中 文 信 息 学 报
第 20卷 第 3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 No. 3

文章编号: 1003 - 0077(2006)03 - 0070 - 08

面向信息检索需要的网络数据清理研究*

刘奕群,张 敏,马少平

(清华大学 计算机系 智能技术与系统国家重点实验室,北京 100084)

摘要: Web数据中的质量参差不齐、可信度不高以及冗余现象造成了网络信息检索工具存储和运算资源的极大浪费,并直接影响着检索性能的提高。现有的网络数据清理方式并非专门针对网络信息检索的需要,因而存在着较大不足。本文根据对检索用户的查询行为分析,提出了一种利用查询无关特征分析和先验知识学习的方法计算页面成为检索结果页面的概率,从而进行网络数据清理的算法。基于文本信息检索会议标准测试平台的实验结果证明,此算法可以在保留近 95%检索结果页面的基础上清理占语料库页面总数 45%以上的低质量页面,这意味着使用更少的存储和运算资源获取更高的检索性能将成为可能。

关键词:计算机应用;中文信息处理;网络信息检索;数据清理;机器学习

中图分类号: TP391 **文献标识码**: A

Web Data Cleansing for Effective Information Retrieval

LIU Yi-qun, ZHANG Min, MA Shao-ping

(State Key Lab of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

Abstract: The existence of low quality Web pages affects the effectiveness and efficiency of Web search. In this paper, we define the Web page quality estimation as a learning problem. First, several query-independent features are investigated which can separate search target page from ordinary ones. Bayes estimation based on these features is then used to train a model to assign importance scores to Web pages. In TREC based experiments, the top-scored set reduces 45% low quality pages as well as retains 95% high quality ones. It shows the possibility to gain better performance with less storage and computing resource for search engines.

Key words: computer application; Chinese information processing; Web information retrieval; data cleansing; machine learning

1 引言

随着 Internet技术的发展和普及,Web信息迅速成为了社会成员获取信息的主要渠道之一,面对海量的网络数据资源,信息检索工具已经成为人们使用这些资源的必要手段。尽管已经取得了很大进展,但检索工具还无法充分满足人们快捷高效获取网络信息的要求:网络搜索引擎返回的查询结果依然数目浩繁,让人难以着手;结果网页与用户查询主题的贴切程度以及网页本身的质量也良莠不齐。

网络信息检索的发展水平不尽如人意是由许多方面的原因造成的,其中繁杂的 Web信息

* **收稿日期**: 2005 - 11 - 03 **定稿日期**: 2006 - 02 - 23

基金项目:国家重点基础研究 (973)资助项目 (2004CB318108);自然科学基金资助项目 (60223004, 60321002, 60303005, 60503064);教育部科学技术研究重点项目资助 (104236)

作者简介:刘奕群 (1981—),男,山东济南人,博士生,主要研究方向为信息检索. 机器学习.

环境和海量的用户需求是技术发展上的两个主要障碍。网络信息检索工具已经面临着巨大的存储和计算压力,但 Web环境中的低质量、冗余乃至垃圾数据却进一步增加了处理的难度:尽管这些数据几乎不会为用户所利用,但检索系统无法分辨它们,因此也不得不花费大量的资源存储这些数据,并计算用户查询与这些页面的相关性。任何一个不涉及数据清理和数据质量评估模块的搜索引擎,将在当今的网络数据环境中寸步难行,更谈不上为用户提供高质量的检索服务。

按照网络数据清理的粒度不同,现有的解决思路大致分为两类,即 Web页面级别的数据清理和基于页面内部元素级别的数据清理,前者以 Google公司提出的 PageRank算法[4]和 IBM公司 Clever系统的基石 HITS算法[5]为代表;而后面一个思路则集中体现在作为 MSN搜索引擎核心技术之一的 VIPS算法[6]上。这些算法被各自企业投入到大规模网络搜索引擎的应用中,并取得了一定的成功。但这些算法大都是从网络数据本身的特性而非检索用户的真正需求出发进行清理;数据清理的依据基本也只是网页的超链接结构信息。面对当前为提高搜索引擎排名而层出不穷的网页作弊方式,仅仅利用这些清理算法已经无法很好的起到过滤低质量数据的目的。

