第33卷第3期 华中科技大学学报(自然科学版) Vol.33 No.3
2005年 3月 J. Huazhong Univ. of Sci. & Tech.(Nature Science Edition) Mar. 2005

一种基于VSM文本分类系统的设计与实现

李 凡 林爱武 陈国社

(华中科技大学 计算机科学与技术学院,湖北 武汉 430074)

摘要:阐述了一个基于改进向量空间模型的中文文本分类系统的设计与实现,包括对该系统的结构、预处理、特征提取、训练算法,分类算法等关键技术的介绍.通过引入结构层次权重系数来改进文本特征项权重,同时提出一种新的训练算法和文本相似度域值计算方法.实验结果证明:该分类系统能有效地提高文本分类效果,开放性测试的平均准确率在80 %以上,且平均查全率达到了86 %.

关 键 词:文本分类;向量空间模型;特征提取;结构层次权重系数;训练算法;分类算法

中图分类号:TP391 **文献标识码**:A **文章编号**:1671-4512(2005)03-0053-03

A Chinese text categorization system based on the improved VSM

Li Fan Lin Aiwu Chen Guoshe

Abstract: A Chinese text categorization system was developed based on the improved vector space model, including the important aspects of system structure, text preprocessing, feature selection, training algorithm, and recognition algorithm. The system introduced the structure-layer weight coefficient to improve the term weighting, and a new training algorithm and a way of computing text similarity threshold were described. The test result illustrated the effectiveness of the system for categorizing Chinese text. The average precision was over 80 % and the recall was 86 %.

Key words: text categorization; vector space model; feature selection; structure-layer weight coefficient; training algorithm; recognition algorithm

Li Fan Prof.; College of Computer Sci. & Tech., Huazhong Univ. of Sci. & Tech., Wuhan 430074, China.

文本自动分类是指在给定的分类模型下,根据文本的内容自动确定文本类别的过程.本文主要讨论一个基于改进向量空间模型(VSM)中文文本分类系统的设计与实现,在设计中引入了结构层次权重系数来改进传统TFIDF权重,使具有明显分类特征的特征词发挥了较好的分类效果,从而抑制了权重小的特征词的干扰,同时提出了一种新的训练算法和文本相似度域值计算方法.从实验结果来看,本系统在分类效果上相对传统分类方法有显著提高.

1 文本向量空间模型

向量空间模型(VSM)[1]将文档 d_i 看作是由一组特征项($t_{i1}, t_{i2}, \cdots, t_{im}$)和相应的权重($w_{i1}, w_{i2}, \cdots, w_{im}$)构成的.包含 n 个文档的文档集用VSM表示为

$$\boldsymbol{D} = (w_{ij})_{n\times m} = \begin{pmatrix} w_{11} & \cdots & w_{1m} \\ \cdots & \ddots & \cdots \\ w_{n1} & \cdots & w_{nm} \end{pmatrix}. \quad (1)$$

本研究采用的算法基础是"简单向量距离法",此算法将文档 d_i, d_j 相似度 S 定义为文档向量之间的夹角余弦[2],

收稿日期:2004-06-04.

作者简介:李 凡(1943-),男,教授;武汉,华中科技大学计算机科学与技术学院(430074).

E-mail:lifan@hotmail.com

基金项目:国家高性能计算基金资助项目(00303).

$$S(d_i,d_j)=\cos\varphi=\sum_{k=1}^{m}w_{ik}w_{jk}\Big/\Big[\Big(\sum_{k=1}^{m}w_{ik}^2\Big)\Big(\sum_{k=1}^{m}w_{jk}^2\Big)\Big]^{1/2}. \quad (2)$$

式中:m 为文档向量维数;w_{ik}为第k维特征项权重.

2 基于改进 VSM 的文本分类系统模型

本研究所采用的分类模型由训练模块与分类模块两大模块构成,具体模型略.

2.1 文挡结构分析

特征项权重 w_{ik}是指特征项 t_k 代表文本文档 d_i 的能力大小.目前存在多种权值计算方式,本研究采用流行的 TFIDF 公式来计算权重[3].为了使权重处于区间[0,1],抑制文本由于不同长度造成的负面影响,通常对 w_{ik}做范化处理[4].

