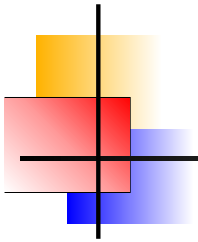
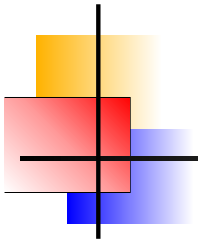


基于改进语言模型的网页排序问题研究

杨波

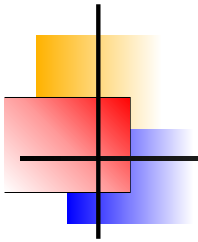
指导教师：黄亚楼教授
研究方向：数据挖掘

January 3, 2010

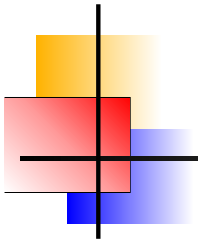


概要

题目阐述
选题依据
研究内容
研究方法
总结与展望
附录



概要



题目阐述

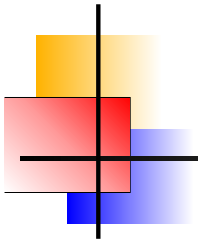
论文题目：基于改进语言模型的网页排序问题研究

关键词：信息检索、语言模型、网页排序、变长多元组

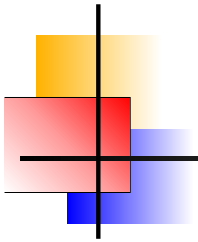
研究领域：信息检索领域

关注问题：信息检索中网页排序问题

论文目标：对经典语言模型进行改进，提高排序性能



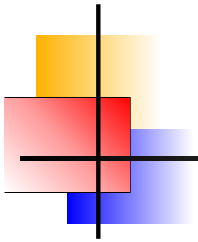
概要



选题依据-背景

信息检索

- 随着互联网的普及，搜索已经在人们的生活中占有了重要的地位
- 搜索结果的排序是信息检索中关键的问题



选题依据-背景

信息检索

- 随着互联网的普及，搜索已经在人们的生活中占有了重要的地位
- 搜索结果的排序是信息检索中关键的问题

信息检索模型

- 布尔模型
- 向量空间模型
- 概率检索模型(TF*IDF, BM25)
- 统计语言检索模型(LMIR)

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

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

各种各样的语言模型

- 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$p(d)$

选题依据-统计语言模型

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

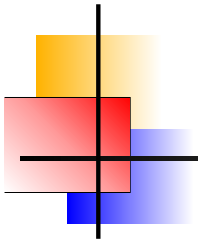
计算$p(w_i|d)$

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

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

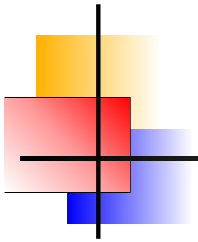
- 未见词会造成0概率问题
- 通过使用候选数据集做backoff，即在网页中有未出现的查询词的时候，使用平滑数据集中的词频代替网页的词频
- 通过减少文档中可见词的概率来平滑语言模型
  - The Jelinek-Mercer method
  - Bayesian smoothing using Dirichlet priors
  - Absolute discounting
  - Two-stage Smoothing model



选题依据-统计语言模型

存在问题

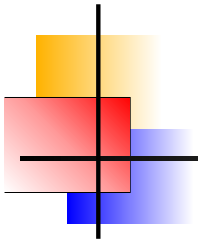
- 简单的统计网页中出现每个单词的频率会受到网页中各种噪声词语的干扰
- 直接使用整个互联网的corpus去平滑文档中未出现的词，忽略了网页所属的网站以及目录带来的信息量



选题依据-统计语言模型

估计$p(d)$

- 假设$p(d)$为均匀分布
- 使用PageRank来做$p(d)$(Lemur)



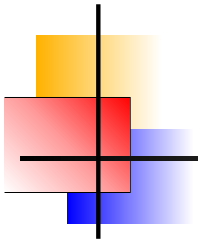
选题依据-统计语言模型

估计$p(d)$

- 假设$p(d)$为均匀分布
- 使用PageRank来做$p(d)$(Lemur)

存在问题：无法利用更多的结构信息

- 文档长度、URL长度
- 指向文档的链接数、PageRank值、HITS值
- ...

