**硕士学位论文**

**基于改进语言模型的网页**

**排序问题问题研究**

论文作者： 杨 波

学 号： 2120070360

培养院系： 软件学院

专 业： 计算机软件与理论

研究方向： 信息检索

指导教师： 黄亚楼 教授

南开大学研究生院

2010年4月

MASTER’S DISSERTATION

**Research on Web Page Ranking Based on Improved Language Model**

**By:** Yang Bo

**Supervisor:** Prof. Yalou Huang

Nankai University

April 2010

**南开大学学位论文版权使用授权书**

本人完全了解南开大学关于收集、保存、使用学位论文的规定，同意如下各项内容：按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版；在不以赢利为目的的前提下，学校可以适当复制论文部分或全部内容用于学术活动。

学位论文作者签名：

年 月 日

经指导教师同意，本学位论文属于保密，在 年解密后适用本授权书。

|  |  |  |  |
| --- | --- | --- | --- |
| **指导教师签名：** |  | **学位论文作者签名：** |  |
| **解 密 时 间：** | 年 月 日 | | |

各密级的最长保密年限及书写格式规定如下：

内部 5年（最长5年，可少于5年）

秘密★10年（最长10年，可少于10年）

机密★20年（最长20年，可少于20年）

**南开大学学位论文原创性声明**

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年 月 日

# 摘 要

随着Internet技术的突飞猛进，Web信息量爆炸性增长，人们越来越习惯使用搜索引擎查找所关心的信息。但浩瀚的信息资源却给搜索引擎的发展提出了新的挑战。如何有效、快捷、准确地将查询结果返回给用户，提高Web信息检索效果，已变成一项迫切而有意义的研究课题。

在现阶段信息检索领域的研究中，基于监督学习的排序学习逐渐成为排序研究的热点。基于监督学习的排序学习需要大量的人工标注的样本，为了减少人工标注样本的标注量，产生了一些基于“选择最值得标注的样本进行标注”思想的所谓主动排序学习算法。通过主动排序学习算法，用户不需要一开始标注所有的样本，而是开始只标注一部分样本，先学习得到一个排序模型；然后每次从剩下的未标注样本中选择一个最值得标注的样本进行标注，把这个新标注的样本放入训练集中，重新训练得到新的排序模型；然后在剩下的未标注样本中再重新选择一个样本进行标注，加入训练集，如此类推直到得到最终的排序模型。主动学习减少了排序学习的样本标注量，但此方法有一个问题是每次只选择一个样本标注，之后又要重新训练，训练需要很多时间，同时标注人员标注下一个样本需要等待很长时间。如果每次可以选择多个样本，则可以减少整个主动排序学习的时间，降低标注人员的工作量，即标注代价，同时，如果有多个标注人员的话，还可以实现并行标注，提高主动排序的效率。

针对上述问题，本文提出批量主动排序学习的思想，主动排序学习的时候，一次能够找到多个值得标注的样本给用户标注，这多个标注的样本对排序模型性能的提升有很大的价值。

本文提出了两种批量主动排序学习算法，一种是基于夹角差异的批量主动排序学习算法，该算法通过加入批量选择的样本之间的夹角差异度量，来减少批量选择的样本之间的相似度，提高批量主动排序的性能。另一种是基于损失函数的批量主动排序学习算法，该算法直接从提高排序模型性能的损失函数入手，批量选择能够使损失函数达到最小值的那些样本进行标注。

本文在不同数据集上进行实验评价以上两种批量主动排序学习算法，同时与单样本主动排序学习算法，原始的批量主动排序学习算法（直接用单个主动排序学习算法选择多个样本）等进行比较分析。实验结果表明，本文提出的批量主动排序学习算法具有很好的性能。

**关键字：**排序学习，主动学习，批量主动，夹角差异，损失函数

# Abstract

Along with the rapid growth of Internet technology, there is much information in Web environment. People are used to gain information with search engine. But there is new challenge for search engine. How to rapidly and exactly return search results to users, and how to enhance the effect of information retrieve, have become research topics and hotspot with impendency and significance.

In recent years, supervised rank learning gradually becomes research hotspot. Supervised rank learning needs many samples manually labeled by person, for reducing the size of samples needs labeled, there bring some algorithms based on the idea of “select the most valuable sample to label”, its named active learning. Active learning has one issue, it only select one sample to label, then needs put this sample to training set to retrain, retraining costs many time, and also person must wait retraining for much time to label next sample. If active learning once selects more samples with batch mode to label, that will reduce both the time of active learning and the workload of person for label, namely label expense. Additionally, if there exists more people for label, people can label sample in parallel, boosting the efficiency of active learning.

To solve the problems mentioned above, in this paper, the idea of batch active learning is proposed, it solve how to select more samples with batch mode to label, these samples have good value for rank model.

In this paper, two batch active rank learning algorithm are proposed, one is batch active based on angle difference, this algorithm reduce similarity of the samples batch selected by adding angle difference of the samples. Another is batch active based on loss function, this algorithm directly searches loss function to enhance the capability of rank model, the samples batch selected are those which can make loss function to minimum.

