

信息获取中两类不确定性的风险表达模型研究*

杜文华

(中南民族大学管理学院,湖北武汉,430074)

摘要:分析了搜索引擎在信息获取中效率低下的原因,在 Lafferty 等人提出的风险表达形式化模型基础上,提出了风险表达的认知模型,并对该模型中风险函数 R 的可计算问题进行了分析和讨论。

关键词:信息获取;风险认知模型;风险函数

中图分类号:TP18

文献标识码:A

搜索引擎是当前互联网上最常用的信息获取工具,也是在互联网上最先商业化的一个应用服务,它产生的经济价值非常巨大。但是,在目前的信息获取中,其性能和效率变得十分低下。原因之一是海量信息和非结构化信息日益增多使得信息的有效管理日益艰难;另外一个重要的原因是信息获取中的两类不确定性因素的影响。这里,两类不确定性因素主要是指:第一,查询者所要表达的真实查询意图或者文档作者所要表达文档主题的真正含义不能被系统所获知,这种不确定性被称为语义不确定性;第二,信息获取中,查询模型与文档模型之间相关性是模糊的、不确定的,这种不确定性被称为相关不确定性。

本文将从风险的角度来探讨这两类不确定性。

1 风险表达形式化模型

在本文中,风险定义描述如下:风险是针对用户查询相关性预期结果估计中的较为不利的一面。这里“较为不利”是相对于我们预期达到的目标结果而言的。风险表示一种观点,即用各种可能性的统计的观点来观察、研究信息获取中的两类不确定问题,使得考虑更全面、决策更合理。

20 世纪 70 年代,研究人员就开始从不同角度研究和讨论信息获取领域中的决策问题。2000 年,Harman, Lafferty 和 Zhai 等人提出查询模型和文档模型中的两类不确定问题,并给出相应风险决策的形式化描述。

Harman, Lafferty, Zhai 等人在描述信息获取中两类不确定问题时定义两个模型,一个模型是查询模型,另外一个模型是文档模型。查询模型和文档模型分别通过两个独立的 Markov 链 $u \rightarrow \theta_q \rightarrow q$ 和 $S \rightarrow \theta_d \rightarrow d$ 加以表达。这里, θ_q 代表查询模型中基于用户某一次查询的特征参数, $u \rightarrow \theta_q \rightarrow q$ 表示一个查询模型, $\theta_d = \{\theta_{d1}, \dots, \theta_{dn}\}$ 代表文档模型中的基于某一篇文档的特征参数, $S \rightarrow \theta_d \rightarrow d$ 表示一个文档模型。由于参数 u, q, S, C 代表真实的状态往往不能直接观察,随机变量 θ_q, θ_d 是可以被系统观测到的。 Θ_q 和 Θ_d 表示随机变量 θ_q, θ_d 的样本空间。

现在,假定有一个可供选择的方案集 $A, A = \{(D_i, \pi_i)\}$, 其中 $a_i = (D_i, \pi_i) \in A$, 这里, D_i 表示针对某一查询策略 π_i 获得的结果集, $D_i \in \Pi(C)$, $\Pi(C)$ 是 C 的幂集, $C = \{d_1, d_2, \dots, d_n\}$; 策略 $\pi_i = \Sigma(i=1, 2, \dots)$, Σ 表示查询策略空间。

Lafferty 和 Zhai 等人在查询模型和文档模型基础上给出最小期望风险下的最优决策形式化描述为:

$$a^* = (D^*, \pi^*) = \underset{a \in A}{\operatorname{argmin}} R(D_i, \pi_i, q, S, C) \quad (1)$$

其中, $D_i \in \Pi(C)$, $\pi_i = \Sigma(i=1, 2, \dots)$ 。此时,关于方案 a_i 的风险函数可以描述为:

$$R(D_i, \pi_i, q, S, C) = \int_{\Theta} L(D_i, \pi_i, \theta, F(u), F(s)) p(\theta | u, q, S, C) d\theta \quad (2)$$

其中, $\Theta = (\Theta_q, \Theta_d)$, $\theta \in \Theta$ 。

在式(2)中,由于两个模型是两个独立的 Markov 链,所以 $p(\theta | u, q, S, C)$ 还可以进一步表示为:

$$p(\theta | u, q, S, C) = p(\theta_q | u, q) \prod_{i=1}^n p(\theta_{di} | S, d_i) \quad (3)$$

