Journal of Computer Applications
计算机应用,2012,32(2):392-394,415
ISSN 1001-9081
CODEN JYIIDU
2012-02-01
http://www.joca.cn

文章编号:1001-9081(2012)02-0392-03
doi:10.3724/SP.J.1087.2012.00392

基于规则与统计的 Web 突发事件新闻多层次分类

夏华林*,张仰森
(北京信息科技大学 计算机学院,北京 100192)
(* 通信作者电子邮箱 xiahualin888666@163.com)

摘　要:为了适应 Web 新闻以指数趋势增长,传播迅速,且 Web 突发事件新闻在互联网上散布等特点,同时针对传统文本分类方法准确率和效率低,寻找特定主题的突发事件新闻信息难等问题,提出一种基于规则与统计相结合的 Web 突发事件新闻多层次自动分类方法。首先提取类别关键词形成规则库,然后利用分类规则将突发事件分成四大类,再用朴素贝叶斯分类方法将各大类突发事件新闻进行细分,从而形成了基于规则与统计的两层分类模型。实验结果表明,该分类方法的准确率和召回率都达到 90% 以上,分类效率也普遍高于传统的分类方法。

关键词:规则;统计;突发事件新闻;多层次分类
中图分类号:TP181　　**文献标志码**:A

Multiple-layer classification of Web emergency news based on rules and statistics

XIA Hua-lin*, ZHANG Yang-sen
(*College of Computer, Beijing Information Science and Technology University, Beijing* 100192, *China*)

Abstract: The Web news grows in index tendency and disseminates rapidly, and the Web emergency news widely spreads on the Internet. While the traditional text classification is of low accuracy and efficiency, it is difficult to locate the emergency news and information of specific topics. The paper proposed a multiple-layer classification method for Web emergency news based on the rules and statistics. First, it extracted category keywords to form the library of rules. Second, the emergencies would be classified into four major categories by the rules, and then these major categories would be classified into small categories by the Bayesian classification method, thus a two-tier classification model based on rules and statistics was established. The experimental results show that the classification accuracy rate and the recall rate have reached over 90%, and the classification efficiency is generally higher than the traditional classification methods.

Key words: rule; statistics; emergency news; multiple-layer classification

0　引言

随着计算机信息技术的快速发展,网络已经成为最重要的新闻媒体之一。在互联网的众多新闻当中,那些难以准确预测而突然爆发的,对国家和社会产生重大影响的突发事件新闻是人们普遍关心的焦点。Web 突发事件新闻文本自动分类[1]的研究对突发事件新闻信息抽取[2]、新闻信息检索[3]、个性化服务推荐[4-5]、国家关于突发事件信息建设以及突发事件预警管理等领域都具有一定的理论意义和应用价值。

目前,Web 文本自动分类方法[6]大致可以分为两类:基于规则的方法[7]和基于统计的方法。基于规则的方法是指由专家为每个类别定义一些规则,这些规则代表了这个类别的特征,自动把符合规则的文档划分到相应的类别中,该分类方法不用提供训练样本。基于统计的方法主要是在训练统计的基础上进行学习,形成分类模型,从而进行分类测试。该方法减少了大量的人工参与,能够达到与知识工程方法相似的精确度。本文在充分考虑了 Web 突发事件新闻表达特征的基础上,构造了两级分类器。第一级是基于规则的分类器,首先明确 Web 突发事件新闻分类体系,然后通过制定规则构造第一级分类器;第二级主要是利用统计学习的方法构造分类器。因此,这两级分类就构成了 Web 突发事件文本两层分类模型。

本文介绍了几种统计的自动分类方法,分析了各自的优势与不足,同时还阐述了基于规则的自动分类方法的优缺点,并在此基础上给出了本文的基于规则与统计相结合的多层次自动文本分类方法。

