DF 还是 IDF?主特征模型在 Web 信息检索中的使用*

张 敏[1,2+], 马少平[1,2], 宋睿华[1,2]

[1](清华大学 计算机科学与技术系,北京 100084)

[2](清华大学 智能技术与系统国家重点实验室,北京 100084)

DF or IDF? On the Use of Primary Feature Model for Web IR

ZHANG Min[1,2+], MA Shao-Ping[1,2], SONG Rui-Hua[1,2]

[1](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[2](State Key Laboratory of Intelligent Technology and Systems, Tsinghua Univeristy, Beijing 100084, China)

+ Corresponding author: Phn: +86-10-62783191, E-mail: z-m@tsinghua.edu.cn, http://www.tsinghua.edu.cn

Received 2004-04-14; Accepted 2004-11-22

Zhang M, Ma SP, Song RH. DF or IDF? On the use of primary feature model for Web IR. *Journal of Software*, 2005,16(5):1012−1020. DOI: 10.1360/jos161012

Abstract: In Web information retrieval (IR), input queries are too short and fuzzy to describe user request, which leads to mismatch problem between user query and the documents full of redundancy and noise. This paper first studies the feature of web documents information, proposes concepts of primary feature word, primary feature field and primary feature space (PFS). Then a new PFS query term weighting scheme has been proposed, which takes document frequency (DF) into account instead of traditional IDF factor. Finally, a combination strategy of term weighting is given. Using this PFS Model, three groups of experiments have been performed on 10G and 19G large scale Web collections with TREC9, TREC10 and TREC11 standard tests of Web tracks. Comparative studies indicate that the new DF-related PFS term weighting improves system performance consistently and effectively in terms of recall, top n precision and mean average precision. At most 18.6% improvement has been made.

Key words: Web IR, primary feature model, term weighting, document frequency

摘 要: Web 信息检索的难点之一就是简短模糊的用户查询与存在大量冗余和噪声的文档之间的不匹配.对 Web 文档信息特征进行分析,提出 Web 文档主特征词、主特征域和主特征空间的概念,在该空间上使用文档频度 DF 信息而非传统上的 IDF 信息进行权值计算,并给出一个改进的相似度计算模型.使用该模型在 10G 和 19G 的两个大规模 Web 文档集合上进行了 3 组标准测试.比较实验表明,与传统 IDF 思想相比,在各项评价指标上,DF 相关的主特征权值计算方法都能始终较大幅度提高系统性能,最大达到 18.6%的性能改善.

关键词: Web 信息检索;主特征模型;权值计算;文档频度

中图法分类号: TP309 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant Nos.60223004, 60321002, 60303005 (国家自然科学基金),and the Chinese National Key Foundation Research & Development Plan Grant No. 2004CB318108 (国家重点基础研究 973)

作者简介: **张敏**(1977−),女,宁夏银川人,博士,讲师,主要研究领域为信息检索,机器学习;**马少平**(1961−),男,教授,博士生导师,主要研究领域为汉字识别,古籍数字化,信息检索;**宋睿华**(1978−),女,硕士生,主要研究领域为信息检索.

信息检索的关键,在于将用户空间需求信息的描述与文档空间已知信息的描述匹配起来.但是当前的信息检索应用中,用户的需求描述往往模糊而简短.已有的用户行为研究表明,在 Web 搜索环境下,用户键入的查询语句通常为 1~3 个单词,同时用户很少考虑如何精确地表示查询,有 78%的人不会根据结果修改他们第一次提交的查询语句[1−3].另一方面,一个 Web 文档通常由几千个特征项来描述.因此,信息检索的难点之一就是简短模糊的用户查询与存在大量冗余和噪声的文档之间的不匹配问题[4].解决该问题可以从文档和用户查询两个角度入手.本文从文档空间的角度开展研究,对 Web 文档信息特征进行分析和研究,寻求更合理的文档特征表示及特征权值计算方法,从而使文档空间的信息描述更精确,以改进 Web 信息检索的效果.

对于传统的文本信息检索来说,从一篇文档中提取出更重要的部分作为文档的特征是很困难的.因为首先目前自然语言理解技术还不能达到实际要求,对文档不同部分的重要性判断无法起到指导作用;其次由于不同类型文档在书写和表达上的差异,普通文档的自然段落结构也不能用来区分特征[5];最后普通文档不具有任何附加结构信息.因此,这一想法在传统的文本信息检索中一直无法实现.

而 Web 信息检索中 Web 文档的特殊性为我们提供了解决这一问题的途径.在 Web 中,网页作为文档,除了普通文档具有的内容信息以外,还有一些结构性的特征标记来表示不同的部分,通常称为域(field).它们虽然都用来描述网页的内容,但是重要性却有区别.例如,直观地说,出现在题目(〈TITLE〉),各种大小标题(一号标题〈H1〉,二号标题〈H2〉,…)以及强调性的文本(粗体〈B〉,…)中的词,可能在表现力上比正文内容更重要.这些特征都为 Web 文档信息的特征提取提供了有利条件,也成为本文工作的基础.

