信息检索的概率模型*)

邢永康　马少平
(清华大学计算机系智能技术与系统国家重点实验室　北京100084)

Survey on Probability Models of Information Retrieval

XING Yong-Kang　MA Shao-Ping
(The State Key Lab. of Intelligent Tech. and System, Tsinghua University, Beijing 100084, China)

Abstract The study of mathematical models on information retrieval is an important area in the Information Retrieval community. Because of the uncertainty characteristic of IR, the probability model based on statistical probability is a promising model from recent to future. Those models can be classified into classical models and probability network models. Several famous models are introduced and their shortcomings are pointed out in this paper. We also clarify the relationship of these models and introduce a new models based on statistical language model curtly.

Keywords Information retrieval, Probability models, Classical models

1　引言

信息检索是以文档为主要的处理对象，对其结构化和非结构化数据包括多媒体信息进行储存、索引、查询和管理的方法和技术。给定一批文档 $D=\{d|d_1,d_2,\cdots,d_j,\cdots,d_m\}$ 及用户的检索要求 q(简称检索)。我们将构成文档的基本项，如构成文本文档的单词、短语等，构成图像文档的纹理特征项、图像内容标注的单词等，统称为索引项，用集合表示为 $T=\{t|t\in T\}=\{t_1,t_2,\cdots,t_n\}$。同理，将构成检索的基本项统称为检索项。一般情况下，检索项和索引项是同一个集合。信息检索的目标是快速而准确地找出满足检索的 q 文档(称这些文档与检索 q 相关)。在大多数情况下，还需要根据相关程度对这些文档排序。评价一个检索系统的主要量化指标是查全率和查准率。

信息检索的数学模型，简称信息检索模型，是对信息检索任务及实现方法的一种抽象描述。作为信息检索研究的一个主要内容，信息检索模型是对信息检索任务的数学抽象，它避开了对具体实现细节如数据存储、数据结构等的描述，而主要从以下两个方面抽象地研究信息检索方法：

1. 确定在模型中，如何表示构成检索系统的两个要素——文档和检索。

2. 确定在模型中，如何定义和计算文档和检索之间的关系。主要任务是定义并计算表示任意文档 $d\in D$ 与用户检索 $q\in Q$ 的相关程度的函数 $f_r(q,d)$。相关程度函数又称相关度排序函数，或信息检索函数，该函数值的大小表示文档满足检索要求的程度。

依据各个模型的数学基础，可以将现有的信息检索模型分为三类。一类是基于集合理论的模型。该类模型将每个文档表示为索引项集合，通过集合运算来判定文档与检索的相关度。典型的模型有：布尔模型、扩展的布尔模型及基于模糊集的模型。第二类是基于代数理论的模型。该模型将文档和检索请求表示为所有索引项张成的向量空间中的点，通过向量的空间关系来定义和计算相关度函数。典型的模型有：向量空间模型、扩展的向量空间模型及潜在语义空间模型。第三类是基于概率统计理论的模型，即信息检索的概率模型。与其它模型相比，概率模型具有如下的优点：

1. 信息检索过程中存在着大量的不确定性，如用户检索目的的不确定，检索目标的不确定，以及检索结果的不确定等。因此用概率理论来描述这些不确定性，就具有自然、直观的特点。概率模型是从理论上研究信息检索问题的基本模型。

2. 信息检索过程其实是检索系统对用户检索需求的逐步了解的过程。大量实践证明，相关反馈(Relevance Feedback)是提高检索结果的有力技术。而基于 Bayes 概率的信度修正理论，不但为该过程提供了理论支持，而且已经为此建立了大量的实用方法。

3. 信息检索与自然语言处理之间有着天然的联系。近年来以大规模真实语料库的处理为基本方法的语料库语言学(Corpus Linguistics)取得了较大的进展，由此建立起来的统计语言模型是一个主要成果。基于两者共同的概率本质，如何将这些研究成果用于信息检索，建立基于语言模型的检索模型就成了一个必然的想法，并且也有望获得较好的检索结果。

