.cn

Systems(4)


== 这是 中国WEB信息博物馆 （Web InfoMall） 2008年08月10日 存储的网页 ==
击这里查看本网页的其他版本 请选择:
隐藏 | 禁止
北京大学
PEKING UNIVERSITY
110周年校庆 ENGLISH 旧版主页
新闻资讯 北大概况 教育教学 科学研究 招生就业 合作交流 图书档案
北大新闻 详细 »
校领导视察校园管理与安全保卫工作 2008-07-29
【2008学生骨干训练营系列报道之八】寻访社会精英 砥砺自... 2008-07-29
【奥运与北大（人文篇系列之二）】龙协涛：我为奥运唱颂歌 2008-07-29
北京大学志愿者参加国家体育场运行团队誓师大会 2008-07-29
北京大学体育馆场馆团队举行奥运誓师动员大会 2008-07-29
有一种奉献源于热情——访国家体育场志愿者吴天琪 2008-07-29
搜索北大 请输入关键字 搜索 奥运搜索
通知公告 详细 »
停 电 通 知（7月22日） 2008-07-23
停电通知（7月19日） 2008-07-14
2008年宽禁带半导体暑期学校通知 2008-07-10
北大校内信息 今天有0条新信息 详细 »
图文热点
北京大学举行2008年本科生毕业典礼
"忆昔长别，阳关千叠，狂歌曾竞夜，收拾山河待百年约。"四年时光匆匆，湖光塔影间，初入燕园时求知若......
院系设置 医 学 部
管理服务 基 金 会
人才招聘 校 友 网
校园文化 产 学 研
校内门户 | 网络服务 | 电子邮箱 | 未名BBS | 北大故事 | 相关链接 |
本站地图
版权所有©北京大学 | 地址：北京市海淀区颐和园路5号 | 邮编：100871 | 邮箱：webmaster@pku.edu.cn | 建站献策

Summary

- Thank you for your attention:-)
- Thank Prof. Yao and Prof. Li for giving me such a chance.

- Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai: Topic sentiment mixture: modeling facets and opinions in weblogs. WWW 2007
- Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, Thomas L. Griffiths: Probabilistic author-topic models for information discovery. KDD 2004
- Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai: Topic modeling with network regularization. WWW 2008
- Chenghua Lin, Yulan He: Joint Sentiment/Topic Model for Sentiment Analysis. CIKM 2009
- Andrew McCallum, Andres Corrada-Emanuel, and Xuerui Wang. Topic and role discovery in social networks. IJCAI 2005

- D. Blei and P. Moreno. Topic segmentation with an aspect hidden Markov model. SIGIR 2001
- Wei Jin, Hung Hay Ho: A novel lexicalized HMM-based learning framework for web opinion mining. ICML 2009
- Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen, Hang Li: Towards context-aware search by learning a very large variable length hidden markov model from search logs. WWW 2009
- T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. NIPS, 2005
- Günes Erkan, Dragomir R. Radev: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. JAIR 2004

- Xiaojun Wan, Jianwu Yang, Jianguo Xiao: Manifold-Ranking Based Topic-Focused Multi-Document Summarization. IJCAI 2007
- Jian Zhang, Zoubin Ghahramani, Yiming Yang: A Probabilistic Model for Online Document Clustering with Application to Novelty Detection. NIPS 2004
- Xiaoyong Liu, W. Bruce Croft: Cluster-based retrieval using language models. SIGIR 2004
- Xiaojun Wan, Jianwu Yang: Multi-document summarization using cluster-based link analysis. SIGIR 2008
- Zoltán Gyöngyi, Hector Garcia-Molina, Jan O. Pedersen: Combating Web Spam with TrustRank. VLDB 2004
- Thomas Hofmann: Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning 42(1/2) 2001

- D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 2003
- Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, Yihong Eric Zhao: Analyzing patterns of user content generation in online social networks. KDD 2009
- D. Blei and J. Lafferty. Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning, 2006
- J. Chang, J. Boyd-Graber, and D. Blei. Connections between the lines: Augmenting social networks with text. KDD 2009
- Deng Cai, Xiaofei He, Ji-Rong Wen and Wei-Ying Ma. Block-level Link Analysis. SIGIR 2004
- Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma. Block-based Web Search. SIGIR 2004

- Jure Leskovec: Tutorial summary: Large social and information networks: opportunities for ML. ICML 2009
- Ding Zhou, Xiang Ji, Hongyuan Zha, C. Lee Giles: Topic evolution and social interactions: how authors effect research. CIKM 2006
- ChengXiang Zhai, John D. Lafferty: Model-based Feedback in the Language Modeling Approach to Information Retrieval. CIKM 2001
- Victor Lavrenko, W. Bruce Croft: Relevance-Based Language Models. SIGIR 2001
- Yue Lu, Chengxiang Zhai: Opinion integration through semi-supervised topic modeling. WWW 2008
- T. L. Griffiths and M. Steyvers: Finding scientific topics. PNAS 2004

- Loulwah AlSumait, Daniel Barbará, Carlotta Domeniconi: On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. ICDM 2008
- Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y. Chang: PLDA: Parallel Latent Dirichlet Allocation for Large-scale Applications. AAIM 2009
- J. Chang, J. Boyd-Graber, S. Gerris, C. Wang, and D. Blei. Reading tea leaves: How humans interpret topic models . NIPS, 2009