In this paper, test of two batch algorithm is performing on two different and authoritative data sets, comparing analysis is conducted between the proposed algorithms and single select active learning, general batch active learning and so on. Experiment result shows the algorithms mentioned in this paper have nice capability.

**Keywords:** Rank learning, Active learning, Batch Active, Angle difference, Loss Function

# 目 录

[摘 要 I](#_Toc255068258)

[Abstract III](#_Toc255068259)

[目 录 V](#_Toc255068260)

[第一章 绪 论 1](#_Toc255068261)

[1.1 引言 1](#_Toc255068262)

[1.2 研究现状 1](#_Toc255068263)

[1.2.1 信息检索中的排序模型 2](#_Toc255068264)

[1.2.2 统计语言模型 2](#_Toc255068265)

[1.3 本文主要研究内容 3](#_Toc255068266)

[1.3.1 多维多元组算法研究 3](#_Toc255068267)

[1.3.2 多层数据平滑算法研究 3](#_Toc255068268)

[1.3.3 结构模型估计算法研究 4](#_Toc255068269)

[1.4 本文结构 4](#_Toc255068270)

[第二章 相关工作综述 5](#_Toc255068271)

[2.1 信息检索 5](#_Toc255068272)

[2.1.1 布尔模型 5](#_Toc255068273)

[2.1.2 向量空间模型 6](#_Toc255068274)

[2.1.3 概率检索模型 7](#_Toc255068275)

[2.1.4 统计语言模型 8](#_Toc255068276)

[2.1.5 基于监督学习方法的排序学习模型 9](#_Toc255068277)

[2.2 信息检索评价指标 10](#_Toc255068278)

[2.2.1 MAP 11](#_Toc255068279)

[2.2.2 NDCG 12](#_Toc255068280)

[2.3 统计语言模型详述 12](#_Toc255068281)

[第三章 改进内容模型生成算法 17](#_Toc255068282)

[3.1 引言 17](#_Toc255068283)

[3.2 网页噪声和多元组的影响 17](#_Toc255068284)

[3.3 多维多元组内容模型估计算法 17](#_Toc255068285)

[3.4 实验及分析 17](#_Toc255068286)

[3.5 多层数据平滑算法 17](#_Toc255068287)

[3.6 实验及分析 17](#_Toc255068288)

[第四章 结构模型生成算法 18](#_Toc255068289)

[4.1 引言 18](#_Toc255068290)

[4.2 文档结构知识对排序的影响 18](#_Toc255068291)

[4.3 文档先验知识的融合 18](#_Toc255068292)

[4.4 算法描述 18](#_Toc255068293)

[4.5 实验及分析 18](#_Toc255068294)

[第五章 结束语 19](#_Toc255068295)

[5.1 本文工作总结 19](#_Toc255068296)

[5.2 未来工作展望 19](#_Toc255068297)

[参考文献 20](#_Toc255068298)

[致 谢 22](#_Toc255068299)

[个人简历 24](#_Toc255068300)

# 第一章 绪 论

## 1.1 引言

随着互联网规模的不断扩大，它已经成为了全球最大、最广泛使用的信息库。同时，越来越多的人开始使用互联网，据统计全球2008年底已经有10亿网民。并且，随着Web 2.0时代的到来，越来越多的网民不只停留在浏览信息的阶段，他们开始创造信息。这样，网络给我们带来了纷繁复杂、多种多样的信息。那么，如何有效检索这些海量数据获取感兴趣的信息成为当前重要的研究课题。信息检索是解决上述问题的关键技术。众多的网络搜索引擎如*Google、*百度等，应用信息检索技术，帮助人们快速访问网络信息。

Web信息检索是指从大量网页集合中找出与用户给定查询相关的网页子集，是处理海量文本的重要手段。Web信息检索，基本的过程包括：爬取互联网中的网页、对网页进行索引、根据用户请求返回相关的文档。由于信息量巨大，Web页面繁多，检索系统检索出的相关文档数量相当的多，为便于用户尽快地获得最相关的文档，大多数检索系统都把检索结果以其与用户提交的查询的相关程度进行排序后返回给用户。因此，如何准确、高效地为检索出的文档进行排序已成为信息检索研究的核心问题之一。

检索系统网页排序的主要任务是将和用户输入的查询最相关的网页排到最前面，为了完成这个任务，研究人员提出了各种排序模型。这包括，布尔模型、向量空间模型、概率检索模型和统计语言模型。当然，还有最近几年提出的基于监督学习方法的各种排序学习模型。由于监督模型方法，使用的特征数据基本都来自非监督方法，所以，本文着眼于基于非监督方法的网页排序问题研究。

## 1.2 研究现状

本文的工作基于信息检索领域已有的很多关于信息检索，统计语言模型的相关研究成果。在本节中，将对这些相关的研究成果做简单介绍。

### 1.2.1 信息检索中的排序模型

自从信息检索模型被提出以来，排序就一直是信息检索中的核心问题之一。为了解决排序问题，研究人员提出了各种模型。按照模型提出的时间顺序，包括布尔模型、向量空间模型、概率检索模型和统计语言模型。