1　基于统计的自动分类方法分析与选择

目前,国内主要的文本分类方法有:K 最近邻(K-Nearest Neighbor,KNN)、朴素贝叶斯、支持向量机、Rocchio 和神经网络。

KNN 分类模型[8]方法的基本思想是:计算待分类文档与系统训练集中与其最近的 K 个邻近文档之间的距离,通过这些 K 个邻近文档所属的类别来进行候选类别评分。KNN 分类方法的主要优点是分类简单,不需要预先进行学习;缺点是分类速度受到训练文档的数量很大影响,因为对于每一个待分类的文档,都需要与训练集合中的每一个训练文档进行相似度计算,这样就影响了分类的效率。

Rocchio 分类器是情报检索领域经典的算法,该分类模型的基本思想是,首先为每一个训练文本 C 建立一个特征向

收稿日期:2011-07-28;修回日期:2011-09-23。　基金项目:国家自然科学基金资助项目(60873013,61070119);北京大学计算语言学教育部重点实验室开放课题基金资助项目(KLCL-1005);北京市属市管高等学校人才强教计划项目(PHR201007131)。

作者简介:夏华林(1984-),男,湖北黄冈人,硕士研究生,主要研究方向:中文信息处理;　张仰森(1962-),男,山西临猗人,教授,博士,主要研究方向:中文信息处理。

量[9-10]，然后使用训练文本的特征向量为每个类建立一个原型向量(类向量) 。当给定一个待分类文本时，计算待分类文本与各类别的原型向量之间的距离，其距离可以是向量点积、向量之间夹角的余弦值或者其他相似度计算函数，根据计算出来的距离值决定待分类文本属于哪一类别。该分类方法的最大特点是计算简单、操作易行，缺点是分类准确率较低。

对于 Web 突发事件新闻文本而言，本文选择朴素贝叶斯方法作为统计分类方法主要原因如下:

1) 突发事件种类较多，利用 *KNN* 分类效率很低;

2) 没有足够质量较高的突发事件语料进行训练，难以构建高质量的原型向量，利用 Rocchio 分类方法准确率较低;

3) 朴素贝叶斯方法实现简单，并且通过实验证明朴素贝叶斯方法的准确率和效率较高。

朴素贝叶斯是一种基于概率的学习算法，该模型的基本思路是: 首先根据基于假设的先验概率，通过给定条件下得到的不同数据的概率以及观察到数据的本身，计算得到显式的假设概率。

朴素贝叶斯模型是文本分类模型中相对比较简单但分类效果较好的分类模型，在概率论中根据全概率公式:

$$P(C,X) = P(C \mid X)P(X) = P(X \mid C)P(C) \tag{1}$$

通过变换即可得到贝叶斯公式如下:

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \tag{2}$$

其中: $P(X \mid C)$ 表示条件概率，$P(C)$ 表示先验概率。从以上公式可以看出，后验概率 $P(C \mid X)$ 取决于 $P(X \mid C)$ 与 $P(C)$ 的乘积，这就是贝叶斯分类算法的核心思想。因此，在分类中来说，主要任务是考虑候选假设类别集合 C，根据其给定的训练数据而找到文本 X 属于的可能性最大的那个类别 C。换句话说，就是根据已经分类好的训练样本数据，如何利用训练样本数据去学习，当遇到新的文本时，可以将新文本分类到正确的类别当中，这是朴素贝叶斯分类模型的基本思想。

2　基于规则的分类方法和多层分类模型

2.1　基于规则的分类方法及规则制定的途径

通过以上三种常用分类方法的介绍与分析可以看出，基于统计的自动文本分类主要是经过大量语料的训练来得到各类别的模板，从而利用模板进行分类。其优点是训练过程简单方便，通常情况下分类精度较高。其主要缺点是:

1) 在各类别之间交叉现象比较严重的情况下，分类器的精度会大大降低，尤其是在多层分类中，部分子类之间的特征交叉更为严重;

2) 基于统计的分类方法对训练语料的数量和质量均有较高要求，如果语料不全面，代表性不强，则会直接影响自动分类的精度。

基于规则的自动文本分类[11]的主要思想: 用户直接为每个类别制定分类规则形成类别模板，规则分类器依据类别模板统计测试样本中满足的规则条数及规则出现的次数信息，同时利用规则在测试文本结构中的位置信息，来衡量测试样本所属的类别。