式中: $p(\theta | u, q, S, C)$ 为 θ 的条件概率分布; $p(\theta_q | u, q)$ 是 θ_q 的条件概率分布; $p(\theta_{di} | S, d_i)$ 是 θ_{di} 的条件分布函数; $p(\theta_q | u, q)$ 和 $p(\theta_{di} | S, d_i)$ 分别表示查询者在表达查询语义时的不确定性和文档作者在表达文档主题时的不确定性。

在以上讨论基础上, Lafferty 和 Zhai 还给出基于期望风险最小下的相关文档集合描述为:

$$D^* = \underset{D_i}{\operatorname{argmin}} R(D_i, q, S, C) =$$

$$\underset{D_i}{\operatorname{argmin}} \int_{\Theta} L(D_i, \theta, F(u), F(s)) p(\theta | u, q, S, C) d\theta \quad (4)$$

同样,基于期望风险最小下的最优决策策略 π^* 可以描述为:

$$\pi^* = \underset{\pi_i}{\operatorname{argmin}} R(\pi_i, q, S, C) =$$

$$\underset{\pi_i}{\operatorname{argmin}} \int_{\Theta} L(\pi_i, \theta, F(u), F(s)) p(\theta | u, q, S, C) d\theta \quad (5)$$

相对于查询者的某一查询而言,可以有不同的查询策略(查询的风险策略空间)供 Robot 系统作选择,由它在其中寻找一种期望风险最小方案,也是与用户查询意图最相关的文档集合返回给用户。

下面讨论对风险表达形式化模型中损失函数的计算,并以驻留概率分布(Stop Probability Distribution)下的损失函数计算为例。驻留分布是这样表述的,假定针对某一个用户查询需求,由 Robot 按照某一种查询策略 π_i 执行查询任务获得 N 个相关文档。某用户浏览了其中 k 个文档,最后停留在第 k 个文档上,且有 $\sum_{i=1}^N p_i = 1$ 的值由 $F(u)$ 决定。

这时式(4)中的损失函数可以表示为:

$$L(D_i, \theta, F(u), F(s)) = \sum_{j=1}^k p_j(d_{j1}, d_{j2}, \dots, d_{jn}, \theta, F(u), F(s)) \quad (6)$$

式(6)表示用户按照某种查询策略 π_i 获得 N 个相关文档,但是用户只浏览了其中 k 个文档时的风险损失; $l(d_{j1}, d_{j2}, \dots, d_{jn}, \theta, F(u), F(s))$ 表示假定某一位用户在浏览 d_{jn} 时,已经浏览过 d_{j1}, \dots, d_{jn-1} , 在浏览 d_{jn} 时所获得的风险损失。

2 风险表达形式化认知模型

在信息获取中,查询模型和文档模型都是围绕着特定的查询者进行构建的,为了使查询模型或者生成文档模型的构建更具普遍意义,本文在 Lafferty 等人提出的风险表达形式化模型基础上,从认知层次讨论这个问题,引入了一个隐藏变量 H (H 与认知、情绪、感情等因素有关),提出风险表达的认知模型。

在 Helmholtz(赫尔姆霍茨)机器学习中,用 $p(\theta_q | H)$ 表示生成用户查询模型,而用 $p(\theta_d | H)$ 表示生成文档模型。将 $p(\theta_q | H)$ 和 $p(\theta_d | H)$ 分别视为对 θ_q 和 θ_d 的认知过程。在文档特征 θ_d 为已知的前提下,基于隐藏变量 H 的生成查询模型形式化描述为:

$$p(\theta_q | \theta_d) = \int_{\Xi} p(\theta_q, H | \theta_d) dH = \int_{\Xi} p(\theta_q | H) p(H | \theta_d) dH \quad (7)$$

其中, Ξ 是隐藏变量 H 的样本区域空间。

同样, $p(\theta_d|\theta_q)$ 表示从信息获取角度寻找与 θ_q 特征最相似的 θ_d 特征, 其形式化描述为:

$$p(\theta_d|\theta_q) = \int_{\Xi} p(\theta_d|H)p(H|\theta_q)dH \quad (8)$$

可以利用样本数据集 $p(\theta_d|\theta_q)$ 对 $p(\theta_d|H)$ 和 $p(H|\theta_q)$ 进行建模。

将 Helmholtz (赫尔姆霍茨) 机器学习视作一个 Bayesian 学习网络, CPT 中的初始值随机给定。具体步骤表述如下:

步骤 1: 给出某用户基于某一次查询 q 的特征参数 θ_q 的先验分布, 用 $p(\theta_q)$ 表示。

步骤 2: 执行信息获取任务, 根据式(8)计算获得 $p(\theta_d|\theta_q)$ 。

步骤 3: 对 $p(\theta_d|\theta_q)$ 进行评价。并按照 EM 算法对 Bayesian 学习网络模型(CPT 中的值)进行调整、修改。

步骤 4: 重复步骤 1~3。

按照上面步骤, 通过 Bayesian 网络学习方法同样可以得到一个生成查询模型。

引入隐藏变量 H 以后, 在策略 $\pi, \epsilon \in \Sigma$ 下的风险函数可以描述为:

$$R(D|u, q, S, C) = \int_{\Theta} \int_{\Xi} L(D, \theta, H, F(u), F(s)) p(H|\theta, q, S, C) dH d\theta \quad (9)$$

式(9)可以进一步描述为:

$$R(D|u, q, S, C) = \int_{\Theta} \int_{\Xi} L(D, \theta, H, F(u), F(s)) p(H|\theta) p(\theta|u, q, S, C) dH d\theta \quad (10)$$

3 风险函数 R 的可计算性讨论

本节将对上一节引入隐藏变量 H 以后, 风险函数 $R(\cdot)$ 的可计算性问题进行分析和讨论。

对式(10)做进一步变换:

$$R(D|u, q, S, C) = \int_{\Theta} \int_{\Xi} L(D, \theta, H, F(u), F(s)) p(H|\theta) p(\theta|u, q, S, C) dH d\theta = \int_{\Theta} \left\{ \int_{\Xi} L(D, \theta, H, F(u), F(s)) p(H|\theta) dH \right\} p(\theta|u, q, S, C) d\theta \quad (11)$$

$$\text{并令 } \psi(D, \theta, F(u), F(s)) = \int_{\Xi} L(D, \theta, H, F(u), F(s)) p(H|\theta) dH \quad (12)$$

这时, 式(11)可以描述为:

$$R(D|u, q, S, C) = \int_{\Theta} \psi(D, \theta, F(u), F(s)) p(\theta|u, q, S, C) d\theta \quad (13)$$

实际在对式(13)的计算求解时还是要把积分中的表达式进行离散化。如果对于不同 $\theta \in \Theta$ 可以得到相应 $\psi(\cdot)$ 的值, 可以认为引入隐藏变量 H 以后的风险函数 R 是可计算的。这样, 我们就将对风险函数 R 的可计算问题研究转化为对 $\psi(\cdot)$ 的可计算性问题的研究。

下面讨论 $\psi(\cdot)$ 中损失函数 $L(\cdot)$ 。在信息获取中, 损失函数 $L(\cdot)$ 表达查询模型和文档模型之间相关性。并且:

$$L(D, \theta, H, F(u), F(s)) = \sum_{d \in D} k(d, \theta, H, F(u), F(s)) \quad (14)$$

$$\text{其中, } k(d, \theta, H, F(u), F(s)) = k(\theta_d, \theta_q, H) \quad (15)$$

$L(\cdot)$ 主要依赖于 θ_q, θ_d, H 几个参数。这里 $\theta_q = \theta_d, \theta_d$ 表示 d 的文档模型特征参数。

由式(19)和式(20), 式(17)可以进一步描述为:

$$\begin{aligned} \psi(D, \theta, F(u), F(s)) &= \int_{\Xi} L(D, \theta, H, F(u), F(s)) p(H|\theta) dH = \\ &= \int_{\Xi} \sum_{d \in D} k(d, \theta, H, F(u), F(s)) p(H|\theta) dH = \\ &= \sum_{d \in D} \int_{\Xi} k(d, \theta, H, F(u), F(s)) p(H|\theta) dH \end{aligned} \quad (16)$$

假定 θ_d 在区域 Θ_d 内取值, 这时式(16)可以进一步描述为:

$$\begin{aligned} \psi(D, \theta, F(u), F(s)) &= \int_{\Xi} L(D, \theta, H, F(u), F(s)) p(H|\theta) dH = \\ &= \sum_{\theta_d \in \Theta_d} \int_{\Xi} k(\theta_d, \theta_q, H) p(H|\theta_d, \theta_q) dH \end{aligned} \quad (17)$$

现在, 利用效用函数去定义损失函数 $k(\theta_d, \theta_q, H)$,

$$u(\theta_d, \theta_q, H) = -k(\theta_d, \theta_q, H) \quad (18)$$

效用函数 $u(\cdot)$ 表示 θ_d 和 θ_q 之间的效用相关性, 效用相关性用概率率方法进行描述。

$$u(\theta_d, \theta_q, H) = p(\theta_d|\theta_q) \quad (19)$$

由式(17)、式(18)和式(19)可以得到:

$$\begin{aligned} \psi(D, \theta, F(u), F(s)) &= \int_{\Xi} L(D, \theta, H, F(u), F(s)) p(H|\theta) dH = \\ &= - \sum_{\theta_d \in \Theta_d} \int_{\Xi} p(\theta_d|\theta_q) p(H|\theta_d, \theta_q) dH \end{aligned} \quad (20)$$

