计算机科学1999Vol. 26№. 9

# 文本信息检索技术*)

A Survey of Text Retrieval

邹 涛 王继成 杨文清 张福炎

(南京大学多媒体计算机研究所 软件新技术国家重点实验室 南京 210093)

**Abstract** Information Retrieval is an important technology in information service. This paper discusses the technology of full text retrieval which adopts the approach of "indexing characters" and "indexing words" first, afterwards the technology of content based text retrieval is described.

**Keywords** Retrieval, Full text retrieval, Retrieval model

## 1. 引言

随着信息技术的发展,人们已经从信息缺乏的时代过渡到了信息极大丰富的时代,于是也就出现了"富数据穷信息"的问题。如何迅速、有效地从大量数据中找到所需的信息已经成为信息服务领域中的重要的、亟待解决的问题,信息检索技术就是针对这一问题所发展起来的。尽管与多媒体信息相比,文本信息显得比较平凡,但它是人们用于信息记载和信息传播的最重要的媒体之一,也是人们最熟悉、使用最多的媒体,因此,文本信息检索技术既是很重要又是很容易取得技术突破并获得广泛应用的技术。国内外对信息检索技术特别是文本信息检索技术的研究已经开展了多年,在理论上和应用上都取得了很大的进展,并开发出了多个成功的信息检索系统,如Massachusetts 大学的 INQRERY 系统、Cornell 大学的 SMART 系统、易宝北信的 TRS、北大方正的 MIRS 等。文本信息检索包括两方面的核心技术:一是如何建立和维护检索索引库;二是如何提供快速有效的检索机制。本文针对文本信息的检索介绍几种主要的、具有代表性的技术。

## 2. 全文检索技术

全文检索是指以文档的全部文本信息作为检索对象的一种信息检索技术。全文检索的核心技术是文档的索引,即如何将源文档中所有基本元素的出现信息以适当的形式记录到索引库。在中文文档中,"基本元素"可以是单个汉字,也可以是词或词组,根据索引库中索引的元素不同,可以将全文检索分为基于字表的全文检索和基于词表的全文检索两大类。

### 2.1 字表法

2.1.1 *索引的组织* 字表法是对每个单字的出现位置进行索引,并依据单字的位置信息进行检索的文本检索方法。字表法索引库的主要部分是每个汉字的字表信息,索引库中的字表结构如图1所示,其中字符 $i$ 对应的字表记录了该字符在源文档中的所有出现位置 $P_{ik}$,出现位置通常用字符相对于文档头的偏移字节数表示。建立字表索引时,需要扫描整个源文档,对所出现的每一个有效字符,计算其在文档中的出现位置并将该位置值加入到对应的字表中。

2.1.2 *字符串的检索* 字表记录了对应字符在源文档中的所有位置信息。考察一个字符串,例如两个字的字符串 $XY$(其中 $X$、$Y$ 表示任意的汉字字符),假设 $X$ 的位置为 $P_x$,如果字符串 $XY$ 在源文档中出现,则 $Y$ 的位置 $P_y$ 必定等于 $P_x+2$(2为两个汉字间的字节距离),在索引库中,$X$ 的字表中将包含 $P_x$,而 $Y$ 的字表中也必然包含 $P_x+2$。进行检索时,扫描 $X$ 和 $Y$ 各自对应的字表,若文档中有该词的出现,则必定有 $X$ 对应的字表中存在位置值 $P_x$,$Y$ 对

*) 本文受到江苏省科委95科技攻关项目:"面向电子报刊电子图书馆的多媒体网络出版系统"的资助。邹 涛 博士生,研究方向为中文信息处理、计算机网络、多媒体技术。张福炎 教授,博士导师,主要研究方向为多媒体技术、中文信息处理、计算机网络、计算机图形学。

应的字表中存在位置值 $P_y$，使得 $P_y=P_x+2$ 成立，每查到一对这样的位置值，就是检索出字串 $XY$ 的一次出现，扫描完两字字表，就可以检索出字符串的所有出现。

| … | … |
|---|---|
| 啊 | $P_{11}P_{12}P_{13}\cdots$ |
| 阿 | $P_{21}P_{22}P_{23}\cdots$ |
| … | … |
| 网 | $P_{i1}P_{i2}P_{i3}\cdots P_{im}\cdots$ |
| … | … |
| 络 | $P_{j1}P_{j2}P_{j3}\cdots P_{jn}\cdots$ |
| … | … … |