因此，本文将主要对信息检索的概率模型进行分析研究。根据这些模型提出的时间顺序，我们将其分为两类：

1. 经典概率模型：其中有代表性的模型有：二元独立模型(Binary Independency Model)[1]；二元一阶相关模型(Binary First-order Dependency Model)[2]；双 Poisson 分布模型(Two-Poisson Model)[3,4]等。

2. 概率网模型：其中有代表性的模型有：推理网络模型(Inference network)[9]及信度网模型(Belief network)[10]。

2　信息检索的经典概率模型

经典的概率模型基于如下思想：根据用户的检索 q，可以

*)本项目受到国家重点基础研究(973)(G1998030509)、自然科学基金项目(60223004)以及863高科技项目(No. 2001AA114082)资助。邢永康博士，主要研究领域为机器学习，模式识别，网络数据挖掘。马少平　博士，教授，博士导师，主要研究领域为模式识别，信息检索，网络数据挖掘。

将 D 中的所有文档分为两类，一类与检索需求 q 相关(集合 R)，另一类与检索需求不相关($\overline{R}$)。在同一类文档中，各个索引项具有相同或相近的分布；而属于不同类的文档中，索引项应具有不同的分布。因此，通过计算文档中所有索引项的分布，就可以判定该文档与检索的相关度。

对于检索 q，任意文档与其相关和不相关的概率分别表示为：$P(R|d)$和 $P(\overline{R}|d)$。根据贝叶斯公式：

$$P(P|d)=P(d|R)P(R)/P(d)$$
$$P(\overline{R}|d)=P(d|\overline{R})P(\overline{R})/P(d) \quad (1)$$

上式中的后两项只与检索需求 q 有关，而与每个文档 d 无关，可以不计算，则将计算 $P(R|d)$转化为计算 $P(d|R)$。同理，对 $P(\overline{R}|d)$的计算也将转化为对 $P(d|\overline{R})$的计算。

由于索引项的数目很大，因此常常在计算中引入一些假设，以简化计算。对应不同的假设，就形成了三种不同的经典概率模型，分别是：二元独立模型(binary independent model)、二元一阶依赖模型(binary first order dependent model)和双 Poisson 分布模型(two Poisson independent model)。

2.1 二元独立概率模型(Binary independent model)

二元独立模型对文档中索引项的分布做了如下两个假设：

假设1(二元属性取值假设) 任意一个文档 d 可以表示为 $d(x_1,x_2,\cdots,x_i,\cdots)$，其中二元随机变量 x_i 表示索引项 t_i 是否在该文档中出现，如果出现，则 $x_i=1$；否则，$x_i=0$。

假设2(索引项独立性假设) 在一个文档中，任意一个索引项的出现与否不会影响到其它索引项的出现，它们之间相互独立。

根据假设1和2有：

$$P(d|R)=P(x_1,x_2,\cdots|R)=\prod_{i=1}^{n}P(x_i|R) \quad (2)$$

$$P(d|\overline{R})=P(x_1,x_2,\cdots|\overline{R})=\prod_{i=1}^{n}P(x_i|\overline{R}) \quad (3)$$

至此，我们可以定义文档 d 与检索 q 的相关度排序函数 $f_r(q,d)$为(相关性排序函数还可有不同的定义方式，如文[1]就根据 Bayes 决策理论，给出了另外一种函数形式。但分析可以看出，两者在本质上都是一样的。所以本文只讨论一种形式)：

$$f_r(q,d)=\frac{P(R|d)}{P(\overline{R}|d)} \quad (4)$$

该值越大，表示文档 d 与检索 q 越相关。将式(2)(3)代入(4)式，去掉常数并整理后有：

$$f_r(q,d)=\frac{\prod_{i=1}^{n}p_i^{x_i}(1-p_i)^{1-x_i}}{\prod_{i=1}^{n}q_i^{x_i}(1-q_i)^{1-x_i}} \quad (5)$$