布尔模型将一个文档组织成以关键词或者主题词为标识的倒排文本。倒排文本以倒排索引的形式呈现，索引的主键是关键词，索引的值是唯一标示文档的id。以倒排文本为基础，布尔模型在处理用户查询的时候，对于查询中的每个单词，去倒排索引中查找响应的单词，然后将结果进行逻辑操作，得到最后的返回列表。布尔模型是一种相当原始的检索方法，返回文档的排序手段也很单一。

向量空间模型是目前广泛应用于各个工业信息检索系统中的一个模型。其基本思路是：在信息检索中，文档或者查询的基本其所包含的词（检索单元）来表述的，可以定义由检索单元组成的向量来描述每一篇文档和每一条检索，再通过计算文档与查询之间的相关程度来判断文档与查询是否相关，与某一特定的查询的相关程度越高者被认为是与该查询越相关的文档。对于向量空间检索模型，需要定义向量来描述文档和检索的含义。通常的做法是，以所有包含在文档和查询中的检索单元为检索空间，将文档和查询以向量的形式表示出来。

概率检索模型(Probabilistic Retrieval Model)是另一种普遍使用的信息检索算法模型，它应用文档与查询相关的概率来计算文档与查询的相似度。通常，利用检索单元作为线索，通过统计得到每个检索单元在相关的文档集（对应于某查询）中出现和不出现的概率以及其在与该查询不相关的文档集中出现和不出现的概率，最终，利用这些概率值，计算文档与查询的相似度。这类模型的典型代表就是BM25算法。

### 1.2.2 统计语言模型

## 1.3 本文主要研究内容

本文针对经典语言模型中，对文档模型多元组元组间关系考虑不足、平滑算法对数据层次考虑不足、对结构信息利用不足这三个问题，开展研究。从真实的数据出发，揭示经典方法中可以改进的地方。同时，在理论上提出三种改进的算法：多维多元组算法、多层数据集平滑算法和结构模型估计算法。同时，本文还探讨了以上三种算法的实现方式和效率问题。

### 1.3.1 多维多元组算法研究

多元组算法是统计语言模型的基本算法，传统算法中，将元组独立考虑，分别统计，然后进行综合。多维多元组算法，在传统算法的基础上对元组之间的关系加以考虑，将连续元组的权重适当提高，从而获得更准确的文档模型估计。

将多重元组纳入考虑，随之带来的就是计算上的巨大负担。传统语言模型在线计算文档模型的时候，能够利用倒排索引非常快速的获得结果。但是，考虑多维多元组之后，速度必然会有很大下降。本文在提出算法之后，还对算法的高效实现提出了相应的解答方法。

为了验证本文提出的算法，本文在lemur的基础上，实现了相应的算法，并且在Trec Web数据集之上进行了实验。分别使用Trec的3个task来对比经典语言模型。对结果进行了研究和分析。

### 1.3.2 多层数据平滑算法研究

未见词的平滑是维持语言模型性能稳定的关键之处，以往的平滑算法，只使用单一的数据集进行平滑，这样的状况下，带来的副作用是，对于未见词较多的查询，其排序性能有所下降。

鉴于互联网是由人为组织起来的网络，其中任何一个页面它所在的目录和网站都有其人为分类的因素在起作用。本文利用这一特点，引入页面所在的目录和网站本身的信息来进行为见词平滑。

本文在提出算法之后，同时设计了如何存储这一平滑算法所需要的多层数据的结构，并且进行了大规模真实网络数据实验。实验显示，该算法本身具有很好的平滑性能。

### 1.3.3 结构模型估计算法研究

除了对于传统的语言模型研究热点p(q|d)的改进，本文对于文档先验概率p(d)的估计也进行了研究。文档先验概率的估计，与用户查询无关，本文将考虑多种可能影响排序的节后信息。并引入一套系统的方法，结合多种不同的结构特征对p(d)进行估计。

## 1.4 本文结构

本文共分为五章，各章节内容和结构安排如下：第一章是绪论，概括介绍本文的研究背景、研究内容和研究目标。第二章综述相关工作，介绍信息检索，排序模型和统计语言模型。对信息检索，排序模型、统计语言模型现有算法进行综述，作为后续章节工作的基础。第三章详细介绍文档模型算法和平滑算法，并在此基础上介绍一种新的文档模型算法和平滑算法。描述多维多元组算法和多层数据平滑算法以及如何高效实现这两种算法。第四章对经典语言模型没有加以利用的文档间结构信息加以整合利用，提出了基于bayes信念网络结构模型估计算法。在第五章，对本文的工作进行了总结与展望，总结本文工作，并展望下一步的研究工作。

# 第二章 相关工作综述

本文的研究得益于统计学习领域和信息检索领域的相关研究成果，本章详细介绍与本文相关的研究工作和成果，包括信息检索模型、统计语言模型作为后续章节工作的基础。

## 2.1 信息检索

信息检索是指从大量非结构化的文档集合中找出与用户给定查询相关的文档子集，是处理海量文本的重要手段。

一个文档检索的基本过程通常包括以下三个步骤：首先，用户可以从某一终端将其查询输入到检索系统中；之后，检索系统针对用户的查询，通过适当的算法，在已经建立了索引的文档集中进行检索，获得与用户查询相关的文档集；最后，检索系统为用户提供与其查询相关的文档集。通常，检索系统将所提供的相关文档集按照与用户查询的相关程度进行排序，最相关的文档排在最前面。