图1　字表结构

| … | … |
|---|---|
| 计算机 | $P_{11}P_{12}P_{13}\cdots$ |
| 多媒体 | $P_{21}P_{22}P_{23}\cdots$ |
| … | … |
| 网络 | $P_{i1}P_{i2}P_{i3}\cdots P_{im}\cdots$ |
| … | … |
| 通信 | $P_{j1}P_{j2}P_{j3}\cdots P_{jn}\cdots$ |
| … | … |

图2　词表结构

### 2.2　词表法

2.2.1　*索引的组织*　与字表法相似，词表法是以能表达一定意义的词为基本检索单位，并根据词的出现位置进行索引和检索的文本检索方法。词表的结构如图2所示，词表中记录了词条 $i$ 在源文档中的所有出现位置 $P_{ix}$，出现位置通常也采用词条相对于文档头的偏移字节数表示。建立索引时，首先需要引用切分词表对源文档进行词条的切分，然后对切分后的文档词条进行统计，记录每一个出现的词条及其出现的位置。除了切分词表外，基于词表的检索系统一般还要建立同义词表、反义词表、关联词表等多个辅助词表，用于进行同义词、反义词等的概念检索。

2.2.2　*字符串的检索*　当检索单个词条 $X$ 时，只需直接在词表中进行查找，如词表中包含 $X$，则说明文档中也包含词条 $X$，词表中记录的出现位置也即为词条 $X$ 在文档中出现的位置；当需要检索例如 $XY$ 形式的词组（$X$、$Y$ 表示任意的词条）时，检索方法与字表法相同：对于词条 $X$ 的所有出现位置 $P_x$，如果在词表中存在词条 $Y$ 的记录且存在 $P_y=P_x+4$（假设 $Y$ 为双字词）的出现位置，则检索到了词组 $XY$ 的一次出现。

### 2.3　字表法与词表法的比较

字表法和词表法各有优缺点，有各自适用的场合和处理对象。字表法是以单字为基础进行检索的方法，其缺点是生成的索引库庞大（索引文件的长度往往大于源文档的长度），检索速度低，错检率高，例如检索"华人"一词时，会检索出"中华人民共和国"这样的错误结果；其优点是适应性强，应用范围广、索引的生成简单，比较适用于内容复杂、新词汇和特殊词汇多的文档的检索。词表法是以词为基本元素进行索引与检索的，它需要使用大规模的切分词表对被索引文档进行词条切分，切分词表和索引的建立较复杂，漏检率较高，且不能进行单字和任意字符串的检索；其优点是对于大规模应用，索引库规模小，检索的处理速度快，同义、反义等概念检索的实现较为简单，比较适用于特定领域中或内容相对固定的文档的全文检索。因此在应用时应根据应用的具体情况和应用对象采用适当的检索方法。

## 3. 基于内容的文本检索技术

上述基于字表和基于词表的全文检索方法，是不考虑文档的具体内容而仅判断是否包含被检词条的检索方法。基于内容的检索是能够根据文档的内容处理类似"检索出属于多媒体类且包含通信内容的文档"等涉及文档内容查询的检索技术。检索模型的构造是基于内容检索的核心技术，检索模型包含三个方面的内容：文档与用户查询的表示；查询匹配策略；匹配结果的相关度表示。下面介绍几种典型常用的信息检索模型。

### 3.1　布尔模型

布尔模型是一种简单而常用的严格匹配模型，它定义了一个二值变量集合来表示文档，这些变量对应于文档中的特征项，一般是由训练文档集中的词条或短语组成，如果词条对文档内容有贡献则赋予 True，否则置为 False。检索时，根据用户提交的检索条件是否满足文档表示中的逻辑关系将检索文档分为两个集合：匹配集和非匹配集。因匹配结果的二值性，所以无法在匹配结果集中进行查询结果的相关性排序。布尔模型实现简单，检索速度快，在许多检索系统中得到应用，例如 Yahoo!，InfoSeek 等诸多网络检索站点均采用了布尔检索模型。但布尔模型的文档表示能力差，无法区分特征项对文档内容贡献的重要程度，并且逻辑表达式过于严格，往往

