基于本体的信息检索系统提高检索结果相关性的研究

张 伟

(南京大学工程管理学院 南京 210093)

黄 奇

(南京大学国家信息资源管理南京研究基地 南京 210093)

【摘要】 主要介绍基于本体的信息检索技术的基本思想,并依据基本思想提出一个系统模型和一种用于提高检索结果和检索目标相关性的算法。

【关键词】 本体 信息检索 【分类号】 TP391

Research on Enhancing the Relativity of Outcome of Ontology - based IR

Zhang Wei

(*School of Management and Engineering, Nanjing University, Nanjing 210093, China*)

Huang Qi

(*National Center for Information Resource Management, Nanjing University, Nanjing 210093, China*)

【**Abstract**】 This paper introduces the basic idea of the Ontology - based IR, and puts forward a model of system according to the idea. An algorithm to enhance the relativity between the result and the aim is also given.

【**Keywords**】 Ontology Information Retrieval (IR)

当今,几乎所有搜索引擎的技术思想都是基于关键词匹配或者基于内容分类目录,这两种方法对信息的处理都还处在语法的层次上,还没有达到语义层次,所以检索结果的精度不高。由于本体对领域和任务进行了良好的描述,具有较好的概念层次结构和对逻辑推理的支持,从而在信息检索,特别是在基于知识的检索中得到了很好的应用[1]。在基于本体的信息检索系统中,通过对历史资料的学习和本体的推理机制,可以进一步提高检索结果和检索目标的相关性,从而使得检索结果更加符合人们的要求。

1 基于本体的检索实验系统研究

1.1 基于本体的检索系统的基本思想

(1)在领域专家的帮助下,按照一定的方法论建立相关领域以及顶级的本体[2]。

(2)收集信息源中的数据,并参照已建立的本体把收集来的数据按规定的格式存储在元数据库(关系数据库、知识库等)中[3]。通过本体对数据进行划分,确定文档所属的领域和概念。基于本体的分类法主要有字频统计法和句法分析。字频统计法首先查找本体中是否含有当前数据的高频词,有则说明这个数据属于该本体,通过这种方法可以过滤掉一些不相关的本体,然后在相关本体中进行精确查找。句法分析根据语法知识对句型结构进行分析,用本体代替词典得到的句法树能够有效消除句子的歧义,如"杨树打人"这个句子通过本体分析可以得到"杨树"是人名,是句子的主语,而不是树名[1]。

(3)查询转换器按照本体把从用户检索界面获取的查询请求转换成规定的格式[2],然后通过本体进行模糊判断,判断关键词属于哪几个领域,确定好相关概念和子概念,在本体的帮助下从元数据库中匹配出符合条件的数据集合。

(4)在本体的帮助下,可以很好地学习用户的偏好,通过用户的个性化偏好对结果进行过滤,从而进一步精练搜索结果。检索的结果经过定制处理后,返回给用户。

1.2 关于语义相关性的研究

在基于本体的信息检索系统中一个很重要的研究方向就是对语义相关性进行评估。Rensik 等人开发出了"is - a"的分类方法。该方法认为,在层次中衡量语义相关性的一个直接的方法就是测量两个节点之间的距离,一

收稿日期:2007 - 05 - 28
收修改稿日期:2007 - 07 - 18

个节点到另一个节点的路径越短,它们越相似,如果两点之间有多条路径,取最短者的长度。尽管对边的计量有时会不一致,但在概念上是非常直观的,只需对分类网中节点间的边进行计量,将结果作为它们的距离。通过对分类中概念间的关联概率的计算,可以避免对边距离计量的不可靠性[4]。

除了基于分类的相似性衡量,还有一些基于潜在语义分析的方法。Landauer 和 Dumais 提出的潜在的语义标引(LSI),是一种高维的线性相关模型,用来生成一种能够把握词汇和文本相似性的表述。他们认为,在相似的文本语境中的两个词语应该在语义空间中是最接近的。那么,任何对词汇间相关相似性的初始估计就能够通过单值分解的统计技术来分析,在此技术的应用中,词语和语境可表示高维空间中的点或向量[5]。

