基于量子的交互式信息检索模型分析

徐连杰　胡德华

(中南大学湘雅医学院　湖南长沙　410013)

摘　要:阐述信息检索的概念,简要介绍比较常见的信息检索模型,详细介绍新的信息检索模型——基于量子的交互式信息检索模型。

关键词:检索模型　量子物理　交互式信息检索　搜索引擎

中图分类号:G354　　文献标识码:A　　doi:10.3969/j.issn.1005-8095.2012.01.008

随着互联网技术的发展以及网络信息的快速膨胀，人们在日常生活中对网络信息的获取也日益倚重。但是,互联网上在给人们带来前所未有的海量信息源的同时，也给人们在浩如烟海的网络信息中找到最合适、最准确的信息带来了巨大困难。此时,搜索引擎的出现，大大缓解了人们对网络信息快速,准确获取的急迫需求,并已逐渐成为人们日常生活中不可或缺的一部分[1]。搜索引擎技术——信息检索,不断研究和发展,已变得日益成熟。

1　信息检索的概念

信息检索（简称 IR）作为一个正式的学术概念,在 1945 年由美国学者 Mooers 在其硕士学位论文中首次提出[2]。经过半个多世纪的研究,信息检索的发展演变可以看作是不断消除一道道信息存取障碍的过程,形成了一种检索理论和方法,在实际应用方面,也出现了各具特点的信息检索系统。

信息检索是指将信息资源按照一定方式组织和存储,并能从信息资源集合 D 中,根据用户信息需求集合 Q,利用匹配处理框架 F 和匹配计算函数 R(qi,dj),返回满足用户所需的相关信息的过程。其研究内容包括信息的获取（Acquisition）、表示(Representation)、存储(Storage)、组织(Organization)和访问(Access)等几个方面。信息检索系统可以用一个四元组表示:$S=\{D,Q,F,R(q_i,d_j)\}$[3]。

2　常见的信息检索模型

随着信息和知识环境的不断变化，对信息检索的要求越来越高，对信息检索模型的研究也一直在进行之中。比较常见的信息检索模型分类见图 1。

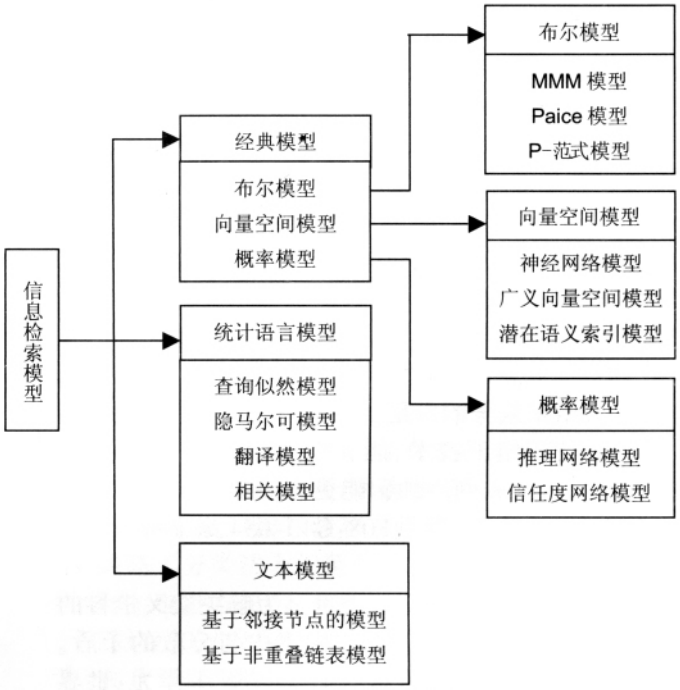

图 1　信息检索模型分类

总之,《中图法》第五版 F 类在类目的规范化方面虽取得了显著成果,但仍然存在着遗漏,仍有待于进一步的规范,使其具备通用性更高、理论更成熟、知识分类体系更完整、逻辑性更严格等优势,能更好地为广大用户服务。

