维普资讯 http://www.cqvip.com

信息检索相关性理论的语义基础分析*)

杨志峰[1] 王 斌[2] 李素建[3]

(中国科学院计算技术研究所 北京 100080)[1,2] (北京大学计算机系 北京 100080)[3]

摘 要 目前信息检索系统的评估方式主要是黑箱方法,无法了解系统内部情况。逻辑方法为比较不同类型的信息检索系统的内部功能提供了途径。为此一些基本的定义和假设已经被建立,并被初步应用于理论研究,但目前的逻辑框架尚未完善。本文通过分析逻辑框架的语义基础,重新定义了基本语义概念,为逻辑框架提供了语义支持,使逻辑结果更逼近实际语义。

关键词 信息检索,相关性,性能评测,语义,Aboutness

A Study of Semantics in Aboutness Theory Framework

YANG Zhi-Feng[1] WANG Bin[2] LI Su-Jian[3]

(Software Devision, Institute of Computing, The Chinese Academy of Sciences, Beijing 100080, China)[1,2]

(Department of Computer Science, Peking University, Beijing 100080, China)[3]

Abstract The most popular used method of evaluating information retrieval systems in the experimental approach at present. This kind of method cannot learn the internal function of the systems. The new logic method of information retrieval has provided a approach that can make functional comparison among different kind of systems. A basic framework, including definitions and postulations has been established and used in recent research. But the framework is not perfect yet. This paper analyzes the semantics basis of the logic framework, and redefines basic definitions and operations. These concepts provide semantics support for the logic reasoning. The reasoning results will be more accurate in semantics than that in original framework.

Keywords Information retrieval, Relevance, Performance evaluating, Semantics, Aboutness

1 引言

对于信息检索系统来说,它的基本任务是对用户的查询请求给出解答。用户根据信息需求N,按照信息检索系统的界面所要求的形式给出查询Q,信息检索系统根据Q得到它认为符合要求的文档D(一般不止一个),并把所有的结果按一定顺序返回给用户。

信息检索领域有两个主要的性能指标,即查准率(precision)和查全率(recall)。信息检索系统的性能评价一直是信息检索领域的重要课题之一。目前主要的评测方法之具体过程为:首先指定一个标准的测试数据集,通常足够大,以包括尽量丰富的文档;同时指定一个标准的问题集合(信息需求N),作为测试集合。每个问题都有确定的"正确的"答案。每个被评测系统对这些问题进行处理(得到查询Q),在指定的数据集中进行搜索,给出自己的响应(文档D)。然后评测系统根据这些系统给出的结果,比照预先给定的"标准"答案,进行统计计算,可以围绕查准率和查全率这两个核心指标给出一系列的衍生性能指标。目前比较权威的信息检索系统评测会议TREC(Text Retrieval Conference)即采用此法,但细节有所变化[1]。这种方法属于黑箱方法。它不关心被评测系统的内部细节和具体功能,只关注最后结果。因此,测评过程本身无法揭示一个性能良好的系统的优越之处,也无法发现较差系统的缺陷所在。此外,目前的信息检索方法常常依赖某些重要的常数,但这些常数并非固定的值,它们需要针对特定的数据集合进行调整,才能得到满意的结果。因此,这又为系统评测工作带来了额外的变量。

作为补充,另一种信息检索系统的评价方法已经得到了发展。van Rijsbergen在20世纪80年代建立了基于逻辑的信息检索方法[2]。这种方法不仅提供了新的信息检索思路,也为信息检索系统的比较提供了新的途径。利用基于逻辑的方法,可以尝试使用代数逻辑进行推理论证,从而避免使用基于实验的黑箱方法。

2 基于逻辑的理论框架

无论使用何种方法,相关性问题是信息检索领域最根本的问题,是无法回避的。在基于逻辑的方法框架中,Bruza等人提出了如下定义[3]:

定义 1(相关性,Aboutness) 如果由信息载体j所产生的信息在信息载体i中得到保留,那么i是有关j的(i is about j)。

信息载体(Information Carrier)指包含了一定信息的实体,它可以是文档,也可以是查询,甚至可以是单个的关键词,在基本逻辑的方法中,它们被统一对待。定义1可以从van Rijsbergen对相关性的观点推得。Bruza认为,这个定义体现

*)基金项目:国家重点基础研究发展规划973资助项目(G1998030413,G1998030510);计算所领域前沿青年基金(20026180-24)。**杨志峰** 博士生,主要研究领域为信息检索,知识挖掘;**王 斌** 博士,副研究员,主要研究领域为网络信息处理,自然语言处理;**李素建** 博士后,主要研究领域为自然语言处理,知识挖掘,机器翻译。

维普资讯 http://www.cqvip.com

的思想是：信息载体 i 作为一个环境，信息载体 j 在此环境中解释。

相关性一般用符号"⊨"表示。例如：ship⊨sea。

定义 2（信息包含，Information Containment） 信息包含用符号"→"表示。令 i、j 为信息载体，i→j，当且仅当 i 蕴含的信息已经包括了 j 所蕴含的信息。

从直观来看，i 和 j 的包含关系说明 i 包含的信息比 j 多。例如，一个广为引用的例子是 salmon→fish。

这说明 salmon 不仅包含了 fish 所包含的信息，而且含有更多的语义信息。如果用通常意义的文氏图来表示概念之间的包含关系，"fish"对应的区域将包含"salmon"所对应的区域；但在这里，情况恰恰相反，由于"salmon"包含了更多的语义信息，它对应的区域应包含"fish"的区域。

0 是一个特殊的信息载体，它包含所有其它载体。

定义 3（信息合成，Information Composition） 信息合成的一个较明显的例子是用户查询。假定某位用户在第一步使用关键词 horse 进行检索，得到最初的结果；然后，又用关键词 white 执行增量检索，得到更为精细的结果。在第二个步骤中，用户的信息需求可用 white∧horse 表达，这就是 white 和 horse 这两个信息载体进行合成得到的结果。信息合成的含义，就在于操作的结果要包括所有参加合成的信息载体组合之后所传达的信息。

信息合成用符号"⊕"表示。⊕是不可交换的。上面的举例可以表示为：white⊕horse（示例 1）。

定义 4（信息互斥，Information Preclusion） 并不是所有的信息载体都可以进行信息合成。由于语义的原因，某些载体是互相排斥的。例如给定两个载体 white⊕horse 和 black⊕horse，从直观上来看，一匹马只能是黑色或者白色的，不可能二者都成立。Bruza 认为，互斥是信息检索过程中的常见现象。他的 preferential preclusion 理论运用了互斥概念[3,7]。

信息互斥用符号"⊥"表示：i、j 互斥，即 i⊥j。

定义 5（信息域，Information Field） 信息域是一个五元结构(Ξ,→,⊕,⊥,0)，它满足以下性质：

1)Ξ 是一个信息载体的非空集合；

2)(Ξ,→)是一个 poset（部分有序集）；

3)0∈Ξ，且任给 i∈Ξ，都有 0→i；

4)⊥定义在 Ξ×Ξ 上；

5)任给 i,j∈Ξ：①i⊕j→i，i⊕j→j。②i⊕i→i。③i⊕j∈Ξ，或 j⊕i∈Ξ。

为了使用以上的概念对不同的检索系统进行比较，需要某些假设（并非公理，实际系统不一定满足）。这些假设是有关系统性质的描述。Bruza 等所利用的假设有：

假设 1 Reflexivity(R) $i \models i$

假设 2 Containment(C) $\dfrac{i \rightarrow j}{i \models j}$

假设 3 Right Containment Monotonicity(RCM) $\dfrac{k \models i \quad i \rightarrow j}{k \models j}$

假设 4 Context-Free And(C-FA) $\dfrac{k \models i \quad k \models j}{k \models i \oplus j}$

假设 5 Negation Rationale(NR) $\dfrac{k \not\models i}{k \not\models i \oplus j}$

假设 6 Left Composition Monononicity(LM) $\dfrac{i \models k}{i \oplus j \models k}$

