JIYU ZIRAN YUYAN CHULI DE XINXI JIANSUO

基于自然语言处理的信息检索

李卫疆　李东军　王玲玲　著

云南大学出版社
YUNNAN UNIVERSITY PRESS

基于自然语言处理的
信息检索

李卫疆　李东军　王玲玲　著

图书在版编目（CIP）数据

基于自然语言处理的信息检索 / 李卫疆，李东军，王玲玲著. --昆明：云南大学出版社，2014
ISBN 978-7-5482-2071-8

Ⅰ. ①基… Ⅱ. ①李… ②李… ③王… Ⅲ. ①自然语言处理-应用-情报检索-研究 Ⅳ. ①G252.7

中国版本图书馆 CIP 数据核字（2014）第 165201 号

基于自然语言处理的信息检索

李卫疆　李东军　王玲玲　著

责任编辑：叶枫红
封面设计：刘文娟
出版发行：云南大学出版社
印　　装：昆明市五华区教育委员会印刷厂
开　　本：787mm×1092mm　1/16
印　　张：16.875
字　　数：347 千
版　　次：2014 年 7 月第 1 版
印　　次：2014 年 7 月第 1 次印刷
书　　号：ISBN 978-7-5482-2071-8
定　　价：52.00 元

地　　址：昆明市翠湖北路 2 号云南大学英华园内
邮　　编：650091
网　　址：http://www.ynup.com
E-mail：market@ynup.com

目　录

第一章　导　　论

基于全文索引的信息检索发展至今已有十几年的历史。在这十几年里，研究者们不断尝试着将自然语言处理应用到信息检索中，试图提高信息检索的效果。自然语言处理包括自然语言处理技术和自然语言处理资源。在信息检索中使用自然语言处理技术的尝试大部分没有获得好的效果。尽管在小部分实验中信息检索效果有了一些提高，但改进的程度往往很小，为此而使用的复杂的自然语言处理技术则有着巨大的计算消耗，很难被认为是值得的[1]。在信息检索技术中结合自然语言处理资源，例如词典，实验结果也不能令人满意[2]。

信息检索中常常使用到的自然语言处理技术包括去除停止词、取词根、词性标注、词义消歧、句法分析、命名实体识别、指代消解等，自然语言处理资源包括的则是 WordNet 和 HowNet 这样的词典。

自然语言处理技术被用来对自然语言进行处理，目的是让计算机“理解”自然语言的内容。而信息检索中所涉及的文档和查询都是用自然语言描述的，因此，在信息检索中使用自然语言处理以提高其效果的想法被寄予了厚望。信息检索可以看作是用查询和文档内容进行匹配的过程，匹配的单位通常是查询和文档中的词。基于词匹配的信息检索中存在着与自然语言特点相关的问题，同样促使研究者们求助于自然语言处理[3]：

- 不同的词可以表达同一个意思
- 同一个词可以表达多种意思
- 对一个概念的描述可以有不同的角度
- 同一个词在不同的领域会有不同的意思

自然语言处理技术最大的难点在于自然语言中有各种级别的歧义难以消除，包括词汇级别、句法级别和语义级别[3]。歧义的存在使计算机在“理解”自然语言时发生了困难，并很可能出现错误。这无疑为自然语言处理没能为信息检索带来较大帮助提供了一个解释。然而事实上这个解释并不全面。因为和信息检索的效果相比，自然语言处理的很多技术实际上已经有了很高的准确率—— 尽管直接用两者的准确率进行比较并不科学。

因此，本书对信息检索中使用自然语言处理的研究工作进行综合分析，总

结出哪些自然语言处理技术和资源对信息检索有帮助，需要达到怎样的精度才能使信息检索的效果有较大提高，并试图对未来自然语言处理在信息检索中的使用方向进行归纳和展望。

信息检索(Information Retrieval，IR)和自然语言处理(Natural Language Processing，NLP)已经共存了几十年。传统的检索模型不能使用NLP技术，因为过去的自然语言处理技术不是鲁棒的、可靠的，或者实践中不足以处理大型语料。因此，传统的信息检索系统性能低下。当前，很多研究者相信各种NLP技术能够容易地应用到IR中。本书的目的就是把语言学家、自然语言处理和信息检索系统的想法融合到一起，并且把它们应用到信息检索中去。