本文第 1 节介绍相关研究工作.第 2 节提出主特征词、主特征域和主特征空间的概念.第 3 节对基于主特征空间的文档权值计算方法进行描述.第 4 节给出实验结果并进行分析.最后总结本文的研究工作.

1 相关研究工作

网页中的不同域在描述网页的主题和内容上有不同的重要性,这一特点较早就被人们注意到了.目前为止这方面研究可以分为两类:

一类是先得到初次检索的结果,如果某个文档的特殊域中出现了用户的查询词,那么把该文档的结果排名提前.一些搜索引擎通过这种方法来改进相关文档的排序.例如在 AltaVista(Digital Equipment Corporation, ALTA VISTA: Main Page, http://altavista.digital.com/cgi-bin/query/)和 Yahoo(Yahoo Inc., Yahoo Search, http://www.yahoo.com/search.HTML)中,如果一个网页的题目中出现了用户查询中的某个词或者短语,那么这个网页的相似度评分会被提高[6].Lycos[7]则在排序函数中引入了位置信息,考虑用户查询中的特征词在网页的题目、正文或者文档前 100 个最相关的词的集合中出现的情况.

另一类则是把 HTML 文档结构中不同的域分别进行考察[8],通过增加词频因子 *tf* 的权值来区分不同域的作用,是一种简单加权的方法.Cutler 对这种增加特定域中词项的 *tf* 权值来改进检索的效果进行了比较研究[9].他把 HTML 的记分为 6 类:纯文本(plain text),题目(title),大标题(H1~H2),小标题(H3~H6),强调的文本(strong)以及链接文字(anchor text).研究结论是链接文字和强调的文本两种特征应赋以更高的 *tf* 权重.

在 TREC9(the 9th Text Retrieval Conference, 2000)中,Information Space 系统对使用题目和 H1~H3 的标题文本改进检索做了研究[10].研究中只对这 4 个标记标引出的文本建立索引,并得到了更高的检索精度.他们的研究结论是:对网页的题目或者其他的关键标记(key tags)中的内容应给以更多的考虑.

一般的信息检索中,通过两个因素来影响文档与用户查询之间的相似度计算[11]:(1) 在文档中出现的词频越高,则这个词越重要,应给以更高的权重,即 *tf* 因子;(2) 包含该特征词的文档数越多,则这个特征词越不重要,这也就是目前被广泛应用的 IDF(inverse document frequency)因子.现有的所有针对 HTML 文档结构信息来改进检索的研究,都是通过前者——对词在网页的不同域中出现的 *tf* 因子赋不同的权重来进行计算的.对于普通文档,IDF 的思想从本质上表达了一个特征项在区分不同文档方面的重要性.但是对于网页文档结构中的一些特殊域,例如粗体字部分以及网页的题目等,情况则有所不同.本文提出主特征域的概念,并在其构成的主特征空间上从第 2 个因素着手,使用与传统 IDF 思想不同的 DF 相关的相似度计算方法.

2 主特征词、主特征域与主特征空间

在 Internet 上,虽然网页作者不同,但是有一些词经常被大多数作者用来突出表示他们的网页,这些词通常属于某个特定的标记.它们可以分为两类:一类是功能性的或者问候式的语言,例如“版权所有”,“欢迎来到 xxx 的主页”等;另一类则是用来强调网页内容的,例如正文中被加粗的文字等.第 1 种文本在信息检索中很难带来更多信息,而第 2 类突出表示的词,则反应了 HTML 文档中更能突出内容或主题的词,在信息描述上具有更强的表现力,因而也是本节的研究中所关心的部分.

定义 1. 在整个 Web 中,有一些词在被使用时,被不同的作者认为具有更丰富的主题或内容表现力,因而经常在一些特定的域中出现.这样的词被称作主特征词(primary feature term,简称 PFT),相应的域被称作主特征域(primary feature field,简称 PFF),整个网络中,所有网页的主特征域构成了主特征空间(primary feature space,简称 PFS).

主特征域中的每个特征项都是主特征空间中的一维.整个文档集合中的词在这个空间中的分布是不均匀的.空间中的每一维都带有一定的特征权值,可以用该词在整个文档集合构成的主特征空间中被不同作者使用的次数来决定.即那些被大多数作者都认为具有主题或内容表现力的词更有可能作为主特征词,因此它们作为主特征的权值也比较大.而只被少数作者在主特征域使用的词,则可以认为不具有代表性.当文档集合确定以后,主特征空间及空间中每一维的权重也就确定了.

容易看到,在主特征空间思想下建立模型,主特征域的选取是关键问题之一.网页文本与传统文本的一个重要区别就是网页文本具有 HTML 标记信息.这些标记信息在表达一定的结构信息以外(例如列表、分段等),也能在一定程度上表现出主题和内容的不同重要性.例如更大的字体、强调的文字例如粗体等由于视觉效果突出而可能引起读者特别的注意,因此网页的作者往往通过这些标记及相应的信息来表达文档的重要或中心内容.结合前人对网页文档中不同域的考察结果,在实验中,我们重点考察以下 HTML 标记对应的内容作为主特征域的效果:粗体字(〈b〉)、题目(〈Title〉)、标题(〈h1〉等)和斜体字(〈i〉).其中通常认为与正文部分相比,网页作者可能使用粗体字和斜体字来突出表现文档的中心或者重要内容,而标题和题目部分则具有更多的主题表现力.将上述不同域作为主特征域进行检索的实验效果比较及分析将在后面的第 4.3 节中给出.