通过对大规模网络语料库和检索用户需求的分析,我们发现检索目标页面与普通页面之间存在着与用户查询无关的特征差异。从这些差异出发,本文提出一种基于机器学习方法的数据清理算法,与以上提到的数据清理算法不同,此算法从用户检索需求出发,基于包括超链接特征、页面内部结构特征在内的多种查询无关特征计算页面成为检索目标页面的先验概率,进而实现面向检索需要的网络数据清理。本文的主要贡献包括:

1. 通过对检索目标页面与普通页面区别的分析,总结出检索目标页面的查询无关特性。
2. 提出一种利用检索目标页面的查询无关特性和机器学习算法进行数据清理的方法。
3. 验证这种基于先验概率学习算法的数据清理方法的可行性。

2 相关工作概述

2.1 面向信息检索的网络数据清理算法

网络数据清理按其施行粒度不同可以分为两类:页面内部元素级别的数据清理目标在于清理页面内部的无用信息(包括广告信息、版权信息等);而页面级别的数据清理主要定位在从整体页面集合中选择出高质量页面。

第一类数据清理在网络信息检索技术中较早的就得到了应用,类似 AltaVista和 Yahoo等搜索引擎在创立之初就对 HTML文档中的不同部分采用加权的方式进行相关度计算。Cutler等人总结了前人对超文本文档内部结构分析的工作并将根据标签将文档划分为六个部分(如标题、链接文字等),并使用不同的重要性因子对某些部分进行强调。尽管早期研究中,页面内部结构特征得到了充分重视,但这时的技术是基于加权模型,会为参数调整问题所困扰,难于总结出合理的物理含义,也不便于进一步研究的开展。2003年,微软公司的 Cai等人指出了一条利用页面内部结构特征的新途径,他们从页面的视觉分块特征出发,提出将页面划分为语义上相对独立的单元(block,块),并以这种单元为基本单位进行检索(VIPS算法[6])。他们还进一步提出了基于块结构的链接分析和查询扩展算法[7~9],并使这系列算法成为 MSN搜索引擎主要的技术特征之一。

Henzinger等人在文献[10]中指出了第二类数据清理的重要意义,他们认为独立于用户查询需求判断网页质量将是对搜索引擎有重大意义的进展。尽管现有的超链接结构分析算法如

PageRank, HITS等能够独立用户查询判定页面的质量,而且这些算法也在当前的几乎所有搜索引擎中得到应用,但总有人提出类似的疑问:PageRank/HITS数值真的代表了他能够满足检索用户需求的程度么? Amento等人在文献[11]中利用实验验证了至少在小规模数据上,包括PageRank/HITS在内的各种链接结构分析算法都无法有效的提高纯文本检索的效果。尽管PageRank数值的可靠程度在数据规模不同的情况下会有较大的差异,但是算法的设计初衷并非为反映检索用户的需求这一点是无法改变的。

尽管现有的网络数据清理算法已经得到了较为广泛的应用,但它们也远非尽善尽美,主要存在的问题包括:第一,面对搜索引擎极高的实时计算的要求,不少算法不但没有减轻,反而加重了线上计算的压力。各种基于页面内部结构的清理方式不仅增加索引规模,而且需要引入额外的权重计算、合并的工作量;HITS算法更是需要在线迭代计算与某个查询主题相关的几千篇文档的链接评价数值。第二,两种类型数据清理算法都是从赋予不同页面,或者页面的不同部分之间不同的权重来实现对高质量页面的区别对待,这种操作尽管保留了网络环境的所有信息,但实际上没有减轻存储的负担,搜索引擎仍然需要保存包括各种低质量页面在内的所有页面。最重要的,上述各种数据清理算法都没有从检索用户实际需要的角度出发进行清理,网络信息检索的需求千差万别,但检索目标页面是否存在独立于用户查询需求的特征呢?具有这些特征的页面又是否被传统的数据清理算法给予较高评价呢?这都是我们需要在本文中考察的问题。