根据对相关文档判定方法的不同，信息检索模型可以分为以下五类经典模型：布尔模型、向量空间模型、概率统计模型、统计语言模型、基于有监督学习的检索模型等，以下将分别介绍这些检索模型。

### 2.1.1 布尔模型

布尔模型[24]（Boolean Model）是一种建立在集合论和布尔代数上的比较简单的检索模型。在经典布尔模型中，文档被表示成词项的集合。每个词项在每篇文档中只有两个值1 和0：“1”表示该词项出现在该文档中；“0”表示未出现

在该文档中。用户的查询用布尔表达式的形式来进行描述，支持逻辑与（AND）、逻辑或（OR）和逻辑非（NOT）的操作，只有满足该布尔表达式的文档才被认为是相关的文档，否则就是不相关的。

经典布尔模型存在着一些缺陷，主要问题包括：没有提供词项的权重信息，检索系统无法区分文档中不同的词项对相关性的贡献；检索系统不能提供相关度的排名；对于相关性的二值判定过于严格等。

为解决经典布尔模型的诸多问题，研究者们提出了扩展布尔模型（Extended Boolean Model）。在该模型中，文档和查询的相关度不再是0 和1，而是区间[0,1]中的一个实数，从而使得对文档的相关度排名成为可能。

### 2.1.2 向量空间模型

向量空间检索模型[37][38]（Vector Space Model）是信息检索领域中广泛使用的一种信息检索模型。其基本思路是：在信息检索中，文档或者查询的基本含义都是通过其所包含的词（检索单元）来表述的，可以定义由检索单元组成的向量来描述每一篇文档和每一条检索，再通过计算文档与查询之间的相关程度来判断文档与查询是否相关，与某一特定的查询的相关程度越高者被认为是与该查询越相关的文档。对于向量空间检索模型，需要定义向量来描述文档和检索的含义。通常的做法是，以所有包含在文档和查询中的检索单元为检索空间，将文档和查询以向量的形式表示出来。

向量空间检索模型通常使用基于文档集合的统计频率的权值，也被称为

tf-idf 权值。tf-idf 权值由两部分组成，一部分是检索单元在文档中出现的频率（term frequency, tf），另一部分则被称为文档频率的反转（inverse document frequency, idf），通常，对于一个给定的检索单元tf-idf 权值是tf 与idf 的乘积。

为了方便说明问题，作如下定义：

*m*: 整个检索空间的大小。

*d*: 文档集合中文档的总个数。

*tfij*: 检索单元*tj*在文档*di*中出现的次数（*tf*）。

*dfj*: 在整个文档集合中，包含检索单元*tj*的文档的个数（*df*）。

则文档频率的反转定义为：

 （2.1）

对于给定的某一个文档，描述该文档的向量由m个元素组成，分别对应着文档中出现的m个检索单元。每一个元素的权值根据其所对应的检索单元在文档中出现的频率以及该检索单元在整个文档集中出现的频率两项因素共同决定

 （2.2）

使用公式（2.2）中的作为向量中各元素的权值，对前面所述的向量进行进一步调整，这样的向量更精确地描述了文档和查询的内容。

对于向量空间检索模型，不仅需要定义向量来描述文档和查询的含义，还需要选择适当的方法来计算文档与查询的相关程度以判断文档与查询是否相关。

原则上讲，只要是能够判断出描述文档与查询的各个向量方向的接近程度各种计算方法都可以用来作为文档与查询相关程度的判断依据。很自然的，可以考虑使用向量夹角的余弦作为文档与查询相关程度的判断依据。如前所述，在检索空间Ω 中，定义文档D 和查询Q 的相似度（Similarity Coefficient, SC）为：

 （2.3）

其中，是对该文档含义的一系列描述。当检索单元*ti*出现在文档*D*中时，为1，否则为0；当检索单元出现在查询*Q*中时，为1，否则为0。都采用*tf-idf*权值。

### 2.1.3 概率检索模型

概率统计检索模型[35]（Probabilistic Retrieval Model）是另一种普遍使用的信息检索算法模型，它应用文档与查询相关的概率来计算文档与查询的相似度。通常，利用检索单元作为线索，通过统计得到每个检索单元在相关的文档集（对应于某查询）中出现和不出现的概率以及其在与该查询不相关的文档集中出现和不出现的概率，最终，利用这些概率值，计算文档与查询的相似度。

BM25检索算法是一种经典的概率统计检索算法，由Roberston 1994年在TREC-3上提出，BM25计算文档D和查询Q的相似性。对查询Q中的每一个检索单元，一共有三个权值与之相关：

 （2.4）

其中：是由用户指定的参数，是检索单元在Q中出现的频率qtf （within query frequency）。

 （2.5）

其中：k和b是用户指定的参数，是检索单元在D中出现的频率tf （within document frequency），L是正则化之后的文档长度，计算方法为原始文档长度除以文档集合中平均的文档长度。