当一些类别没有充足的训练语料或者语料的质量不高，统计的分类方法效果不好时，可以采用基于规则的分类方法来代替基于统计的分类方法; 对于类别特征词比较明显，特征词表比较小，规则的制定比较容易时，也可以采用基于规则的方法分类。

类别模板的形成过程: 每个类别模板可以由很多条规则组成，在类别模板中每一行信息代表一条规则，每条规则可以由多个项组成，每一项都是该类的关键词。规则支持与、或、非和异或等逻辑运算。

制定规则的途径主要有以下方面:

1) 通过相关领域的专家来制定规则，这样能够保证规则的正确性和科学性;

2) 首先利用统计模型算法来提取各类别关键词，然后通过人工进行筛选之后制定规则;

3) 利用各种分类主题词典来获得类别主题词信息从而制定规则。

基于规则的自动分类与基于统计的自动分类的主要区别是基于规则的分类不需要提供训练语料。基于规则的分类的优势是分类精度高，对规则模板可以随时修改，灵活方便; 缺点是规则要全面，要有代表性，否则就会直接影响分类器的性能，当类别规模增大时，需要确定的规则数量增多，而且规则的维护比较麻烦。

2.2　基于规则与统计相结合的突发事件分类模型

经过对大量 Web 突发事件新闻信息进行分析和实验探索，充分考虑了 Web 突发事件新闻文本的标题和内容的用词特点，提出了一个基于规则和统计的两层 Web 突发事件新闻文本分类模型，以下是模型建立的步骤。

第一层设计了一个基于规则的分类器。基于规则的分类步骤如下:

步骤 1　利用抽取系统及人工参与的方法从新闻门户网站上抽取质量较高的突发事件新闻文本信息[12-14]，并且进行人工分类。

步骤 2　对抽取出来的新闻标题、新闻正文内容进行分词，词频统计，类别关键词提取，形成类别特征词库，再结合 Web 突发事件新闻信息的类别特点形成类别规则。

步骤 3　每个类别规则由多个项组成，每一项都是该类的关键词，将这些项用逻辑运算符 AND、OR、NOT 连接起来形成规则。

步骤 4　判断逻辑表达式的真假即可知规则是否成立，从而将待测文本归入相应的类别之中。

例如，对于公共卫生事件，新闻标题或者内容中满足规则: 名称(病毒传播 OR 食物中毒 OR 艾滋病 OR AIDS OR 猪流感 OR 甲型 H1N1 OR 甲流 OR 禽流感 OR SARS …) AND 症状(呕吐 OR 头痛 OR 发烧 …) AND 机构(卫生部 OR 医院 OR 门诊 …) 。因此，自然灾害、事故灾难、社会安全事件这三类 Web 突发事件新闻也按照这样的方法制定相应的规则。

第二层设计了基于统计的分类器。本文采用的统计方法是朴素贝叶斯分类法，其算法步骤如下:

步骤 1　将待分类文本分词、去除停用词及无用词性的词等预处理后，余下的词条作为其特征项，将文本表示成特征向量 $\boldsymbol{X}(X_1, X_2, \cdots, X_n)$ 。

步骤 2　循环遍历待分类文本向量以及类别特征词库，需要计算待分类文本中的每个词 X_i 在 C_j 类别中出现的次数，因此只需用 select 语句从数据库中查找每个词，如存在则读

出相对应的类别文档频率(每个词 X_i 在 C_j 类别中出现的次数),循环得到每一个词的对应该类别 C_j 的值 $N(X = X_i, C = C_j)$。

步骤 3 将循环遍历得到的值根据式(1)(2)计算 $P(X_i \mid C_j) = \dfrac{N(X = X_i, C = C_j) + 1}{N(C = C_j) + M}$ 即可得到待分类文本对类别 C_j 的朴素贝叶斯值。