1.3 基于本体的信息检索系统所面临的一些问题

现在虽然基于本体的研究已经得到了快速发展,但是许多还只是试验性质的,即使本体一般都是在领域专家的帮助下通过人工构建的,因此,现阶段本体在信息检索中不能大范围应用,只有实现了本体的自动或半自动的构建,此问题才会得到解决。另外,现在领域本体得到快速发展,它集中于某一具体的领域,基于这种本体的检索系统都应用于在特定的领域进行信息检索,能够大范围应用的本体检索系统还很少,效率也很低下[6]。

笔者在参考基于本体的信息检索一般模型的基础上,吸收参考文献[7]的有关思想,提出了一个新模型,同时在参考了文献[8]所提出的算法的基础上提出了一种进一步提高检索相关性的算法。

2 基于本体的检索系统的模型

现有的一些实验性系统已在一定程度上证明了本体在检索系统中的作用,通过本体的引入使得在查准率和查全率上都得到了极大的提高,图 1 为笔者在基于本体的信息检索系统一般模型的基础上提出的一个新模型。

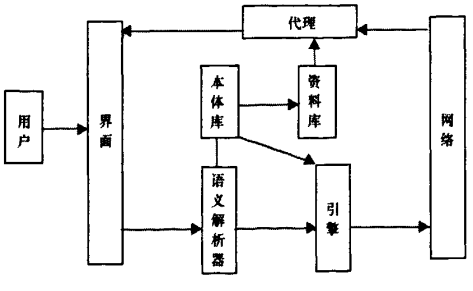

图 1 基于本体的搜索系统模型

系统的运行过程如下:

(1)用户向系统发出检索命令,语义解析器通过对用户检索命令的分析提取出它的关键词,并把关键词分别传送给引擎和本体库。在本体库中,通过本体的推理机制找出该关键词属于哪个或哪几个领域,同时找出和这个关键词具有一定程度相关性的相关概念或子概念,并分别传到引擎和资料库。

(2)引擎在接到语义解析器和本体库传来的关键词及相关概念之后,到网络上搜索和这些概念相关的网页,当然这个搜索也是基于语义的。所有的网页资源都必须进行预处理,首先,划分它们各自所属的领域,提取出内容的核心概念并标引,在基于本体的信息检索过程中,查询表达与信息资源之间的匹配过程仿佛一种"探索"过程,这一过程能依靠查询的表达形式和逻辑理解以不同的方式实现[1]。

(3)将搜索到的网页送到代理中,代理利用资料库送来的相关历史数据计算出所检索到的网页和检索目标的相关系数,将属于一定相关系数范围之内的网页按照相关系数的大小排列,从而进一步缩小结果的数量并提高检索结果的准确性。资料库的信息可理解为:资料库中保存着以前不同用户的搜索主题以及与之相对应的点击资料。笔者认为,当某一用户在对某一主题进行搜索时,他所点击的那些页面应该会是他所感兴趣的页面,这些页面与他所检索的主题的相关性会比那些他没有点击的页面要大很多,而且点击的数量也显示出了这个页面和检索主题相关性的强弱。那么资料库在接受到本体库送来的相关概念之后,就会通过对历史资料的查找,得出以前人们在检索这些主题时对哪些页面感兴趣,再将这些信息传送给代理,代理会在上一步所得到的初步结果的基础上,根据资料库所送来的信息并通过下文中提出的算法对初步结果进行第二次处理,改进向用户所提供的结果。

3 提高检索结果相关性的算法

笔者假设用户的检索主题的关键词为 Ai,同时本体库中推导出了和这个关键词具有一定相关性的几个关键词,假设为 4 个,设为:A_2, A_3, A_4, A_5。利用本体相似度计算可以计算出这几个关键词和 A_1 之间的相似度。为了简化,可以假设 A_2, A_3, A_4, A_5 和 A_1 之间的相似度分别为:0.9,0.8,0.7,0.6。这时可以将这几个关键词组成一个向量:$[A_1\ A_2\ A_3\ A_4\ A_5]$。为了便于统计,可以假设将每一向量限定为二值函数{0,1}。对于检索目标 M 而言,如果以前某一用户在检索某一个关键词时点击了这个目标,那么这个向量上的取值就为 1 ,如果没有点击,那么这个取值就为 0。同理,对于检索结果 $N_i(i=1,2,\cdots)$ 而言,在某一主题之下曾经被用户点击过的取 1,没有打开过的就取 0。那么,可以将检索目标以及每一个检索结果在各分向量上的值构成一个矩阵,这个矩阵称之为点击

