中 文 信 息 学 报
第 15 卷 第 1 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 15 No. 1

基于 N-gram 信息的中文文档分类研究①

周水庚[1] 关佶红[2] 俞红奇[1] 胡运发[1]

(1. 复旦大学计算机系 上海 200433; 2. 武汉大学计算机学院 武汉 430072)

摘要:传统文档分类系统都是基于文档的词属性,分类过程需要庞大的词典支持和复杂的切词处理。本文研究基于 N-gram 信息的中文文档分类,使中文文档分类系统摆脱对词典和切词处理的依赖,从而实现中文文档分类的领域无关性和时间无关性。利用 kNN 分类方法,实现了一个基于 N-gram 信息的中文文档分类系统。测试结果表明该文档分类系统具有和其它同类文档分类系统相当的性能。

关键词:文档分类;N-gram 信息;属性选择;kNN 法

中图分类号:TP 391.1

Chinese Documents Categorization Based on N-gram Information

ZHOU Shui-geng[1] GUAN Ji-hong[2] YU Hong-qi[1] HU Yun-fa[1]

(1. Computer Science Department, Fudan University Shanghai 200433;
2. School of Computer Science, WTUSM Wuhan 430072)

E-mail: sgzhou@fudan.edu.cn

Abstract: Traditional document classifiers are based on keywords in the documents, which need dictionaries support and efficient segmentation procedures. This paper explores the problem of utilizing N-gram information to categorize Chinese documents so that the classifiers can shake off the burden of large dictionaries and complex segmentation procedures, and subsequently be domain and time independent. Such a Chinese documents categorization system is implemented with kNN classification method. Experimental results show that it can achieve comparable performance to other classifiers of the same type.

Keywords: text categorization; N-gram information; feature selection; kNN method

一、概述

文档分类已经成为处理和组织大规模文档数据的关键技术。现有文档分类技术基本上是

① **收稿日期**:2000-03-16;**修改稿收到日期**:2000-05-16
基金项目:国家自然科学基金(69933010);国家 863 计划(863-306-ZT04-02-2)
作者周水庚,男,1966 年生,博士,副教授,主要研究方向为数据库、数库仓库、数据挖掘和信息检索.

基于词信息[1~3],这使得文档分类需要借助于词典和使用专门的词提取技术。就中文文档分类而言,词提取的关键为切词(或分词)技术[4]。切词是一项复杂的工作,现有切词系统一般都比较复杂和庞大,切词速度慢,准确度也不高。考虑到语言的领域相关性和随时间变化特性,一个文档分类系统需要不断修正和扩充词典并改进切词技术,以适应语言环境变化和语言发展的需要。如果能够在无需词典支持和切词处理的情况下进行中文文档分类,则文档分类系统无疑具有更广泛的适用性和更持久的生命力。

为此,本文提出基于中文文本的 N-gram 信息进行中文文档分类。由于 N-gram 信息获取简单,用 N-gram 信息进行分类可使文档分类系统摆脱对复杂切词处理程序和庞大词库的依赖。利用 kNN 法建立了中文文档分类系统。测试结果表明:用 N-gram 信息进行分类可以得到和其它同类分类器相当的分类结果。

二、N-gram 信息提取

2.1 N-gram 信息与中文文档分类

对于长度为 L 的中文文档 d,若不考虑标点符号和其它各种字符,也就是说,文档是长度为 L 的汉字序列,那么,这个文档中包含的 N-gram 信息项总共为 $L(L+1)/2$。由此可见,文档中包含的 N-gram 信息项非常丰富。这提醒我们,在用 N-gram 信息进行文档分类时,必须有所选择。从另一方面来看,文档分类是面向语义的操作,因此,用于文档分类的文档属性应该能够尽可能地表现文档的语义。通常情况下选择词进行文档分类就是因为词是表征语义的最小语言单元。

显然,并不是所有出现在文档中的 N-gram 信息项都对分类有用。一个 N-gram 信息项对分类的有用性(或者分辨度)可以从三个方面来衡量:频度、分散度和集中度。下面分别给出它们的定义。

