基于聚类和索引技术的语言模型压缩方法

祁斌川[1],杨端端[2],丁建国[1]
(1.中国科学院上海应用物理研究所 束测控制部门,上海 201800;
2.盛大创新研究院 语音主题部门,上海 201203)

摘 要:由于训练语料的庞大,SRILM训练生成的ARPA统计语言模型数据文件体积过大,导致查找效率低下以及消耗大量的存储空间。针对该问题,借鉴聚类和索引查找的思想,提出了一种基于K均值(K-means)聚类算法的对语言模型中的转移概率和回退概率压缩,并通过多级索引技术提高查找速度的压缩方法。理论分析和实验表明,该方法可以在减少压缩造成的数据失真对选词影响的同时,取得非常好的压缩效果,同时提高了对语言模型文件查找效率,并且输入法的反应速度得到了明显的提升。

关键词:语言模型;压缩方法;聚类算法;多级索引
中图分类号:TP319.14 **文献标识码**:A **文章编号**:1673-629X(2012)12-0025-04

Compression Method of Language Model Based on Clustering Algorithm and Multistep Indexing

QI Bin-chuan[1], YANG Duan-duan[2], DING Jian-guo[1]
(1. Dept. of Beam Instrumentation & Control, Shanghai Institute of Applied Physics, Chinese Academy of Science, Shanghai 201800, China;
2. Dept. of Speech Research, SNDA Institute of Innovation Research, Shanghai 201203, China)

Abstract: Because of the large-scale training corpus, the language model data file of the ARPA format produced by SRILM toolkit usually takes too much space and reduces the search rate. For the problem, learning from the idea of unsupervised clustering analysis and multilevel index, proposed a compression method of N-Gram Chinese language model file based on K-means clustering algorithm and multilevel index technology to increase search speed. Theoretical analysis and experiments show that the method can promptly obtain an outstanding compression ratio and effectively reduce the redundant search times, showing a good performance.

Key words: language model; compression method; K-means clustering algorithm; multilevel index technology

0 引 言

应用在移动电子平台上的输入法,要实现句子级别的智能输入,必须借助语言模型。统计语言模型在对特定的语料库进行统计和学习后,推测出自然语言的规律性。现在最常用的语言模型是N-gram语言模型。理论上,N-gram语言模型的N越大,提供的语境信息也越多,但代价就越大,且需训练语料多;N较小时,提供的信息比较少,但计算代价小,且无需太多训练语料。目前输入法中的常用语言模型有Tri-gram($N=3$)和Bi-gram($N=2$),其中微软拼音、智能狂拼使用的是Tri-gram,而谷歌拼音、搜狗拼音和紫光则使用的是Bi-gram[1-3]。

采用大量的训练语料,在保证训练得到的语言模型的性能的同时,也导致了语言模型的体积过于庞大,常用的Bi-gram与Tri-gram所占的空间占整个输入法软件大小的90%以上。并且手机等移动设备的空间和运算资源有限,因此对应用于移动平台上的语言模型进行压缩显得非常必要。

文中所介绍的是针对SRILM[1-3]工具训练语料所得的N-gram语言模型文件的压缩方法,基于K均值聚类算法对语言模型中的转移概率和回退概率压缩,并通过多级索引技术提高查找速度。实现数据表明,在兼顾语言模型的困惑度和整体性能的同时,获得了良好的压缩效果,并提高了查找速率。

1 N-gram模型和SRILM

现在输入法中应用最普遍的语言模型是统计语言模型中的N-gram模型。诸如微软拼音、谷歌拼音、搜

收稿日期:2012-03-17;**修回日期**:2012-06-23
基金项目:国家"973"重点基础研究发展计划项目(2011CB808300)
作者简介:祁斌川(1986-),男,江苏新沂人,硕士研究生,研究方向为输入法设计。

狗拼音等主流输入法均使用该模型。SRILM(Stanford Research Institute Language Modeling Toolkit)是著名的约翰霍金斯夏季研讨会(Johns Hopkins Summer Workshop)的产物,诞生于1995年,由SRI实验室的Andreas Stolcke负责开发维护,用来构建和应用统计语言模型[4]。

1.1 N-gram语言模型

统计语言模型建模方法中,由于N-gram其简单有效,得到了广泛的应用。N-gram模型基于这样一种假设:第n个词的出现只与前面$n-1$个词相关,而与其它任何词都不相关。用$w_1,\cdots,w_n$来表示这n个词,那么词w_n出现的概率就可以写为$p(w_n \mid w_1^{n-1})$,这里w_1^{n-1}表示词串$w_1,\cdots,w_{n-1}$。

