文本信息检索技术
Text Information Retrieval Technology

李广原 陈 丹
Li Guangyuan Chen Dan

(广西师范学院计算中心 南宁 530001)
(Computer Centre, Guangxi Teacher College, Nanning, 530001)

摘要 论述 3 种常用的文本信息检索技术，即布尔检索、向量空间检索和概率检索，对它们的优缺点进行评价，并对文本信息检索技术进行了展望。

关键词 信息检索 文本信息 检索技术
中图法分类号 TP 391.3

Abstract Text Information Retrieval is a common information retrieval. The Bullean retrieval, vector space retrieval and probabilistic retrieval are introduced and evaluated on their merits and demerits. A prospect for the Text Information Retrieval is discussed.

Key words information retrieval, text information, retrieval technology

面对信息社会浩如烟海的信息数据，如何快速有效地找到需要的信息，是一个十分重要的课题。在信息检索中，最常见的一类检索是文本信息检索，人们对它研究最早，成果也最为显著。在实践中，研制了许多成功的信息检索系统，如美国 Massachusetts 大学的 INQRER 系统；Cornell 大学的 SMART 系统，中国北大方正的 NARS 系统等。文本信息检索采用的技术主要有布尔检索、向量空间检索和概率检索。目前，人们对文本信息检索技术的研究正在深入，新的检索技术不断出现。

1 文本信息检索的基本原理

信息检索是人们借助某种检索工具，运用某种特定的检索策略从待检的信息源中查找出自己需要的信息。在人们日常生活、工作、学习所获取的信息，文本信息占据很大的比例，它主要以文字、或辅以图片呈现在人们面前。信息检索是一种不确定性检索，用户在检索信息时，并不知道信息源里是否有符合需要的东西，检索出来的信息并不一定完全符合用户的需要。信息检索过程是信息源中的信息和用户需求相互之间匹配的过程，信息源就是某个信息检索系统，我们用下列一个四元式表示：

2000-12-12 收稿，2001-04-06 修回。

$S=(D,Q,T,f)$,

其中 $D=(D_1,D_2,D_3,\cdots,D_m)$ 为标引词集合,用于对文本进行标识;

$Q=(Q_1,Q_2,Q_3,\cdots,Q_l)$ 为用户查询集;

$f=Q\times D\rightarrow R$ R 为某一值集,通常 $R\in[0,1]$.

2 文本信息检索技术

2.1 布尔检索

布尔检索就是采用布尔表达式来表示用户提问,通过对文本标识与用户给出的检索式进行逻辑比较来检索文档。用户表达式是把用户给出的检索词用布尔运算符“∧”(and),“∨”(or)连结起来的式子。

设文本集 $D=(d_1,d_2,d_3,\cdots,d_n)$,$d_i(i=1,2,\cdots,n)$ 为文本集中某一文档;又设 $T_i=(t_1,t_2,\cdots,t_m)$ 为 d_i 的标引词集,则对于形如 $Q=W_1\wedge W_2\wedge\cdots\wedge W_l$ 的检索式,如果 $W_1\in T_i$,$W_2\in T_i,\cdots,W_l\in T_i$,则 d_i 为检索到的文本,我们称 d_i 为命中文档,否则 d_i 为不命中文档;而对于形如 $Q=W_1\vee W_2\vee\cdots\vee W_l$ 的检索式,如果至少存在某个 $W_k\in T_i(k=1,2,\cdots,l)$,则 d_i 为命中文档,如果不存在任何一个 $W_k\in T_i(k=1,2,\cdots l)$,则 d_i 为不命中文档。