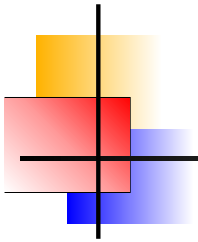


概要

语言模型的分隔

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

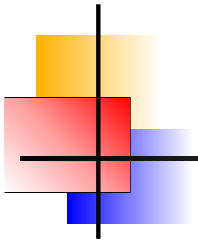
- $p(q|d)$和query词相关，称为内容模型
- $p(d)$和query词无关，称为结构模型



研究内容

主要研究工作

- 可变多元组内容生成算法研究(Variable N-gram)
- 多层数据集平滑算法研究
- 结构模型估计算法研究
- 将以上方法在语言模型的框架下整合



概要

1-gram

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

问题

Table: 相关文档中平均包含1-gram的个数

数据集	平均包含1-gram个数
HP2004	13.5
NP2004	22.1
TD2004	7.6

问题-续

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%

考虑2-gram和3-gram的情况

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%

Table: 考虑3-gram时，被错误计算的文档的个数

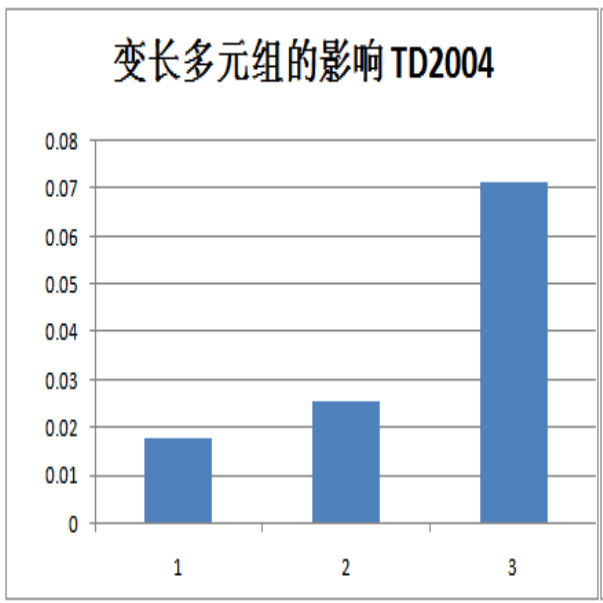
数据集	平均文档个数	百分比
HP2004	246	0.3%
NP2004	351	0.4%
TD2004	13	0.02%