 （2.6）

其中：N表示文档集合中文档的总数；R表示与查询q相关的文档总数；n表示含有检索单元的文档总数；r表示与q相关的文档中，含有检索单元的文档数。

这样，在BM25公式中，查询Q和文档D的分值为：

 （2.7）

近年来， Robertson[36]等提出了一种简单的基于BM25的改进，改进算法能够同时计算具有多个域的文档和查询的相似度，克服了BM25在这方面的不足。

### 2.1.4 统计语言模型

近些年来，统计语言模型（Language Model）在信息检索领域取得了令人瞩目的效果。1998年，Ponte和Croft在SIGIR 会议上发表了一篇名为“A Language Modeling Approach to Information Retrieval”[31]的论文，由此开创了一个新的研究课题：统计语言模型在信息检索中的应用。随后几年，众多研究人员不断加入到该课题的研究工作中来，取得了丰硕的研究成果。许多实验结果显示：基于统计语言模型的方法在检索性能上普遍优于以前普遍采用的向量空间模型方法。

利用语言模型进行信息检索，查询*Q*一定时，检索出的文档*D*根据后验概率*P*（*D*|*Q*）来排序。根据贝叶斯公式可得：

 （2.8）

这便对应了一个信源-信道模型：信源模型*P*（*D*）和信道模型*P*（*Q* | *D*），该方法认为：如果文档*D*和查询*Q*相似度越高，则通过观察信道的输出（*Q*）便能获得信源（*D*）的更多信息，这也是通过语言模型方法进行信息检索的一个基本假设。

利用语言模型进行信息检索的基本过程是：对每一篇文档均建立一个模型，计算每一个模型产生某主题（查询）的概率值（相当于其他模型中文档-查询相似度），然后对这些概率值进行排序，返回排序结果，即为该主题的检索结果。

对于统计语言模型，在后面章节还会进行更加详细的介绍。

### 2.1.5 基于监督学习方法的排序学习模型

近些年，机器学习方法在各个领域取得了很好的成绩。在信息检索领域，机器学习方法也被引入，发展出了很多适用于文档排序学习的模型。

下面，给出排序学习问题的形式化定义。给定一个输入向量的集合X

 （2.12）

和其对应的标号

 （2.13）

其中：m 表示训练样本的数目，n 表示输入向量的维度。S =（X, Y）为某一分布的独立同分布（iid）的样本集合，也称为训练集合。独立同分布即意味着任意一个样本既不依赖于其他样本，也不依赖于其下标i。

排序学习的目的是寻找一个能够精确预测数据的未知标号y 的决策函数，也就是说，排序学习所学习的预测函数将最小化对排序的预测错误，预测错误的定义为的概率。



图2.1 排序模型

图2.1 （a） 用于排序的效应函数*U*（*x*），通过把数据点投影到*U*（*x*）上可以得到对数据的排序，是各个序类别之间的边界。（b） 数据在*U*（*x*）上的投影

如图2.1所示，在排序学习中，其对象按照效应函数（排序函数）进行排序：，每一个对象的效应值为其在上的映射值。

综上所述，排序学习中学习目的是从决策函数集合寻找一个最优决策函数，能够精确的预测数据点的未知标号*y*。

排序学习在近年来引起了机器学习研究者们的极大兴趣，有很多研究成果发表。用机器学习的方法来解决排序学习问题大致可以分为以下三类：基于回归的排序学习、基于分类的排序学习和基于顺序回归的排序学习。基于顺序回归的排序算法是当前排序学习研究的热点，根据对训练数据处理手段的不同又可以分成三大类别：基于数据点的排序学习算法、基于有序对的排序学习算法以及最新提出的基于列表的排序学习算法。

## 2.2 信息检索评价指标

在信息检索的研究工作中，检索性能的评价是一个非常重要的问题，它是衡量各种检索模型好坏的量化指标。

由于用户查询条件中所固有的模糊性，信息检索系统检索出来的文档集合不一定全是用户所希望的，因此有必要对这些文档集合根据其与用户查询条件的相关性进行排序，相关程度越高的文档排得越靠前为好，并以此来判定信息检索系统检索出的文档集合满足用户查询条件的程度，这种评测就是检索系统的检索性能的评测。

在本文的实验中，我们主要使用MAP 和NDCG 来评价排序结果序列的性能。

### 2.2.1 MAP

MAP（Mean Average Precision）是用来衡量算法对多个查询的平均排序结果。MAP 的计算公式为：

对于某一个查询，其平均查准率计算公式为：

 （2.9）

其中：j 表示排序的位置，M 是检索到的文档总数，Precision（j）是前j 个检索到的文档的查准率，pos（j）是一个0-1 函数，如果排在第j 个文档是相关的，其值为1，否则为0。这样平均查准率的均值MAP 的计算公式为：

 （2.10）

在计算MAP 时，由于其要求文档被标注成两个等级：相关和不相关，因此把标注为相关的文档（definitely relevant）看成相关的文档，其他两个级别的文档（部分相关（partially relevant）和不相关（not relevant））都看成不相关文档。

