基于 n－Gram 的中文文本示例检索方法研究

刘晓丽　张佳骥

（信息产业部电子第 54 研究所）

摘要　用从文本中提取出的 n－Gram 统计特性来表示文本的内容特征，采用向量空间模型进行检索。所采用的技术方法简单易行，避免了许多自然语言处理的复杂问题。由于采用示例检索，用户只需提供感兴趣的一篇文章作为输入，无需构造查询式，减轻了用户负担。

关键词　文本处理　示例检索　n－Gram　向量空间模型

本文将统计学的方法 n－Gram 技术、VSM，与示例检索的思想结合起来应用于中文文本的检索，不但避免了许多自然语言处理的复杂问题，而且方便了用户的使用。实验结果表明本文采用的技术方法是行之有效的，性能指标达到了实用要求。

1　基本概况与组成

本文采用一种基于 n－Gram 统计特性的文本示例检索方法。用从文本中提取出的 n－Gram 来表示文本的内容特征。采用向量空间模型进行检索。首先提取所有文本的所有 n－Gram 项，从中提取出特征项，计算每个特征项在每个文本中的权重，将存储的所有文本都表示成特征向量，检索时将用户提供的样本文本也表示成特征向量（即为查询向量）。计算查询向量与每个文本特征向量的相似度，将相似度大于阈值的文本输出即为检索出的文本。基本组成如图 1 所示。

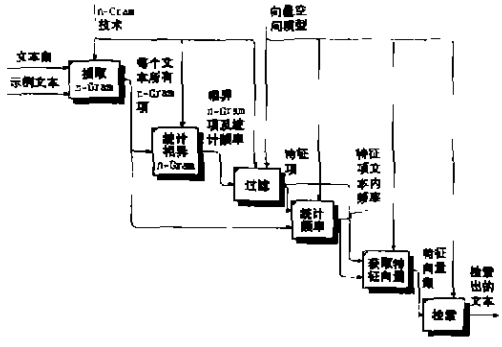

图 1　基本组成

2　n－Gram 技术

n－Gram 是目前最常用的统计语言学模型。最近几年国内才开始使用，一般只是应用于统计语言学方面，如汉语信息熵的计算，而用于文本检索方面极少。本文将 n－Gram 技术开创性地应用于中文文本处理，取得了很好的效果。

从文本字符流中连续截取 n 个字节，便得到该文本的一个长度为 n 个字节的字符串，称为该文本的 n－Gram。若有一宽度为 n 个字节（设 n 不超过文本的长度）的截取窗口置于文本上，截取的连续 n 个字节构成一个 n－Gram，则当窗口从文本的开头以 s 字节步长移到文本末尾时，把得到的该文本的所有 n－Gram 记作 Gram(n,s)。

文本的 Gram（n，s）显然与文本所包含的字、词、常用搭配以及相邻字、词之间的先后次序（n＞1 时）都有关系。因此，可以用 n－Gram 在文本中的分布特性来表示文本的特征。

可以看出，统计 Gram(n，s)时既不需要任何词典，也不需要对文本进行分词预处理，避免了切分歧义及未登录词等问题，而且可同时处理汉语字符与非汉语字符。n－Gram 包含一定的上下文信息，特别是包含有一定的词组以及相邻词的搭配信息，上下文相关性在文本检索中很重要。因此本文采用一种基于 n－Gram 统计特性的文本示例检索方法。

2.1　n 与 s 的选取

Gram(n，s)中 n 和 s 的选择，一般取 s＝1，这样可以包括文本中所有的 n－Gram，其中无效的项可以过滤掉。对于双字节编码的汉语文本，可以取 n＝4 统计 n－Gram，这样可以把占有较大比重的二字词包括在内。从实验结果也可以看出 n 取 4，s 取 1 时，检索性能最好，如图 2 所示。

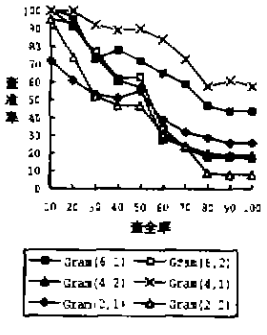

图 2　Gram(n,s)中 n,s 取不同值时的 P-R 曲线

2.2 文本中 n - Gram 总数

实际中 n - Gram 的数目至多与文本的长度成线性关系(当文本的长度为 L 个字节时,s = 1 时,文本中 n - Gram 的总数为 L - n + 1 个)。随着统计文本的不断加大,重复出现的项数会不断增多,因而不同 n - Gram 的数目将逐渐趋于饱和。本文滤除在所有文本中只出现一次的 n - Gram,过滤后不同 n - Gram 的数目是很小的,约为过滤前的 1/7。

