文章编号:1008-0570(2012)10-0264-02

基于 KNN 的话题跟踪研究

Study on Topic Tracking Based on KNN

(1.牡丹江师范学院; 2.廊坊燕京职业技术学院) 李树平[1] 夏春艳[1] 李胜东[2] 亓智斌[1] 赵 杰[1]
LI Shu-ping XIA Chun-yan LI Sheng-dong QI Zhi-bin ZHAO Jie

摘要: 话题跟踪任务的关键技术是文本分类算法,难点在于话题/报道表示模型。根据话题跟踪的定义,对比常用的文本分类算法和文本表示方法,选择 KNN 文本分类算法作为话题跟踪关键技术,利用向量空间模型设计话题/报道表示模型,结合话题检测与跟踪评测方法实现了话题跟踪系统,试验结果证明 KNN 作为话题跟踪关键技术,系统具有较稳定的话题跟踪性能。

关键词: KNN; 话题跟踪; 话题/报道表示模型; 特征选择
中图分类号: TP391.1 **文献标识码:** A

Abstract: The key technology of topic tracking task is text classification algorithm, its difficulty is topic / reports representation model. According to the definition of topic tracking, contrast to commonly used text classification algorithms and text representation methods, this paper selects KNN text classification algorithm as key technology of topic tracking, uses Topic vector space model to design topic / reports representation model, combines topic detection and tracking evaluation method to achieve the topic tracking system. Experimental results prove that the system has stable topic tracking performance when key technology of topic tracking is KNN.

Keywords: KNN; Topic tracking; Topic / reports representation model; Feature selection

技术创新

1 引言

在话题检测与跟踪研究中,话题跟踪是它的一个子任务,被定义为在给定同一个话题的几篇新闻报道的前提下检测出该话题的后继新闻报道。从定义可以看出,话题跟踪研究在本质上等价于一种受监督的分类研究, 它的关键技术就是文本分类算法,难点在于话题/报道表示模型。文本分类算法一般包括 KNN 算法,Rocchio 算法,支持向量机(SVM),简单贝叶斯算法和决策树算法,其中最常用的是 KNN 算法。它也是目前分类效果最好且应用最广泛的文本分类算法。

2 基于 KNN 的话题跟踪系统

基于 KNN 的话题跟踪系统由话题/报道表示模型,KNN 文本分类方法和话题检测与跟踪评测方法三个模块组成。

2.1 基于向量空间模型的话题/报道表示模型

话题/报道表示模型一般采用文本表示方法实现,向量空间模型由 Salton 教授于 1968 年提出的,是最简便而又高效的文本表示方法之一,在海量文本信息处理方面具有非常强的优势。因此,本文采用向量空间模型实现话题/报道表示模型。在向量空间模型实现话题/报道表示模型时,存在高维特征空间问题,这个问题可以通过特征选择算法解决。

2.1.1 特征选择

当使用向量空间模型设计话题/报道表示模型表示报道时,最大的困难是高维的特征空间。特征选择也是构造特征评分函数的过程。本文选择信息增益作为特征评分函数,实现特征选择过程。信息增益值越大,表示分布越集中,被选取的可能性也越大。

2.2KNN 算法

KNN 文本分类算法最初由 Cover 和 Hart 于 1968 年提出,是一个理论上比较成熟的方法。该算法的基本思想是: 根据话题/报道表示模型,文本内容被形式化为特征空间中的加权特征向量,即 $D=D(T_1,W_1;T_2,W_2;\cdots;T_M,W_M)$ (M 为特征空间维数)。对于一篇测试报道,计算它与训练报道集中每个报道的相似度,找出 K 个最相似的报道,根据加权距离和,判断测试报道所属的话题类别。

定义 1: 测试报道向量 d_i 与每一个训练报道向量 d_j 的相似度 $Sim(d_i,d_j)$,定义为:

$$Sim(d_i,d_j)=\frac{\sum_{k=1}^{M} w_{ik}\times w_{jk}}{\sqrt{\sum_{k=1}^{M} w_{ik}^2}\times\sqrt{\sum_{k=1}^{M} w_{jk}^2}} \quad (1)$$