尽管MAP已广泛用作信息检索系统中检索算法的评测方法，但也有其限制：

1、MAP 将查询和文档的相关简化成为0-1 关系，一个查询和一个文档要么相关，要么不相关。而实际上相关是一个程度的量，0-1 关系并不能准确的反映查询和文档的相关关系，例如在“相关”和“不相关”之间还可能存在着“部分相关”的文档。

2、在实际的检索中，用户往往只是浏览位于序列头部的结果，因此位于序列头部的结果对排序性能的影响越大。然而MAP 并没有很好的解决这两个问题，因此，我们又使用NDCG 来评价排序结果中顶部序列的准确性。

### 2.2.2 NDCG

NDCG（Normalized Discounted Cumulative Gain）对传统的评价标准做出了改进，这些改进基于以下两个原则：1、在信息检索中，相关可以分为多个级别，高度相关的文档比部分相关的文档更有价值，其在评价中应该赋予更大的权值；2、文档在序列中的位置越靠后，这个文档的价值越小，从用户的角度考虑，由于时间、精力以及从已经阅读过的文档中所得到了信息等原因，用户可能根本不会去看这些文档。NDCG 用来评价排序结果中顶部序列的准确性。

能根本不会去看这些文档。NDCG 用来评价排序结果中顶部序列的准确性。在这种评价方法中，每一个文档都对它所在的位置有一定的贡献，其贡献值与文档的相关度有关，然后，从1 到n 的所有的位置上的贡献值都被加起来作为最终的评价结果。这样，一个一定长度的文档序列被转换成了一个相关分值的序列。

给定一个排序后的文档序列，在第r 位的NDCG 值NDCG@r 的计算公式为

 （2.11）

其中：是第j 个文档的级别，是归一化参数，它使得最优的排序的NDCG@r的值始终为1；如果结果序列中文档的个数n 要少于r，则计算公式返回NDCG@n的值。

在计算NDCG 时，我们把相关映射为数值2、部分相关为1、不相关映射为0。

## 2.3 统计语言模型详述

统计语言模型最早应用于语音识别领域，目的是根据已经识别的语音元素，预测语音序列中下一个单词。如公式1-1所示：



公式1-1：用于语音识别的语言模型

统计语言模型通过把预测问题转化为简单的条件概率求值而解决这些问题。首先，模型认为第i个单词是什么，依赖于其前n个单词的状态。那么，通过统计前n个单词序列，计算条件概率而得到第i个单词的似然值。能产生最大的似然值的单词w，就是我们估计的结果。

由于语言模型利用了统计学的基本方法，将问题转化为条件概率和Bayes公式能够处理的方式，能够最大限度的去拟合真实数据，并且最大限度的去利用先验信息，其应用非常的广泛。除了在语音识别领域，语言模型又广泛的应用于机器翻译、自然语言处理等学科，并且都取得了相当不错的效果。

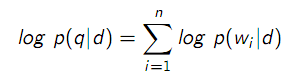
在1998年，由Ponte和Croft首次将语言模型方法引入到信息检索领域。其背后的思想和语音识别中的语言模型略有不同，如公式1-2所示：

公式1-2：信息检索中的语言模型

其中，d为文档，q为用户输入的查询，模型左侧是给定用户查询q，计算每个文档d和q相关的概率。根据Bayes公式，得到右侧的结果。由于每个查询q的分布和文档d无关，直接忽略p(q)而得到了最后的模型形式。

统计语言模型背后的思想很简单，它为每个文档d建立一个模型，计算由该文档产生某个查询q的概率。最后，利用文档对于查询的似然率来排序文档。

在应用于信息检索的统计语言模型提出之后，相关领域的研究人员对其进行了大量的研究。其中大部分的工作都是围绕p(q|d)的计算来进行的。一个查询q由其中的n个单词w1,w2…wn组成,那么计算p(q|d)的方法如公式1-3:



公式1-3

对于在文档d中出现的单词，可以通过统计其频率直接得到相应的p(wi|d)，对于在文档d中未出现的单词，就需要模型通过某种算法来做平滑处理，以使得任何查询q的生成概率不为0.那么，如何准确的计算p(q|d)包括两个重要的部分，一是在文档d中出现单词的计算方法，另一个是对于未出现单词的平滑方法。对于前者，以往的研究人员提出了基本的n-gram模型以及很多改进n-gram模型。

[1]、n-gram模型

对于出现单词的统计，以往的学者基本采用每个单词独立分析的做法，即分别统计用户查询q中的n个单词在文档d中的频率，然后取和来计算p(q|d)。

语言模型模型本身在语音识别和机器翻译领域，也有学者采用同时考虑多个单词的方法，这种方法统称为“n-gram”方法。N-gram方法同时考虑由n个单词组成的单词组，来分析其出现的联合概率，从而能够更准确的估计p(q|d)。

随着n-gram方法提出之后，研究人员很快就发现了其中的一些问题，比如由于数据稀疏造成的估计偏差。随后，在自然语言处理和机器翻译领域，很多学者通过引入语义信息，来丰富n-gram，获得了更好的识别性能。这其中包括基于主题的n-gram模型和跳跃式n-gram模型。

