

一种基于全文检索系统的文档关联研究与实现^{*}

饶 伟 郭 辉 蔡庆生

(中国科学技术大学计算机科学技术系 合肥230027)

Research on Document Relevancy Based on Full-Text Retrieval System

RAO Yi GUO Hui CAI Qing-Sheng

(Department of Computer Science and Technology, University of Science and Technology of China, Hefei 230027)

Abstract As a important application of the Full-Text retrieval system, document relevancy has powerful function. In this paper, a document relevancy method based on the Full-Text retrieval system is presented, which is deeply discussed from two aspects, content relevancy and properties relevancy. This system is proved to have good response time and precision by tests. It has great prospects in application area.

Keywords Full-text retrieval, Document relevancy, Content relevancy, Properties relevancy

1. 引言

办公自动化(Office Automation, 简称 OA)自70年代引入我国,目前已经成为一个大众化的概念。办公自动化的核心是对文档的管理,在企业和政府部门中存在着海量的文档,如何有效的对文档进行管理成为一个越来越重要的问题。

在此基础上,基于全文检索的技术也有了很大的发展空间,文档关联技术正是在这样的背景下产生的。在庞大的文档信息被储存、被查询、被检索的同时,人们也希望给这些杂乱无章的文档赋予一定的关系,使得这些文档能被更有效地利用。

目前中英文的全文检索系统已经有不少产品投入使用,这些产品有着各自的特点和不同的用户对象,有些是大型出版社的书刊检索系统,有些专门针对网页检索,然而这些全文检索系统大多只针对用户提出的关键词信息给出搜索结果,没有充分考虑到全文检索系统中文档的相关性,因此,本文提出使用文档关联算法,提高全文检索系统的性能。

目前大部分全文检索系统都不具有文档关联功能,即使有些全文检索系统这样的具有属性检索功能的系统,实现的属性检索功能也不同于本文讨论的属性关联技术,只是该技术的一部分。将标题、日期和发表时间具体化到检索内容,不仅要求用户多次检索,而且不能全面反映文档间的属性关联度,本文讨论的属性关联并不局限到某个属性,只要文档的各个属性有关系,并且其关联度的加权和大于阈值,就表示文档的属性关联。显然,文档关联算法的引入,可以提高全文检索查全率和查准率。

为此,我们开发一个能够实现对以 TXT 文本为对象的海量文档进行全文检索和文档关联的 COM 组件,能够方便地集成到办公室自动化软件系统之中,以实现用户对海量文档的检索需要和关联查阅。

2. 全文检索系统的实现

全文检索(Full-Text Retrieval)是指以全文本信息为主要检索对象,允许用户以自然语言根据资料内容而不是外在特征来实现检索的先进查询手段。全文检索有以字为单位进

行索引和以词为单位进行索引之分。按体系机构来分:有基于文件系统、基于数据库系统和两者相结合三种。

我们的全文检索系统主要负责以下两个功能:

1. 根据用户输入的检索词和限制条件,在索引文件中检索出所有满足条件的文档;
2. 对录入的新文档建立全文索引,将过期文档从索引文件中删除。

考虑到索引文件空间的使用,我们采用按关键词为单位的索引,使用文件系统和数据库系统结合的方法存储索引信息。对于每一篇新添加的文档,系统会根据分词程序得到该文档的所有关键词,以及每个关键词出现的词频和位置信息,使用压缩技术存储,系统中的关键词信息使用倒排表存储到文件系统。倒排文件是对每种属性都设立一个索引,每种属性值都对应一个索引项。将具有相同属性值的记录链接字都保存在与之对应的索引项中。倒排文件中的各个属性的索引为倒排表。在倒排文件结构中,处理布尔询问的方法是对索引项进行并、交、差的逻辑运算,然后再按结果集合中的指针去取记录。这种结构的优点在于检索记录较快,它对于任何复杂的布尔询问,均只需访问一次满足询问的记录,就能获得肯定或否定的回答,这样有效地提高了系统的检索速度。

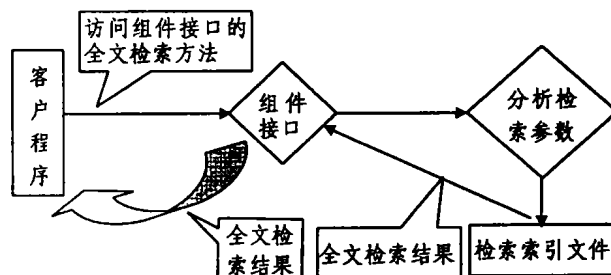


图1

索引文件分为三类:汉字、英文和数字。索引文件采用结构化存储技术,其中存在两类对象:

1. 存储对象:存放索引文件内部结构信息,可以有子对象,其子对象可以是存储对象、也可以是流对象。
2. 流对象:存放索引文件内部数据信息,即文档信息及汉

^{*} 本文获国家自然科学基金(60075015, 90104030)资助。饶 伟 硕士研究生,研究方向:全文检索和文档关联研究。郭辉 硕士研究生。蔡庆生 教授,博士生导师,主要研究方向:人工智能,机器学习,知识发现。

字、英文或数字位置信息。

三种不同的索引文件独立存储,以方便数据查询和更新,这样的存储方式同样能够提高系统的响应速度。

图1是全文检索系统的流程图。

从中可以表明,系统通过 com 接口通讯,任何的布尔询问,均只需访问一次满足询问的记录,就能获得肯定或否定的回答。

另外,系统中还存储了每篇文档的属性信息,例如,作者、出版日期、标题、出版社。这样就为文档关联中的属性关联算法的实现提供了可能,使得文档关联的准确度得到提高,并且增强了系统的功能。

3. 文档关联算法

正如本文第三部分指出的,传统的全文检索系统,只是对文档内容中出现的字或词进行检索,用以发现是否有用户需要的关键词,并将其作为结果提交给用户,当然也有部分系统用到了一些中文语义理解的技术,如上文提到的 ISEARCH 全文检索等系统,加入了同音、同义词的检索功能;但这些依然是针对关键词本身的,不能够提高到文档的角度考虑。

正是考虑到传统全文检索系统在这方面的局限,本文提出了使用文档关联算法提高全文检索性能的方法。使用关键词索引技术得到的文档,以及这些文档的关联文档,构成用户检索的结果。文档关联算法分为内容相关和属性相关两部分实现。

(1) 内容相关度计算

a. 如何衡量相关度 Sim 的大小:

相关度取值在 $[0, 1]$ 区间, $\text{Sim} = 0$, 两篇文档完全不相关; $\text{Sim} = 1$, 两篇文档完全相关。

以索引信息作为衡量内容。

b. 内容相关度计算公式:

文档可以看作由关键词信息组成的集合,在这里我们扩展集合定义,即允许集合中含有相同元素,我们将这样的集合称为扩展集。定义:文档 A 记为扩展集 $A^* = \{a_1, \dots, a_1, \dots, a_n, \dots, a_m\}$, a_m 为关键词, a_m 重复的次数即为在文档 A 中该关键词的词频, $m = 1, \dots, n$ 。

设文档 A、B 对应的扩展集为 A^* 、 B^* , 则文档 A、B 的相似度 $\text{sim}(A, B)$ 定义为:

$$\frac{\text{card}(A^* \cap B^*)}{\text{card}(A^* \cup B^*)}$$

注:扩展集的交集和并集的定义与基本集合的定义大致相同。

输入:新文档关键词;输出:关联文档信息

```
begin
    按关键词计算数据库中每篇文档的分数(sim);
    文档汇总,将分数大于某一阈值的文档加入关联文档;
    返回;
end
```

算法评估:

计算公式满足定义: $\text{Sim} = 0$, 两篇文档完全不相关; $\text{Sim} = 1$, 两篇文档完全相关。

内容关联计算不受文档添加先后次序的影响。

内容关联只需比较两篇文档的字,提高了内容相关度计算的效率。并且,算法的时间复杂度为 $O(n)$, n 为数据库中文档的总数。另外,利用全文检索功能,算法的时间复杂度可降低为 $O(k)$, $k < n$ 。

(2) 属性相关度计算

属性相关度的计算与内容相关度的计算类似。我们可以

将每个属性值看作一个句子。因此,属性相关度的计算可以采用与内容相关度计算相同的方法。由于不同属性条目的文档其属性相似度难以定义。因此,属性相关度计算只对具有相同属性的条目进行比较。因此,属性相关度算法的时间复杂度为 $O(k)$ (k 为具有相同属性条目的文档数)。

首先,判断两篇文档的属性 $(X_1, X_2, X_3 \dots)$ ($Y_1, Y_2, Y_3 \dots$) 中有无相同部分(这里 X_i, Y_i 是属性的类别);