3 基于主特征空间的文档权值计算方法

通常,一个用户查询与文档之间的相似度关系,可以分解为文档与查询中的每一个特征项之间的相似度之和.而用户查询中,并不是所有的词都具有同等重要的作用.因此相似度的计算可以表示为

$$sim(D,Q)=\sum_{t\in Q}\lambda_t\cdot w_t\cdot sim(D,t) \tag{1}$$

其中 Q 是用户查询,t 为查询中的一项,w_t 为查询项本身的权重,而 $sim(D,t)$则是该查询项与文档之间的相似度.对于一次检索来说,λ_t=1.对相关反馈得到的扩展查询,则可以对原查询项和新扩展出来的查询项给以不同的权值.

3.1 传统的查询项权重计算方法区别

在 Robertson 和 Sparck Jones 提出的著名的概率检索模型中,用户查询与一篇文档的相似度可以用下面的式 2(称为 BM2500 公式)来表示[12]:

$$sim(D,Q)=\sum_{t\in Q}w^{(1)}\frac{(k_1+1)tf}{(K+tf)}\frac{(k_3+1)qtf}{(k_3+qtf)},\ K=k_1\left((1-b)+b\times\frac{dl}{avdl}\right) \tag{2}$$

其中 tf 和 qtf 分别为查询项 t 在观察文档和查询中出现的次数,dl 和 $avdl$ 分别为观察文档的长度和平均文档长度,通常以词或者词组作为单元来表示.$w^{(1)}$就是查询项本身的权重,一般被称作 Robertson/Sparck Jones(RSJ)权重因子,可以用下面的式(3)来计算:

$$w^{(1)}=\log\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \tag{3}$$

其中 N 是集合中的文档总数,n 是出现该项的文档数,R 是与该查询主题相关的文档数,r 是相关文档中含有该检索项的文档数.通常在第 1 次检索时,因为缺少相关性信息,R 和 r 取值为 0,这时查询项的权重因子就简化成为文档集合频度权重:

$$w^{(1)} = \log \frac{N - n + 0.5}{n + 0.5}.$$

这就与向量空间模型中的 IDF(inversed document frequency)因子反映了同样的思想:出现一个词的文档总数越多,则这个词越不重要.

3.2 引入主特征域的查询项权重计算

当引入主特征域的概念后,一个文档可以分为两个部分:文档的正文部分,以及文档的主特征域部分.这两部分内容信息的重要性上是有区别的,其中主特征域中的文字具有更丰富的主题和内容表现力.因此,考虑到网页文档主特征域所携带的信息,用户查询中的词项本身的权重 w_t(见式(1))应该由两部分组成:该查询项作为网页普通文本(下面称作正文)所具有的权重;以及该查询项作为主特征词所具有的权重.

如果沿用传统的特征词权重计算方法,则文档与查询的相似性可以用这样的公式来计算:

$$sim(D,Q) = \sum_{t \in Q} (\lambda w_{body}^{(1)} + (1-\lambda) w_{pff}^{(1)}) \frac{(k_1+1)tf\,(k_3+1)qtf}{(K+tf)(k_3+qtf)} \tag{4}$$

其中 $w_{body}^{(1)}$ 和 $w_{pff}^{(1)}$ 分别表示查询项在正文和主特征域上的 Robertson/Sparck Jones 权重因子.

3.3 主特征域上的改进权值计算

在前文的定义 1 中指出主特征域中的主特征词具有更丰富的内容或者主题的表现力.因此主特征域与正文部分不同,它含有更丰富的信息,是作为文档的一部分特征而出现的.并非文档中所有的词都可以作为主特征词.在整个文档集合构成的主特征空间中,只有那些被大多数作者都认为可以作为特征的词才有意义,因此它们作为主特征的权值也比较大.而只被少数作者在主特征域使用的词,则可以认为不具有代表性.

因此,主特征词的权值可以表示为已知文档集合中该特征项在主特征空间上的词频来衡量:出现的词频越高,则该特征项就越多地被整个文档集合的作者使用作为主特征词,更有普遍意义.即特征词 i 的重要性因子 D_i 可以表示为

$$D_i = \sum_{k=1}^{N} tf_{ik}, i = 1,2,...,n \tag{5}$$

其中 tf_{ik} 是特征项 i 在文档 k 的主特征域上出现的频度,N 为集合中的文档总数,n 为主特征空间的维数,它等于文档集合中所有的主特征域中不同的词项的个数(通常是过滤了停用词 stopwords 之后的结果).