实现布尔检索,首先要对文本集中每个文档进行标识,标引词可以采用关键字、自由词、作者、篇名等能反映文档特征的词,其次,要对文档进行合理的组织,建立文档的索引,通常把文档组织成倒排文档结构,就是把与某标引词有关的所有文档的号数通过索引集中在一起,当通过该标引词查找文档时,可以立即找到文档所在的位置,从而检索到文档。布尔检索具有简单,易理解,容易在计算机上实现且检索速度快等优点,故在许多检索系统中得到应用,例如 Yahoo! Inforseek 等诸多网络检索站点均采用布尔检索技术。目前,几乎所有的商用检索系统都采用布尔检索作为主要的检索方式或提供布尔检索,虽然布尔检索有着许多优点,但它的缺陷是明显的:(1) 布尔逻辑式的构造不易全面反映用户的需求;(2) 匹配标准存在某些不合理的地方,例如,在响应某个用“∧”连接的检索式时,系统把只含有检索式中的一个或数个但非全部检索词的文档看作与那些不含有检索式中的任一检索词的文档一样无用,同样加以排除;(3) 检索结果不能按照用户定义的重要性排序输出。系统检索输出的文档中,排在第一位的文档不一定是文本集中最适合用户需要的文档,用户只能按照检索结果的顺序浏览才能知道文档中那些更适合自己的需要。

为了克服上述缺陷,人们对布尔检索理论进行了改造,一种方法是对标引词引进权值,权值的大小即反映标引词在文档中的重要程度,由此,形成了所谓的加权布尔检索,或称扩展布尔检索,如 Bookstein 检索模型,Salton 模型等[1]。

2.2 向量空间检索

在向量空间检索中,把文档和用户查询均用一组相互独立的词条组成,设在文本集 D 中,共使用了 n 个词条 $t_1,t_2,\cdots,t_n$,文本集 D 中某一文档 D_i 可表示为:

$D_i=(w_{i1},w_{i2},\cdots,w_{in})$,

其中 $w_{i1},w_{i2},\cdots,w_{in}$,分别为词 $t_1,t_2,\cdots,t_n$ 在文档 D_i 中的权值。权值越大,表示该词在文档中的份量越大,即该词越能反映 D_i 的内容;权值越小,该词的份量越小,越不反映 D_i 的内容。权值的取值范围是[0,1]。

同样地,用户的查询可表示为:

$Q_j = (w_{j1}, w_{j2}, \cdots, w_{jn})$，

其中 $w_{j1}, w_{j2}, \cdots, w_{jn}$ 分别为给出的 $t_1, t_2, \cdots, t_n$ 的权值。把几个词看作为 n 维坐标系中的坐标，权植的对应的坐标值。这样，文档和用户查询均可看成是由这坐标轴组成空间中的一个点，或称为一个矢量。文档和用户之间的比较，用相似度大小来表示，计算相似度有多种方法，一般常用下式计算：

$$S(D_i,Q_j) = \frac{\sum_{k=1}^{n} W_{ik}W_{jk}}{\sqrt{\sum_{k=1}^{n}(W_{ik})^2 \sum_{k=1}^{n}(W_{jk})^2}}.$$

这种计算方法实质上就是计算 n 维空间中，文本向量和提问向量之间夹角的余弦。式中涉及词的权值，词的权值设计是基于这样一个假说，即词的权值与在文档中出现的频率成正比，与在文本集中出现该词的文档频数成反比。一般地，可把权值设计成如下式：

$$W_{ik} = tf_{ik}\log(\frac{N}{n_k} + 0.5),$$

其中，tf_{ik} 表示词 t_k 在文档 D_i 中出现的频数，N 表示文本集中文档的总数，n_k 表示词 t_k 的文档频数。

向量空间检索具有如下优点：(1) 为标引词引进权值，通过调节标引词对应权值的大小来反映标引词与被标引文档的相关程度，它部分地克服了传统布尔检索的缺陷；(2) 检索通过计算文档之间的相似度，使属性相似的文档尽量聚拢在一起，以提高检索效率；(3) 满足用户需求多样化以及检索手段多样化的需要。用户可以根据需求特点选择一组可供使用的检索手段。

向量空间检索存在的缺点：(1) 相似度计算量大，影响检索速度；(2) 标引词的权值较难确定；(3) 对标引词的相互独立的假设不符合实际情况[2]。对第 (3) 点所指出的缺陷，人们又研究基于词的相依性的向量空间检索，例如 S. K. M 旺格等人在 1985 年提出用一组经过挑选的正交基向量来表示词向量，词间关系可直接由其向量表示，给出较为精确的计算，这种模型称为广义向量空间模型。李广原[4]对相似度计算进行研究，提出一个新的计算方法，该计算方法的实验效果跟传统方法一致。