作者假定读者没有或者仅有很少的语言学或信息检索的背景知识。因此，读者能够在书中了解到非常完整的信息检索的介绍。Van Rijsbergen[4]、Frakes和Yates[5]提供了很好的信息检索的综述。Van Rijsbergen对统计方法给出了详细的描述，而Frakes和Yates的综述对应用于信息检索的各种算法进行了深入的剖析，并且给出了各个算法的实现源代码。

1.1 信息检索概述

搜索服务是所有互联网用户使用最多的服务之一，随着互联网上信息的日益丰富，人们越来越依赖搜索引擎来寻找所需信息。在搜索引擎的背后，是信息检索技术在起作用。信息检索这一术语最早是由Calvin N. Mooers在1950年的Zator Technical Bulletin (NO. 48)[6,7]中公开提出的。信息检索最初主要是应用于图书馆中的文献检索，1954年美国海军兵器中心(HOTS)图书馆在IBM 701型号计算机上成功建立了世界上第一个计算机文献检索系统。随着计算机技术与互联网络的发展，信息检索系统也从批处理方式的文件检索发展到70年代后的联机情报检索，以至于到现在的大规模的互联网信息检索和数字图书馆等领域。现今乃至未来，信息检索技术都将对我们的科学研究和日常生活产生积极而又重要的影响。

人们对信息检索的研究伴随着应用已经有很多年的历史。在20世纪50年代，当计算机被图书馆等部门用于存储管理文档时，信息检索作为一个研究领域也随之诞生；到80年代，信息检索领域在索引模型、文档内容的标识、匹配策略以及排序算法方面取得了大量的研究成果；今天，随着社会信息化程度的快速提高，因特网日益普及，数字图书馆和各种各样的电子信息载体不断涌现，信息的总量以惊人的速度不断地膨胀，信息处理技术迫切需要更有效的理论和方法来处理如此海量的信息。信息检索适应这一要求并成为当前信息处理研究

领域中的研究热点，布尔模型、向量空间模型、概率模型、统计语言模型、基于监督学习的检索模型等先后被提出并取得了良好的应用效果。目前，对检索模型的研究仍然是信息检索研究的热点，各种新的检索模型不断涌现。

互联网的发展，使它正成为人们生活的一部分。由于互联网上信息太过庞大，因此，人们必须通过搜索引擎来寻找所需要的信息。搜索引擎以一定的策略在互联网中搜索、发现信息，对信息进行理解、提取、组织和处理，并为用户提供检索服务，从而起到信息导航的作用。对使用搜索引擎频度的调查结果显示：有19.40%的网民经常使用搜索引擎。根据中国互联网络信息中心(China Internet Network Information Center，CNNIC)最新发布的《2007中国搜索引擎市场调查报告》显示，截至2007年9月，我国网民数量已经达到1.62亿，其中44.71%(超过7000万)的网民经常使用(每天多次使用)搜索引擎，可见近半数网民高度依赖搜索引擎提供的服务。另外，每天使用一次搜索引擎的用户比例占到17.2%，这意味着每日使用搜索引擎用户所占比例高达61.91%。由此可见，网民的搜索依赖性呈现增强趋势。搜索引擎成为网民的第二常用的网络服务，仅次于浏览新闻。随着人们对搜索引擎服务的日益依赖，搜索引擎市场将是一个巨大的潜在市场。在众多搜索引擎的背后，是信息检索技术在起作用。

信息检索的基本任务是基于用户的查询在数据集中找到与用户查询相关的文档。在本书中我们谈到信息检索主要指文本检索，所谓文本检索，就是在文档集合中检索信息。一个常见的文档集合的例子就是Internet上的可获得的信息。