定义 1 (频度)在文档 d 中,N-gram 信息项 t 的频度用它在 d 中出现的次数 tf 表示。

定义 2 (分散度)在文档类 c 中,N-gram 信息项 t 的分散度用 c 中包含 t 的文档数目 df 表示。df 越大,则 t 在 c 中越分散;反之,越不分散。

定义 3 (集中度)在文档集 D 中,N-gram 信息项 t 的集中度用 D 中包含 t 的文档类数目 cf 表示。cf 越小,则 t 在 D 中的越集中;反之,越不集中。

直观地,对于 N-gram 信息项 t,其频度越高、分散度越大、集中度越强,则对分类越有用,即分辨度越强。不过,目前还没有找到很好的数学方法来综合频度、分散度和集中度这三个因素,使得选出的文档属性能够获得最优分类效果。为了减少提取不必要的 N-gram 信息,在提取 N-gram 信息项时加如下两个约束条件:

约束 1 对于预先给定的最小频度值 $min\text{-}tf$,在文档 d 中某一 N-gram 信息项 t 被提取的先决条件是它在 d 中的 $tf \geq min\text{-}tf$。

约束 2 对于预先给定的最小分散度值 $min\text{-}df$,在文档类 c 中某一 N-gram 信息项 t 被提取的先决条件是它在 c 中的 $df \geq min\text{-}df$。

测试中,我们一般取 $min\text{-}tf$ 和 $min\text{-}df$ 为 2。

2.2 N-gram 信息提取方法

一种直接的 N-gram 信息项提取方法是只扫描一遍文档,一次性将所有满足上述两个约束条件的 N-gram 信息项取出。由于只需扫描一遍文档,这种方法对于较小的训练文档库是

有效的。但对于大训练库,则需要很大的内存空间。否则,就得在内存和外存之间不断交换中间结果。这里采用一种分步提取的方法,其基本思想是:先提取符合约束条件的 1-gram 信息项;从选择得到的 1-gram 信息项构造候选 2-gram 信息项,剔除其中不符合约束条件的候选项,得到真正需要的 2-gram 信息项。依此方法,提取其它 N-gram(N = 3,4,...)信息项。限于篇幅,具体算法略。

2.3 参数 N 的选取

使用 N-gram 信息进行文本分类,最基本的要求是所选择的 N-gram 信息项能够覆盖文档中的词。因此,并非 N-gram 信息项越多越好。这就涉及如何选择参数 N 的问题。统计分析表明,在中文文档中,主要词条为 1-字、2-字、3-字和 4-字词条[5]。因此,用这些词条可较完整地表达文档语义。这意味着在用 N-gram 进行文档分类时,只需取 1-gram、2-gram、3-gram 和 4-gram 信息项,因为这些 N-gram 信息项可以覆盖文档中所有 1-字、2-字、3-字和 4-字词条。

三、文档属性选择

我们把从文档中提取的 N-gram 信息项都称作文档的属性。尽管提取文档属性时加了约束,但文档的属性量仍将很大,特别是训练文档库比较大的时候。有些属性对文档分类的作用可能并不大。因此,我们有必要对前面提取到的文档属性进行筛选,选出那些最能代表文档类别概念的属性。这一过程就是文档属性的选择。

3.1 属性选择方法

我们使用三种方法进行文档分类属性选择。这三种方法分别是:信息增量、互信息和 χ^2 统计。目的是通过测试比较,找到比较好的文档属性选择方法。其中,前两种基于信息论;后一种基于统计分析。这些方法已经在西文文档分类种得到普遍应用[6]。

属性选择的具体步骤为:

1)按第 2 节中的 N-grams 提取方法,从训练文档库中取得所有 N-gram(N = 1～4)项,构成文档属性集合 F;

2)对集合 F 中的每一项用下列某一种方法进行打分。譬如,选用信息增量方法,则对 F 中的任意 N-gram 项 f,求 $IG(f)$。当 F 中的所有项都打分完成后,按分值由高到低进行排序;

3)对 F 按性质 2 进一步削减冗余项,得到属性集合 F_1;