参考文献

[1]　国家图书馆. 中国图书馆分类法（第五版）[M]. 北京:北京图书馆出版社,2010:7-8

[2]　肖桂山. 西方经济学[M]. 大连:东北财经大学出版社,2005:11-13

[3]　刘家顺,杨洁. 产业经济学[M]. 北京:中国社会科学出版社,2006:1-5

[4]　安应民. 旅游学概论[M]. 北京:中国旅游出版社,2007:1-5

[5]　付菊,徐沈新. 保险业概论[M]. 北京:电子工业出版社,2007:10-11

[6]　日本经济新闻社. 景气一百题[M]. 上海:上海远东出版社,1994:43-47

[7]　宋玉华. 世界经济周期理论与实证研究[M]. 北京:商务印书馆,2007:2-3

[8]　徐大均. 互助级怎样解决耕地作业上的矛盾[M]. 福州:福建人民出版社,1955:2-3

[9]　薛荣久. 世界贸易组织概论[M]. 北京:高等教育出版社,2006:6-7

[10]　曹军,孙福春. 现代企业经营管理基础[M]. 北京:中国农业大学出版社,2006:2-3

布尔模型是以简洁易懂的方式表示查询和文档,相似度以是否满足布尔表达式为依据;向量空间模型是以向量的形式来表示查询和文档，通过计算向量相似度来判断查询与文档的相似度，并将查询返回的结果文档集按相似度进行排序；概率模型是基于概率排序原理，考虑词条、文档之间的内在联系，利用词条之间和词条与文档之间的概率相似度进行信息检索；统计语言模型是假设每个文档都存在一个语言模型的基础上，从文档的语言模型抽样产生检索概率来表示文档与查询的相似度；基于本体的信息检索模型克服了基于统计的信息检索模型缺乏语义理解的问题，在信息检索模型中通过引入本体对领域知识进行表示，通过定义和共享共同的领域知识来建立人机交流平台，促进用户和信息系统对领域知识的共同理解,提高知识检索层次[4-5]。

3　基于量子的交互式信息检索模型

无论是经典的布尔模型、向量空间模型、概率模型，还是随着环境变化和技术发展涌现出来的语言模型、基于本体的信息检索模型,都从不同层面丰富了信息检索模型的研究内容。即使是最好的信息检索模型也无法时刻给予用户查询的最佳答案。特别是在网络搜索系统,不同背景、不同知识水平的用户所提交查询的质量也参差不齐。在这种情况下,不可能，也不应该对用户查询的方式、方法做统一的要求。用户对于信息需求描述的模糊性总是存在的,是不可避免的。因此,对检索系统的要求应该是接受并适应这种信息描述的模糊性，并不断完善系统自身的处理机制,以提高检索的准确性。

英国学者 Benjamin Piwowarski 和 Mounia Lalmas 认为用户和检索系统之间存在着交互作用,网站搜索引擎可以根据量子物理概率形式构建交互信息检索(基于量子的交互式信息检索模型)以向用户查询提供较好答案。交互式信息检索(IR)系统是指用户通过一系列搜索系统的交互作用来访问信息[6]。

3.1　信息需求空间

信息用户是指具有某种信息需求并利用信息资源的个人和社会团体。信息需求是信息用户对于信息内容和信息载体的一种期待状态。用户信息需求可被表示为量子物理中的一个系统,即在 Hilbert 空间[7]中作为一个单位矢量,同时当用户与系统交互时,此状态就会转变。

依据量子概率形式，信息需求向量形成了 Hilbert 空间不同子空间中的一个概率分布。模型假设:在其他可能的使用中,子空间能被联系到文档的相关性,因此使用相关分数计算文档,并尽可能地开发这些交互作用。

从几何角度，利用子空间来描述信息需求“区域”已经(有时隐式)被研究并受到了以向量空间表征为基础的一些工作的支持。在搜索过程的开始阶段,用户信息需求处于一种特殊情况,而且是所有可能的纯粹信息需求的混合。那就是此模型没有任何有关用户的信息,只能知道用户是在一定的概率下,所有可能信息需求中的一种，这种概率取决于信息需求受欢迎的程度。

此模型认为，由于用户在搜索过程中改变了他们的观点，所以使用信息需求空间可以构建交互信息检索，而且时事性的相关性和相反性预计会在搜索过程中改变。更确切地说,此模型可以在搜索过程中找出两种不同类型的动力学:(1) 从系统观点角度,信息需求变得越来越特别,例如当一个用户确定一些关键词或点击一些文件以降低不确定性。(2)信息需求以用户的观点角度加以改变。当用户阅读一些文档时,信息需求变得更有针对性,或者当涉及用户利益时,信息需求也可以稍微漂移。

第一种类型的标准概率框架很容易地被描述(此模型将信息需求限制于整体空间的子空间中),当信息需求可以从两个重叠子空间漂移时，后者将受益于量子概率形式。此模型假定,经典概率框架阐述有关检索过程中系统观点的不确定性，而量子概率框架阐述用户内部状态的变化。当量子概率框架是一种广泛概率框架时，此模型就可以用同样的表达式和演化算法来构建这两个过程。

