一种实用的信息检索方法*)

白田恬　邢永康
(重庆大学计算机　重庆 400044)

摘　要　本文依次介绍了信息检索的三类数学模型——集合模型、代数模型和概率模型，并对这三类信息检索模型的检索效果进行了分析。在此基础上提出了一种实用的信息检索方法，我们称为二次检索方法。该方法基于布尔模型和向量空间模型，综合了两者的特点，从而有效地提高了信息检索的效果。文章最后通过实验，对二次检索方法、布尔模型、向量空间模型的查全率、查准率进行了比较，验证了二次检索的优点。

关键词　信息检索，信息检索模型，二次检索，查全率，查准率

1　引言

信息检索是在大量的信息中找出所需要的信息，而如今各类信息日益剧增，并且由于网络的存在推动了信息的加速传播。所以当我们在面对如此海量信息的时候，如何才能精确地检索到所需要的信息，成了大家最关注的问题。

现阶段我们所采用的信息检索模型一般都基于布尔表达式的集合模型或向量空间的代数模型。集合模型由于它模拟的是传统的手工检索方法，因此最先用于计算机信息检索系统。集合模型符合传统的检索习惯，容易理解与掌握；检索过程不需要占据很大的存储空间；检索运算是文献顺序号的与、或、非代数运算，检索速度快。而代数模型将向量空间的概念引入到信息检索中，大大地提高了检索的效率，在理论上具有较高的先进性。所以本文在以提高检索效率为目的的情况下，将集合模型和向量模型结合在一起，提出了一种实用的信息检索方法。

本文各节内容如下：第2节将当前信息检索系统所采用的模型分为三类，并分别对其进行了介绍；第3节对这三类信息检索模型进行了对比分析，尤其对各种模型的检索特点进行对比；第4节提出了一种实用的信息检索方法；第5节通过实验，验证了该检索方法的有效性；最后进行了总结。

2　信息检索模型介绍

信息检索的数学模型，简称信息检索模型，是对信息检索任务及实现方法的一种抽象描述，信息检索模型的选择确定了信息检索的效果。对信息检索模型的研究现在已经是信息检索领域的主要研究内容之一。

依据各个模型的数学基础，可以将现有的信息检索模型分为三类。第一类是基于集合理论的模型，这种模型至今仍是信息检索系统各种数学模型的基础。第二类是向量模型、第三类是概率模型。我们现在使用的搜索引擎，比如平常生活中最常用的google、百度等都是在这三种模型的基础上设计出来的。下面我们分别对这三种模型做简单介绍。

2.1　集合模型

集合模型是基于集合理论的模型，其典型的模型有：布尔模型、扩展的布尔模型及基于模糊集的模型。而最常用的是布尔模型，所以我们常常也把布尔模型当成是集合模型的代表，以此来介绍集合模型。布尔模型是基于集合论和布尔代数的一种简单的过滤模型，它是将每个文档表示为索引项集合，通过集合运算来判定文档与检索的相关度。系统要求用户输入查询的内容，如果有多项内容，那么用户的查询词就用布尔运算符“与”(and)，“或”(or)，“非”(not)进行连接，查询串一般以语义精确的布尔表达式的方法输入，然后通过对文献标记与查询串的逻辑比较获取文献。

2.2　代数模型

检索系统的代数模型是检索系统中所有数学模型中相对来说较有想像力和创造性的一种模型，较能较好地揭示文献之间的关系，但使用最复杂、要求最高的模型。典型的模型有：向量空间模型、扩展的向量空间模型及潜在语义空间模型。

向量空间是由一组线性无关的基本向量组成，向量维数与向量空间维数一致，并可以通过向量空间进行描述。在向量空间模型中，文档D是泛指文档或文档中的一个片段(如文档中的标题、摘要、正文等)。而特征项t是指出现在文档中能够代表文档性质的基本语言单位(如字、词等)，也就是通常所指的检索词，这样一个文档D就可以表示为$D(t_1, t_2, \cdots, t_n)$，其中n就代表了检索字的数量。向量空间模型中的文档就被形式化为n维空间中的向量，

