

文章编号:1001-9081(2008)12-3248-03

基于概念向量空间模型的电子邮件分类

曾超, 吕钊, 顾君忠

(华东师范大学 信息科学技术学院, 上海 200241)

(czeng@ica.stc.sh.cn)

摘要: 提出了一个基于概念向量空间模型的电子邮件分类方法。在提取电子邮件特征向量时, 以 WordNet 语义本体库为基础, 以同义词集合概念代替词条, 同时考虑同义词集合间的上下位关系, 从而建立电子邮件的概念向量空间模型作为电子邮件的特征向量。使用 TF * IWF * IWF 方法对概念向量进行权值修正, 最后通过简单向量距离分类方法来确定电子邮件的类别。实验结果表明, 当训练集合数目有限时, 该方法能够有效提高电子邮件的分类准确率。

关键词: 电子邮件分类; WordNet; 概念向量; 向量空间模型

中图分类号: TP393.098 文献标志码:A

E-mail classification based on concept vector space model

ZENG Chao, LU Zhao, GU Jun-zhong

(Institute of Computer Applications, East China Normal University, Shanghai 200241, China)

Abstract: A new approach of e-mail classification based on the concept vector space model was proposed. In this approach, the eigenvector of the e-mail was extracted during training process by replacing terms with synonymy sets in WordNet and considering hypernymy-hyponymy relation between synonymy sets. Then, TF * IWF * IWF method was used to revise the weight of the concept vector. In the end, the type of e-mail was determined using the simple vector classification method. Compared with the term-based VSM approach, the results show that this approach can improve the accuracy of e-mail classification especially when the size of training set is small.

Key words: E-mail classification; WordNet; concept vector; Vector Space Model (VSM)

0 引言

电子邮件是人们在网络环境下实现信息交流的主要手段之一。在 Internet 网络已经普及的今天, 收发电子邮件几乎已成为相当一部分人正常生活的一部分。但是, 当人们在享用网络信息交流快捷的同时, 电子邮件的存在和泛滥也给人们带来较大干扰。然而电子邮件往往反映出社会当前的热点问题和公众的舆论焦点, 倘若能将电子邮件进行自动分类, 那么人们就可以准确、全面、迅速地获取到自己关心的内容, 大大提高工作效率, 从而减少了人力、财力、物力等方面的损失。由此可见对电子邮件进行研究将其分类具有重要意义和应用价值。

针对电子邮件产生手段的不断变化, 电子邮件过滤技术也随着在不断发展。现在很多反电子邮件技术方案都不会只采用一种技术, 而是多种技术的综合体。目前的反电子邮件产品所采用的技术主要还是黑名单、白名单、DNS 识别、速率控制、OCR 识别和分析、病毒扫描、全面信誉系统、基于规则的评分系统和数据挖掘等技术。除数据挖掘技术外的过滤技术属于事后防备型, 而且名单内容需要专人不断添加, 所依赖的规则也要根据电子邮件的发展状况不断改写, 在阻隔效果上受技术影响有滞后性。数据挖掘技术能够依据历史数据自动预测未来趋势和行为, 使事后防备型变为主动预防型。目前基于数据挖掘技术的电子邮件分类技术有贝叶斯分类法、人工智能、文本聚类和决策树等。

1 相关研究背景

通常, 电子邮件分类主要有如下三个环节: 电子邮件文本的预处理、特征选择和构造分类器。

预处理环节由文本分词、特征表示和特征提取三部分组成。特征表示目前按照是否进行语义理解可以分为两类: 基于关键词的表达模型——向量空间模型 (Vector Space Model, VSM) 和基于词义理解的概念表达模型。虽然 VSM 模型不考虑语义信息并且部分丢失了文本中词和词的相互关联, 但是相比后一种复杂尚未成熟的方法, 它简单易处理, 并且对文本处理 (主要是分类) 可以得到很好的效果, 因此是目前最常用的方法。