3.2　量子观点

量子概率可以被视为经典概率理论的延伸,并且依据 Hilbert 空间的线性代数。逻辑命题的等价物或者事件 A 的等价物是一个子空间或相当于其相关推理 O_A(被称为是/否观测值)。

有关概率分布的所有信息可以包含于一个密度算子 ρ 中,也可以被表示为:Hilbert 空间的任何概率分布中存在着相应的密度算子。一个密度算子 ρ 可以写成推理的混合 $\rho=\sum_o Pr(O)O$,其中推理 O 和推理 Pr(O)总数的合计为 1。注意纯净状态可以被定义为一种密度，此密度等于一维推理。指示 tr 微量算子,密度算子 ρ 的事件 O_A 概率被表示为:

$$Pr\rho(O_A)=tr(\rho O_A) \quad (1)$$

从实用的角度来看,上述以 Hilbert 空间对标准概率的描述,通过几何关系解开定义概率的潜在性。此模型假设，在此阶段能够构建搜索过程的第一个组件,相对应的找到信息需求适宜的子空间,即从经典的术语角度发现信息需求样本空间的子集。然而,人们直觉地认为信息需求并不是相互排斥。此模型做出的假设是这种非排斥性可以通过信息需求空间的几何形状来获取，而且可以在量子概率形式的基础上构建。

3.3　叠加、混合和信息需求

此模型介绍混合和叠加的概念，并列出它们与此模型交互信息检索模型中它们使用之间的关系。简单地说，叠加与本体不确定性相关(系统状态完全是已知的，但只有在一个给定概率的情况下，一些事件是真实的)，而混合与标准概率不确定性相关(系统是处于给定概率的一种状态)。混合是一个显著的量子概率特征，而且重要的是它给予此模型一种方法来代表几何新型的信息需求，同时量子概率框架确保此模型仍然可以计算新的信息需求概率。混合和叠加以此模型能够代表目前信息需求知识状态的方式来给予此模型更大的灵活性。

举例说明：假设 $\omega_T=(10)^T$ 和 $\omega_L=(01)^L$ 构成了信息需求空间的基础(T 表示矩阵的转置)。假设前者代表查询有关老虎(T)信息的用户信息需求，而后者是查询有关狮子(L)信息的用户信息需求。为了表达用户查询虎狮(老虎和狮子所生的后代)，此模型认为这可以由空间 $\omega_{TL}=\frac{1}{\sqrt{2}}(\omega_T+\omega_L)$ 表示，此空间是两个信息需求的叠加，其中 $\frac{1}{\sqrt{2}}$ 因素确保 ω_{TL} 规范。这是一个强烈的假设，即模型研究什么时候测试此框架。美国学者 Gabora 和 Aerts 研究了如何在(量子)向量空间中使用增加维数空间来结合概念。作为对信息需求叠加的最后评论，此模型强调，复杂的数字能被用来结合信息需求，例如从狮虎（狮子是父亲)中区分虎狮(老虎是父亲)，而且叠加不限于时事性。例如，假设此模型知道如何代表一个用户搜索段落和寻找章节。

假设的信息需求 ω_{TL} 与对老虎或狮子感兴趣的用户的信息需求不同。后者可以表示为 ω_T 和 ω_L 信息需求的混合。信息需求与密度算子的表达式 $\rho_{T\vee L}=\frac{1}{2}(\rho_T+\rho_L)$，其中 ρ_T 和 ρ_L 分别表示与 ω_T 和 ω_L 相关的推理，例如 $\rho_T=\omega_T\omega T^{T_0}$ 密度算子 $\rho_{T\vee L}$ 以 50%概率解释为信息需求是与老虎有关的(或同样有关狮子的)。

如果此模型以(ω_T,ω_L)为依据的矩阵来表示密度，那么此模型可以发现其差异。此模型中信息需求混合 $\rho_{T\vee L}=\frac{1}{2}\begin{pmatrix}1&0\\0&1\end{pmatrix}$，其不同于纯信息需求 $\rho_{T\vee L}=\frac{1}{2}\begin{pmatrix}1&1\\1&1\end{pmatrix}$。个重要的发现是这些不同密度暗示了不同概率。让此模型假设文档的相关性对应是/否观察，而且有关狮子文档的相关性(分别是老虎，虎狮)由 ω_T,ω_L 和 ω_{TL} 分别产生的与子空间相关的推理(是/否观察)O_L,O_T 和 O_{TL} 所表示。根据公式(1)，此模型能够计算不同文档间相关性概率，即：

$$\mathrm{Pr}\rho_{TL}(O_L)=\mathrm{Pr}\rho_{T\vee L}(O_L)=\frac{1}{2};\mathrm{Pr}\rho_{TL}(O_{TL})=\mathrm{Pr}\rho_{T\vee L}(O_{TL})=\frac{1}{2}$$