2.2 网络信息检索用户行为分析研究概述

在分析检索目标页面的查询无关特征之前,我们首先需要对网络信息检索的查询需求有一个全面的认识。在实际应用中,用户检索的需求多种多样,但根据Broder等人对AltaVista搜索引擎的用户日志分析工作,检索可以根据查找信息目的不同分为三类,即导航类查询、信息类查询和事务类查询。2004年,Yahoo公司的Danny等人在Broder工作的基础上将检索类型进行了细化,但总体分类结构则基本保持不变,这也说明了这种查询分类体制的可靠性。

导航类查询的目标是查找一个用户已知的网页(帮助其找到对应的URL),例如:“软件学报主页”、“清华大学招生简章”都属于导航类检索的范畴。导航类检索按照查找目标页面的不同细分为特殊需求页面定位任务和主页定位任务两类,主页定位任务的目标页面是站点/子站点的主页,而特殊需求页面定位任务的目标页面则是主页以外的页面。

信息类检索的目标是查找关于某个查询主题的相关信息,如“加强党的执政能力”就可以算作信息类查询;事务类检索则是用于查找关于某个内容的网络服务,如购物服务、查询服务、下载服务等,典型的样例如“mp3下载”等。

依据Broder和Danny的分类体系,我们可以将网络信息检索的目标页面分为特殊需求页面和关键资源页面两类。关键资源页面是指对用户获取某方面(通常是一个较为广泛的主题)最有帮助的页面,而特殊需求页面则是指关键资源页面之外的能够满足用户需求的页面。按照查询类型与目标页面的对应关系,导航类检索的主页定位任务以及信息类检索的目标页面可以认为是关键资源方面,而特殊需求页面定位任务的目标页面则是特殊需求页面。在利用检索目标页面分类进行数据清理的工作中,我们会分别考察这两种检索目标页面的特性。

3 检索目标页面与普通页面的差异分析

对于检索目标页面和普通页面差异的分析是必要的:一方面,按照上文的论述,检索目标页面与普通页面的差异是本文数据清理工作的出发点;另一方面,这种分析对于我们了解检索

目标页面的实质也是大有裨益的。

根据前人的工作[13～15],查询无关特征可以用来提高某些特殊类型检索的性能(如主题过滤、主页查找等);而我们之前的工作[16,17]也说明一些查询无关特征可以用于关键资源页面的挑选。在本文的分析中,我们借鉴了这些工作所采用的特征,并引进了一些新特征以发现检索目标页面和普通页面之间的差异,这些特征包括:

1. 文档长度:以词数(英文单词数/中文字数等)计算的文档长度;

2. 链接文本长度:链接到该页面的链接文字的长度总和,反映了此页面被多少个页面所链接,以及这些页面对该页面的重视程度;

3. PageRank:根据文献[4]算法计算出的反映用户访问该页面概率的数值;

4. 入链接个数:有多少个页面链接到该页面;

5. 出链接个数:该页面链接到多少个页面;

6. 站点内出链接个数:该页面链接到多少个本站点内部的页面;

