基于K－近邻算法的网页自动分类系统的研究及实现

张高胤,谭成翔,汪海航
(同济大学,上海 201804)

摘 要:随着网络信息量的爆炸式增长,人们查找信息越来越难。Web搜索引擎的出现在一定程度上解决了这种矛盾。然而现行的搜索引擎无法根据用户所指定的主题进行针对性的搜索,因此,必须在搜索后对结果是否属于目标主题进行判断,以提高搜索的准确性,文中提出了一种基于K－近邻机器学习算法的信息自动分类的方法,能够对搜索到的网页自动地判定是否属于目标主题,并在实验的基础上验证了其在提高搜索准确性上的作用。

关键词:K－近邻算法;机器学习;网页分类
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673－629X(2007)01－0021－03

Design and Implementation of Web Page Automation Classification System Based on K－Nearest Neighbor Algorithm

ZHANG Gao-yin, TAN Cheng-xiang, WANG Hai-hang
(Tongji University, Shanghai 201804, China)

Abstract: The capacity of information in the network is growing like explosion, and it is much harder for people to find information that they want to precisely and quickly. The appearance of Web search engine resolves that problem in some level. However the popular search engine nowadays can not search in a special category. To improve the accuracy of search, decisions for whether a Web page belongs to the target category needs to be made automatically. Introduce a Web page automation classification method based on one of machine learning algorithm named K－nearest neighbors(KNN) algorithm, which can be used to decide if a Web page is relative to a special category. This page also proved the improvement of this method by experiments.

Key words: K－nearest neighbors algorithm; machine learning; Web pages automation classification

0 引 言

目前对于网页的搜索一般是基于关键词进行的,由于语言中的一词多义和多词一义现象的普遍存在。对于网页的搜索的效果差强人意,特别是基于某一主题领域的搜索,搜索结果往往会存在许多与该主题无关的结果,使得用户浪费大量时间浏览自己并不感兴趣的内容,目前,所有主流的搜索引擎普遍存在着搜索的准确率低、召回率低、冗余度高的问题,针对这一现象,现行的解决方法一般有两种:

1)开发基于针对某一特定主题的搜索工具;利用已知的知识使得搜索限定在某一特定主题的范围内。

2)采用机器学习的方法通过对一部分已经经过人工分类的网页进行学习,总结出其中的规律,提取能够尽可能准确分类网页的目标函数,来对未知网页进行分类。由于开发针对特定主题的搜索引擎成本较高,需要充分了解与该主题相关的领域知识,对于主题领域的发展和变化也无法进行自动的适应,因此,应用前景很有限,而基于机器学习的方法,却具有前者无法比拟的通用性和自适应性。具有更为广泛的应用前景[1]。文中所提出的就是一种基于机器学习方法的网页自动分类的方法。

1 K－近邻算法简述

目前,机器学习的算法有很多种[2],文中采用K－近邻算法。下面对这一算法作简单描述。K－近邻算法假定所有的实例对应于N维空间 R^n 中的点。一个实例的最近邻是根据标准欧氏距离定义的。更精确地讲,把任意的实例表示为下面的特征向量:$(a_1(x), a_2(x), a_3(x), \cdots, a_n(x))$,其中 $a_r(x)$ 表示实例 x 的第 r 个属性值。那么两个实例 x_i, x_j 间的距离定义为 $D(x_i, x_j)$,其中:

$$D(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

目标函数为离散值的K－近邻算法的基本思想为:训练时对于每个训练样例 $(x, f(x))$,把这个样例加入列表training_ examples中,在对具体的实例 x_q 进行分类时,从training_ examples中选出最靠近 x_q 的 K 个实例,并用 x_1,

收稿日期:2006－04－20

作者简介:张高胤(1981－),男,浙江嘉善人,硕士研究生,研究方向为计算机网络安全与电子商务;谭成翔,教授,研究员,博士,主要研究方向为计算机网络安全与电子商务;汪海航,教授,博士生导师,主要研究方向为电子商务、网络与信息安全。

$x_2,\cdots,x_k$ 表示，返回 $f'(x_q)\leftarrow\arg\max_{v\in V}\delta(v,f(x_i))$，其中，如果 $a=b$ 那么 $\delta(a,b)=1$，否则 $\delta(a,b)=0$。即，在训练集中选取在欧氏空间中距离 x_q 最近的 K 个实例，算法所返回的 $f'(x_q)$ 为对目标函数 $f(x_q)$ 的近似值，它就是最靠近 x_q 的 k 个训练样例中最普遍的 $f(x)$ 值。

2 网页自动分类基本步骤

对于一个基于机器学习(Machine Learning)方法的文本自动分类系统，其主要组成部分有：类别特征词抽取模型（key words fetch Model），训练和学习模型(Training and Learning Model)、分类处理模型(Categorization Processing Model)以及结果评估模型(Performance Evaluation Model)[3]。训练和学习模型只需在分类前使用一次即可；结果评估模型在训练学习模型中需要使用到，并且是对最终分类效果进行衡量的方法；而在文本自动分类的一个流程中，需要使用的就是分类处理模型。