其中，$p_i=P(x_i=1|R)$，$q_i=P(x_i=1|\overline{R})$。为了简化计算，对(5)式右边取对数并整理，就得到相关度排序函数的计算公式为：

$$f_r(q,d)=\sum_{i=1}^{n}x_i*\log\frac{p_i(1-p_i)}{q_i(1-q_i)} \quad (6)$$

(6)式中需要确定的参数为 p_i,q_i，它们分别表示索引项 t_i 在两类文档 R 和 $\overline{R}$ 中的出现概率。如果能够预先得到一定数量的带有标记(相关性标记)的文档，则可以通过最大似然估计法来确定参数 p_i,q_i 的值。假设对给定的文档集的统计结果如表1所示，则：

$$p_i=P(x_i=1|R)=n_R(x_i=1)/[n_R(x_i=1)+n_R(x_i=0)] \quad (7)$$

$$q_i=P(x_i=1|\overline{R})=n_{\overline{R}}(x_i=1)/[n_{\overline{R}}(x_i=1)+n_{\overline{R}}(x_i=0)] \quad (8)$$

表1 二元独立模型的参数估计表

	相关	不相关
$x_i=1$	$n_R(x_i=1)$	$n_{\overline{R}}(x_i=1)$
$x_i=0$	$n_R(x_i=0)$	$n_{\overline{R}}(x_i=0)$

在实际应用中，一般无法预先给出带有相关性标记的文档集，所以常常通过相关反馈(Relevant Feedback)技术来获取标记文档：即先采用其它检索技术，如全文检索技术等获得一批文档，并由用户对这些文档进行相关性标记，然后再将这些标记后的文档作为确定参数的文档集。

2.2 二元一阶相关概率模型

索引项独立性假设只是为了数学上计算处理方便，并不符合实际情况。可以看到，一些索引项在文档中的出现往往并不是相互独立，而是存在某种关系，如某些索引项经常会同时出现在一篇文档中。因此要想获得更好的检索结果，就必须考虑各个索引项之间的相互依赖关系这一信息。这就是建立二元依赖模型的背景。与二元独立模型相比，后者在假设1上与前者完全一致(也就是对文档的表示两者一致)，唯一的区别在于后者不承认假设2，从而对 $P(d|R)$和 $P(d|\overline{R})$的计算与前者不同。这里我们主要研究 $P(d|R)$的计算，即在相关文档中各个索引项的分布。同理也可以计算出 $P(d|\overline{R})$。

为了实际地表示文档中各个索引词的相互关系，我们可以假设在相关文档中，各个索引项之间存在统计相关性。统计相关性不同于逻辑相关性，它是两个或多个索引项在文档中出现频率之间所表现出的一种相关性，而并不考虑各个索引项在文档中出现的先后次序。根据统计相关性有：

$$P(d|R)=P(x_1,x_2,\cdots|R)=P(x_1|R)P(x_2|x_1,R)\cdots P(P|x_n|x_1,x_2,\cdots,x_{n-1},R)) \quad (9)$$

上式尽管可以准确地表示各个索引项之间的相关性，然而它包含的参数数目非常大。一种简化计算的假设是：假设对于每一个索引项 t_i，有且只有一个索引项 $t_{j(i)}$，使得索引项 t_i 与其余索引项之间条件独立，即：$P(x_i|x_1,x_2,\cdots,x_{j(i)},\cdots,R)=P(x_i|x_{j(i)})$。该假设称为一阶相关性假设。根据该假设，我们必须找到分布 $P(d|R)$的一个近似分布 $P'(d|R)$，它满足如下分解性质：

$$P'(d|R)=P(x_1,x_2,\cdots,x_n|R)=P(x_1|x_{j(1)},R)R(x_2|x_{j(2)},R)\cdots P(x_n|x_{j(n)},R)$$

并且该近似分布 $P'(d|R)$与分布 $P(d|R)$越接近越好，这可以通过它们之间的交叉熵来表示。Chow 等给出了一个算法(Maximum Spanning Tree)，并证明了该算法可以求得符合上述要求的近似分布 $P'(d|R)$[6]。

如果设 $s_i=P'(x_i=1|x_{j(i)}=1,R)$；$r_i=P'(x_i=1|x_{j(i)}=0,R)$则可以推导出：

$$P'(x_i|x_{j(i)},R)=[s_i^{x_i}(1-s_i)^{1-x_i}]^{x_{j(i)}}[r_i^{x_i}(1-r_i)^{1-x_i}]^{x^{j(i)}} \quad (10)$$

