一种无需语句分割的中文文档自动分类方法研究及实现

何 涛 黄国兴

(华东师范大学计算机科学与技术系 上海200062)

摘 要 目前,对于外文文档的自动分类,已有许多有效的方法。但是,中文的特殊性使得这些方法对于中文文档不是很有效。本文提出了一种比较简单的中文文档的自动分类方法,即不用语句分割,只需要计算出文档中各个单字出现的频率,对照已经训练过的模版,就可以比较准确地对其分类。

关键词 中文文档分类,Apriori性质,语句分割,N-Gram,词素解析

An Approach for Chinese Document Classification with No Splitting

HE Tao HUANG Guo-Xing

(Department of Computer Science, East China Normal University, Shanghai 200062)

Abstract Some effective approachs for English Document Classification have been researched and implemented. But these approachs do not fit in with Chinese Document because Chinese language is special from English language. The paper puts forward a simple Document Classification needn't to split the sentence. Only to sum up the frequency of each word, then contrast to the pattern, we can get the class of document.

Keywords Chinese document classification, Apriori property, Split sentence, N-Gram, Parse word

1 引言

随着Internet的普及以及Google等类型的搜索引擎的普遍使用,人们能够方便地从万维网中寻找到自己需要的文章。但面对着搜索出来的一大堆文章,如何过滤掉那些自己不需要的类别?如果人工分类,工作量非常大,因此如何利用计算机进行快速的自动分类,就越来越引起重视。

2 目前的文档分类方法

文档的分类,其实就是一种映射过程:$f:D\rightarrow C$。其中,D是待分类的文档集,C是某个类别。

目前,各种文档分类方法主要是基于词信息。通过对文档中的特征词的提取,推断出该文档属于哪个类别。向量空间模型(Vector Space Model)是应用比较广泛且效果也比较好的一种模型,它把文档看作是一个词的序列,每个词都有一个对应的权值,这样,文档就被映射成为一个向量 $K=\{w_1,w_2,\cdots,w_n,\cdots\}$。词的权值计算用得比较多的是如下形式的TF-IDF公式[7]:

$$W(t,d)=\frac{tf(t,d)\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in d}[tf(t,d)\times\log(N/n_t+0.01)]^2}}$$

其中,$tf(t,d)$表示词t在文档d中出现的频率,N表示训练文档的总数,n_t表示在训练集中出现词t的文档数,$W(t,d)$表示词t在文档d中的权重。由TF-IDF公式,一批文档中某词出现的频率越高,它的区分度则越小,权值则越低;而在一个文档中,某词出现的频率越高,区分度则越大,权重也越大。

文档分类的一般方法是:首先由一组预先分好类别的文档组成训练集,提取出具有能表征此文档类别的一组特征词,由此得出该类别的分类模式。然后,利用该分类模式,就可以对其他未知类别的文档进行分类。设 $K=\{w_1,w_2,\cdots,w_i,\cdots\}$ 是文档d的特征词集合,它是一个向量,其中,w_i是第i个特征词的权重,$f:K\rightarrow C$。那么,对于一篇类别未知的文档,如果它的特征词集合K'与K的距离最短,就可以把它划分到类别C去。最常见的距离度量方法是欧几里得距离:

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2}$$

另外,还可以用曼哈坦距离或明考斯基距离。

可见,文档分类的关键,在于如何建立从文档的特征词集合到分类的映射。现在用得比较多的分类方法主要有:NB(Naive Bayes)[5]、KNN(K-Nearest Neighbor)[2]等,一般都是基于统计学或者机器学习方法。而怎样从整句中分割出特征词(语句分割),对于文档自动分类是非常重要的。

3 语句分割方法

目前,主要的语句分割方法有两种:N-Gram[8]和词素解析[8]。

3.1 词素解析

所谓词素解析,是指对文字序列按字典意义上的最小单位进行分解处理。按字典意义,是指分解出来的词是字典中存在的有文字意义的词。用词素解析对文档进行语句分割后,就可以利用上文所提到的一些方法对文档进行自动分类。

