计算机科学2004Vol. 31№. 9

文本信息检索的精确匹配模型*)

李孝明　曹万华
(武汉数字工程研究所　武汉430074)

摘　要　本文分析了搜索引擎技术存在的不足，并在现有信息检索模型的基础上，提出了一个精确匹配模型。实验表明，该模型比较适合企业内部应用。

关键词　信息检索，搜索引擎，检索模型，精确匹配模型

A Precise-Matching Model of Text Information Retrieval

LI Xiao-Ming　CAO Wan-Hua
(Wuhan Digital Engineering Institute, Wuhan 430074)

Abstract　In this paper, the shortage of search engine technique is analyzed, and a precise-matching model is proposed based on existed information retrieval models. The experiment shows that the precise-matching model is more adaptable for enterprise-application.

Keywords　Information retrieval, Search engine, Retrieval model, Precise matching model

1　引言

随着信息技术的发展，人们已经从信息匮乏的时代过渡到了信息极大丰富的时代，于是，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一个重要的、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。

多年来，国内外对信息检索技术特别是文本信息检索技术进行了多方面的研究，在理论上和应用上都取得了很大的进步，并开发了多个成功的信息检索系统(搜索引擎)，国外具代表性的有 Google、Fast/AllTheWeb、AltaVista、Inktomi、Teoma、WiseNut 等，国内著名的有百度(Baidu)[2]。

但是这些搜索引擎背后的技术是一般用户看不到的，其信息检索搜索引擎一般由专业的搜索引擎服务商提供，而且这些搜索引擎所采用的技术和实现情况都被视为私有财产而不开放。因此，很少有文章披露搜索引擎的详细设计和实现，这在很大程度上制约了搜索引擎技术的继承、交流和发展。

2　搜索引擎技术现状

近年来，互联网的普及大大促进了信息检索技术的发展和应用，正如上文提到的一批搜索引擎产品已经产生，为用户提供了很好的快速信息获取和网络信息导航工具，目前最著名的搜索引擎包括 Google、AltaVista 等，国内百度的中文搜索引擎也取得了很好的成绩。

搜索引擎服务和搜索引擎技术是完全不同的两个概念，每个门户网站都会提供搜索引擎服务，但背后的搜索引擎技术是一般用户看不到的。搜索引擎普遍采用了全文搜索技术，但互联网的信息和一般企业内部信息是不同的，有两个关键问题需要解决：一是速度，一般传统信息检索系统的索引库规模度在 G 级，但互联网网页搜索需要处理成千上万的网页；二是相关性，信息太多，查准和排序就特别重要。解决第一个问题的基本策略都是采用检索服务器群集技术，解决第二个问题的方法包括像 Google 和百度等发展了的链接分析技术。

目前搜索引擎面临两个主要挑战：一是检索的质量仍然需要提高。常常检索的是大量的无用的结果，真正有用的结果却被淹没在其中不容易发现。搜索引擎的索引和以前相比已经有了极大的增长，一般检索时都会返回大量的结果。但是人们查看和选择结果的能力与耐心没有得到相应的提高，通常还是只会注意最前面的部分。因此，搜索引擎的"精度"，尤其是检索结果排在前面的部分对于用户的有用性，是非常重要的，有时候相对于查全率来说显得更加突出。

二是如何转向企业应用。搜索引擎技术的一些优势在企业应用中常常不起作用，甚至变为劣势，变成查不全、查不准、查不稳的搜索技术。如排序技术，企业应用的检索要求基于内容的相关性排序。就是说，和检索要求最相关的信息应排在检索结果的前面，一些搜索引擎所谓的链接分析专利技术以及内容聚类(门户搜索引擎的有效手段)对查询结果的排序基本不起作用，链接分析是根据一个网页被连接次数的多少作为重要性评判的依据，而一个网站内部的网页的链接是由网站内容采编发布系统决定，其链接次数完全是偶然因素，不能作为判别重要性的依据。又比如企业应用中要求搜索结果是稳定的，但搜索引擎常常做不到，而在许多搜索引擎应用中，为了在大规模网页下提高检索速度所采用的检索策略(可以说是技巧)以及索引方法常常导致检索结果的不稳定和不可理解。