至此，我们可以将 $P(d|R)$的计算表示如下：

$$P(d|R)\propto\log P(d|R)\approx\log P'(d|R)=\sum_{i=1}^{n}[x_i\log r_i+(1-x_i)\log(1-r_i)]+\sum_{i=1}^{n}[x_{j(i)}\log\frac{1-s_i}{1-r_i}+x_i x_{j(i)}$$

$$\log\frac{s_i(1-r_i)}{r_i(1-s_i)}]+C \tag{11}$$

只要假设 $t_i=s_i$，则(11)式就可以转化为(2)式，这表明二元独立模型是一种特殊的二元一阶依赖模型。该模型中的参数包括 s_i 和 r_i，对它们的估计要比二元独立模型复杂一些。假设给定一批带有相关性标记的文档，其统计情况如表2，则利用最大似然法估计参数如下：

$$s_i=n(x_i=1,x_{j(i)}=1)/[n(x_i)=1,x_{j(i)}=1)+n(x_i=0,x_{j(i)}=1)] \tag{12}$$

$$r_i=n(x_i=1,x_{j(i)}=0)/[n(x_i)=1,x_{j(i)}=0)+n(x_i=0,x_{j(i)}=0)] \tag{13}$$

表2 二元独立模型的参数计算

	$x_{j(i)}=1$	$x_{j(i)}=0$
$x_i=1$	$n(x_i=1,x_{j(i)}=1)$	$n(x_i=1,x_{j(i)}=0)$
$x_i=0$	$n(x_i=0,x_{j(i)}=1)$	$n(x_i=0,x_{j(i)}=0)$

2.3 双泊松分布概率模型

双泊松分布模型最先是由 Harter 在研究文档索引时提出的。该模型的基本思想来源于如下的实验观察：文档中的单词可分为两类：一类单词与表达文档的主题相关，称为内容词(conent-bearing words)；另一类只完成一些语法功能，称为功能词(functional words)。统计实验发现[5]：功能词在文档中的分布与内容词不同，前者出现的频率比较稳定，其波动情况可以近为泊松分布，即如果用 x 表示某个功能词在文档中的出现频率，则：

$$P(x)=\frac{u^x}{x!}e^{-x}$$

其中，u 为该分布的均值，表示该功能词的平均出现频率。

可见内容词在文档中的出现频率，在一定意义上反映了一个文档的主题。因此 Harter 假设：根据一个内容词可以将文档从主题上分为两类，同时该内容词在两类文档中的出现频率也会很不相同：一类文档的主题与该内容词相关，那么该内容词在其中的出现频率应该比较高，其波动特征可以用一个 Poisson 分布表示；而另一类文档的主题与内容词不相关，所以内容词在其中的出现频率应该比较低，其波动特征也可以用一个 Poisson 分布表示；综合起来，一个内容词在文档中的出现频率 x 可以表示为两个 Poisson 分布的加权组合：

$$P(x)=\pi\frac{u^x}{x!}e^{-x}+(1-\pi)\frac{v^x}{x!}e^{-x} \tag{14}$$

其中，u,v 分别为内容词在两类文档中出现频率的均值。π 表示了任意一个文档属于第一类的概率，该假设被称为双 Poisson 分布假设。只要将所有的索引项看作是内容词(其实，在实际的检索中，索引项一般都是内容词)，则它们也满足2-Poisson 模型，则就形成了双 Poisson 分布模型。与二元独立模型相比，2-Poisson 模型的不同在于不承认假设1，其余都相同，所以可以与二元独立模型一样来定义相关性排序函数(参见公式(4))。

$$f_r(q,d)\propto\frac{P(d|R)}{P(d|\overline{R})}==\frac{P(x_1,x_2,\cdots,x_i,\cdots|R)}{P(x_1,x_2,\cdots,x_i,\cdots|\overline{R})} \tag{15}$$

根据2-Poisson 分布假设得：

$$=\frac{\prod_{i=1}^{n}u_i^{x_i}e^{-x_i}/x_i!}{\prod_{i=1}^{n}v_i^{x_i}e^{-x_i}/x_i!} \tag{16}$$