其次,分别计算这些有类别相同的属性的相似度 $\text{Sim}(i), \text{Sim}(j), \text{Sim}(k) \dots$

最后,计算得到属性关联度 Sim-properties 为

$$\frac{K(i) * \text{Sim}(i) + K(j) * \text{Sim}(j) + K(l) * \text{Sim}(l) + \dots}{K(i) + K(j) + K(l) + \dots}$$

其中 $K(i)$ 是各个属性在关联度计算中的权重,根据系统的初始设置决定,一般情况设为 1。属性相关度和内容相关度的复合就是整个关联文档的相关度。记为:

$$k * \text{Sim-content} + (1-k) * \text{Sim-properties}$$

$$(0 < k < 1)$$

其中 k 是内容相关度和属性相关度的复合参数,可以由系统指定,也可以允许用户修改。计算得到的文档相关度 Sim 在 $[0, 1]$ 之间,系统取相关度大于某个阈值(例如 0.5)的文档为关联文档,存储到数据库中;用户在检索某个关键词时,被检索到的文档,以及关联文档,将会按照关键词出现的次数 $\text{Count}(X)$ 以及文档的关联度的复合,进行排序,得到返回结果。

$$\text{Count}(X) + \text{Sim}$$

4. Com 组件的实验及结果分析

我们就文档关联功能的使用做了如下实验(测试环境:配置:CPU:P3 800, RAM:256M, 软件:Windows2000 Server, SQL Server 2000)

4.1 系统添加、删除时间测试

文档添加是按文档号顺序,随着库存文档数量的增多,每 k 的添加时间呈增长趋势。个别文档由于文档容量过大,将使添加时间有较大增长。

文档删除是倒序进行的,基本在 1.0~3.0(秒/ k)

添加文档:内存使用:10.8M~15.7M, CPU 占用:4~35%

删除文档:内存使用:12.9~13.8M, CPU 占用:5~34%

表1 全文检索系统添加删除文档时间

文档号	添加时间(秒)	删除时间(秒)	文档大小(K)
1	8(8)	1(1)	1.021
2	13(10)	2(2)	1.56
3	31(22)	2(2)	2.09
4	23(15)	2(2)	1.80
5	31(22)	3(2)	2.09
6	45(32)	3(2)	2.83
7	30(9)	2(2)	1.35
8	119(51)	6(3)	5.95
9	83(29)	4(2)	2.71
10	179(72)	7(3)	7.69
11	72(9)	2(2)	1.30
12	165(44)	5(2)	3.27
13	95(22)	2(2)	1.04
14	130(23)	3(2)	1.97
15	97(16)	2(2)	1.52

(下转第135页)

心坐标为(149.0000,149.0000),半径为69.0000(均以像素为单位)。实验中给出了对原始图像进行处理的过程和定位结果,下面我们以列表的形式给出测量算法对于仿真图像分别在无噪声、高斯噪声($\mu=0, \sigma=0.02$)、脉冲噪声(脉冲概率为20%)和乘性噪声(方差为0.1)下的各项指标和定位结果。

表1 仿真实验数据

图像 指标	原始图像	高斯噪声 图像	脉冲噪声 图像	乘性噪声 图像
初始点	(137,217)	(139,217)	(99,195)	(127,214)
进缘像素数	388	388	389	389
中心坐标	(149.0000, 149.0000)	(148.9860, 149.0762)	(149.0826, 149.1241)	(148.9117, 149.0904)
半径	68.7800	68.8425	68.6892	68.7528
中心偏差	(0.0000, 0.0000)	(0.0140, 0.0762)	(0.0826, 0.1241)	(0.0883, 0.0096)

从表1中数据可以看出,对于原始仿真图像,定位目标中心没有产生偏差。表1实验数据仍然表明系统算法对于各种噪声不敏感。虽然噪声的加入使得实际的边缘点数发生细微变化,但这并不影响定位精度,在常见的三种噪声干扰下,目标中心的偏差总是不超过0.15个像素。同时也注意到,系统检测目标半径的偏差还是较大,最大是接近0.3个像素。然而由于ICF图像测量系统的特殊要求,在实际应用中我们更为关心的是圆形目标物体的中心坐标,半径偏差不是我们所考虑的,它也不会对后面的控制产生影响。