3　现有信息检索模型及分析

信息检索模型在文本信息检索中具有相当重要的地位，它直接影响到搜索引擎的质量，对数据的查全率、查准率方面有着重要的关系。

*)本文工作得到十五国防预研项目资助(项目编号4010601010301)。李孝明　硕士研究生，主要研究方向：软件工程。曹万华　研究员，主要研究方向：软件工程技术及应用。

现有的信息检索模型主要有三种，即“布尔模型”、“向量空间模型”和“概率模型”。下面作简单的分析和介绍。

3.1 布尔模型[1]

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的不满足而忽略了其它全部特征，造成了大量的漏检。

P 范数模型[1]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：$D=(d_1,d_2,d_3,\cdots,d_n)$，用户查询可表示为：$Q=(q_1,q_2,q_3,\cdots,q_n)$，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 和 q_i 在[0,1]的区间上取值，定义文档和查询间的相似度为 $Sim(D,Q)$，即：

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

其中 $p\in[1,\infty)$，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。但 $p\to\infty$，且 d_i 的取值为0或1，$q_1=q_2=q_3=\cdots\cdots=q_n=1$时，$P$ 范数模型则退化为简单的布尔模型。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[1,4]

向量空间模型(VSM)是近几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$(T_1,T_2,T_3,\cdots,T_n)$构成，对每一词条 T_i，都根据其在文档中的重要程度赋予一定的权值 W_i，并将 $T_1,T_2,T_3,\cdots,T_n$ 看成一个 n 维坐标系中的坐标轴，$W_1,W_2,W_3,\cdots,W_n$ 为对应的坐标值。这样由$(T_1,T_2,T_3,\cdots,T_n)$分解而得到的正交词条矢量组就组成了一个文档向量空间，文档则映射成为空间中的一个点。对于所有的文档和用户查询都可映射到此文本向量空间，用词条矢量$(T_1,W_1;T_2,W_2;T_3,W_3;\cdots;T_n,W_n)$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，而被检索文档为 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，用户查询 Q 与被检索文档 D 相似度计算公式如下：

$$sim(Q,D)=\frac{\sum_{k=1}^{n}W_{q_k}W_{d_k}}{\sqrt{\sum_{k=1}^{n}W_{q_k}^2}*\sqrt{\sum_{k=1}^{n}W_{d_k}^2}}$$

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。

3.3 概率模型[1,3]

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(BIM)是一种实现简单且效果较好的概率检索模型。在 BIM 中，假设文档 D 和用户查询 Q 都可用二值词条向量$(x_1,x_2,x_3,\cdots,x_n)$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到文档与用户查询间的相关函数：

$$sim(D,Q)=\sum_{i=1}^{n}\log\frac{p_i(1-q_i)}{q_i(1-p_i)}$$

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 表示个相关文档中包含词条 T_i 的文档数。

3.4 模型分析

上文提供的“布尔模型”、“向量空间模型”、“概率模型”当中，首先也是最关键的一步是提取文档表示词条，即特征项的提取，特征项一般是由训练文档集中的词条或短语组成。自然语言文档中，各词条在不同内容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。一种较好的理想的文档特征项提取结构图如图1所示。

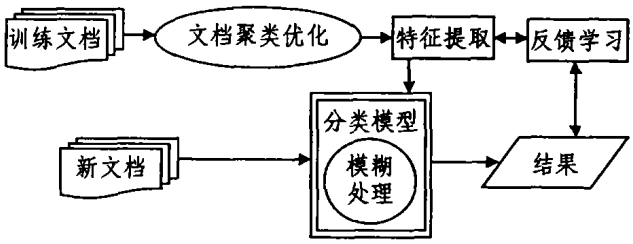

图1 文档特征项提取结构图

从这个文档特征项提取结构图可以看出，特征项的提取是一个相当复杂、艰巨的过程，需要运用先进的计算机技术、人工智能技术辅助完成。由于文本本身的复杂性、不规律性等特点，文本的特征项提取涉及到多方面的综合技术。