假设 7 Right Composition Monotonicity(RM) $\dfrac{i \models j}{i \models j \oplus k}$

假设 8 Preclusion(P) $\dfrac{i \perp j}{i \not\models j \text{ and } j \not\models i}$

假设 9 Close World Assumption(CWA) $\dfrac{i \not\models j \quad j \perp k}{i \models k}$

除了上述的假设，还有针对相关性的非单调性质提出的若干假设，包括 Guarded Left Composition Monotonicity (GLM)和 Guarded Right Composition Monotonicity(GRM)等。

3 逻辑框架的语义分析

目前对于 Aboutness 理论的研究主要集中在符号逻辑上，即试图按照代数逻辑的要求，从少数定义和公理（包括可以选择的假设）出发，得到可供使用的定理。但是，作为逻辑起点的基本定义和基本运算，在某种程度上受到了忽视。这些定义和运算建立在对部分事实的观察基础上，对现实情况的概括是不完全的。也就是说，从语义角度来看，这些定义和运算缺乏完整的规定。这种情况的结果，就是在将理论成果映射到现实系统的过程中，映射规则不得不根据具体情况进行语义调整。于是，理论上的形式一致性不能保证实际应用过程中的语义一致性。

信息载体是 Aboutness 理论中最重要的基础元素，因为它提供了理论框架的运算对象。但是作为基础的成分，它没有明确的定义，仅从直观出发，规定它是一部分信息的承载者，从现实情况来看，它可以是文档，也可以是查询，甚至可以是单独的概念或词。无论实际情况如何，信息载体被统一对待。在形式上，这样做是合理的，因为可以得到统一的形式表达式。对于定义 1，语义上是没有问题的，因为一个文档可以和另一个文档相关，也可以和一个用户查询相关，或者与一个概念（表现为词或词组）相关。但是，一旦进入计算阶段，语义问题就会出现。这个问题表现为两方面。

首先，代数运算的结果在理论框架中没有语义定义。从形式上看，信息载体的运算结果仍然是信息载体；但从语义角度观察，问题比较复杂。如果两个参与运算的信息载体属于同一类型，那么运算结果仍然是这一类型，这是容易理解的；如果这两个载体的类型不一致，运算结果的语义则无法定义。仍以上文的示例 1 讨论，j⊕k 表示信息载体 j（概念"white"）和信息载 k（概念"horse"）的合成，其结果可以视为一个合成得到的信息载体，代表概念"white horse"，这是有语义结果的。现在假定存在一个文档 d，按照定义，它是一个信息载体，那么 j⊕d 在形式上有确切的结果，但在语义上没有明确定义。因此，这种运算结果无法得到。

其次，在实际的操作对象发生变化时，运算所对应的实际规则常常也随之变化，这是因为语义发生了变化。设信息载体 j 和 k 均为查询，j 由单个的概念"white"构成（注意和上文的 j 不同，那里 j 直接代表概念），k 由单独的概念"horse"构成。那么 j⊕k 代表一个合成后的查询，由一个单独的概念"white horse"构成。从语义上观察，这里 j、k 之间是 AND 关系，因为用户的意图是用这个查询表示一个既需要"white"的概念，又需要"horse"概念的信息需求。只有完全包含这两方面语义的信息载体，才是最相关的（在实际系统中，无论是向量空间模型还是概率模型，相关性均是一种概率，因此可以比较）。如果 j、k 均为文档，则 j⊕k 的语义不同于查询的情况。两个子文档的合成结果是一个新的文档，包含了原有文档的所有内容，因此容易判断，j、k 的合成过程类似于两个集合求并集的运算，是一种 OR 关系。假设 j 文档是有关概念"white"的，k 文档是

维普资讯 http://www.cqvip.com

有关概念"horse"的，那么文档 j⊕k 中有一部分内容是有关"white"的，另一部分是有关"horse"的。这种不统一的情形，在推理得到的定理中将导致不同的语义结果。考虑定理 $\frac{a\models b\oplus c}{a\models b}$（示例 2）。