进一步地,因为 Web 中的网页通常由不同的作者所建立,这些网页作者的书写习惯,表达方式都各不相同.而主特征空间表达的是整个网络对于哪些词可以被用来更准确地描述网页内容或者主题的普遍观点,因此有必要把同一个作者的观点对主特征空间特征词权值的影响进行归一化处理.也就是说:令

$$tf_{ik} = \begin{cases} 1, & \text{if term } i \text{ occurs in the primary field of doc k} \\ 0, & \text{otherwise} \end{cases}.$$

因此特征词的重要性因子的计算可以简化为

$$D_i = \sum_{k=1}^{N} tf_{ik} = n_{ip},$$

其中 n_{ip} 就是在整个文档集合的主特征域中包含词项 i 的文档的个数.对这一重要性因子进行平滑处理,得到主特征词的权值计算见式(6):

$$w^{(2)} = \log(n_{ip} + 1) \tag{6}$$

其中 n_{ip} 是主特征域中含有该查询项的文档的个数,即文档频度 DF(document frequency),它与传统模型中使用的 IDF 思想在本质上完全不同,表示了该查询项在主特征空间的密度.我们把 $w^{(2)}$称为主特征域权重因子.

由此,得到考虑主特征域的用户查询与文档的相似度的改进计算式(7),其中λ是网页正文内容的影响因子:

$$\left.\begin{aligned}\sum_{t\in Q}(\lambda w^{(1)}_{body}+(1-\lambda)w^{(2)}_{pff})\frac{(k_1+1)tf}{(K+tf)}\frac{(k_3+1)qtf}{(k_3+qtf)}\\ w^{(1)}_{body}=\log\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)}\\ w^{2}_{pff}=\log(n_{ip}+1)\end{aligned}\right\} \quad (7)$$

其中 $w^{(1)}_{body}$ 是正文中的 IDF 相关的权值因子,与传统的计算方法相同, $w^{(2)}_{pff}$ 是主特征域上 DF 相关的权值因子.当 λ=0 时,查询项的权重因子完全由该项的主特征信息权重因子决定,而 λ=1 时这个公式就退化为传统的概率模型相似度计算方法.

4 实验结果及分析

4.1 实验数据和条件

对该基于主特征空间的 DF 因子相关的权值计算方法的考察,使用了 3 年的 TREC Web Track 的测试数据.TREC(text retrieval conference)是文本信息检索领域影响最大最著名的国际标准评测会议,提供标准测试数据集.其中 TREC9(2000)和 TREC10(2001)的 Web Track 使用 10G 大小的 WT10g 数据[13],包括 169 万篇 Web 文档.TREC11(2002)使用约 19G 的.GOV 数据集[14],包括 125 万篇文档.这两个集合分别有 50 个测试查询以及相应的相关文档集(称为 qrels),都是从 Internet 上抓取的真实网页文档.

对于每个查询,系统检索返回前 100 篇文档,分别用检索到的相关文档数(反映了前 100 篇文档的召回率,后文中用*#rel_ret* 表示),平均精度(后文中用 *ave_P* 表示,TREC 评测中的标准评价指标),前 5 篇和前 10 篇文档的精度(分别表示作 *p*@5 和 *p*@10)4 个指标来进行评价.

实验的主要目的是对基于 Robertson/Spack Jones 权值因子(后简写作 RSJ 权值)的相似度计算方法(式(4))和我们提出的基于主特征空间权值(后简写作 PFS 权值)的相似度计算方法(公式 6)的检索效果进行比较.其中前者使用了传统的 IDF 的思想,后者则在主特征空间上使用与传统方式完全不同的 DF 因子.在实验中,主要考察以下特征域:粗体字(bold)、题目(Title)、标题(head 〈h1〉~〈h3〉)和斜体字(〈i〉).一方面在前人的工作中曾经对这些标记对应的内容进行一些考察(主要集中在对这些域赋以不同的权值以加强其 *tf* 值的作用);另一方面在文本提出的主特征域和主特征空间思想中,通常我们认为网页作者可能使用粗体字和斜体字来突出表现其中心或者重要内容,而标题和题目部分则具有更多的主题表现力.

4.2 PFS权值计算方法的有效性实验

图 1 和图 2 中显示了在 TREC10 Web Track 的 50 个查询测试集上使用粗体字作为主特征域时的系统检索性能,分别考察检索返回的相关文档数及平均精度.

图中当 λ=1 的时候,表示不考虑主特征域的特殊作用,整个文档内容都作为普通正文.注意到,如果区分主特征域,并且在主特征域上使用传统的 RSJ 权重计算相似度,则检索性能变化曲线随着 λ 的增加而呈单调增长;当 λ=1 的时候,检索到的相关文档数最多,这时,粗体域的影响因子为 0.即使用传统的 RSJ 权值计算方法时,考虑粗体域的特殊性则只能为检索召回率性能带来损害.