其次，上面三种模型在查准率、查全率方面存在较大的误差，由于文本词条的切分、特征词条的提取没有一个严格的标准，在技术处理方面往往采用一种模糊策略。而且文档与用户查询间的相关函数计算也存在一定误差。

4 精确匹配模型

我们在上述“布尔模型”、“向量空间模型”和“概率模型”基础之上，提出一种信息检索精确匹配模型。采用国标汉字字符集 GBK/2：GB2312中包含的6763个汉字作为文档特征项。

①文档特征项可以表示为向量形式：$T=\{t_1,t_2,t_3,\cdots,t_n\}$。其中 $n=6763$，$t_i(1\leqslant i\leqslant n)$代表国标汉字字符集 GBK/2：GB2312中包含的6763个汉字中的某个特定汉字。汉字编码是采用双字节形式，编码分为9个区，高字节分别为：B0～B7，B8～BF，C0～C7，C8～CF，D0～D7，D8～DF，E0～E7，E8～EF，F0～F7；相应的低字节编码均为：A1～FE。

②假设被检索文档为 D，其向量表示形式为：$D=\{d_1,d_2,d_3,\cdots,d_n\}$。其中 $n=6763$，$d_i(1\leqslant i\leqslant n)$在集合{0,1}中取值。如果特征分项 t_i 在文档 D 中出现，即 $t_i\in D$，则相应 $d_i=1$；如果特征分项 t_i 在文档 D 中没有出现，即 $t_i\notin D$，则相应 $d_i=0$。

③用户查询可表示为 Q，其向量形式为：$Q=\{q_1,q_2,q_3,\cdots,q_m\}$。这里的 m 值理论上不受限制，但从实际出发考虑 m 应该小于等于 n，为了方便两个向量的运算，在此取 $m=n$(差项用数字0填补)。同样 q_i 在集合{0,1}中取值。如果特征分项 t_i 在文档 Q 中出现，即 $t_i\in Q$，则相应 $q_i=1$；如果特征分项 t_i 在文档 Q 中没有出现，即 $t_i\notin Q$，则相应 $q_i=0$。

④将向量 D、Q 做数量积运算，设向量 D、Q 的数量积为 R。即，$R=D\cdot Q=\{d_1,d_2,d_3,\cdots,d_n\}\cdot\{q_1,q_2,q_3,\cdots,q_m\}=d_1*q_1+d_2*q_2+d_3*q_3+\cdots\cdots+d_n*q_m$，其中 $m=n$。

i)如果 $R=0$，则说明用户检索条件在被检索的文档中不存在。

ii)如果 $R\neq 0$，下面继续判断。

计算 $\cos\alpha=(D\cdot Q)/|Q|^2$，($|Q|^2$表示向量 Q 的模的平方)。

ii)如果$0<\cos\alpha<1$，则说明用户检索条件不完全存在于被检索文档中。

如果 $\cos\alpha=1$，则说明用户检索条件完全存在于被检索文档中，但不能说明是否连续存在。要判断是否连续存在，则需要采集到被检索文档中汉字的位置关系，这可以依据数据索引结构和数据索引方式来完成，其结构图如图2所示。

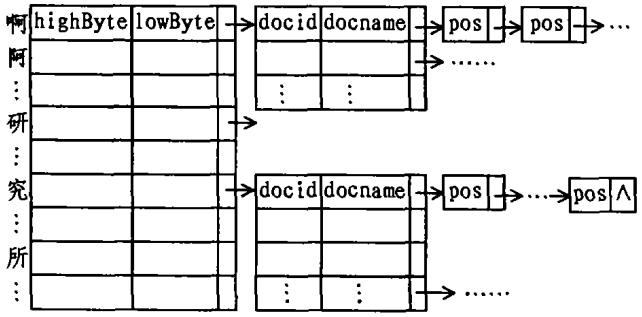

图2　数据索引结构

同样假设用户查询可表示为 $Q=Q_1Q_2Q_3\cdots Q_i\cdots Q_n$。

i)取出 Q_i 的地址链 P_{qi} 中的一个地址值 p。(一般要首先取 $i=1$，Q_1作为基点)

