LIBRARY AND INFORMATION SERVICE
Vol. 48, No. 8, August, 2004

一种基于 N－Gram 改进的文本特征提取算法

于津凯　王映雪　陈怀楚
清华大学计算机与信息管理中心　北京 100084

〔摘要〕介绍一种改进的文本特征提取及匹配算法。该算法基于 N－Gram 算法思路进行文本处理和特征提取，设计了 gram 关联矩阵用于统计与合并特征词，从而在固定长度 N－Gram 算法的基础上能够提取出不同长度的特征词。实验证明，该特征提取算法能够更为准确地描述文本特征，可应用于文本检索、Web 挖掘等信息处理领域。

〔关键词〕文本特征提取　N－Gram 算法　gram 关联矩阵
〔分类号〕TP391

An Improved Text Feature Extraction Algorithm Based on N－Gram

Yu Jinkai　Wang Yingxue　Chen Huaichu
Computer & Information Management Center, Tsinghua University, Beijing 100084

〔**Abstract**〕This paper introduces an improved text feature extraction algorithm based on N－Gram theory. It designs a gram correlative matrix to unite the consecutive bigrams into a multigram and breaks the limit of N－Gram which has fixed－length gram extractions and forms the multigram features.

〔**Keywords**〕text feature extraction　N－Gram algorithm　gram correlative matrix

1　引　言

文本特征提取与匹配是文本检索和文本挖掘任务中基础性和关键性的工作。文本特征提取抽取信息的特征，并表示为统一的方式，可以有效地降低文本向量空间维数，简化计算，防止过分拟合，是文本类共性与规则的归纳过程。

常见的文本特征提取算法，包括基于自然语言理解的文本语义理解技术、基于关键词列表和专业词典的分词匹配技术以及基于纯统计学方法的无意义文本分解技术。由于汉语中字词分隔不明显，歧义较多，词序、语序的自由度较高，因此语义理解技术和分词技术在中文文本环境中的应用都存在一定困难，而基于纯统计学的 N－Gram 算法，可以绕过分词的障碍，具有较高的实用性。

本文讨论 N－Gram 算法的特征及优缺点，并在此算法基础上提出一种改进的文本特征提取算法，通过统计并合产生多字特征词，从而较好地解决了 N－Gram 算法在多字词方面的缺陷，使特征提取过程能够获取更为准确有效的特征向量。

2　N－Gram 算法

N－Gram 算法的基本思想，是将文本内容按字节流进行大小为 N 的滑动窗口操作，形成长度为 N 的字节片断序列，每个字节片断称为 gram，对全部 gram 的出现频度进行统计，并按照事先设定阈值进行过滤，形成关键 gram 列表，即为该文本内容的特征向量空间，列表中的每一种 gram 均为一个特征向量维度。

N－Gram 算法具有如下优点：①语种无关性，可以同时处理中英文、繁简体文本。② 不需对文本内容进行语言学处理。③ 对拼写错误的容错能力强。④ 勿需词典和规则。

根据语言学方面的统计，约 70% 左右的中文词汇是双字词，因此在进行中文文本处理中，大多采用双字词进行分解，称之为 bigram，下文中所指 N－Gram 算法，均采用 bigram 切分方式。由于汉字是双字节字符，因此取 N＝4，即以 4 字节为单位进行字节片断划分。首先要对文本语料按中英文和语段标点进行切分，将原文由大段文本切分为语段序列，即相对逻辑独立的单句或区段；再对每一个语段进行 bigram 切分，即可获得 gram 列表，如图 1 所示。

由于 N－Gram 算法的采用长度固定为 N 的窗口进行切

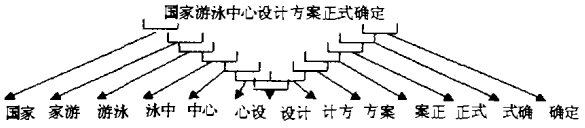

图1 bigram 切分示例

分,因此对长度恰好为 N 的特征词处理较好(即 bigram 方式对双字词的处理具有较高的准确性),但对长度大于或小于 N 的特征词,就会进行切分处理,从而造成语义和语序方面的一些偏差,在后续的检索或分类过程中就会产生错误的结果。就 bigram 方式而言,三字词、四字词的处理就会存在上述问题。