对其取对数，并去掉不变量有：

$$f_r(q,d)=\sum_{i=1}^{n}(v_i-u_i)+\sum_{i=1}^{n}x_i\log(u_i-v_i)\propto\sum_{x_i\in q}x_i\log(u_i/v_i) \tag{17}$$

对于2-Poisson 分布独立模型，一般采用矩法(method of moments)计算[4]估计该模型的所有参数。可以证明，对于一个索引项，参数 u,v 的值是以下二次方程的两个根(其中值较大的根对应参数 u)：

$$ax^2+bx+c=0$$

其中：$a=M_1-L;b=K-LM;c=L^2-MK;M_1,M_2$和 M_3分别为三个样本原点矩(moment about origin)，且 $L=M_2-M_1;K=M_3+2M_1-3M_2;$。

解出方程后，参数 π 的值由以下公式计算：

$$\pi=\frac{M_1-v}{u-v} \tag{18}$$

为了更精确地描述索引项的频率分布，还可以采用 n-Poisson 分布假设[7]。实验证明，n 越大，对索引项的频率分布描述也越精确，但参数的数目也会急剧增加，将会增加需要的学习文档数目以及参数估计的复杂性[8]。

3 概率网信息检索模型

概率网信息检索模型的产生背景是：

1. 从60年代开始信息检索研究以来，人们已经建立了许多检索模型，这些模型具有各自的特性和优点，在不同的具体应用中很难相互代替，因此，寻找一种能综合多个模型优点的模型，就成为一种必然的想法。

2. 人们越来越认识到，改变以往基于关键词匹配的检索方法，发展基于概念的信息检索才是信息检索可望取得突破性进展的方向。

基于这两种考虑，1990年 Turtle 在他的博士论文中提出并建立了推理网信息检索模型。在此基础上，1996年 Berthier 等人建立了具有更坚实的理论依据和更强的表达能力的信度网检索模型。这两个模型的主要思想都是基于对概率的主观性理解，将求解信息检索问题的一般过程转化为一个基于给定证据的推理过程。具体方法是为检索过程中的各个要素，包括文档、用户检索、索引项等建立随机变量，并利用概率网模型表示这些随机变量之间的概率依赖关系，从而建立起一种解决信息检索问题的通用框架。这类模型具有以下优点：1. 可以综合多个经典模型的特点，从而获得更好的检索结果。2. 利用概率网的优点，综合利用各种信息(证据)来进行信息检索。如历史检索记录就可以作为证据，加入网络中，从而提高信息检索的效果。3. 通过网络关系的设置，与传统的经典模型相比，部分地具有从关键词匹配到概念推理的信息检索能力。

3.1 推理网模型(Inference Network Model)

一个推理网检索模型分为两部分(见图1)，图中虚线上面的部分称为文档网络(Document network)，下面部分称为检索网络(Query network)。网络中所有结点变量都是二值变量(在概率网中，每个节点必然对应一个随机变量，所以称为节点还是随机变量，都是一个意思，但侧重点不同)，其值域为{0,1}。

文档网络由文档结点 $d_1,d_2,\cdots$和文档表示结点 $x_1,x_2,\cdots,x_n$ 构成。每一个文档结点 d 对应文档集合 D 中的一个实际文档，它的取值分别表示该文档是否被观察到。每个表示结点 x 对应文档的索引项，它的取值分别表示某个文档是否包含了该索引项。因为每一个文档结点与它包含的索引项之间

存在因果关系，所以用一个有向边来连接，因果关系的强度用结点的条件概率表来表示。我们知道，不同的信息检索模型的一个主要区别是对文档的不同表示方法。可以想象，对于不同的文档表示方法，都可以定义不同的表示结点及条件概率表，所以文档的网络模型可以同时将多种文档的表示方法综合在一起。

检索网络是对用户信息检索需求的结构化表示。与文档的表示不同，它是以用户的信息检索需求是否被满足作为构成网络的因果关系，因此与文档网络的方向相反。在该网络中，结点 I 表示用户的信息检索需求，对应事件——该检索需求是否被满足。由于该需求是一种用户的内在需求，一般无法明确理解，因此采用不同的表示方案，可以形成多个格式化的检索，如图中的 q，q_1 等结点。该结点的取值分别表示该检索是否被满足。与文档一样，每一个检索又可以采用多种不同的表示方案，每一种表示方案通过一批不同的表示结点来表示。这些表示结点的取值分别表示它们是否被观察到。