7. URL长度:按文献[15]的描述,根据长度不同将URL分为四个类型,分别为ROOT型(域名),SUBROOT型(一级目录),PATH型(多级目录),FILE型(其它)。

上述都是在以往工作中被证明可以一定程度反映页面质量的查询无关特征,为了比较检索目标页面与普通页面在这些特征上的差异,必须考察大规模语料库上两者之间的分布情况。.GOV是2002年从.gov域抓取的规模为125万页面的基于真实网络环境的语料库,共包括近20G的数据,并在2002～2004年的TREC评测中使用。我们之所以选择.GOV而不是其它更大规模的语料库,原因在于基于.GOV已经进行了多年的信息检索评测研究,有充分的被TREC评测人员手工标注的高质量页面充当后续工作的训练与测试集合。例如在考察查询目标页面的查询无关特征方面,我们就采用了如表1所示的训练集合。

表1 查询目标页面训练集合的组成情况

	关键资源训练集合	特殊需求页面训练集合
页面集合来源	TREC 2002 - 2003主题过滤任务	TREC 2002 - 2003导航类检索任务
页面集合大小	327	860

根据对表1的训练集合以及.GOV页面的查询无关特征分析,我们发现查询目标页面与普通页面之间存在着较明显的差异,具体的实验结果如表2所示。

表2 查询目标页面与普通页面的相关系数比较

查询无关特征	关键资源页面 vs 普通页面	特殊需求页面 vs 普通页面	关键资源页面 vs 特殊需求页面
文档长度	0.3491	0.4878	0.9775
链接文本长度	0.7185	0.8198	0.9793
PageRank	0.5684	0.0695	0.8198
入链接个数	0.6517	0.5289	0.8858
出链接个数	0.2640	0.6728	0.8725
站点内出链接个数	0.5888	0.7226	0.9221
URL长度	0.9639	0.9888	0.9923

表2中,相关系数是统计学研究中常用的数值标准,这里用来表示待比较的两组页面数据之间的相似程度,相关系数越小,说明两组页面之间的相似程度越小;相关系数越接近1,则说明两组页面在这个查询无关特征上的表现越相似。由表2可得以下结论:

(1)关键资源页面与普通页面的最大差异在于与出链接相关的几个特征上,关键资源中

出链接 /站内出链接个数较多的页面比例明显多于普通页面,相关系数的数值比较也说明关键资源页面在出链接特征方面与普通页面有非常大的不同。这是由关键资源页面在用户获取信息方面所发挥的作用决定的,根据我们之前的工作[16,17],关键资源页面本身不一定提供大量可用信息,但一般都能够提供较多高质量页面的链接。

(2)特殊需求页面与普通页面的最大差异在于 PageRank特征上。不同特征之间的相关系数比较中可以看出,PageRank是区别特殊需求页面与普通页面的最重要特征。根据 Page等人在文献[4]中的描述,PageRank反映的是互联网浏览用户访问到某个页面的概率大小。由于特殊需求页面一般都是用户曾经访问到的页面,因此它在 PageRank数值上与普通页面有较明显的区别也就是合理的了。

(3) 特殊需求页面与关键资源页面的相似性远大于它们各自与普通页面之间的差异,如在文档长度、链接文字长度、URL长度等特征上两类页面都表现得非常相似。在 PageRank和出链接特征上两类页面有一定差异,这也验证了这两个特征分别是特殊需求页面与关键资源页面的主要特征。

4 基于先验概率学习的数据清理

根据上文阐述,查询目标页面与普通页面之间存在着查询无关特征的差异,因此利用这些特征进行查询目标页面的分类,进而实现面向检索需要的网络数据清理是可能的目标。然而,使用何种机器学习算法进行特征综合,并实现分类还是需要进一步考虑的问题。这是由于查询目标页面分类问题与一切大规模数据上的网页分类问题一样,面临着训练样例获取的问题。任何机器学习算法必须依赖于一定数量的训练样例才能实现,但是对于面向网络数据的学习算法而言,大量的研究如文献[18]则证明主要的困难集中在如何获取有充足代表性的训练样例上。