步骤 4 重复步骤 2、3,直至完成待分类文本对相应类别的贝叶斯值计算,将计算所得的值存在数组中,最后,通过比较得到最大的值即为文本所对应的类别。

突发事件可分成自然灾害、事故灾难、公共卫生事件和社会安全事件四大类。由于这四大类中的小类较多,各个小类别之间的交叉现象严重,并且找不到足够全面且质量较高的突发事件语料;此外,Web 突发事件新闻主题类别非常明显,便于规则的制定。因此,本文第一层对这四类进行分类时采用的是基于规则的方法。经过第一层的四大类分完之后,再对这每个大类中的小类之间利用统计的方法分类,这样就避免了四大类中各小类之间的交叉现象,而同一个大类中的各小类之间,可以通过加入类别关键词和去掉小类的共现词,从而提高统计分类器的精度。本文设计并实现了该两层基于规则与统计的分类系统。该系统经过一段时间的测试与优化,性能稳定,分类精度较高。

3 规则与统计相结合的突发事件多层分类系统

3.1 突发事件新闻分类体系

本系统主要是对近年来经常发生的八类突发事件进行分类,主要语料是通过各大新闻门户网站抽取获得的。根据国家 2007 年发布的《中华人民共和国突发事件应对法》,可以得到 Web 突发事件新闻分类体系[15] 如图 1 所示。

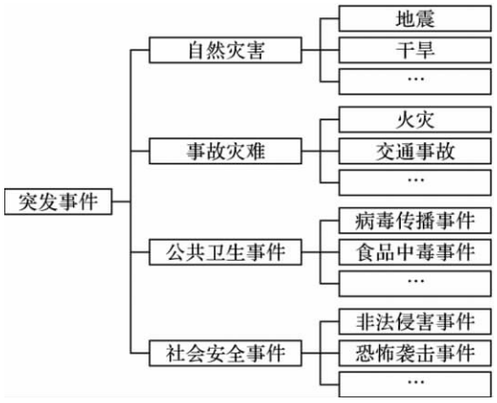

图 1 Web 突发事件新闻信息分类体系

3.2 Web 突发事件新闻多层自动分类系统设计

通过以上分类体系和分类核心算法的分析,本文设计了一个基于规则和统计的 Web 突发事件新闻文本多层次自动分类系统,其分类流程如图 2 所示。

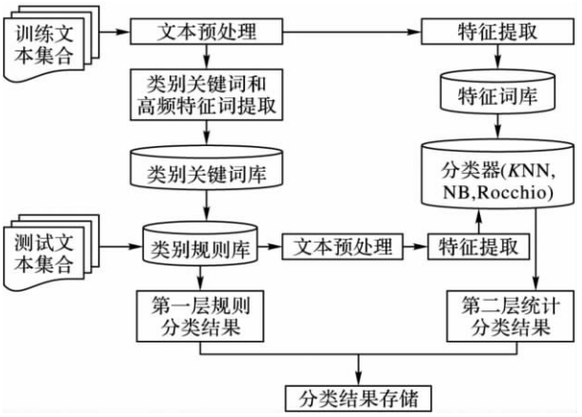

图 2 系统分类流程

4 实验结果及分析

本文所使用的 Web 突发事件新闻的训练语料和测试语料主要来源于新浪网、搜狐网、新华网和人民网四个大型新闻门户网站。首先,通过 Web 新闻文本抽取技术从以上四个门户网站抽取 Web 突发事件新闻 1 200 篇,然后利用人工将其分成地震、干旱、火灾、交通事故、病毒传播事件、食品中毒事件、非法侵害事件、恐怖袭击事件等八类。其中 800 篇用来训练,400 篇用于测试。本文使用的评价指标是常用的召回率(*Recall*)和准确率(*Precision*)以及 F_β 值。

$$Recall = \frac{\text{实际分类正确的文本数}}{\text{应该分到此类的文本数}} \times 100\% = \frac{\text{正确分类文本数量}}{\text{类别文本数量}} \times 100\%$$

$$Precision = \frac{\text{实际分类正确的文本数}}{\text{实际分到此类的文本数}} \times 100\% = \frac{\text{正确分类文本数量}}{\text{系统分类文本数量}} \times 100\%$$