但是,如果使用新的 DF 相关的 PFS 权值计算方法,则粗体字部分表现出很强的网页内容信息表达能力,从而有效提高检索性能.从图 1 中可以看到,如果使用适当的网页正文内容影响因子 λ,则本文提出的基于 DF 信息的主特征空间权重因子的相似度计算方法(PFS 方法)的检索性能,总是优于传统的基于 IDF 信息的相似度计算方法(RSJ 方法).且 λ 的适用范围很广,只要影响因子不小于 0.3,就能检索到更多相关文档.图 2 中,在平均精度上的实验效果与图 1 的结论类似.只要用传统的 RSJ 权值计算方法对粗体域的内容进行加权,则检索的精度都会下降.但是使用 PFS 权值计算方法,则会在一定取值范围内改善系统的检索精度.

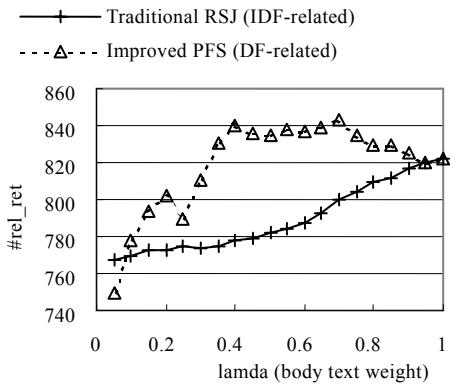

Fig.1 Performance on TREC10: Returned relevant documents number (PFF: Bold text)

图 1 TREC10 上的检索性能:返回的相关文档数(主特征域:粗体字)

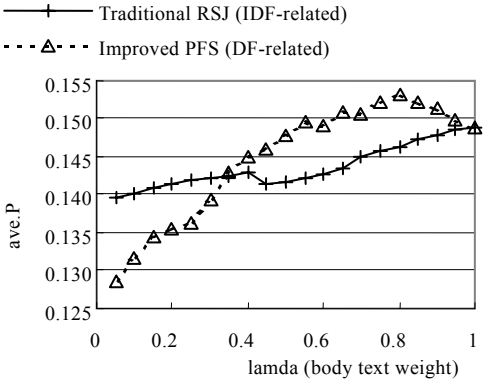

Fig.2 Performance on TREC10: Mean average precision (PFF: Bold text)

图 2 TREC10 上的检索性能:综合平均精度(主特征域:粗体字)

上述两个两组实验结果反映了在检索精度和召回率上检索性能的提高,也验证了文中提出的基于 DF 因子的主特征空间及改进的相似度计算模型的有效性:区分特征域,使用 DF 相关的特征项权值计算方法,其性能优于基本的不区分特征域的方法,也优于区分特征域但使用传统的 IDF 相关的权值计算方法.

图 3 和图 4 中显示了同样以粗体字部分作为主特征域使用 PFS 权值在 TREC9 测试查询上的检索效果,分别用检索到的相关文档数和前 10 篇文档的精度来评价.

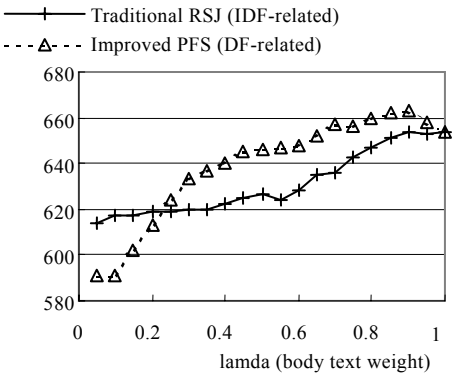

Fig.3 Performance on TREC9: The number of relevant documents returned (PFF: Bold text)

图 3 TREC9 上的检索性能:返回的相关文档数(主特征域:粗体字)

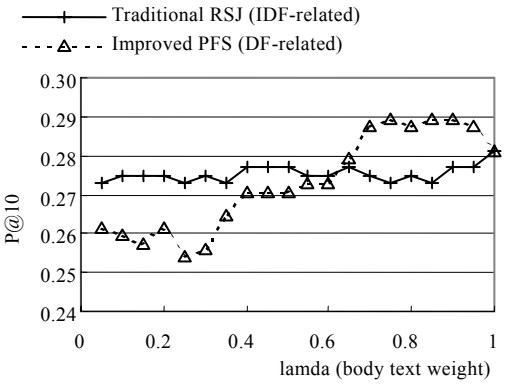

Fig.4 Performance on TREC9: Average precision of top 10 documents (PFF: Bold text)

图 4 TREC9 上的检索性能:前 10 篇结果文档的精度 (主特征域:粗体字)

可以看到图 3 和图 1 有非常接近的曲线变化趋势,这表明了使用 PFS 权值方法进行检索的结果具有一致性.图 4 中的比较结果则表明了 PFS 权值方法不仅能够提高检索的平均精度,同时也能够提高返回结果中排在前面的文档的相关性,这一部分文档也正是用户获取信息的最重要的来源.

由于 TREC9 和 TREC10 使用了同样的文档数据集和不同的测试查询,因此从图 1~图 4 的比较实验结果的一致,说明了使用 DF 相关的 PFS 权值计算方法在以粗体字作为主特征域时的可靠性和有效性.