在 $\psi(\cdot)$ 中, 由于

$$\begin{aligned} p(\theta_d|\theta_q) p(H|\theta_d, \theta_q) &= \frac{p(\theta_d|\theta_q) p(\theta_d, H, \theta_q)}{p(\theta_q, \theta_d)} = \\ &= \frac{p(\theta_d|\theta_q) p(\theta_d|H, \theta_q) p(H|\theta_q) p(\theta_q)}{p(\theta_d|\theta_q) p(\theta_q)} = \\ &= p(\theta_d|H, \theta_q) p(H|\theta_q) = p(\theta_d|H) p(H|\theta_q) \end{aligned} \quad (21)$$

所以

$$\begin{aligned} \psi(D, \theta, F(u), F(s)) &= - \int_{\Xi} L(D, \theta, H, F(u), F(s)) p(H|\theta) dH = \\ &= - \sum_{\theta_d \in \Theta_d} \int_{\Xi} p(\theta_d|H) p(H|\theta_q) dH = - \sum_{\theta_d \in \Theta_d} p(\theta_d|\theta_q) \end{aligned} \quad (22)$$

$$\text{其中, } p(\theta_d|\theta_q) = \int_{\Xi} p(\theta_d|H) p(H|\theta_q) dH$$

在式(8)中, 假定在样本数据集 $p(\theta_d|\theta_q = \theta_q)$ 已经存在, 并根据 EM 算法可以对 $p(\theta_d|H)$ 和 $p(H|\theta_q)$ 进行建模, 因此, 由式(22)可知, $\psi(\cdot)$ 是可计算的, 从而风险函数 R 也是可计算的。

4 结语

风险表达模型是一种很好的分析和处理不确定性问题的工具。为了构建通用生成查询模型或者生成文档模型, 在 Helmholtz 机器学习基础上, 引入一个与认知、情绪和感情等因素有关的隐藏变量 H , 提出基于隐藏变量 H 的生成查询模型和生成文档模型的形式化描述, 以及相应的风险表达描述, 并分析和讨论因此而产生的风险函数 R 的可计算性问题。

* 本文为高校自然科学基金资助项目(编号: YZQ05009)论文。

参考文献

- [1] Harman. Overview of the first TREC conference [G]//Korfhage R, Rasmussen E, Willet P, Proceedings of the 16th ACM SIGIR. New York: [s. n.], 1993: 36-47.
- [2] Stephen P Harter. A probabilistic approach to automatic keyword indexing. Part II. An algorithm for probabilistic indexing [J]. Journal of the American Society for Information Science, 1975(9-10): 280-289.
- [3] Cooper W S, Maron M E. Foundations of probabilistic and utility theoretic indexing [J]. Journal of the Association for Computing Machinery, 1978, 25(1): 67-80.
- [4] Binder J, Koller D, Kanazawa K, et al. Adaptive Probabilistic Networks with Hidden Variable [J]. Machine Learning, 1997(29): 213-244.
- [5] Motomura Y, YOSHIDA K, Fujimoto K. Generative user models for Adaptive Information Retrieval [G]//Proc. of 2000 IEEE International Conference on System, Man, and Cybernetics, Tennessee: [s. n.], 2000: 665-670.
- [6] 傅京孙, 蔡自兴. 人工智能及其应用 [M]. 北京: 清华大学出版社, 1987.
- [7] Nils J Nilsson. 人工智能 [M]. 北京: 机械工业出版社, 2000.

(实习编辑: 张 瑛)

第一作者简介: 杜文华, 女, 1976年7月生, 2005年毕业于武汉大学(博士), 讲师, 中南民族大学管理学院, 湖北省武汉市, 430074.