有趣的是，当信息需求是有关老虎和狮子或者有关狮虎时，此模型无法分辨有关狮子文档的概率。其中涉及两个原因：在前者中，$\frac{1}{2}$ 概率是由信息需求和文档之间的差异所产生的，而在第二种情况下的概率是由于文件只涵盖一部分信息需要所决定；由于量子形式，当此模型评估有关虎狮文档的相关性时，相同信息需求的概率是不同的。因此此模型通过二维空间来区分不同的信息需求，这些需求同样地表达于标准向量空间模型。结果是，如果此模型查询满足 T 或 L 文档集，那么此模型会有两种类型的文档(有关老虎和狮子的，假设每个文件只涵盖信息需求)，而一个文档会满足 TL。

当此模型不知道是哪个地方用户时，混合也可用来表示信息检索过程开始阶段的信息需求密度算子 ρ_0。此模型要定义初始密度算子为 $\rho_0=\sum_i \mathrm{Pr}_i\mathrm{R}_i$，其中 i 涉及所有可能的纯信息需求 R_i 而且 Pr_i 是个概率，即当随机用户开始搜索时，他可能有信息需求。混合的使用也是受到此模型处理非决定论的事实的启发。这种混合也可以被作为一组描述所有可能信息需求的向量(每个向量与概率相关)。此模型显示信息需求 ρ_0 通过交互作用被转化。

3.4　测量和交互作用

除了区分混合和叠加外，量子形式也有条件概率的计算结果。这些结果与量子物理中测量的方式相关。此模型使用测量来构建交互作用，并描述测量如何改变密度算子以及此模型如何把测量与不同的交互作用连接起来。

为了简单起见，此模型现使用 O_A 以指示相关的是/否观测值，子空间或推理。由于它们之间存在一一对应关系，所以即使在不同的情况下，它们也可以被用来表示同一件事。给定一个系统密度算子 ρ，如果此模型观察到 O_A，那么新的密度算子 $\rho\triangleright O_A$ 可以表示为：

$$\rho\triangleright O_A=O_A\rho O_A/\mathrm{tr}(\rho O_A) \qquad (2)$$

限制 ρ 于 O_A 所定义的子空间的数量，以及确保 $\rho\triangleright O_A$ 仍是个密度算子。限制的效果是把每个混合 ρ 的信息需求定位于 O_A 所定义的子空间中(用重整化以确保概率总和为 1)。人们很容易确定对应 $\rho\triangleright O_A$ 的 O_A 概率是 1。这意味着当 A 刚刚被测量时，此模型知道在深度交互改变密度算子之前它是真的。测量可以作为条件的一个特例，正如此模型可以在给定 O_B 的情况下计算 O_A 的条件概率，或者更准确地测量 O_A，知道此模型已经测量 O_B，当 $\mathrm{Pr}\rho(O_B|O_A)=\mathrm{P}\rho\triangleright O_A(O_B)$。

在量子理论中，测量的顺序很重要，因为一般情

况下密度 $\rho \triangleright O_A \triangleright O_B$(两次使用公式(2),先是计算 O_A,再计算 O_B)与 $\rho \triangleright O_B \triangleright O_A$ 是不同的。当后续测量系统给出不同结果时,交互作用序列表明用户信息需求的改变应该被考虑进来。从一种信息需求(如巴塞罗那的酒店)漂移到另一个信息需求(如巴塞罗那的博物馆)的用户与反过来是不一样的,这说明丰富的量子形式来处理这类漂移。它生动地展示了先测量 O_B(酒店)再测量 O_A(博物馆)与顺序反过来测量的不同,因为在一个案例中,信息需求向量存在于子空间 C 中,而它们存在于另一个案例的子空间 B 中。

此模型以初始密度算子 ρ_0 假设信息检索系统和用户之间的每个隐式或显式交互作用回应每个测量,也就是说,每一次交互作用与是/否观测值 O 有关。交互作用之后,此模型使用公式(2)重新计算信息需求的密度算子。例如,用户的内部情况是与 $O_{用户}$ 相关联的,其查询相关的 O_{q1} 被视为相关文档(与 O_{d1} 相关),这可以用密度算子 $\rho_0 \triangleright O_{用户} \triangleright O_{q1} \triangleright O_{d1}$ 表示。在其他用户中,此密度算子可以被用来预测其他文档的相关性。