$$F_\beta = \frac{(\beta^2 + 1) \times p \times r}{\beta^2 \times p + r}$$

其中 β 是调整正确率和召回率在评价函数中所占比例的参数,通常取 $\beta = 1$,这时,评价指标变为

$$F_1 = \frac{2 \times p \times r}{p + r}$$

传统的基于统计的分类结果和本文提出的基于规则与统计相结合的多层次分类结果如表 1 所示。

表 1 传统方法与本文方法的分类结果对比

类别	传统的基于统计的分类方法			基于规则与统计结合的分类方法		
	准确率/%	召回率/%	F_1/%	准确率/%	召回率/%	F_1/%
地震	93.62	88	90.72	94.34	100	97.09
干旱	97.56	80	87.91	100.00	94	96.91
火灾	93.75	90	91.84	100.00	90	94.74
交通事故	58.82	100	74.07	90.91	100	95.24
病毒传播事件	97.87	92	94.84	100.00	96	97.96
食品中毒事件	70.42	100	82.64	96.16	100	98.04
非法侵害事件	100.00	34	50.75	100.00	98	98.99
恐怖袭击事件	100.00	88	93.62	98.04	100	99.01

(下转第 415 页)

[2] CHUA L O, YANG L. Cellular neural networks: application[J]. IEEE Transactions on Circuits and Systems, 1988,35(10):1273 – 1290.

[3] HOPFIELD J J. Neural networks and physical systems with emergent collective computational abilities [M]// Neurocomputing: Foundations of Research. Cambridge: MIT Press, 1988: 457 – 464.

[4] BRUCOLI M, CARNIMEO L, GRASSI G. Discrete-time cellular neural networks for associative memories with learning and forgetting capabilities [J]. IEEE Transactions on Circuits System Ⅰ: Fundamental Theory and Application, 1997, 44(7):646 – 650.

[5] LIU DERONG, MICHEL A N. Sparsely interconnected neural networks for associative memories with applications to cellular neural networks [J]. IEEE Transactions on Circuits System Ⅱ: Analog Digit Signal Processing, 1994,41(4):295 – 307.

[6] LIU DERONG. Cloning template design of cellular neural networks for associative memories [J]. IEEE Transactions on Circuits System Ⅰ: Fundamental Theory and Application, 1997,44(7):646 – 650.

[7] LIU DERONG, LU ZANJUN. A new synthesis approach for feedback neural networks based on the perceptron training algorithm[J]. IEEE Transactions on Neural Networks, 1997,8(6):1468 – 1482.

[8] LU ZANJUN, LIU DERONG. A new synthesis procedure for a class of cellular neural networks with space-invariant cloning template [J]. IEEE Transactions on Circuits System Ⅱ: Analog Digital Signal Processing, 1998,45(12):1601 – 1605.

[9] BRUCOLI M, CARNIMEO L, GRASSI G. Heteroassociative memories via cellular neural networks [J]. International Journal of Circuit Theory and Application, 1998,26(3):231 – 241.

[10] ZENG ZHIGANG, WANG JUN. Analysis and design of associative memories based on recurrent neural networks with linear saturation activation functions and time-varying delays [J]. Neural Computation, 2007,19(8):2149 – 2182.

[11] ZENG ZHIGANG, WANG JUN. Design and analysis of high-capacity associative memories based on a class of discrete-time recurrent neural networks [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2008,38(6):1525 – 1536.

[12] ZENG ZHIGANG, WANG JUN, LIAO XIAOXIN. Stability analysis of delayed cellular neural networks described using cloning templates [J]. IEEE Transactions on Circuits and Systems Ⅰ: Regular Papers,2004, 51(11):2313 – 2324.

[13] GIROLAMO F, ANTONIO G. Adaptive particle swarm optimization for CNN associative memories design [J]. Neurocomputing, 2009, 72(16/17/18):3851 – 3862.