由于查询目标页面有较为明确的定义,人们对这类页面也有较明确的感性认识,因此正例相对较易获得,特别是 TREC的网络信息检索任务提供了大量的标准答案页面作为正例页面的可靠来源。相比较而言,反例的获得就困难得多,Web页面多种多样,作为 Web页面中绝大部分的非查询目标页面更是数目繁多,情况繁杂。内容不可信,内容太少,没有高质量的入链接或者出链接,有太多广告信息等等,都可能是页面无法成为关键资源页面的理由。 Google公司的 Henzinger等人近年来多次指出:Web页面的均一采样是当前在理论和应用上都无法克服的困难[10]。因此获得高质量的反映 Web真实情况的反例绝非易事。在之前的工作[16]中,我们提出了可以利用改进的决策树算法实现这种基于正例样例的学习,并利用此算法获得了较好的关键资源页面定位效果。但是,当时采用的算法需要对网络语料库中的正例页面比例进行估计,对于一般的页面分类问题而言,这个比例比较难于给出,因此本文提出的先验概率学习方式尽量避免这方面问题的出现。

根据贝叶斯公式,一个具有查询无关特征 A的页面 p成为查询目标页面的概率为:

$$P(p \in Target\ page \mid p\ has\ feature\ A) = \frac{P(p\ has\ featrue\ A \mid p \in Target\ page)}{P(p\ has\ feature\ A)} \times P(p \in Target\ page) \quad (1)$$

其中,$P(p \in Target\ page)$表示语料库中,正例页面即查询目标页面所占的比例,这个比例相对比较难于估计,但是由于各个页面之间的概率值仅仅用于比较相对大小,而各个页面概率计算中此值保持不变,因此可以忽略它的影响,仅仅关注 (1)式的前半部分。根据先验概率的

定义,又有:

$$\frac{P(p\ has\ feature A \mid p \in Target\ page)}{P(p\ has\ feature A)} = \frac{\#(p\ has\ feature A \cap p \in Target\ page)}{\#(Target\ page)} \Big/ \frac{\#(p\ has\ feature A)}{\#(\cdot GOV)} \tag{2}$$

由于我们认为查询目标页面的采样是均一可靠的(只要我们所选取的查询主题涉及范围足够广,可以认为这个假设是成立的),因此可以将(2)式中的分子部分进一步细化:

$$\frac{\#(p\ has\ feature A \cap p \in Target\ page)}{\#(Target\ page)} = \frac{\#(p\ has\ feature A \cap p \in Target\ page\ training\ set)}{\#(Target\ page\ training\ set)} \tag{3}$$

因此,将(2)、(3)二式代入(1),则有:

$$P(p \in Target\ page \mid p\ has\ feature A) \propto \frac{\#(p\ has\ featrue A \cap p \in Target\ page\ training\ set)}{\#(Target\ page\ training\ set)} \Big/ \frac{\#(p\ has\ feature A)}{\#(.GOV)} \tag{4}$$

(4)式的分子与分母都可以通过对页面语料库以及查询目标页面训练集合的统计分析得到,因此语料库中任意一个页面成为检索目标页面的概率都可以使用此式计算得到。

5 实验与结果分析

数据清理实验的测试集使用了 TREC2004网络信息检索任务的查询主题及标准答案,此任务一共提供了 225个查询主题和对应主题的标准答案,查询主题包括主页查找类、特殊需求查找类和主题过滤类各 1/3,主题内容来源于真实网络搜索引擎的用户查询,包含的内容领域涉及社会政治、经济生活的方方面面,而标准答案的标定也经过了多位评测人员的反复验证。因此具有较高的权威性与可靠性。根据查询目标概率分布情况,我们采用不同的参数进行了数据清理实验,具体的实验结果参见表 3。

表 3 基于先验概率分析的数据清理实验结果

概率条件	关键资源页面测试集覆盖率	特殊需求页面测试集覆盖率	页面保留率
log(p(NP)) >0	98.47%	99.07%	70.77%
log(p(NP)) >0.1	96.33%	98.25%	63.10%
log(p(NP)) >0.2	94.49%	94.63%	51.28%
log(p(KEY)) >0.1	77.07%	92.75%	46.15%

