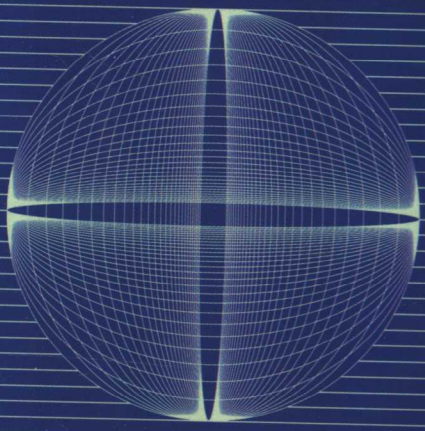
普华经管

正略钧策 ADFAITH

普华经管学术文丛

信息检索与信息融合

李培 编著



Information Retrieval and Information Fusion

人民邮电出版社
POSTS & TELECOM PRESS

普华经管学术文丛

国家社会科学基金项目资助(项目编号:05BTQ026)

信息检索与信息融合

李 培 编著

人民邮电出版社
北 京

图书在版编目（CIP）数据

信息检索与信息融合 / 李培编著. —北京：人民邮电出版社，2007.12

（普华经管学术文丛）

ISBN 978-7-115-17379-9

Ⅰ. 信… Ⅱ. 李… Ⅲ. ①情报检索②信息处理 Ⅳ. G252.7 G202

中国版本图书馆 CIP 数据核字（2007）第 196273 号

内 容 简 介

本书以国内外理论研究和技术开发的最新成果为基础，以信息检索与信息融合两大信息管理手段的紧密结合为主线，借鉴已有研究先进之处的同时，在多个方面实现了创新。全书在介绍信息检索和信息融合基本原理方法的基础上，对国外的相关研究进行了梳理，在信息检索的查询处理、分布式多代理检索、分布式检索中的资源选择、检索结果排序、自动标引、多媒体检索等领域对信息融合的应用进行了系统深入的研究。

本书的研究突出理论的严谨性、成果的先进性、方法的新颖性，可供信息管理、计算机科学、情报学等领域的研究者进行相关研究时参考借鉴，并可作为信息管理、信息系统、情报学等专业的研究生、本科生的教材或教学参考书。

普华经管学术文丛

信息检索与信息融合

◆ 编　　著　李　培

责任编辑　许文瑛

◆ 人民邮电出版社出版发行　　北京市崇文区夕照寺街 14 号

邮编 100061　　电子函件 315@ptpress.com.cn

网址 http://www.ptpress.com.cn

北京铭成印刷有限公司印刷

新华书店总店北京发行所经销

◆开本：700×1000　1/16

印张：15　　2007 年 12 月第 1 版

字数：165 千字　　2007 年 12 月北京第 1 次印刷

ISBN 978-7-115-17379-9/F

定　价：35.00 元

读者服务热线：（010）67129879　　印装质量热线：（010）67129223

反盗版热线：（010）67171154

“普华经管学术文丛”出版说明

在市场竞争日益激烈、科学技术迅猛发展、全球化浪潮日益迫近的今天，经济学、管理学的学术成果正在不断转化为生产力，对我国社会主义市场经济的发展以及国内各种规模、各种类型的企业的管理产生了深远的影响。

20世纪90年代以来，中国的出版界引进了大量国外最新的经济管理理论和研究成果，中国的学术界和产业界因此得以大开眼界。尤其是一大批的经济理论、管理工具纷纷涌入我国，如全面质量管理、流程再造、平衡计分卡（六西格玛）等管理手段已被很多国内企业在实践中加以应用，这些思想和工具极大地提高了我国企业的管理水平。

可喜的是，近十几年来我国的学者在经济管理领域不断探讨，不断实践，一大批学术成果涌现出来。这些成果不仅吸收了国外最新的理论和思想，而且很好地与中国国情相结合，较之国外的理论，它们更符合我国企业的管理特点与要求，针对性和指导性也更强。但多少令人遗憾的是，由于种种原因，国内学者的很多研究成果难以找到好的出版平台，在出版、发表方面存在着诸多困难，从而使这些研究成果的效益大打折扣。

“普华经管学术文丛”其宗旨就是为学者搭平台，为读者出好书，希望通过我们的努力推动我国经济社会协调发展，健全和创新我国企业的管理体制及管理方法，最终实现和谐社会的美好愿望。

