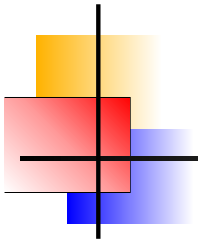
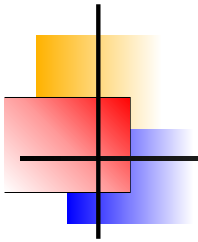


基于改进语言模型的网页排序问题研究

杨波

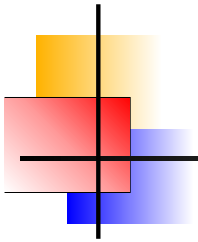
指导教师：黄亚楼教授
研究方向：信息检索

November 29, 2009



概要

统计语言模型
内容模型估计算法
结构模型估计
.GOV、*Lemur*及*Baseline*
附录



概要

统计语言模型的历史

- 来自speech recognization领域
- 用来估计语音序列中下一个词的概率

$$p(w|h_1, h_2, \ldots, h_n)$$

统计语言模型的历史

- 来自speech recognization领域
- 用来估计语音序列中下一个词的概率

$$p(w|h_1, h_2, \ldots, h_n)$$

basic n-gram

$$p(w_i|w_{i-n}\ldots w_{i-1}) = \frac{C(w_{i-n}\ldots w_{i-1}, w_i)}{C(w_{i-n}\ldots w_{i-1})}$$

语言模型的历史

- topic-based n-gram

$$P(w_i|w_{i-n+1}...w_{i-1}) = \sum_{Topic=i} \lambda_i P_i(w_i|w_{i-n+1}...w_{i-1})$$

- skip-based n-gram

$$P_{Mixed}(w_i|w_{i-n+1}...w_{i-1}) = \sum_{k=1}^{n-1} \lambda_k \times P(w_i|w_{i-k})$$

- 指数级语言模型

信息检索中的语言模型

- 查询产生相应文档的概率

$$p(d\ is\ relevent|q) = \frac{p(q|d\ is\ relevent)p(d\ is\ relevent)}{p(q)}$$

- 信息检索中的语言模型

$$p(d|q) \propto p(q|d)p(d)$$

其中d是待排序文档, q用户的查询词, 返回的文档以$p(d|q)$大小排序

- 关键在于估计$p(q|d)$和$p(d)$

估计$p(q|d)$

$$log\ p(q|d) = \sum_{i=1}^{n} log\ p(w_i|d)$$

其中w_i为用户查询中的单词

计算$p(w_i|d)$

$$p(w|d) = \frac{c(w;d)}{\sum_{w' \in V} c(w';d)}$$

其中$c(w;d)$为文档d中出现单词w的次数；V为词汇表

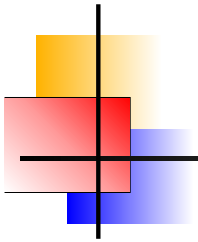
对未见词进行估计

- 未见词会造成0概率问题
- 通过使用候选数据集做backoff，即在网页中有未出现的查询词的时候，使用平滑数据集中的词频代替网页的词频
- 通过减少文档中可见词的概率来平滑语言模型
  - The Jelinek-Mercer method
  - Bayesian smoothing using Dirichlet priors
  - Absolute discounting
  - Two-stage Smoothing model