定义 2: 测试集中的报道向量 x 在 K 最近邻报道中的权重,定义为:

$$P(x,C_j)=\begin{cases}1 & \sum_{d_i\in kNNDoc} Sim(x,d_i)\,y(d_i,C_j)-b\geq 0\\ 0 & otherwise\end{cases} \quad (2)$$

在公式(2)中,x 为测试文本的特征向量; C_j 为第 j 类话题; $Sim(x,d_i)$ 为相似度计算公式;b 为阈值,有待于优化选择;而 $y(d_i,C_j)$ 的取值为 1 或 0,如果 d_i 属于 C_j,则函数值为 1,否则为 0。

3 试验结果与分析

在实验中, 所采用的分词程序是中科院计算所软件室提供的 ICTCLAS; 语料是中科院计算所谭松波博士提供 14150 篇中文新闻报道文本文档,共分两个层次,第一个层次是 12 个主题,第二个层次是 60 个话题。

李树平: 副主任 教授

基金项目: 基金颁发部门: 黑龙江省教育厅; 项目名称: 话题跟踪关键技术研究 (2011.1-2012.12); 基金项目编号:(12511580); 基金申请人: 李树平.
牡丹江师范学院科研项目(NO.KY201001)牡丹江科技局科研项目(NO.Z2011n061)的资助

3.1 试验结果

在实验中,话题跟踪关键技术为 KNN 算法,特征空间维数为 1000。通过调整 K 最近邻值,得到不同 K 最近邻值条件下的话题检测与跟踪(TDT)评测结果。本次试验分别使用 60 个话题进行测试,共测试了 60 次,然后按照每个主题进行评测,共评测了 12 次,最终得到了 12 个 TDT 评测结果 (归一化检测开销 $(CDet)_{Norm}$),根据实验的评测结果评估这个算法作为话题跟踪关键技术的性能,如表 1 所示。

表 1 在实验 1 中的 TDT 评测结果

话题类别	K=10	K=20	K=30	K=40	K=50	K=60
人才	0.2100	0.2181	0.2097	0.2198	0.2174	0.2104
体育	0.0613	0.0694	0.0707	0.0702	0.0726	0.0770
卫生	0.2971	0.2894	0.2821	0.2762	0.2643	0.2549
地域	0.6526	0.6400	0.6787	0.6920	0.7122	0.7792
环境	0.2486	0.2554	0.2521	0.2547	0.2623	0.2650
房产	0.0904	0.1016	0.1086	0.1050	0.1029	0.1094
教育	0.3008	0.2946	0.2925	0.2922	0.2888	0.2773
汽车	0.0961	0.1196	0.1122	0.1080	0.1030	0.0974
电脑	0.2026	0.2181	0.2250	0.2205	0.2252	0.2279
科技	0.5967	0.5877	0.5921	0.5829	0.5765	0.5885
艺术	0.7009	0.7027	0.7192	0.7292	0.7309	0.7414
金融	0.3188	0.3307	0.3475	0.3400	0.3513	0.3493
平均 $(C_{Det})_{Norm}$	0.3147	0.3189	0.3242	0.3242	0.3256	0.3315

为了便于分析 KNN 算法中 K 最近邻值对话题跟踪性能的影响,用 Excel 2003 中的图表向导工具把表 1 中的数据映射成图 1。

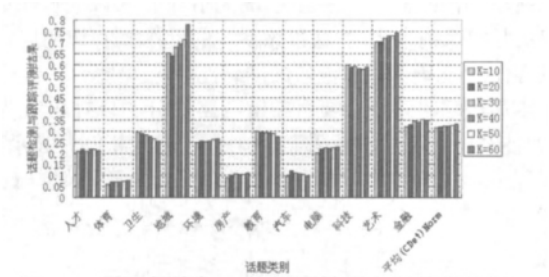

图 1 K 最近邻值与 TDT 评测结果之间的趋势图

在实验中,根据 TDT 评测结果评测话题跟踪性能,然后根据话题跟踪性能评估 KNN 算法作为话题跟踪关键技术的性能。TDT 评测结果越小,说明话题跟踪性能越好,也表明 KNN 作为话题跟踪关键技术的性能越好。根据表 1 和图 1,当 K 最近邻值为 10 时,平均的 TDT 评测结果最小,其值为 0.3147。这说明了 KNN 算法作为话题跟踪关键技术在 K 最近邻值为 10 时有最好的性能。