4)假设需要选取 N 个分类属性,则从 F_1 中的选取分值高的 N 个项,构成最终的分类属性集 F_s。F_s 将用于分类训练和测试。

3.2 属性的进一步削减

通过上述方法选出的文档属性,还可以做进一步的削减。在文档中,有些约定俗成的词和专用名词,组成这些词的字都是同时出现的。比如"巴基斯坦",只要"巴基斯坦"被选中,则"巴基"和"巴基斯"肯定也被选中。因此,它们之中有两个是冗余的,应该抛弃掉。下面给出性质 2,说明如何进一步削减冗余文档属性。

性质 2 若有两个 N-gram 项 t_i 和 t_j,满足 $t_i \supset t_j$ 和 $score(t_i) = score(t_j)$,则 t_i 和 t_j 中有一个是冗余的,只需取 t_i。

上面,score(.)表示(1)～(6)中的任何一个表达式。譬如选式(1),则有 $score(t_i) = IG(t_i)$; $score(t_j) = IG(t_j)$。

四、基于 kNN 法的文档分类

在 kNN 文档分类方法中,所有文档均用向量空间模型表示。因此,一个文档就是文档向量空间中的一个向量,这个向量也称为文档向量。文档向量中各个维对应于用于表征文档的各个词(词组),这也就是文档属性。对于某一具体文档,其向量中各个维的值为该向量维对应的词在文档库中的权值。

对于文档库 D,假设对应的文档属性集为 V, $V=\{W_i\}(i=1\sim n)$。现有一文档 d,用向量模型表示为:

$$\vec{d}=(w_1, w_2, \cdots, w_n) \tag{1}$$

上面, $w_i(i=1\sim n)$ 为属性 W_i 对应的权值。权值计算一般采用 $TFIDF$ 估算方法,即

$$w_i=\frac{tf_i \times \log(N/n_i)}{\sqrt{\sum_{i=1}^{n}(tf_i)^2[\log(N/n_i)]^2}} \tag{2}$$

这里, N 为训练文档总数; tf_i 表示 W_i 在训练文档 d 中的频数; n_i 为训练文档中包含有 W_i 的文档数。很显然,用式(2)计算文档属性权值得到的文档向量(1)为一单位向量。这样,文档 d_i 和 d_j 间的相似度可用余弦公式表示为:

$$Sim(d_i, d_j)=\vec{d_i}\cdot\vec{d_j} \tag{3}$$

kNN 方法进行文档分类的过程如下:对于某一给定的测试文档 d,在训练文档集中,通过相似度找到与之最相似的 k 个训练文档。在此基础上,给每一个文档类打分,分值为 k 个训练文档中属于该类的文档与测试文档之间的相似度之和。也就是说,如果在这 k 个文档中,有多个文档同属于一个类,则该类的分值为这些文档与测试文档之间的相似度之和。对这 k 个文档所属类的分值统计完毕后,即按分值进行排序。还应当选定一个阈值,只有分值超过阈值的类才予以考虑。测试文档属于超过阈值的所有类。形式化表示为:

$$score(\vec{d}, c_i)=\sum_{\vec{d}\in kNN} Sim(\vec{d}, \vec{d_j})\, y(\vec{d_j}, c_i) - b_i \tag{4}$$

式中 $y(\vec{d_j}, c_i)=\begin{cases}1 & \vec{d_j}\in c_i\\ 0 & \vec{d_j}\notin c_i\end{cases}$

b_i 为阈值;

$score(\vec{d}, c_i)$ 为测试文档 d 属于 c_i 类的分值。

对于某一特定类来说, b_i 是一个有待优化选择的值。一般, b_i 可以通过一个验证文档集来进行调整。验证文档集是训练文档集的一部分。根据式(4)的结果,可以确定测试文档的类别。很显然,对于每一个测试文档,必须求解它和训练文档库中所有文档的相似度。因此, kNN 方法的时间复杂度为 $O(|D|\, n_t)$ ($|D|$ 和 n_t 分别为训练文档总数和测试文档总数)。