例如,某篇介绍专用设备"潜水箱"的文章,按照 bigram 划分,产生特征词"潜水"和"水箱",而这两个特征词与"潜水箱"的语义存在较大差距。再例如,"数据挖掘"是一种数据处理的专用算法,但切分后产生的特征词为"数据"和"挖掘",有可能检索(或分类出)"煤矿挖掘过程监测数据"这样的内容,从语义和语序两方面看都有错误。

虽然仅有不到 20% 的中文词汇属于多字词(即字数大于 2 的词汇),但在专业研究领域内,往往这些多字词才是文章的核心特征,对它的错误处理会导致较大的负面影响。

3 基于 N-Gram 的改进算法

鉴于 N-Gram 算法,特别是中文应用中主要采用的 bigram 切分方法,在处理多字词时存在上述问题,为了在特征提取过程中获取更为准确有效的特征向量,我们提出一种基于 N-Gram 的改进算法。该算法通过统计与合并双字特征词,产生多字特征词,较好地弥补 N-Gram 算法在多字词方面的缺陷。

3.1 算法的基本思想

基于 N-ram 的改进算法,是在进行 bigram 切分时,不仅统计 gram 的出现频度,也要统计某个 gram 与其前相邻 gram 的情况,将此记录在 gram 关联矩阵中。全部文本处理完毕后,通过处理 gram 关联矩阵,发现哪些 gram 是经常接续出现的,如果接续出现频率大于事先设定的阈值,则将其合并成为多字特征词。

由于绝大多数的多字词属于三字词和四字词,五字以上的词汇很少出现,即使出现也可以划分为两个以上的四字以下词汇而并不损耗信息量,所以本算法中仅处理三字词和四字词的情况:分别采用两个 gram 关联矩阵进行记录,处理三字词的 gram 关联矩阵为 A,是二维矩阵;处理四字词的 gram 关联矩阵为 B,是三维矩阵。

本算法中使用到两个阈值 α 和 β,其中 α 代表特征词频度限制,需要根据文本长度进行确定。β 代表合并阈值比例,即如果两个 gram 接续出现的频度达到这两个 gram 各自出现频度的规定比例 β,则认为这两个 gram 应该合并为一个特征词。经实验检验,β 取值为 70% 效果较好,α 取值取决于应用方向和需求。目前,α 采用 4 倍的平均频度作为特征词阈值。

在 N-Gram 算法的使用过程中,由于切分出的 bigram 数量极为可观,而 gram 关联矩阵属于二维或三维矩阵,维度向量即为切分出的 bigram,空间复杂度很高,因此在本算法实现过程中,采用稀疏矩阵方式存储 gram 关联矩阵,从而有效地降低了空间占用量。算法流程如图 2 所示:

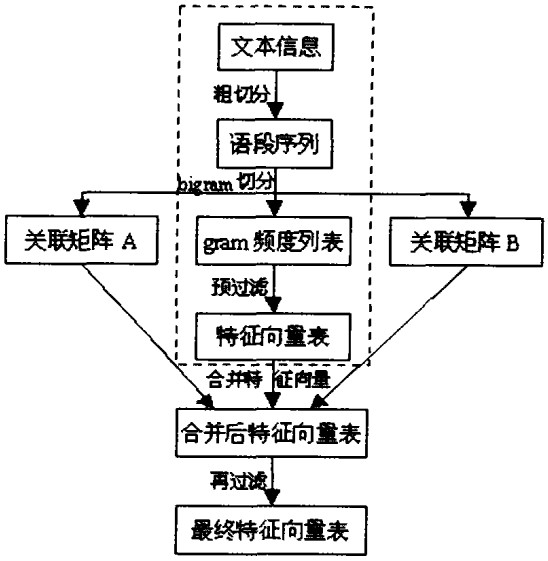

图2 基于 N-Gram 的改进算法流程

需要说明的是,图中虚线框内为一般的 N-Gram 算法流程,本算法在此流程后增加了两个环节。

3.2 具体的算法流程

- 对文本语料按中英文和语段标点进行粗切分,将原文由大段文本切分为语段序列,即相对逻辑独立的单句或区段。
- 对语段进行 bigram 切分,产生的 gram 片断记之为 G_i,同时统计其出现频度,记为 F_i。
- 如果当前语段内,G_i 之前存在一个 gram(G_{i-1}),则记入 gram 关联矩阵 A(G_{i-1},Gi,m),其中 m 代表 G_{i-1} 与 G_i 接续出现的频度。
- 如果当前语段内,G_i 之前存在两个 gram(G_{i-2},G_{i-1}),则记入 gram 关联矩阵 B(G_{i-2},G_{i-1},G_i,n),其中 n 代表 G_{i-2},G_{i-1},G 接续出现的频度。
- 重复过程 2-4,直至全部文本语段处理完毕。
- 预过滤:选择出现频度 F_i 大于设定阈值 α 的 gram 作为特征向量,删除所有不符合条件的 gram,同时在矩阵 A、B 中删除出现这些 gram 的记录,此时形成特征向量表。
- 合并特征向量:从矩阵 A(x, y, m)中,找出符合以下条件的记录:

$m > criteria$ 且 $m > \alpha$,其中 $criteria =$

$$\begin{cases} \beta \times F_x, & x \text{ 为多字词} \\ \beta \times F_y, & y \text{ 为多字词} \\ \beta \times \max(Fx, Fy), & x, y \text{ 均为双字词} \end{cases}$$

- 对符合条件的记录进行如下操作：

a. 生成一个新的特征向量，记为 xy，频度 $F_{xy} = m$；

b. 原特征向量 x 频度 F_x 变更为 $F_x' = F_x - m$；

c. 原特征向量 y 频度 F_y 变更为 $F_y' = F_y - m$；

d. 从矩阵 B 中，寻找 gram 内容与排列次序相同的记录并移入矩阵 A，即所有 B(x, y, z, n) → A(xy, z, n)，B(w, x, y, n) → A(w, xy, n)；

e. 从矩阵 A 中删除本记录。

- 重复此前两个过程（合并特征向量和对符合条件的记录进行 a－e 步骤的操作），直至从矩阵 A(x, y, m) 中找不到满足条件的记录，此时形成合并后特征向量表。

- 再过滤：由于上述过程改变了特征向量的频度，因此需要再次过滤。从现有特征向量中，保留出现频度 F_i 大于设定阈值 α 的特征向量，并按出现频度排序，即产生最终的特征向量表。

4 实验结果分析

算法实验过程共处理各种类型的文本 5 172 篇，内容包含约 300 万个汉字，其中新闻内容最长为 1.6 万字，最短为 23 字，平均为 580 字。结果抽样分析证明，多字特征词获取正确率（即经人工判断语义有效的多字特征词比例）平均为 91.27%，特征向量维度平均减少 13.4%。

而采用普通的 N－Gram 算法，双字特征词的正确率平均仅为 68.47%。

在此，我们选择一篇公告《北京奥运行动规划》（1.6 万字）、一篇论文《SLIQ 高速可伸缩数据挖掘分类模型在 AMINER 系统中的应用》（1 万余字）和一篇综述《决策支持系统的发展》（1.2 万字）来比较本算法与 N－Gram 算法，结果如下：

从比较结果来看，在采用相同阈值的前提下，改进算法在没有损耗信息量的同时减少了特征词向量维度，使特征词更具准确性和代表性。同时，从结果可以看出，对于专业文献，多字特征词的出现频率和代表性更高，采用本算法的效果更好。