3.2 试验分析

根据表 1 和图 1,K 最近邻值不同,导致平均的 TDT 评测结果不同。当 K 最近邻值从 10 增加到 60 时,平均的 TDT 评测结果从 0.3147 增加到 0.3315。此时,随着 K 最近邻值增加,从训练集中选择的最相似的报道向量也增加,这时也增加了一些虽然与测试集中某报道向量的相似度比较高但与该报道不是同一话题的训练集报道向量的风险,这种风险导致了 K 最近邻向量中干扰数据比较多,最终的结果不但不能增加话题跟踪性能,还有可能使话题跟踪性能降低。例如,当 K 最近邻值为 10 时,K 最近邻向量中包含了几乎所有与测试报道向量最相似的训练集向量,此时的话题跟踪性能较好,其平均评测结果为 0.3147,随着 K 近邻值增加,干扰数据越来越大,话题跟踪性能也越来越不理想,当 K 最近邻值增加到 60 时,平均评测结果为 0.3315,这个值比 K 近邻值为 10 时增加了 5.338%,也就是说,KNN 作为话题跟踪关键技术具有较稳定的系统性能。

4 结论

话题跟踪任务是话题检测与跟踪研究中的一个子任务,其关键技术是文本分类算法,其难点在于话题/报道表示模型。通过对比,本文选择 KNN 文本分类算法和向量空间模型设计并实现了话题跟踪系统,试验结果证明了 KNN 作为话题跟踪关键技术具有较稳定的话题跟踪性能。

本文无抄袭,作者全权负责版权事宜。

参考文献

[1] 郑伟, 张宇,邹博伟等. 基于相关性模型的中文话题跟踪研究 [C]. 第九届全国计算语言学学术会议. 大连: 大连理工大学, 2007. 558~563.

[2] 丁伟莉. 中文 Blog 热门话题检测与跟踪技术研究 [D]. 哈尔滨:哈尔滨工业大学,2007.

[3] 苏力华. 基于向量空间模型的文本分类技术研究 [D]. 西安:西安电子科技大学,2006.

[4] 李慧,李存华,王霞. 文本分类中基于差值思想的多特征选择算法研究 [J]. 微计算机应用,2009,30(10):1~5.

[5] 刘科. 基于 KNN 算法的文本分类 [J]. 科技经济市场,(06), 2009:12~13.

[6] 胡佳妮,徐蔚然,郭军等. 中文文本分类中的特征选择算法研究 [J]. 光通讯研究,2005,(3):44~46.

[7] 杨丽华, 戴齐, 郭艳军. KNN 文本分类算法研究?[J]. 微计算机信息,2006,7–3:269~270.

[8] 中科院计算所. 基于多层隐马模型的汉语词法分析系统 ICTCLAS. http://www.nlp.org.cn/project/project.php?proj_id=6.

[9] 谭松波, 王月粉. 中文文本分类语料库–TanCorpV1.0. http://www.searchforum.org.cn/tansongbo/corpus.htm.

[10] Tan, S.B., et al. A Novel Refinement Approach for Text Categorization[C].ACM CIKM2005, 2005.

作者简介:李树平(1964–),女(汉族),安徽临泉人,牡丹江师范学院计算机科学与技术系副主任,教授,主要从事数据挖掘研究。

Biography: LI Shu–ping (1964–), Female (Han),Linquan in Anhui province,Working in Mudanjiang Normal University, Deputy Director in Department of Computer Science and Technology , Professor, her major field of study mainly engaged in data mining.

(157012 黑龙江省 牡丹江市 牡丹江师范学院 计算机科学与技术系)李树平　夏春艳　赵　杰

(065200 河北省 三河市 廊坊燕京职业技术学院 计算机工程系)李胜东

(157012 黑龙江省 牡丹江市 牡丹江师范学院 人事处)亓智斌

通讯地址:(157012 黑龙江省 牡丹江市 牡丹江师范学院 计算机科学与技术系) 李树平

(收稿日期:2011.10.28)(修稿日期:2012.01.28)