矩阵。假设在检索中存在如下的点击矩阵：

	A_1	A_2	A_3	A_4	A_5
M	1	0	1	1	1
N_1	1	1	0	0	0
N_2	0	0	1	1	0
N_3	0	0	0	1	1
N_4	1	0	1	1	0

进一步，可以把页面的点击量这个因素考虑进来，从而可以构建第二个矩阵，这个矩阵称之为分布矩阵。在所有的页面中，把页面按照点击量的大小排列，设点击量最大的那个页面在相对应主题下的值为 1，某一点击量以及小于这个数量的页面在相对应的主题下的值为 0。比如：M 在 A_1, A_2, A_3, A_4, A_5 主题下的点击量分别为：1000，0，500，350，600，N_1 在 A_1, A_2, A_3, A_4, A_5 主题下的点击量分别为：700，120，0，0，0。最大的为 1000，相对应的 M 在 A_1 处的值为 1，设点击量小于 200 的值为 0。那么从 1000 到 200 之间的点击量就可以按照一个泊松分布来计算它们各自的权重。比如：M 在 A_2, A_3, A_4, A_5 向量下的值为：0，0.3，0.2，0.5。N_1 在 A_1, A_2, A_3, A_4, A_5 向量下的值为：0.6，0，0，0，0。根据这两组值可以得出相应的矩阵：

	A_1	A_2	A_3	A_4	A_5
M	1	0	0.3	0.2	0.5
N_1	0.6	0	0	0	0

那么，可假设在整个检索结果中，存在如下的分布矩阵：

	A_1	A_2	A_3	A_4	A_5
M	1	0	0.3	0.2	0.5
N_1	0.6	0	0	0	0
N_2	0	0	0.8	0.7	0
N_3	0	0	0	0.4	0
N_4	0.9	0	0.3	0	0

从点击矩阵中可以得到检索结果与检索目标相匹配的向量个数 $l(l \leqslant m, l \leqslant n)$，其中 m 为 M 中取值为 1 的个数，n 为 N 中取值为 1 的个数。设点击矩阵为 A，分布矩阵为 B，关键词之间的相关系数组成的向量为 P。通过如下公式可以计算出 $N_i(i = 1, 2, 3, \cdots)$ 和 M 之间的相关性：

$$R = \frac{1}{2}\left(\frac{\sum B_{ij} * P_j}{m} + \frac{\sum B_{i'j} * P_j}{n}\right) \qquad (1)$$

其中，B_{ij} 和 $B_{i'j}$ 为矩阵 B 中相关的值。可以得出：在矩阵 A 中，能够清楚地看出 M 和 $N_i(i = 1, 2, 3\cdots)$ 之间相匹配的向量，那么矩阵 B 中相对应位置的值就是上面的 B_{ij} 和 $B_{i'j}$。为了进一步解释计算过程，下面以计算 M 和 N_1 之间的相关性为例具体说明。

(1)从矩阵 A 中看出，M 和 N_1 相匹配的值为：A_{11} 与 A_{21}，其他的值都不相匹配。那么相对应的 B_{11} 和 B_{21} 就是所谓的 B_{ij} 和 $B_{i'j}$，同时还可以得出 m = 4，n = 2。

(2)找出相对应的 P_j，在这个例子中，M 和 N_1 相匹配的值只有在 A_1 列，那么 P_j 只有一个，即 $P_1 = 1$。

(3)将上面的各值代入计算公式，可以计算出 M 和 N_1 之间的相关系数：

$$R(M, N_1) = \frac{1}{2}\left(\frac{1 * 1}{4} + \frac{1 * 0.6}{2}\right) = 0.275$$

同上，可以计算出其他的相关系数：

$$R(M, N_2) = \frac{1}{2}\left(\frac{0.8 * 0.3 + 0.7 * 0.2}{4} + \frac{0.8 * 0.8 + 0.7 * 0.7}{2}\right) = 0.3725$$

$$R(M, N_3) = \frac{1}{2}\left(\frac{0.2 * 0.7 + 0.5 * 0.6}{4} + \frac{0.7 * 0.7}{2}\right) = 0.1775$$

$$R(M, N_4) = \frac{1}{2}\left(\frac{1 * 1 + 0.3 * 0.8 + 0.2 * 0.7}{4} + \frac{0.9 * 1 + 0.3 * 0.8}{3}\right) = 0.3625$$