语言模型的分隔

$$p(d|q) = p(q|d)p(d)$$

- $p(q|d)$和query词相关，称为内容模型
- $p(d)$和query词无关，称为结构模型



概要

多维多元组

- 同时考虑1元组、2元组和3元组
- 为大的元组赋予较大的权重

$$s(q|d) = \alpha_1 p(q_1|d) + \alpha_2 p(q_2|d) + ... + \alpha_n p(q_n|d)$$

其中$q_1 \ldots q_n$分别表示查询中的1元组到n元组，$\alpha_1 \ldots \alpha_n$ 为相应的权重

多维多元组

- 同时考虑1元组、2元组和3元组
- 为大的元组赋予较大的权重

$$s(q|d) = \alpha_1 p(q_1|d) + \alpha_2 p(q_2|d) + ... + \alpha_n p(q_n|d)$$

其中$q_1 \ldots q_n$分别表示查询中的1元组到n元组，$\alpha_1 \ldots \alpha_n$ 为相应的权重

α_i的估计问题

- 相应的权重如何计算，怎么从训练样本中估计

估计α_i

Table: 只考虑1-gram时，被错误计算的文档的个数

数据集	平均文档个数	百分比
HP2004	8998	12.1%
NP2004	5603	7.6%
TD2004	12563	17.2%

Table: 考虑2-gram时，被错误计算的文档的个数

数据集	平均文档个数	百分比
HP2004	4006	5.4%
NP2004	2861	3.9%
TD2004	3674	5.1%

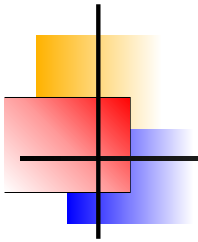
估计α_i-续

Table: 考虑3-gram时，被错误计算的文档的个数

数据集	平均文档个数	百分比
HP2004	246	0.3%
NP2004	351	0.4%
TD2004	13	0.02%

Table: 利用错误文档百分比的比例估计α_i

数据集	α_1	α_2	α_3
HP2004	1	$2.2 = 12.1 \div 5.4$	$40.3 = 12.1 \div 0.3$
NP2004	1	$1.9 = 7.6 \div 3.9$	$19.0 = 7.6 \div 0.4$
TD2004	1	$3.4 = 17.2 \div 5.1$	$860.0 = 17.2 \div 0.02$



多层数据集平滑算法

URL的意义

http://host.part/directory/part/file-name

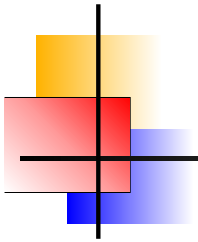
数据集定义

- 目录数据集
  URL中目录部分相同的文档，所组成的数据集，如:
  http://www.nasa.gov/missions/future/ 下的所有文档
- 站点数据集
  属于同一个站点的文档，所组成的数据集
- 互联网数据集
  整个.GOV所有文档，组成的数据集

线性插值平滑

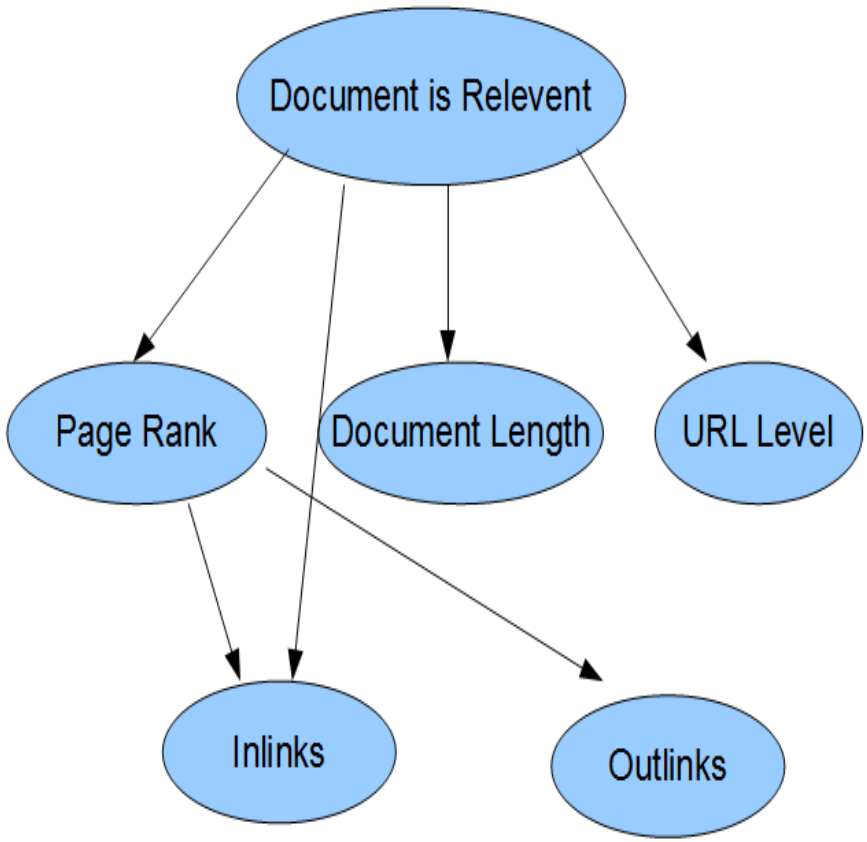
$$s(w|d) = (1-\lambda-\mu-\omega)p(w|d) + \lambda p(w|D) + \mu p(w|S) + \omega p(w|C)$$