3.5　构建观测值之间的交互作用图

为了构建观测值之间的交互作用图,此模型限制主题相关性,并认为向量空间中的维度与术语相关。此模型知道如何计算初始密度算子 ρ_0——可能近似于使用文档表示后续描述。

考虑到目前的信息需求密度算子 ρ_t,此模型可以计算文档 d 的相关 $Pr\rho_t(O_d)$的概率,提供的 O_d 是与文档 d 相关的观测值。为了构建这样一种观测值,并作为一级近似值,此模型可以假设每一段的 p 相应于一个确切的 ω_p,因此它表示为一维子空间。然后尽可能地计算各向量$\{\omega_p\}$跨度的子空间来回应不同段落,并使用此空间来表示文档关联性。当一个用户认为一个文档相关时,那么此模型可以用同样的表达来更新目前的信息需求 ρk。在这种情况下,此模型就有了一个新的密度算子 $\rho_{t+1}=\rho_t \triangleright O_d$。

此模型联系到给定的查询子空间/观测值 O_q,并且更新现有的概率密度算子 $\rho_t \rightarrow \rho_{t+1}=\rho_t \triangleright O_q$。例如查询的表达式可以通过提供的伪相关反馈计算,此模型知道如何代表文档:与 O_q 相关的子空间就是代表顶置文档的观测值所跨度的子空间。例如,图 2 中,如果 A 和 B 回应两种不同顶置文档的给定查询,那么 O_q 符合整个三维空间(如子空间 A 和 B 的结合)。另一种不依赖外部模型的计算查询观测值 O_q 的方法将是子空间集合,其中子空间代表每个查询术语出现的段落。

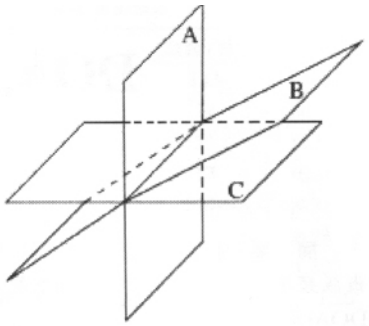

图 2　三维空间中的三个二维子空间

此模型给予交互信息检索框架量子形式有效性的例子。如果用户改变太重要而不能成为一个简单的漂移,那么查询观测值 O_q(或文档观测值 O_d)可以用来检测一个重要特征(交互信息检索系统应具有此特征)。在量子理论框架内,此模型用相同的几何模型表示来更新知道此事件的密度算子,并且计算此事件概率。实际上,在 t 时刻当用户确定一种新的查询 q′ 类型时,此模型可以根据当前的信息需求密度算子 ρ_t,即计算 $Pr_{pk}(O_{q'})$来计算查询概率。此模型的信息查询系统根据此价值确定转移到新的信息需求的用户,并进行相应调整。

4　总结

基于量子的交互式信息检索模型探讨了量子概率形式中几何和概率之间的强烈关系。此模型的框架将用户交互作用和几何图整合于检索模型中,演示如何处理点击/相关反馈和查询转化。到目前为止,除了提供查询推荐之外,如何使用后者信息研究还不是十分明确。但其他形式的交互作用(如导航)已符合此模型的框架。

参考文献

[1]　蔡柯柯. 基于查询特征上下文的检索模型研究[D]. 杭州:浙江大学,2007

[2]　吴丹,齐和庆. 信息检索模型及其在跨语言信息检索中的应用进展[J]. 现代情报,2009,29(7):215-221

[3]　陈圣兵. 基于商空间理论的海量信息检索模型的研究[D]. 安徽大学,2010

[4]　孙坦,周静怡. 近几年来国外信息检索模型研究进展[J]. 图书馆建设,2008(3):82-85

[5]　焦玉英,温有奎,陆伟,等. 信息检索新论[M]. 武汉大学出版社,2008

[6]　Piwowarski B,Lalmas M. A Quantum-based Model for Interactive Information Retrieval (extended version)[M]. USA:ArXiv e-prints,2009

[7]　Aerts D,Gabora L. A theory of concepts and their combinations II:A Hilbert space representation[J]. Kybernetes,2005,34(1):192-221

收稿日期:2011-05-13

作者简介:徐连杰(1987—),女,2010 级硕士研究生,研究方向为信息研究与信息服务;胡德华(1972—),男,副教授,研究方向为信息检索、开放存取期刊、信息检索语言、生物信息学、期刊评价。