*)本研究得到国家自然科学基金青年基金(60403009)资助。白田恬　研究生。邢永康　副教授，硕士生导师。

空间的一维是倒排表中的一个元素。这样文档 D 的向量可以表示为 $D(w_{i1}, w_{i2}, \cdots, w_{in})$，其中 $w_{i1}, w_{i2}, \cdots, w_{in}$ 分别代表了特征项 $t_1, t_2, \cdots, t_n$ 对文档 d_i 的贡献程度，称之为权重。特征项权重 $w_{i1}, w_{i2}, \cdots, w_{in}$ 代表特征项 $t_1, t_2, \cdots, t_n$ 能够代表文档 d_i 能力的大小，体现了特征项在文档中的重要程度，其取值范围是[0,1]。对于所有文档和用户查询都映射到向量空间，从而将文档的分类过程简化为空间向量的运算，文档信息的匹配问题转化为向量空间的矢量匹配问题，大大减小了问题的复杂度。

2.3 概率模型

这个模型的基本思想是：根据事前检索过程中得到的相关性的先验信息，计算文献集合中每篇文献成为相关文献的概率，然后根据用统计决策理论（即贝叶斯决策准则）决定的输出标准确定哪些文献作为命中文献输出。

概率检索有几个的假设前提和理论：

(1)、相关性独立原则假设。文献对一个检索项的相关性与文献集合中的其他文献是独立的。

(2)、词的独立性假设。索引项和检索项中词与词之间是相互独立，任何一个词的出现与否都不会影响到其它词的出现，它们是相互独立的。

(3)、文献相关性是二值的，即只有相关和不相关两种。

(4)、概率排序原则，是 Robertson 1977 年提出的。该原则认为，如果一个检索系统对用户的每个检索提问的反应是以文献集合中的文献按相关性递减的顺序排列，那么系统的总体效果将是最好的。

(5)、贝叶斯(Bayes)定理，用公式表示为：

$$p(R|d)=\frac{p(d|R)\times p(R)}{p(d)}$$

概率信息检索的目的是估计 $p(R|q,d)$，即文献 d 对检索式 q 来说被用户判断为相关的概率。概率检索模型基本方法是：每一篇文献根据有没有特征项将文献表示为二值向量 $d=(d_1, d_2, \cdots, d_n)$，n 是特征项的数量，$d_i=0$ 或 1 表示文献中没有或有第 i 个特征项。再由文献相关性独立假设：用 R 表示文献相关，$\overline{R}$表示文献不相关，对每一篇文献计算 $p(R|x)$和 $p(\overline{R}|x)$来决定哪个是相关的，哪个是不相关的。由于我们不能直接估计 $p(R|x)$和 $p(\overline{R}|x)$的值，因此要用已知的量来进行估计，根据贝叶斯理论 $p(R|d)=\frac{p(d|R)\times p(R)}{p(d)}$ 和 $p(\overline{R}|d)=\frac{p(d|\overline{R})\times p(\overline{R})}{p(d)}$，这里 $p(R)$和 $p(\overline{R})$是相关和不相关的先验概率，$p(R|x)$，$p(\overline{R}|x)$正比于给定的文献 d 相关或不相关的概率。为了决定文献相关的阈值，需要一个决策依据，最简单的决策判断是：$p(R|x)>p(\overline{R}|x)$即文献相关程度大于不相关程度则认为文献 d 是相关的，否则认为文献 d 不相关。在两者相等时，人为地认为它是不相关的。为了使检索结果能够排序，还要确定排序函数。

3 三类模型对比分析评价

在对各种类型的模型的简单介绍，我们可以来看看各个模型的优缺点：

(1)集合模型的优点是简单、易理解、易实现，故在检索系统中得到了广泛的应用。尽管布尔模型有着种种优点，但它还是存在明显的局限性：布尔模型是基于二值判定标准的，文献要么相关、要么不相关，并没有一个相关级别的概念，因此很难有好的检索效果；构造布尔逻辑式不是一件容易的事情，对于一般用户而言，很难用 AND、OR、NOT 运算符的结合来准确地表达一个检索语句，并且检索词的简单组配不能完全反映用户的实际需要；检索输出完全依赖于布尔提问与文献的匹配情况，很难控制输出量的大小；检索结果不能按用户定义的重要性排序输出，用户只能从头到尾浏览输出结果才能知道哪些文献更适合自己的需要。