* 本实验所用计算机配置为：P III－800MHz，256M 内存，40G 硬盘。

5 结 论

从实验结论来看，本算法能够较好地解决 N－Gram 算法在多字词方面的缺陷，使特征提取过程能够获取更为准确有效的特征向量，也就是能够更为准确地描述文本特征。特征向量包括双字特征词和多字特征词，其中多字特征词的正确率和代表性要好于 N－Gram 算法，特别是对学术论文、技术类文章等专业文献的处理效果更好。在文本检索、Web 挖掘等信息处理领域，文本特征提取算法是其中的关键算法，可以成功地应用于实际处理过程中。

表 1 N－Gram 算法与改进算法实验结果比较

算法	文章类型	公告	论文	综述
N－Gram 算法	特征词数量	181	120	140
	前 10 个特征词	奥运 (177) 运会 (114) 建设 (111) 城市 (68) 技术 (50) 北京 (49) 设施 (47) 环境 (47) 发展 (46) 文化 (41)	数据 (132) 算法 (61) 一个 (55) 节点 (52) 决策 (46) 策树 (41) 分类 (38) 系统 (38) 模式 (35) 采掘 (33)	数据 (117) 处理 (34) 程序 (29) 分析 (29) 系统 (27) 一个 (27) 体系 (27) 结构 (25) 系结 (24) 文件 (23)
改进算法	特征词数量	173	100	120
	其中多字词数量	15	20	22
	特征词减少比例	4.4%	16.7%	14.3%
	前 10 个特征词	奥运会 (114) 建设 (59) 城市 (52) 发展 (41) 环境 (36) 文化 (36) 国际 (34) 北京市 (33) 奥林匹克(33) 现代化 (32)	数据 (49) 一个 (44) 决策树 (41) 系统 (38) 算法 (33) 模式 (30) 节点 (27) 数据采掘(27) 分类 (26) 训练集 (24)	数据 (54) 系统 (23) 结构 (17) 应用程序(17) 一个 (16) 自然演化(16) 体系结构(16) 抽取程序(14) 问题 (13) 可信性 (13)
	多字特征词占前 10 个特征词比例	40%	30%	50%
	多字特征词占全部特征词比例	8.7%	20%	18.3%
	多字词正确率	93.3%	95%	90%

但是，本算法也同样具有基于统计的特征提取算法的通病，即无法从语义角度识别和区分文本特征。例如，“球员”这个特征词在不同的上下文环境中代表的是不同的球员个体，甚至不同种类运动的参与者，而本算法无法对此进行区分，在检索和分类时可能造成一定混淆，需要借助其他特征词加以区分。

此外，本算法由于在 N－Gram 算法基础上增加了两个环节，同时使用了两个稀疏矩阵，因此算法效率比 N－Gram 算法低（经测试，处理速度约低 13%）。下一步的重点工作，就是进一步提高算法效率。同时，还需研究阈值取值对效率和准确性的影响，进而寻找最优阈值。

参考文献：

1 M. Damashek, Gauging similarity with N－Grams: language－independent categorization of text. Science, 1995(267): 843－848

2 A. Chen et al., Chinese text retrieval without using a dictionary. See: Proceedings of the 20th Annual International ACMSIGIR Conference on Research and Development in Information on Retrieval, 1997: 42－49

（下转第 43 页）

4.3 各成员馆要分工明确

在联盟内部,分工与协作有利于各图书馆间优势互补,形成更为有效的专业化分工,发挥规模效益,以使信息产品的整体成本降低,从而使图书馆联盟成员实现各自的"低成本"和"专业化"的发展战略[9]。

4.4 建立相互信任机制

对一个成功的图书馆联盟,其内部成员之间的相互信任被视为必要的前提。沙贝尔(Sabel)认为:"相互信任就是联盟各方坚信,没有一方会利用另一方的脆弱点去获取利益"。这就充分的说明了,图书馆联盟是各方在面向不确定的未来时所表现出的彼此间的信赖,它的形成需要长期的努力培养[10]。

- 设计联盟内部信任的评审体系。在缔结图书馆联盟的过程中,必须通过一套经常性的、持续的内部评估审核分析体系对每个成员馆的过去、现在和将来等一系列要素进行综合评估,以建立可靠的相互信任机制。