管理咨询类企业绩效考核模式探析

刘景生

(迁西县委党校,河北迁西,064300)

摘要:在总结我国咨询业发展现状及特征的基础上,介绍了咨询业绩效考核体系总体设计思路,并对咨询业绩效考核指标进行了说明。

关键词:咨询业;绩效考核;指标体系

中图分类号:C932

文献标识码:A

随着中国的人世,相比制造业的国民化待遇而言,服务业的全面开放更是我国经济发展的软肋,我国服务性企业将会直接面对更加残酷的竞争。要提升我国企业的竞争力,必须拥有更加规范的管理,尤其是在以人为主要因素的管理咨询行业,建立科学的绩效考核体系更是我们面临的一项紧迫任务。

1 中国咨询业发展现状和特点

1.1 中国咨询业发展现状

随着市场经济不断走向成熟和发展,企业对管理咨询的需求越来越大。2003年11月新华社在《2003年度中国管理咨询行业市场发展报告》中指出,2002年我国管理咨询公司的新增客户数量增长80%,市场渗透率增长32.5%,咨询产业将是我国21世纪最具希望的朝阳产业。近几年我国咨询市场规模在迅速扩张,从1996年的21.85亿元增长到2002年的302亿元,6年间增长了13.8倍,年均增长速度为69.09%,超过了同期我国快速发展的电信产业。

1.2 中国咨询业发展特点

总的来说我国的咨询产业已经初具规模,并呈现出以下特点:

(1)市场化程度较高。我国经过数十年的改革开放,市场经济程度不断提高,使得企业面对的市场竞争压力较大,需要管理咨询专业的服务,外资咨询公司开始进入,带动了我国管理咨询业发展,出现了一批管理咨询公司,通过借鉴国外和香港管理咨询的理论、方法和经验,按市场规律运作,为企业提供规范化咨询服务,并培育了自身的核心竞争力。

(2)逐步形成了专业化的管理咨询体系。我国的管理咨询企业在十几年的发展过程中,逐步形成不同市场定位的咨询服务,形成了如政策咨询、战略咨询、财务顾问、企业管理、人力资源管理、生产管理商业调查和市场研究、营销策划、CIS策划等专业化方向,打造一大批有影响、有实力、有品牌的管理咨询机构。随着市场经济的发展,这些管理咨询企业以其高质量的专业化服务赢得市场,并将业务拓展到了全国各地。

(3)逐步形成了现代化的咨询服务模式。我国的管理咨询企业,通过学习国外出色咨询企业的咨询经验与程序,结合自己企业的专业优势,形成了符合我国咨询企业实际的一套服务手段、技术方法、服务模式,并不断加以改善,从单一的提供咨询报告(方案)的服务方式发展为围绕客

户的问题开展各种培训、辅助实施、辅助决策、决策后的实施执行和客户委托管理等,注重观念的创新和方法的领先,注重方案的有效执行和咨询效果评价的增值服务。

(4)管理咨询机构的规模较小。尽管我国管理咨询机构的自身能力、服务质量有所提高,但从总体看大部分管理咨询机构的规模较小。

(5)行业发展环境亟待改善。作为一个新兴的行业,目前我国的管理咨询业仍然存在着缺乏行业规范、管理落后、人员素质参差不齐等问题,政府在行业宏观指导和调控方面缺位,使得行业发展无章可循,制约了整个行业良性发展。

2 咨询业绩效考核体系总体设计思路

设计具体的绩效考核体系时,需要对考核对象公司进行具体的调研,获得企业足够的背景资料,对企业的发展愿景、经营状况、业务流程以及今后企业发展思路等相关问题进行较为深入的了解。本研究所设计的咨询业绩效考核指标体系只是对咨询业共性的问题进行描述,对于同属咨询业不同企业的差异性而言,由于管理基础不同,员工素质不同,企业文化不同,同时面临的竞争环境差距较大,因此在各自的发展进程中将采取不同的战略,所设定的目标必须要有针对性。

2.1 绩效考核体系的设计目标

抛开具体企业带来的差异性,对于绩效考核体系设计的本身而言,应实现以下几个方面的目标。

(1)与战略挂钩,使企业的绩效考核无论是组织绩效考核还是员工绩效考核都是源于战略目标进行系统思考确定的。

(2)建立绩效管理循环,加入检验、修订等过程,从而增加考核指标制定的针对性。

(3)建立系统的考核体系,促进各个环节工作的良性循环,从而提高组织的绩效和员工的绩效。

(4)合理地建立各层员工之间的考核关系,传递层级式的压力和动力,激发各级员工的工作积极性。

2.2 绩效考核设计体系的设计流程

基于以上提出的各种设计目标,具体设计流程见图1。

2.3 绩效考核设计体系的关键因素

Research on the Risk Representation Model of Two Kinds of Uncertainties in the Information Retrieval

DU Wen-hua

ABSTRACT: This paper analyzes on the reasons of the lower efficiency of search engineer in the information retrieval, based on the risk representation formalization model advanced by Lafferty and others, advances the cognition model of the risk representation, and analyzes and discusses on the calculable problem of risk function R in this model.

KEY WORDS: information retrieval; cognition model of risk; risk function