1.1.1 信息检索中的重要概念

信息检索(Information Retrieval，IR)是一个具有多重含义的术语[8]。在本书中，我们谈到的信息检索主要指文本检索。这一用法已经被广泛接受[9]。信息检索可以指检索其他类型的信息，如图像、声音等。本书中，我们用信息检索(IR)描述这样一个系统：用户给出一个查询，系统返回一个排序的文档列表。排序是根据文档与查询的相关度而定的。当然这种相关只是系统的一种推断，并不一定与用户的真正需求相吻合。

信息检索问题的形式化的描述为：给定一个由n个文档d1，d2，d3，…，dn组成的文档集D和一个查询q，在文档集D中查找与给定查询q相关的所有文档，并对查到的文档按已给定查询的相似度度量函数RSV(q，di)排序，然后返回给用户。这里$1 \leqslant i \leqslant n$，并且函数值越高意味着相关程度越高。在信息检索应用中，我们谈到两个概念：词(terms)和文档(documents)。文档有时也

被称为条目(items)。文档是检索单元的统称，比如段落、章节、网页、文章、书等。一个索引词是预先选择的词，可以用来表示文档的内容。一般说来，索引词是名词或名词短语[10]。

信息检索的主要研究内容包括对信息的表示、存储、组织和访问，其目的在于让用户更加容易地访问到所需要或者感兴趣的信息[11,12]。信息检索的过程可以简单地描述为：用户根据其信息需求，组织一个查询字符串提交给信息检索系统，信息检索系统在文档集中检索出与查询相关的文档子集返回给用户(如图 1-1)[13]。

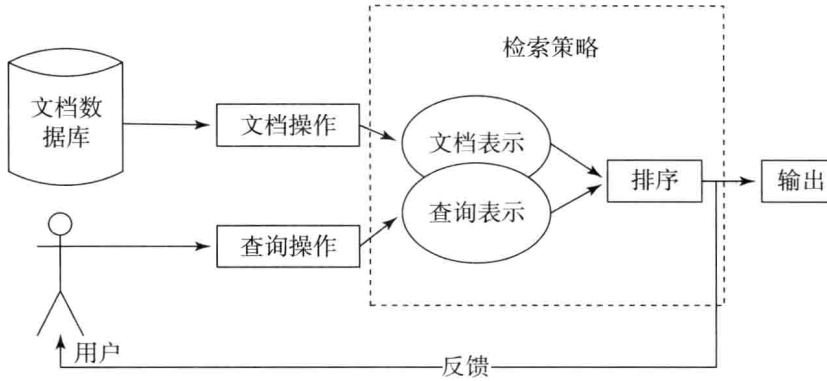

图 1-1　信息检索的过程图示

信息检索研究的对象——信息，可以有多种表现形式，包括数字、图形、图像、语言和文本等，其中文本是最普遍的信息表达方式[14]。本书所研究的对象是文本。

检索问题的研究基本上可分成两大类：基于语义的检索方法和基于统计的检索方法。基于语义的方法试图从语法和语义上理解自然语言来解决检索问题，但是这种方法需要投入较多的资源，如分类体系、语义词典、推理规则等，这些资源的完善程度受人力限制。目前在信息检索中占统治地位的仍是基于统计的方法，它不强求从语义上理解自然语言，只是简单地观察自然语言的特征，从统计学的角度寻找某些可以利用的信息。例如，通过观察一个查询词在文档中的出现频率以及它在整个文档集中的出现频率，可以直观地认为，如果这个词在某个文档中出现的频率高，那么这个文档很有可能是和查询代表的信息需求相关的；如果包含这个查询词的文档数目很少，那么进行检索时这个词的区分程度可能会很高[15]。在目前的技术水平下，基于统计的方法是我们最好的选择，它代价较小，却可以在一定程度上满足 IR 系统的性能需求。为了从文档集

中找出与用户查询相关的文档，检索系统必须根据用户给出的查询对文档集中的每篇文档做出是否与查询相关的判断，这种判断依赖于文档排序(Ranking)算法。文档排序算法依据各个被检索的文档与查询的相关程度，建立一个排列顺序，相关度高的排在前面。可以看出，设计文档排序算法是信息检索系统的关键技术。