其中，D为当前文档（网页）所在的目录中所有单词组成的数据集，S为对应的站点的数据集，C为整个互联网的数据集



概要

各种先验概率之间的关系



贝叶斯信念网络估计

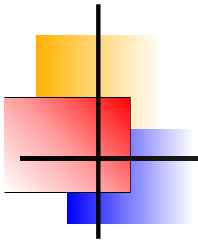
假设：每一个先验概率特征为x_i

已知：

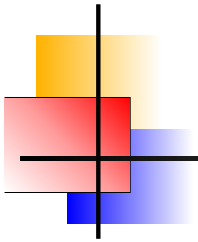
- $p(x_i|d)$，即相关文档出现特征x_i的概率
- $p(x_i|x_j)$，即不同的特征i,j统计不独立时，二者之间的条件概率

求解：

- $p(d|x_1, x_2, \ldots, x_n)$



概要



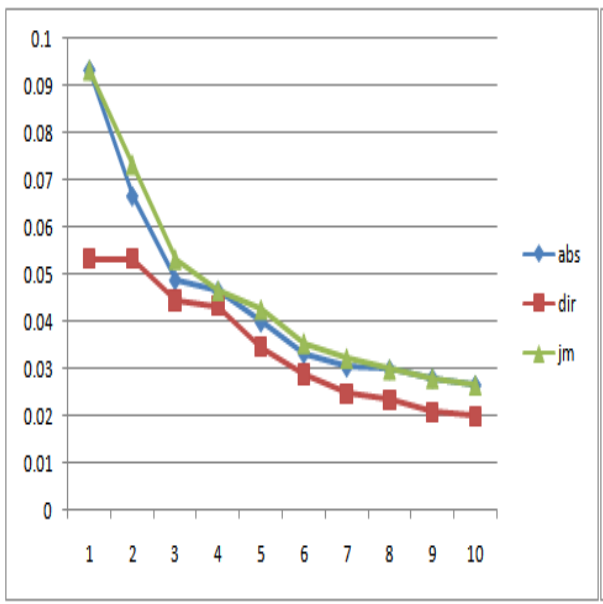
.GOV文本数据集

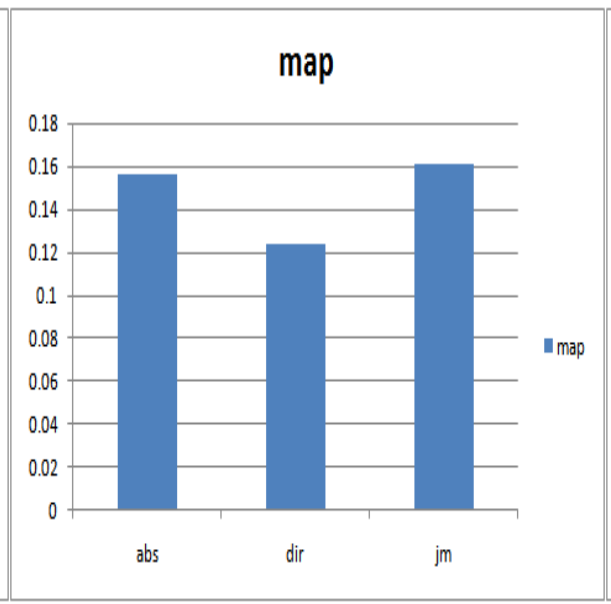
- Document: 100万+
- Words: 900万+
- Unique Words: 16万+
- Queries: Trec Web Track 2004 3个task(td,np,hp)，575