现有的文本特征选择方法通常可分为两类: 一类是过滤的方法, 另一类是清除的方法。前者是把特征的选择作为文本预处理的一步来做, 通过一系列的规则对特征进行加权, 而后取权值最大的前 K 个特征向量构成降维后的文档向量空间。比如: 文档频度、互信息和 $2\chi^2$ 统计法等。这一类方法的缺陷是: 认为各个特征维度之间是相互独立的, 忽略了有些只有和其他特征维度同时出现时对分类才有贡献的特征维度, 从而降低了分类精度。现在针对文档频度和 TF * IDF 算法有不少的改进, 如概率 TF * IDF 算法^[1] 和 TF * IWF * IWF 算法^[2] 等。清除的方法主要是将分类算法看作是一个黑箱来进行特征选择。这类方法基于特征对分类准确率的贡献大小来进行特征选择, 而且已经被验证比过滤的方法更有效。但是, 它的计算量过大, 尤其是在文本集合的特征向量太多时, 其实用性不强。

收稿日期: 2008-06-27; 修回日期: 2008-09-07。

作者简介: 曾超(1984-), 女, 江西上饶人, 硕士研究生, 主要研究方向: 信息安全; 吕钊(1970-), 女, 四川江油人, 副教授, 博士, 主要研究方向: 软硬件协同设计、信息安全、计算机支持的协同技术; 顾君忠(1949-), 男, 上海人, 教授, 博士生导师, 主要研究方向: 多媒体、CSCW;。

分类器的构造有简单向量距离分类、基于统计^[3-4]、基于连接^[5]和基于规则^[6]等方法。基于统计的方法有 Naive Bayes、KNN、支持向量机等;基于连结的有神经网络;基于规则的有决策树等,一些研究者已经验证了这些算法有效性。同时,对包括 KNN、决策树、朴素贝叶斯和神经网络等 14 种分类算法的性能比较实验表明:经过大数据量的训练集后,KNN 等算法分类准确率是令人满意的。但是,它们存在一个共同的问题:由于不考虑词之间的语义关系,往往会出现向量空间的高维性,大大降低了分类的性能;对于训练文本集合有限的情况,由于表示类别的信息量过小,层次过低,造成这些算法的分类精度大大下降。简单向量距离分类法适用于文本表示简单,向量维数较低的情况。而由于电子邮件文本基本上都是小文本,内容不会很多,因此从算法复杂程度及分类效果来看简单向量距离分类法比较适合本电子邮件分类系统。

通过分析已有技术,本文提出一种电子邮件特征提取方法,在提取电子邮件特征向量时,以 WordNet^[7]语言本体库为基础,以同义词集合概念代替词条,同时考虑同义词集合间的上下位关系,从而建立电子邮件的概念向量空间模型作为电子邮件特征向量,使得在训练过程中能够提取出代表类别的高层次信息。使用 TF * IWF * IWF 方法对概念向量进行权值修正,最后通过简单向量距离分类方法来确定电子邮件的类别。

2 电子邮件的概念向量空间生成及分类

电子邮件分类由训练阶段和分类阶段组成:训练阶段使用已标注分类的电子邮件集合训练分类器,得到各个分类的特征向量空间;分类阶段将待分类电子邮件输入,得到分类结果。如图 1 所示。

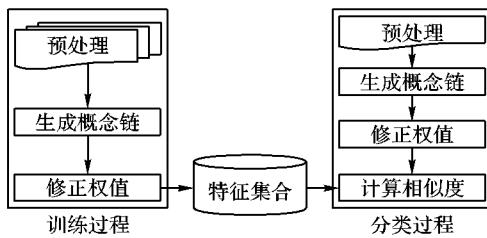


图 1 电子邮件分类流程

2.1 训练过程

训练集电子邮件的训练过程如下:

- 1) 电子邮件的预处理模块;
- 2) 表示电子邮件内容的概念链生成模块;
- 3) 概念链的权值修正模块。

2.1.1 预处理

预处理的过程如下:

输入:一封电子邮件 d_j 。

1) 将电子邮件 d_j 进行分词,除去标点、高频词。然后除去词根、词缀,将 d_j 表示为: $\{w_1 w_2 \dots w_i \dots w_k\}$ 词集的形式,计算词 w_i 在电子邮件 d_j 中出现的次数 $N(w_i)$ 。

2) 计算训练集中属于类 C_k 的电子邮件中词 w_i 出现的次数, $N(w_i; C_k) = \sum_{j=1}^{n_k} N(w_i)$, 其中 $C_k = \{d_1 d_2 \dots d_j \dots d_{n_k}\}$, 即类 C_k 中有 n_k 篇训练电子邮件。同时建立一个词的集合 $Q_L(C_k)$ 用于保存该类训练电子邮件中出现的词。

3) 计算整个训练电子邮件集合中词 w_i 出现的次数, $N(w_i) = \sum_{k=1}^n N(w_i; C_k)$, n 为类别数。同时建立该训练电子邮件

件集合的词库 Q_J 。

2.1.2 生成概念链

生成概念链的过程如下:

初始化: $Q_L(C_k) = Q_L(C_k); Q_J = Q_J$ 。

1) 判断词 w_i 是否在 WordNet 中出现:若 w_i 不在 WordNet 中出现,则 $Q_L(C_k) \leftarrow Q_L(C_k) - w_i; Q_J \leftarrow Q_J - w_i$ 。

2) 取 $Q_L(C_k)$ 中的词 w_i ,查询 WordNet,得到词 w_i 的概念链。

3) 将概念链顺次存入 VectorId 中,同时将 $N(w_i; C_k)$ 作为词 w_i ($i = 1, 2, 3, 4, \dots, m$, 其中 m 为 $Q_L(C_k)$ 中词的个数) 对应的概念链中其同义词集合和其直接上位词对应的同义词集合的权值,并存入 VectorCValue。若 VectorId 已存在则不必再存储。

4) 顺次取 $Q_L(C_k)$ 中的词 w_i ($i = 2, 3, 4, \dots, m$), 查询 WordNet, 得到词 w_i 的概念链。首先判断此概念链中其同义词集合(可能有多个)是否在 VectorId 中,若是存在,则修改 VectorCValue 中相应的同义词集合的权值 V_i ,即: $V_i \leftarrow V_i + N(w_i; C_k)$, 同时修改这些同义词集合的直接上位词集合的权值;若是不存在,则转到 3)。

5) 如果 $Q_L(C_k)$ 中的词均处理了,则构建出了类 C_k 对应的训练电子邮件集的概念向量空间。

6) 最后是将各个类对应的 VectorId 和 VectorCValue 进行合并,得到整个训练电子邮件集的 VectorId 和 VectorSValue。

2.1.3 修正权值

修正权值的过程如下:

1) 计算概念的反文本频度 $IDF(V_i) = \left[\log \left(\frac{N(V_i)}{N} \right) \right]^2$,

其中 N 为整个电子邮件集合中出现概念 V_i 的训练电子邮件数, $N(V_i)$ 为概念 V_i 在类 C_k 中出现的训练电子邮件数。

2) 计算概念的反类别频度 $F_{ICF}(V_i) =$

$$\sqrt{\frac{\sum_j (U_{ij} - \bar{V}_i)^2}{\sum_j V_{ij}}}, \text{ 其中 } U_{ij} \text{ 为概念 } V_i \text{ 在类 } C_k \text{ 中的权值, } \bar{V}_i \text{ 为概念 } V_i \text{ 在整个训练电子邮件集合中的权值除以训练电子邮件集合的类别数。}$$

3) 修正概念权值:

$$\text{VectorCValue}(V_i) = \text{VectorCValue}(V_i) * IDF(V_i) * F_{ICF}(V_i)$$