TFIDF 公式并没有考虑文本的结构特性对特征项权重的影响,事实上,同一个关键词出现在文档中的不同位置,它所能表达文档内容的能力是有差别的.本研究引入了结构层次权重系数的概念,对一个文本文档,可以按照其结构分层,依次为标题、摘要、正文、参考文献(或者链接)等,按照不同的结构域在文档的重要程度,对不同域的特征项给予不同程度的加权(见表 1).

表 1 不同域的特征项的权值

	权重系数
标题	α
摘要	β
正文	1
参考文献或链接	γ
其他	0

表 1 中,$\alpha,\beta,\gamma>1$ 且 $\alpha>\beta>\gamma$,把正文层次结构权重系数设为标准 1,标题、摘要、参考文献(或者链接)的层次结构权重依次减小,但重要性都大于正文,具体的取值要通过实验调整,其他层次对文本分类影响有限,因此一律设为 0,这样有助于减少向量空间的维数.

2.2 特征提取

步骤 1 假设训练集为 $\boldsymbol{D}$,其中的一篇文本为 T,通过结构分析,分词,过滤等处理后,最终文本 T 可表示为

$$\boldsymbol{D}(T)=((a_1,b_1,c_1,\theta_1),(a_2,b_2,c_2,\theta_2),\cdots), \quad (3)$$

式中:a_i 代表文本 T 中的词项;b_i 表示词语 a_i 在文本 T 中出现的文档频率;c_i 表示反向文档频率;θ_i 代表文本 T 的结构层次权重系数,取值参考表 1,若 a_i 同时属于多个层次结构域,规定 θ_i 取最大的那个层次结构权重系数.

步骤 2 根据如下改进的 TFIDF 公式计算文档集合 $\boldsymbol{D}$ 中每篇文档词条的权重,

$$w_{ik}^{'}=w_{ik}\theta_k\Big/\Big[\sum_{k=1}^{m}(w_{ik}\theta_k)^2\Big]^{1/2}, \quad (4)$$

式中 θ_k 取值参考式(3),其他参数含义同式(2).

步骤 3 根据每个词条的权重 $w_{ik}^{'}$进行排序,抽取排在前面一定数量的词条作为特征项,对于具体多少数目,需要在实验中不断地调整达到最优效果为止.

步骤 4 根据提取的特征重新表示训练文本,以达到空间降维的目的.

2.3 训练算法

步骤 1 对于训练集 $\boldsymbol{D}$ 中的任一文档 T_i,经过特征提取后得到文档 T_i 的向量表示

$$\boldsymbol{D}(T_i)'=((a_1,w_{i1}^{'}),(a_2,w_{i2}^{'}),\cdots,(a_m,w_{im}^{'})), \quad (5)$$

式中 $w_{ik}^{'}$为特征项a_k 的权重.

步骤 2 计算整个训练集中某类文档集 $\boldsymbol{D}$ 的类特征向量,

$$\boldsymbol{D}=\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{D}(T_k)'=\frac{1}{n}\sum_{k=1}^{n}((a_1,w_{k1}^{'}),(a_2,w_{k2}^{'}),\cdots,(a_m,w_{km}^{'}))=((a_1,w_1),(a_2,w_2),\cdots,(a_m,w_m)), \quad (6)$$

式中:n 代表类文档集 $\boldsymbol{D}$ 中的文档数量,通过对本类所有文档向量取平均值,得到本类的类别特征向量,作为新文档分类的依据.

步骤 3 假设训练集总共分为 C 个类,对这 C 个类分别通过训练算法得到 C 个类特征向量的集合,保存,即得到该训练集的类特征向量库.

2.4 分类算法

步骤 1 根据式(1),计算每一类文档集 $\boldsymbol{D}$ 中每一个文档 T_k 与类特征向量 $\boldsymbol{D}$ 的相似度

$$S(T_k,\boldsymbol{D})=\sum_{i=1}^{m}w_{ki}w_i\Big/\Big[\Big(\sum_{i=1}^{m}w_{ki}^2\Big)\Big(\sum_{i=1}^{m}w_i^2\Big)\Big]^{1/2}, \quad (7)$$

式中 $k=1,2,\cdots,n$.