在有大量训练语料保证的前提下,根据最大似然准则,可以得到:

$$p(w_n \mid w_1^{n-1}) = \frac{c(w_1^n)}{c(w_1^{n-1})}$$

$c(w_1^n)$和$c(w_1^{n-1})$分别表示词串$w_1,\cdots,w_n$和$w_1,\cdots,w_{n-1}$在训练语料中出现的次数。对于n个词构成的句子W,那么这句话的先验概率是:

$$p(W) = \prod_{i=1}^{n} p(w_i \mid w_{i-n+1}^{i-1})$$

之所以把这种模型称为N-gram模型,就在于其反映了连续n个词之间的相关信息。当$n=1$、2和3时,分别称为Uni-gram、Bi-gram和Tri-gram模型[5]。

1.2 SRILM

SRILM的主要目标是支持语言模型的估计和评测。其最基础和最核心的模块是N-gram模块,这也是最早实现的模块,包括两个工具:ngram-count和ngram,相应地被用来估计语言模型和计算语言模型的困惑度[2,6,7]。

以生成的语言模型文件ime_chars.2g.lm(8,372,523 Byte)为例,为ARPA文件格式。

为了说明方便,文件中的括号是笔者加上的注释:

```
\data\
ngram 1=7165 (注:一元词有7165个)
ngram 2=458382 (注:二元词有458382个)
\1-grams:(注:以下为一元词的基本情况)
-5.116223 (注:log(概率),以10为底)!
-5.212389% -0.6116316(注:回退权重log,以10为底)
…
\2-grams:
-0.1648012% </s>
-2.008501% 中
-2.161202% 人
…
```

2 K-means聚类算法和多级索引技术

2.1 K-means聚类算法

K-means算法是很典型的基于距离的聚类算法,使用误差平方和准则函数来评价聚类性能。给定数据集X,其中只包含描述属性,不包含类别属性。假设X包含k个聚类子集$X_1,X_2,\cdots,X_K$;各个聚类子集中的样本数量分别为$n_1,n_2,\cdots,n_k$;各个聚类子集的均值代表点(也称聚类中心),分别为$m_1,m_2,\cdots,m_k$。则误差平方和准则函数公式为:

$$E = \sum_{i=1}^{k} \sum_{p \in X_i} \| p - m_i \|^2$$

算法把得到紧凑且独立的簇作为最终目标,即通过迭代需求符合某个准则函数的聚类中心。

K-means算法的工作过程说明如下:首先从n个数据对象任意选择k个对象作为初始聚类中心;而对于所剩下其它对象,则根据它们与这些聚类中心的相似度(距离),分别将它们分配给与其最相似的(聚类中心所代表的)聚类;然后再计算每个所获新聚类的聚类中心(该聚类中所有对象的均值);不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数。k个聚类具有以下特点:各聚类本身尽可能的紧凑,而各聚类之间尽可能的分开[8]。

2.2 多级索引

多级索引是将多个不同或相同的索引方法组合使用,对单级索引空间或者空间范围进行多级划分,解决大型数据量的检索、分析、显示的效率问题[9]。

文章中介绍的多级索引方法是基于单字对双字词建立位置索引表,根据双字词对三字词建立索引列表。

3 压缩方案设计

在SRILM生成的ARPA格式的N-gram语言模型文件中,记录的内容主要是词组和其所对应的概率值,因此压缩语言模型也就是对概率值和词组进行压缩。

具体是通过采用聚类算法中的K-means算法对N-gram中的转移概率和回退概率进行聚类,通过存储聚类中心和每个概率值所对应的聚类中心索引序号来代替原来的表示浮点数的字符串,从而起到压缩的目的。另外,双字词中存在一定的冗余信息,滤除双字中的字首,并做相应的地址标记等一些技巧性的处理,在起到压缩作用的同时,使查找速率提高[10]。

3.1 对概率值的聚类压缩

语言模型中的概率分为转移概率和回退概率,处理的方式相同。整个处理算法分为二个模块:

1.将ARPA格式文件中的表示概率的字符串转换为整型。首先读入字符串,将表示浮点概率的字符串

转换为浮点数。将浮点数(0 ~ -99)乘以 1024 转换为整型。

2. 利用表示概率的整型数据获得 255 个聚类中心,对每个概率值计算出所对应的聚类中心索引,由于总共有 255 个聚类中心,所以每个索引可以用 8bits 的字符类型表示。

以 Bi-gram 为例:分别对 1-ngram、2-ngram 中的转移概率和回退概率进行处理,各自得到 255 个聚类中心和每个词所对应的索引值。聚类的流程如图 1 所示。