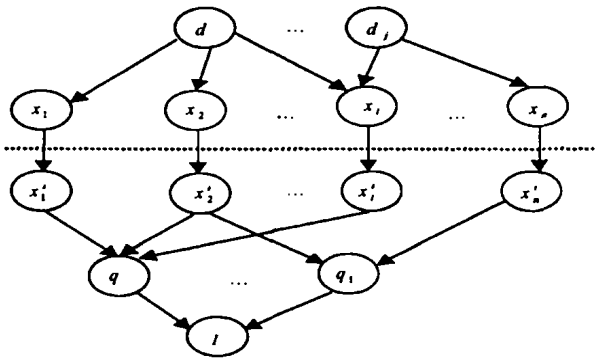

图1　推理网检索模型

将文档网络与检索网络结合在一起的是两者的表示结点各自构成的空间。这两个空间之间可以根据实际的应用背景，建立起多种映射关系。最简单和常用的映射是一一映射（如图1所示）。如在文本检索中，如果每个表示结点对应一个单词，就会产生这种映射关系。利用推理网模型，就将信息检索过程表示为一个基于证据的推理过程：一次指定一个文档变量的值为1，即将它作为证据，计算出检索结点的后验概率；对于所有的文档分别作如上的计算，就可以根据后验概率值对这些文档与检索的相关度进行排序。

定义1　推理网模型中，一个文档 d 与检索 q 的相关度定义为给定证据 $d=1$ 条件下，检索 $q=1$ 的后验概率，即相关度排序函数为：

$$f_r(q,d)=P(q=1|d=1) \tag{19}$$

下面我们针对简化的推理网结构（将结点 x_i 和 x'_i 合并为结点 x_i，并将所有表示结点构成的向量表示为 $\vec{x}=(x_1,x_2,\cdots,x_n)$），来推导相关性排序函数式(19)的计算公式：

$$P(q|d)\propto P(q,d)=\sum_{\forall \vec{x}} P(q,d,\vec{x})$$

$$=\sum_{\forall \vec{x}} P(q|d,\vec{x})P(d,\vec{x}) \tag{20}$$

根据该网络结构中包含的两类条件独立关系可以有效地简化计算：

1. 由于结点 q 与结点 d 被结点 $\vec{x}=(x_1,x_2,\cdots,x_n)$ 分割开，因此它们之间条件独立：

$$P(q|d,\vec{x})=P(q|\vec{x}) \tag{21}$$

2. 由于结点 $x_1,x_2,\cdots,x_n$ 等都是结点 d 的子结点，因此当变量 d 的值确定时，各个子结点之间相互独立：

$$P(\vec{x}|d)=P(x_1,x_2,\cdots,x_n|d)=\sum_{i=1}^{n} P(x_i|d) \tag{22}$$

将式(21)(22)代入式(20)并整理：

$$P(\vec{x}|d)=\sum_{\forall \vec{x}} P(q|\vec{x})[\prod_{i=1}^{n} P(x_i|d)]P(d) \tag{23}$$

上式中的参数有 $P(d)$、$P(x_i|d)$ 及 $P(q|\vec{x})$，在推理网中称为结点的条件概率表。只要为这些参数指定不同的取值方式，就可以模拟多个已有的经典信息检索模型，如 Bool 模型、向量空间模型等。

推理网模型具有以下的优点：

·该模型用文档网络和检索网络分别表示文档和检索，它将多种文档表示方案及检索表示方案综合在一个统一的网络中。因此从理论上讲，它可以同时综合多种信息检索模型的优点。

·推理网模型对用户的信息需求做了结构化的表示，从而可以综合各种检索生成与扩展技术。如图2所示，节点 I 表示用户的信息需求，它可以看成多个检索的逻辑组合，如果假设 $I=q\vee q_1$，则相关度排序函数可定义为：$f_r(I,d_j)=P(I,d_j)$