使用词素解析的好处在于:可以排除一些无意义的词序列的干扰。有些词序列可能有部分文字是一致的,但是这部分文字可能并没有任何的文字意义。比如,"词素解析的好处"与"这是一种分析的好方法"两个词序列中,都存在"析的好"这个子序列,但这个字序列没有任何的字典意义,使用词素解析法就可以把它排除掉。

但是,使用词素解析法也有其缺点。字典中没有的词,就会认为是没有意义的词序列而加以排除。一方面,很难把现有的大量的词条收录完全;另一方面,随着社会的发展,各种新

何 涛 硕士研究生,研究方向:数据挖掘;黄国兴 教授,研究方向:智能信息系统,数据挖掘。

词的也在不断产生。所以，如果字典的更新速度更不上，那么这种检索遗漏的可能性还是很大的。

3.2 N-Gram 方法

与词素解析相对应，N-Gram 则不考虑文字序列的实际意义，只以长度为 N 个字的固定单位进行语句分割，分割出来的每一个字序列称为 N-Gram 信息项。比如，对于“中文文档自动分类方法”这句话，取 $N=2$，可以分解为{中文、文文、文档、档自、自动、动分、分类、类方、方法}，当然 N 越大，N-Gram 信息项就越多。

和词素解析相反，N-Gram 不会出现检索遗漏，但是会出现干扰项。像上文的语句分割中，“文文”、“档自”等明显是无意义的序列，属于干扰项。

这两种语句分割方法各有优缺点，精度也不相上下。2002年12月，在日本国立信息学研究所主办的第3届搜索引擎评价型国际会议 NTCIR(NII Test Collection for Information Retrieval Systems)上，认为词素解析法优于 N-Gram 方法[8]。

4 中文文档的特殊性

中文、英文文档有着许多不同的特性。在英文中，一个单词往往能够表示一个明确的意义。比如 computer，表示计算机；而在中文中，计算机需要用“计、算、机”3个相连的字，才能表示确定的意义，这样的一组字，称为词条。这种情况在中文中非常普遍，根据一些资料统计，中文中2字以上的词条占了绝大多数，当然，N-Gram 可以解决这个问题。中文的另外一个特点就是语法的模糊性，这从目前还没有一个好的全文翻译软件就可以看出来。英文的语法是非常明确的，一句话往往可以确切地表示出时态、意愿；而中文的语法则非常灵活，句式也多变，这就给词素解析带来了很大的困难。因此，在中文文档的自动分类模型中，如果使用目前常用的方法，那么如何有效地进行语句分割，显得尤其重要。

5 Apriori 性质

5.1 涉及的几个概念

对于一个关联模式，它的支持度[1]定义为使该模式为真的任务相关的元组(事务)所占的百分比。比如，对于关联规则 $A\Rightarrow B$，它的支持度定义为：

$$\sup(A\Rightarrow B)=\frac{\text{包含 } A \text{ 和 } B \text{ 的元组数}}{\text{元组总数}}$$

其中，A 和 B 是项目的结合。上式用概率统计的方法表示，就是

$$\sup(A\Rightarrow B)=P(A\cup B)$$

项的集合称为项集，项集的出项频率简称为频率、支持计数或计数。如果项集的出现频率大于或等于最小支持度(min_sup)与事务总数的乘积，称项集满足最小支持度，这样的项集就称为频繁项集[1]。