文档排序算法的关键是文档排序函数 RSV (D, Q)，该函数给出了当前文档 D 与给定查询 Q 之间的相关度的测度值，在信息检索领域这种测度值称为相关状态值(Relevance Status Value, RSV)。RSV 越大，表明当前文档与该查询越相关。在基于统计的检索方法中，一个关键的问题就是怎样用形式化的方法对文档排序函数 RSV (D, Q)进行数学上的定义，这种定义依赖于检索系统对文档和查询的内在表示，依赖于检索系统对“相关度”做出何种解释并根据文档和查询的表示来对这种“相关度”进行相应的测度。

影响一个检索系统的性能有很多因素，其中最关键的是信息检索策略，包括文档和查询条件的表示方法、评价文档和查询相关性的匹配策略、查询结果的排序方法和用户进行相关反馈的机制等。经过相关科研人员近半个世纪的努力，陆续提出了一些有效的信息检索模型并被逐渐应用到相关的系统中。其中影响比较大的检索模型包括：布尔逻辑模型、向量空间模型、概率模型、语言模型以及新近提出来的基于监督学习的检索模型[16-18]。

最初，检索系统借用数据库的查询方法，采用布尔逻辑模型[19,20]，使用语词的布尔逻辑组合作为查询条件，从文档数据库中检索出满足查询条件的文档。布尔模型的优点在于它表达形式简单且形式化，易于理解；然而布尔逻辑模型起源于数据库管理系统，只能查找结构化、精确的数据信息，不能很好地解决无结构的文本信息检索，也不能实现查询条件与文档的部分匹配。

Salton 等人在 20 世纪 60 年代末提出了向量空间模型[21]，使用由语词构成的向量来表示文档与查询条件中的信息，并研制了基于向量空间模型的 SMART 实验检索系统[22]。用户无需构造布尔逻辑组合的查询条件，只需输入重要语词、短语、语句甚至一段文章，检索系统就能根据用户提交的查询条件构造查询向量，按照在检索词构成的向量空间中查询条件与文档向量的余弦相似性排序后得到检索结果，这样也就实现了查询条件与文档间的部分匹配。实践证明，尽管向量空间模型在许多方面依然和“现实”不符合，但实际效果比布尔逻辑模型改进了许多。

Roberson 和 Sparck 在 1976 年提出经典概率模型[23-25]，在概率的框架下解决信息检索问题。概率检索模型按照文档与查询“相关”和“不相关”的关系对文档进行排序，检索系统中文档与查询条件的相似性计算是基于概率排序原理，

即通过估计文档与用户查询条件的相关概率对文档集合进行排序[26-28]。基于以上原理，研究者们提出了二值独立检索模型(Binary Independent Retrieval, BIR)[29]、双泊松模型(2 - Poisson Model)[30]、BM25 检索模型[31]以及改进的 BM25 模型[32]。

语言模型研究最初是利用统计技术计算词汇间的依赖关系以帮助语音识别系统提高识别率。在 20 世纪 80 年代后期，语言模型开始被应用到其他相关领域。在 1998 年，Ponte 和 Croft 首次提出了将统计语言模型和信息检索相结合的新思路[18]。Ponte 和 Croft 最初提出的语言模型检索方法现在经常被称为“查询条件生成的概率模型”。这个模型假设用户头脑中有一个能够满足他的信息需求的理想文档，用户从这个理想文档中抽取词汇作为查询条件，用户所选择的查询条件词汇能够将这个理想文档同文档集合中的其他文档区分开来。这样查询条件可以看作是由理想文档生成的能够表征该理想文档的文本序列。Ponte 和 Croft 给出的研究思路是：首先估计每篇文档的词汇概率分布，然后计算从这个分布抽样得到的查询条件的概率，并按照查询条件的生成概率来对文档进行排序。在这个模型中，一些统计信息比如词频信息(Term Frequency, TF)和文档频率(Document Frequency, DF)等信息成为语言模型检索方法中的有机组成部分。这一点与传统检索模型不同，在传统检索模型中，这些信息都是作为启发规则性质的计算因子引入的。另外，文档长度(Document Length, DL)归一化因素成为不必单独计算的因子，因为它已经隐含在语言模型中的概率参数中了。