- 建立相互信任的产生机制。相互信任的产生机制主要有3种类型:过程型、特征型和规范型。要建立该机制,就要做到与联盟的属性和要求相互吻合,相互适应,以确保整个联盟形成协同效应。其中,过程型机制,是从联盟的创建、成长到成熟阶段,联盟各方的相互信任关系始终贯穿其中。特征型机制,是具有不同背景特征的图书馆构成不同特性的联盟,必须通过管理培训、鼓励非正式接触、提高行为和策略的透明度等措施来努力消除彼此的隔阂。规范型机制,就需要在联盟内建立一套阻止相互欺骗和停止机会主义行为的规范机制,以加强它们之间的合作。

4.5 全面有效的沟通

若图书馆联盟获得成功,应该确保各种信息能在联盟图书馆之间沟通与传播,努力做到完全信息下的博弈。这样可以提高图书馆对联盟的兴趣,求得对联盟的支持。而且通过信息沟通,也能够促进相互知识的增长,形成学习的新优势。同时,图书馆联盟领导者要尽可能促进信息在联盟内部沟通。

- 互相理解。联盟中友好相处的最重要原则就是从合作方的角度去分析他们的行为或立场,而不应从自己的角度出发,不要轻易地做出对合作方价值观的任何判断,要尊重对方的观点。

- 组织交流。要有意识地进行双向交流,达成对某些问题的共识,这是降低由差异性产生的理解障碍的有效手段。为长久地维持联盟关系,应组建一系列委员会来维护新的合作关系,如指导委员会、协调委员会等,以此来缓解双方的矛盾。

参考文献:

1 于湖滨,张 军. 图书馆共享联盟的形成与发展. 图书馆杂志,2001(9):60-61
2 戴龙基,张红扬. 图书馆联盟——实现资源共享和互利互惠的组织形式. 大学图书馆学报,2000(3):36-39
3 白君礼,李志俊. 文献信息资源共享的博弈分析. 图书情报工作,2003(8):35-38
4 张汉江,马超群等. 信贷行为中的不完全信息动态博弈. 系统工程理论与实践,1999(5):7-8
5 邵祖峰,潘祥杰. 企业战略联盟有效实施的博弈分析. 科技进步与对策,2002(6):97-98
6 周明华,谢春枝. 美国大学图书馆联盟研究. 中国图书馆学报,2003(5):77-82
7 林 嘉. 网络环境下图书馆联盟建设的思考. 中国图书馆学报,2003(2):32-34,44
8 屠 航. 动态联盟:图书馆馆际合作的新途径. 情报杂志,2003(3):72-73
9 董 军. 非零和博弈与战略联盟. 东方企业家,2001(2):52-56
10 邱 毅. 跨国公司战略联盟的博弈诠释. 贵州财经学院学报,2001(6):77-80

〔作者简介〕蒋丽艳,女,1978年生,硕士研究生,发表论文12篇。

(上接第50页)

3 W B Cavnar. N-Gram based text filtering. See: Proceedings of the Second Text Retrieval Conference (TREC2, D. K. Harman Edition), Place National Institute of Standards and Technology, 1993: 171-179
4 W B Cavnar, J M Trenkle. N-Gram based document categorization. See: Proceedings of the Third Symposium on Document Analysis and Information Retrieval. LasVegas, 1994:161-176
5 何 浩,杨海棠. 一种基于N-Gram技术的中文文献自动分类方法. 情报学报,2002(4):421-427
6 J D Cohen. Recursive Hashing functions for N-Gram. ACM Transaction Information Systems, 1997, 15(3):291-320
7 Kneser R, Ney H. Improved backing-off for N-Gram language modeling. See: Proc ICASSP 95[C] 1995:181-184
8 周 强. 规则与统计相结合的汉语词类标注方法. 中文信息学报,1995,9(2):1-10

〔作者简介〕于津凯,男,1978年生,硕士研究生。
王映雪,女,1945年生,研究员,发表论文25篇。
陈怀楚,男,1971年生,工程师,发表论文16篇。