(2)从向量空间模型的特点可以看出，在特征项确定的情况下，特征项的权重计算是文档分类的关键，特征项权重计算常用的方法有布尔函数、开根号函数、对数函数、TFIDF 函数等。其中 TFIDF 函数应用最为广泛，其基本思路体现出了查询内容与文档的相关度大小，一般采用使用出现频率的倒数来计算，但是 TFIDF 函数也存在缺点，它虽然考虑了出现特征项的文本在整个文档集中的比例，却不能很好地把握特征项在文本集合中分布的差异，所以影响了分类的最终效果。

向量模型的也有两个缺点，第一个就由于特征项在文档中的不同位置会代表不同的权重，而不同的关键词长度也会影响权重的大小。在传统的 TFIDF 函数中，每增加一个文档都要重新计算向量，导致查询速度降低，同时由于使用频率因子，在扩大查询范围时，不可避免地会影响到查询的准确性。

向量模型的另一个问题在于查询和文档向量间是依靠链接来判断的，而且判断的依据中简单的两者相同关键词的比较，但实际情况是，大量的关键词具有相同的语义，同一关键词也会有多种语义的解释描述（即产生了语义分歧）。

(3)只有概率模型才能反映出文献与提问的相关性的大小，从而对相关文献进行排队，集合模型和代数模型都做不到这一点。但是概率模型不好单独应用于检索系统中，所以我们可以把概率模型和其它两种检索模型总和起来使用，以达到最佳的检索效果。

4 一种实用的信息检索方法

通过对上述模型优缺点的对比，为了提高检索的效率，本文采用了一种新的实用信息检索方法：将

布尔模型和向量模型结合起来，进行两次检索，由此来提高信息检索的效率。我们称这种检索方法为二次检索方法。在此过程中，我们先用布尔模型对原始文献的数据集 $D=(d_1,d_2,\cdots,d_n)$ 进行第一次检索，得到文献数据集 D_1。这里我们之所以采用布尔模型的原因是在目前所有的检索模型中布尔模型是最简单、易理解、易实现的。再采用向量模型对 D_1 进行二次检索得到最后我们需要的相关文献 D_2，D_2 就是我们得到的最终结果。

(1)首先我们使用布尔模型进行一次检索：用户输入查询的内容的，如果有多项内容，那么用户的查询词就用布尔运算符“与”(and)，“或”(or)，“非”(not)进行连接得到查询串 q，查询串一般以语义精确的布尔表达式的方法输入，然后通过对文献检索项与查询串的逻辑比较获取文献。查询串 q 是一个传统的布尔表达式，文档与查询串的相关度定义为查询串 q 是一个传统的布尔表达式，文档与查询串的相关度定义为：

$$sim(d_j,q)=\begin{cases}1, q\in d_j\\ 0, q\notin d_j\end{cases}(j=1,2,\cdots,n)$$

如果 $sim(d_j,q)=1$，布尔模型表示文档 d_j 与查询串 q 相关，否则就表示文档 d_j 与查询串 q 不相关。我们将 $sim(d_j,q)=1$，也就是的用布尔模型进行检索后的相关文档进行归类得到 $D_1=(d_1,d_2,\cdots,d_m)$，我们可以知道在这里有 $m\leqslant n$。

(2)我们用向量模型对得到 D_1 进行二次检索。在这里查询词就不再是用查询串来表示，而是查询向量 Q 来表示。相似度 S 来代表两个文档内容的相关程度，所以文档 d_j 和查询向量 Q 均以一个 n 维向量来表示时，Q 的权重向量就表示为 $(w_{q1},w_{q2},\cdots,w_{qn})$，$d_j$ 和 Q 的相关度就是相似度，一般使用内积或夹角 θ 的余弦来计算，两者夹角越小，余弦值越大说明相似度越高。图 1 示意了文档与文档，文档与查询之间的相关度。

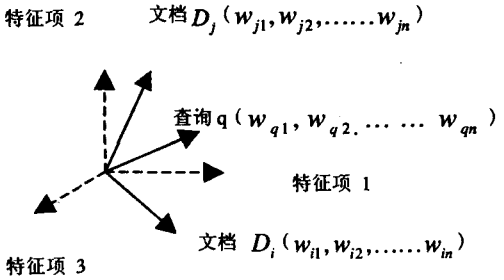

图 1　文档的相关度图示

如下的公式就表示了文档与查询之间相关度的计算。

$$sim(d_j,q)=com(d_i,q)=\frac{d_i\times q}{|d_i|\times|q|}$$

$$=\frac{\sum_{j=1}^{n}w_{ij}\times w_{qj}}{\sqrt{\sum_{j=1}^{n}w_{ij}^2}\times\sqrt{\sum_{j=1}^{n}w_{qj}^2}}$$ 排序这个结果后与设立的阈值进行比较，如果大于阈值则文档 d_j 与查询相关，保留该文档 d_j 的查询结果，如果小于则不相关，过滤此文档 d_j，这样就可以控制查询结果的数量，加快查询速度。最后我们就得到了查询的最后结果文献集合 D_2。