检索模型面临的主要挑战包括：语言的模糊性和相关性概念。语言的模糊性是指语言本身存在这样一个现实：描述一个概念不是唯一的或者意义是模糊的或者用户理解是模糊的。同样一个词语可能描述两个概念。同一个概念可能由两个互不相关的词语集来描述。相关性的概念面临同样的尴尬问题：对于一个查询和文档集，不同的专家来判定文档集中的文档对于这个查询的相关性的大小会得出不完全一致的结果集，甚至对相关性概念本身的认识也存在分歧。更糟糕的是，同一个专家当被问到多次时可能给出不同的答案，这使得判断两个相关结果集的优劣变得非常困难。文本检索会议(TREC)正是在这种背景下设立的，它试图创造一个环境，在这个环境中，由几个专家手工评估集中起来的结果集(每个参加者评估每个查询的前 100 个文档)，由此获得一个对于给定的查询和文档集的估计的适合的答案集。合适的答案集是满足相关性的一个简单的文档集合。之所以说简单，是因为它并不是要成为一个全面的和无错误的结果集，它只是为考察不同排序算法的性能优劣提供一个独立的度量。很多技术可以用于信息检索问题，这些技术独立于模型之外的应用。这些技术通过运用更精确的文档和查询概念的表示来改善检索效率和效果。这些技术包括：相关

反馈、利用词典和语义网络的查询扩展、停用词表、词干提取等。

1.1.2 经典检索模型

从数学角度描述信息检索模型，可以定义一个四元组{D，Q，F，R(qi，dj)}[14]，其中

· D 是文档集中的一组文档逻辑表示；

· Q 是一组用户信息需求的逻辑表示，也称为查询；

· F 是一种构建文档表示、查询以及它们之间关系的模型；

· R(qi，dj)是排序函数，它输出一个与查询 qi∈ Q 和文档表示 dj ∈ D 有关的实数，这样就根据查询 q 在文档之间定义了一个顺序。

因此，信息检索的任务就是定义 D，Q，F，R，以达到符合用户需求的检索要求。围绕 D，Q，F，R 的不同定义，可以建立不同的信息检索模型，其中最常用的有四大模型，即布尔模型、向量空间模型、概率模型和语言模型。下面将就它们的具体定义作详细介绍。不论采用哪种模型，信息检索的任务都决定了它必须对信息的存储、表示、检索方式进行大量的研究与实践。目前，信息检索的研究已经取得了一定的成果，人们提出了多种信息检索的检索方法和算法，并实现了一些实用化的系统。

但我们应该注意到：目前的信息检索的效果并不乐观。当用户输入查询后，一般的信息检索系统会在一两分钟内相对高效率地为用户返回一批所需信息的候选，而在这些候选信息中，用户真正需要的信息往往较少，用户不得不花费相当数量的时间在众多候选中进行人工筛选。用户提交的查询不能准确表示用户需求，检索库中信息的存储、表示方法都是造成这种检索精度低下的原因，因此，信息检索在相关技术领域仍然需要进一步的研究工作来满足人们日益增长的检索需求。

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四类经典模型：布尔模型、向量空间模型[12,33]、概率模型[28,34]、语言模型方法[18]。随着信息检索系统的深入研究与发展，又从这四类经典模型中派生了许多扩展模型。以下分别从模型概念、相关文档判定方法即相似性计算两方面对经典模型进行介绍，并对它们各自主要的派生模型做简单概述。本文不打算对这些模型进行详细和深入的讨论，对此感兴趣的读者可从相关文献中找到一些资料。[12,13,16,35]

1. **布尔模型**。布尔模型是基于集合理论和布尔代数的一种简单的检索模型，在其检索过程中要进行标引词的严格匹配。布尔模型是出现最早的 IR 模型，也是应用最广泛的 IR 模型，在 20 世纪六七十年代得到了较大的发展，出现了许多基于布尔模型的商用检索系统，如 DIALOG，STAIRS，MEDLARS 等。