会因为一个条件未满足而忽略了其它全部特征，造成大量的漏检。

$p$ 范数模型是对布尔模型的扩展，它克服了简单布尔模型匹配函数过于严格而导致漏检率高的致命缺陷。在 $p$ 范数模型中，假设文档 $D$ 可表示为：$D=(d_1,d_2,\cdots,d_n)$，用户查询可表示为：$Q=(q_1,q_2,\cdots,q_n)$，其中 $d_i$ 和 $q_i$ 分别表示第 $i$ 个特征词条对文档内容和查询内容的贡献程度，$d_i$、$q_i$ 在[0,1]的区间上取值。定义文档与查询间的相似度：

$$Sim(D,Q)=1-\left[\frac{\sum_{i=1}^{n}q_i^p(1-d_i)^p}{\sum_{i=1}^{n}q_i^p}\right]^{1/p}$$

其中 $1\leqslant p\leqslant\infty$，根据具体应用改变 $d_i$、$q_i$ 和 $p$ 的取值即可达到不同的检索效果。当 $p=\infty$，且 $d_i$ 的取值为0或1，$q_1=q_2=\cdots=q_n=1$ 时，$p$ 范数模型则退化为简单布尔模型。在实际使用中 $p$ 的取值由实验得出，取值范围一般为[2,5]。

### 3.2 向量空间模型

向量空间模型(VSM)是近些年使用较多且效果较好的一种信息检索模型。在 VSM 中，将文档看作为是由相互独立的词条组$(T_1,T_2,\cdots,T_n)$构成，对于每一词条 $T_i$，都根据其在文档中的重要程度赋以一定的权值 $W_i$，并将 $T_1,T_2,\cdots,T_n$ 看成一个 $n$ 维坐标系中的坐标轴，$W_1,W_2,\cdots,W_n$ 为对应的坐标值。这样由$(T_1,T_2,\cdots,T_n)$分解而得的正交词条矢量组就张成了一个文档向量空间，文档则映射成为空间中的一个点。对于所有文档和用户查询都可映射到此文本向量空间，用词条矢量$(T_1,W_1;T_2,W_2;\cdots;T_n,W_n)$来表示，从而将文档信息的匹配问题转化为向量空间中的矢量匹配问题处理。假设用户查询为 $Q$，被检索文档为 $D$，两者的相似程度可用向量之间的夹角来度量，夹角越小说明相似度越高，相似度计算公式如下：

$$Sim(Q,D)=\cos(Q,D)=\left(\sum_{k=1}^{n}W_{qk}\cdot W_{dk}\right)\Big/\sqrt{\sum_{k=1}^{n}W_{qk}^2}\cdot\sqrt{\sum_{k=1}^{n}W_{dk}^2}$$

表示矢量中词条 $T_i$ 及其权值 $W_i$ 的选取称为特征提取，特征提取是利用向量空间模型进行信息检索的关键步骤。自然语言文档中，各词条在不同内容的文档中所呈现出的频率分布是不同的，因此可根据词条的频率特性用统计的方法进行特征提取。文档中，词条的重要性正比于词条的文档内频数，反比于训练文档集中出现该词条的文档频数，因而可构造词条权值评价函数：

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

其中 $tf_{ik}$ 表示词条 $T_k$ 在文档 $D_i$ 中的出现频数，$N$ 表示用于进行特征提取的全部训练文本的文档总数，$n_k$ 表示词条 $T_k$ 的文档频数。在实际应用中，为避免因文档长度引起的频数变化，还应对词条权值评价函数作规范化处理：

$$W_{ik}=\frac{tf_{ik}\cdot\log\left(\frac{N}{n_k}+0.5\right)}{\sqrt{\sum_{k=1}^{n}(tf_{ik})^2\cdot\log^2\left(\frac{N}{n_k}+0.5\right)}}$$

### 3.3 概率模型

布尔模型和向量空间模型都将文档表示词条视为是相互独立的项，忽略了表示词条间的关联性，而概率模型则考虑到了词条、文档间的内在联系，利用词条间和词条与文档间的概率相依性进行信息的检索。

二值独立检索模型(BIR)是一种实现简单且效果较好的概率检索模型。在 BIR 中，假设文档 $D$ 和用户查询 $Q$ 都可用二值词条向量$(x_1,x_2,\cdots,x_n)$表示，如果词条 $T_i\in D$，则 $x_i=1$，否则 $x_i=0$。利用 Bayes 公式并经过简化后可得文档与用户查询间的相关函数：

$$Sim(D,Q)=\sum\log\frac{p_i(1-q_i)}{q_i(1-p_i)}$$

其中 $p_i=r_i/r$，$q_i=(f_i-r_i)/(f-r)$，$f$ 表示训练文档集中文档总数，$r$ 表示训练文档集中与用户查询相关的文档数，$f_i$ 表示在训练文档集中包含词条 $T_i$ 的文档数，$r_i$ 表示 $r$ 个相关文档中包含词条 $T_i$ 的文档数。