5.2 Apriori 性质

Apriori 性质[1]，是指频繁项集的所有非空子集都必须也是频繁的。也就是说，如果一个项集 A 不是频繁的，那么所有的它的超集也都不是频繁的。

把 Apriori 性质用于中文文档的分类，就可以得出一个结论：一个词条在文档中有相当的权重，即，这个词条是频繁的，那么，组成它的各个单字也必定是频繁的。例如，假设“计算机”是计算机分类的特征词，也就是它满足最小支持度，则“计”、“算”、“机”也必定满足最小支持度。虽然像“计程车”这样的词条中不仅有“计”字，而且“程”字也可能是计算机类文档的频繁项集，因为在计算机文档中，“程序”词条出现的频率也是比较高的，但是，“车”字在计算机的相关文档中未必是频繁项集。设 $K1=\{a1,a2,\cdots,an\}$ 是文档 A 的映射向量中权重最大的前 n 个属性，$K2=\{b1,b2,\cdots,bn\}$ 是文档 B 的映射向量中权重最大的前 n 个属性，$K3=\{c1,c2,\cdots,cn\}$ 是文档 C 的映射向量中权重最大的前 n 个属性，已知 A 属于类别 Class1，$K2$中有 i 个维与 $K1$相同，$K3$中有 j 个维与 $K1$相同，其中 $i<j$，则可以计算出 $K2$、$K3$分别与 $K1$的相似度，具有代表性的是余弦计算法[1]：

$$sim(k_1,k_2)=\frac{k_1\cdot k_2}{|k_1||k_2|}$$

通过计算相似度，可以得出文档 C 比 B 更可能属于列别 Class1。所以，问题的关键在于选取什么样的最小支持度和向量的维数。由此，我们的想法是，能不能利用这个性质，不需要语句分割，就能对中文文档进行自动分类。

6 系统模型

图1所示结构图是我们测试系统的基本构架。由于使用的是单字统计，因此，系统测试使用的是 NB(Naive Bayes)方法，这是由于这种方法的类独立条件假设[1]比较符合这个系统的环境(只考虑单字而不考虑其间的相关性)。整个系统模型分为4层：输入层、特征字提取层、模式识别层和数据库层。

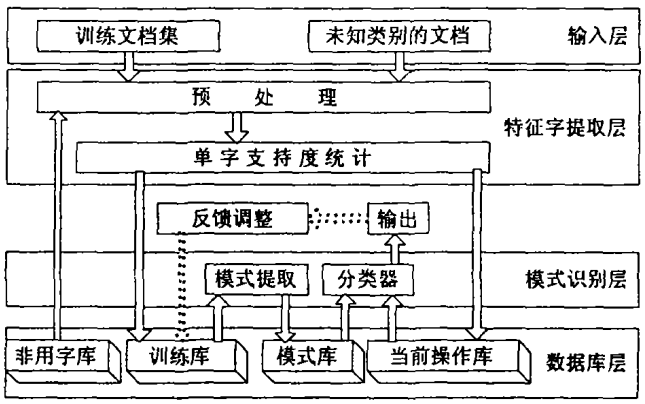

图1 测试系统结构图

输入层提供与用户的接口，有两种类型的输入，已知类型的训练文档集和未知类型的待分类文档。

特征字提取层主要统计出单字的支持度。首先，应该对文档进行预处理，目的是过滤掉那些没有意义的单字，如“的”、“啊”等，这些称为非用词，存放在非用词库中。去除非用词后，就可以对文档中的各单字出现的频率进行统计并计算出它们各自的支持度。

模式识别层则是识别出存在的模式以及利用现有的模式对未知文档进行分类。

数据库层存放整个系统的中间数据及模式信息。其中，训练库存放着训练文档集的中间结果，当前操作库则存放对未知类型文档的单字支持度统计等中间结果。这两个库的结构大体上是相同的，之所以不合二为一，是因为训练量越大，则训练的结果可信度更大，每当系统正确地对一篇未知类型的文档分类后，都要把该文档的统计结果存入训练库，并对训练库进行相应的更新，同时重新对相应的模式进行提取，更新模式库的信息。

(下转第158页)