在布尔模型中，文档的逻辑表示被定义为：$d_i = (w_{i1}, w_{i2}, \cdots, w_{it})$，$w_{i,j} \in \{0, 1\}$。而查询 q 是一个常规的布尔表达式，由标引词和逻辑运算符“AND”“OR”以及“NOT”组成。文本与查询的匹配规则遵循布尔运算的法则。也就是说查询作为布尔表达式，其运算结果的文档集合作为检索结果。

布尔模型主要优点是：速度快，易于表达一定程度的结构化信息。如同义关系(电脑 OR 微机 OR 计算机)或词组(文本 AND 过滤 AND 系统)，符合经过数据库系统使用培训的人们的习惯，因此许多搜索引擎仍利用这种方式。然而经典布尔模型有以下限制：

(1) 对于布尔模型，查询条件与文档间的相似性是二元的：1 为相关，0 为不相关，没有部分相关的情形；

(2) 布尔模型没有提供评分函数，不能对所检索到的文档进行排序。一般来说，用户总是认为排在相关文档集中的前列文档与用户的查询条件更加相关；

(3) 基于布尔模型的检索系统可能返回过多的或者过少的结果文档。例如对于查询条件“A and B and C and D”，系统可能会产生过少的结果，而实际上用户可能希望得到一些相关的文档，并不一定要求文档都必须包含检索词“A，B，C，D”；

(4) 布尔模型没有提供对查询检索词赋权重值的机制。实际上用户可能知道他所提交的查询条件中，哪些检索词是比较重要的，哪些检索词并不重要。

2. **向量空间模型**。早在 20 世纪 50 年代，Luhn 就提出了把文本表示成带权重信息的词项向量的思想[36]，这种思想正是向量空间模型(Vector Space Model，VSM)的精髓所在。

由 Salton 等提出的向量空间模型[33,37]是近十几年来信息检索领域应用最为广泛的检索模型，著名的 SMART 系统[38]即是基于向量空间模型而构建的原型检索系统。其基本思想是假设词与词之间不相关，以向量来表示文本，从而简化了文本中的关键词之间的复杂关系，文档用十分简单的非二进制的权重来实现。权重应该能够体现关键词的重要程度，是对整个文档内容的描述能力和区别其他文档的区别能力的量化。特征项的权重计算是人为赋予的，因此随意性较强，但多数情况下，可利用统计方法获得。通常使用词频来表示。词频又分为绝对词频和相对词频：绝对词频，即使用词在文本中出现的频率表示文本；相对词频为归一化的词频，其计算方法主要运用 TF * IDF 公式，目前存在多种 TF * IDF 公式，以下是一种比较普遍的 TF * IDF 公式：

$$W(t, d) = tf(t, d) * idf(t)$$

其中，$W(t, d)$为特征词 t 在文档 d 中的权重，而 $tf(t, d)$为特征词 t 在文

档 d 中的词频，idf(t)是特征词 t 在整个文本集 D 中的反比文档频数，idf(t) = $\frac{N}{n_t}$，N 是整个文档集的数目，n_t 是包含特征词 t 的文本个数，有时，采用 lg(N/n_t)来表示。应该考虑到文本的长度，否则文本长度越长，被检索的概率越大。因此，将上面的权重公式做归一化处理，得到：

$$W(t,d) = \frac{tf(t,d) * \log(N/n_t + 0.01)}{\sqrt{\sum_{t \in V} tf(t,d) * \log(N/n_t + 0.01)}}$$

文本与查询的匹配程度，由表示文本和查询的向量的夹角余弦来表示，其计算公式为：

$$sim(D_i,Q) = \frac{\sum_{k=1}^{n} w_{ik} * q_k}{\sqrt{\sum_{j=1}^{n} w_{ij}^2 \sum_{j=1}^{n} q_j^2}}$$

3. **概率模型**。概率检索模型最早由 Maron 和 Kuhns 于 1960 年提出[39]，经过 Maron、Cooper、Robertson、Van Rijsbergen、Croft 和 Turtle 等人的发展，概率检索模型已从理论走向实际应用。基于概率检索模型的 OKAPI 检索系统[30]在多次 TREC 评测中取得了优异的成绩，另一个概率检索系统 INQUERY[40]也有着不错的声誉。本节仅对概率检索模型的基本原理作一简要介绍，更为详细的综述可参阅相关文献[35]。