如果一个实际的检索系统支持假设 3(RCM)，那么结合定义 5 可以得到示例 2 的证明。

$$\frac{a\models b\oplus c\ \ b\oplus c\rightarrow b}{a\models b}$$

可以证明，传统布尔模型支持假设 3[4]，因此该定理在逻辑框架中是成立的。假设 b、c 都是查询，必须把"⊕"映射为"∧"。从语义上看，这个定理显然成立，因为只要"b∧c"所包含的信息在 a 中得到保留，b 的信息显然也在 a 中有保留。但是，如果把 b、c 设为文档，则"⊕"必须修改为"∨"。假设 b、c 所牵涉的主题不同，且 a 与 c 相关，a 与 b 不相关，这时 a⊨b⊕c 是成立的，但根据示例 2 的形式证明，可得 a 与 b 相关，这是一个矛盾的结果。因此，同一个定理，针对不同的操作对象，可能并不总成立。

信息载体的统一表示，也会给定义 1 在逻辑推理中的应用造成困难。从语义角度观察，相关关系应是可交换的。信息载体 a 与信息载体 b 相关，应该可以直观地得到 b 和 a 也是相关的。考虑假设 7，如果某个实际系统支持 LM，那么 $i\models_J\Rightarrow j\models i\Rightarrow_J\oplus k\models i\Rightarrow i\models_J\oplus k$。

即 RM 也是成立的。这使 RM 变成了一个可证明的定理，而非可选择的基本假设。

为了保持逻辑框架表达式的统一性，同时避免本文所讨论的弊端，可以在框架中增加对信息载体的定义，同时将相关性概念(Aboutness)和应用(Aboutness 在推理中的表示)区别对待。已有框架中，将信息载体规定为可以直接映射到实际系统中的实体的概念，这导致了一系列问题，因此，应该进一步划分框架中信息载体这一概念的粒度。假设信息可以用可数的基本概念集来表达，任何一个信息载体都包含了若干基本概念。在这个前提下，可将文档、查询等原信息载体均视为信息对象 O，定义信息载体为 O 的函数，此函数的计算结果为一个集合，其中包含了 O 中的基本概念。

语义定义 1(信息载体) 令 Γ 为基本概念集(可数集合)，P(Γ)为 Γ 的幂集，O 为信息对象，C 为 O 的集合。则信息载体为函数 $f:C\rightarrow P(\Gamma)$

在下文中出现的信息载体，均指计算结果，作为参数的信息对象是隐含的。

语义定义 2(Aboutness) 令 a，b 为信息载体，如果 $|a\cap b|>0$，那么 a 和 b 相关。

语义定义 3(⊨) 令 a，b 为信息载体，b 表示一定的信息需求，如果 a 和 b 相关，则记为 a⊨b。⊨不可交换。

此定义明确规定⊨是不可交换的。实际上，这个定义的语义含义是 a 部分或全部满足信息载体 b 所指定的信息需求。重新定义⊨之后，逻辑表达式中拥有此符号的子表达式的格式为一般信息载体⊨信息需求载体。

语义定义 4(信息包含，Information Containment) 令 a，b 为信息载体，若 $a\supseteq b$，则 a 包含 b，记为 $a\rightarrow b$。

这里使用的符号与定义 2 一致。

语义定义 5(语义判定函数，Semantics Judgment Function) 针对概念集合 Γ 的语义函数。形式为 $F:P(\Gamma)\times P(\Gamma)\rightarrow P(\Gamma)$。

它的语义定义是：如果信息载体 a，b 所承载的信息在语义上是不可调和的，则 Γ((a，b))的值为空，否则为(a，b)在语义上所对应的新的信息载体。注意这里参数 a 和 b 是有序的。