Bi9/7。因此，多小波的理论及应用研究是有一定价值的。

参考文献

1 Daubechies I. Ten lectures on wavelets. Philadelphia: SIAM Publ, 1992
2 Goodman T N T, Lee S L. Wavelets in wandering subspaces. Trans. Amer. Math. Soc, 1993, 338(1): 639～654
3 Goodman T N T, Lee S L. Wavelets of multiplicity r. Trans. Amer. Math. Soc, 1994, 342(1): 307～324
4 Geronimo J S, Hardin D P, Massopust P R. Fractal functions and wavelet expansions based on several scaling functions. J. Approx. Theory, 1994
5 Chui C K, Lian J. A study on orthonormal multiwavelets. Applied Numerical Mathematics, 1996, 20(3): 273～298
6 Shen L, Tan H H, Tham J Y. Symmetric-antisymmetric orthonormal multiwavelets and related scalar wavelets. Applied and Computational Harmonic Analysis, 2000(8): 258～279
7 Tan H H, Shen L, Tham J Y. New biorthogonal multiwavelets for image compression. Signal Processing, 1999(79): 45～65
8 Jiang Q T. On the design of multifilter banks and orthonormal multiwavelet bases. IEEE Trans. Signal processing, 1998, 46(12): 3292～3303
9 Strela V, Heller P N, Strang G. The application of multiwavelet filterbanks to image processing. IEEE Trans. Signal Processing, 1999, 8(4): 548～563
10 Cotronei M, Lazzaro D. Image compression through embedded multiwavelet transform coding. IEEE Trans. Image Processing, 2000, 9(2): 184～189
11 Chen J Z, Zhou J L. A kind of symmetric orghogonal multiwavelets and the associated prefilter. Chinese Journal of computers, 2003, 26(6): 701～707
12 Strang G, Strela V. Short wavelet and matrix dilation equations. IEEE Trans. Signal Processing, 1995, 43(1)108～115
13 Xia X G, Geronimo J S, Hardin D P, Suter B W. Design of prefilters for discrete multiwavelet transforms. IEEE Trans. Signal Processing, 1996, 44: 25～35
14 Xia X G. A new prefilter design for discrete multiwavelet transforms. IEEE Trans. Signal Processing, 1998, 46: 1558～1570
15 Hardin D P, Roach D W. Multiwavelets prefilters I: Orthogonal prefilters preserving approximation order p≤2. IEEE Trans. Circuits Syst. II, 1998, 45: 1106～1112
16 Attakitmongcol K, Hardin D P, Wilkes D M. Multiwavelet Prefilters-Part II: Optimal Orthogonal Prefilters. IEEE Trans. Image Processing, 2001, 10: 1476～1487
17 Lebrun J, Vetterli M. Balanced multiwavelets theory and design. IEEE Trans. Signal Processing, 1998, 46: 1119～1125

（上接第138页）

在本系统中，为对文档进行训练，从 Yahoo!、网易、搜狐、中国期刊网等网站下载了分属于交通、教育、体育、军事4个类别共两百多篇已经明确了分类的文档作为测试。每个属性（即每个单字）都用一个4元组 $A=\{W,A,C,S\}$ 表示。其中，W 表示该字，A 是文档所属类别，C 是计数，S 是支持度。训练时，不是每个类别分别训练，而是把所有的训练文档全部输入，这样，就可以过滤掉那些在所有文档中出现频率都很高的字，然后，根据 TF-IDF 公式，计算出每个字的权重，并计算出支持度。

有了训练库，就可以从中提取出各分类的对应模式。模式可以根据权重或者支持度进行提取。权重或支持度的阈值既能由用户确定，也可由系统自动调整，也就是，如果系统分类错误的概率大于某值，则认为阈值太小，系统自动减小阈值。但是，阈值也不能太小，否则会导致干扰项太多。确定了阈值，就可以得到各个分类的模式，存放到模式库中。

系统对未知类型文档的处理，在特征字提取层是和训练文档一样处理的，不同的是，每个属性用三元组 $\{W,C,S\}$ 表示，中间结果存放在当前操作库中。与对训练库的处理一样，也要对当前操作库中的各个属性计算出权重和支持度。然后，分类器就可以对照模式库，输出该文档的可能分类。分类器采用 NB(Naive Bayes)分类方法。设 X 是待分类的文档，对于每个类别 C_i，计算出 $P_i=P(X|C_i)P(C_i)$，令 P_i 最大的类别就是 X 的类别。对分类结果的准确性度量，可以用查准率(precision)[9]和查全率(recall)[9]来衡量。