五、系统测试

5.1 测试文档库

目前国内还没有普遍接受的、标准的中文文档分类测试库,我们使用自己建立的测试文档库测试我们的文档分类系统。表 1 列出了这些测试文档类的类名及其包含的文档数目。

进行实验时,取 70 %的文档为训练文档,30 %为测试文档。采用普遍接受的查全率(recall,简记为 r)、准确率(precision,简记为 p)来评价文档分类系统的性能。

表 1　测试文档库

类名	政治	体育	经济	农业	环境	航天	艺术	教育	医药	交通
文档数目	617	350	226	86	102	119	150	120	104	116
总共	1990									

5.2　**测试结果**

从两个方面考察基于 N-gram 信息的文档分类性能:

1) 取不同 N-gram 信息项和不同数量的信息项时分类器的性能;

2) 和用不同选词方法时分类器的性能。

图 1 所示为选择不同的 N-gram 项和取不同的 N-gram 项总数时分类器的查全率和查准率。这里,1-gram 表示只取 1-gram 项,2-gram 同此;1/ 2-gram 表示既取 1-gram 项又取 2-gram 项,1/ 2/ 3-gram 和 1/ 2/ 3/ 4-gram 与此类似。从图 1 中可以得到如下结论:

® 使用 1-grams 也能取得不错的分类效果,但随着 1-grams 数量的增加,经过一个性能高峰值后,很快会衰落下去。因为,文档中不同的 1-grams 数量是有限,而和类概念相关的并不是很多,这时过于增加数量,相当于引入噪声,反而不利于分类;

® 当属性数目选择 2000 以内时,采用 1 + 2-grams 和 1 + 2 + 3 + 4-grams 比较好,而且后者优于前者。受 1-grams 影响,1 + 2-grams 和 1 + 2 + 3 + 4-grams 都表现出先扬后抑的态势;

® 2-grams 表现出一直上升的趋势,这是因为不同的且和文档类别语义相关的 2-grams 数量特别庞大。所以,在一定的数量范围内,增加 2-grams 数量总会有好处;

® 在相同的属性数量下,2-grams 比 2 + 3-grams 性能优越。原因在于 3-grams 可以用 2-grams 来表示,因此一定数量的 2-grams 比相同数量的 2 + 3-grams 包含的信息量要大。

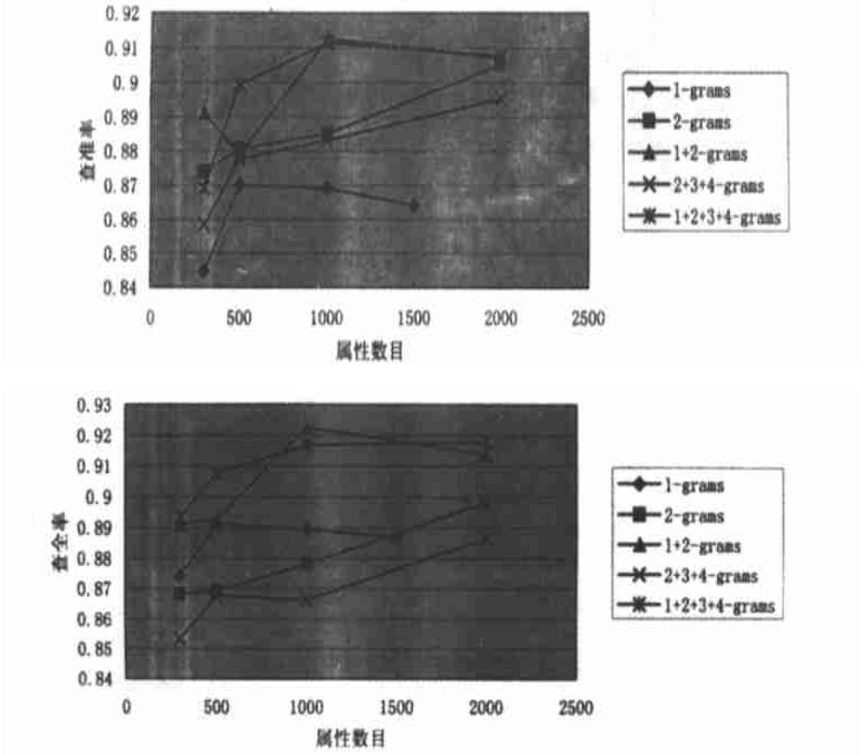

图 1　N-gram 信息项及其数量对分类性能的影响

图 2 所示为采用不同选词方法时分类器的性能测试结果。用 χ^2 统计和信息增量法进行文档属性选择得到的结果明显优于互信息法。特别是当属性数目较小时,用互信息法选词得到的分类结果比较差。χ^2 统计方法相对较优。