我们可以通过下面的实验，将这种实用的信息检索方法与传统的信息检索方法相比较，得到它们的检索效果。

5　实验及分析

信息检索效率是评价一个检索系统性能优劣的质量标准，衡量检索效率的指标有查全率、查准率、漏检率、误检率、响应时间等。目前，人们通常主要以查全率和查准率这两个指标来衡量。查全率和查准率用公式可以表示为：

$$\text{查全率(R)}=\frac{\text{检索到的相关信息量(W)}}{\text{系统中存储的所有相关信息量(X)}}\times 100\%$$

$$\text{查准率(p)}=\frac{\text{检索到的相关信息量(W)}}{\text{检索出的信息总量(M)}}\times 100\%$$

我们也将用这查全率和查准率两个指标来进行衡量仅用布尔模型、仅用向量模型和本文提出的使用的二次检索方法我们所选的实验数据为 786 篇信息提取和数字图书馆方面的文章，数据来源主要为 CIIR 所发表的文章和数字图书馆杂志中的文章。如果我们对文章的全部内容进行分析，提取检索项的话，这样工作量就太大了，所以我们只是将文献的标题作为检索项，这样的存储量和计算复杂度并不大，这些文档的标题中包括检索项 1816 个。然后我们选定了十个查询关于信息检索和数字图书馆方面信息的用户。在以上实验数据的基础上采用了基于布尔模型和向量模型的二次信息检索方法进行检索，并与仅用布尔模型、仅用向量模型的检索方法进行了对比，结果如表 1 所示。

表 1　三种模型的检索效率（保留小数点后一位）

所采用的方法	查全率（%）	查准率（%）
仅用布尔模型的方法	58.9	25.6
仅用向量模型的方法	80.1	32.2
实用的二次检索方法	89.2	35.1

从上面的实验结果可以得出以下两点：

(1)、采用二次检索的方法得到的检索结果比仅用布尔模型的查全率和查准率分别高了 30.3%、9.5%。比仅有向量模型的查全率和查准率高了 9.1%、2.9%。说明二次检索的效率比较好，实用性比较高。

(2)、我们也很容易就可以看到二次检索的查准率虽然相对布尔模型和向量模型是有所提高，但是 35.1%还是比较低的，特别是对一些本身文档数量就比较少的集合进行检索，可能效果会很不理想。而之所以查准率提不高的原因是代数模型和集合模

型都不能反映文献与提问的相关性的大小。只有概率模型才能反映出文献与提问的相关性的大小，从而对相关文献进行排队，所以想要提高查准率还要采用概率模型才能得到实现。本文就不在这方面再作详细的讨论了。

总结 本文首先从信息检索数学模型做了简单的介绍入手，提出了现有的三种有代表性的信息检索模型：集合模型，代数模型，概率模型，并对这三种检索模型的优缺点作了分析，在此基础上提出了一种实用的二次检索方法，并且详细说明了此方法的检索过程，指出了采用二次检索的目的是为了提高检索的效率。文章最后通过对一组具体数据的采集和实验，用查全率和查准率来作为检索效率的评价标准，可以知道采用基于二次检索方法的查全率和查准率高于仅用布尔模型和向量模型的查全率和查准率。

各种模型的混合使用是多变、复杂的，结合各种模型的优势，采用一种混合模型就能够很好地提高检索效率。由于集合模型的成熟研究，现在的研究多在代数模型和概率方面，比如以概率论和模糊数学为补充手段的文献向量加权上，对比各种加权方法的优劣，选择出一种或几种好的加权方法也是现在的研究热点。总之以代数模型和概率模型为基础的综合性模型的研究将成为将来信息检索模型的研究重点。