$precision=\frac{\Sigma correct_i}{\Sigma all_j}$，($0\leqslant i\leqslant j$，其中，$i,j$ 表示某个文档类别)

$recall=\frac{|correct_j|}{|c_j|}$，($0\leqslant j\leqslant m$，其中，$m$ 是文档类别数)

其中，$correct_i$ 表示分类 i 中分类正确的文档数，$\Sigma correct_i$ 则是全部正确分类的文档数，all_j 是分类 i 的全部文档数，Σall_j 则是参加分类的全部文档数；$|correct_j|$ 是分类 j 中正确分类的文档数，$|c_j|$ 是属于分类 j 的全部文档数。

为了进行准确度测试，又从网上另外下载了43篇文档作为待分类文档输入到模型中，得到的结论是：分类结果的准确度与阈值关系密切，阈值太大或太小都会影响分类的精度。目前，最高的查准率和查全率都是在70%到80%之间，与一些使用 N-Gram(N 大于2)的分类方法[4,5]相比(85%左右)，略微低了一些。究其原因，一则，在中文文档中，大部分词条都是由2个以上的词组成，如果采用了 N-Gram 方法，那些对文档区分度有很大权重的词条都可以被检索到，并且，通过设定阈值，可以过滤掉大部分干扰项，另一方面，这个模型还是有许多改进的余地。和 N-Gram 相比，它的速度更快，比较适用于需要即时分类的场合，如作为搜索引擎的搜索结果的快速分类器。

总结 本文提出了使用单字建立分类模型的一些初步构想和一个简单模型的实现。这种模型的优点在于：相对于词素解析，它不需要词典的支持；相对于 N-Gram，它的速度比较快。虽然准确度比其它的分类方法略低，但70%以上的分类准确度还是令我们振奋的，下一步的工作，一是要研究阈值和准确度的关系，希望可以得到一个比较准确的阈值。另外，由于没有找到比较权威的非用词库，现有系统的非用词数据很不全面，需要进一步调整。不同类别的文档，它的非用词可能不尽相同，因此，对非用词进行分类，也是提高精度的手段。在上面的模型图中，虽然画出了反馈调整的模块，但目前还没完全实现，下一步需要完善它。另外，对于 Web 文档，是否可以利用它的超链接给我们提供的信息作为分类的参考？如何做？还有对于孤立点如何处理，这些都是我们下一步的研究目标。相信这种模型在经过进一步的改进后，精度还可以进一步得到提升。

参考文献

1 Han Jiawei, Kamber M. Data Mining Concepts and Techniques. 2001
2 Duda R O, Hart P E, Stork D G. Pattern Classification (Second Edition). 2003
3 Wang K, Zhou S, Liew S C. Building hierarchical classifier using class proximity. In: Proc. 1999 Int. Conf. Very Large Data Bases (VLDB'99), Edinburgh, UK, Sept. 1999. 363～374
4 周水庚，等. 基于 N-Gram 信息的中文文档分类研究. 中文信息学报，15(1): 34～39
5 周水庚，等. 基于相邻字对信息的中文文档分类研究. 小型微型机算机系统，2001，22(4)
6 孙宏林，等. 浅层句法分析概述. http://ling.ccnu.edu.cn/message/yyxlwx/collection-4/25qianceng.htm
7 常毅，等. 基于关键词表达式模型的文本自动分类系统的研究与实现. http://www.software.ict.ac.cn/seminar/members/tjl/keyExpr
8 NII Test Collection for Information Retrieval Systems. http://research.nii.ac.jp/~ntcadm/index-en.html
9 王汉萍，等. 文本自动分类在搜索引擎上的应用. http://www.ahcit.com/200305/27.doc