4.3 主特征域选取效果的实验及分析

第 4.2 节中给出的实验结果均以粗体字部分作为文档的主特征域.在本节的实验中,考察选取不同的主特征域对检索结果的影响.表 1~表 3 分别列出了以题目(⟨title⟩)、斜体(⟨i⟩)、大小标题(⟨h1⟩,⟨h2⟩,…)作为主特征域时,在 TREC10 和 TREC9 数据集上,使用传统的 RSJ 权值计算方法和文中提出的 PFS 权值计算方法所能达到的

最好检索性能比较.结果表明使用题目作为主特征域,在几乎所有的评价指标上,PFS 方法都同样能够达到一定的性能提高,但是提高的幅度小于使用粗体字作为主特征域的结果.而使用斜体和标题部分作为特征域的效果相似:都能够在一定程度上提高检索的召回率(即返回结果的相关文档数),但是在平均精度、前 *n* 选精度上的效果与传统方法几乎相当,没有带来较大幅度的性能提高.

对结果进行分析如下:

首先,用粗体字作为主特征域的检索效果优于使用题目作为主特征域的结果是合理的.主要原因有二:

(1) 观察 HTML 网页发现,大量的网页文档的题目无法很好地描述网页的主要内容,甚至有很多网页的题目是由 HTML 编辑工具自动生成的.尤其同一个站点之下的网页,往往使用相同的题目,并且经常是该站点的名称或者欢迎信息.但是粗体部分则不同.粗体字通常都是处在文档的正文中间的部分,是被网页作者强调说明的内容,因此带有具体的文档内容信息,也更能够表达网页内容上的主要特征.

(2) 一个网页最多只有一个题目,且非常简短甚至只是一个短语而非整句,极易发生与查询词项不匹配的情况,这时这种基于主特征域的方法则无法对改进系统性能起到作用.但是大多数情况下网页中都有多处被加粗了的部分,信息比较完整且容易匹配.因此网页题目在内容上的描述能力差于粗体部分.

其次,用斜体字作为主特征域与传统方法相比效果相当的原因主要在于斜体字部分内容的复杂性.仔细观察和分析网页文档,不难发现,网页作者使用斜体字主要出于两种原因:(1) 用来强调表现某个概念或者某部分内容,这种情况下斜体字部分所起到的作用与主特征域的定义一致:可以用来突出表现网页的内容;(2) 用来表示一些不重要甚至于文档中心内容无关的信息,例如在网页的下部给出版权信息(copyright),或者作者的联系方式等,这时斜体字部分无法起到主特征域的作用.由于上述两种情况都较大量存在,因此表现出的效果就是对检索没有特别的作用.如果能够对网页进行进一步处理,尽量过滤掉第 2 类信息,则斜体字部分有希望带来一定的性能提高.

最后,用各大小标题部分作为主特征域也没有带来性能提高,主要原因在于网页作者对标题的使用缺乏一致的观点.有大量网页作者不对网页内容添加大小标题,因此抽取出来的标题部分内容少于粗体部分的内容,也少于网页题目(title)的内容,数据稀疏性更加严重,丢失了大量信息.另外,即使使用了标题标记,不同的网页作者对标题的使用观点也不一致,有的将网页中最大的字体设为⟨h1⟩,有的则使用⟨h3⟩作为最大标题,从而使得更多的小标题内容无法用 HTML 文档中的“head”标记突出出来.总之,分析得到,由于网页文档存在的书写不规范性,使得标题无法起到更好的作用.

Table 1 Comparison of best retrieval performances on TREC9 and TREC10 data, using Title as PFF

表 1 使用题目(⟨title⟩)作为主特征域在 TREC9 和 TREC10 上的检索最佳性能比较

	TREC10		TREC9	
	RSJ (IDF-related)	PFS (DF-related)	RSJ (IDF-related)	PFS (DF-related)
#rel_ret	807	**832**	658	**670**
Ave.P	0.146	**0.149**	0.184	**0.183**
P@5	0.348	**0.364**	0.342	**0.354**
P@10	0.326	**0.326**	0.273	**0.279**

Table 2 Comparison of best retrieval performances on TREC9 and TREC10 data, using italic as PFF

表 2 使用斜体(⟨i⟩)作为主特征域在 TREC9 和 TREC10 上的检索最佳性能比较

	TREC10		TREC9	
	RSJ (IDF-related)	PFS (DF-related)	RSJ (IDF-related)	PFS (DF-related)
#rel_ret	815	**841**	663	**679**
Ave.P	0.140	**0.141**	0.181	**0.179**
P@5	0.328	**0.329**	0.320	**0.317**
P@10	0.310	**0.308**	0.266	**0.267**

Table 3 Comparison of best retrieval performances on TREC9 and TREC10, using head (⟨h1⟩, etc)as PFF

表 3 使用大小标题(⟨h1⟩等)作为主特征域在 TREC9 和 TREC10 上的检索最佳性能比较

	TREC10		TREC9	
	RSJ (IDF-related)	PFS (DF-related)	RSJ (IDF-related)	PFS (DF-related)
#rel_ret	814	**840**	665	**682**
Ave.P	0.136	**0.138**	0.179	**0.180**
P@5	0.324	**0.324**	0.319	**0.317**
P@10	0.303	**0.302**	0.264	**0.264**

4.4 主特征空间(PFS)权值计算方法稳定性和适用性实验

为了进一步验证使用粗体字部分作为主特征域使用 PFS 模型的有效性,在 TREC11 上也进行了实验.这是因为与 TREC9 和 TREC10 相比,TREC11 不仅改变了查询,而且改变了文档数据集,同时任务的要求也发生了变化:从以前的查找内容上相关的文档变为查找关键资源,因而检索任务更难也更接近实际情况.