由表 3的实验结果可以得到如下结论:首先,利用查询目标概率分类的方式,可以在保留 51.28%语料库页面的同时仅仅损失 5%左右的查询目标页面;在保留 70.77%页面的情况下,则数据清理过程几乎可以不损失查询目标页面。其次,从清理效果来看,与关键资源先验概率相比,特殊需求页面概率更适合用于数据清理任务,因为它在进行清理的同时保留了更多的查询目标页面。

6 结论与未来工作

面对纷繁复杂的网络数据环境,如何能够独立用户查询主题实现低质量页面的过滤是网络信息检索研究的前沿热点。本文试图利用网页特征分析和机器学习方法实现面向检索需要

的网络数据清理,我们的主要结论包括:

(1)查询目标页面与普通页面之间存在着查询无关特征的差异。

(2)利用先验概率学习的方法,可以在清除页面集合中近一半页面的同时保留几乎所有的检索目标页面。由于检索目标页面才是用户使用检索工具中唯一关心的对象,因此可以利用这种方式实现不降低检索效果基础上的数据清理。

未来可能的工作方向包括:在更大规模的语料库上考察此数据清理算法的效果;挖掘更多的可以用于查询目标页面挑选的查询无关特征;考察将数据清理算法与检索算法加以结合的可能性等。

参 考 文 献:

[1] Lyman, Peter and Hal R. Varian, How Much Information 2003 [EB/OL]. http://www.sims.berkeley.edu/how-much-info-2003 on 2005-06-18, 2003-10-30/2005-06-18.

[2] Danny Sullivan, Search Engine Sizes [EB/OL]. From search engine watch web site http://searchenginewatch.com/reports/article.php/2156481, 2005-01-28/2005-06-18.

[3] Danny Sullivan, Searches Per Day [EB/OL]. From search engine watch web site http://searchenginewatch.com/reports/article.php/2156461, 2003-02-25/2005-06-18.

[4] Sergey Brin and Lawrence Page, The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks and ISDN Systems, 1998, 30(7): 107-117.

[5] Jon M. Kleinberg, Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 1999, 46(5): 604-632.

[6] Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma. VIPS: a Vision-based Page Segmentation Algorithm [R], Microsoft Technical Report (MSR-TR-2003-79), 2003.

[7] Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma. Block-based web search [A]. In: proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval SIGIR 04 [C]. New York, NY: ACM Press, 2004, 456-463.

[8] Deng Cai, Xiaofei He, Ji-Rong Wen and Wei-Ying Ma. Block-level Link Analysis [R], Microsoft Technical Report MSR-TR-2004-50, 2004.

[9] Ruihua Song, Haifeng Liu, Ji-Rong Wen and Wei-Ying Ma, Learning Block Importance Models for Web Pages [A]. In: proceeding of the Thirteenth World Wide Web conference [C], New York, NY: ACM Press, 2004, 203-211.

[10] Monika R. Henzinger, Rajeev Motwani and Craig Silverstein, Challenges in Web Search Engines [A], Georg Gottlob, Toby Walsh eds. IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence [C]. San Francisco: Morgan Kaufmann Press, 2003. 1573-1579.

[11] B. Amento and L. Terveen and W. Hill. Does authority mean quality? Predicting expert quality ratings of Web documents [A]. Nicholas J. Belkin, Peter Ingwersen and Mun-Kew Leong, eds. In: proceedings of the 23rd Annual International ACM SIGIR Conference [C]. New York: ACM Press, 2000. 296-303.

[12] Andrei Broder, A taxonomy of Web search [J]. SIGIR Forum, 2002, 36(2): 1-8.