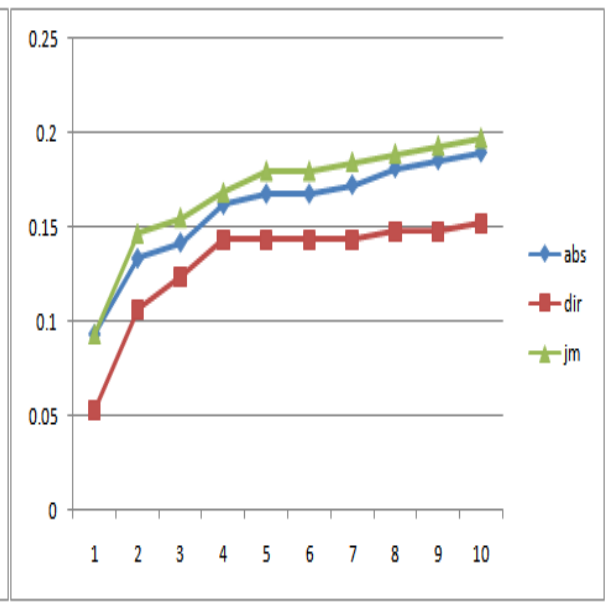
主要功能

- 提供两种索引结果，能对大规模TREC格式的文档进行索引
- 提供TF*IDF, BM25, Basic LM和常用的三种平滑算法的实现
- 通过C++扩展平台，可以实现新的算法