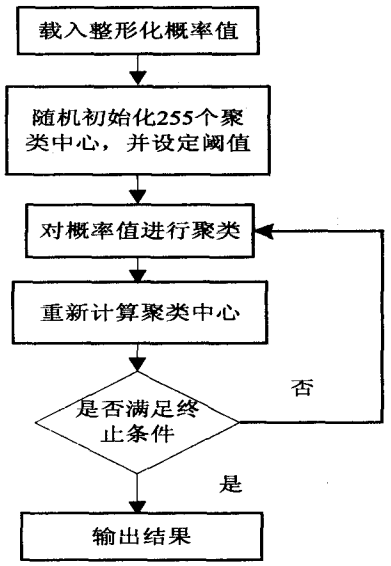

图 1　整形化概率值聚类流程

3.2　多级索引

对 ARPA 格式的数据文件进行分析,发现 N-gram 的字符数据是按照其首字的 utf-8 的编码大小进行顺序排列的。

在 Tri-gram 的 ARPA 文件中,2-gram 二元词的首字在 1-gram 中都有记录,3-gram 三元词的前两个字构成的双字词在 2-gram 中也已经记录。因此 2-gram 中的首字和 3-gram 中的前两个字是冗余的,同时文件中空格、换行、'Tab' 等格式符也可以过滤掉。

为了消除这种信息冗余,并且提高对语言模型文件的查找效率,以 Bi-gram 语言模型为例,对文件中一元词和二元词建立的数据结构如下:

```
typedef struct Unigram
{ unsigned short gram;
unsigned char Unipro;
unsigned char Unibow;
unsigned int begposition;
unsigned int endposition;
} Unigram;
typedef struct Bigram
{ unsigned short gram;
unsigned char Bipro;
} Bigram;
```

Uni-gram 数据结构中的 gram 为无符号短整型,两个字节用来表示对应字符的 utf-8 编码;Unipro 为 unsigned char 类型,取值范围是 0 ~ 255,用来表示所对应一元词的转移概率的索引,同理 Unibow 用来表示回退概率的索引。要注意的是当字符没有回退概率时,Unibow 索引号被置为 0,对应的索引值为 101376(99×1024)。短整型的 beginposition 和 endposition 分别表示以当前字符开头的双字开始位置和结束位置,位置用距离开头的字节数来表示。所以当搜索双字词时,由于单字是按照 utf-8 编码的从小到大顺序排列的,采用二叉搜索在 7165 个单字中寻找到双字的第一个字,然后根据 beginposition 和 endposition 可以迅速地定位到所要寻找的双字。

Bi-gram 中的短整型的 gram 表示双字中第二个字符的 utf-8 编码,Bipro 表示该双字的转移概率索引(注:Bi-gram 模型中的双字词不存在回退概率)。

图 2 说明对 Bi-gram 建立多级索引的结构:每个一元词都记录一期开头的二元词的起始和终止位置,并且二元词也是按照字的 utf-8 编码顺序排列的,所以在查找时可以使用二分搜索,加快查找速度。

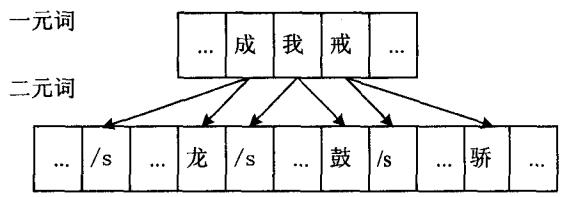

图 2　Bi-gram 模型建立索引机构示意

包含 7165 个单字符和 458382 双字词的 Bi-gram 语言模型文件压缩后的数据组织格式如图 3 所示。

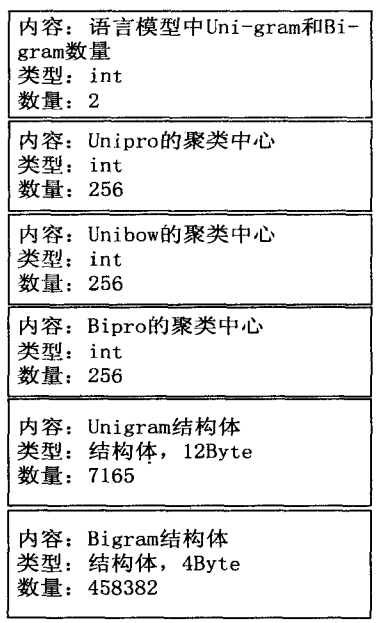

图 3　压缩后的语言模型组织结构

4 实验结果

通过对8M的Bi-gram ARPA格式的语言模型的压缩试验,压缩结果为1.4M,压缩率达到0.175。查找效率方面,同样以此语言模型文件进行理论上的分析。该语言模型统计的一元词有7165个,二元词有458382个。