概率推理网络是近些年被提出并深受重视的一种新型检索模型。推理网络模拟人脑的推理思维模式，将文档内容与用户查询匹配的过程转化为一个从文档到查询的推理过程。基本的文本检索推理网络包含文本网络与用户查询网络两部分，如图3所示。图中每个节点表示一个文档、一个查询或者一个概念，其中 $D_i$ 为文本节点，$T_i$ 为文本表示节点，$R_i$ 为文本概念节点，$C_i$ 为用户提问概念节点，$Q$ 为用户查询节点，有向边表示节点间的概率相依性。网络中文档节点与查询节点间的相关性可以表示为：给定文档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。如要估算用户查询 $Q$ 与文档 $D_i$ 间的概率相关性 $P(R|Q,D_i)$，先将文档

节点 $D_i$ 置为 $True$，然后依次计算使 $P(Q=True)$ 的相依节点的概率即可。推理网络同其它概率模型相比有很多优点、推理网络能够将其它许多概率模型映射到网络中、并能够采用多种检索形式和由其它知识源得到的统计数据或经验数据，进行综合检索。

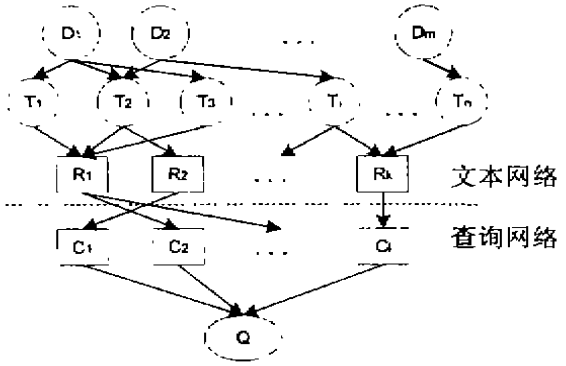


图3 推理网络

**结束语** 以上介绍的是几种较为成熟且得到广泛应用的文本信息的检索方法，信息检索技术的研究已经经历几十年的历史，检索模型不断完善，检索技术研究也获得了巨大的进展，实际应用中的检索效率和查全率、准确度等技术指标也都得到了大幅度的提高，但随着电子图书馆的出现和由 Internet 普及所导致的信息爆炸，对信息检索技术特别是对基于网络的大容量文本信息检索技术的要求越来越高，因此还需对信息检索技术做更进一步的探索，特别是在分布式信息检索、信息提取、检索标准、相关反馈等多个方面的进行深入的研究。

## 参 考 文 献


1 Salton G, et al. A Vector Space Model for Automatic Indexing. CACM, 1975(18): 613～620

2 Gudivada V N Information Retrieval on the World Wide Web. IEEE Internet Computing, 1997(5): 58～68

3 Salton G, Buckley C. Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 1988, 24(5): 513～523

4 Broglio J, Callan J P. INQUERY System Overview. Available at: http://www.cs.umass.edu/~croft

5 邹涛、张福炎. 网络信息搜寻技术与发展. 计算机工程与科学, 1998, 20(4): 33～37

6 吴立德. 大规模中文文本处理. 复旦大学出版, 1997.7

7 潘谦红、等. 文本信息检索模型. 中国计算机报, 1998.19

8 潘谦红, 等. 全文检索的发展. 中国计算机报, 1998.19


(上接第57页)

## 参 考 文 献


1 Braden R, et al. Integrated services in the Internet architecture: an overview. RFC 1633, Jan. 1994

2 Crawley E, Nair R. A Framework for QoS-based Routing in the Internet. RFC2386, Aug. 1998

3 TCP/IP and Related Protocols, Black. U

4 ATM: Internetworking with ATM, UYLESS BLACK, 清华大学出版社

5 Flanagan P. This year's 10 Hottest Technologies in Telecom. Telecommunications, 1998(5)

6 Nelson G. Test techniques for next-generation IP Networks. Telecomm ASIA, Oct. 1998

7 G. malkin Xylogics. RIPng for IPv6, RFC2080, Jan. 1997

8 Nichonls K, et al. A Two-bit Differentiated Services Archietecture for the Internet. IETF Internet Draft, Nov. 1997

9 Braden R, et al. Resource ReSerVation Protocol (RSVP)-Version 1 Functional Specification. rfc2205, Sep. 1997