3 向量检索

向量检索基于向量空间模型。向量空间模型(Vector Space Model, VSM)是关于文档表示的一个统计模型,是最简便高效的文本表示模型之一,近年来使用较多且效果也较好。

文本的内容特征常常用它所含有的基本语言单位(字、词、词组等)来表示,这些基本的语言单位统称为特征项,本文中特征项为过滤后相异 n - Gram 项。文本可以用特征项表示为 $D(T_1, T_2, \cdots, T_n)$。

对于含有 n 个特征项的文档,每一特征项 T_i 都根据其在文档中的重要程度赋予一定的权重 W_i,简记为 $D(W_1, W_2, \cdots, W_n)$。

给定一文档 $D(T_1, W_1; T_2, W_2; \cdots; T_n, W_n)$,将 $T_1, T_2, \cdots, T_n$ 看成一 n 维坐标系中的坐标轴,$W_1, W_2, \cdots, W_n$ 为对应的坐标值。这样由 $(T_1, T_2, \cdots, T_n)$ 分解而得的正交特征项向量组就张成了一个文档向量空间,所有文档和文档类都可映射到此文档向量空间。称 $D(W_1, W_2, \cdots, W_n)$ 为文档的向量表示或向量空间模型。

3.1 特征项的权重

本文利用特征项的统计信息,采用目前被广泛采用的、效果较好的 TFIDF 方法计算权重:

$$W_{ik} = tf_{ik} * idf_k \qquad (1)$$

其中 tf_{ik}(Term Frequency)表示第 k 个特征项在第 i 个文档中的出现频率;idf_k(Inverse Document Frequency)表示第 k 个特征项的逆文档频率,一般采用 $idf_k = \log(N/n_k)$。N 表示文档集中的文档数量,F_k 表示文档集中包含有第 k 个项的文档数。

考虑到文本长度的影响,将上式归一化:

$$W_{ik} = \frac{tf_{ik} * \log(N/F_k)}{\sqrt{\sum_{k=1}^{t} tf_{ik}^2 * \log(N/F_k)}} \qquad (2)$$

3.2 文本之间的相似度

两个文档 D_1 和 D_2 之间的(内容)相关程度(Degree of Relevance)常常用它们之间的相似度 $Sim(D_1, D_2)$来度量。当文档被表示为 VSM 时,可以用向量之间的内积来计算:

$$Sim(D_1, D_2) = \sum_{k=1}^{N} W_{1k} * W_{2k} \qquad (3)$$

4 文本示例检索实现方法

基于 n - Gram 统计特性进行示例检索功能模型如图 1 所示,具体实现方法如下:

①从文本开始按 n = 4,s = 1 产生每个文本的 Gram(n, s);

②统计出所有文本中包含的不同的 n - Gram 项,并统计各个 n - Gram 项在文本集中所出现的文本数;

③滤掉只在一个文本中出现的所有 n - Gram 项,剩余的相异 n - Gram 项即为特征项;

④统计各个特征项在每个文本中的出现频率;

⑤按式(2)计算每个特征项在每个文本中的权重,得到所有文本的特征向量表示;

⑥根据式(3)求出文档向量与查询向量之间的相似度,然后选出相似度大于阈值的文档作为检索结果。

5 实验结果

要计算检索系统的性能指标查全率(recall ratio)和查准率(precision ratio),必须预知每个示例文本的相关文本。实验中通过不同人对同一示例文本分别进行检索,查找文本集中该示例文本的相关文本,将检索结果汇集起来,加以审查,筛选出相关文本,最后生成每个示例文本的相关文本总集,从而根据检索结果计算出每个示例文本的查全率与查准率。

对于每一查询,都有一条查全率 - 查准率曲线与之对应,最后得到所有样本查询的平均曲线,将该曲线作为评价检索性能的标准。图 3 为查全率分别为 10、20、30、40、50、60、70、80、90、100 时,查准率的曲线。从图中可以看出,曲线下降缓慢,基本保持水平,说明检索性能很好,平均精度达到 88.8%。

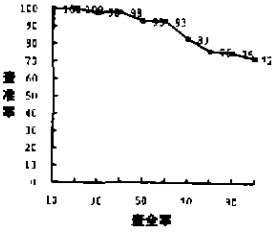

图 3 检索性能曲线

6 结论

本文把对文档内容和查询要求的处理简化为向

量空间中向量的运算，而权重的计算通过统计方法自动完成，使问题的繁杂性大为减低。另外本文采用向量空间模型统一描述文档表示和查询表示过程，可以有效地基于示例样本进行检索。检索结果可以按相似度大小排序输出，便于用户根据需求控制检索量。由于以相似的程度作为检索的标准，可以从量的角度判断命中与否，改变了布尔检索非1即0的简单判断，显然具有模糊检索的特点。