[14] COLORNI A, DORIGO M, MANIEZZO V. An investigation of some properties of an "Ant algorithm" [C]// PPSN 92: Proceedings of the Parallel Problem Solving from Nature Conference. Brussels: Elsevier Publishing, 1992: 509 – 520.

[15] DORIGO M, MANIEZZO V, COLORNI A. The ant system: Optimization by a colony of cooperating agents [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 1996, 26(1): 29 – 41.

[16] CHEN C-Y, YE F. Particle swarm optimization algorithm and its application to clustering analysis [C]// IEEE International Conference on Networking, Sensing and Control. Piscataway: IEEE, 2004,2: 789 – 794.

(上接第 394 页)

通过实验结果分析可以看出,对于 Web 突发事件新闻分类来说,准确率、召回率、F_1 值都较高,这是因为突发事件新闻比较特殊并且测试语料的类别特征相当明显。总体来看,本文所提出的基于规则与统计相结合的两层分类方法的准确率、召回率和 F_1 值都明显提高,由于前一级基于规则分类方法的使用,使得分类的整体性能和分类效率都远远高于传统的基于统计的方法。因此,本文提出的方法对于 Web 突发事件新闻文本自动分类具有较好的性能。

5　结语

本文分别介绍并分析了基于统计和基于规则的自动文本分类方法的优势和不足,然后在充分分析了 Web 突发事件新闻文本特点的基础上,结合规则与统计各自的优点,提出了一种基于规则和统计相结合的 Web 突发事件新闻两层自动文本分类方法。设计并实现了该分类系统,经实验证明,系统性能良好。

参考文献:

[1] 张永奎, 李红娟. 基于类别关键词的突发事件新闻文本分类方法[J]. 计算机应用, 2008,28(6):139 – 143.

[2] 谷文. 基于概念树的 Web 信息抽取技术研究[D]. 长春: 长春工业大学, 2010.

[3] 马晖男. 信息检索中浅层语义模型的研究[D]. 大连: 大连理工大学, 2007.

[4] 陈炯. Web 突发事件新闻个性化推荐方法的研究[D]. 太原: 山西大学, 2005.

[5] 高峰. 基于兴趣分类的用户行为分析系统的研究[D]. 济南: 山东大学, 2010.

[6] MARKOV A, LAST M, KANDEL A. The hybrid representation model for Web document classification[J]. International Journal of Intelligent Systems, 2008, 23(6): 654 – 679.

[7] 李渝勤,孙丽华. 基于规则的自动分类在文本分类中的应用[J]. 中文信息学报,2004,18(4):9 – 14.

[8] 沈志斌,白清源. 基于加权修正的 KNN 文本分类算法 [C]// 第二十五届中国数据库学术会议论文集. 重庆: 计算机科学, 2008: 123 – 126.

[9] 王维娜,康耀红,伍小芹. 文本分类中特征选择方法研究[J]. 信息技术,2008(12):29 – 31.

[10] 徐燕,李锦涛,王斌,等. 文本分类中特征选择的约束研究[J]. 计算机研究与发展,2008, 45(4):596 – 602.

[11] 刘红梅. 基于关联规则的分类方法初探[J]. 电脑知识与技术, 2009,5(3):535 – 536.

[12] 李蕾,王劲林,白鹤,等. 基于 FFT 的网页正文提取算法研究与实现[J]. 计算机工程与应用,2007,43(30):148 – 151

[13] de BOER V, van SOMEREN M, WIELINGA B J. A redundancy-based method for the extraction of relation instances from the Web [J]. International Journal of Human-Computer Studies, 2007, 65(9): 816 – 831.

[14] PENG XIAOGANG, MING ZHONG, WANG HAITAO. Text learning and hierarchical feature selection in Web page classification[C]// ADMA '08: Proceedings of the 4th International Conference on Advanced Data Mining and Applications. Berlin: Springer-Verlag, 2008: 452 – 459.

[15] 杨丽英,李红娟,张永奎. 突发事件新闻语料分类体系研究[C]// 中文信息处理前沿进展. 北京: 清华大学出版社,2006: 403 – 409.