由于词是按照其utf-8编码进行排序的,所以可以采用近似的二分查找方法进行匹配搜索。假设每个元素的查找是等概率的,则二分查找成功的平均比较次数为$\log_2(n+1)-1$[11],其中n为树的结点数。在语言模型中,由于n取值比较大,所以近似为$\log_2 n-1$。所以压缩之前采用平均比较次数为:

$$N_1=\frac{n_1}{n_1+n_2}(\log_2 n_1-1)+\frac{n_2}{n_1+n_2}(\log_2 n_2-1+\frac{1}{2}\times\frac{n_2}{n_1})$$

将数据代入公式,得到平均比较次数约为47.617。

采用在压缩过程中引入多级索引技术后,无需再在二元词中进行搜索,通过一元词的索引可直接定位。所以平均比较次数为:

$$N_2=\frac{n_1}{n_1+n_2}(\log_2 n_1-1)+\frac{n_2}{n_1+n_2}(\log_2 n_1-1+\log_2\frac{n_2}{n_1}-1)$$

代入数据得平均比较次数为16.714。搜索语言模型时所需要的平均比较次数是原来的0.35。当采用Tri-gram语言模型时,多级索引技术所带来的搜索效率上的提高将更加显著。

5 结束语

文中借鉴K-means聚类思想和多级索引方法,提出一种基于聚类和多级索引技术的ARPA格式的N-gram语言模型压缩方法。

根据ARPA格式N-gram语言模型的特点,对其组织方式进行了优化。对概率值进行聚类较好地解决了ARPA格式的N-gram语言模型所占空间过大的问题,而且多级索引技术的应用使模型获得了优良的检索效率。通过实验也验证了该方法的有效性和实用性。

参考文献:

[1] 李晓光,王大玲,于 戈.基于统计语言模型的信息检索[J].计算机科学,2005,32(8):124-127.

[2] Manning C,Schutze H.统计自然语言处理基础[M].苑春法,李庆中译.北京:电子工业出版社,2005.

[3] 殷芳刚,吴建国,吴海辉.Windows Mobile平台下智能手机输入法研究[J].计算机技术与发展,2011,21(5):75-78.

[4] Rosenfeld R. The CMU Statistical Language Modeling Toolkit [C]//Proc of ARPA Spoken Language Technology Workshop.[s.l.]:[s.n.],1995.

[5] Jelinek F,Mercer R L. Interpolated Estimation of Markov Source Parameters from Sparse Data[C]//Proc of Workshop on Pattern Recognition in Practice. Amsterdam:North-Holland,1980.

[6] Lafferty J D,Sleator D,Temperley D. Grammatical Trigrams:A Probabilistic Model of Link Grammar[C]//Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language. Cambridge,MA:[s.n.],1992:89-97.

[7] Ye Z X, Berger T. Information Measures for Discrete Random Fields[M]. Beijing:Science Press,1998.

[8] Kaufnman L,Rousseeuw P J. Finding group in data:an introduction to cluster analysis[M]. New York:Wiley,1990:83-88.

[9] 段小斌,林 雯,阮百尧,等.一种基于三级索引词库结构的中文分词方法研究[J].计算机与数字工程,2007,35(7):47-49.

[10] Brown P F,deSouza P V,Mercer R L,et al. Class-based n-gram models of natural language[J]. Computational Linguistics,1992,18(4):153-157.

[11] 王海涛,朱 洪.改进的二分法查找[J].计算机工程,2006(5):60-62.

(上接第24页)

[s.l.]:[s.n.],2010:677-683.

[5] Habib M A. 2010 International Conference on Internet Technology and Secured Transactions (ICITST)[C].[s.l.]:[s.n.],2010:1-6.

[6] Li Ninghui,Wang Qihua. Beyond separation of duty:An algebra for specifying high-level security policies[J]. Journal of the ACM,2008,55(3):1-4.

[7] 胡金柱,陈娟娟. RBAC模型中角色的继承与互斥问题的研究[J].计算机科学,2003(11):160-163.

[8] 付志峰,张焕国. RBAC系统中职责分离的实现[J].计算机工程,2003(6):61-63.

[9] 张 雷,向 宏,胡海波.基于语义的RBAC模型权限冲突检测方法[J].计算机工程与应用,2011(26):74-78.

[10] Sandhu R,Coyne E,Feinstein H,et al. Role-based Access Control Model[J]. IEEE Computer,1996,29(2):38-47.

[11] 段隆振,文 锋,黄水源,等.一种描述RBAC角色层次关系和互斥关系的模型及实现[J].南昌大学学报(理科版),2006(6):601-604.

[12] 卢 昱,王 宇,吴忠望.信息网络安全控制[M].北京:国防工业出版社,2011.