步骤 2 求训练集中每一类文档集 $\boldsymbol{D}$ 的类相似度均值 $E(S(T_k,\boldsymbol{D}))$与类相似度均方差 $\boldsymbol{\delta}(S(T_k,\boldsymbol{D}))$,

$$E(S(T_k,\boldsymbol{D}))=\frac{1}{n}\sum_{k=1}^{n}S(T_k,\boldsymbol{D});$$

$$\boldsymbol{\delta}(S(T_k,\boldsymbol{D}))=$$

$$\left[\frac{1}{n}\sum_{k=1}^{n}\left(T_k-\frac{1}{n}\sum_{k=1}^{n}S(T_k,\boldsymbol{D})\right)^2\right]^{1/2}.$$

步骤3　对任一待分类新文档特征向量 $\boldsymbol{d}_i$，计算其与每个类特征向量 $\boldsymbol{D}_j$ 的相似度

$$S(d_i,D_j)\quad (j=1,2,\cdots,c);$$

$\boldsymbol{d}_i\in\boldsymbol{D}_j$ 的条件是

$$(E(S(T_k,\boldsymbol{D}_j))-\boldsymbol{\delta}(S(T_k,\boldsymbol{D}_j)))\leqslant S(\boldsymbol{d}_i,\boldsymbol{D}_j)\leqslant(E(S(T_k,\boldsymbol{D}_j))+\boldsymbol{\delta}(S(T_k,\boldsymbol{D}_j))),\quad(8)$$

式(8)给出的是文挡 $\boldsymbol{d}_i$ 的相似度域值，即以类相似度均值 $E(S(T_k,\boldsymbol{D}_j))$ 为圆心，类相似度均方差 $\boldsymbol{\delta}(S(T_k,\boldsymbol{D}_j))$ 为半径的圆. 根据式(8)将文档 $\boldsymbol{d}_i$ 分派到相应的一个或者多个类别中.

3　实验与结果分析

文本分类系统的评估指标根据文本检索的度量来定义，常用的评价标准是准确率(Precision)和查全率(Recall)[4].

"北大天网"搜索引擎实现了中文网页的自动分类功能，本系统从"北大天网"自动分类系统中下载了大量的网页作为训练集和测试集. 本次试验层次结构系数量化为：$\alpha=10,\beta=6,\gamma=5$. 实验选择计算机与因特网、政府与政治、医疗与健康、娱乐与休闲、自然科学、社会科学6个类之间进行，收集了600篇中文文本文档，其中每类训练样本50篇，兼作封闭测试样本，其他300篇作为开放待测样本集，每类50篇. 封闭测试准确率和查全率都达到90%以上，表2只给出有实际意义的开放测试结果.

表2　开放测试实验结果

类别	基于传统VSM的分类系统		基于改进VSM的分类系统	
	准确率	查全率	准确率	查全率
计算机与因特网	0.82	0.88	0.88	0.94
政府与政治	0.74	0.78	0.82	0.84
医疗与健康	0.68	0.70	0.86	0.90
娱乐与休闲	0.72	0.78	0.76	0.82
自然科学	0.70	0.80	0.72	0.80
社会科学	0.62	0.74	0.78	0.86

从表2可以看出，通过引入层次结构权重系数后，系统分类效果较基于传统VSM的分类方法有明显的提高，开放性测试的平均准确率在80%以上，且平均查全率达到了86%.

参　考　文　献

[1] Salton G, Yang C S. On the specification of term values in automatic indexing[J]. Journal of Documentation, 1973, 29(4): 351—372

[2] 武　旭，须　德. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 北方交通大学学报，2003，27(2)：38—41

[3] 朱华宇，孙正兴，张福炎. 一个基于向量空间模型的中文文本自动分类系统[J]. 计算机工程，2001，27(2)：15—17，63

[4] Fabrizio Sebastiani. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1): 11—12, 32—33

合成碳酸二苯酯新工艺获863计划立项

从国家科技部获悉，我校化学系李光兴教授申报的"碳酸二甲酯酯交换合成碳酸二苯酯的研究"获得国家863计划立项资助. 该课题是"石油化工关键过程催化新材料及其工业应用"(课题编号：2004AA32G030)的子课题.

碳酸二苯酯是重要的有机碳酸酯，可用于合成许多重要的有机化合物及高分子材料，用途广泛，是"十五"国家攻关项目"酯交换法合成聚碳酸酯"的关键原料之一. 本项目的实施将为我国建立具有自主知识产权的全非光气法合成聚碳酸酯工艺路线打下良好基础.