参 考 文 献

1 Miyamoto S. Information Retrieval Based on Fuzzy Association [J]. Fuzzy Sets and System, 1990, 38(2): 191～205
2 Lee C, Lee G G. Probabilistic Information Retrieval Model for Dependency Structured Indexing System. Information Processing and Management [M]. SIGIR2002, August2002
3 Jones K S, Walker S, Robertson S E. A Probabilistic Model of Information Retrieval : Development and Comparative Experiments Part 2 [J]. Information Processing and Management, 2000, 36(12): 779～808
4 沈一栋，邢永康. 一种新的知识表达模型——信度网[J]. 计算机科学，2000, 9(27): 40～43
5 邓珞华. 信息检索系统数学模型的理论及其评价——谨以此文献给信息检索的先驱杰拉尔德·索顿先生[J]. 大学图书馆学报，2002, 1(3): 6～4
6 邢永康，马少平. 信息检索的概率模型[J]. 计算机科学，2003, 30(8): 13～17
7 康耀红. 现代情报检索理论[M]. 北京：科技文献出版社，1990
8 王继成，邹涛，等. 基于 Internet 的信息资源发现技术与实现. 计算机研究与发展[J]，1999, 36(11): 1369～1374
9 何静，刘海燕. 信息检索与过滤中的信息需求表示方法[J]. 计算机工程与设计，2003, 24(8): 3～8

（上接第 238 页）

空间关系描述发展的基础上，可将其研究结果应用于空间推理，对空间推理的发展产生了正向的推动。由于空间推理的研究对象的转变，极大地扩展了空间推理的应用领域，使空间推理的理论和应用研究近年来有了长足的进展。在国外，近年来成立了许多专门从事空间推理方面研究的协会和联盟，如 NCGIA(National Center for Geographic and Analysis)，USGS(U. S. Geological Survey)，欧洲定性空间推理网 SPACENET 以及匹兹堡大学的空间信息研究组和慕尼黑大学空间推理研究组等等。国际知名期刊 Artificial Intelligence 近年来发表了许多篇空间推理方面的文章，而且呈逐年增长的趋势。

(2)空间拓扑本身方法的创新。由于拓扑空间关系表示是空间关系理论的重要组成部分，也是空间数据库设计的重要基础，其研究将有助于设计有效的空间查询和有效的数据处理方式。而由于对空间关系表示的侧重点不同，也产生了很多的空间拓扑关系表示方法，因此基于不同空间拓扑关系表示方法的空间推理方法一般来说是不可以通用的，这就导致了对同一问题研究的重复和浪费。

因此，在空间拓扑领域的研究中，一方面应该研究出更符合地理信息系统本身特点的表达方式，另一方面应该找到将现有各种表示方法转化为一个统一的较优的空间拓扑关系表达方式，使原来在不同空间拓扑关系条件下的空间推理方法得到新的应用。

参 考 文 献

1 刘亚彬，刘大有. 空间推理与地理信息系统综述. 软件学报，2000
2 Renz J, Nebel B. On the complexity of qualitative spatial reasoning: a maximal tractable fragment of the region connection calculus. Artificial Intelligence, 1999
3 U. S. Geologic Surver URL. 1998. http://nsdi.Usgs.gov/nsdi
4 肖乐斌，等. 三维 GIS 的基本问题探讨. 见：地理信息系统论坛(GIS Forum)-学术论文，2002
5 王康弘，刘 利，钟耳顺. 地籍空间实体的空间拓扑关系和变更类型分析. 2001'中国 GIS 年会论文集，2001. 3
6 肖乐斌，钟耳顺，等. GIS 空间概念模型的研究. 见：地理信息系统论坛(GIS Forum)-学术论文，2002
7 GuoPing Tao Huang-Fu, Research on the Relationship Between 4-intersection and Classifying Invariant Base on the Simple Regions. In: International Conference on Machine Learning and Cybernetics (ICMLC 2003), 2003. 11
8 虞强源，刘大有，等. 空间区域拓扑关系分析方法综述. 软件学报，2003
9 Egenhofer M J, Franzosa R D. On the Equivalence of Topological Relationships. Int. Jour. Of Geographical Information Systems, 1995(9): 133～152.
10 Clementini E, Felice P Di. Topological invariants for lines. IEEE Transactions on Knowledge and Data Engineering, 1998, 10(1): 38～54
11 郭薇，陈军. 基于点集拓扑学的三维拓扑空间关系形式化描述. 测绘学报，1997, 26(2): 122～127
12 廖士中，石纯一. 定性空间推理的研究与进展. 计算机科学，1998, 25(4): 11～13