从实验中可以看出采用n－Gram技术以及向量空间模型进行中文文本的示例检索是行之有效的，性能指标达到实用要求。本文采用统计学的技术方法，避免了许多自然语言处理的复杂问题，简单易行。另外重要一点是，本文采用示例检索，不需要用户构造查询式，大大减轻了用户的负担，方便用户的使用。

参 考 文 献

1 Dunja Mladenic. Text－Learning and Related Intelligent Agents: A Survey. IEEE Intelligent Systems, July/August 1999:44～54

2 林鸿飞，战学刚，姚天顺．文本结构分析与基于示例的文本过滤．小型微型计算机系统．Vol.21.No.4,2000.4

3 William B. Cavnar. N－Gram－Based Text Filtering for TREC－2. Second Text Retrieval Conference (Proc. of TREC－2), 1994

4 周水庚，关佶红，胡运发．基于文档实例的中文信息检索．计算机工程与应用，2000,10:14～29

5 李业丽，林鸿飞，姚天顺．基于示例的用户信息需求模型的获取和表示．计算机工程与应用，2000,9:11～16

6 Jonathan D. Cohen. Recursive Hashing Functions for n～Grams ACM Transaction on Information Systems. Vol. 15. No. 3, July 1997:291～320

(上接第11页)

$$e_t(\tau)=\begin{cases}=0 & \tau=0\text{ 时}\\ >0 & \tau>0\text{ 时}\\ <0 & \tau<0\text{ 时}\end{cases} \quad (3)$$

可见，$e_t(\tau)$刚好反映了抽样定时处在正确、超前、滞后三种状态。误差函数的均值可以表示为

$$E_t(\tau)=\frac{4G_R}{T}\sin 2\pi\frac{\tau}{T} \quad (4)$$

其鉴相特性为正弦曲线。

由于相邻两码元之间的信号波形是斜变的，$e_t(\tau)$正好反映了定时偏差。码元定时的偏差很小时，$e_t(\tau)$与定时误差的偏差τ呈一定的线性关系。可以验证，在偏差不太大的情况下，进行线性调整，就可以获得位定时。

在8DPSK传真信号中，误差信号的检测发生在码元$+1\rightarrow-1$或$-1\rightarrow+1$的跃变的时候。把检测的误差信号通过环路滤波器，进行平滑处理就得到定时误差。用它来矫正判决点的选通时刻，可以实现位定时的跟踪。

1.2 插值算法

插值算法可以采用F.M.Garnder的“Interpolation in Digital Modems－Part Ⅰ，Part Ⅱ”中介绍的数字内插法，实现对定时信号的调整。数字内插的抽样关系如图3所示。在非同步抽样中加入内插，可以得到正确的选通脉冲值

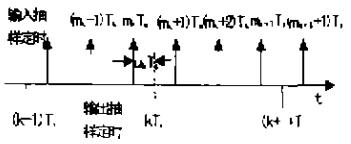

图3 抽样时间关系

内插器滤波器的种类很多，多项式滤波器是其中之一。多项式滤波器可以表示成

$$h_I(t)=\sum_{i=0}^{N}b_I(i)\mu_k^i \quad (5)$$

对于三阶内插，输出为

$$y_{(k)}=\{[V(3)\mu_k+V(2)]\mu_k+V(1)\}\mu_k+V(0) \quad (6)$$

式中 $V(i)=\sum_{i=1}^{n}b_I(i)x(m_k-i)$。 (7)

其中三阶内插系数$b_I(i)$由附表给出。

可见，对于每个内插，只需要确定μ_k就可以完成。而μ_k可以由误差检测的结果计算得出。

附表 三阶内插系数

i	I=0	I=1	I=2	I=3
－2	0	－1/6	0	1/6
－2	0	1	1/2	－1/2
0	1	－1/2	－1	1/2
1	0	－1/3	1/2	－1/6

2 仿真结果

用上述方法对8DPSK传真数据进行定时恢复，取得了较好效果。仿真显示，在训练期间可以恢复位定时，数据传输时实现位定时跟踪，并且运算量不大。

参 考 文 献

1 Floyd M. Gardner, fellow, IEEE. Interpolation in Digital Modems－Part I: Fundamentals. IEEE Transactions on Communications. Vol. 41. No. 3, march 1993:501～507

2 Lars Erup. Member. IEEE, Floyd M. Gardner. fellow. IEEE and Robert A. Harris. Member. IEEE. Interpolation in Digital Modems－Part II: Implementation and Performance. IEEE Transactions on Communications. Vol. 41. No. 6, June 1993: 998～1008

3 朱向东，王士林．可变速率QPSK Modem中码元定时恢复电路的研制．通信工程学院学报，1992.6,6(1):48～56