图 5~图 7 列出了使用基于 DF 因子的 PFS 权值计算方法进行检索的实验结果,分别用找到的相关文档数,综合平均精度和前 10 篇文档的精度(TREC11 的官方评价指标)3 个指标来衡量.表 4 则给出了传统的 RSJ 方法和文中的 PFS 方法所能达到的最佳效果的比较实验结果.

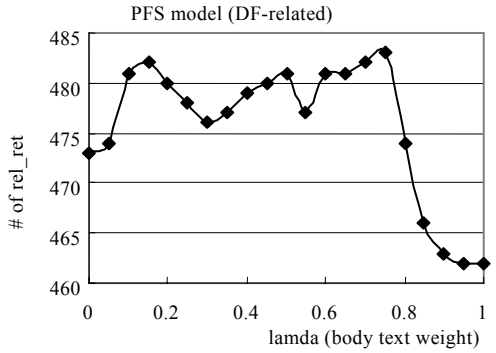

Fig.5 Performance on TREC11: The number of relevant documents returned (PFF: Bold text)

图 5 TREC11 上的检索性能:返回的相关文档数(主特征域:粗体字)

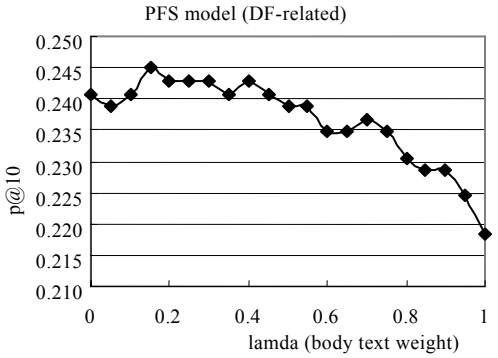

Fig.6 Performance on TREC11: The average precision of top 10 documents (PFF: Bold text)

图 6 TREC11 上的检索性能:前 10 篇文档的精度(主特征域:粗体字)

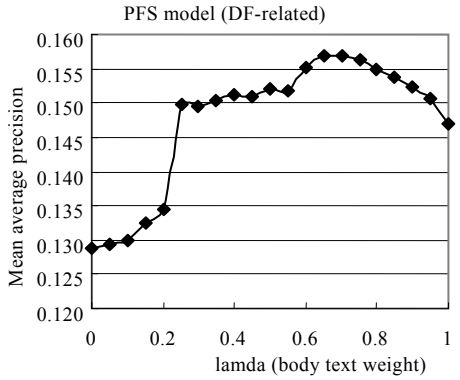

Fig.7 Performance on TREC11: Mean Average Precision (PFF: Bold text)

图 7 TREC11 上的检索平均性能:综合平均精度(主特征域:粗体字)

Table 4 Comparison of best retrieval performances on TREC11 data, using bold text as primary feature field

表 4 使用粗体字作为主特征域在 TREC11 上的检索最佳性能比较

	#rel_ret	Ave.P	P@5	P@10
Baseline	462	0.147 0	0.240 8	0.218 4
RSJ (IDF-related)	478	0.150 0	0.277 6	0.240 8
PFS (DF-related)	483	0.157 0	0.285 7	0.244 9
RSJ improvement	+3.5%	+2.0%	+15.3%	+10.2%
PFS improvement	+4.5%	+6.8%	+18.6%	+12.1%

从图 5~图 7 这 3 个结果图中可以看到,在检索到的相关文档数即前 100 篇文档的召回率和前 10 篇文档的精度方面,只要使用主特征域和主特征空间的概念,无论λ取何值,检索结果都会大大好于不区分 HTML 域(即正文影响因子λ为 1)的效果.当λ取 0.1~0.75 之间时,检索性能的改进最大.而在综合平均精度上,只要λ取值在 0.3~1 之间,则考虑主特征域都会为检索带来很大的提高.这个适用范围是很广的,反应了 PFS 权值计算方法的稳定性和适用性.

在 TREC11 的 web 信息检索评测任务中,共有 17 个不同的机构参加标准评测,使用.GOV 数据集,共提交 71 组结果.这些结果能够反映目前国际上 web 信息检索领域的主要方法和最高的水平.在所有使用了 HTML 文档结构信息的方法中,能够得到的最好结果的前 10 选平均精度(p@10)为 24.08%,这与表 4 中我们实验中的数据一

致.在这些方法中,链接文本信息(anchor text),正文部分(body),题目(title),地址信息(URL),大字体(big font)等都被考察和使用.而本文实验中使用了文档主特征模型 PFS 的方法能够达到 24.49%的前 10 选精度,比其他所有方法有 10.2%的提高,显示了该方法的有效性和结果的可信性.