从语义角度来看，F((a，b))非空时必然包括 a 及 b 所承载的信息。

语义定义 6(互斥，Information Preclusion) 令 a，b 为信息载体，若 F((a，b))为空，则称 a 与 b 互斥，记为 a⊥b。

语义定义 7(信息合成，Information Composition) 令 a，b 为信息载体，若 F((a，b))非空，则信息合成的结果定义为 $a\oplus b=F((a,b))$。

语义定义 8(信息域，Information Field) 信息域是一个六元结构(Γ，⊨，→，⊕，⊥，0)，它满足以下性质：

1)Γ 是一个基本概念的可数集合，P(Γ)是它的幂集，F 为 Γ 的语义判定函数；

2)⊨不可交换；

3)(P(Γ)，→)是一个 poset(部分有序集)；

4)0∈P(Γ)，且任给 i∈P(Γ)，都有 0→i；

5)⊥定义在 P(Γ)×P(Γ)上；

6)任给 i，j∈P(Γ)：①i⊕j→i，i⊕j→j。②i⊕i→i。③i⊕j∈P(Γ)，或 j⊕i∈P(Γ)。

上述定义增加了不可交换的⊨和语义判定函数 F，以及修改了运算对象所在集合，其它性质同定义 5。通过语义定义 1 至 8 的重新定义，新的逻辑基础框架实现了信息载体的统一定义，避免了原框架的语义问题，而且保持了原框架的概念界面。建立在这些概念基础上的推理，会自动获得正确的语义支持。但是，统一定义的代价是引入了语义判定函数 F，这是逻辑框架之外的知识，无法形式化给出，因此，破坏了这个框架理论上的完备性。但这也恰恰反映了客观现实。人的语言符号反映了人的思想(语义)，特定的语言符号的组合所代表的语义或者存在于人的思想中，或者可以在一定的范围内由机器在它可读的知识库中搜索得到，因而都是不能形式化的。由于人工智能技术(尤其是自然语言理解)发展水平的制约，当前主流的信息检索系统一般采用统计方法，利用各种方式计算查询与文档的匹配程度，从而避免了使用语义判定函数。这种方法一方面取得了巨大的成功，但另一方面，缺乏语义支持也使信息检索系统无法真正理解用户的信息需求，甚至也无法理解语义环境更为丰富的文档，实际效果很难进一步提高。为了在一定程度上解决语义问题，信息检索领域的研究者们通常利用查询扩展和相关反馈技术，前者一般是利用外部的概念知识库，后者则利用查询对象本身得到一定的相关语义信息，这些都可以视为语义判定函数的简化。

结束语 目前使用的逻辑框架在形式化描述信息检索系统的检索特性方面取得了较大进展，但是缺乏对基础定义和运算的准确语义定义。本文在逻辑框架的语义方面进行了探讨，提出了基于语义的基础定义，在不影响现存逻辑的条件下，实现了基础定义的语义定义。在对逻辑框架的后继研究中，由于获得了一致的语义支持，研究成果将更逼近实际语义环境，对信息检索方面的研究与评估将更有指导意义。

目前的逻辑框架尚未完善，如上所述，此前的工作是实际系统的抽象和概括，是否可以完全反映各种实际系统的语义仍需要实践验证。本文后继的研究工作将继续从现有实际信息检索方法中抽象出事实和方法，使逻辑框架充分反映实际语义。

致谢 中国科学院计算技术研究所白硕研究员对本文的

维普资讯 http://www.cqvip.com

计算机科学 2004Vol. 31№. 3

网格环境下的数据库系统

汪锦岭 金蓓弘 李 京
(中国科学院软件研究所软件工程中心 北京 100080)

摘 要 目前网格上的各类应用系统几乎都是使用文件来保存数据,因此人们对如何将数据库系统集成入网格这个问题研究甚少。如果网格要在将来支持更大范围的应用系统,那么必须要解决将数据库系统集成入网格这个问题。本文在分析网格环境对数据库系统的需求的基础上,总结了将数据库系统集成入网格的两种方法,并分别对它们做出评价。

关键词 网格,数据库系统,需求,集成

The Database System in the Grid Environment

WANG Jin-Ling JIN Pei-Hong LI Jing
(Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences, Beijing 100080 China)

Abstract Until now, almost all of the Grid applications use files to store data, so little has been done on the integration of database systems into the Grid. If the Grid is to support a wide range of applications, we must solve the problem of integrating database systems into the Grid. This paper firstly examines the requirement of the Grid environment to database systems, and then summarizes two methods to integrate database systems into the Grid. Finally, we give a short evaluation on these two methods.