ii)取出 Q_{i+1}的地址链 P_{qi+1}。

iii)若 $p+2*i\in P_{qi+1}$(每个汉字由两字节组成)，继续第iv步，否则返回第1步。

iv)$i=i+1$，循环运行第ii、iii。

v)若 $i=n$，则说明成功检索到一篇目标文档，取完地址链 P_{qi} 中的所有地址值进行判断，则可以计算出用户检索条件在该篇目标文档中出现的所有次数(出现频率)。

注意，出现频率在这里是相当关键的一个参数，在“结果排序机制”中将得到应用。

对于“结果排序机制”，在这里简单分析一下。结果排序也是文本信息检索中一个重要组成部分，它直接关系到搜索引擎工具的质量，好的排序机制有助于提高用户查找信息的效率。

我们采用的排序机制主要依据两个参数：

1)文档和查询条件的相关性；

2)文档的质量(重要性)。

相关性主要依据精确匹配模型，在这个模型当中，我们能够检索出检索串在相应文档中的出现次数(出现频率)，以该参数作为标准，如果检索串在相应文档中出现次数越多，则其相关性越高。

在进行全文数据索引之前，给每篇文档赋予不同的贡献值，检索的时候我们就能够检索出相应文档的贡献值，贡献值越大，重要性越大。

定义　“相关性权值”为 R，“重要性权值”为 W，其中$0<R<1$，$0<W<1$，$R+W=1$。

假设一检索串 AB 在相应的文档 D 中出现的次数为 C，文档 D 的贡献值为 Z，则排序值 *Criterion*(作为文档排序的一种度量)为：

$Criterion=R*C+W*Z$

然后就依据 *Criterion* 值大小进行文档排序。

图2数据索引结构说明，数据索引结构分为三级，第一级为字级索引结构，包含 highByte(高字节)、lowByte(低字节)和指向下一级的结构指针，这里分为高字节和低字节是因为汉字编码为双字节编码，结构中包含6768个单元，是按照国标汉字字符集 GBK/2:GB2312实现的；第二级为文档级索引结构，包含 docid(文档编号)、docname(文档路径和文件名)两部分；第三级主要为地址链结构，用来存放汉字在文档中位置(位置用相对于文档开头的偏移量表示)。

从这整个匹配模型可以看出，用户检索条件与被检索文档匹配是采用一种精确匹配模型。查全率基本上不存在问题，查准率也有很大的提高(因为有“出现频率”这个关键参数作为保证)。

结论　以“信息检索精确匹配模型”为基础，我们实现了一个面向企业应用的搜索引擎工具，该工具在查全率方面不存在问题，在查准率方面质量也很高(当然这里要结合“结果排序机制”来完善)。缺陷是速度偏慢，这可以利用多线程并行技术来弥补，或者采用服务器群集技术来实现。

参 考 文 献

1　邹涛，王继成，等. 文本信息检索技术. 计算机科学，1999，26(9)：20～24

2　彭洪汇，林作铨. Internet 上的搜索引擎和元搜索引擎. 计算机科学，2002，29(9)：1～12

3　邢永康，马少平. 信息检索的概率模型. 计算机科学，2003，30(8)：13～17

4　庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现. 计算机应用研究，2001，10(9)：23～26

5　Yang Yiming. An Evaluation of Statistical Approaches to Text Categorization. Journal of Information Retrieval，1999，11(2)：11～14

6　孙晋文，肖建国. 自动文本分类中的智能处理技术. 计算机科学，2003，30(8)：18～20

7　Salton G，Buckley C. Term-weighing approaches in automatic text retrieval. Information Processing & Management，1988，24：513～523