得到的结论与前面实验的完全一致:与不区分 Web 文档的特征域以及根据传统的 RSJ 计算方法区分特征域的性能相比,使用基于 DF 因子的主特征空间权值计算方法总是能够最好的检索性能,实现最大 18.6%的性能改善.

5 结 论

本文对 Web 文档信息特征进行分析和研究,找到更合理的文档特征表示方法及特征权项值计算方法,从而使文档空间的信息描述更精确,以改进 Web 信息检索的效果.从 TREC9(2000),TREC10(2001)和 TREC11(2002)这 3 个大规模测试集的比较实验结果中,可以得到如下结论:

首先,提取主特征域和主特征空间的方法,与普通文本检索中不区分主特征域和正文的方法相比,总是能够得到较大的系统检索性能改进;其次,考虑 HTML 文档结构,基于 DF 因子的主特征空间权值计算方法(PFS 方法)进行检索,与使用传统的 IDF 相关的 RSJ 权值计算方法相比,在各种评价指标上,总是能够有更优的检索效果;最后,粗体字域是 HTML 文档中一个有效的内容表述部分,非常适合作为网页文档的主特征域.

在未来的工作中,将进一步考察更多的 HTML 信息作为主特征域的效果,并结合 HTML 文档内容来选取主特征域.

References:

[1] Anick PG. Adapting a full-text information retrieval system to computer the troubleshooting domain. In: W. Bruce Croft, C. J. van Rijsbergen eds. Proc. of the 17th Annual Intl ACM-SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'94). Ireland: ACM Press, 1994. 349~358.

[2] Croft WB, Cook R, Wilder D. Providing government information on the Internet: Experience with THOMAS. In: Proc. of the 2nd Int'l Conf. in Theory and Practice of Digital Libraries (DL'95). Texas, 1995.19~24. (http://csdl.tamu.edu/DL95/papers/croft/croft.html)

[3] Stefan K, Armin H, Markus J, Andreas D. Improving document retrieval by automatic query expansion using collaborative learning of term-based concepts. Lecture Notes in Computer Science 2423, 2002. 376~387.

[4] Zhang M. Study on Web text information retrieval [Ph.D. Thesis]. Beijing: Tsinghua University, 2003 (in Chinese with English abstract).

[5] Moffat A, Davis R, Wilkinson R, Zobel J. Retrieval of partial documents. In: Harman D, ed. Proc. of the 2nd Text REtrieval Conf. (TREC-2). National Institute of Standards and Technology Special Publication, Gaithersburg, MD, 1994. 181~191.

[6] Srinivasa S, Bhatt PCP. Introduction to Web information retrieval: A user perspective. Journal of Science Education, 2002,7(6): 27~38.

[7] Meng M, Yu C, Liu KL. Building efficient and effective metasearch engines. ACM Computing Surveys, 2002,34(1):48~89.

[8] Glover E, Tsioutsiouliklis K, Lawrence S, Pennock D, Flake G. Using Web structure for classifying and describing Web pages. In: Proc. of the Int'l World Wide Web Conf. (www 2002). Hawaii: ACM Press, 2002. 562~569. (http://www2002.org/CDROM/refereed/504/index.html)

[9] Cutler M, Shih Y, Meng W. Using the structure of HTML documents to improve retrieval. In: Proc. of the USENIX Symp. on Internet Technologies and Systems (NISTS'97). 1997. 241~251. (http://www.usenix.org/publications/library/proceedings/usits97/full_papers/cutler/cutler.pdf)

[10] Newby GB. Information space based on HTML structure. In: Vorhees E, ed. Proc. of the 9th Text REtrieval Conf. (TREC-9). National Institute of Standards and Technology Special Publication, Gaithersburg, MD, 2000. 601~610.

[11] Ricardo BY, Berthier RN. Modern Information Retrieval. New York: Addison-Wesley, ACM Press, 1999. 19~34.

[12] Robertson SE, Walker S. Microsoft cambridge at TREC-9: Filtering track. In: Vorhees E, ed. Proc. of the 9th Text REtrieval Conf. (TREC-9). National Institute of Standards and Technology Special Publication, Gaithersburg, MD, 2000. 25~33.

[13] Bailey P, Craswell N, Hawking D. Engineering a multi-purpose test collection for web retrieval experiments. Information Proceeding and Management. 2003,39(6):853871.

[14] Craswell N, Hawknig D. Overview of the TREC-2002 Web track. In: Vorhees E, ed. Proc. of the Text REtrieval Conf. 2002 (TREC-2002). National Institute of Standards and Technology Special Publication, Gaithersburg, MD, 2002. 6168.

附中文参考文献:

[4] 张敏.Web 文本信息检索方法研究[博士学位论文].北京:清华大学,2003.