目 录

第 1 章 信息检索模型与算法

不同信息检索设施（或系统）获取信息的方式与途径各异，但是，它们的基本原理是相同的：即检索系统对用户信息需求与系统存储的信息资源所进行的匹配。为了进一步严密地表述和论证这一原理，需要建立相应的信息检索模型。所谓检索模型就是对信息检索任务的数学抽象，它避开了对具体实现细节如数据存储、数据结构等的描述，而主要从两个方面抽象地研究信息检索方法：一是确定在模型中如何表示构成检索系统的两个要素，即文档和检索条件；二是确定在模型中如何定义和计算文档和检索条件之间的关系。典型的信息检索模型包括布尔模型、向量空间模型、概率模型、模糊模型等；而其中的一些模型在应用到 Web 信息环境时，结合 Web 资源的特点，又形成了独特的 Web 检索算法。

1.1 布尔检索模型

布尔检索模型是最早也是最简单的一种检索模型，其理论已基本成熟，许多检索系统都以这种检索模型作为工作原理。

1.1.1 布尔模型的基本原理

布尔模型在解释信息检索处理过程时，主要遵循以下两条基本规则。

（1）系统标引词集合中的每一个标引词在一篇文档中只有两种状态：出现

或者不出现。

（2）检索提问式 q 由三种布尔逻辑运算符“and”、“or”、“not”连接检索词来构成。

根据布尔逻辑的运算规定，提问式 q 可以被表示成由合取子项（Conjunctive Components）组成的析取范式（Disjunctive Normal Form，缩写为 dnf 或 DNF）形式。例如，传统的布尔提问式

$q = k_1$ and （$k_2$or not k_3）

可以写成如下等价的析取范式形式：

$\bar{q}_{dnf}$ = （k_1 and k_2 and k_3） or （k_1 and k_2 and not k_3）

or （$k_1$and not $k_2$and not k_3）

这里，$\bar{q}_{dnf}$为提问式 q 的主析取范式。

布尔模型中的相关参数定义如下：

t：系统中标引词的数量。

k_i：第 i 个标引词，$K = \{k_1, \ldots, k_t\}$ 代表所有标引词的集合。

w_{ij}：文档 d_j中标引词 k_i的权重；对于经典布尔模型，每个标引词的权值 $w_{ij} \in \{0, 1\}$。

$\bar{d}_j = (w_{1,j}, w_{2,j}, \cdots, w_{t,j})$：文档 d_j的标引词向量。

g_i：函数，定义为 $g_i(d_j) = w_{ij}$，它返回标引词 k_i在向量 $\bar{d}_j$ 中的权重。

$\bar{q}_{cc}$：$\bar{q}_{dnf}$的析取成分。

基于上述规则与假定，布尔模型对于任一篇文档 $d_j \in D$，定义 d_j与用户提问 q 的匹配函数为

$$sim(d_j, q) = \begin{cases} 1 & if\ \exists \bar{q}_{cc} \mid (\bar{q}_{cc} \in \bar{q}_{dnf}) \cap (\forall k_i, g_i(\bar{d}_j) = g_i(\bar{q}_{cc})) \\ 0 & otherwise \end{cases} \tag{1-1}$$

若 $sim(d_j, q) = 1$，则文档 d_1与提问 q 相关；否则，不相关。

假设文档集合 D 中存在两篇文档 d_1和 d_2，其中，d_1含有标引词 k_1和 k_2，d_2含有标引词 k_1和 k_3，则它们的文档向量分别为

$$\bar{d}_1 = (1, 1, 0)$$

$$\bar{d}_2 = (1, 0, 1)$$

根据匹配函数 $sim(d_j, q)$ 的定义，我们可以发现，文档 d_1 与提问式 $q = k_1$ and $(k_2$or not $k_3)$ 的匹配函数值为 1，即文档 d_1 与提问 q 是相关的；而文档 d_2 与提问 q 的匹配函数值为 0，表明文档 d_2 与提问 q 是不相关的。

1.1.2 布尔模型的分析

传统的布尔检索是将用户查询与文献进行逻辑的（而非数值的）比较而获得结果的检索。布尔检索模型的突出优点在于这种结构化的提问方式与用户的思维习惯相一致。同时，这种模型把复杂的检索过程简单化，能够将较复杂的情报提问按其概念组面的逻辑关系描述出来，从而变成可以由计算机执行的逻辑运算，变成机器根据事先确定的程序进行自动匹配的过程，这种运算上的简单易行是布尔检索系统的又一突出特征。此外，用布尔检索进行操作的某些系统允许用户通过给他使用的一个有结构的词典来缩小或扩大检索。所谓有结构的词典是指对任何一个给定的标引词都存储了与之相关的更一般的（上位）或更精确的（下位）关键词的词典。布尔检索很容易利用这些相关项来改进检索。