在概率检索模型中，我们感兴趣的问题是：给定一个查询，当我们观察到一篇文档的时候，它和查询相关的概率是多少？概率检索模型中两个著名的检索方法是 Robertson 和 Sparck Jones 提出的 BIR 检索模型[23]以及 Croft 和 Harper 提出的检索模型[41]。

文本信息相关性判断的不确定性和查询信息表示的模糊性，促使人们使用概率的方法解决这方面的问题。信息检索的概率模型是基于概率排序原则：对于给定用户查询 Q，对所有文本计算概率，并从大到小进行排序，概率公式为 P($R \mid D$, Q)。其中，R 表示文本 D 与用户查询 Q 相关，另外，用 $\bar{R}$ 表示文本 D 与用户查询 Q 不相关，有 P($R \mid D$, Q) + P($\bar{R} \mid D$, Q) = 1，也就是用二值形式判断相关性。把文本用特征向量表示：$x = (x_1, x_2, \cdots, x_n)$。其中，$n$ 为特征项的个数，x_i 为 0 或 1，分别表示特征项 i 在文本中出现或不出现。在信息检索中，估计参数是困难的，一般并不直接计算 P，而是把计算 P($R \mid d_i$, q_k)换为计算 P($R \mid x$, q_k)。这样处理略去了公式中与文本无关的特征项，计算的结果可能与实际不符。为了容易计算，假设包含相同特征项的文本，经过计算后，它们的可能性是相同的。将所有文本按相关概率 P 进行排序，等价于将所有文

本按特征向量排序。任一文本 D 的概率相关性的计算为：

$$P(R \mid D,Q) = \sum_i x_i * \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

其中，$p_i = P(x_i = 1 \mid R, Q)$，$q_i = P(x_i = 1 \mid \bar{R}, Q)$。参数 p_i 和 q_i 主要通过相关反馈进行估计，简单的方法如：

$$p_i = r_i / r;\ q_i = (n_i - r_i)/(n - r)$$

其中，n 为反馈文本集所含文本总数，r 为与用户查询相关的文本个数，n_i 为特征 i 出现的文本个数，r_i 为特征 i 出现且与用户查询相关的文本个数。

在该模型中，文本向量只采用简单的二值形式，没有利用文本的更多信息，比如特征在文本中出现的频率。

在该模型的基础上，又扩展出许多模型。Fuhr 提出了概率索引模型，没有更多的参数估计问题，对文本的表示也更加详细。Croft 模型体现了面向描述的索引思想，其公式为

$$V(R \mid D,Q) = \sum_i u_i * \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

其中，$u_i = P(x_i = 1 \mid d)$，u_i 的获取主要是使用概率索引模型，如传统的 2 - Poisson 模型。

上面讨论的模型都属于概率相关模型，这种模型对所处理的文本集依赖性过强，而且处理问题过于简单。鉴于概率相关模型存在这些弱点，人们又提出了改进模型。

Van Rijsbergen[28] 把信息检索看成是一个非确定性的推理过程，把查询和文本的内容表示为逻辑形式，并利用推理规则进行演绎。该模型把文本与用户查询之间的相关性判断看成是一个从文本命题到查询命题、描述命题的不确定的推断过程。

还有一种概率模型使用推理网络。网络中的一个节点代表一个文本、一个查询或一个概念，网络中节点间的弧表示节点间的概率相关性。其基本思想是：在计算 P（D → Q）时，把文本节点设置为 TRUE，计算与该文本节点相邻的节点的概率，直至得到 P（Q = TRUE）的值为止。