3.2　信度网模型（Belief Network Model）

信度网检索模型是由 Berthier 在推理网模型的基础上提出的。与前者相比，它明确定义了模型中概率的样本空间，因此具有更坚实的理论基础，并且比前者有更强的表达能力。

定义2（样本空间）　文档集合 D 中所有文档的所有索引项构成的集合 $S=\{t_1,t_2,\cdots,t_n\}$，称为模型的样本空间（sample space）。

定义3（概念）　定义在样本空间 S 上的一个概念（concept）c 是集合 S 的一个子集，即：

$$c=\{t_1,t_2,\cdots,t_m\}\subset S$$

可以看出，概念就是一个集合。为了方便处理和表示概念之间的关系（即集合关系），可以采用随机变量来表示这些概念（集合）：对 S 中的每一个索引项 t_i 分别设置值域为{0,1}的随机变量 x_i，且用 $x_i=1$（$x_i=0$）表示该索引项包含（不包含）在相应的概念中，则一个概念 c 就可以表示为集合 $c=\{x_1,x_2,\cdots,x_m\}$ 且 $\forall x_i=1$。

根据定义3，我们可以将每一个文档 d 看作是样本空间 S 上的概念，即：$d=\{x_1,x_2,\cdots\}$。同样，检索 q 也可看作是样本空间 S 上的概念。

定义4（概率分布 P）　对于样本空间 S 上的任意一个概念 c，它的概率分布 $P(c)$ 定义为概念 c 对样本空间 S 的覆盖度，并通过下式计算：

$$P(c)=\sum_{u\in U} P(c|u)P(u) \tag{24}$$

U 表示样本空间 S 上的所有概念构成的集合。

基于以上的定义，我们将信息检索问题转化为在样本空间 S 上的概念匹配问题，即：

定义5　文档 d 与检索 q 的相关度定义为在样本空间 S 上，概念 d 对概念 q 的覆盖程度：

$$f_r(q,d)=P(d=1|q=1) \tag{25}$$

下面我们推导相关度函数式(25)的计算公式。

$$P(d|q)=P(q,d)/P(q)=\alpha P(q,d)$$

α 是一个常数，所以只需计算 $P(q,d)$。根据定义4引入基本概念 u 有：

$$P(q,d)=\sum_{u}P(q,d|u)P(u) \tag{26}$$

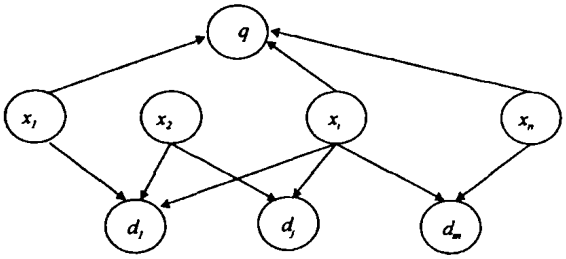

图2 信度网检索模型

进一步计算需要利用这些变量之间的概率依赖关系。我们可以将这些关系表示为一个信度网结构，如图2所示。该结构包含三类结点，每个结点变量都是值域为{0,1}的二值变量：