布尔检索在理论上存在的一些缺陷也是不容忽略的，具体包括下列几个方面。

（1）布尔逻辑式的构造不易全面准确反映用户的需求。

（2）匹配标准存在不合理的地方，严格的匹配可能导致检出的文档过多或过少，难以控制结果输出量的大小。

（3）对检索结果平等对待，不能按照用户定义的重要性排序输出。

（4）对用户的检索技能有较高的要求。

1.2 向量空间模型

20 世纪 70 年代中期，Salton 提出了检索系统的向量模型（Vector Space Model，VSM），定义了文档向量、提问向量、文档提问相关系数以及属性文档

相关矩阵、标引词相关矩阵与文档相关矩阵。把文档和查询用向量来表示，这是建立向量空间模型的基本前提。

1.2.1 向量空间模型的基本原理

1. 文档向量的构造

对于任一文档 $d_j \in D$，我们可以把它表示为如下 t 维向量的形式：

$$\overline{d_j} = (w_{1j}, w_{2j}, \cdots, w_{ij})$$

其中，向量分量 w_{ij} 代表第 i 个标引词 k_i 在文档 d_j 中所具有的权重，t 为系统中标引词的总数。在布尔模型中，w_{ij} 的取值范围是 $\{0, 1\}$；在向量空间模型中，由于采用“部分匹配”策略，w_{ij} 的取值范围是一个连续的实数区间$[0, 1]$。

在检索的前处理中，一篇文档中会标引出多个不同的标引词，而这些标引词对表达该篇文档主题的能力往往是不同的。也就是说，每个标引词应该具有不同的权值。如何计算文档向量中每个标引词的权值，不仅关系到文档向量的形成，也关系到后续的检索匹配结果。

标引词权重的大小主要依赖其在不同环境中的出现频率统计信息，相应的权重就分成局部权重和全局权重。

局部权重（Local Weight）l_{ij} 是按第 i 个标引词在第 j 篇文档中的出现频率计算的权重。它以提高查全率为目的，对在文档中频繁出现的标引项给予较大的权重。

全局权重（Global Weight）g_i 则是按第 i 个标引词在整个系统文档集合中的分布确定的权重。它以提高查准率为目的，对在许多文档中都出现的标引项给予较低的权重，而对仅在特定文档中出现频次较高的标引项给予较大的权重。计算全局权重的典型方法就是逆文档频率 IDF（Inverse Document Frequency）加权法。

$$g_i = \log (N / n_i)$$

其中，N 为系统文档总数，n_i 为系统中含有标引词 k_i 的文档数。

目前，标引词权值计算方案已有很多种，比较常用的是“TF - IDF（Term

Frequency—Inverse Document Frequency)”方法，即词频—逆文档频率加权法。

设$freq_{ij}$为标引词k_i在文档d_j中的出现次数；idf_i表示标引词k_i的逆文档频率；maxtf_j表示文档d_j中所有标引词出现次数的最大值。那么，对于文档d_j中标引词k_i的权重w_{ij}可按如下方法计算：

（1）计算局部权重 $tf_{ij} = freq_{ij} / \max tf_j$

（2）计算全局权重 $idf_i = \log (N / n_i)$

（3）计算标引词权重 $w_{ij} = tf_{ij} * idf_i$

2. 提问向量的构造

在向量空间模型中，用户的信息需求被转换为提问向量，并用与文档向量类似的表示形式表示，即

$$\bar{q} = (w_{1q}, w_{2q}, \cdots, w_{tq})$$

这里，t为系统中标引词的总数，向量分量w_{iq}表示第i个标引词k_i在提问q中的权值，且有$w_{iq} \geqslant 0$。对于查询语词的权值，Salton 和 Buckley 认为可以采用如下的方法：

$$w_{iq} = \left(0.5 + \frac{0.5 freq_{iq}}{\max tf_q}\right) \times \log \frac{N}{n_i} \tag{1-2}$$

式（1-2）中，$freq_{iq}$为标引词k_i在表述用户信息需求的文本内容中所出现的次数，而maxtf_q则是在表述用户信息需求的文本内容中所使用的所有标引词出现次数的最大值。

3. 文档与提问向量相似度的计算

在文档与提问向量化表示的基础之上，文档与查询提问之间的相关程度（即相似度）就可以由它们各自向量在t维空间的相对位置来决定。

向量间相似程度的度量方法有很多种，主要有内积法（Inner Product）、Dice 法（Dice Coefficient）、Jaccard 法（Jaccard Coefficient）和余弦法（Cosine Coefficient）。