概率模型的优点在于无需经验性的权值计算公式，完全从理论上推断出排序，可以很好地支持用户反馈，并且具有渐进的优化检索效果。

它的缺点主要有：

（1）需要假定初始的相关和不相关文档集合；

（2）没有考虑文档内部索引检索词的频率信息，检索词的权重值是二元的；

（3）假定索引检索词是互相独立的；

（4）先验分布很难得到。

4. **语言模型**。Ponte 和 Croft 最初提出的语言模型检索方法现在经常被称为“查询条件概率模型”[18]。这个模型假设用户头脑中有一个能够满足他的信息需求的理想文档，用户从这个理想文档中抽取词汇作为查询条件，用户所选择的查询条件词汇能够将这个理想文档同文档集合中其他文档区分开来。这样查询条件可以看作是由理想文档生成的能够表征该理想文档的文本序列。由这个假设我们可以看出信息检索系统的任务被转化为判断文档集合中每个文档与理想文档哪个最接近的问题。也就是说，我们需要计算：

$$\arg\max_{D} \mathrm{P}(D \mid Q) = \arg\max_{D} \mathrm{P}(Q \mid D) P(D)$$

其中，Q 代表查询条件，D 代表文档集合中某个文档。先验概率 $\mathrm{P}(D)$ 对于文档集合中每篇文档来说都是相同的。所以关键是估计每篇文档的语言模型 $\mathrm{P}(Q \mid D)$。换句话说，我们首先需要估计每篇文档的词汇概率分布，然后计算从这个分布抽样得到查询条件的概率，并按照查询条件的生成概率来对文档进行排序。

这个经典的基于语言模型的信息检索模型，为信息检索领域开辟了一个很有前景同时也具有相当挑战性的方向。与传统检索模型相比，语言模型检索方法有下列优点[42-44]：

（1）能够利用统计语言模型来估计与检索有关的参数，是语言模型信息检索系统的一个优点。

（2）使用语言模型的另外一个好处是，我们可以通过对语言模型更准确的参数估计或者使用更加合理的语言模型来获得更好的检索性能。这样，与传统的模型相比较，在如何改善检索系统性能方面有更加明确的指导方向。

（3）另外，语言模型方法对于文档中的子主题结构和文档间的冗余度建立统计模型也是有帮助的。

尽管实验表明该方法检索性能优于一些传统的检索模型，但是其本身还是存在一定的缺点：

（1）该方法隐含着词汇相互独立关系，没有考虑词汇间的相互影响。

（2）传统检索模型中常用的查询反馈技术在概念层面融入语言模型框架比较困难。

1.1.3 信息检索中的相关技术

1. **索引技术**。索引技术的目的是理解文档信息，从中抽取索引项，用于表

示文档以及生成文档库的索引表。索引项有客观索引项和内容索引项两种。客观项与文档的语义内容无关，如作者名、更新时间、编码、长度等；内容索引项是用来反映文档内容的，如关键词及其权重、短语、单字等。内容索引项可以分为单索引项和多索引项(或称短语索引项)两种。单索引项对于英文来讲是英语单词，比较容易提取，因为单词之间有天然的分隔符(空格)；对于中文等连续书写的语言，必须进行词语切分。索引算法对索引技术的性能有很大的影响。

信息检索中索引的组织结构有两种，即正排表和倒排表[39]。正排表是以文档的 ID 为关键字，表中记录项记录文档中每个字或词的位置信息，查找时扫描表中每个文档中字或词的信息直到找出所有包含查询关键字的文档。正排表结构如表 1 - 1 所示。这种组织方法在建立索引的时候结构比较简单，建立比较方便且易于维护，但是在查询的时候需对所有的文档进行扫描以确保没有遗漏，这样就使得检索时间大大延长，检索效率低下。

表 1 - 1　正排表

文档 1	关键词 1	关键词 2	……
文档 2	关键词 1	关键词 2	……
文档 3	关键词 1	关键词 2	……
……	……	……	……
文档 n	关键词 1	关键词 2	……

表 1 - 2　倒排表

关键词 1	文档 1	文档 2	……
关键词 2	文档 1	文档 2	……
关键词 3	文档 1	文档 2	……
……	……	……	……
关键词 n	文档 1	文档 2	……

倒排表结构如表 1 - 2 所示。倒排表以字或词为关键字进行索引，表中关键字对应的记录表项记录了出现这个字或词的所有文档，一个表项就是一个字表段，记录该文档的 ID 和字或词在该文档中出现的位置情况。由于每个字或词对应的文档数量在动态变化，所以倒排表的建立和维护都较为复杂，但是在查询