2.1 类别特征词抽取模型

对于网页归类时所用的比较属性一般都采用关键词，因此确定比较哪些关键词，即确定比较属性是对网页进行正确分类的基础。类别特征词抽取模型的主要任务就是通过已对某一特定主题进行人工分类过的网页进行学习，获取该类别的特征关键词及其相对应的权值。获取机器学习所必须的比较属性。其基本步骤为：首先对归档网页进行取词，并进行文本预处理，从中去掉一些无法应用于分类的词汇、文档和一些文本格式标记、停用词等。然后对属于目标主题的所有网页和取词取到的所有词语计算其 w_{ik} 组成网页词汇矩阵，针对特定词语 t_k 计算：$w_k=\sum_i w_{ik}$。按 w_k 从大到小取出 N 个词，作为该类别的特征词，并把 w_k 作为该词的权值，用于 K－近邻算法中。

词语权重的计算：对于词权重的计算，经典的方法考虑两个因素[4]：

1) 词语频率 tf(Term Frequency)：词语在文档中出现的次数。

2) 词语倒排文档频率 idf(Inverse Document Frequency)：该词语在文档集合中分布情况的一种量化，常用的计算方法是 $\log_2(N/n_k+0.01)$，其中 N 为文档集合中的文档数目，n_k 为出现该词语的文章数。

根据以上两个因素，可以得到公式：

$$w_{ik}=tf_{ik}\times\log_2(N/n_k+0.01)$$

其中 tf_{ik} 为词语 t_k 在文档 D_i 中出现的次数，w_{ik} 为词语 t_k 在文档 D_i 中的权值，$k=1,2,\cdots m$（m 为词的个数）。现有的方法为根据词语的权值进行排序[5]，选取权值最高的几个词作为该类文档的关键词，由于 idf 值的计算是考虑选取在其他文档中出现次数较少的词语作为标识这篇文档的特征词汇，如果在学习用例中有较多的正例文档，而真正的标识该类文档的特征词在学习集的正例中也应有较高的出现率，如果统一采用 idf 计算，由于 n_k 较高，则会得到较低的 w_i 值，因此可能会忽略了真正的核心特征词，考虑到一类文档的特征词应为在属于这一主题的文档中有较高的出现率，而在非该主题的文档中较少出现，基于这一想法，文中提出了一种计算关键词权值的改进算法：若

$$\log_2(N_{反}/n_k+0.01)<idf_{\min}$$

则表明该词汇在反例中有较高的出现率，说明该词可能是该目标主题的特征词汇，则忽略该词。其中 $idf_{\min}$ 表示一个阈值。

若不满足上述条件则计算该词的权值：

$$w_{ik}=tf_{ik}\times\frac{\log_2((N_{反}+1)/n_{k反})+0.01)}{\log_2((N_{正}+1)/n_{k正})+0.01)}$$

其中 $N_{反}$ 表示训练集中反例总数，$N_{正}$ 表示训练集中正例总数，$n_{k正}$ 表示所有出现词 t_k 的正例文档总数，$n_{k反}$ 表示所有出现词 t_k 的反例文档总数，该公式的提出是考虑到用于表征特定主题的词应在正例文档中有较高的出现率，而在反例文档中较少出现。

类别学习的过程是一个动态的过程，系统在运行时，能够定期地对用户选取的新网页进行学习，不断地调整目标主题的特征关键词及其权值，以自动适应该主题领域的发展和变化。

2.2 训练和学习过程

训练和学习过程是指对现有的已经人工分类的文档，按照由特征抽取模型获取的特征词，转换成相对应的相量值。即对任一网页 D_i，计算每一特征词的权重，构成文档的特征相量 $(a_1(d_i),a_2(d_i),\cdots,a_k(d_i))$，其中 $a_j(d_i)$ 为特征词 t_j 在文档 d_i 中的权值，其值根据 t_j 在网页中出现的位置和频率综合而定。最后将该相量归一化，将相量值和判断结果（是否属于目标主题）一起进行保存。作为训练结果[6]。

2.3 对新文档的自动分类

对新文档的自动分类是指运用计算机自动对搜索引擎抓取的网页进行判别，确定其是否属于用户期待的目标主题。

主要步骤为：

在搜索引擎抓取的新网页中，对目标主题的每一个特征词，分别计算权值，得到该网页的特征向量，在训练集中选取与该特征向量距离最近的 K 个训练样例，根据训练样例值的分布，确定该网页是否属于目标主题。在应用 K－近邻算法时，经常会碰到一个称之为“维度灾难”的实践问题，由于在分类网页时，某些属性的作用要高于其余的属性，但一般的 K－近邻算法在计算实例间距离时等价考虑所有的属性，这样就必然会造成近邻间的距离会被大量的不相关属性所支配。针对这一问题，文中所采用的解决方法为，在计算两个实例间距离时，对每个属性加权，权值即在特征词抽取时获得相对于特定属性的权值。即：

$$d'=\sqrt{\sum_{r=1}^{n}(w_r(a_r(x_i)-a_r(x_j)))^2}$$

通过对特定属性距离的加权，其效果等同于伸展坐标轴，

以突出重要属性的作用,消减不重要属性的影响,对 K－近邻算法进行了优化,提高了准确率。

3　实验结果及分析

同济大学信息安全实验室选取了安全领域和电子商务这两个特定主题作为实验的目标主题,在同样的实验环境下选取了不同 K 值和 N 值对文中所提出的网页归档算法进行了实验,训练集通过选取网络上 2000 多页网页先进行人工分类,再供系统进行自动学习,记录分类结果,并人工检验实验结果,判断其准确性,具体的实验结果见表 1。