结论 在介绍图像测量系统原理的基础上,提出了一种新的系统设计方案。对新系统所涉及的各种算法进行了深入分析,并就算法在测量系统中的应用做了介绍。仿真实验结果

表明了新系统的合理性,可行性和先进性。

参考文献

- 1 吴晓波. 图像边缘特征分析. 光学精密工程, 1999, 7(1)
- 2 Canny J. A Computational approach to edge detection. PAMI, 1986, 8(6): 699~698
- 3 张小洪, 杨丹, 刘亚威. 基于 Canny 算子的改进型边缘检测算法. 计算机工程与应用, 2003, 12(待发)
- 4 李云, 鲍苏苏, 杨璐. 二值化图像中物体区域的选定及外边缘跟踪技术. 华南师范大学学报(自然科学版), 2002, 8(3): 27~29
- 5 周丽娟, 柳池, 郝瑛. 图像边界自动跟踪的边缘提取算法. 哈尔滨理工大学学报, 1997, (5): 50~53
- 6 惠毓伦, 施泽生. 任意形状二值线图边缘追踪的实现. 青岛大学学报(自然科学版), 1994, 7(2): 58~63
- 7 黄宏一, 丘悦, 范滇元. 靶瞄准中的图像处理. 中国激光, 1998, A25(7): 649~652
- 8 王建国, 浦昭邦, 刘国栋. 提高图像测量系统精度的细分算法的研究. 光学精密工程, 1998, 6(4): 44~50
- 9 Lyvers E P, Mitchell O R. Subpixel measurements using a moment-based edge operator. IEEE Trans. Pattern Anal. Machine Intell, 1989, 11(12): 1293~1308
- 10 王建国, 浦昭邦, 尹继学. 空间矩亚像素细分算法的研究. 光学技术, 1999(4)
- 11 刘亚威, 杨丹, 张小洪. 基于空间矩的亚像素边缘定位技术的研究. 计算机应用, 2003, 23(2): 47~49
- 12 杨耀权, 施仁, 等. 用 Hough 变换提高激光光斑中心定位精度的算法[J]. 光学学报, 1999, 19(2): 1655~1660
- 13 Thomas, Chan. A Simple Approach For The Estimation Of Circular Arc Center And Its Radius. CVGIP, 1989, 45: 362~370
- 14 孔兵, 王昭, 谭玉山. 基于圆拟合的激光光斑中心检测算法. 红外与激光工程, 2002, 31(3): 275~279
- 15 孔丹. 基于空间矩的灰度边缘亚像素度量精度分析. 红外与激光工程, 1998, 27(2): 6~10

(上接第79页)

添加删除时间里,前者使用了文档关联算法,括号内则是未使用文档关联算法测出的时间,可见文档关联算法的使用,对添加文档有了较大的影响,而对于删除文档则影响不大。

4.2 查询结果测试

在系统中添加了200篇文档后,我们对系统测试结果如下。

表2 全文检索系统查询结果

编号	结果1	结果2	相关文档数
1	6	6+3	5+3
2	2	2+4	2+4
3	10	10+9	9+8
4	27	27+8	25+7
5	12	12+1	11+1
6	5	5+1	4+1
7	7	7+2	6+2
8	3	3+0	3+0
9	5	5+2	5+2
10	3	3+1	3+1
合计	80	80+31	73+29

这个测试是针对用户给出的关键词信息进行搜索,对于

使用和不使用文档关联算法,得到了两个结果(搜索得到的文档数),对于使用文档关联算法的测试,以形如:X+Y的方式表示,其中X是直接搜索关键词得到的文档数,而Y是这X篇文档的关联文档数(去掉了X篇文档中的重复部分)。

表2显示,系统在使用了文档关联算法后,增加了29/80=36.25%关联文档,显著提高了系统的查全率,查准率也得到了提高。

小结 本文介绍了一种基于全文检索的文档关联算法的实现,提高了全文检索系统的查全率和查准率,尤其是属性关联度的应用,提高了关联度计算的准确性,而且丰富了关联度计算的方式,目前该系统已经投入使用。

参考文献

- 1 曹元大,等. 中文 Web 文档全文检索系统的设计及实现. 北京理工大学学报, 2002
- 2 Mladenec D, Stefan J. Institute Text-learning and Related intelligent Agents: A Survey IEEE 1999
- 3 范焱, 郑诚, 王清毅, 蔡庆生, 刘洁. 用 Naïve Bayes 方法协调分类网页. 软件学报, 2001
- 4 Lawrence S, Bollacker K, Gils C L. Indexing and Retrieval of Scientific Literature 1999
- 5 William Hersh, Paul Over TREC-9 Interactive Track Report 2001