4) 将 VectorCValue 进行归一化:

$$\text{VectorCValue}(V_i) \leftarrow \frac{\text{VectorCValue}(V_i)}{\sqrt{\sum_{i=1}^n (\text{VectorCValue}(V_i))^2}}$$

2.2 分类过程

测试电子邮件的分类过程如下:

1) 对测试电子邮件进行预处理,并构建表示该电子邮件内容的概念链,得到该电子邮件的 VectorId 和 VectorCValue。

2) 利用概念链的权值修正模块,得到修正了的该电子邮件的 VectorId 和 VectorCValue。

3) 计算电子邮件的概念向量空间 $X = (x_1, x_2, \dots, x_m)$ 与各个类别的特征向量空间 $Y = (y_1, y_2, \dots, y_m)$ 之间的相似度,注意计算电子邮件与类 C_k 的相似度时只考虑在类 C_k 特征向量空间含有的概念,不考虑电子邮件空间中特有的概念:

$$F_{sim}(X, Y) = \frac{\sum_{k=1}^m (x_k \cdot y_k)}{\sqrt{\sum_{k=1}^m x_k^2 \cdot \sum_{k=1}^m y_k^2}}$$

4) 取最大的 $F_{\text{sim}}(X, Y)$ 所对应的类别定为该测试电子邮件的类别。

3 实验及结果分析

3.1 实验数据描述

本文采用了标准文档集 20_newsgroups, 20_newsgroups 中的文档放在 20 个目录下面, 每个目录就是新闻组的一个类, 每个类一般包含 1000 篇文章。抽取 20_newsgroups 文档集的某些类进行实验, 将数据集的一部分作为训练集, 另一部分作为测试集。分别运用传统的向量空间模型和本文所提出的基于 WordNet 概念向量空间模型的方法对这些类进行实验。

3.2 实验结果分析

实验选取了常用的三种评估方法作为对分类系统性能的评价, 它们分别是查准率 (Precision)、查全率 (Recall) 和 F1 值。查准率是被正确分类的电子邮件数与实际被分到该类的电子邮件数的比例, 它主要反映的是分类器准确查找电子邮件的能力。查全率是被正确分类的电子邮件数与该类应有的电子邮件数的比例, 它主要反映的是分类器查找电子邮件范围的能力。F1 值则是一种综合考虑查全率和查准率的评价方法。

本文做了两个实验: 1) 固定训练电子邮件数和测试电子邮件数时, 对运用传统的向量空间模型和本文所提出的基于 WordNet 概念向量空间模型的方法进行性能比较; 2) 测试训练电子邮件规模对运用基于 WordNet 概念向量空间模型的方法的性能影响。

实验 1 中选取了 20_newsgroups 中的三个类, 包括 alt. atheism、comp. graphics 和 rec. autos 类, 在每个类中选取 200 个电子邮件, 共 600 个电子邮件进行实验; 对于测试样本, 在每个类中选取 10 个电子邮件作为测试样本, 采用了查准率、查全率和 F1 值三种评价指标来对分类系统的性能进行评估。

表 1、2 分别列出了采用改进和传统算法所得的分类结果。从中可以看出, 基于 WordNet 概念向量空间模型算法计算所得的分类结果优于传统的向量空间模型所得结果。改进方法的查准率和查全率的平均值大于 83%, 而传统的向量空间模型这两项的平均值却未超过 74%。

表 1 传统的向量空间模型算法

Type	rec. autos	Alt. atheism	comp. graphics
Precision	0.78	0.70	0.73
Recall	0.70	0.70	0.80
F1	0.74	0.70	0.76

表 2 基于 WordNet 概念向量空间模型算法

Type	rec. autos	alt. atheism	comp. graphics
Precision	0.80	0.89	0.82
Recall	0.80	0.80	0.90
F1	0.80	0.84	0.86