[13] Nick Craswell, David Hawking and Stephen Robertson. Effective Site Finding using Link Anchor Information [A]. W. Bruce Croft, David J. Harper, Donald H. Kraft, Justin Zobel eds. In: proceedings of the 24th Annual International ACM SIGIR Conference [C]. New York: ACM Press, 2001. 250-257.

[14] Nick Craswell and David Hawking. Query-independent evidence in home page finding [J]. ACM Transactions on Information Systems (TOIS), 2003, 21(3): 286-313.

[15] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search

[A]. Ricardo Baeza-Yates ed. In: proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. New York: ACM Press, 2002. 27 - 34.
[16] Yiqun Liu, Min Zhang, Shaoping Ma, Effective Topic Distillation with Key Resource Pre-selection [J], Lecture Notes in Computer Science, Volume 3411, 129 3/ 140.
[17] Yiqun Liu, Canhui Wang, Min Zhang, Shaoping Ma, Web Data Cleansing for Information Retrieval using Key Resource Page Selection [A]. In: proceedings of the 14th International World Wide Web conference [C], New York: ACM Press, 2005, 1136 - 1137.
[18] Hwangjo Yu et al. PEBL: Web Page Classification without Negative Examples. IEEE Trans. On Knowledge and Data Engineering [J], 2004, 16(1).

(上接第 54页)

供了难得的契机,国家标准藏文编码字符集(扩充集 A)结束了长期以来藏文编码标准不统一的混乱局面,开发藏文办公套件应以此编码标准为基础,可在 OpenOffice.org强大的软件国际化框架下,加入对藏文本地环境的支持。另外,支持藏文的排版习惯是藏文办公套件的重中之重,通过将藏文的断行习惯转换为断行规则,解决了藏文自动断行的问题,并实现了用音节分隔符填充在藏文文本断行后留下的空白的功能。

参 考 文 献:

[1] 江荻,周季文. 论藏文的序性及排序方法[J]. 中文信息学报, 2000, 14(1): 56 - 64.
[2] 林河水,程伟,等. 一种 ISO14651语义的藏文排序实现方法[J]. 中文信息学报, 2000, 18(5): 36 - 41.
[3] 达哇彭措,尕藏草. 网络媒体中藏文版式规则[A]. 见:鲍怀翘、金星华、宗成庆主编,少数民族语言信息技术研究进展:中国少数民族语言信息技术与语言资源库建设学术研讨会论文集[C]. 中国北京, 2004年 4月, 84 - 87.
[4] 胡书津. 简明藏文文法[M]. 云南民族出版社, 1995.
[5] 国家质量技术监督局, GB16959 - 1997信息技术 —信息交换用藏文编码字符集 —基本集[M]. 北京:中国标准出版社, 1998年 5月.
[6] 国家质量技术监督局, GB/T16960.1 - 1997信息技术 —藏文编码字符集(基本集) 24X48点阵字型 —第1部分:白体[M]. 北京:中国标准出版社, 1998年 4月.
[7] 国家质量技术监督局, GB/T17543 - 1998信息技术 —藏文编码字符集(基本集)键盘字母数字区的布局[M]. 北京:中国标准出版社, 1999年 6月.
[8] Joan Aliprand, Julie Allen, Joe Becker, Mark Davis, Michael Everson, Asmus Freytag, John Jenkins, Mike Ksar, Rick McGowan, Eric Muller, Lisa Moore, Michel Suignard, and Ken Whistler. The Unicode Standard Version 4.0[M]. Addison-Wesley, Aug 2003.
[9] Unicode Standard Annex #14 Line Breaking Properties. http://www.unicode.org/reports/tr14/.
[10] Unicode Standard Annex #29 Text Boundaries. http://www.unicode.org/reports/tr29/.
[11] Universal I18n Framework for Office Applications. http://l10n.openoffice.org/Universal_i18n_framework.pdf.
[12] Text Element Boundary Analysis. http://oss.software.ibm.com/icu/userguide/boundaryAnalysis.html