1. 索引项结点 $x_1,x_2,\cdots,x_m$。由索引项对应的随机变量构成。变量值为1表示该索引项包含在当前的概念中。

2. 文档结点。一个实际的文档对应有一个文档结点，结点变量 d 的值为1表示概念 d 完全覆盖样本空间。

3. 检索结点：由于将文档和检索都看作是统一样本空间上的概念，因此索引结点与文档结点的处理完全一样。

根据信度网检索模型，对概念 u 表示为向量 $\vec{x}$，则(26)式转化为：

$$P(q,d)=\sum_{\vec{x}}P(d,q|\vec{x})P(\vec{x}) \tag{27}$$

由图2可知，索引结点给定后，文档结点和检索结点之间相互独立，所以：

$$P(q,d)=\sum_{\vec{x}}P(d|\vec{x})P(q|\vec{x})P(\vec{x}) \tag{28}$$

与推理网模型一样，该公式中的参数也是网络中各结点的条件概率表，通过为它们指定不同的函数，就可以模拟各种经典的信息检索模型。

与推理网模型相比，信度网检索模型具有两个明显的优点：

·具有充分的理论基础。尽管与推理网模型一样，它也是基于对概率的主观理解，但它明确定义了样本空间，因此信度网模型中的各个概率的定义及语义比推理网中的定义较易把握。

·模型的表达能力强于前者。比较式(23)和(28)可以看出，两个模型的主要区别是各自包含的函数 $P(\vec{x}|d)$ 和 $P(d|\vec{x})$，这是两个模型网络结构之间差别的反映。根据推理网中的条件独立性关系，函数 $P(\vec{x}|d)$ 不但应该是一个概率函数，而且还必须是一个具有可分解性质的特殊函数式(29)。而信度网中的函数 $P(d|\vec{x})$ 就可以表示任何一个满足概率函数条件的函数，因此，它可以表示函数 $P(\vec{x}|d)$，而后者却无法表示它。所以可以认为推理网模型是信度网模型的子集。

$$P(\vec{x}|d)=\prod_{i=1}^{n}P(x_i|d) \tag{29}$$

总结与展望 信息检索起源于 H. P. Luhn 在20世纪50年代对文献进行的统计学分析。直到20世纪70年代才开始在理论上进行了大量的研究。经典的概率模型则是在20世纪70年代后期和80年代初期建立起来的。这些模型具有简单、直观等特点，为信息检索的实践提供了有利的指导。同时在信息检索实践中，一些对经典模型的修正模型也被提了出来。1990年建立的推理网模型，为各种模型的综合提供了一个统一的框架。1996年建立的信度网模型则是对推理网模型从理论上的进一步修改和泛化。

概率模型存在的主要问题是检索中对带有相关性标注的学习文档的依赖性。因此，研究在没有预知的相关性文档、或相关性文档很少的情况下，如何修正模型的参数、提供较好的检索结果，是目前的一个主要研究问题。各种形式的相关反馈技术是解决该问题的基本方法之一。

近年来，以大规模真实语料库的处理为基本方法的语料库语言学(Corpus Linguistics)取得了较大的进展，由此建立的统计语言模型为信息检索模型的研究提供了新的思路。一类新的信息检索模型——基于统计语言模型的信息检索模型[11]得到了关注。这类模型假设每个文档都存在一个语言模型，以从文档的语言模型抽样产生检索的概率表示文档与检索的相关度。因此这类模型与传统概率模型的基本思想完全不同，借助于统计语言模型的研究成果，有望部分解决传统信息检索模型中存在的问题，使信息检索结果的质量有质的提高。

参考文献

1 Robertson S E, Jones S K. Relevance Weighting of Search Terms. JASIS, 1976, 27: 129～146

2 Van Rijsbergen C J. A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval. Journal of Documentation, 1977, 33: 106～119

3 Bookstein A, Swanson D R. Probabilistic models for automatic indexing. Journal of the American Society for Information Science, 1974, 25: 312～319

4 Hatter S P. A probabilistic approach to automatic keyword indexing. Information Science, 1975, 26: 197～205 (Part I), 280～289 (Part II)

5 Sparck J K, Jackson D M. The use of automatically - obtained keyword classifications for information retrieval. Information Storage and Retrieval, 1970, 5: 175～201

6 Chow C K, Liu C N. Approximating discrete probability distributions with dependence trees. IEEE Transactions on information theory, 1968, IT-14: 462～467

7 Margulis E L. Modeling documents with multiple Poisson distributions. Information Processing and Management, 1993, 29(2): 215～227

8 Titterington D M, Markov U E, Smith A F M. Statistical Analysis of Finite Mixture Distributions. John Wiley and Sons, 1985

9 Turtle H, Croft W B. Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems, 1991, 9(3): 187～222

10 Ribeiro-Neto B A, Muntz R. A belief network model for IR. In: Proc. of the 19th annual int. ACM SIGIR conf. on research and development in information retrieval, Zurich, Switzerland, 1996. 253～260

11 Ponte J, Croft W B. A language modeling approach to information retrieval. In: Proc. of the ACM SIGIR. 1998. 275～281