2.3　概率检索

概率检索考虑词与词的相关性，把文本集中的文档分为相关文档和无关文档。以数学理论中的概率论为原理，通过赋予标引词某种概率值来表示这些词在相关文档和无关文档之间出现的概率，然后计算某一给定文档与查询式相关的概率，系统据此概率作出检索决策。

概率检索有多种形式，常见的一种称之为第二概率检索模型，其基本思想是，标引词的概率值一般是对检索作业重复若干次，每重复一次，用户就对检出文档进行相关性判断。然后利用这种反馈信息，根据每个词在相关文档集合和无关文档集合的分布情况来计算它们的相关概率，在该模型中，词的权值设计为：

$$\log \frac{P(1-P)}{P'(1-P')},$$

式中 P, P' 分别表示某词在相关文档集和无关文档集中出现的概率。某一文档的权值(决定它在排序中的位置) 则是它所含标引词权值之和，于是，文档与用户查询相关概率可定义为：

$$S(D,Q) = \sum_{i=1}^{n} \log \frac{P_i(1-p_i)}{p_i'(1-p_i')}.$$

概率检索的优点是：(1) 采用严格的数学理论为依据，为人们提供了一种数学理论基础来进行检索决策；(2) 采用相关反馈原理，可开发出理论上更为坚实的方法。它的主要缺点是：(1) 增加存贮和计算资源的开销；(2) 参数估计难度较大。

近年来，人们提出了一种新型的概率检索模型——概率推理网络[5]，概率推理网络由文本网络与查询网络构成，其中文本网络由文本节点 D_i，文本表达节点 T_i 和文档概念节点 R_i 组成，文本节点对应于抽象文本，文本表达节点对应于某一文本，而文本概念节点则对应于文本的特征表示，查询网络由查询节点 Q 和查询概念节点组成，查询节点 Q 表示某一用户查询，它是对查询概念节点的相关性描述，而查询概念节点包含了查询概念对查询概念节点概率相关性描述。检索过程是给定文本节点的先验概率和中间节点的条件概率就可计算出查询节点的后验概率。概率推理网络在概率论相关理论的基础上进行推理，具有较坚实的理论基础，不过，要事先给出文本节点的先验概率，这是不容易的。

3 结语

上述三种传统的文本信息检索技术随着文本信息检索技术日趋成熟，已得到进一步的完善。近年来，新的检索技术也不断涌现，出现诸如并行信息检索系统、演绎信息检索系统、基于超文本技术的信息检索系统、分布式检索系统、智能信息检索系统等[6,7]。虽然它们还有待于进一步的发展和完善，但却代表着文本信息检索技术的发展方向。并行信息检索能缩短用户的检索时间，提高检索效率；演绎信息检索系统、基于超文本信息检索系统、智能检索系统的查询界面友好，用户操作更方便。

信息检索需要多学科提供技术支持，它涉及到计算机科学、心理学、认知科学、人机工程学等多领域的研究。充分吸取其它学科的研究成果，开发有效的并行查询语言，研究分布式检索算法以及自然语言的理解，是当前的研究热点。

参考文献

1 康耀红. 现代情报检索理论. 北京：科学技术文献出版社，1990. 3.

2 赖茂生等编著. 计算机情报检索. 北京：北京大学出版社，1993. 3.

3 GERARD SALTON. Developments in automatic text retrieval. Cornell University，1991.

4 李广原. 属性论在文本相假度计算中的应用. 广西师院学报（自然科学版），2000，(3).

5 潘谦红. 文本信息检索模型. 中国计算机报，1998，19.

6 贾同兴. 人工智能与情报检索. 北京：北京图书馆出版社，1997. 7.

7 俞字琴. 90年代我国情报检索理论研究述评. http：//www. north. cetin. net. cn/ hgingbao/969.

（责任编辑：邓大玉）