实验 2 选取了 20_newsgroups 的 6 个类, 包括 alt. atheism、comp. graphics、misc. forsale、rec. autos、sci. crypt 和 sci. med。对于每个类分别选取相同数量的电子邮件进行实验。训练样本的规模分别为每个类选取 5 个样本, 共 30 个训练样本; 每个类选取 50 个样本, 共 300 个训练样本; 每个类选取 100 个样本, 共 600 个训练样本; 每个类选取 150 个样本, 共 900 个样本。对于测试样本, 则是在 6 个类中分别选取 10 个样本作为测试集。这里的评价指标计算都采用了宏平均计算方法。

从图 2 可以看到, 训练样本的规模大小对分类性能是有

影响的。随着训练样本数的增多, 分类性能也随之增强, 这是由于当训练样本数较少时, 组成电子邮件向量的特征出现次数都较少, 这样就不易选择出能较好代表电子邮件向量的特征, 从而使得分类效果不理想。随着训练样本数目增多, 则可以较容易选择出较好的特征来代表电子邮件向量。当训练样本数达到一定规模时, 如本实验中训练数达到 900 时, 分类性能趋于稳定状态, 这是由于在训练集中, 已经可以较容易地区别出特征对于类别的的重要性。

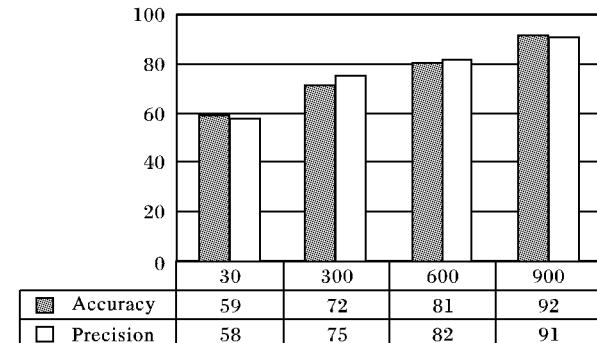


图 2 训练样本规模对分类性能的影响

4 结语

本文提出一种电子邮件特征提取方法, 在提取电子邮件特征向量时, 以 WordNet 语言本体库为基础, 以同义词集合概念代替词条, 同时考虑同义词集合间的上下位关系, 从而建立电子邮件的概念向量空间模型作为电子邮件特征向量, 使得在训练过程中能够提取出代表类别的高层次信息。使用 TF * IWF * IWF 方法对概念向量进行权值修正, 最后通过简单向量距离分类方法来确定电子邮件的类别。实验结果表明, 当训练集合数目有限时, 本文方法能够有效地提高电子邮件的分类准确率。

我们未来的研究工作是利用本文提出的算法所得到的概念向量进行电子邮件的层次分类; 同时, 对训练电子邮件集合很大时, 提高该算法的分类精度以及对特征向量降维的效果。

参考文献:

- [1] JOACHIMS T. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization [C]// ICML'97: Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997: 143 - 151.
- [2] BASILI R, MOSCHITTI A, PAZIENZA M. A text classifier based on linguistic processing [C]// Proceedings of IJCAI-99, Machine Learning for Information Filtering. 1999.
- [3] SCHNEIDER K-M. A comparison of event models for Naive Bayes anti-spam e-mail filtering [C]// Proceedings of the 10th Conference European Chapter of the Association for Computational Linguistics. New York: ACM, 2003: 307 - 314 .
- [4] NOUALI O, BLACHE P. A semantic vector space and feature-based approach for automatic information filtering [J]. Expert Systems with Application, 2004, 26(2): 171 - 179.
- [5] MAZUROWSKI M A, HABAS P A, ZURADA J M, et al. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance [J]. Neural Netw, 2008, 21(2/3): 427 - 436.
- [6]AITKENHEAD M J. A co-evolving decision tree classification method [J]. Expert Systems with Applications: An International Journal, 2008, 34(1): 18 - 25.
- [7] FELBAUM C. WordNet: An electronic lexical database [M]. Cambridge, Massachusetts: MIT Press, 1998.