基于主题的n-gram（Topic-based N-gram）

Gildea在1999年[22]提出了一种基于主题的n-gram方法来用于在线预测语音序列中下一个词的概率。论文首先将不同topic的语料库进行统计，在进行预测时，分别按照一个固定的权值向量λ来进行加权，求得最后的结果。其主要思想可以有公式1-4表示。类似的方法还包括，Peter在其论文[23]中提出的基于分类n-gram的方法。



公式 1-4

跳跃式n-gram（Skip-based N-gram）

跳跃式n-gram模型打破了以前n-gram模型对于约束的限制，及该模型认为，一个单词出现与否，不止依赖于和它相连的n个单词，而有可能依赖于距离它有一定距离的单词，所以命名为跳跃性n-gram，其思路可以由公式1-5表示。



公式 1-5

但是，n-gram方法也只考虑固定长度的单词相关信息。当纳入考虑的单词过短的时候，不能反映数据的真实分布情况。当纳入考虑的单词长度过长的时候，会带来计算复杂度的上升和受到数据稀疏性的影响。为了处理这些问题，Bimbot在论文[19]中提出了multigrams算法。该算法应用于语音识别中，计算一个句子出现的概率。其采用不等长序列建模方法，将一个句子中的单词，按照不超过n元组的方法，任意分割相邻词，组成句子的分段，然后分别计算这些分段组合在一起出现的概率，将最大的概率作为该句子出现的概率。以包含四个单词(w1,w2,w3,w4)的句子***W***为例，其生成概率在n-multigrams（n<=3）中，用公式1-6计算得出。作者在论文中论述，该多元组方法取得了高于一般n-gram方法的性能。但是，该方法并不能直接引入到信息检索模型当中，同时信息检索相关领域也没有多维多元组成功应用的论文。



公式1-6

[2]、平滑算法

语言模型中对为简单词的平滑算法吸引了更多学者的注意，其大致分为两类。一种为backoff算法，一种为插值算法。所谓backoff算法，就是指当某个单词未在文档d中出现的时候，使用一个backoff数据集中统计得到的结果来代替该单词在文档d中的频率。相对于backoff方法，插值算法通过减少可见词的频率来补偿未见词。经典的插值算法主要有三个，它们是：Jelinek-Mercer算法，基于Dirichlet先验概率的贝叶斯平滑算法、Absolute discounting算法。Zhai在他关于语言模型平滑算法综述的论文中又结合了Jelinek-Mercer算法和基于Dirichlet先验概率的贝叶斯平滑算法，提出了两阶段语言模型平滑算法。

[3]、先验结构信息

除了计算p(q|d)，如何准确的估计文档的先验概率p(d)也是语言模型实用性的重要课题之一。但是这方面获得的研究和关注都不如语言模型的另一半：平滑算法。大部分的研究工作，都假设p(d)为一个一致分布，而将其忽略，主要集中考虑p(q|d)。然而，也有学者考虑使用网页中的非内容信息来估计p(d)，以图提高语言模型的性能。这其中，包括Kraaij在其论文[24]中提出的尝试用PageRank和URL等信息计算文档的先验概率的方法和通过收集各种结构信息来分析先验概率的方法。但是对于，如何系统的引入结构信息来估计先验概率p(d)和引入何种结构信息更为重要这方面，没有深入的研究。

# 第三章 改进内容模型生成算法

## 3.1 引言

内容模型p(q|d)的估计，直接影响到语言模型的准确性。传统语言模型中，同等对待查询中的所有一元组，这样做的原因有两个。其一，使用一元组已经能够比较准确的估计内容模型；其二，多元组的引入会造成性能的大幅下降。本章前第二节讨论了，在大规模网络数据集的条件下，一元组本身会被噪声极大程度的干扰，所以引入多元组会对语言模型本身的性能带来提升。在第三节，本文提出了基于多元组的新的算法，同时考虑到了计算性能问题，设计优化算法对计算性能进行提升。

之后，对于语言模型中对未见词进行平滑的方法进行了讨论。利用互联网本身的机构信息，来进行多层次的语言模型平滑。在提出算法之后，对算法使用的数据结构和实现策略进行讨论，提出了一个行之有效的实际文件结构来保存多层平滑数据。最后一节的实验分析，验证了这种平滑算法的效果。

## 3.2 网页噪声和多元组的影响

## 3.3 多维多元组内容模型估计算法

## 3.4 实验及分析

## 3.5 多层数据平滑算法

## 3.6 实验及分析

# 第四章 结构模型生成算法

## 4.1 引言

## 4.2 文档结构知识对排序的影响

## 4.3 文档先验知识的融合

## 4.4 算法描述

## 4.5 实验及分析

# 第五章 结束语

## 5.1 本文工作总结

## 5.2 未来工作展望

# 参考文献

[1] Ponte J. and Croft W. B. A language modeling approach to information retrieval. Proceedings of the ACM SIGIR ‘98

[2] Miller D. H. A hidden Markov model information retrieval system. Proceedings of the ACM SIGIR ‘99

[3] C. Zhai and J. Lafferty. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. ACM Transaction of Information System 2004.