较常用的度量方法是提问向量和文档向量间的内积法，其计算公式如下：

$$\sum_{1}^{N} QT_i * DT_i \tag{1-3}$$

其中，QT_i是检索提问中检索项i的权值，DT_i是文档中标引项i的权值，N

为总的项数。

当每个向量都通过余弦法进行加权后，则内积法转换为余弦法。余弦法采用的相似度计算指标是两个向量夹角的余弦函数（见图 1-1）。

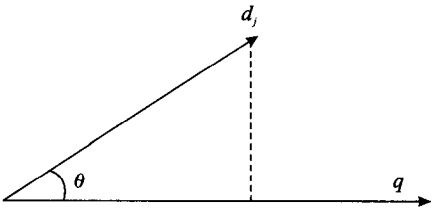


图 1-1　文档向量与提问向量的夹角及余弦值

设提问向量 $\bar{q}=(w_{1q}, w_{2q}, \cdots, w_{nq})$，文档向量 $\bar{d}_j=(w_{1j}, w_{2j}, \cdots, w_{nj})$，按照两个向量夹角余弦的计算含义，文档 d_j 和提问 q 的相似度值就可以通过下面的计算公式获得。

$$sim(d_j, q)=\cos(\bar{d}_j \cdot \bar{q})=\frac{\bar{d}_j \bar{q}}{|\bar{d}_j| \times |\bar{q}|}$$

$$=\frac{\sum_{i=1}^{n} w_{ij} * w_{iq}}{\sqrt{\sum_{i=1}^{n} w_{ij}^2} * \sqrt{\sum_{i=1}^{n} w_{iq}^2}} \tag{1-4}$$

公式（1-4）中，$|\bar{d}_j|$ 和 $|\bar{q}|$ 分别表示文档向量 $\bar{d}_j$ 和提问向量 $\bar{q}$ 的模，因子 $|\bar{q}|$ 对所有文档来说都是一样的，因此它不影响排序的结果；而 $|\bar{d}_j|$ 是文档集合中的一个标准化因子。

因为 $w_{ij}\geqslant 0$ 和 $w_{iq}\geqslant 0$，因此有 $0\leqslant sim(d_j, q)\leqslant 1$。因此，检索处理不仅能判断文档是否相关，而且还可以定量地计算所有文档与某一提问的相关度大小，并能够按照相关度值对结果文档进行排序。即使文档与查询只是部分匹配，该文档仍有可能被检索出来。

为了更有效地控制检索结果的规模，可以设定一个相关度的阈值（threshold）λ，凡与提问向量的相关度值大于 λ 的文档，都将作为检索结果提供给用户。或者，先把所有文档按其与提问的相似度值降序排列，再将前 n 篇文档作为检索结果输出，其中 n 为用户所希望得到的检索结果数量。

1.2.2 向量空间模型的分析

向量空间模型最早起源于文本信息检索实践，对揭示信息检索的基本原理做出过重要贡献。在 VSM 中，研究人员成功地将非结构化的文本信息表示成向量形式，为随后的各种文本信息处理操作奠定了数学计算的基础。向量空间模型在检索处理中所具有的先进技术特征主要表现在以下几个方面。

（1）对标引词的权重进行了改进，其权重的计算可以通过对标引项出现频率的统计方法自动完成，使问题的复杂性大为降低，从而改善了检索效果。

（2）将文档和查询简化为标引词及其权重集合的向量表示，把对文档内容和查询要求的处理简化为向量空间中向量的运算。

（3）采用部分匹配策略，使得在算法层面上基于多值相关性的判断处理得以实现。

（4）根据文档和查询之间的相似度对检索结果进行排序，使对检索结果数量的控制与调整具有相当的弹性与自由度，有效地提高了检索效率。

当然，向量空间模型理论也存在着明显的缺陷，具体包括以下几个方面。

（1）从文档中抽取出的各标引词之间的关系做了相互独立的基本假定，这会失掉大量的文本结构信息，如文档句子中词序的信息，因此降低了语义的准确性。

（2）相似度的计算量较大，当有新文档加入时，必须重新计算标引词的权重。

（3）在标引项权重的计算中，对不同语言单位构成的项都只考虑其统计信息，而仅以该信息来反映标引项的重要性，显然缺乏全面性。

1.3 扩展布尔模型

为了克服布尔检索的缺陷，Waller 和 Kraft 在 1979 年提出了加权布尔检索模型，Salton 于 1983 年提出了扩展布尔检索模型。扩展模型是传统布尔检索模型完全匹配的严格性和向量模型提问的无结构性的折中，在保持布尔检索的结