Keywords Grid, Database system, Requirement, Integration

1 引言

网格作为一种主要用于高级科学和工程领域的分布式计算结构,自20世纪90年代中期出现以后,便引起人们广泛的研究兴趣。迄今为止,网格上的各类应用系统几乎都是使用文件而非数据库系统来保存数据,因此,关于数据网格的研究主要集中在文件的处理和服务方面,如文件复制、文件传输、元数据管理等,而对如何将数据库系统集成入网格这个问题所做的研究则甚少。但数据库系统在数据的存储、组织、访问、权限控制等多方面都比文件系统占优势,如果网格要支持更大范围的应用系统(包括科学、工程、商业等各个领域),那么必须要解决将数据库系统集成入网格这个问题。

与文件系统不同,数据库系统提供了非常丰富的操作,而且不同数据库管理系统之间的差异也远比不同文件系统之间的差异大。因此,仅仅对现在的面向文件的数据网格服务加以扩充,是无法将数据库系统集成入网格的。有鉴于此,本文在分析网格环境对数据库系统的需求的基础上,提出了两种将数据库系统集成入网格的方法,并分别对它们做出评价。

2 网格环境对数据库系统的基本需求

网格是一种用来支持大规模信息共享的分布式计算结构。这种环境所具有的新特性,必然对数据库系统提出一系列新的需求。通过分析已有的数据网格项目,并综合 Paul Watson、Malcolm P Atkinson 等人的研究成果[1,2],我们认为

汪锦岭 博士研究生,研究方向为分布式计算。**金蓓弘** 博士,副研究员,研究方向为面向对象数据库、分布式计算。**李 京** 博士,研究员,博士生导师,研究方向为软件工程、分布式计算。

工作给予了悉心指导,程学旗博士对本文的完成提供了很多支持和帮助,在此一并表示感谢。

参 考 文 献

1 Voorhees E, Harman D. Overview of the Ninth Text REtrieval Conference (TREC-9). In: The Ninth Text REtrieval Conf. (TREC-9), 2001
2 van Rijsbergen C J. A new Theoretical Framework for Information Retrieval. In: Proc. of the Ninth Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1986. 194～200
3 Bruza P D, Huibers T W C. A study of aboutness in information retrieval. Artificial Intelligence Review, 1996, 10: 1～27
4 Wong K F, Song D W, Bruza P D, Cheng C H. Application of aboutness to functional benchmarking in information retrieval. In revision for ACM Transactions on Information Systems. 1998
5 Huibers T W C, Lalmas M, van Rijsbergen C J. Information retrieval and situation theory. SIGIR Forum, 1996, 30(1): 11～25
6 Song D W, Wong K F, Bruza P D, Cheng C H. Fundamental properties of the core matching functions for information retrieval. In: Proc. of the 13th Intl. Florida Artificial Intelligence Society Conf. (FLAIRS'2000), Orlando, Florida, USA, May 2000
7 Bruza P D, Huibers T W C. How nonmonotonic is aboutness?: [Technical Report UU-CS-1995-09]. Department of Computer Science, Utrecht University, The Netherlands, March 1995
8 Maron M E. On indexing, retrieval and the meaning of of about. Journal of the American Society for Information Science, 1977, 28: 38～43
9 van Rijsbergen C R. A non-classical logic for information retrieval. The Computer Journal, 1986, 29(6): 481～485
10 Song D W, Wong K F, Bruza P D, Cheng C H. Towards functional benchmarking of information retrieval models. In: Proc. of 12th Intl. Florida Artificial Intelligence Conf. 1999. 389～393