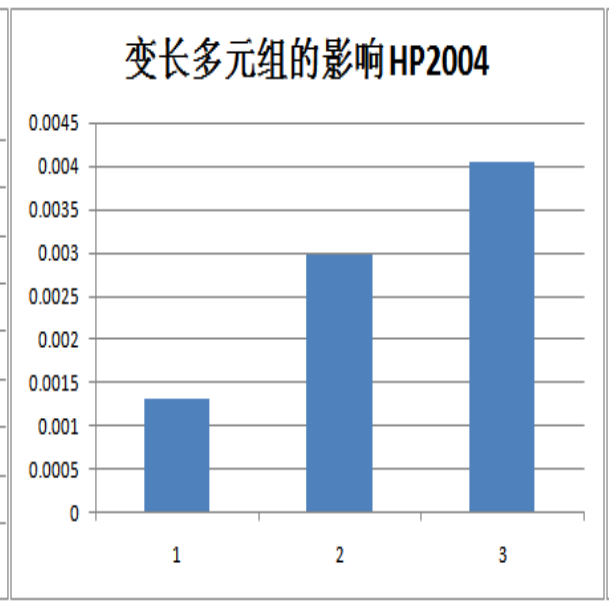
多元组的影响

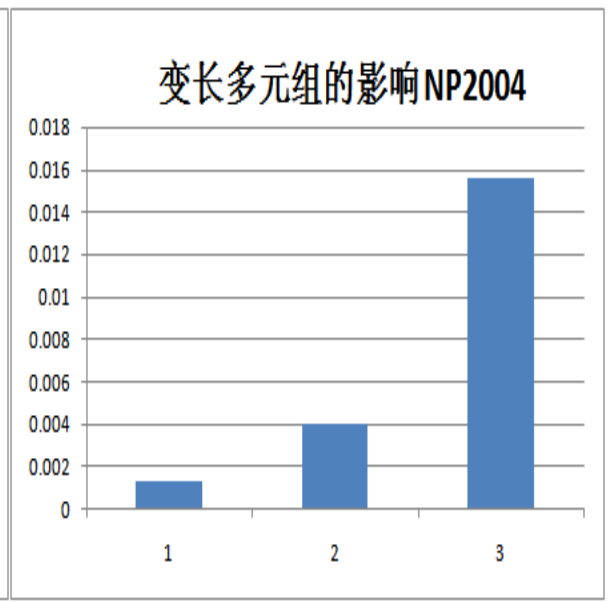
随着纳入考虑的元组数目增多，包含相应元组数目的文档成为相关文档的概率也跟随增大

多元组的影响

随着纳入考虑的元组数目增多，包含相应元组数目的文档成为相关文档的概率也跟随增大







横坐标：1元组，2元组，3元组　　纵坐标：包含了相应元组的文档中相关文档的比例

一个更准确合理的估计

$$log\ p(q|d) = \sum_{i=1}^{n} log\ p(w_1 \ldots w_k|d)$$

k：query中单词的个数

问题

- 数据过于稀疏，造成所有$p(q|d)$都为0

变长多元组

- 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

变长多元组

- 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

新的问题

- 相应的权重如何计算，怎么从训练样本中估计
- 是否需要保证$p(q|d)$仍然是一个概率分布,如何保证$p(q|d)$是一个概率分布

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

线性插值平滑

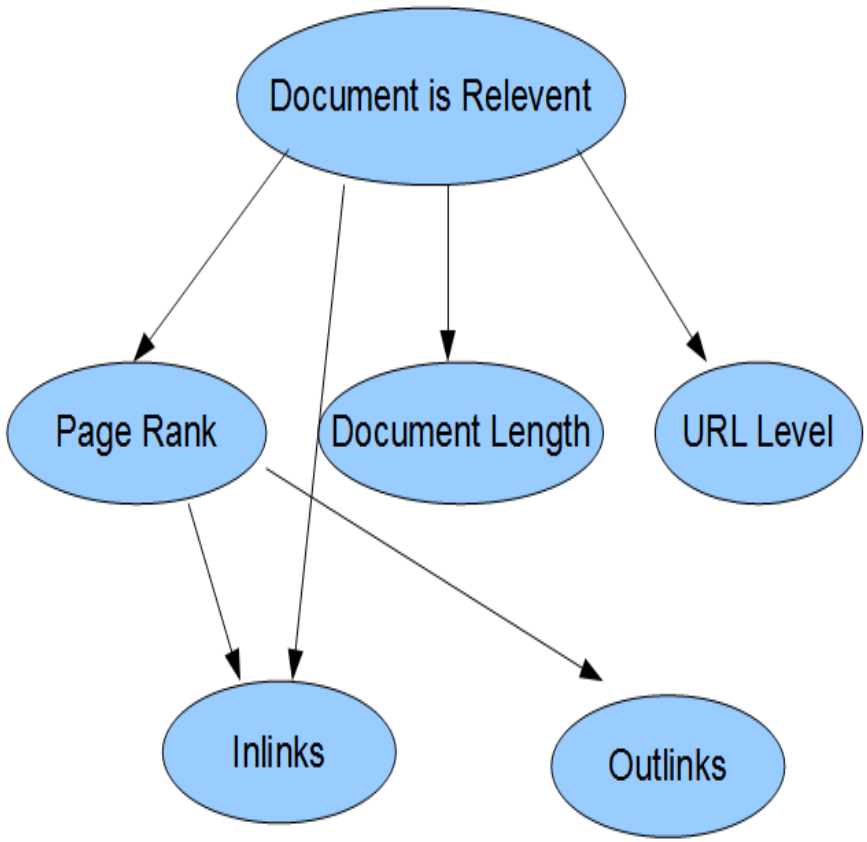
$$s(w|d) = (1-\lambda-\mu-\omega)p(w|d) + \lambda p(w|D) + \mu p(w|S) + \omega p(w|C)$$