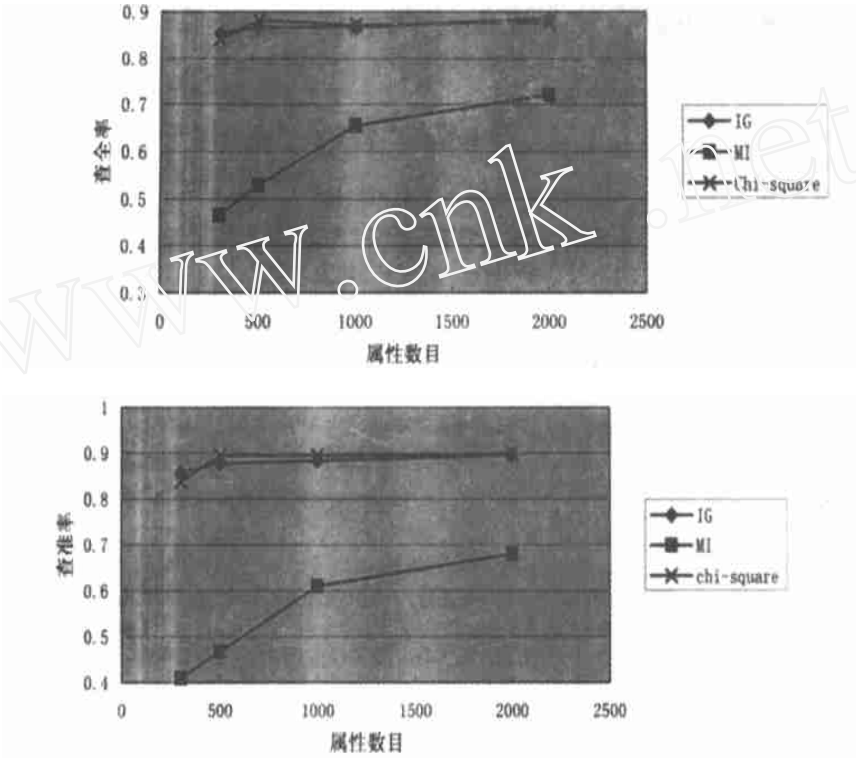

图 2　采用不同选词方法时的分类器性能

由于测试训练库的规模有限,所获得的结果还不能全面反映基于 N-gram 信息的文档分类方法的特性。下一步,我们准备在更大的测试库上,采用更多的选词方法和分类模型进行实验。

六、结束语

本文研究了采用 N-gram 信息进行中文文档分类,目的是使中文文档分类器摆脱对词典和切词处理的依赖,实现文本分类的领域无关性和时间无关性。实验结果表明这种方法是可行的。为了进一步改进本文的文档分类系统,今后,一方面,可以采用更有效的文档属性选择方法;另一方面,使用分类性能更好的分类方法。

参 考 文 献

[1] 黄萱菁,吴立德. 基于向量空间模型的文档分类系统. 模式识别与人工智能,1998,11(2)

[2] 邹淘等. 中文文档自动分类系统的设计与实现. 中文信息学报,1999,13(3):26 - 32

[3] 战学刚等. 中文文献的层次分类方法. 中文信息学报,1999,13(6):20 - 25

[4] 刘源,谭强,沈旭昆. 信息处理用现代汉语分词规范及自动分词方法. 北京:清华大学出版社,1994

[5] 赵珀璋,徐力. 计算机中文信息处理(下册). 北京:宇航出版社,1987

[6] Yang Y, Pederson J. Feature selection in statistical learning of text categorization. In: ICML-97, 1997, 412 - 420