表 1　网页自动分类系统实验结果

目标主题	网络安全	电子商务
K	20	20
N	10	15
训练集大小	1500	2000
测试集大小	1000	1000
准确率	86.54%	80.34%
目标主题	网络安全	电子商务
K	40	40
N	20	30
训练集大小	1500	2000
测试集大小	1000	1500
准确率	89.33%	85.32%

通过比较可以发现,应用该算法时,采用较高的 K 值和 N 值,将会获得更高的准确率,这是由于 K－近邻算法的准确率与选取的属性数量,和比较时选取的邻结点的数量成正比,与实验结果相符。通过实验结果,可以发现,该算法在确定主题的搜索中能够获得较高的准确率。具有较高的实用价值。

4　结束语

文中描述了一种基于机器学习的文档自动分类系统,该系统能够在学习现有的已分档的网页的基础上,总结目标主题的特征,同时把它应用于新网页的自动分类过程中,该算法对现有的文档分类方法进行了一定的改进,并通过实验证明,该算法能够显著地提高针对于特定主题的网页搜索的准确率,为用户提供更好的搜索性能。同时,基于机器学习的方法还能通过不断学习来主动适应特定主题领域的发展,具有很强的适应性。

参考文献:

[1] 陈立孚,周　宁.基于机器学习的自动文本分类模型研究[J].现代图书情报技术,2005(10):80－85.

[2] Dong Yan Shi, Han Ke Song. A Comparison of Several Ensemble Methods for Text Categorization[C]// Proceedings of the 2004 IEEE International Conference on Service Computing. [s.l.]: IEEE Computer Society, 2004.

[3] 何　清,史忠植.机器学习与概念语义空间生成,中文信息处理若干重要问题[M].北京:科学出版社,2003:266－277.

[4] Sebastiani F. Machine Learning in Automated Text Categorization[J]. ACM Computing Surveys, 2002,34(1):1－47.

[5] 赵国涛,柯软铭.基于本体的异构文本分类系统[J].计算机工程,2004(30):21－25.

[6] Hotho A, Madche A, Staab S. Text Clustering Based on Good Aggregations[J]. Kiinstliche Intelligenz, 2002(2):4－9.

(上接第 20 页)

参考文献:

[1] Kalyanpur A, Hendler J, Parsial B. SMORE－Semantic Markup, Ontology and RDF Editor[EB/OL]. 2004－02－28. http://mindswap.org/papers/SMORE.pdf.

[2] Ciravegna F, Dingli A. User－System Cooperation in Document Annotation based on Information Extraction[C]// In Proc. of the 13th International Conference on Kowledge Engineering and Knowledge Management (EKAW02). Siguenza, Spain:[s.n.], 2002.

[3] Handschuh S, Staab S. Authoring and Annotation of Web Pages in CREAM[C]// In Proc. of WWW2002. Honolulu, Hawaii, USA:[s.n.], 2002.

[4] State of the art on Semantic Web languages. IST Project IST－2001－34373 Esperonto Services D2.1[EB/OL]. 2003. http://www.esperonto.net/semanticportal/jsp/frames3.jsp/.

[5] Handschuh S, Staab S, Volz R. On deep annotation[C]// In Proc. of WWW2003. Budapest, Hungary:[s.n.], 2003.

[6] Handschuh S, Staab S, Volz R, et al. Deep Annotation for Information Integration[C]// In Proc. of IJCAI－03 Workshop on Information Integration on the Web (IIWeb－03). Acapulco, Mexico:[s.n.], 2003.

[7] Volz R, Handschuh S, Staab S, et al. Unveiling the hidden bride: deep annotation for mapping and migrating legacy data to the Semantic Web[J]. J. Web Sem, 2004(2):187－206.

[8] 廖述梅,徐升华,陶　皖. 带模板的结构化 HTML 文档深度标注框架研究[C]//中国系统工程学会.信息系统协会中国分会第一界学术年会论文集(B 集).北京:清华大学出版社,2005:111－115.

[9] Mukherjee S, Yang G, Ramakrishnan I V. Automatic Annotation of Content－Rich HTML Document: Structural and Semantic Analysis[C]// In Second International Semantic Web Conference (ISWC2003). Sanibel Island, Florida, USA:[s.n.], 2003.

[10] Hyvonen E, Salminen M, Junnila M. Annotation of Heterogeneous Database Content for Semantic Web[EB/OL]. 2006－02－11. http://www.seco.tkk.fi/publications/2004/hyvonen－salminen－et－al－annotation－of－heterogeneous－2004.pdf.