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

待解决的问题

- 参数λ, μ, ω的估计
- 简单线性插值是否适合多层数据集的平滑，考虑Dirichlet插值和其他更复杂的插值的算法

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：

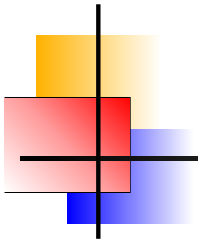
- $p(d|x_1, x_2, \ldots, x_n)$

实验平台

- 数据集：.GOV文本数据集
- 算法测试数据集：Letor 3.0 Topic Distillation, Home Page Finding, Named Page Finding
- 评价指标：Precision@10, Recall@10, MAP, NDCG@10
- 参考程序：Lemur平台
- 程序环境：MSYS脚本系统，Windows 2003 Server， Linux

主要实验

- 改写Lemur，创建适用于计算变长多元组内容生成概率和多层数据集平滑算法的索引结构
- 在新的索引结构上，完成变成多元组算法，对比不同的参数的效果和不同的参数估计方法的效果
- 在变长多元组算法中加入平滑算法，对比不同的参数对算法性能的影响
- 设计实现结构模型生成算法，对各种结构先验概率的效果进行分析
- 综合计算改进后的语言模型，对其性能就行评价



概要

总结

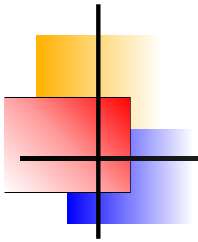
- 通过对现有语言模型研究的调研，分别提出针对内容模型生成算法，平滑算法和结构模型$p(d)$的改进算法

总结

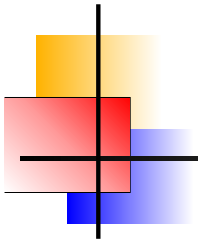
- 通过对现有语言模型研究的调研，分别提出针对内容模型生成算法，平滑算法和结构模型$p(d)$的改进算法
- 通过实验说明了基于variable n-gram的内容模型生成算法和多层数据集平滑算法的可行性和必要性

展望

- 实现已提出算法，并进行实验验证
- 对试验中算法不同的参数进行测试，改进算法
- 尝试寻找更多对排序有利的结构模型特征，验证并引入到最终的结构模型



概要



进度安排 I

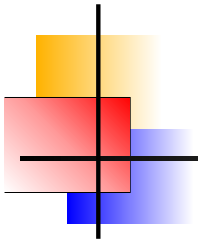
2009年5月-2009年7月：完成了相关理论工作的调研，大量阅读相关论文文献；

2009年8月-2009年9月：搭建实验平台，复现已有的经典算法；

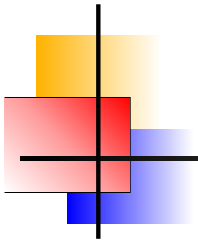
2009年10月-2009年11月：完成内容模型生成算法的设计与实现；

2009年12月-2010年1月：完成结构模型生成算法的设计与实现；

2010年2月-2010年3月：完成实验及结果评价。



论文大纲 I



论文大纲 II

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 II

[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]

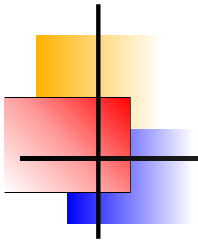
Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams

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！