[4] Tom Mitchell. Machine Learning. McGraw Hill, 1997. Pang-Ning Tan, Michael Steinbach and Vipin Kumar. Introduction to Data Mining. Pearson Education, 2005.

[5] Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval. New York, NY, USA: Addition Wesley, 1999.

[6] Richard O. Duda, Peter E. Hart, David G. Stork. Pattern Classification. John Wiley and Sons, 2000.

[7] Lancaster, F. W. Information retrieval systems: characteristics, testing and evaluation. 2nd Ed., New York: John Wiely and Sons, 1979.

[8] Salton, G., Wong, A., and Yang, C. S. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11):613~620.

[9] Robertson, S. E., S. Walker, M. Hancock-Beaulieu and M. Gatford. Okapi in TREC3. In Proceedings of Text REtrieval Conference, Gaithersburg, USA. U.S. National Institute of Standards and Technology, NIST Special Publication 500-225: 1994. 109~126.

[10] 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 on Research and development in information retrieval, pages 111-119, 2001.

[11] Freund, Y., Iyer, R., Schapire, R., and Singer, Y. An efficient boosting algorithm for combining preferences. Journal of Machine Learning. Research 4, 2003, 933–969.

[12] Nallapati, R. Discriminative Models for Information Retrieval. In: Proceedings of the27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom, 2004. 64~71.

[13] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. Overview of the TREC 2003 web track. In TREC, pages 78–92, 2003.

[14] TREC：http://trec.nist.gov/

[15] Lemur：http://www.lemurproject.org/

[16] 徐君. 用于信息检索的代价敏感排序学习算法研究: [博士学位论文]. 天津：南开大学，2005.

[17] Tie-Yan Liu, Jun Xu, Tao Qin, Wenying Xiong, and Hang Li. LETOR: Benchmarking “Learning to Rank for Information Retrieval”. Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval, Amsterdam, The Netherlands, 2007.

[19] F. Bimbot, R. Preraccini, E. Levin and B. Atal. Variable-Length Sequence Modeling: Multigrams. IEEE Signal Processing Letter, VOL. 2, No. 6, June, 1995

[20] S. Deligne and F. Bimbot. Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams. Acoustics, Speech and Signal Processing, 1995.

[21] R. Kneser and V. Steinbiss. On the dynamic adaption of stochastic language models. Acoustics, Speech and Signal Processing, 1993.

[22] D. Gildea and T. Hofmann. Topic-Based Language Models Using EM. European Conference on Speech Communication and Technology, 1999

[23] Peter F. Brown and Peter V. deSouza. Class-Based n-gram models of natural language. Computational Linguistics, Volume 18, Issue 4, 1992

[24] Wessel Kraaji, Thijs Westerveld and Djoerd Hiemstra. The Importance of Prior Probabilities for Entry Page Search. SIGIR 2002

# 致 谢

在本文工作的进行过程中，我得到了很多人的帮助，我要向他们表示感谢。

感谢我的导师黄亚楼教授，在黄老师的指导和帮助下我才得以完成学业。三年来黄老师让我学到了很多知识，提供了很多实践的机会，并给予精心的指导。特别是在本文的进展过程中,黄老师多次指导、督促，保证了我的论文工作的进度和工作质量。可以说，没有黄老师对我的教导就不会有我今天的进步。

感谢实验室的谢茂强老师，在进行信息检索项目时，不断地给予我谆谆教诲和悉心指导，让我在高水平的项目中锻炼自己，提升自己的研究能力和水平。同时，谢老师在他在我做研究遇到问题的时候督促我的学习并给予我做人处事方法的指导。谢老师他精湛的技术和踏实的研究态度是我今后学习的榜样。

感谢实验室的博士师兄王扬。王扬师兄在我的毕业论文工作中给予了大量的帮助，提供了丰富的意见和建议，对很多细节之处都提出了建设性的意见。在王扬师兄的帮助下我才得以完成本文。同时感谢信息检索项目组其他成员，李栋、倪维健、刘杰博士，刘金莉、郑楠等同学。非常感谢大家在我撰写论文工作过程中给与的帮助和鼓励。

感谢武涛、李超、邹剑、刘洋、任莉媛、刘金莉，在三年的同窗生活中，他们给与我莫大的帮助，给我在生活、学习中带来快乐，这段欢乐时光是我永远不会忘记的。

感谢杨波，郑楠，刘辛，杨俊丽，这些我可爱师弟师妹们。

衷心感谢我的父母。感谢你们含辛茹苦地哺育我长大、教会我做人并给了我一个宽松自由的成长环境。祝愿我的父母身体健康、生活快乐。

最后，祝愿培养我的南开大学软件学院和南开大学智能信息处理实验室蒸蒸日上，更创辉煌！

蒯宇豪

2009年4月

南开大学智能信息处理实验室

# 个人简历

**一、个人信息**

姓名：杨波 性别：男 出生日期：1985-1-10 籍贯：天津宝坻

**二、学习经历**

2007-09至2010-06 南开大学软件学院 计算机软件与理论 工学硕士

2003-09至2007-06 南开大学软件学院 软件工程专业 工学学士

**三、在学期间参与的项目**