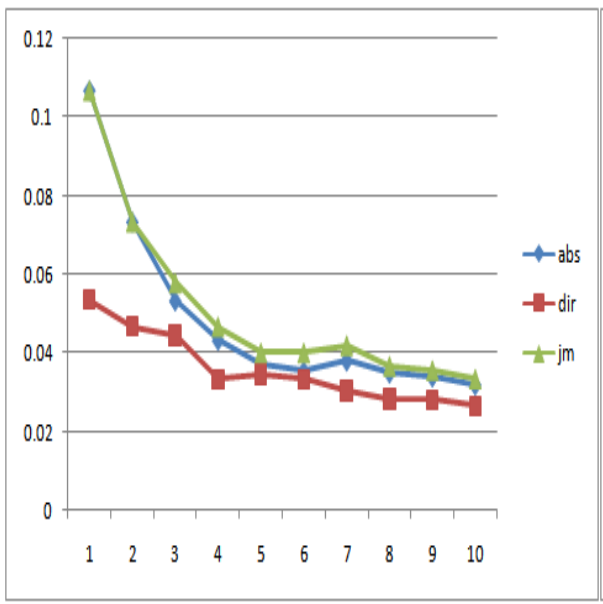
Name Page Finding

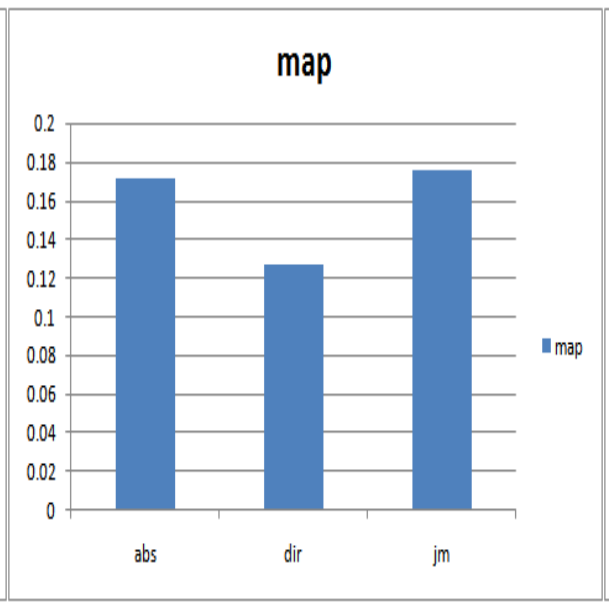


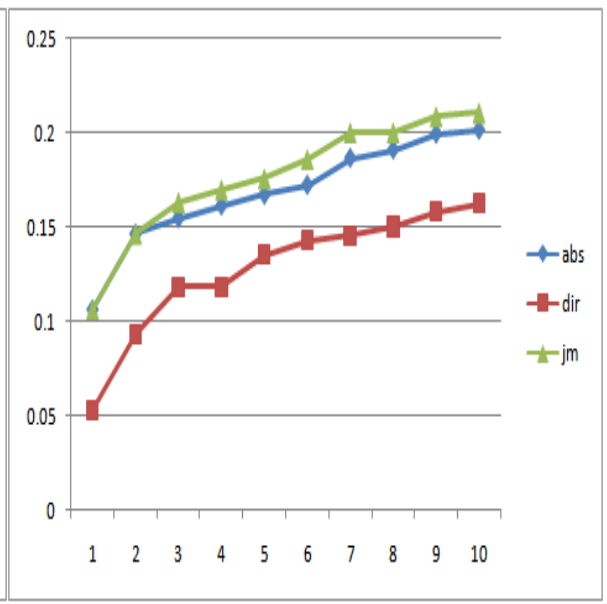




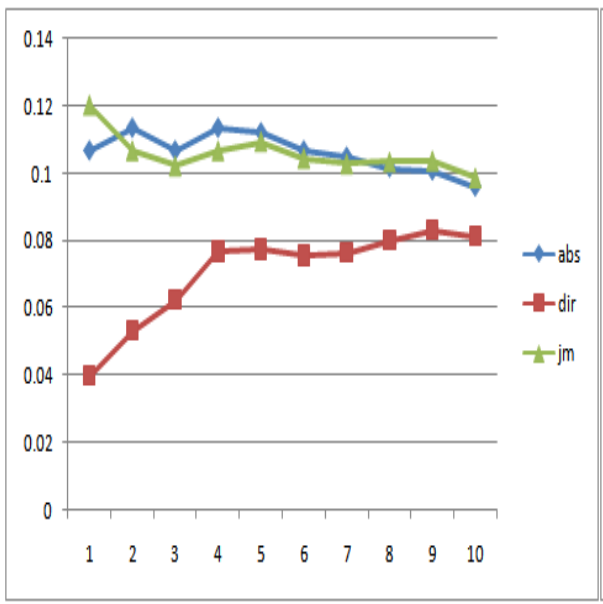
Home Page Finding

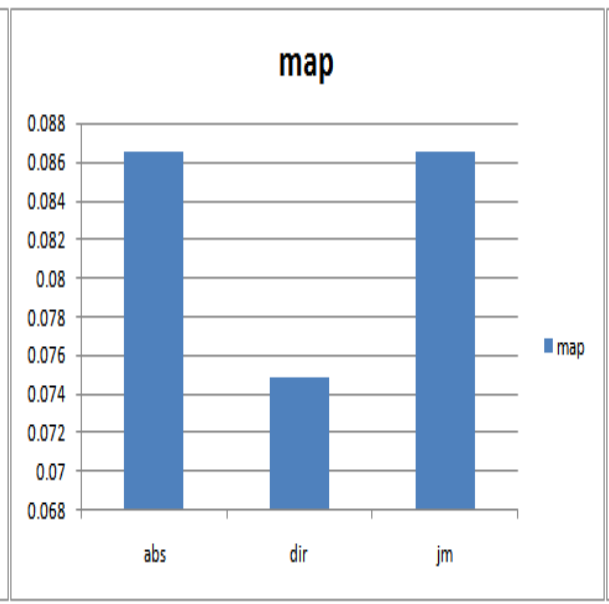


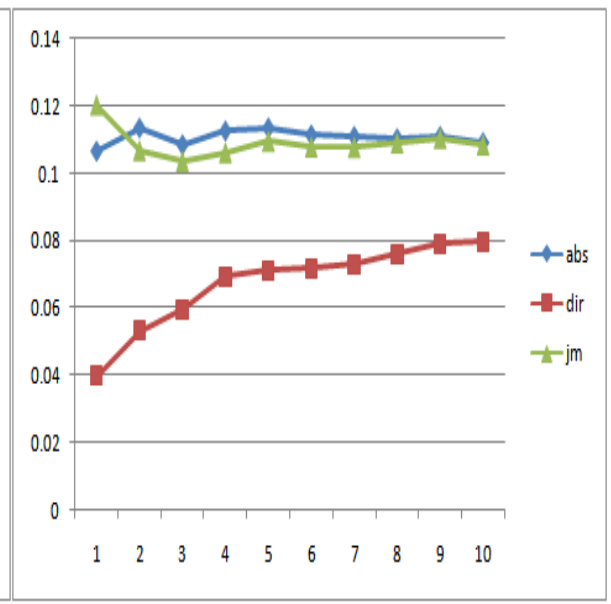




Topic Distillation







参考文献 I

[C. Zhai and J. Lafferty]

A Study of Smoothing Methods for Language Models Applied to Information Retrieval

ACM Treansaction of Information System, 2004.

[Miller D. H.]

A hidden Markov model information retrieval system

Proceedings of the ACM SIGIR, 1999.

[C. Zhai and J. Lafferty]

A Study of Smoothing Methods for Language Models Applied to Information Retrieval

ACM Transaction of Information System, 2004.

[Tom Mitchell]

Machine Learning

McGraw Hill, 1997.

[Baeza-Yates, R. and Ribeiro-Neto B.]

Modern Information Retrieval

New York, NY, USA: Addition Wesley, 1999.

[Lancaster F. W.]

Information retrieval systems: characteristics, testing and evaluation

2nd Ed., New York: John Wiely and Sons, 1979.

参考文献 //

[Salton, G., Wong, A., and Yang, C. S.]

A vector space model for automatic indexing

Communications of the ACM, 1975.

[Robertson, S. E.]

Okapi in TREC3

NIST Special Publication, 1994.

[J. Lafferty and C. Zhai]