按照上述计算结果，可以得出 $R(M, N_2) > R(M, N_4) > R(M, N_1) > R(M, N_3)$，为了使得结果更加符合用户的要求，可以将网页按上面相关系数的大小排列，同时还可以按照用户的喜好，设定一个门槛值，比如说 0.2，将一些相关性较小的页面排除，那么在上面的几个页面中，N_3 由于相关系数只有 0.1775，小于 0.2，因此将会排除在结果之外。

从公式(1)中可以看出，当 $1 = m = n, P_j = 1, B_{ij} = B_{i'j} = 1$ 时，这时是最理想的情况，即检索结果和检索目标的相关性为 1 。

上面的算法只适合于很理想的情况，可以在如下几个方面对它进行进一步优化：

(1) 在上面的计算公式中，前面的系数取值为 $\frac{1}{2}$，这是一种很粗糙的处理，使得只有在最理想的情况下，相关系数才会是 1。可以将这个系数在使用过程中通过学习进一步优化。

(2) 对于门槛值，可以通过对历史资料的学习，来了解用户对某一主题的喜好，从而设定一个变化的门槛值，随主题的改变、用户的使用情况而改变。

4 结 语

信息检索从基于叙词表发展到基于本体是知识工程领域的一大进步。从现在已经创建成功的几个系统来看，虽然还有各种各样的问题，但是已经极大地提高了信息检索的效率。但是正如前文所说，现在基于本体的信息检索的应用还只是处于试验阶段，能够实际应用的信息检索系统还不多，这是因为本体的构建现在绝大部分还是由人工来完成，还没有能够大规模应用的通用本体，这些都限制了系统的实际应用，因此，要实现能够实际应用的基于本体的信息检索系统将还有许多问题有待解决。

参考文献：

[1] 吴丹. 本体在信息检索中的作用及实例研究[J]. 情报杂志，2006,25(6):72 - 75.

[2] 杨建林. 基于本体的文本信息检索研究[J]. 情报理论与实践，2006,29(5):598 - 601.

[3] 苏晓路，钱平，赵庆龄，等. 农业科技信息导航知识库及其智能检索系统的构建[J]. 情报学报，2004,23(6):677 - 682.

[4] Resnik P. Semantic Similarity in Taxonomy: An Information - Based Measure and Its Application to Problems of Ambiguity in Natural Language[J]. Journal of Atificial Intelligence Research, 1999(11):95 - 130.

[5] Landauer T K, Dumais S T. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge[J]. Psychological Review, 1997, 104(2): 211 - 240.

[6] 曹树金，马利霞. 论本体与本体语言及其在信息检索领域的应用[J]. 情报理论与实践，2004,27(6):632 - 637.

[7] 郭祥文，刘惟一，钱民，等. 基于本体论的信息检索[J]. 云南大学学报(自然科学版)，2003,25(4):324 - 327.

[8] 陆宝益，李保珍. 基于本体的检索质量的语义相关度评价[J]. 情报杂志，2006,25(10):63 - 65.

(作者 E - mail: zwfnc@yahco.com.cn)

动 态

大英图书馆保管中心开放

大英图书馆保管中心拥有完备的世界级图书保管设施，同时也是国家音频档案馆的技术保管基地，保管中心为大英图书馆宝贵的馆藏提供了很好的保护。保管中心占地2 600 平方米，这一耗资 13.25 万英镑的项目于2005 年 8 月开始动工，2007 年 1 月 17 日完成，2007 年 5 月 17 日起对公众开放。

大英图书馆的历史上，保管中心首次将保管图书的人员和设施集中起来，而以前这些人员和设施分散在伦敦多个地方，并分别专责保管特定类型的收藏品。同时保管中心还为音频档案馆的技术操作提供用于档案标准的保存复印和专业的重新灌录等便利设备。图书馆现在能够给保管专员提供大量的培训机会，并且让公众参观工作室，给他们提供示范和讲座。

大英图书馆收藏中心主任 Helen Shenton 说："在预定时间和预算内能够完成建造和使用的大英图书馆保管中心是一项伟大的工程。这标志着一个图书馆保管的新时代。"

(编译自：Saved for the Nation: New British Library Centre for Conservation. [2007 - 05 - 17]. http://www.bl.uk/news/2007/pressrelease20070517.html.)

(本刊讯)