Document Language Models, Query Models, and Risk Minimization for Information Retrieval

Proceedings of the 24th annual international ACM SIGIR conference, 2001.

[Freund, Y., Iyer, R., Schapire, R., and Singer, Y.]

An efficient boosting algorithm for combining preferences

Journal of Machine Learning, 2004.

[Nallapati, R.]

Discriminative Models for Information Retrieval

Proceedings of the27th Annual International ACM SIGIR Conference, 2004.

[N. Craswell, D. Hawking, R. Wilkinson, and M. Wu.]

Overview of the TREC 2003 web track

TREC, 2003.

[Lemur Project]

http://www.lemurproject.org/

参考文献 III

[Tie-Yan Liu, Jun Xu, Tao Qin, Wenying Xiong, and Hang Li]

LETOR: Benchmarking "Learning to Rank for Information Retrieval"

SIGIR 2007 Workshop on Learning to Rank for Information Retrieval, 2007.

[F. Bimbot, R. Preraccini, E. Levin and B. Atal]

Variable-Length Sequence Modeling: Multigrams

IEEE Signal Processing Letter, 1995.

[S. Deligne and F. Bimbot]

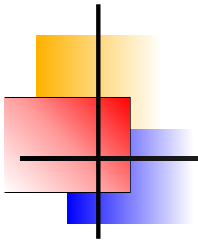
Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams

Acoustics, Speech and Signal Processing, 1995.

[R. Kneser and V. Steinbiss]

On the dynamic adaption of stochastic language models

Acoustics, Speech and Signal Processing, 1993.



谢 谢！